UNIVERSIDADE FEDERAL DE MINAS GERAIS

Luciano B. Domingos Neves

Desenvolvimento de um Método para a Quantificação da Associação Instantânea Multivariável

Belo Horizonte

Abril/2018

Luciano B. Domingos Neves

Desenvolvimento de um Método para a Quantificação da Associação Instantânea Multivariável

Dissertação submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais como requisito parcial à obtenção do título de mestre em Engenharia Elétrica. Linha de Pesquisa: Inteligência Computacional

Orientador: Adriano Vilela Barbosa

Belo Horizonte

Abril/2018

Resumo

Este trabalho apresenta um método para quantificar a associação instantânea entre grupos de variáveis. Tal medida pode ser estabelecida de diferentes maneiras, de acordo com a aplicação desejada. Neste estudo, a associação foi definida de três formas. A primeira delas (v) descreve a variância compartilhada entre os grupos. Para quantificar o impacto de cada variável separadamente na associação, estas foram transformadas em componentes ortogonais por meio de Análise em Componentes Principais (PCA - *Principal Component Analysis*) ou de Análise em Componentes Canônicas (CCA - *Canonical Component Analysis*). A CCA possibilita ainda o cálculo de outras duas medidas de associação: uma que estima probabilidade dos grupos estarem descorrelacionados (h) e a outra que avalia a máxima correlação entre os grupos (c). O comportamento variante no tempo foi capturado através de um filtro média móvel com fator de esquecimento exponencial utilizado para se estimar matrizes de covariância instantâneas a partir das variáveis. O método desenvolvido foi aplicado a três bases de dados: as duas primeiras consistem de dados adquiridos a partir de experimentos de produção de fala humana enquanto a terceira apresenta séries temporais de preços de ações. O método foi capaz de detectar variações nos valores da associação ao longo do tempo, descrever o impacto de cada variável na relação entre os domínios e detectar atrasos entre os grupos.

Abstract

This work presents a method for quantifying the instantaneous association between two groups of variables. This association can be assessed in different ways and it is usually dependent on each application's specific goals. In this work, three measures are defined. The first one (v) captures the shared variance between the groups of variables by mapping the linear relationship between them. In order to establish each group's total variance, Principal Component Analysis (PCA) and Canonical Component Analysis (CCA) are used to remove redundant information by diagonalizing the covariance matrix. The use of CCA provides two additional definitions of association: one that estimates the probability of the two groups being independent (h) and another one where the association is defined as the maximum correlation found between the groups (c). Time-varying fluctuations are captured by using an exponential moving average filter to estimate the covariance between variables. The proposed method was tested on three databases; two collected during speech production experiments and one consisting of time series of stock prices. The method was able to detect how the association changes over time, to establish the impact of each variable over the global association measure, and to detect delays between the domains.

A mamãe e Nani, por todo o suporte para me fazer chegar até aqui, e Débora, pelo carinho, apoio, paciência e companheirismo.

Agradecimentos

Primeiramente, gostaria de agradecer ao meu orientador, Prof. Adriano Vilela Barbosa, pelo suporte e pela paciência ao longo do trabalho. Gostaria também de agradecer ao Prof. Hani Camille Yehia por me auxiliar no aprendizado sobre processamento audiovisual da fala (e dar algumas broncas necessárias), ao Prof. Adriano César Machado Pereira, do Departamento de Ciência da Computação, pelas conversas e pela assistência na matéria de mercado financeiro e, por fim, ao Prof. Eduardo Mazoni Mendes por tirar minhas dúvidas de estatística, por mais simples que fossem. Da mesma forma, gostaria de agradecer aos colegas do CEFALA e aos que se aventuravam no mercado financeiro comigo (Rafael e Paulo) por fazer esta jornada ser mais prazerosa.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Lista de Símbolos

$\boldsymbol{T_1^{EVB}}$	Matriz que contém os valores que representam o movimento do trato vocal do locutor EVB (base 1).
$\boldsymbol{T_1^{CTB}}$	Matriz que contém os valores que representam o movimento do trato vocal do locutor CTB (base 1).
$\boldsymbol{O_2}$	Matriz que contém os valores que representam o movimento da face (base 2).
$\boldsymbol{T_2}$	Matriz que contém os valores que representam o movimento do trato vocal (base 2).
$\boldsymbol{A_2}$	Matriz que contém os coeficientes LSP (base 2).
$\boldsymbol{P_2}$	Matriz que contém os valores da amplitude do sinal de voz (base 2).
$\boldsymbol{PM}$	Matriz que contém os valores dos preços das ações VALE3 e VALE5 coletados a cada minuto (base 3).
$\boldsymbol{PD}$	Matriz que contém os valores dos preços das ações VALE3 e VALE5 coletados a cada dia (base 3).
$\boldsymbol{MPD}$	Matriz que contém os valores dos preços de diversas ações coletados diariamente (base 3).
ρ_{xy}	Correlação real entre as variáveis x e y.
σ_{xy}	Covariância real entre as variáveis x e y.
σ_{xx}	Desvio padrão real da variável x.
σ_{yy}	Desvio padrão real da variável y.
r_{xy}	Correlação estimada entre as variáveis x e y.
t	Valor do teste de *T-Student*.
N	Número de amostras existentes na base.
$\boldsymbol{p}_i$	i-ésima componente principal.
$\boldsymbol{d}_i^{pca}$	i-ésimo vetor base utilizado para a construção da i-ésima componente principal.
$\boldsymbol{D}$	Matriz que contém os autovetores da PCA.
p_i^r	Razão da variância da i-ésima componente principal pela variância total.
$\boldsymbol{\Lambda}$	Matriz diagonal que contém os autovalores da matriz de covariância.
$\boldsymbol{\lambda}_i$	Autovalor da matriz de covariância de número i.

$\boldsymbol{a}_i$	Vetor base utilizado no cálculo da componente canônica u_i.
$\boldsymbol{b}_i$	Vetor base utilizado no cálculo da componente canônica v_i.
$\boldsymbol{u}_i^x$	i-ésima componente canônica calculada a partir do grupo $\boldsymbol{x}$.
$\boldsymbol{u}_i^y$	i-ésima componente canônica calculada a partir do grupo $\boldsymbol{y}$.
s	Número de componentes canônicas.
n_x	Número de variáveis do grupo $\boldsymbol{x}$.
n_y	Número de variáveis do grupo $\boldsymbol{y}$.
$L_{i,k}$	Correlação entre a i-ésima variável e a k-ésima componente canônica.
$v_{x\|y}^{cca}$	Razão da variância do grupo $\boldsymbol{x}$ que é explicada a partir do $\boldsymbol{y}$ a partir da CCA.
$v_{y\|x}^{cca}$	Razão da variância do grupo $\boldsymbol{y}$ que é explicada a partir do $\boldsymbol{x}$ a partir da CCA.
$v_{x\|y}^{pca}$	Razão da variância do grupo $\boldsymbol{x}$ que é explicada a partir do $\boldsymbol{y}$ a partir da PCA.
$v_{y\|x}^{pca}$	Razão da variância do grupo $\boldsymbol{y}$ que é explicada a partir do $\boldsymbol{x}$ a partir da PCA.
h	Medida de associação baseada na probabilidade dos grupos estarem descorrelacionados.
c	Medida de associação baseada na máxima correlação entre os grupos.
η	Parâmetro de ajuste do fator de decaimento para o filtro média móvel exponencial.
c_{filt}	Fator de normalização para o filtro média móvel exponencial.
$cov(k)$	Valor da covariância instantânea estimada.
$\boldsymbol{C_{xx}}(k)$	Matriz de covariância das variáveis dentro do conjunto $\boldsymbol{x}$ no instante k.
$\boldsymbol{C_{yy}}(k)$	Matriz de covariância das variáveis dentro do conjunto $\boldsymbol{y}$ no instante k.
$\boldsymbol{C_{xy}}(k)$	Matriz de covariância cruzada entre as variáveis do conjunto $\boldsymbol{x}$ e do conjunto $\boldsymbol{y}$ no instante k.
P	Matriz intermediária utilizada para o cálculo da matriz de covariância instantânea.
Q	Matriz intermediária utilizada para o cálculo da matriz de covariância instantânea.
d	Atraso considerado entre os grupos.
d_{max}	Valor máximo de atraso considerado entre os grupos.
h_{filt}	Resposta ao impulso do filtro média móvel exponencial.
$s_f(t)$	Sinal de voz.
$u_f(t)$	Sinal de excitação do sistema de produção de fala humana.
$v_f(t)$	Resposta ao impulso do filtro no sistema de produção de fala humana.
$W(t)$	Janela utilizada no cálculo da *short-time fourier transform*.

$\boldsymbol{\alpha}$	Vetor que contem os coeficientes LPC.
$F_1(z), F_2(z)$	Polinômios utilizados para se determinar os coeficientes LSP.
A_p	Polinômio determinado pelo LPC.
$r^{\%}(k)$	Retorno de um ativo dado um instante de tempo k tendo como base variações percentuais.
$r^{log}(k)$	Log-retornos de um ativo para um determinado instante k.
$s_p(k)$	Valor do *spread* de um portfólio em um determinado instante k.
β	Valor do peso atribuído a um determinado ativo na construção de um portfólio.
δ	Limite de perdas aceitas pelos métodos *VaR* e *C-VaR*.

Capítulo 1

Introdução

1.1 Motivação e Justificativa

A correlação é, talvez, o método mais conhecido e utilizado para descrever o grau da associação entre duas variáveis. As aplicações na área da ciência são diversas, onde pode-se destacar a extração da relação linear entre grandezas físicas e a detecção de informação redundante comparando séries temporais com suas versões defasadas (autocorrelação). Os tipos de correlação mais conhecidos são a correlação de *Spearman* e a correlação de *Pearson* [3]. A primeira é uma medida ordinal de relação entre as variáveis considerada, o que apresenta uma certa limitação em seu uso, pois além de necessitar de um algoritmo de ordenação das amostras, ainda pode apresentar um erro de quantização. Desta maneira, a correlação de *Pearson* costuma ser mais utilizada. Outro motivo para esta escolha é que ela consegue descrever a relação linear entre variáveis, o que cria uma ponte para ser utilizada com a teoria de algebra linear. Tal conexão não é possível com a correlação de *Spearman*, uma vez que, esta é capaz de capturar relações não lineares.

Todavia, em algumas aplicações (ou talvez na maior parte delas) os sistemas são multivariados e é necessária uma análise entre grupos de variáveis. Existem algumas formas de estabelecer o grau de associação entre dois domínios. A mais intuitiva delas pode ser feita determinando a

relação entre todos os pares possíveis de variáveis, que é desaconselhável por dois motivos. Primeiramente, o número total de combinações pode ser muito elevado dependendo do número de variáveis que cada grupo possuir. O segundo motivo está relacionado com o fato de que caso se deseje estimar se os grupos são independentes (ou descorrelacionados) entre si com base em um teste de hipótese, a análise par a par pode levar a um resultado falso [4].

Neste trabalho serão utilizadas três medidas de associação entre domínios. A primeira delas define o grau de associação como sendo o valor da variância compartilhada entre os grupos, todavia no caso multivariado é necessário que a variância total de cada um dos conjuntos seja quantificada. Uma abordagem natural seria calcular o traço da matriz de covariância de cada grupo, entretanto, se existir correlação entre as variáveis, esta operação pode descrever um montante maior que o real. Para evitar este erro, foi realizada uma diagonalização prévia das matrizes de covariância de cada grupo com o auxílio da Análise em Componentes Principais (PCA - *Principal Component Analysis*) [5] e da Análise em Componentes Canônicas (CCA - *Canonical Correlation Analysis*)[6, 7]. A probabilidade de dois conjuntos de variáveis estarem descorrelacionados também pode ser utilizada como uma medida de associação entre grupos pois, quanto menor a probabilidade dos grupos estarem descorrelacionados, maior a associação entre eles [5]. A CCA ainda abre espaço para se encontrar uma terceira forma de definir associação: encontrar a máxima correlação entre os grupos [5].

Quando a associação entre dois grupos de variáveis é reavaliada ao longo do tempo, ela pode ser utilizada para quantificar sincronismo entre domínios, sendo um ganho de sincronismo (ou coordenação) representado por um aumento no valor da associação [2]. Tal análise pode ser feita para casos onde os grupos de sinais estão em fase ou não. Caso a primeira hipótese seja verdadeira, um gráfico de como a associação varia ao longo do tempo é suficiente para representar a coordenação entre os domínios. Se existir atraso entre os grupos, deve-se calcular além das associações com atraso nulo, as associações entre os grupos defasados.

Para simplificar a visualização dos dados, ao invés de criar uma série de gráficos pode-se gerar um mapa de calor onde o eixo da abscissa representa o instante de tempo, o eixo das ordenadas o valor do atraso e a cor o valor da associação [2]. Tal figura também é capaz de ilustrar flutuações no valor do atraso e será referenciada neste trabalho como mapa de associação.

As utilizações possíveis para o método são diversas. Na área de processamento audiovisual, por exemplo, uma aplicação surge quando deseja-se analisar as relações entre a acústica da fala, o movimento do trato vocal e o movimento da face, onde cada domínio é multivariado [8]. Outros usos possíveis são encontrados nas ciências do comportamento (Behavioral Sciences), como estimar o nível de coordenação de uma pessoa interagindo com um sinal de referência (como o nível de coordenação existente entre instrumentistas e metrônomos e a coordenação de pessoas dançando com a batida de uma música [9]) bem como o estudo da coordenação entre os movimentos do trato vocal e da cabeça de dois locutores enquanto eles interagem [1]. Na área de macroeconomia, o método desenvolvido pode ser utilizado para quantificar coordenação entre ciclos econômicos presenciados em diferentes países, caracterizando cada um destes por um conjunto de indicadores macroeconômicos (e.g. Produto interno bruto, taxa de inflação e taxa de desemprego) [10]. Em finanças quantitativas, o método pode ser utilizado em diversas aplicações em arbitragem (e.g. entre preços e fatores, entre dois grupos de ativos e entre indicadores técnicos e ativos) [11, 12, 13].

1.2 Objetivo

Neste trabalho será proposto a criação de um método que quantifica a associação instantânea entre domínios multivariados. Tal objetivo será alcançado por meio do desenvolvimento matemático, pelo qual serão estabelecidas três medidas de associação instantânea: uma que quantifica a variância compartilhada entre os grupos $v(k)$, uma que estima a probabi-

lidade dos grupos serem independentes $h(k)$ e uma terceira que busca encontrar uma máxima correlação entre os grupos $c(k)$. Os métodos serão desenvolvidos computacionalmente e testados em três aplicações distribuídas nas áreas de processamento audiovisual da fala e finanças quantitativas.

Inicialmente, deseja-se estudar a coordenação entre os movimentos do trato vocal de dois locutores durante o diálogo [1]. Como até o presente momento os estudos descreviam a coordenação entre pares de variáveis, pretende-se por meio da expansão para o cenário multivariado encontrar novos padrões e realizar uma comparação com os resultados anteriores.

Em um segundo momento, deseja-se analisar a relação entre os movimentos da face, do trato vocal e acústica da fala, buscando além de quantificar a relação entre os domínios entender como ela varia ao longo do tempo [8]. Espera-se realizar uma comparação do mapeamento dinâmico entre os domínios com o mapeamento estático apresentado na referência e destacar os ganhos resultantes do uso da ferramenta aqui desenvolvida.

Por fim, a associação entre grupos será avaliada na área de finanças quantitativas. Neste trabalho será estudado como encontrar uma combinação linear de ativos que maximize a correlação entre dois grupos de ações, um com ativos na posição vendida e outro com os papéis na posição comprada. Espera-se por fim que os resultados auxiliem no desenvolvimento de um algoritmo de arbitragem estatística.

1.3 Organização do texto

O texto está organizado da seguinte maneira. No capítulo 2 será descrito o processo de aquisição de dados o pré-processamento dos mesmos.

No capítulo 3 será apresentada uma revisão de estatística multivariada e descrito o desenvolvimento do método que irá quantificar a associação instantânea entre grupos de variáveis assim como um detalhamento de como este foi implementado.

No capítulo 4 será apresentada uma revisão teórica do processo de

síntese da fala humana, assim como os principais algoritmos de codificação de fala. Neste capítulo, também será apresentado como remover a informação redundante existente entre amostras que caracterizam os movimentos da face e do trato vocal.

No capítulo 5 será exposta uma breve introdução sobre arbitragem estatística e teoria de modelagem de séries temporais financeiras para por fim descrever como as medidas de associação podem ser aplicadas no desenvolvimento de um algoritmo de negociação para fins de arbitragem estatística.

No capítulo seis serão apresentados e discutidos os resultados para as três bases de dados selecionadas.

Por fim, a conclusão é apresentada no capítulo sete.

Capítulo 2

Bases de dados

Neste capítulo, em cada seção será apresentada uma das fontes de dados utilizadas para se conduzir os estudos, tendo em vista que, deseja-se demonstrar a aplicação do método de associação entre grupos de variáveis nas áreas de processamento audiovisual da fala e finanças quantitativas.

No caso das aplicações na área de processamento audiovisual da fala, duas bases de dados foram utilizadas. A primeira consiste em dados que descrevem os movimentos do trato vocal de dois locutores em um experimento no qual eles estão interagindo (Seção 1). A segunda descreve os movimentos da face e do trato vocal quando locutores estão pronunciando um texto predefinido (Seção 2). A escolha por esta base se deu pelo fato de que dados que caracterizam os movimentos do trato vocal são de difícil acesso e que o laboratório onde o trabalho foi desenvolvido disponibiliza tais bases. Outro motivo pela escolha é que como o objetivo deste trabalho é desenvolver uma ferramenta, a possibilidade de compará-la com métodos consolidados em trabalhos anteriores ajuda a elucidar os prós e os contras do método desenvolvido.

Nas aplicações na área de finanças, os dados consistem em séries temporais de preços e volumes negociados de ações da BM& FBOVESPA (seção 3). O motivo da escolha por estes dados se deu pelo interesse em aplicar o método no mercado financeiro. Além disso, dados do mercado financeiro são atraentes por conta do alto volume de informação que

é disponibilizado a cada instante, possibilitando em trabalhos futuros a implementação da ferramenta em uma aplicação online.

2.1 Base 1 – Dados de produção audiovisual da fala – Experimento 1

A base apresentada nesta seção foi a mesma utilizada por [1] e foi gentilmente cedida pelos autores da referência. No experimento, dois locutores, um do sexo masculino (EVB) e outro do sexo feminino (CTB), foram posicionados um de frente para o outro a uma distância de 2 metros. A captura dos movimentos do trato vocal foi feita por articulografia eletromagnética (*Eletromagnetic articulography* - EMA). O sistema EMA é dividido em duas partes. A primeira consiste em um gerador de campo eletromagnético posicionado próximo à cabeça do locutor. A segunda parte consiste em sensores de posição dispostos ao longo da língua, cujas posições relativas são estimadas por indução eletromagnética e transmitidas por fios. Foram utilizados equipamentos diferentes para a coleta de dados: um Carstens AG500 no caso da locutora CTB e um NDI WAVE no caso do locutor EVB. O motivo pela escolha de dois equipamentos distintos se faz pelo fato de que os aparelhos utilizam frequências de transmissão e princípios de funcionamento diferentes, o que garante que os sinais capturados pelos sensores não sofram interferência. A distribuição dos sensores ao longo da língua é ilustrada na figura 2.1 e foi utilizada para ambos os locutores.

O processo de aquisição dos dados consiste em duas partes. Na primeira delas, nove *trials* foram realizadas onde, quando um locutor pronunciava o monossílabo *top*, por exemplo, o outro pronunciava um monossílabo que contrastava com o primeiro, *cop*. Variações como *topper* e *copper* também foram utilizadas. Nenhuma restrição foi aplicada aos locutores, de forma que não eram obrigados a sincronizar as suas falas e nem a cadenciar a pronúncia em determinado ritmo. Na segunda parte do experimento, os locutores conversavam livremente sobre os temas que

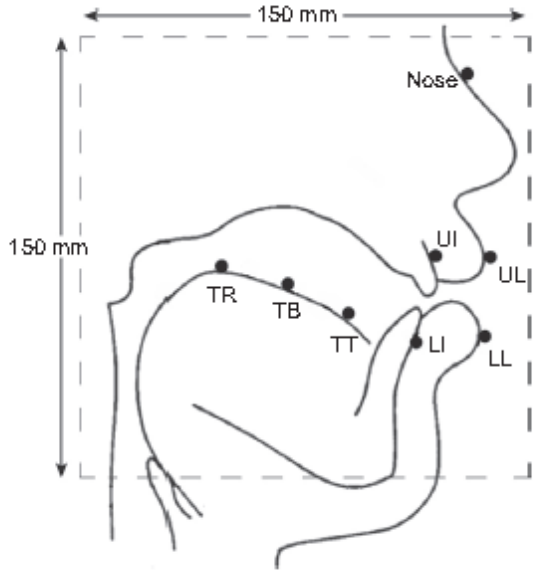

Figura 2.1: Distribuição dos sensores utilizados para capturar o movimento do trato vocal no experimento realizado em [1].

Tabela 2.1: Descrição da duração e monossílabos pronunciados nas *trials* 1 e 2

Trial	**EVB**	**CTB**	**Duração (s)**
1	top	cop	60
2	cop	top	20

desejassem. Neste trabalho serão apenas apresentados resultados para os dados das duas primeiras *trials* do experimento, nomeadas aqui como *trial 1* e *trial 2* a exemplo de [1]. As durações das *trials* 1 e 2 são apresentadas na tabela 2.1.

Cinco séries temporais (duas que representam a rotação do sensor e três que representam a posição do sensor) são extraídas para cada um dos sensores, e posteriormente, estas foram condensadas em um único sinal, que representa a distância do sensor em relação a um determinado ponto de referência. Desta maneira, cada sensor será representado neste trabalho por um único sinal, amostrado a uma frequência de $100Hz$. Nos testes realizados neste trabalho, uma matriz $\boldsymbol{T_1^{EVB}}$ contém em cada coluna

os valores de posição (em relação a um determinado referencial) para cada um dos sensores *TR, TB, TT, LI, LL, UL* (Figura 2.1) do locutor EVB, enquanto a matriz $\boldsymbol{T_1^{CTB}}$ foi definida pelo mesmo conjunto de sensores para o locutor CTB.

2.2 Base 2 – Dados de produção audiovisual da fala – Experimento 2

A base de dados utilizada para o estudo da associação entre o movimento da face, do trato vocal e a acústica da fala foi gentilmente disponibilizada pelos autores de [8]. Para captar os movimentos, foram distribuídos sensores de posição ao longo da face e da língua. A localização dos sensores é ilustrada na Figura 2.2. No experimento, foram coletados dados de dois locutores, o primeiro nativo de língua inglesa, *EVB*, e o segundo nativo de língua japonesa, *TK*.

O processo de aquisição do movimento da face foi realizado por um Optotrak, produzido pela *Northern Digital*, que é um equipamento usado para fazer o rastreamento de marcadores ativos (mais especificamente, marcadores que emitem luz infravermelha) em tempo real. A frequência de amostragem foi de $60Hz$ para o sujeito TK e 125 Hz para o sujeito EVB e no caso deste último, os dados foram reamostrados para $60Hz$. A precisão dos sensores é superior a 0.02 mm e a posição de cada sensor é caracterizada por três sinais de saída, cada uma correspondente a uma dimensão (e.g. x, y e z). Foram utilizados 12 sensores no experimento realizado com *EVB* e 18 no realizado com *TK*, assim totalizando 36 e 54 séries temporais, respectivamente. Todas os sinais, tanto para o experimento realizado com o *EVB* quanto para o *TK*, foram reamostrados para $60Hz$.

Os movimentos do trato vocal foram coletados por meio de articulografia eletromagnética (EMA) e dado o desconforto do experimento (causado pelos sensores colados ao longo da língua) e sua possível influência na expressão facial do locutor, em [8] os experimentos de captura de

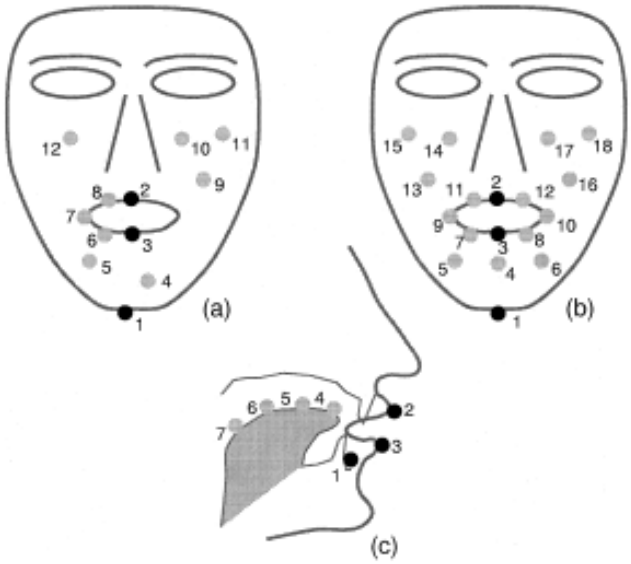

Figura 2.2: A figura que foi retirada de [8], demonstra a posição dos sensores que monitoram os movimentos do trato vocal e dos músculos faciais, sendo que foram posicionados de forma diferente ao longo da face. O primeiro é nativo de língua inglesa (a esquerda) e o segundo de língua japonesa (a direita).

movimento da face e do trato vocal foram realizados em sessões diferentes. A posição dos sensores é representada em um espaço bidimensional, diferente do processo realizado para a face, onde um posicionamento tridimensional é utilizado. Para ambos os sujeitos (TK e EVB), 7 sensores foram distribuídos ao longo da língua e 14 séries temporais foram geradas.

Para possibilitar a comparação dos movimentos da face e do trato vocal, em [8] os autores detalham como foi realizado o alinhamento entre os sinais da face e do trato vocal. Como pode-se observar na Figura 2.2 três sensores são posicionados no mesmo lugar para as sessões de aquisição de dados com EMA e Optotrak, o que garante que os dois grupos possuem sensores em comum. O alinhamento dos sinais da face e trato vocal foram realizados via do *Dynamic Time Warping*, método que é descrito detalhadamente em [14, 15].

No experimento também foram armazenados dados sobre a acústica da fala. A última foi quantificada por meio de coeficientes *Line Spectral Pairs (LSP)*, sendo utilizados 10 coeficientes. Como os sinais de fala foram capturados junto com a coleta dos movimentos faciais, não foi necessário um alinhamento deste com os grupos anteriores.

Neste trabalho serão utilizados somente os dados do locutor de língua inglesa (EVB). Para este, a base adquirida consiste de várias repetições de duas sentenças: *When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow* (Sentença 1) e *Sam sat on top of the potato cooker and Tommy cut up a bag of tiny potatoes and popped the beet tips into the pot* (Sentença 2). Para cada uma das sentenças, existem 5 repetições de dados de acústica da fala, de movimentos do trato vocal e da face. A base contém todas as possíveis combinações de movimentos da face e do trato vocal ($5 \times 5 = 25$) para cada sentença. Assim, cada série temporal consiste em 50 repetições de sentença concatenadas, sendo as 25 primeiras da sentença 1 e as 25 últimas da sentença 2.

Os dados foram armazenados em quatro matrizes, que aqui serão chamadas de $\boldsymbol{O_2}$, $\boldsymbol{T_2}$, $\boldsymbol{A_2}$ e $\boldsymbol{P_2}$, onde cada coluna representa uma série temporal. Como a posição dos sensores da face são representadas em um

espaço tridimensional, é necessário três sinais para cada sensor. Assim, a matriz $\boldsymbol{O_2}$ possui $12 \times 3 = 36$ colunas. No caso dos dados do trato vocal, são utilizados 7 sensores cujas posições são representadas em um espaço bidimensional, assim, a matriz $\boldsymbol{T_2}$ possui um total de 14 colunas. Para caracterizar os dados de acústica da fala foram utilizados 10 coeficientes LSP a tabela $\boldsymbol{A_2}$ possui 10 colunas. Todas as tabelas de dados possuem 13866 linhas. No vetor $\boldsymbol{P_2}$ estão armazenados os valores médio quadráticos (*Root Mean Square - RMS*) do sinal de fala, que foi calculado com base em quadros de amostras do sinal de fala, assim como realizado para os coeficientes LSP. Isso é possível pois a taxa de amostragem aplicada sobre o sinal de fala é muito superior a utilizada pelos equipamentos que capturam os movimentos do trato vocal e da face.

2.3 Base 3 – Dados do mercado financeiro

Os dados do mercado financeiro foram adquiridos pela plataforma *Metatrader 5* (https://www.metatrader5.com/en), uma das plataformas de *trading* automático mais utilizadas atualmente. O sistema, que é desenvolvido pela *MetaQuotes Software Corp.*, é ofertado aos clientes por corretoras de valores mobiliários presentes no mercado brasileiro. Pela plataforma, é possível obter acesso ao servidor da BM& FBOVESPA, a bolsa de valores do estado de São Paulo, que armazena séries temporais de preços e volumes negociados dos ativos, com atualização em tempo real. Os dados são disponibilizados para os usuários em duas estruturas, *Candlestick* e *Tick-by-tick*.

O sistema *Candlestick* foi criado inicialmente para prever o preço do arroz durante o período feudal no Japão. Cada *Candle* armazena quatro informações para um determinado intervalo de tempo:

- *High* - O preço mais alto do ativo registrado no período.
- *Low* - O preço mais baixo apresentado durante o período.
- *Close* - O último preço registrado ao fim do período.

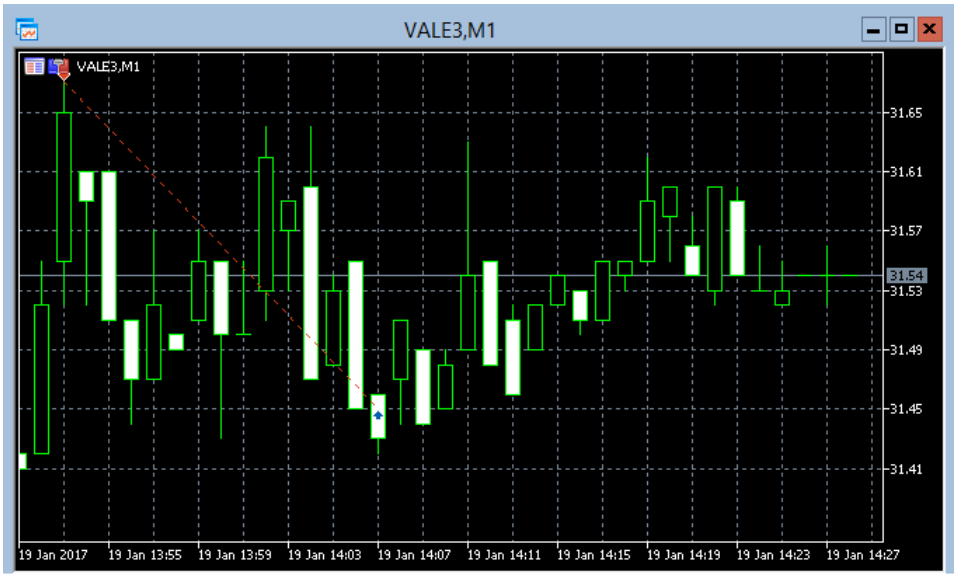

Figura 2.3: Exemplo de uma representação de preços em *Candlestick* para os preços do papel *VALE3* negociado na BM& FBOVESPA onde o período de amostragem é de um minuto

- *Open* - O preço do ativo no início do período.

Um exemplo de tal gráfico é apresentado na figura 2.3 em que a cada minuto um novo *Candle* é gerado. Um *candle* cheio indica que o preço de abertura, *Open*, é mais alto que o preço de fechamento, *Close*, demonstrando uma queda no preço do ativo no período analisado. Um *Candle* vazio indica o comportamento inverso, quando o preço de fechamento é mais alto que o preço de abertura, indicando aumento no preço do ativo durante o período considerado.

O *Tick-by-tick* é uma outra maneira de armazenar a informação onde os dados são coletados em tempo real. Um *Tick* é uma mudança de preço no ativo que está sendo monitorado pelo investidor. Toda vez que um *Tick* é detectado, um *TimeStamp* é associado a esse *Tick* e tal informação é adicionada à base de dados. O *TimeStamp* representa o instante de tempo em que um determinado evento acontece.

Todavia, o volume de *ticks* do mercado financeiro é muito alto e, por tal motivo, optou-se por trabalhar com amostragem uniforme. Neste

caso, será utilizada a captura do preço do valor do ativo em intervalos de tempo fixos. Os motivos pelos quais tal decisão de projeto foi tomada são simples. Primeiramente, utilizar todos os ticks possíveis acrescenta ruído de alta frequência nas séries temporais. Além disso, uma taxa de amostragem variável é incompatível com algumas ferramentas de visualização de dados que serão utilizadas neste trabalho, como a análise por mapas de associação por exemplo.

A grande vantagem de se trabalhar com uma fonte de dados tão ampla como o mercado financeiro para testar o método proposto é que é possível escolher entre séries de preços de ações e períodos de amostragem diferentes. Na figura 2.4 é apresentada uma visualização da interface do *Metatrader*. No gráfico, em verde, é apresentada a série temporal do preço de uma determinada ação (PETR4 - Ação preferencial da Petrobras). No painel, pode-se selecionar com qual frequência se deseja amostrar os dados. A granularidade é representada por uma letra e um número. A primeira descreve se os dados são amostrados em unidades de minutos (M), horas (H), dias (D), semanas (W) ou até mesmo meses (MN). O dígito indica o número de unidades utilizadas. Por exemplo, selecionar *M4* significa que amostras serão coletadas de 4 em 4 minutos.

Apesar de todas as facilidades, algum pré-processamento sobre os dados deve ser realizado para remoção de *outliers*. Deve-se garantir, primeiramente, que ao longo do período de análise nenhuma operação de *Split* e *Join* tenha ocorrido. A primeira ocorre quando uma ação sofre uma valorização muito expressiva e, para melhorar liquidez, decide-se dividir esta em um grupo de novos ativos com preço reduzido. Isso permite que não seja necessário um capital muito alto para se investir, incentivando investidores de pequeno porte. A segunda ocorre em no cenário inverso da primeira, quando as ações sofrem uma alta desvalorização e lotes de ações são substituídos por papéis com valores mais elevados.

Outra correção que deve ser realizada previamente sobre as séries consiste nos dividendos. Isso ocorre porque quando a entidade emissora do papel provém dividendos para seus acionistas, o valor é deduzido do valor da ação e investidores que adquirem o papel após a data onde o de-

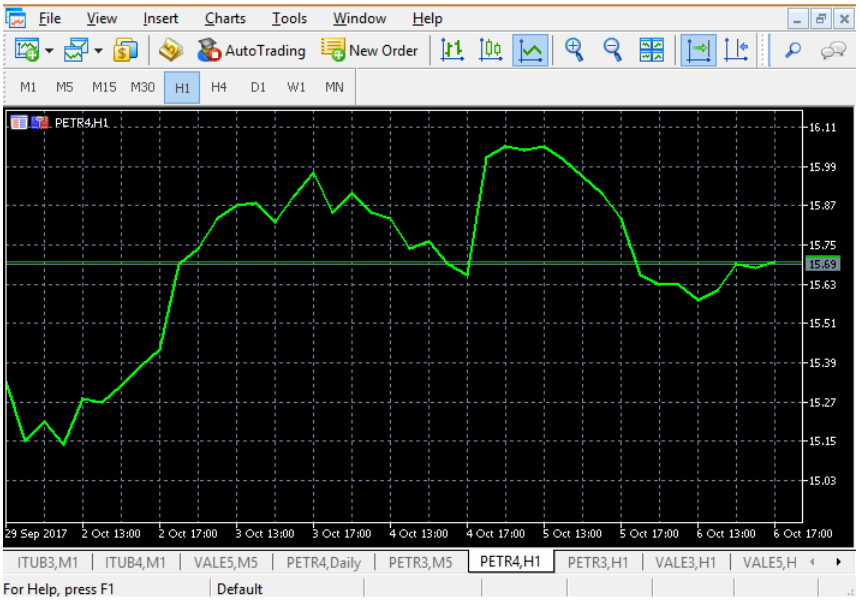

Figura 2.4: Apresentação da interface do software Metatrader, utilizado para adquirir dados em tempo real do mercado financeiro.

créscimo ocorre não tem mais direito a receber o benefício. Logo, deve-se desconsiderar tal decréscimo da análise que será feita.

O estudo de associação entre grupos de variáveis financeiras pode resultar em estratégias de investimento. Neste caso, deve-se levar em consideração a liquidez de cada ação, ou seja, qual facilidade de se comprar/vender determinado ativo. Neste trabalho serão utilizados papéis que são negociados constantemente, reduzindo a ocorrência do fenômeno conhecido pelo termo *Slipage*. Este ocorre quando indica-se que uma compra/venda de ação deve ser realizada em certo momento, mas não é possível realizar a operação por ausência de vendedores/compradores. Neste cenário, a operação fica suspensa e pode ser executada a um preço diferente do desejado, levando o investidor a eventuais perdas. Um exemplo de ação com baixa liquidez é o papel *ITUB3*, ação ordinária do banco Itaú (figura 2.5). Como pode-se observar, alguns *Candlesticks* são reduzidos a apenas um único ponto, o que indica que as negociações são quase um acordo bilateral entre vendedor e comprador e não uma operação de mercado.

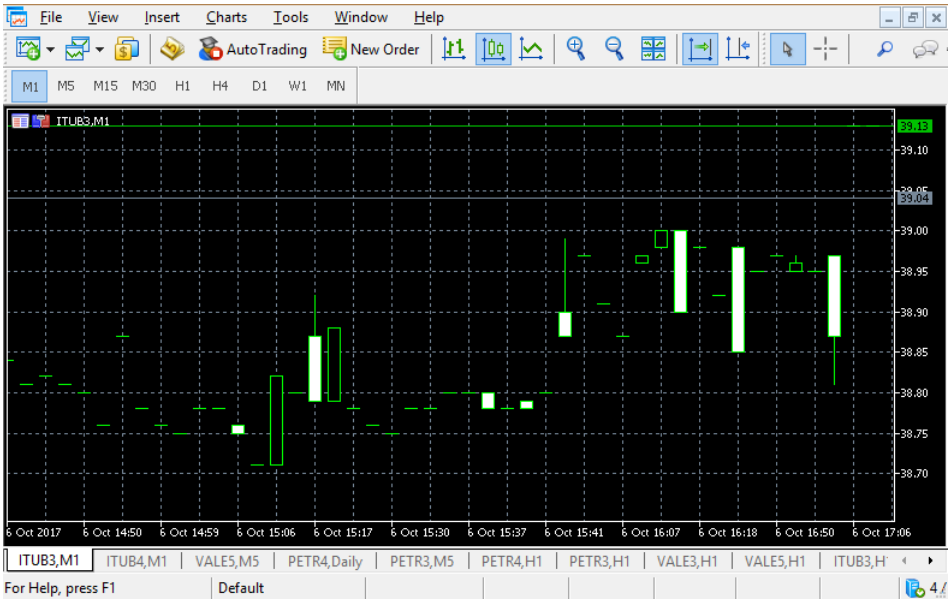

Figura 2.5: Exemplo de série temporal no formato *Candlestick* para um papel de baixa liquidez.

Para as simulações entre pares de ativos, foi utilizada uma base de dados composta de duas tabelas, ambas com dados das ações *VALE3* e *VALE5*. Neste trabalho, elas serão nomeadas ***PD*** e ***PM***, a primeira com dados com granularidade diária e a segunda com os dados coletados a cada minuto. Ambas as tabelas apresentam as seguintes colunas:

- *Timestamp* da amostra: dia e hora;
- Preço ao final do dia de *VALE3* e *VALE5*;
- Volume negociado em lotes para *VALE3* e *VALE5*;
- Diferença em minutos entre os *timestamps*.

Para a primeira tabela, as séries temporais apresentam os preços dos ativos de Junho/2010 até Dezembro/2016 enquanto para a segunda os dados vão de Janeiro/2016 até Dezembro/2016. O número de amostras armazenados é 1623 e 10700, respectivamente.

Para as simulações no cenário com mais de dois ativos, foram armazenados os dados dos ativos *PETR4*, *PETR3*, *ITUB3*, *ITUB4*, *VALE3*, *VALE5*,

OIBR3, OIBR4, BBDC3, BBDC4. Nesta base, não foram armazenadas informações sobre volumes negociados, somente os valores dos preços de fechamento. A granularidade dos dados foi diária e foram consideradas 500 amostras. Os valores dos preços das ações foram armazenados na tabela ***MPD***.

Comentários Finais

Neste capítulo foram descritas as formas de aquisição dos dados utilizados nas simulações. A primeira base de dados consiste em dados capturados dos movimentos do trato vocal quando dois locutores se comunicavam, por meio de EMA. A segunda base de dados caracteriza os movimentos da face, do trato vocal e o comportamento da acústica da fala. A terceira base apresenta séries temporais do preço de ativos negociados na BM& FBOVESPA e estes dados foram adquiridos diretamente do servidor da bolsa.

As fontes de dado passaram por uma etapa de pré-processamento, onde são removidos *outliers* e, quando necessário, realizado alinhamento prévio de séries temporais.

Capítulo 3

Associação Instantânea Multivariável

Este capítulo tem por objetivo realizar uma revisão dos métodos estatísticos necessários para estimar a associação instantânea entre grupos, bem como descrever os principais aspectos do método desenvolvido neste trabalho.

Por ser o caso mais simples possível (uma variável em cada grupo), primeiramente é descrito como é estimada a associação bivariada e as diferentes interpretações possíveis sobre o coeficiente de correlação de Pearson. Em seguida são caracterizadas as ferramentas matemáticas que possibilitam expandir a associação do caso bivariado para o multivariado e definidas as medidas de associação entre grupos de variáveis. Dentre as formas de se estabelecer o grau de associação multivariada, a primeira delas é baseada no conceito de variância compartilhada, onde é pressuposto que a associação entre dois grupos está diretamente relacionada com a capacidade de um grupo estimar as variáveis do outro. A segunda forma de se definir associação entre grupos é definida pela máxima correlação entre os mesmos. Tal correlação é calculada entre duas novas variantes estabelecidas a partir de uma combinação linear das variáveis de cada grupo. A terceira forma de se encontrar a associação entre os grupos é baseada em testes de hipótese. Nesse caso, o método estima a

probabilidade de duas variáveis ou grupos de variáveis serem descorrelacionados.

Todas as medidas de associação descritas neste capítulo encontram a relação entre variáveis a partir de matrizes de covariância. Logo, caso os coeficientes de tais matrizes sejam variantes no tempo, as medidas de associação também se tornam instantâneas. Por esta razão, este capítulo também apresentara como os coeficientes das matrizes de covariância são atualizados recursivamente utilizando um modelo média móvel exponencial [2]. Também é descrito como o método pode ser adaptado para se capturar eventuais avanços e atrasos existentes entre os grupos e apresentado como o método desenvolvido foi implementado em *software* (MATLAB®).

O capítulo está estruturado da seguinte maneira. Na seção 3.1 é apresentado um estudo para o caso de associação entre um par de variáveis. Na seção 3.2 são apresentadas as técnicas que possibilitam expandir o problema do caso bivariado para o multivariado assim como definidas as medidas estáticas de associação entre grupos. Na seção 3.3 é descrito como os coeficientes das matrizes de covariância foram estimados recursivamente, tornando a medida de associação variante no tempo além de apresentada uma ferramenta para detecção de eventuais atrasos entre os grupos de séries temporais. Por fim, uma síntese do método desenvolvido é apresentada na seção 3.4.

3.1 O caso estático bivariado

O caso de associação entre grupos de variáveis de mais fácil análise é aquele onde cada grupo apresenta uma variável, ou seja, o caso bivariado. Existem diferentes medidas de dependência entre duas variáveis, entre as quais podemos citar as medidas de correlação de Spearman, de Kendall e de Pearson [3]. A última define em um cenário bivariado o grau de relação linear entre duas variáveis e será uma das bases deste capítulo.

Por este motivo, a partir deste momento o termo correlação fará men-

ção a correlação de Pearson, definido como [2, 16]

$$\rho_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_{xx}\sigma_{yy}}} \tag{3.1}$$

A correlação normalmente é estimada a partir de um número finito de amostras. Por este motivo, sempre existirá uma incerteza sobre a estimava. Assim, o grau de associação entre variáveis pode ser visto como a probabilidade de que o valor da correlação entre ambas seja nula. Caso uma estimação de correlação entre duas variáveis atinja um valor muito elevado, a probabilidade desta ser igual a zero é baixa [4]. Todavia, o número de amostras utilizado para se estimar a correlação tem relação direta com a precisão. Quanto mais alto for o número de amostras, mais precisa é a estimativa da grandeza. O teste-T ou teste de Student sintetiza estes dois aspectos em um parâmetro definido como [4]

$$t = \frac{r}{\sqrt{\frac{(1-r^2)}{N-2}}} \tag{3.2}$$

onde N é o valor de amostras utilizado na estimação e r o valor estimado da correlação. Pode-se dizer que quanto maior for o valor do parâmetro t, maior a associação entre as duas variáveis, ou vendo por outro ângulo, menor a probabilidade das variáveis estarem descorrelacionadas.

Caso o número de amostras seja alto existe uma outra forma de se avaliar a associação entre duas variáveis. No contexto de uma regressão linear bivariada, ou seja, quando utiliza-se um modelo linear para se estimar o valor de uma variável dependente (y) a partir de uma variável independente (x), um baixo valor residual pode ser visto como um indicador de elevada dependência entre as variáveis [17].

A relação entre a correlação e a regressão linear é destacada no caso onde os dados estão normalizados e o modelo de regressão é definido por

$$y = r_{xy}x + \epsilon \tag{3.3}$$

O valor de correlação pode apresentar valores positivos e negativos e ambos representam uma alta associação. A grandeza ρ_{xy}^2 representa o montante da variância de y que pode ser representada a partir de x, porém como ela é desconhecida na maioria dos casos, pode ser aproximada por r_{xy}^2. Tal coeficiente também é chamado de *coeficiente de determinação* [4]. Este pode ser visto como um estimador de associação.

3.2 O caso estático multivariado

3.2.1 Análise em Componentes Principais

A Análise em Componentes Principais (PCA) é uma técnica que busca aglomerar informação redundante entre variáveis para representar a informação de um determinado grupo a partir de uma série de variantes, chamadas de componentes principais. Estas são obtidas a partir de combinações lineares das variáveis a priori.

Considere um vetor de variáveis aleatórias, $\boldsymbol{x}$ com comportamento gaussiano (média $\boldsymbol{\mu}_x$ e matriz de covariância $\boldsymbol{C}_{xx}$). Se for detectado uma correlação nula entre as variáveis, não é possível extrair nenhuma informação redundante sobre elas (por meio de estimadores lineares), caso contrário, existe informação redundante que pode ser agrupada. Mesmo para o caso de as variáveis possuírem um comportamento não gaussiano, a correlação representa a relação linear entre as mesmas e por isso o método apresenta resultados satisfatórios mesmo quando o vetor $\boldsymbol{x}$ não apresenta uma distribuição normal multivariada.

O objetivo da técnica é encontrar um vetor de pesos $\boldsymbol{a}_1$ que possibilite uma combinação linear das variáveis $\boldsymbol{x}$,

$$\boldsymbol{p}_1 = \boldsymbol{d}_1^{pca}\boldsymbol{x} \tag{3.4}$$

onde p_1 possua máxima variância [5]. Tal variável é chamada primeira componente principal. As componentes principais seguintes são encontradas da mesma forma, entretanto, elas devem estar descorrelacionadas

com as componentes principais encontradas anteriormente.

A matriz de covariância de um vetor aleatório é semi-definida positiva, o que garante que ela pode ser decomposta em valores singulares [5]. Tal manipulação consiste em decompor $\boldsymbol{C}$ como

$$\boldsymbol{C} = \boldsymbol{D}\boldsymbol{\Lambda}\boldsymbol{D}^T \tag{3.5}$$

onde $\boldsymbol{V}$ é uma matriz que contém os autovetores da matriz de covariância como colunas e $\boldsymbol{\Lambda}$ é uma matriz diagonal que contém os autovalores da matriz de covariância. A solução do problema é dada de forma que

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{d}_1^{pca} & \boldsymbol{d}_2^{pca} & \cdots & \boldsymbol{d}_{n_x}^{pca} \end{bmatrix} \tag{3.6}$$

ou seja, as novas bases são autovetores da matriz de covariância $\boldsymbol{C}$ e os autovalores $\boldsymbol{\Lambda}$ são as variâncias das componentes principais $\boldsymbol{p}$. A variância total do grupo de variáveis pode ser encontrada como a soma das variâncias de cada componente principal, ou seja

$$tr(\boldsymbol{\Lambda}) = \sum_{i=1}^{n_x} \lambda_i \tag{3.7}$$

e a contribuição de cada componente principal na variância total do vetor aleatório pode ser representada por

$$p_i^r = \frac{\lambda_i}{tr(\boldsymbol{\Lambda})} \tag{3.8}$$

Para demonstrar a *PCA* será considerado um exemplo com dados sintéticos. O vetor de variáveis aleatórias $\boldsymbol{x}$ tem dimensão três e as amostras foram geradas de acordo com as seguintes regras

$$x_1 \sim \mathcal{N}(0,1)$$

$$x_2 \sim \mathcal{N}(0,1)$$

$$x_3(k) = x_1(k) + x_2(k)$$

Como todos os sinais possuem esperança nula, pode-se considerar que

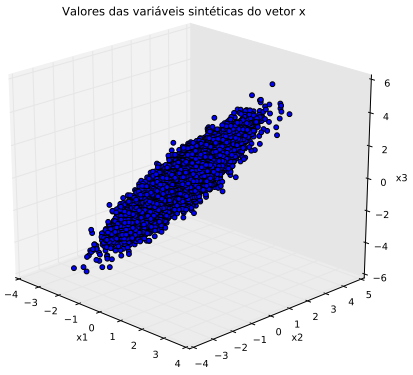

Figura 3.1: Representação das amostras geradas das variáveis x_1, x_2 e x_3 em um plano tridimensional

para um número suficientemente grande de amostras N o vetor que corresponde as médias das variáveis aleatórias será nulo, ou seja, $\boldsymbol{\mu} = 0$. Os pontos gerados são apresentados na figura 3.1. Como é observado, os valores formam um plano, pelo fato de x_3 poder ser escrita em função das outras duas variáveis.

Como esperado, o valor da terceira componente principal é praticamente nulo, pois este é dado por

$$0.58x_1 + 0.58x_2 - 0.58x_3 = 0 \tag{3.9}$$

3.2.2 Análise em Componentes Canônicas

A Análise em Componentes Canônicas tem por objetivo maximizar a correlação entre dois grupos de variáveis. Como no caso do PCA, a técnica consiste em encontrar novas bases de forma a ortogonalizar matrizes de covariância. Matematicamente, o problema consiste em encontrar pares de vetores $\boldsymbol{a}$ e $\boldsymbol{b}$ que sejam solução do seguinte problema de otimização

[6]

$$\rho = max_{\boldsymbol{a},\boldsymbol{b}} \frac{\boldsymbol{a}^T \boldsymbol{C}_{xy} \boldsymbol{b}}{\sqrt{(\boldsymbol{a}^T \boldsymbol{C}_{xx} \boldsymbol{a})(\boldsymbol{b}^T \boldsymbol{C}_{yy} \boldsymbol{b})}} \tag{3.10}$$

onde $\boldsymbol{C}_{yy}$,$\boldsymbol{C}_{xx}$ e $\boldsymbol{C}_{xy}$ são as matrizes de covariância dos vetores aleatórios $\boldsymbol{y}$, $\boldsymbol{x}$ e a matriz de covariância cruzada entre os dois vetores respectivamente. Assim como na análise em componentes principais, o problema consiste em uma mudança de base. Nesse caso, será realizada duas transformações lineares $\boldsymbol{x} \rightarrow \boldsymbol{u}^{\boldsymbol{x}}$ e $\boldsymbol{y} \rightarrow \boldsymbol{u}^{\boldsymbol{y}}$, onde

$$\boldsymbol{u}^{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_{n_x} \end{bmatrix}^T \times \boldsymbol{x} \tag{3.11}$$

e

$$\boldsymbol{u}^{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 & \cdots & \boldsymbol{b}_{n_y} \end{bmatrix}^T \times \boldsymbol{y} \tag{3.12}$$

As variáveis $\boldsymbol{u}$ e $\boldsymbol{u}^{\boldsymbol{y}}$ são chamadas de componentes canônicas. Estas devem ser ortogonais entre si, em termos matemáticos

$$E[\boldsymbol{u}_{\boldsymbol{i}}^{\boldsymbol{x}} \boldsymbol{u}_{\boldsymbol{j}}^{\boldsymbol{x}}] = 0 \qquad \forall i \neq j \tag{3.13}$$

$$E[\boldsymbol{u}_{\boldsymbol{i}}^{\boldsymbol{y}} \boldsymbol{u}_{\boldsymbol{j}}^{\boldsymbol{y}}] = 0 \qquad \forall i \neq j \tag{3.14}$$

O método é capaz de descrever totalmente a relação entre os grupos quando os grupos de variáveis apresentam distribuições normais multivariadas. Entretanto, mesmo quando as variáveis não assumem comportamento normal, o método maximiza e ortogonaliza a relação linear entre os dois grupos.

O problema possui duas soluções diferentes, uma proposta por [6] e a segunda em livros de estatística multivariável, como [5] e [4]. Em ambos os valores das correlações ρ são as mesmas, porém os mesmo apresentam restrições diferentes em relação a variância.

A solução apresentada em [6], as novas bases estão normalizadas, ou seja, $|\boldsymbol{a}_i| = 1$, $\forall i$ e $|\boldsymbol{b}_j| = 1$, $\forall j$. Considerando tal restrição, $\boldsymbol{a}_i$ são au-

tovetores da matriz $\boldsymbol{C_{xx}}^{-1}\boldsymbol{C_{xy}}\boldsymbol{C_{yy}}^{-1}\boldsymbol{C_{yx}}$ e $\boldsymbol{b}_j$ são os autovetores da matriz $\boldsymbol{C_{yy}}^{-1}\boldsymbol{C_{yx}}\boldsymbol{C_{xx}}^{-1}\boldsymbol{C_{xy}}$.

Entretanto, as referências na área da estatística, [5] e [4], fazem algumas exigências a mais sobre a solução analítica apresentada. As novas componentes canônicas devem respeitar a condição de que

$$u_i^x = \boldsymbol{a}_i^T \boldsymbol{x} \tag{3.15}$$

e

$$u_j^y = \boldsymbol{b}_j^T \boldsymbol{y} \tag{3.16}$$

devem possuir variância unitária, ou

$$E[(u^x)^2] = E[(u^y)^2] = 1 \tag{3.17}$$

Tal condição pode ser encontrada aplicando-se uma normalização das variáveis canônicas encontradas no primeiro método. Os valores das correlações ρ_i são iguais para ambos os métodos. O módulo das correlações, $|\rho_i|$, é encontrado como $\sqrt{\lambda_i}$, onde λ_i são autovalores da matriz $\boldsymbol{C_{xx}}^{-1}\boldsymbol{C_{xy}}\boldsymbol{C_{yy}}^{-1}\boldsymbol{C_{yx}}$ correspondentes ao autovetores $\boldsymbol{a}_i$. Os valores das correlações, ρ_j são igualmente encontrados por $\sqrt{\lambda_j}$, onde λ_i são autovalores da matriz $\boldsymbol{C_{yy}}^{-1}\boldsymbol{C_{yx}}\boldsymbol{C_{xx}}^{-1}\boldsymbol{C_{xy}}$ correspondentes ao autovetores $\boldsymbol{b}_j$.

Até o momento, nenhuma menção foi feita em relação ao número de variáveis existentes em cada grupo. Neste trabalho, o número de elementos no vetor $\boldsymbol{x}$ será representado por n_x e o número de variáveis no vetor $\boldsymbol{y}$ será representado por n_y. O método não exige que $n_x = n_y$. Considere o exemplo em que $n_y = 3$ e $n_x = 2$. Nesse caso, existirá um autovetor e um autovalor a mais em uma das matrizes. No caso, o que acontece com o método é que alguns autovalores da segunda matriz serão nulos, e apenas os autovetores que possuírem autovalores não nulos serão utilizados. Matematicamente, isso pode ser descrito como

$$\forall i, j \leq (min(n_x, n_y)), \qquad \lambda_i = \lambda_j \tag{3.18}$$

$$\forall i, j > min(n_x, n_y), \qquad \lambda_i = \lambda_j = 0 \tag{3.19}$$

onde o índice i refere-se aos termos do vetor $\boldsymbol{x}$ e j aos termos de $\boldsymbol{y}$. Por motivos de simplicidade em termos de notação, assim como em [5] vamos definir uma nova variável

$$s = min(n_x, n_y) \tag{3.20}$$

A CCA foi aplicada a dados sintéticos onde os dois vetores de variáveis aleatórias $\boldsymbol{x}$ e $\boldsymbol{y}$:

- Possuem duas variáveis. Em ambos os casos, as amostras são geradas a partir de uma distribuição normal.
- As variáveis estão relacionadas pela seguinte relação linear

$$x_1(k) + x_2(k) = y_1(k) + y_2(k) \tag{3.21}$$

A Figura 3.2 que possui um gráfico do tipo *scatter* dos grupos de variáveis. Como pode-se observar, é muito difícil de se notar alguma relação entre as variáveis em um primeiro momento. Em seguida, foram plotadas as componentes canônicas na Figura 3.3. Neste caso, foi detectada a relação linear entre os dois vetores de variáveis aleatórias, o que era esperado dado a relação $x_1(k) + x_2(k) = y_1(k) + y_2(k)$. Como descrito anteriormente, novas bases ortogonais foram obtidas, e o problema possui uma interpretação vetorial descrita nas figuras 3.4 e 3.5.

No exemplo, o número de elementos dos dois vetores aleatórios, $\boldsymbol{x}$ e $\boldsymbol{y}$, é igual a dois. Mas o que ocorre quando os vetores contêm dimensões diferentes (e.g. $n_y = 3$ e $n_x = 2$)? Como o $min(n_x, n_y) = n_x = 2$, somente dois autovetores referentes a matriz $\boldsymbol{C_{yy}}^{-1}\boldsymbol{C_{yx}}\boldsymbol{C_{xx}}^{-1}\boldsymbol{C_{xy}}$ serão utilizados como base, por possuírem autovalor não nulo. Nesse caso, pode-se afirmar que dois pares canônicos serão formados.

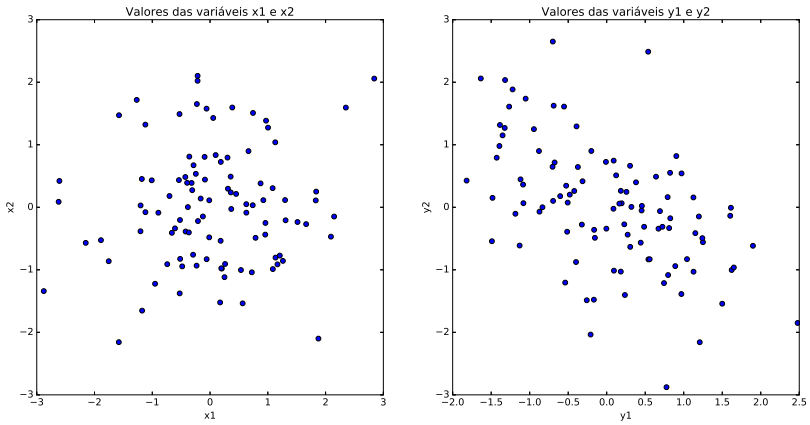

Figura 3.2: Exemplo onde as variáveis são vetores aleatórios gaussianos de dimensão dois. No eixo das abscissas estão os valores para as variáveis de índice 1 e no das ordenadas os valores das variáveis de índice 2

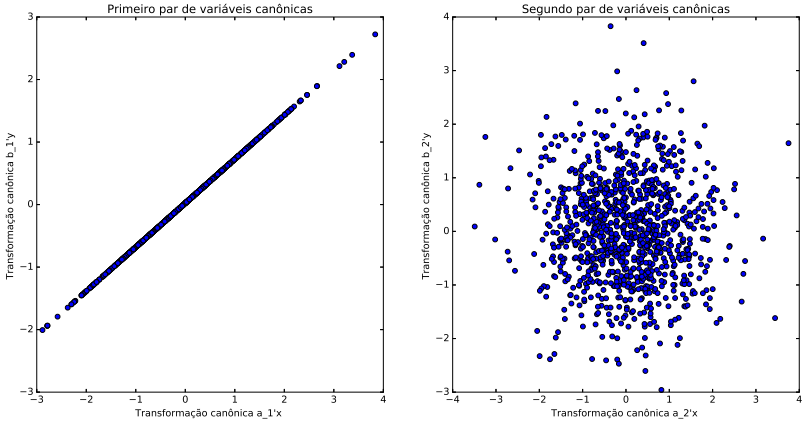

Figura 3.3: Exemplo onde as variáveis são vetores aleatórios gaussianos de dimensão dois. Nesta figura, estão plotados os dois pares canônicos. Como podemos notar, foi encontrada uma relação linear.

3.2.3 Associação definida pela variância compartilhada entre grupos de variáveis

Quando é realizada uma regressão linear entre duas variáveis, pode-se dizer ambas estão fortemente associadas quando a variância relativa do resíduo é baixa. Analogamente, pode-de dizer no caso multivariado que

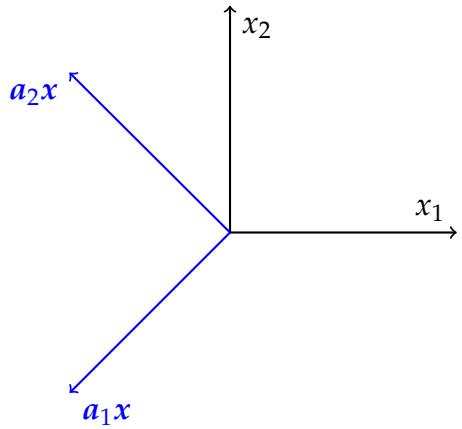

Figura 3.4: Transformação de espaço causada pelas componentes canônicas sobre o vetor aleatório $\boldsymbol{x}$

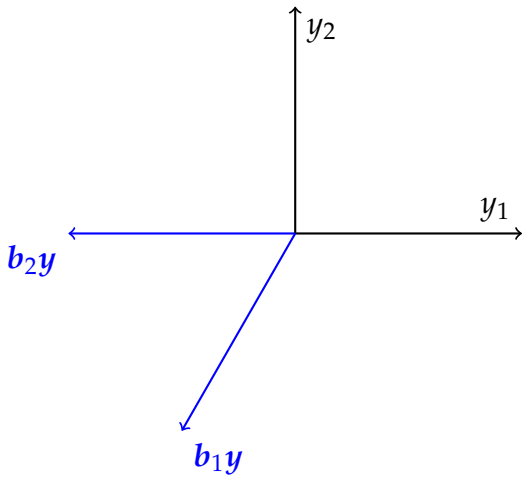

Figura 3.5: Transformação de espaço causada pelas componentes canônicas sobre o vetor aleatório $\boldsymbol{y}$

dois grupos de variáveis estão altamente associados quando uma alta parte da variância de um grupo pode ser explicada a partir do outro, e vice-versa. Por este motivo, nesta seção serão apresentadas medidas de associação baseadas na porcentagem de variância compartilhada entre os grupos. Duas formas diferentes de se encontrar esta medida serão discutidas: a primeira baseada na Análise em Componentes Canônicas e a segunda baseada na Análise em Componentes Principais.

No caso bivariado, o valor da correlação de Pearson nos informa a

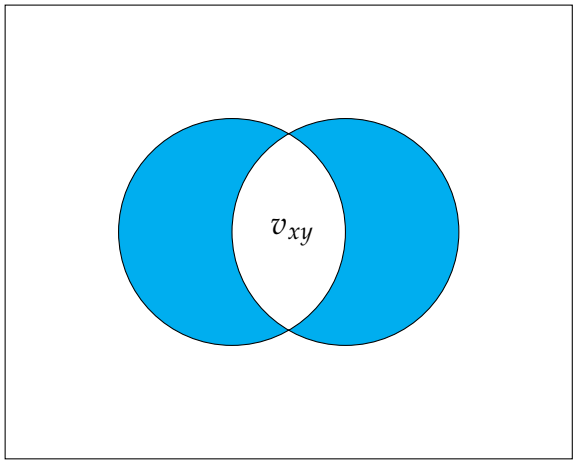

Figura 3.6: Exemplo de variância compartilhada entre as componentes canônicas. Neste exemplo, cada círculo representa a variância total de cada grupo ($\boldsymbol{u}^{\boldsymbol{x}}$ a esquerda e $\boldsymbol{u}^{\boldsymbol{y}}$ a direita). A porcentagem da variância compartilhada é a mesma, pois os dois grupos possuem o mesmo valor de variância total e a matriz de covariância cruzada é diagonal.

porcentagem da variância que está sendo compartilhada pelo coeficiente de determinação, entretanto, no caso multivariado, a variância compartilhada pelos grupos pode ter uma representatividade diferente sobre cada um dos grupos. Para explicar o fenômeno, considere os diagramas de Venn apresentados nas figuras 3.6 e 3.7. Em ambas as figuras estão representados a variância compartilhada e os valores da variâncias totais de cada grupo. Pode-se observar que a razão entre a variância compartilhada e as variâncias de cada grupo podem ser diferentes. Assim, duas medidas de associação serão definidas: a porcentagem da variância do grupo $\boldsymbol{x}$ que pode ser representada a partir do grupo $\boldsymbol{y}$, ($v_{x|y}$), e a porcentagem de variância do grupo $\boldsymbol{y}$ que pode ser representada a partir do grupo $\boldsymbol{x}$, ($v_{y|x}$).

Cálculo da Variância Compartilhada a partir de Componentes Canônicas

Conforme descrito em [7], uma das principais limitações das componentes canônicas é que elas não representam a variância total dos grupos.

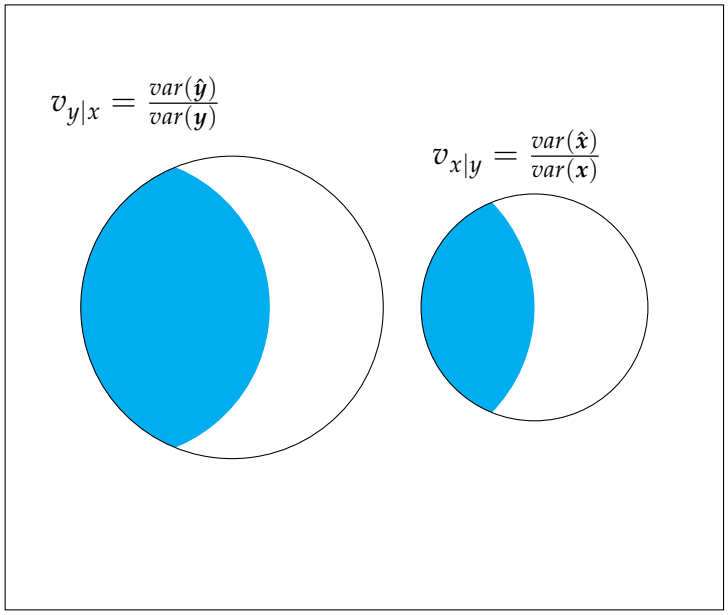

Figura 3.7: Exemplo de associação por variância compartilhada entre dois grupos de variáveis $\boldsymbol{x}$ e $\boldsymbol{y}$. Neste caso, as variâncias totais dos grupos ($var(x)$ e $var(y)$, representadas pelas áreas totais dos círculos) e as variâncias que podem ser representadas pelo outro grupo ($var(\hat{x})$ e $var(\hat{y})$, representadas pelas áreas azuis sobre os círculos) são distintas, assim como os valores das medidas de associação.

Isto ocorre pois caso duas variáveis sejam altamente correlacionadas dentro de um mesmo grupo as componentes canônicas somente levarão em consideração a variância de uma delas. Pode-se também interpretar que a CCA descarta qualquer informação redundante existente dentro dos grupos.

Contudo, uma vez que as componentes canônicas são combinações lineares das variáveis a priori, é possível reconstruir as variáveis de cada grupo a partir das componentes canônicas (somente no caso do conjunto que possui o menor número de variáveis, pois quando um grupo possui mais variáveis que o número de componentes canônicas, existe uma perda natural de informação devido a redução de dimensionalidade). Com base nesse conceito, em [18] é apresentada uma medida de variância compartilhada baseada na análise em componentes canônicas. O autor estabelece esta medida buscando representar a associação entre grupos como o valor médio da variância compartilhada entre cada variável e cada componente canônica.

A porcentagem da variância do grupo $\boldsymbol{x}$ que pode ser representada a partir do grupo $\boldsymbol{y}$ é calculada em duas etapas. Primeiramente, para cada componente canônica encontra-se a média ponderada do quadrado das correlações entre esta e as variáveis iniciais. Em um segundo instante, o valor médio destas correlações é multiplicado pelo valor da correlação canônica ao quadrado, com o objetivo de realizar a ponte entre os grupos. O cálculo pode ser sintetizado pela seguinte equação

$$v_{x|y}^{cca} = \sum_{k=1}^{s} \rho_k^2 \sum_{i=1}^{n_x} \frac{L_{i,k}^2}{n_x} \tag{3.22}$$

e para o caso inverso

$$v_{y|x}^{cca} = \sum_{k=1}^{s} \rho_k^2 \sum_{i=1}^{n_y} \frac{L_{i,k}^2}{n_y} \tag{3.23}$$

onde L_i é o valor da correlação entre a componente canônica i e a variável k. Esta medida, por realizar uma média ponderada das correlações entre as componentes canônicas e a as variáveis iniciais, indiretamente realiza

uma normalização entre as variáveis, ou seja, considera que todas as variáveis possuem variância unitária a priori. Uma descrição do algoritmo utilizado para extrair a medida de associação a partir das matrizes de covariância é descrito no Algoritmo 1.

Algoritmo 1 Cálculo da medida de associação v^{cca}

Input: $\boldsymbol{C_{xx}}$,$\boldsymbol{C_{yy}}$ e $\boldsymbol{C_{xy}}$

$s \leftarrow min(n_x, n_y)$

$(\boldsymbol{A'}, \boldsymbol{\rho}) \leftarrow eig({\boldsymbol{C_{xx}}}^{-1/2}\boldsymbol{C_{xy}}{\boldsymbol{C_{yy}}}^{-1}{\boldsymbol{C_{xy}}}^{T}{\boldsymbol{C_{xx}}}^{-1/2})$

$(\boldsymbol{B'}, \boldsymbol{\rho}) \leftarrow eig({\boldsymbol{C_{yy}}}^{-1/2}{\boldsymbol{C_{xy}}}^{T}{\boldsymbol{C_{xx}}}^{-1}\boldsymbol{C_{xy}}{\boldsymbol{C_{yy}}}^{-1/2})$

$\boldsymbol{A} \leftarrow {\boldsymbol{C_{xx}}}^{-1/2}\boldsymbol{A'}$

$\boldsymbol{B} \leftarrow {\boldsymbol{C_{yy}}}^{-1/2}\boldsymbol{B'}$

$\boldsymbol{C_{u^x u^x}} \leftarrow \boldsymbol{A}^T\boldsymbol{C_{xx}}\boldsymbol{A}$

$\boldsymbol{C_{u^y u^y}} \leftarrow \boldsymbol{B}^T\boldsymbol{C_{yy}}\boldsymbol{B}$

$\boldsymbol{C_{xu^x}} \leftarrow (\boldsymbol{C_{xx}}\boldsymbol{A})./diag(\boldsymbol{C_{xx}})$

$\boldsymbol{C_{yu^y}} \leftarrow (\boldsymbol{C_{yy}}\boldsymbol{B})./diag(\boldsymbol{C_{yy}})$

$v_{x|y} \leftarrow mean(\boldsymbol{C_{xu^x}}.^2)\boldsymbol{\rho}$

$v_{y|x} \leftarrow mean(\boldsymbol{C_{yu^y}}.^2)\boldsymbol{\rho}$

Output: $resultado \leftarrow [v_{x|y}, v_{y|x}]$

Tal propriedade é interessante quando estão sendo comparadas variáveis que possuem ordem de grandeza diferentes, caso onde é difícil de se dizer que a variância é um indicador de precisão de determinada medição. Todavia, em algumas aplicações, quando as variáveis possuem uma mesma ordem de grandeza, o valor da variância delas pode ser um indicador da precisão, no caso de um sensor, ou até mesmo de risco como é feito para preços de ativos no mercado financeiro. Por este motivo, será apresentada uma nova medida de associação, baseada na PCA, que considera a variância das variáveis a priori.

Cálculo da Variância Compartilhada a partir de Componentes Principais

Uma das principais e mais úteis propriedades da análise em componentes principais é que ela é capaz de criar um novo conjunto de variáveis descorrelacionadas a partir das variáveis iniciais. Entretanto, a análise em

componentes canônicas também apresenta tal propriedade. Logo, qual a distinção entre os dois métodos?

No caso da CCA o novo conjunto de bases $\boldsymbol{A}$ e $\boldsymbol{B}$ não são ortogonais entre si, e consequentemente, o novo conjunto de variáveis não representará a variância total do sistema. No caso da PCA a solução apresentada possui características diferentes. Ela assegura que as novas bases criadas são ortogonais, o que garante que novo espaço conserva a variância total das variáveis originais, alocando a maior parte da informação redundante nas primeiras componentes principais. Todavia, a matriz de covariância cruzada não será diagonal. Desta maneira, para representar as componentes principais do grupo $\boldsymbol{x}$, $\boldsymbol{p_x}$, a partir das componentes principais do grupo $\boldsymbol{y}$, $\boldsymbol{p_y}$, será necessário o uso de uma transformação afim

$$\hat{\boldsymbol{p}}_x = \boldsymbol{T_{xy}} \boldsymbol{p}_y \tag{3.24}$$

O valor da variância de cada componente principal é dado pelos autovalores da matriz de covariância $\boldsymbol{C}_{xx}$, que serão representados por λ. De maneira semelhante, a variância das componentes estimadas podem ser encontradas por

$$\hat{\lambda}_{x,i} = E[\hat{p}^2_{x,i}] \tag{3.25}$$

Como as componentes principais são ortogonais, a variância total do grupo é encontrada pela soma dos autovalores da matriz de covariância referente às variáveis presentes dentro de cada grupo. Logo, a variância total do grupo $\boldsymbol{x}$ explicada a partir do grupo $\boldsymbol{y}$ é estimada pela razão

$$v^{pca}_{\boldsymbol{x}|\boldsymbol{y}} = \frac{\sum_{i=1}^{n_x} \hat{\lambda}_{x,i}}{\sum_{j=1}^{n_x} \lambda_{x,j}} \tag{3.26}$$

e a medida $v^{pca}_{y|x}$ pode ser encontrada seguindo o mesmo raciocínio. O método foi implementado em software como descrito no Algoritmo 2.

Algoritmo 2 Cálculo da medida de associação v^{pca}

Input: $\boldsymbol{C_{xx}}$, $\boldsymbol{C_{yy}}$ e $\boldsymbol{C_{xy}}$

$(\boldsymbol{V_x}, \boldsymbol{\Lambda_x}) \leftarrow eig(\boldsymbol{C_{xx}})$

$(\boldsymbol{V_x}, \boldsymbol{\Lambda_y}) \leftarrow eig(\boldsymbol{C_{yy}})$

$\boldsymbol{\Lambda_{xy}} \leftarrow \boldsymbol{V_x}^T \boldsymbol{C_{xy}} \boldsymbol{V_y}$

$\boldsymbol{T_{xy}} \leftarrow \boldsymbol{\Lambda_{xy}} / \boldsymbol{\Lambda_y}$

$\boldsymbol{T_{yx}} \leftarrow \boldsymbol{\Lambda_{xy}}^T / \boldsymbol{\Lambda_x}$

$v^{pca}_{x|y} \leftarrow trace(\boldsymbol{T_{yx}} \boldsymbol{\Lambda_x} \boldsymbol{T_{yx}}^T) / trace(\boldsymbol{\Lambda_x})$

$v^{pca}_{y|x} \leftarrow trace(\boldsymbol{T_{xy}} \boldsymbol{\Lambda_y} \boldsymbol{T_{xy}}^T) / trace(\boldsymbol{\Lambda_y})$

Output: $resultado \leftarrow [v^{pca}_{x|y}, v^{pca}_{y|x}]$

3.2.4 Associação definida por teste de hipótese e máxima correlação entre grupos de variáveis

No início do capítulo, foi apresentado como o teste-T de hipótese pode ser utilizado para determinar a associação entre duas variáveis e que quando deseja-se rejeitar a hipótese nula $H_0 : \rho = 0$, quanto mais alto é o valor do parâmetro t estabelecido pelo teste, maior é a associação entre as variáveis.

Como uma extensão do teste de hipótese sobre uma correlação bivariada, existem testes de hipótese para o medir a descorrelação entre dois grupos de variáveis a partir da análise em componentes canônicas [5, 4]. Como a CCA tende a encontrar a máxima correlação entre os grupos, podemos dizer que os mesmos são descorrelacionados caso

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_s = 0 \tag{3.27}$$

que é equivalente a testar a hipótese nula parar a matriz de covariância canônica $\boldsymbol{C_{u^x u^y}} = \boldsymbol{0}$. Como já foi mencionado, após a decomposição em componentes canônicas, a nova matriz de covariância cruzada $\boldsymbol{C_{u^x u^y}}$ é diagonal e os elementos desta são os valores das correlações canônicas. Uma outra forma de se ver o problema é considerar que caso os grupos

sejam independentes, a seguinte condição é satisfeita

$$\begin{bmatrix} \rho_1 & 0 & \dots & 0 \\ 0 & \rho_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \rho_s \end{bmatrix} = \mathbf{0} \tag{3.28}$$

ou de forma equivalente, considerar que

$$(\boldsymbol{C}_{xx})^{-1/2}\boldsymbol{C}_{yx}\boldsymbol{C}_{xx}{}^{-1}\boldsymbol{C}_{xy}(\boldsymbol{C}_{yy})^{-1/2} = \mathbf{0} \tag{3.29}$$

O critério de *Wilks*, o primeiro dos testes de hipótese propostos para tal problema define um parâmetro para o cenário multivariado onde [19]

$$\Lambda = \frac{|\boldsymbol{C}|}{|\boldsymbol{C}_{u^x u^x}||\boldsymbol{C}_{u^y u^y}|} = \frac{\prod_{i=1}^{s} \lambda_{u^x,i} \prod_{j=1}^{s} \lambda_{u^y,j} - (\prod_{k=1}^{s} \lambda_{u^x u^y,k})}{\prod_{i=1}^{s} \lambda_{u^x,i} \prod_{j=1}^{s} \lambda_{u^y,j}} = \prod_{i=1}^{s}(1-\rho_i^2) \tag{3.30}$$

onde

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{u^x u^x} & \boldsymbol{C}_{u^x u^y} \\ \boldsymbol{C}_{u^y u^x} & \boldsymbol{C}_{u^y u^y} \end{bmatrix} \tag{3.31}$$

assim, como no caso bivariado, podemos determinar uma relação entre o teste de hipótese sobre correlação nula e a associação entre os grupos. No caso do teste de *Wilks*, quanto mais próximo Λ se encontra do valor unitário, maior a probabilidade dos grupos estarem descorrelacionados. Um dos principais problemas deste coeficiente é uma eventual perda da precisão numérica, pois caso a primeira componente canônica seja muito próxima do valor unitário, o valor de Λ já estará muito próximo do valor nulo e o impacto das outras componentes canônicas sobre o coeficiente não será significativo. Neste caso, a máxima correlação entre os grupos (ou primeira correlação canônica) pode ser vista como uma medida de associação

$$c = \rho_1^2 \tag{3.32}$$

Um exemplo do uso de c será discutido em um dos capítulos posteriores deste trabalho onde estaremos estudando análise de risco para ativos no mercado financeiro. Dada a instabilidade sobre o valor da primeira componente canônica [7], para aumentar a robustez, as outras componentes canônicas devem ser consideradas e uma medida que mede a probabilidade dos grupos estarem descorrelacionados (h) pode ser definida como a porcentagem de variância compartilhada entre as componentes canônicas [5, 20]

$$h = \frac{1}{s}\sum_{i=1}^{s}\rho_i^2 \tag{3.33}$$

Tal medida de associação foi implementada computacionalmente conforme descrito no Algoritmo 3.

Algoritmo 3 Cálculo da medida de associação h.

Input: $\boldsymbol{C_{xx}}$, $\boldsymbol{C_{yy}}$ e $\boldsymbol{C_{xy}}$
$s \leftarrow min(n_x, n_y)$
$\boldsymbol{\rho} \leftarrow eig(\boldsymbol{C_{xx}}^{-1/2}\boldsymbol{C_{xy}}\boldsymbol{C_{yy}}^{-1}\boldsymbol{C_{xy}}^{T}\boldsymbol{C_{xx}}^{-1/2})$
Output: $resultado \leftarrow sum(\boldsymbol{\rho}.^2)/s$

3.3 O caso variante no tempo

Para capturar variações temporais no valor da associação, nesta seção será apresentado como as matrizes de covariância que descrevem a relação entre as variáveis passam a ser reavaliadas a todo instante. A partir deste momento, a matriz de covariância instantânea ($\boldsymbol{C_{xy}}(k)$) entre dois vetores aleatórios no instante k, $\boldsymbol{x}(k)$ e $\boldsymbol{y}(k)$, será definida como

$$\boldsymbol{C_{xy}}(k) = E[(\boldsymbol{x}(k) - \hat{\boldsymbol{x}}(k))(\boldsymbol{y}(k) - \hat{\boldsymbol{y}}(k))^T] \tag{3.34}$$

onde

$$\hat{\boldsymbol{x}}(k) = E(\boldsymbol{x}) \tag{3.35}$$

e

$$\hat{\boldsymbol{y}}(k) = E(\boldsymbol{y}). \tag{3.36}$$

Talvez a forma mais simples de se capturar variações no valor da covariância seja por meio de um modelo média móvel. Neste caso, ao invés de utilizar todas as amostras existentes até o instante k, a média do sinal e a variância são estimadas a partir das N_w últimas amostras. A estimação da covariância entre dois sinais, x e y, no instante k pode ser dada por

$$cov_{xy}(k) = \frac{1}{N_w}[\sum_{l=0}^{N_w-1} x(k-l)y(k-l) - (\sum_{l=0}^{N_w-1} x(k-l))(\sum_{l=0}^{N_w-1} y(k-l))] \tag{3.37}$$

Neste trabalho optamos por utilizar a média móvel com decaimento exponencial descrita em [2]. O motivo da escolha desta solução ocorreu em função de seu baixo custo computacional. Todavia, nada impede que outros métodos sejam utilizados para tal objetivo, como por exemplo modelos com viés bayesiano como filtros de Kalman ou filtros de partículas [21] e modelos com base em heterocedasticidade condicional [22].

Para o filtro média móvel exponencial, a covariância instantânea é definida como sendo [2]

$$cov_{xy}(k) = \sum_{l=0}^{\infty} c_{filt} e^{-\eta|l|} x(k-l)y(k-l) - [\sum_{l=0}^{\infty} c_{filt} e^{-\eta|l|} x(k-l)][\sum_{l=0}^{\infty} c_{filt} e^{-\eta|l|} y(k-l)] \tag{3.38}$$

onde c_{filt} é um fator que assegura a normalização dos pesos atribuídos a cada amostra. Os valores de $cov_{xx}(k)$ e $cov_{yy}(k)$ podem ser encontrados da mesma maneira. A estimação de covariância por meio do filtro apresentado pode ser representada como um problema de filtragem onde a resposta ao impulso do filtro é dada por

$$h_{filt}(k) = c_{filt} e^{-\eta|k|} \tag{3.39}$$

Definindo-se o sinal $i(k)$ como

$$\theta(k) = x(k)y(k) \tag{3.40}$$

a estimação da covariância pode ser vista como uma resposta do sistema linear, obtida pelo operador de convolução

$$cov_{xy}(k) = h_{filt}(k) * \theta(k) - [h_{filt}(k) * x(k)][h_{filt}(k) * y(k)] \tag{3.41}$$

Como o objetivo da análise é estabelecer uma matriz de covariância, a operação de filtragem deve ser realizada sobre todos os pares de variáveis possíveis. Como a utilização de um *loop* acabaria aumentando o custo computacional e o tempo de processamento consideravelmente, são criadas duas matrizes (cujas estruturas são demonstradas na Figura 3.8) que auxiliam a realizar todas as combinações possíveis entre variáveis conforme descrito no diagrama da Figura 3.9. Estas matrizes otimizam o cálculo da matriz de covariância, pois quando as funções de filtragem do *MATLAB*® recebem como entrada uma matriz, o filtro é aplicado sobre cada uma das colunas individualmente. Após o cálculo das covariâncias, o resultado é redimensionado de uma matriz $N \times (n_x n_y)$ para um estrutura de dados com três dimensões $N \times n_x \times n_y$, ou seja, uma matriz de covariância cruzada para cada instante de tempo considerado na análise. O processo de cálculo da matriz de covariância instantânea pode ser resumido pelo Algoritmo 4.

Algoritmo 4 $covinst(\boldsymbol{X}, \boldsymbol{Y})$ - Função que calcula a covariância instantânea entre dois grupos de variáveis.

Input: $\boldsymbol{X}$,$\boldsymbol{Y}$

$\mathbf{P} \leftarrow$ repetir cada coluna de $\boldsymbol{X}$ n_y vezes.

$\mathbf{Q} \leftarrow$ replicar cada coluna da matriz $\boldsymbol{Y}$ n_x vezes.

$\mathbf{C} \leftarrow (\boldsymbol{P}.\boldsymbol{Q}) * h_{filt} - (\boldsymbol{P} * h_{filt}). \times (\boldsymbol{Q} * h_{filt})$

Redimensionar o array $\boldsymbol{C}$ com dimensões $N \times (n_x * n_y)$ para $N \times n_x \times n_y$

Output: $\boldsymbol{C}$

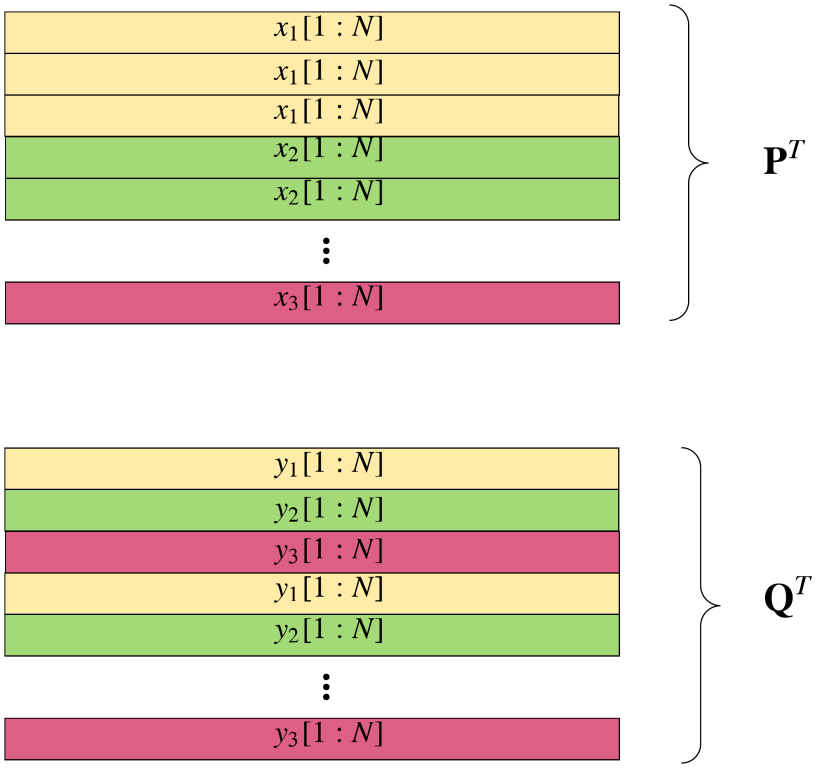

Figura 3.8: Estrutura das Matrizes P e Q que auxiliam na redução do custo computacional durante o cálculo da matriz de covariância. No exemplo, tanto o grupo $\boldsymbol{x}$ quanto $\boldsymbol{y}$ possuem 3 variáveis.

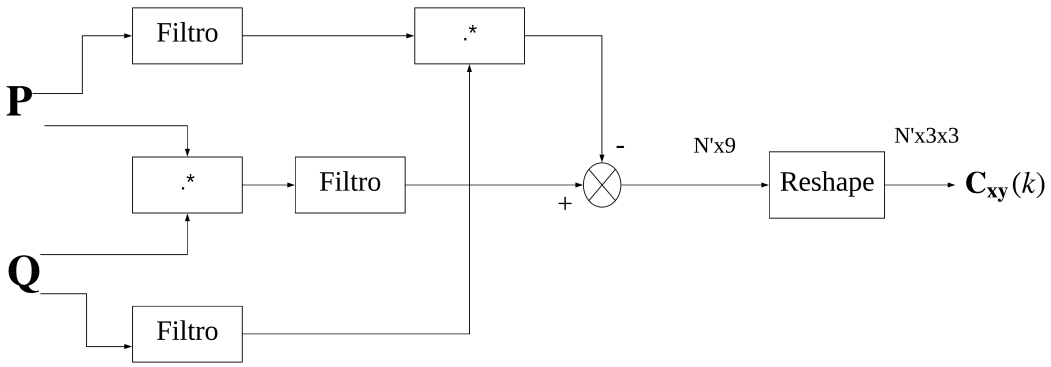

Figura 3.9: Diagrama que representa o processamento do filtro média móvel exponencial sobre as amostras para encontrar o valor da matriz $\boldsymbol{C_{xy}}(k)$ ao longo do intervalo $[1 : N]$. O bloco "Filtro"faz referência ao filtro com fator de decaimento exponencial h_{filt} definido em [2]

Em casos reais, talvez os grupos de séries temporais não estejam em fase e deve-se considerar o valor da associação instantânea considerando um atraso d entre os grupos. Assim, a covariância instantânea passa a ser uma função tanto do instante de tempo k quanto do atraso d definido matemáticamente como [2]

$$cov_{xy}(k,d) = \sum_{l=0}^{\infty} ce^{-\eta|l|} x((k-d/2)-l)y((k+d/2)-l) - \\ [\sum_{l=0}^{\infty} ce^{-\eta|l|} x((k-d/2)-l)][\sum_{l=0}^{\infty} ce^{-\eta|l|} y((k+d/2)-l)]. \tag{3.42}$$

Como a função nos permite avaliar a covariância para todas as combinações possíveis das variáveis independentes (k,d), foi escolhida uma representação por meio de mapa de calor para visualizar o resultado, onde a cor representa o valor da associação, o eixo das abscissas o valor de k o eixo das ordenadas o valor de d [2]. Por meio de uma análise deste mapa de calor, além do valor da associação, é possível encontrar o atraso entre os domínios bem como descrever como este varia ao longo do tempo.

O processo de construção do mapa pode ser ilustrado com o auxílio da Figura 3.10 onde o defasamento entre os grupos varia de $-d_{max}$ até $+d_{max}$ e cada cor representa um valor de atraso. Quando a análise é realizada para os grupos em fase ($d_{max} = 0$), o número de valores de associação possíveis é igual ao número de amostras N disponíveis. Todavia, quando o mapa é gerado e valores de atraso entre grupos são considerados, o número de valores possíveis de associação cai para $N' = N - d_{max}$ uma vez que o número de amostras são finitos, conforme é descrito na Figura 3.10. Por fim, as associações entre os grupos defasados são armazenadas em uma matriz, conforme ilustrado na parte inferior da Figura.

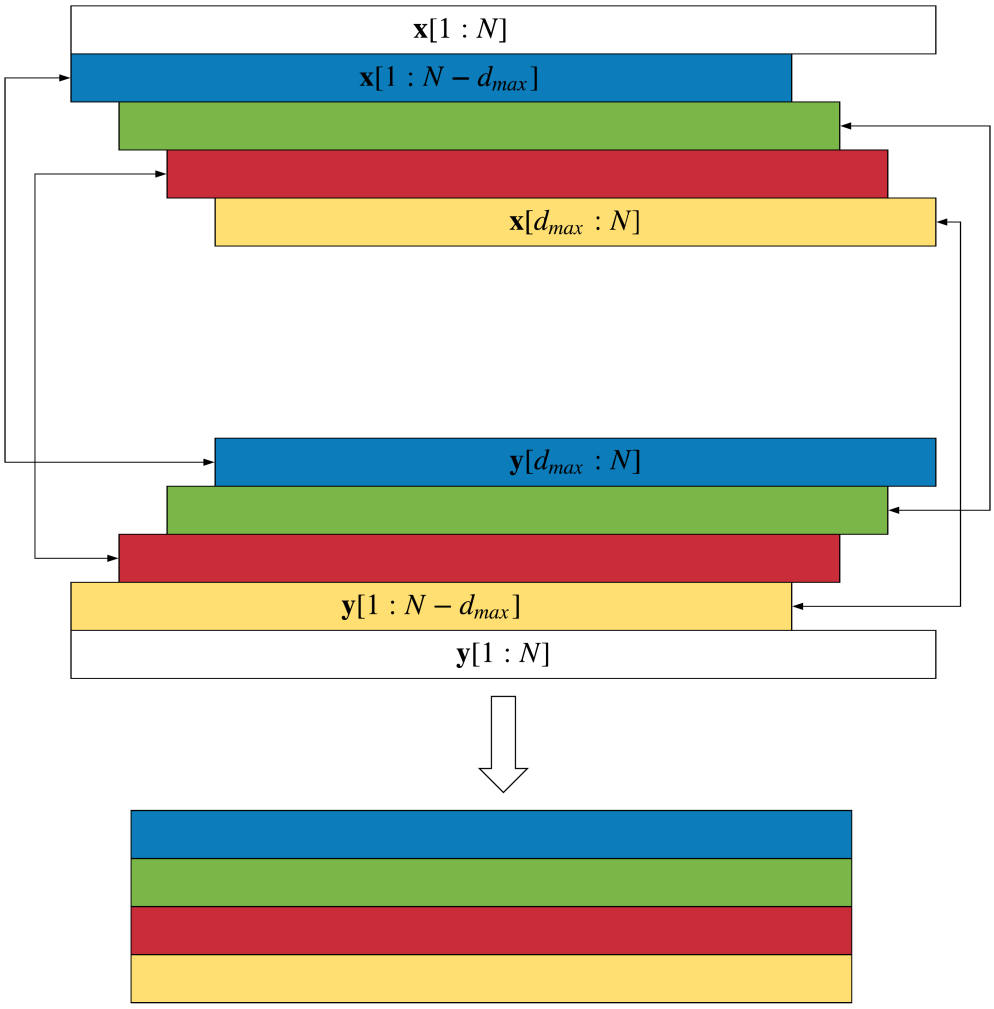

Figura 3.10: Representação de como é obtido um mapa de associação a partir da correlação instantânea entre os grupos.

3.4 A associação entre grupos de variáveis

O método desenvolvido ao longo deste capítulo pode ser consolidado pelo Algoritmo 5, que é composto por três etapas

- Estabelecer os trechos de séries temporais de cada grupo que serão utilizados para avaliar a associação instantânea para cada valor de atraso. Na Figura 3.10, por exemplo, os limites inferiores de $\boldsymbol{x}$ e $\boldsymbol{y}$ vão de $1 \rightarrow d_{max}$ e $d_{max} \rightarrow 1$ respectivamente. Caso $d_{max} = 0$, o mapa se torna um gráfico de associação instantânea e é chamado de *Associação 1D*. Caso a análise considere atrasos entre os grupos, a medida passa a ser representada em um mapa de calor e é chamada de *Associação 2D* [2].

- Para cada valor de atraso, calcular o valor das matrizes de covariância instantâneas $\boldsymbol{C_{xx}}(k)$, $\boldsymbol{C_{yy}}(k)$ e $\boldsymbol{C_{xy}}(k)$.

- Para cada ponto do intervalo de tempo analisado ($1 \rightarrow N'$), onde $N' = N - d_{max}$, estimar o valor da associação. No algoritmo, a função que calcula a associação $assocfunc()$ faz referência a qualquer uma das medidas de associação discutidas ao longo deste capítulo (v^{pca}, v_{cca}, h,c), selecionada de acordo com a aplicação de interesse.

Comentários Finais

Neste capítulo, foram apresentadas algumas formas para se medir a associação entre grupos de variáveis bem como suas implementações. Elas são baseadas em três princípios: variância compartilhada, máxima correlação e probabilidade dos grupos estarem descorrelacionados. Também foi descrito como a associação se torna instantânea quando as matrizes de covariância são variantes no tempo e como a ferramenta pode ser configurada para detectar avanços e atrasos entre os grupos.

Algoritmo 5 Associação Instantânea Multivariada

Input: Matrizes $\boldsymbol{X}$ e $\boldsymbol{Y}$ e o atraso máximo d_{max}

$N' \leftarrow N - d_{max}$
$\boldsymbol{Assoc} \leftarrow zeros[d_{max}][N']$
$\boldsymbol{x}_{init} \leftarrow 1:1:d_{max}$
$\boldsymbol{y}_{init} \leftarrow d_{max}:-1:1$
$\boldsymbol{x}_{fim} \leftarrow N':1:N$
$\boldsymbol{y}_{fim} \leftarrow N:-1:N'$
for $i \leftarrow 1:d_{max}$ **do**
  $\boldsymbol{X'} = \boldsymbol{X}[x_{init}[i]:x_{fim}[i]][]$
  $\boldsymbol{Y'} = \boldsymbol{Y}[y_{init}[i]:y_{fim}[i]][]$
  $\boldsymbol{C}_{x'x'} \leftarrow covinst(\boldsymbol{X'}, \boldsymbol{X'})$
  $\boldsymbol{C}_{x'y'} \leftarrow covinst(\boldsymbol{X'}, \boldsymbol{Y'})$
  $\boldsymbol{C}_{y'y'} \leftarrow covinst(\boldsymbol{Y'}, \boldsymbol{Y'})$
  for $j \leftarrow 1:N'$ **do**
    $\boldsymbol{A}_{xy}[i][j] \leftarrow assocfunc(\boldsymbol{C}_{x'x'}[j][][], \boldsymbol{C}_{x'y'}[j][][], \boldsymbol{C}_{y'y'}[j][][])$
  end for
end for

Output: A_{xy} - Associação instantânea entre os grupos

Capítulo 4

Processamento Audiovisual da Fala

Como mencionado diversas vezes ao longo do trabalho, uma das aplicações de interesse é estimar a interação variante no tempo entre os movimentos da face, do trato vocal e a acústica da fala. Assim, para descrever como os parâmetros que representam a acústica da fala são extraídos a partir do sinal, é imprescindível descrever como funciona a produção da fala humana. Logo, na primeira seção deste capítulo será apresentado o modelo fonte filtro, que é a base para compreender o mecanismo de codificação e síntese da fala.

Na segunda seção são apresentados os Coeficientes de Predição Linear (*Linear Predictive Coding - LPC*). Este é talvez um dos modelos mais conhecidos e bem sucedidos para a representação da acústica da fala. Na terceira seção é apresentado um segundo método para caracterização da acústica da fala conhecido como *Line Spectrum Pairs - LSP*, que apresenta algumas vantagens em relação ao LPC como maior robustez no momento da codificação dos coeficientes [23] e curvas mais suaves quando se acompanha como os valores dos coeficientes variam ao longo do tempo. Este último foi o método escolhido neste trabalho para se extrair os dados de acústica da fala na Base 2.

Ao final deste capítulo, serão apresentados os vetores autoregressivos,

que podem ser utilizados em uma etapa de pré-processamento para extrair a dinâmica dos movimentos do trato vocal e da face. A utilidade de tal remoção, como será explicado, vem do fato de que a mecânica dos movimentos ocasiona uma alta correlação entre as amostras do sinal, o que pode apresentar um certo atraso para o método detectar variações no valor da associação instantânea.

4.1 Produção da fala humana e o modelo fonte-filtro

O sistema de produção de fala humana é composto pelos seguintes órgãos do corpo humano: os pulmões, que são a fonte de ar que excita o trato e as cordas vocais; a traquéia, que conduz o ar do pulmão até as cordas vocais; as cordas vocais, que vibram quando são excitadas pelo fluxo de ar; a cavidade nasal e o trato vocal.

Conforme descrito em [24, 25], o processo de síntese de fala humana pode ser representado por um modelo composto de um sistema linear variante no tempo (filtro) acionado por um sinal de excitação (fonte).

Os tipos de sinal de excitação gerados pela fonte podem ser classificados em três tipos. O primeiro ocorre normalmente quando um sinal vozeado está sendo pronunciado, onde o ar flui dos pulmões e é modulado pela vibração das cordas vocais, resultado em um trem de pulsos periódicos. No segundo tipo, o ar sai do pulmão e é comprimido no trato vocal, resultando em um sinal de excitação parecido com um ruído branco, criando os fonemas conhecidos como não vozeados. Uma terceira forma de excitar o trato vocal consiste no impedimento total de saída do ar. Logo depois, o ar é liberado por uma súbita abertura do trato vocal e um som transiente é criado. Em resumo, o sinal de excitação pode ser aproximado por um sinal de pulsos periódicos, quando o locutor está pronunciando um sinal vozeado e um ruído branco, quando o locutor está pronunciando um sinal não vozeado [24].

O filtro descreve o comportamento do trato vocal, sendo este último

composto pela faringe e pela cavidade oral [24] e uma parte de suma importância no processo de produção da fala humana, pois sua área seccional, definida pelas posições da língua, lábios, mandíbula e palato muscular, varia ao longo do tempo e está condicionada ao fonema que está sendo pronunciado pelo locutor. Esta gera modos de ressonância cujas frequências características são chamadas de formantes. Como será apresentado na seção que descreve o LPC, os formantes caracterizam o filtro.

Assim, seguindo a premissa do modelo, dado o sinal de excitação $u_f(t)$ e a resposta ao impulso descrita pelo trato vocal $v_f(t)$, com suas respectivas transformada de Fourier $V_f(j\omega)$, o sinal de fala pode ser recriado pela convolução entre a fonte e o filtro

$$s_f(t) = u_f(t) * v_f(t) \tag{4.1}$$

ou em uma representação no domínio da frequência, onde a operação de convolução se torna um produto

$$S_f(j\omega) = U_f(j\omega)V_f(j\omega) \tag{4.2}$$

Como mencionado no começo desta seção, assim como o comportamento da fonte varia de acordo com o tipo de fonema pronunciado, a configuração do trato vocal também está relacionada ao fonema. Para captar o comportamento variante no tempo, o sinal de fala costuma ser dividido em quadros de aproximadamente $20ms$ (intervalo de tempo dentro do qual este sinal pode ser considerado estacionário) e uma análise de Fourier é aplicada sobre cada quadro. Tal método também é conhecido como *Short-Time Fourier Transform* (STFT) e é implementado com o auxílio de uma janela [26]. Por este motivo, tal operação pode ser representada pela seguinte equação [24]

$$S_c(\tau, \omega) = \int_{-\infty}^{\infty} w(\tau - t)s_f(t)e^{-j\omega t}dt \tag{4.3}$$

onde $w(t)$ representa a janela que está sendo utilizada. Como neste trabalho os sinais estão representados no tempo discreto, uma melhor re-

presentação seria

$$S_f(e^{j\omega}) = \sum_{-\infty}^{\infty} w[n-m]s_f[m]e^{j\omega m} \tag{4.4}$$

Até o presente momento, nada foi dito sobre como a fonte e o filtro são separados dado um determinado sinal de fala. Na próxima seção será apresentada uma abordagem para tal problema.

4.2 Codificação da acústica da fala e separação fonte-filtro pelos coeficientes LPC

Conforme descrito na seção anterior, o sinal de fala é analisado em quadros com duração de aproximadamente $20ms$. Isto acontece pois considera-se que a configuração do trato vocal é constante ao longo de cada um destes segmentos. Consequentemente, pode-se dizer que a amplitude de um sinal de fala no instante k está correlacionada com as amostras de instantes anteriores. Neste cenário, o valor de um sinal de fala no instante k pode ser dividido em duas parcelas [25]

$$s_f(k) = \hat{s}_f(k) + \epsilon(k) \tag{4.5}$$

onde $\hat{s}_f(k)$ é o sinal de saída de um modelo autoregressivo definido como

$$\hat{s}_f(k) = \sum_{i=1}^{p} \alpha_i s_f(k-i) \tag{4.6}$$

de acordo com [24], $\epsilon(k)$ representa o sinal de excitação com a adição de um ganho G, ou,

$$\epsilon(k) = Gu_f(k) \tag{4.7}$$

se a equação 4.2 for transformada para o domínio Z encontra-se

$$S_f(z) - \sum_{i=1}^{p} \alpha_i z^{-i} = GU_f(z) \tag{4.8}$$

e a resposta do filtro é dada por

$$V_f(z) = \frac{G}{1 - \sum_{i=1}^{p} \alpha_i z^{-i}} \tag{4.9}$$

onde o modelo de predição é definido pelos parâmetros $\boldsymbol{\alpha}$ conhecidos como coeficientes *LPC* e estes caracterizam a acústica da fala para o quadro analisado. Os coeficientes são encontrados por uma solução de mínimos quadrados [17]. O sinal $u_f(k)$, ou resíduo, representa o sinal de excitação que passa posteriormente pelo trato vocal. Quando o fonema que está sendo pronunciado pelo locutor é um sinal vozeado, as cordas vocais são excitadas e o sinal $u_f(k)$ consiste em um trem de pulsos com período quase constante [27]. Nesse caso, considerando que o número de coeficientes p utilizado seja suficiente, o sinal de excitação será praticamente um trem de impulsos unitários e, assim, toda a informação estará no modelo do filtro. Quando um sinal não vozeado é pronunciado, o sinal produzido pela fonte se assemelha muito a um ruído branco. Uma vez que o valor esperado de uma regressão linear seja algo próximo de ruído branco [4], a aproximação por um modelo autoregressivo é perfeitamente plausível.

Apesar de o estimador de mínimos quadrados ter sido definido como a maneira de se encontrar os coeficientes, existem duas maneiras diferentes de definir os regressores para encontrar a matriz de correlação que a regressão linear necessita [24]: o método da autocorrelação e o método da covariância.

O primeiro consiste na aplicação de uma janela sobre o quadro que se deseja extrair os parâmetros. Desta forma, aplica-se o valor nulo para todos os instantes de tempo que se encontram fora do intervalo do quadro. Assim, os parâmetros são estimados sobre o sinal

$$s_f(k) = s_f(k)w(k) \tag{4.10}$$

onde $w(k)$ é uma janela com valores não nulos do instante $k = 0$ até o tamanho da janela L. Por simplicidade de notação, foi considerado que a amostra inicial do quadro é $k = 0$. Tal método nunca apresenta erro

nulo, pois para se estimar o valor do sinal no instante $k = 0$ é necessário que existam amostras de instantes anteriores. Assim, mesmo se o modelo representar totalmente o sinal, a variância do erro nunca será nula.

No caso do método da covariância, para se evitar o tipo de erro apresentado pelo primeiro método, são utilizadas amostras de instantes de fora do quadro para se extrair os parâmetros. Matematicamente, isso significa que para se estimar o valor do sinal no instante $k = 0$, são utilizadas amostras dos instantes $k = -p, ..., -1$, onde p é o número de parâmetros do modelo autoregressivo.

4.3 Coeficientes LSP

Reconsidere o denominador da função de transferência do filtro LPC

$$H_f(z) = \frac{G}{1 - \sum_{i=1}^{p} \alpha_i z^{-i}} \tag{4.11}$$

caso alguns pólos estejam muito próximos do círculo unitário, um erro de quantização pode levar a um polo ficar na região de instabilidade. Por isso, foi criado o *line spectrum pairs* (LSP). Neste caso, um novo conjunto de parâmetros foi criado no domínio da frequência, composto de ângulos

$$(\omega_1, \theta_1, \omega_2, \theta_2, ..., \omega_{p/2}, \theta_{p/2})$$

Estes ângulos são as raízes dos polinômios

$$F_1(z^{-1}) = A_p(z^{-1}) - z^{-(p+1)} A_p(z) = 1 + (\alpha_1 - \alpha_p) z^{-1} + \cdots + \tag{4.12}$$

$$(\alpha_p - \alpha_1) z^{-p} - z^{-(p+1)} \tag{4.13}$$

e

$$F_2(z^{-1}) = A_p(z^{-1}) + z^{-(p+1)}A_p(z) = 1 + (\alpha_1 + \alpha_p)z^{-1} + \cdots + \quad (4.14)$$

$$(\alpha_p + \alpha_1)z^{-p} - z^{-(p+1)} \quad (4.15)$$

Ambos os polinômios somente possuem coeficientes reais, logo os polos são complexos conjugados. Assim, se uma raiz do polinômio é $e^{j\omega}$ a outra é $e^{-j\omega}$. Sobre o módulo das raízes, todas elas se encontram no círculo unitário. Em ambas as equações, A_p é o polinômio definido em [23] como

$$A_p(z^{-1}) = 1 - \sum_{i=1}^{p} \alpha_i z^{-i} \quad (4.16)$$

4.4 Os vetores autoregressivos

Como referido em livros de estatística, para otimizar os resultados, os dados de entrada da PCA e da CCA devem possuir o comportamento próximo de uma gaussiana e isto normalmente ocorre quando existe descorrelação entre amostras. Além disso, uma eliminação da redundância entre as amostras pode levar uma melhoria na performance do método pois reduziria o atraso do método ao detectar variações sobre a associação instantânea. Por este motivo, nesta seção serão apresentados os vetores autoregressivos, com o objetivo de remover informação redundante entre as amostras.

Sistemas físicos apresentam elementos derivativos e integrativos e seu comportamento dinâmico resultante gera informação redundante entre as amostras. Neste caso

$$E[x(k)|x(k-1), x(k-2), x(k-3), \cdots, x(0)] \neq \mu_x \quad (4.17)$$

pois os valores das variáveis nos instantes anteriores influenciam diretamente o valor esperado da variável no instante k. Nessa situação, conforme descrito em [28], pode-se estimar o valor da série no instante k a

partir de instantes anteriores

$$E[x(k)|x(k-1),\cdots,x(0)] = \hat{x}(k) \tag{4.18}$$

e calcular as matrizes de covariância sobre a parte da amostra $\boldsymbol{x}(k)$ que não pode ser estimada pelo instante anterior, $\epsilon_x(k)$, definida matematicamente como sendo

$$x(k) - \hat{x}(k) = \epsilon_x(k) \tag{4.19}$$

Como este trabalho calcula a associação entre grupos de variáveis, o problema se torna encontrar o vetor $\hat{\boldsymbol{x}}(k)$ ao invés de cada variável separadamente

$$\hat{\boldsymbol{x}}(k) = E(\boldsymbol{x}(k)|\boldsymbol{x}(k-1),\cdots,\boldsymbol{x}(0)) \tag{4.20}$$

$$\boldsymbol{x}(k) - \hat{\boldsymbol{x}}(k) = \boldsymbol{\epsilon}_x(k) \tag{4.21}$$

e a nova matriz de covariância é representada como

$$E[\boldsymbol{\epsilon}_x\boldsymbol{\epsilon}_x^T] = \boldsymbol{C}_{\epsilon_x\epsilon_x} \tag{4.22}$$

O problema de se encontrar $\hat{x}(k)$ não é trivial e é estudo da área de identificação de sistemas [17] e dois tipos de representações de sistemas devem ser ressaltadas: representações lineares e não lineares. Modelos não lineares não foram considerados, pelo fato de que os movimentos da face e do trato vocal são mecânicos e, normalmente, são representados por sistemas lineares de primeira ou de segunda ordem. Por este mesmo motivo, foram escolhidos os vetores autoregressivos. O modelo é caracterizado

pelas matrizes de parâmetros $\boldsymbol{\psi}$

$$\begin{vmatrix} y_1(k) \\ y_2(k) \\ \vdots \\ y_{n_y}(k) \end{vmatrix} = \boldsymbol{\Psi}_1 \begin{vmatrix} y_1(k-1) \\ y_2(k-1) \\ \vdots \\ y_{n_y}(k-1) \end{vmatrix} + \boldsymbol{\Psi}_2 \begin{vmatrix} y_1(k-2) \\ y_2(k-2) \\ \vdots \\ y_{n_y}(k-2) \end{vmatrix} + \cdots + \boldsymbol{\Psi}_L \begin{vmatrix} y_1(k-L) \\ y_2(k-L) \\ \vdots \\ y_{n_y}(k-L) \end{vmatrix} + \begin{vmatrix} \epsilon_1(k) \\ \epsilon_2(k) \\ \vdots \\ \epsilon_{n_y} \end{vmatrix} \tag{4.23}$$

que buscam minimizar a variância dos resíduos $\boldsymbol{\epsilon}$. Na equação acima, L é o número máximo de instantes anteriores das séries temporais que são adicionados como entrada ao modelo. A solução do problema é encontrada pelo estimador de mínimos quadrados descrita em [17].

Comentários Finais

Neste capítulo foram apresentados um resumo do processo de produção de fala humana, uma vez que este é importante para a interpretação dos resultados que serão apresentados neste trabalho. Eles também são base para os métodos LPC e LSP, que foram detalhados nas seções seguintes. Eles foram utilizados para a caracterização da acústica da fala na base 2.

Ao fim do capítulo foram apresentados os métodos utilizados para remover a informação redundante entre as amostras que caracterizavam os movimentos da face e do trato vocal, para auxiliar a deixar as séries temporais com um comportamento mais próximo de uma gaussiana, podendo eventualmente melhorar a performance das medidas de associação.

Capítulo 5

Arbitragem Estatística

5.1 Introdução

No mercado financeiro, a previsão de preços de ativos talvez seja um dos problemas mais complexos estudados por estatísticos, engenheiros e economistas. Por conta do comportamento aleatório das séries temporais de preços de ações, alguns pesquisadores começaram a realizar estudos sobre como se comportaria um novo ativo obtido pela combinação linear de outros. A esperança era que este apresentasse um comportamento mais previsível ou que, pelo menos, as perdas dos investidores fossem menos significativas. Para tal combinação de ativos é dada o nome de portfólio. Dentre os tipos de portifólios possíveis, um é de especialmente importante para neste capítulo: os portfólios neutros em relação ao mercado, ou *Long-Short Portfolios*, que são caracterizados por sofrerem uma baixa influência do mercado. Sobre estes, investidores aplicam uma técnica de *trading* conhecida como arbitragem, tema de estudo deste capítulo. Tal método está intimamente ligado ao conceito de *Cointegração*. Dois ativos são considerados cointegrados caso seja possível estabelecer uma combinação linear de forma que o resultado seja um processo estocástico reversível a média (como média móvel, autoregressivo ou ruído branco). Como séries temporais reversíveis a média apresentam previsibilidade em seu comportamento, investidores tentam obter lucro negociando em cima de

tal comportamento.

Como será apresentado, os processos de busca de pesos ótimos para esta combinação linear estão intimamente relacionados com as medidas de associação entre grupos de variáveis descritas em capítulos anteriores. Para instigar o leitor nesse momento, pode-se adiantar que o método mais conhecido na teoria de portfólio moderno encontra os pesos do portfólio por meio da análise da matriz de covariância das séries temporais dos retornos. No caso de um portfólio contendo somente dois ativos, a melhor solução para encontrar os pesos é uma regressão linear, que como apresentado em capítulos anteriores, pode ser vista como uma análise em componentes canônicas onde cada um dos grupos possui somente uma variável.

5.2 Modelagem de preços de ações e teoria do mercado eficiente

Um mercado eficiente é definido por [29] como

A capital market is said to be efficient if it fully and correctly reflects all relevant information in determining security prices. Formally, the market is said to be efficient with respect to some information set, Ω_t , if security prices would be unaffected by revealing that information to all participants. Moreover, efficiency with respect to an information set, Ω_t , implies that it is impossible to make economic profits by trading on the basis of Ω_t .

Dentro desta passagem, o autor realiza menção constante ao conjunto de informações Ω_t. De fato, grande parte da informação do mercado financeiro pode ser obtida por qualquer pessoa em tempo real. Isso gera certa dificuldade em realizar modelos de predição confiáveis, pois é possível que outros analistas estejam realizando modelos com respostas muito semelhantes e que isto leve a uma mudança do sistema no instante seguinte[30], levando os preços de ações a possuírem um comportamento aleatório. Por este motivo, os preços de ações são modelados como um processo Browniano, onde as variações percentuais sobre os preços pos-

suem um comportamento normal. Em outras palavras

$$\frac{dp_x(t)}{p_x(t)} = \sigma(t)dB_t \tag{5.1}$$

que em tempo discreto se torna

$$r_x^{\%}(k) = \frac{\Delta p_x(k)}{p_x(k-1)} = \sigma(k)u(k) + \mu_r \tag{5.2}$$

onde $u(k) \sim \mathcal{N}(0,1)$. Tal processo possui um valor médio de retorno (μ_r) nula e uma determinada variância ($\sigma(k)$), que em séries econômicas e financeiras é denominada volatilidade. Em outra forma de se representar o mesmo problema, o retorno pode ser matematicamente definido como a primeira diferença sobre o logaritmo da série temporal dos preços de uma determinada ação, ou

$$r_x^{log}(k) = ln(p_x(k)) - log(p_x(k-1)) = ln(\frac{p_x(k)}{p_x(k-1)}) \tag{5.3}$$

Considerando que os instantes k e $k+1$ são muito próximos, a seguinte normalização pode ser aplicada

$$r_x^{log}(k) = ln(1 + r_x^{\%}(k)) \approx r_x^{\%}(k) \tag{5.4}$$

Assim, as duas definições são equivalentes. Quando se utiliza a primeira diferença do logaritmo ganha-se precisão, pois o sinal fica mais próximo de uma distribuição normal, entretanto, tovadia ocorrem erros gerados pelo processo de linearização.

Trabalhar sobre as variações pode aumentar o nível de ruído, pois a operação de primeira diferença funciona como um filtro passa-altas. Isto pode ser ilustrado pela a representação no domínio da frequência do operador de primeira diferença dada por

$$H(z) = 1 - z^{-1} \tag{5.5}$$

ou

$$H(e^{j\omega}) = 1 - e^{-j\omega} \tag{5.6}$$

O processo apresenta ganhos próximos a zero para as baixas frequências e ganhos próximos a 2 para a valores de frequência próximas a π, um filtro passa altas. Como a maior parte do rúido se encontra nas componentes de alta frequência, um corte nas componentes de baixa frequência reduz a relação sinal-ruído.

Uma forma de melhorar a relação sinal-ruído é adicionar alguma redundância entre as amostras. Neste caso, mesmo que algumas premissas exigidas pela estatística sejam violadas, o sistema seria capaz de contornar pequenas oscilações temporais existentes entre as séries e, consequentemente, a medida seria mais robusta contra ruído.

Logo, pode-se considerar que se a granularidade for muito baixa, o modelo de mínima variância apresentará resultados mais robustos quando aplicado sobre o produtório acumulado dos retornos, que será nomeado neste trabalho como retorno cumulativo ou retorno geométrico. Pode-se encontrar o valor do retorno cumulativo de forma recursiva por

$$r(k) = r(k-1)(\Delta^{\%} + 1) \tag{5.7}$$

e após n períodos de tempo, o retorno percentual total é encontrado por

$$\Delta\% = \frac{p(n)}{p(1)} - 1 \tag{5.8}$$

e o retorno geométrico médio do período pode ser estimado pela regra de juros compostos

$$\mu = \sqrt[N-1]{\frac{p(N)}{p(1)}} - 1 \tag{5.9}$$

Pode-se facilmente provar que calcular o retorno cumulativo é equivalente ao logaritmo da série temporal de preços e que ambas as abordagens são válidas. Por exemplo, em [12], o autor opta por realizar a primeira abordagem, enquanto em [13], o autor opta por utilizar o logaritmo da série dos preços.

Outro motivo para se comprovar tal escolha é descrito na passagem

presente em [31] :*If you lose one-third, or* 33.33% *of your assets, you will have to make* 50% *on your remaining assets to break even. If you make* 50% *first, a loss of* 33.33% *will bring you back to your starting level.* Tal exemplo demonstra que calcular a média aritmética dos retornos não é a melhor forma de se computar o retorno médio de um portfólio, uma vez que, a ordem na qual os ganhos e as perdas ocorrem influenciam o resultado final.

5.3 Arbitragem Estatística

Arbitragem estatística é uma das técnicas de *trading* mais utilizadas no mercado de capitais. A técnica consiste em extrair uma série reversível à média a partir da combinação linear de ativos e, a partir do uso de tal série, aumentar os lucros sobre operações de compra e venda de ações, pois uma série reversível à média é mais previsível que a série temporal do preço de um determinado ativo individualmente [32, 11]. O comportamento reversível a média somente é possível pelo fato do preço destes ativos serem influenciados por uma série de fatores em comum, exceto por fatores específicos, e assim uma parte das tendências do mercado são eliminadas.

Para exemplificar, considere duas séries temporais de preço nomeadas $p_x(k)$ e $p_y(k)$. A série estacionária, chamada neste trabalho de *spread*, pode ser encontrada pela combinação

$$s_p(k) = p_y(k) - \beta p_x(k) \tag{5.10}$$

onde

$$s_p(k) \sim \mathcal{N}(\mu(k), \sigma(k)) \tag{5.11}$$

O método de como aumentar as chances de lucro a partir de tal propriedade matemática é simples. Quando o *spread* está com um valor um pouco acima do seu valor médio, as duas causas mais provaveis são: $p_y(k)$ se encontra precificado um pouco acima do esperado e/ou $p_x(k)$ se encontra precificado um pouco abaixo do esperado. Neste caso, a expec-

tativa é de que a longo prazo o valor do *spread* retorne a média, ou seja, que o valor relativo de $p_y(x)$ caia e/ou o valor relativo de $p_x(k)$ suba. Assim, o *trader* deve comprar $p_x(k)$ e realizar uma venda descoberta sobre $p_y(k)$.

Caso o leitor não esteja familiarizado com o tema, uma venda descoberta é uma operação no mercado financeiro que o investidor realiza quando acredita que o preço de um papel irá cair. Neste caso, no momento da operação, ele fica com um saldo positivo referente ao preço do papel de ativo e se compromete a comprar o papel para cobrir o saldo negativo após um determinado período de tempo. O processo pode ser visto como uma espécie de empréstimo.

Em [33], é apresentado um modelo de como é possível modelar a série $s_p(k)$ em um modelo autoregressivo. Neste caso, o *spread* é representado como

$$s_p(k+1) - s_p(k) = (a - bs_p(k))\tau(k) + \sigma\sqrt{\tau(k)}\epsilon_{k+1} \tag{5.12}$$

O valor do erro de predição do processo autoregressivo está relacionado com a diferença de tempo entre as amostras, $\tau(k)$. Assim, pode-se dizer que cada vez maior o intervalo de tempo entre as amostras da série temporal, maior o erro de predição.

Outra questão interessante é que como o *spread* possui variância constante a curto prazo, esta propriedade pode ser aproveitada para se extrair informações sobre momentos de compra e venda. Suponha, por exemplo, que a série atingiu o desvio padrão σ_s. Provavelmente, em algum momento posterior, por conta de ser reversível a média, existie uma alta probabilidade do *spread* retornar a sua esperança matemática μ_s. Assim, pode-se realizar uma venda do portfólio e comprar os ativos novamente quando o valor estiver abaixo da média, em $-\sigma_s$, por exemplo.

A escolha dos papéis e a definição do peso β está ligada ao conceito de cointegração, que será apresentado na próxima seção.

5.4 Cointegração

O conceito de cointegração é extremamente importante na área de arbitragem estatística e, por isso, foi criada uma seção somente para descrevê-lo. Tal fenômeno foi primeiramente descrito por Engle e Granger em [34]. Considere duas séries temporais de preços de dois ativos: $p_x(k)$ e $p_y(k)$. As duas séries são ditas cointegradas se existe uma combinação linear que dê origem a uma série $s_p(k)$ que seja estacionária em sentido amplo. Matematicamente, o problema consiste em determinar um coeficiente β onde

$$s_p(k) = p_y(k) - \beta p_x(k) \tag{5.13}$$

e

$$s_p(k) \sim \mathcal{N}(\mu_s, \sigma_s) \tag{5.14}$$

Conforme descrito em [13], séries se tornam cointegradas por terem seus preços influenciados por fatores comuns e por este motivo, a longo prazo, os preços das mesmas continuarão a se movimentar juntas. Engle e Granger propuseram representar tal fenômeno por um modelo definido pelos parâmetros α e β e pelo seguinte código corretor de erro

$$p_y(k) - p_y(k-1) = \alpha_y(p_y(k-1) + \beta p_x(k-1)) + \epsilon(y_k) \tag{5.15}$$

$$p_x(k) - p_x(k-1) = \alpha_x(p_y(k-1) + \beta p_x(k-1)) + \epsilon(x_k) \tag{5.16}$$

Em outras palavras, um *spread* gerado a partir da combinação linear dos ativos $p_x(k)$ e $p_y(k)$ é estacionário em sentido amplo quando se encontram parâmetros dentro dos intervalos $\alpha_x < 1$ e $\alpha_y < 1$. Consequentemente, tal série será reversível a média, ou seja, ela tende a retornar a um valor médio μ_s. Conforme descrito na seção anterior, a série temporal

do preço de uma ação pode ser representada como um processo browniano puro, e, consequentemente, se duas séries temporais de preços são cointegradas, as séries temporais dos logaritmos dos preços também são cointegradas, pois aplicar o logaritmo consiste somente em uma normalização das séries temporais. Assim, os códigos corretores de erros podem ser igualmente aplicados em

$$l_y(k) - l_y(k-1) = \alpha_y(l_y(k-1) + \beta l_x(k-1)) + \epsilon(y_k) \tag{5.17}$$

$$l_x(k) - l_x(k-1) = \alpha_x(l_y(k-1) + \beta l_x(k-1)) + \epsilon(x_k) \tag{5.18}$$

onde

$$l_y(k) = log(p_y(k)) \tag{5.19}$$

e

$$l_x(k) = log(p_x(k)) \tag{5.20}$$

Outra forma de definir cointegração é analisando a possibilidade de cointegração pela abordagem descrita em [35], chamada de *commom trends model*, que afirma que séries temporais financeiras podem ser separadas em duas componentes, uma determinística determinada por uma série de indicadores e uma completamente aleatória, representada matematicamente por um ruído branco, ou

$$l_y(k) = \hat{l}_y(k) + \epsilon_y(k) \tag{5.21}$$

$$l_x(k) = \hat{l}_x(k) + \epsilon_x(k) \tag{5.22}$$

Para existir cointegração e encontrar um *spread* $s(k)$ estacionário, é necessário que os termos determinísticos se anulem. Em outras palavras,

dadas duas séries temporais

$$l_y(k) - \beta l_x(k) = (\hat{l}_y(k) + \beta \hat{l}_x(k)) + (\epsilon_y(k) + \beta \epsilon_x(k)) \tag{5.23}$$

deve-se encontrar um parâmetro β onde

$$(\hat{l}_y(k) + \beta \hat{l}_x(k)) = 0 \tag{5.24}$$

Além disso, outra inferência realizada por [35], que pode ser facilmente provada, é que o coeficiente de cointegração β pode ser encontrado pelos retornos dos processos brownianos puros, ou seja,

$$\hat{r}_y = \gamma \hat{r}_x \tag{5.25}$$

Extrair os termos determinísticos a partir de uma série de fatores não é uma tarefa tão trivial, todavia em [13], é apresentada uma metodologia para estimar estes termos baseada na teoria de precificação por arbitragem. Assim, se for assumido que o erro sobre a precificação do ativo é um ruído branco, uma simples regressão linear pode ser possível para definir o parâmetro β.

Logo, existem duas formas de se realizar a regressão linear. A primeira é diretamente sobre os processos brownianos puros, ou os logarítimos dos preços das ações, conforme feito em [12]. A segunda forma de se performar a regressão linear seria sobre os valores dos retornos.

Cada uma delas apresenta possui seus pontos positivos e negativos. Na regressão sobre o processo Browniano (ou Random Walk), as amostras estão correlacionadas, ferindo as premissas estatísticas necessárias pelo método. Todavia, mesmo que estas não possuam um comportamento gaussiano, a redundância entre as amostras pode acarretar em uma melhoria na relação sinal-ruído.

Quando se trabalha sobre os retornos, pelo fato destes estarem descorrelacionados, as premissas do teorema do limite central são respeitadas, mas há uma queda na relação sinal-ruído, pois o processo apresenta-se como um filtro passa-altas. Tal abordagem também transforma o pro-

blema em encontrar pesos de um portfólio de mínima variância, um tema estudado pela teoria de portfólio. Este será descrito na próxima seção deste capítulo.

5.5 Portfólio neutro em relação ao mercado

Conforme descrito em referências clássicas sobre a ciência do investimento, a função de um portfólio é dividir recursos de um investidor entre diversos ativos para atingir um determinado objetivo, que pode ser maximizar o retorno de uma determinada carteira ou minimizar a volatilidade da mesma. Neste caso, deseja-se dividir o dinheiro do investidor em ativos de forma que a soma do montante em todos os ativos atinja 100% ou

$$\sum_{i=1}^{N} \beta_i = 1 \tag{5.26}$$

Todavia, no caso da arbitragem estatística, o objetivo final é obter um portfólio neutro em relação ao mercado. Neste tipo de portfólio, os ativos são combinados para minimizar a influência do mercado [13]. Para auxiliar a ilustração do conceito, considere a equação da seção anterior, onde foi afirmado que para que dois ativos sejam cointegrados as componentes determinísticas de cada um dos sinais devem se anular, ou

$$(\hat{l}_y(k) + \beta \hat{l}_x(k)) = 0 \tag{5.27}$$

Neste caso, caso estes sejam cointegrados, grande parte da influência do mercado será neutralizada, pois a combinação linear deverá resultar em um ruído branco. Assim, se no mercado brasileiro o índice bovespa possuir uma queda muito brusca, o valor do spread estará descorrelacionado do evento.

Vale a pena ressaltar algumas peculiaridades neste tipo de portfólio. Em primeiro lugar, este tipo de portfólio deve apresentar, normalmente, pesos negativos e positivos. No caso de uma operação do tipo *daytrade*

na BMF & Bovespa, caso o investidor esteja com uma posição aberta de uma ação e uma venda descoberta da outra, os pesos não se anulam. Isso ocorre, pois o trader deve possuir dinheiro na sua conta para garantir uma porcentagem do valor para ambas as operações, mesmo para o caso de o risco ser altamente reduzido. Assim, a restrição dos pesos do portfólio pode ser redefinida como sendo

$$\sum_{i=1}^{N} |\beta_i| = 1 \tag{5.28}$$

No caso bivariado, os pesos do portfólio, pois dois ativos devem ser redimensionados como

$$s(k) = \frac{1}{(1+|\beta|)}\hat{l}_y(k) + \frac{|\beta|}{(1+|\beta|)}\hat{l}_x(k) \tag{5.29}$$

Para o caso multivariado, uma regressão linear simples não é suficiente para encontrar os pesos do portfólio. Por isso, nas próximas subseções, serão apresentados métodos para encontrar os pesos do portfólio quando forem considerados mais de dois ativos.

5.5.1 A teoria do portfólio moderno

O problema de se obter um ótimo portfólio consiste em encontrar o vetor $\boldsymbol{\beta}$ que minimiza a função de custo

$$f(\boldsymbol{\beta}) = \lambda \sum_{i=1}^{N}\sum_{j=1}^{N} \beta_i cov_{i,j} \beta_j + (1-\lambda)(-\sum_{i=1}^{N} \beta_i \bar{\mu}_i) \tag{5.30}$$

onde β_i é o montante investido em cada ativo, geralmente, normalizado de forma que

$$\sum_{i=1}^{N} |\beta_i| = 1 \tag{5.31}$$

$cov_{i,j}$ é a covariância entre os retornos dos ativos i e j, $\bar{\mu}_i$ é o valor esperado do retorno do ativo i e λ é uma constante que descreve o perfil

do investidor. O investidor pode optar por minimizar o risco $\lambda = 1$, ou maximizar o valor esperado dos retornos médios, $\lambda = 0$. A função de custo também pode ser descrita no seguinte formato matricial

$$f(\beta) = \lambda \boldsymbol{\beta}^T \boldsymbol{C} \boldsymbol{\beta} + \boldsymbol{\beta}^T \bar{\boldsymbol{\mu}} \tag{5.32}$$

Para a aplicação desejada neste trabalho, não existe qualquer interesse em maximizar o retorno médio do portfólio. Desta maneira, a partir deste momento, o valor λ será sempre unitário e o problema se concentra somente em encontrar os pesos que caracterizam o portfólio de mínima variância.

5.5.2 Value at Risk e Conditional Value at Risk

Nesta seção, serão descritas duas das principais técnicas de gestão de risco a partir de portfólios. Em [36] são apresentados as premissas e as características de cada método. Em ambos, a variável independente do modelo é o valor das perdas de cada ativo. Estas são obtidas por

$$l(k) = -r(k) \tag{5.33}$$

onde $r(k)$ é o valor do retorno. A equação acima nos diz que uma perda significativa pode ser interpretada como um retorno negativo de alta magnitude. Tal referência também introduz duas novas medidas de custo a serem otimizadas, α_δ e ϕ_δ, que buscam ser minimizadas pelos métodos $\beta - VAR$ e $\beta - CVAR$, respectivamente.

Antes de entrar em detalhes, primeiramente, deve-se introduzir a grandeza δ. Suponha que seja de interesse do investidor encontrar pesos que minimizem os valroes dos 5% das suas perdas mais significaticas. Desta forma, $\delta = 0.95$ e $1 - \delta = 0.05$, ou 5%. Uma vez definido o valor de δ, α_δ pode ser definido matematicamente como sendo

$$\alpha_\delta(\boldsymbol{x}) = min[\alpha \in \mathcal{R} : \Psi(\boldsymbol{x}, \alpha) \geq \delta] \tag{5.34}$$

onde

$$\Psi(\boldsymbol{x}, \delta) = \int_{r_p} p(\boldsymbol{r_p}) d\boldsymbol{r} \tag{5.35}$$

Onde r_p é o valor do retorno do portfólio, enquanto $\boldsymbol{r}$ é o vetor que descreve o valor dos retornos de cada um dos ativos pertencentes ao portfólio, separadamente. Como as amostras dos valores dos retornos são dadas, $\Psi(\boldsymbol{x}, \alpha)$ pode ser encontrado via uma Simulação de Montecarlo. Suponha que $\delta = 0.95$, como descrito anteriormente. Se possuímos um número de amostras $N = 100$, α_δ consiste no ponto de número 95 em uma escala crescente. O parâmetro ϕ_δ é definido matematicamente como

$$\psi_\delta(\boldsymbol{x}) = (1 - \delta)^{-1} \int_{r_p \geq \alpha_\delta(\boldsymbol{x})} r_p p(\boldsymbol{r}) d\boldsymbol{r} \tag{5.36}$$

Tal medida pode ser interpretada como sendo o valor esperado das perdas que se encontram nas $1 - \delta$ perdas mais significativas.

5.6 Portfólio e Análise em Componentes Canônicas

Nesta seção, será descrito como a CCA pode ser utilizada para se estabelecer um portfólio de mínima variância. Desta maneira, define-se uma ponte entre a teoria de portfólio moderno e as medidas de associação propostas neste trabalho.

No portfólio de *Markowitz*, uma vez escolhidos os ativos previamente, o problema que deseja-se resolver é o de encontrar o vetor de pesos $\boldsymbol{\beta}$ que minimiza a sefunção de custo

$$f(\boldsymbol{\beta}) = \boldsymbol{\beta}^T \boldsymbol{C} \boldsymbol{\beta} \tag{5.37}$$

respeitando a condição de que

$$\sum_{i=1}^{n} |\beta_i| = 1 \tag{5.38}$$

Após a solução do problema, os pesos possuem sinais diferentes e podem ser divididos em dois grupos. Os que apresentam sinal positivo (grupo $\boldsymbol{y}$) são aqueles sobre os quais as operações de compra serão efetuadas. Os que apresentam sinal negativo (grupo $\boldsymbol{x}$) são aqueles sobre os quais uma venda descoberta será realizada. Desta maneira, o *spread* pode ser definido como o resultado da associação entre dois grupos

$$s(k) = p_+(k) - p_-(k) \tag{5.39}$$

onde

$$p_+ = \boldsymbol{\beta}_y^T \boldsymbol{y} \tag{5.40}$$

e

$$p_- = \boldsymbol{\beta}_x^T \boldsymbol{x} \tag{5.41}$$

Na solução via CCA, o procedimento pode ser estabelecido da seguinte maneira: primeiramente é realizado um teste aplicando a CCA sobre todas as combinações de grupos de ativos possíveis até encontrar os grupos que apresentam o maior valor da primeira correlação canônica, ou de outra maneira, aqueles grupos que apresentam máxima correlação. Um alto valor de correlação normalmente está associado a um baixo valor de variância, e consequentemente, um baixo risco está associado com uma alta correlação entre os grupos de ativos. Considerando esta premissa, o *spread* pode ser definido como

$$s = u_1 - \rho v_1 \tag{5.42}$$

onde v_1 é o valor da primeira componente canônica relativa ao grupo $\boldsymbol{y}$, u_1 o valor da primeira componente canônica relativa ao grupo $\boldsymbol{x}$ e ρ o valor da primeira correlação canônica.

Entretanto, algumas adaptações devem ser feitas para que os resultados de ambos os métodos sejam comparados, dadas as suas singularida-

des. A primeira delas consiste no módulo dos pesos. No caso da análise em componentes canônicas, o objetivo é que tenham variância unitária. Por este motivo, ao fim de cada teste, os pesos obtidos pela análise em componentes canônicas devem ser normalizados de forma que

$$\sum_{i=1}^{n_x} |a_{i,1}| - |\rho| \sum_{j=1}^{n_y} |b_{j,1}| = 1 \tag{5.43}$$

como ocorre no portfólio de mínima variância.

A segunda restrição está relacionada com o sinal dos pesos e é um tópico mais delicado. Primeiramente, considere o caso onde os pesos são estáticos, ou seja, não variam ao longo do tempo. Na hora de estimá-los, para que o método seja robusto ao ruído, será aplicada uma validação cruzada onde diferentes grupos de treino e de teste serão utilizados. Pode ocorrer soluções onde o valor da correlação canônica é o mesmo, mas os pesos apresentam sinais invertidos. Por este motivo, foi fixado que o peso de um determinado ativo fosse sempre positivo. Tal ativo foi escolhido aleatoriamente dentre os ativos definidos a priori. Tal restrição também foi imposta aos métodos de portfólio utilizados.

Comentários Finais

Neste capítulo foi apresentada uma revisão da teoria de finanças quantitativas necessária para se acompanhar os estudos que serão apresentados nos capítulos de resultados. Primeiramente foi descrito a teoria do mercado eficiente e como ela define as premissas utilizadas para o pré-processamento das séries temporais financeiras.

Em um segundo momento foi descrito o que é arbitragem estatística e como esta técnica é utilizada por analistas de mercado financeiro para se obter lucro. Em seguida foram apresentados os conceitos de cointegração e portfólio neutro em relação ao mercado, que estão relacionados com o desenvolvimento de um bom algoritmo de arbitragem estatística. Por fim, foi descrito como é realizada a ponte entre a teoria de portfólio e a

medida de associação baseada na máxima correlação entre os grupos e como os métodos tiveram que ser modificados para serem comparados.

Capítulo 6

Resultados e Discussão

Neste capítulo, será aplicado o método descrito no Capítulo 3 sobre as bases de dados descritas no Capítulo 2. Cada seção deste capítulo apresenta resultados do método proposto sobre cada uma das base de dados utilizada. Assim, na Seção 6.1 será apresentada uma expansão do estudo realizado em [1] do caso bivariado para o multivariado. Em sequência, na Seção 6.2 serão realizados novos testes para o estudo de associação entre movimentos do trato vocal e da face realizados em [8] com o objetivo de descrever como estes três domínios estão coordenados ao longo do tempo. Por fim, na Seção 6.3 será realizado um estudo sobre como encontrar a mínima variância e a máxima correlação entre dois grupos de ativos e como esta relação varia ao longo do tempo, para futuramente auxiliar no desenvolvimento de um algoritmo baseado em arbitragem estatística.

6.1 Base de dados 1

Neste tópico são aplicados os métodos de associação variante no tempo para estudar a coordenação entre os movimentos do trato vocal enquanto dois locutores estão conversando [1]. Como a análise é realizada entre grupos de variáveis, os sensores são divididos em dois grupos, cada um representando os movimentos da língua de cada locutor. Os grupos de

dados consistem nos sensores distribuídos ao longo da língua do locutor *EVB* (vetor $\boldsymbol{x}$, representando as variáveis contidas nas colunas da matriz $\boldsymbol{T_1^{EVB}}$) e *CTB* (vetor $\boldsymbol{y}$, representando as variáveis contidas nas colunas da matriz $\boldsymbol{T_{CTB_1}}$), conforme descrito no capítulo 2. Assim, são apresentados resultados para a *trial 1* e a *trial 2* (mesma nomenclatura utilizada em [1]). Desta forma, é realizada uma comparação entre os resultados dos trabalhos. Para o cálculo da covariância foi aplicado o filtro média móvel bidirecional com decaimento exponencial com $\eta = 0.025$, um dos valores utilizados no estudo anterior. Consequentemente, também foi rejeitada a implementação de vetores autoregressivos, pois um pré-processamento distinto pode dificultar uma comparação entre os resultados.

Dentre as análises, em um primeiro momento foram calculados os valores das medidas de associação instantâneas v^{pca}, v^{cca} e h. Para a medida v^{cca}, os resultados são apresentados por meio de gráficos de área para descrever a representatividade de cada variável na constituição da medida de associação. Seguindo a mesma lógica, juntamente com a medida de associação h é apresentada a representatividade das componentes canônicas na composição desta. Também são apresentados gráficos que ilustram como cada componente principal contribui para a variância total de seu respectivo grupo ao longo do tempo. Busca-se por meio desta análise validar a hipótese de que existe relação entre a representatividade da primeira componente principal e a associação entre os grupos. Por fim, são apresentados os mapas de associação para os dois grupos e estes são comparados com os resultados encontrados para o caso bivariado na referência.

Nas figuras 6.1 (*trial 1*) e 6.2 (*trial 2*) são apresentados os gráficos da associação instantânea entre os grupos calculadas com base nas medidas h, $v_{x|y}^{cca}$ e $v_{y|x}^{cca}$. Como pode-se observar, além do valor da associação instantânea (soma cumulativa de todas as áreas), os gráficos apresentam como cada uma das componentes canônicas contribui para a medida h assim como a representatividade de cada um dos sensores em $v_{x|y}^{cca}$ e $v_{y|x}^{cca}$. É possível constatar que os sensores distribuídos ao longo da língua contribuem de forma similar para a associação entre os grupos, com

a exceção daquele posicionado no lábio superior do locutor *EVB*. Outra conclusão que pode ser extraída desta figura é que o valor da primeira componente canônica é uma medida de associação consistente, pois sua representatividade (valor da correlação canônica dividido pelo número de componentes canônicas) está altamente coordenada com a medida h.

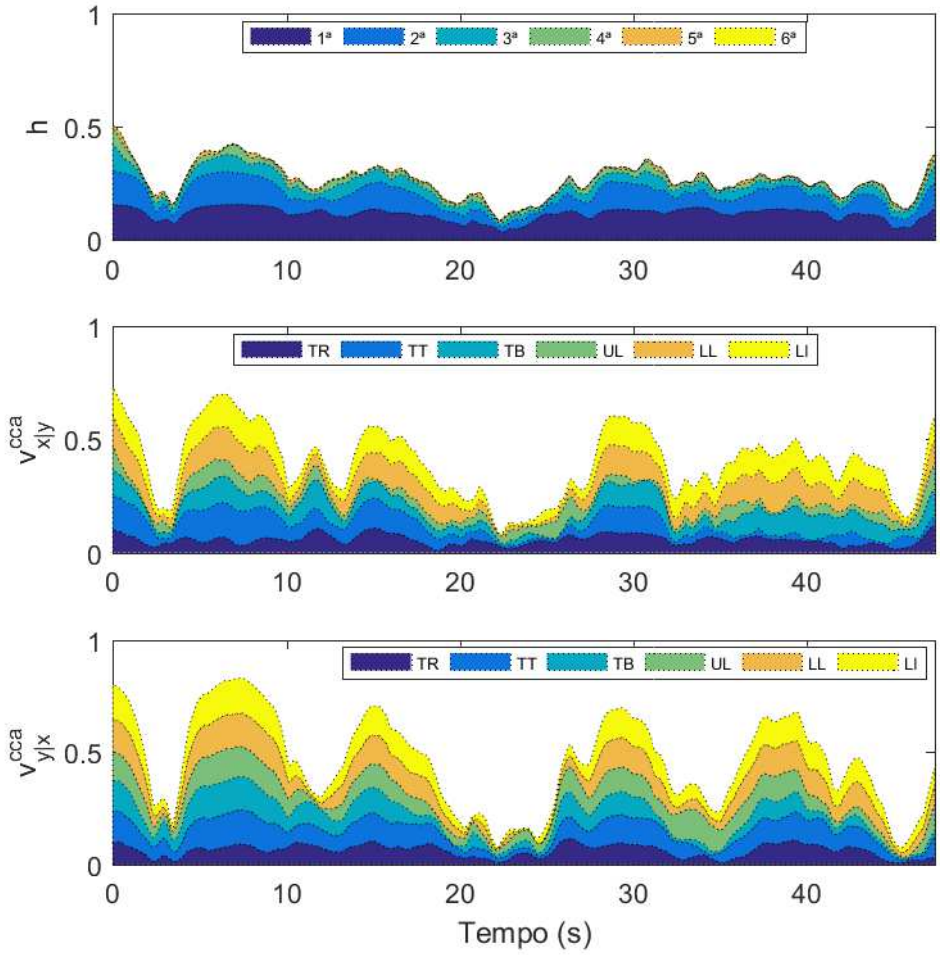

Figura 6.1: Representação de cada componente canônica em h e do peso de cada um dos sensores na medida de associação $v^{cca}_{x|y}$ e $v^{cca}_{y|x}$, respectivamente. Resultados gerados a partir da *trial 1*

Nas figuras 6.3 (*trial 1*) e 6.4 (*trial 2*) são apresentados resultados das simulações para as medidas $v^{pca}_{x|y}$ e $v^{pca}_{y|x}$ e uma análise da representatividade (ou peso) de cada uma das componentes principais dentro da variância total de cada grupo ($p^r_1(k)$, representada no gráfico pela letra p). Esta última análise é explorada por acreditar que caso os grupos estejam descoordenados, o número de componentes principais necessário para representar a variância total de cada grupo será maior. Para quantificar e validar a hipótese, foram calculadas matrizes de correlação entre as diversas grandezas físicas e os resultados são apresentados nas tabelas 6.1 e 6.2 para a *trial 1* e a *trial 2*, respectivamente. Os valores de correlação entre a representatividade da primeira componente principal de *CTB* com as medidas de associação apresentam valores elevados, entre 0.5 e 0.6, sugerindo que, de fato, pode existir uma relação entre os domínios.

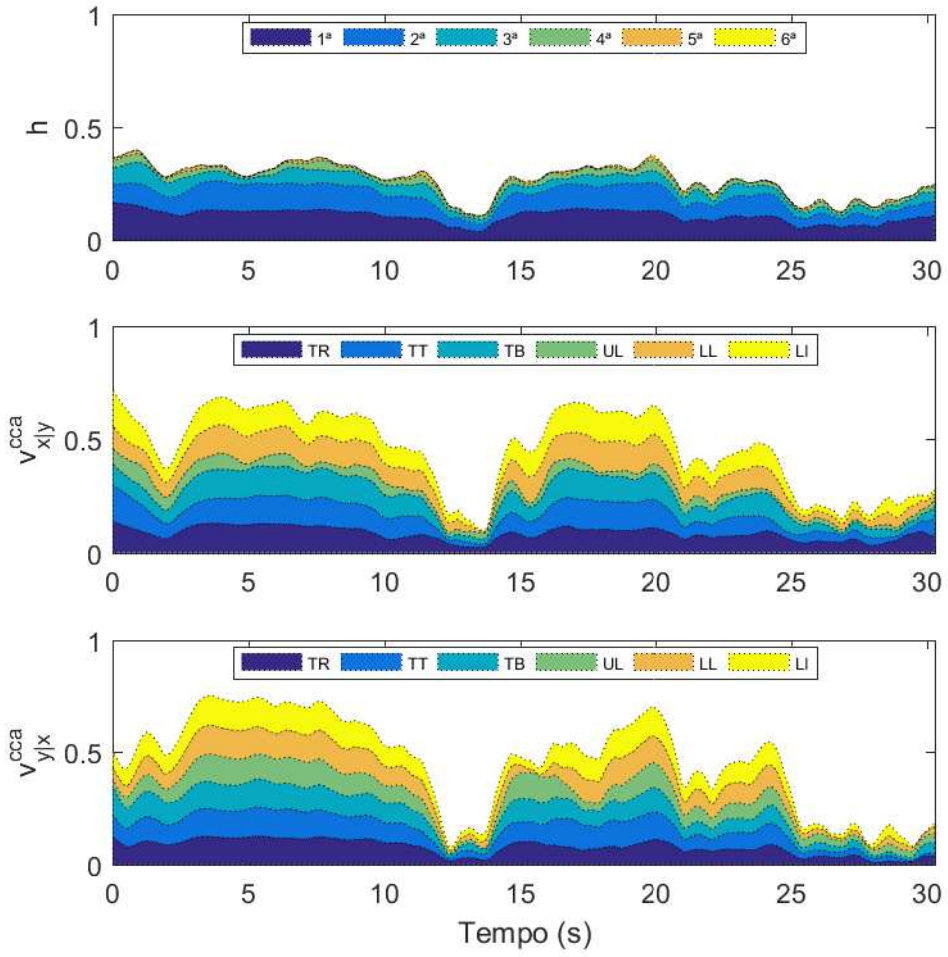

Figura 6.2: Representação de cada componente canônica em h e do peso de cada um dos sensores na medida de associação $v^{cca}_{x|y}$ e $v^{cca}_{y|x}$, respectivamente. Resultados gerados a partir da *trial 2*

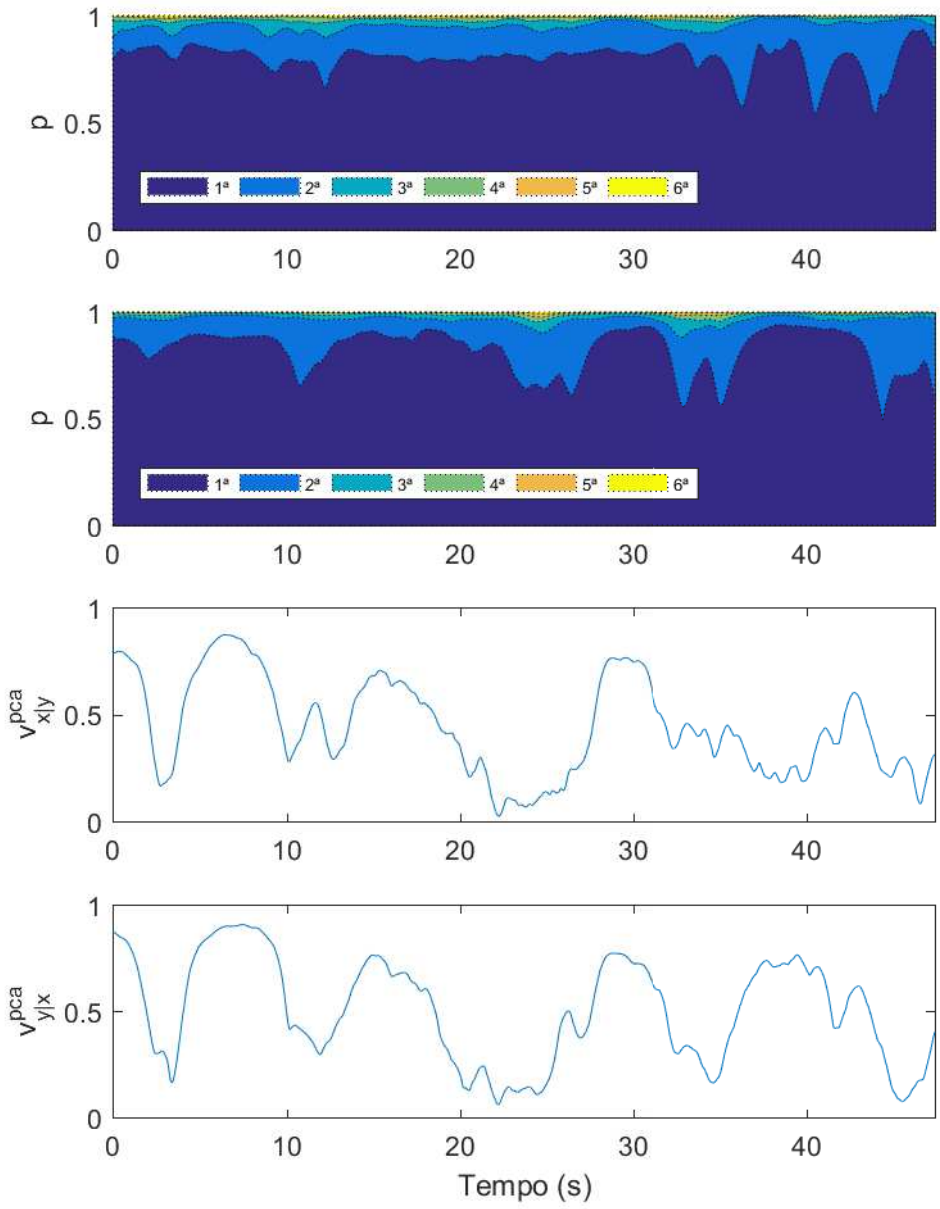

Figura 6.3: Valor da representatividade das componentes principais para o grupo $\boldsymbol{x}$ e $\boldsymbol{y}$ e valor da associação instantânea para as medidas $v^{pca}_{x|y}$ e $v^{pca}_{y|x}$, respectivamente. Resultados gerados a partir da *trial 1*

Tabela 6.1: Matriz de correlação comparando a representatividade da primeira componente principal de cada grupo com as medidas de associação baseadas em variância compartilhada. Resultados extraídos a partir da *trial 1*

	1ª PC x	1ª PC y	$c_{pca}^{x\|y}$	$c_{pca}^{y\|x}$	$c_{cca}^{x\|y}$	$c_{cca}^{y\|x}$
1ª PC x	1	0,1089	0,1224	0,0483	0,0608	0,1274
1ª PC y	0,1089	1	0,4746	0,6699	0,5406	0,6094
$c_{pca}^{x\|y}$	0,1224	0,4746	1	0,7666	0,8585	0,7660
$c_{pca}^{y\|x}$	0,0483	0,6699	0,7666	1	0,9064	0,9840
$c_{cca}^{y\|x}$	0,0608	0,5406	0,8585	0,9064	1	0,9106
$c_{cca}^{y\|x}$	0,1274	0,6094	0,7660	0,9840	0,9106	1

Em sequência, são apresentados os mapas de associação instantânea entre os grupos de séries temporais para os coeficientes h (*trial 1* 6.5 e *trial 2* 6.8), $v_{x|y}^{pca}$, $v_{y|x}^{pca}$, $v_{x|y}^{cca}$ e $v_{y|x}^{cca}$ (*trial 1* 6.6 e *trial 2* 6.9). Para efeito de comparação, também são inseridos dois mapas de correlação discutidos na referência, um para a *trial 1* (6.7) e outro para a *trial 2* (6.10). Todavia, tais mapas foram modificados, pois o valor da correlação foi elevado ao quadrado, passando a representar a variância compartilhada entre as variáveis e consequentemente, facilitando as comparações. O aumento no número de variáveis (de 1 para 6 variáveis em cada grupo) faz com que o caminho de máxima correlação dentro gráfico de associação seja menos claro. Todavia, os momentos de máxima associação são ressaltados, como as regiões com alta coloração vermelha em 30s e 40s na Figura 6.6 que não eram tão nítidas na Figura 6.7.

6.2 Base de dados 2

Nesta seção são apresentados os resultados da associação instantânea entre os movimentos da face e trato vocal [8], com objetivo de capturar variações ao longo do tempo na relação entre os domínios. Ao longo desta seção, o vetor $\boldsymbol{x}$ faz referência às variáveis presentes na matriz $\boldsymbol{O_2}$, $\boldsymbol{y}$ faz referência às variáveis presentes na matriz $\boldsymbol{T_2}$ e $\boldsymbol{z}$ faz referência às

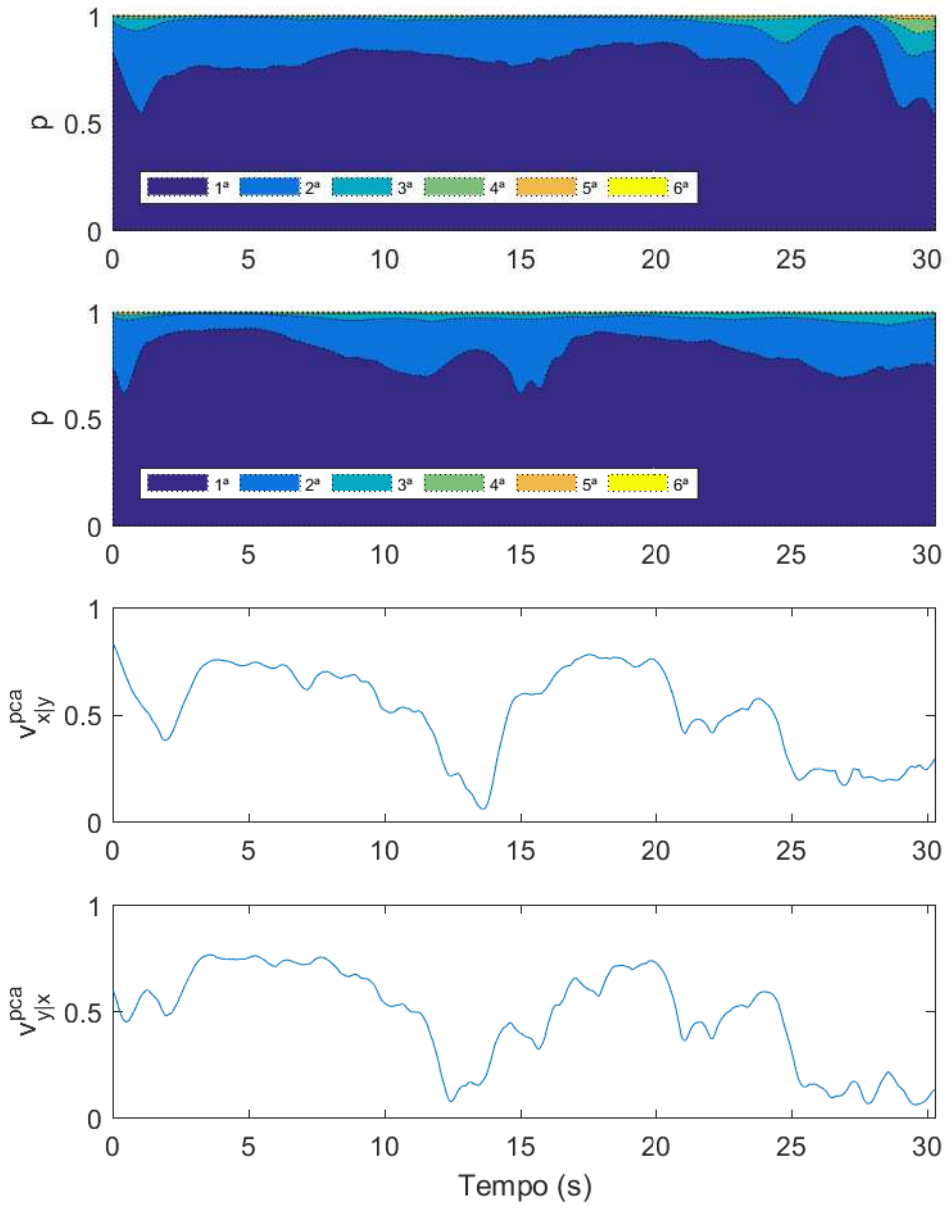

Figura 6.4: Valor da representatividade das componentes principais para o grupo $\boldsymbol{x}$ e $\boldsymbol{y}$ e valor da associação instantânea para as medidas $v^{pca}_{x|y}$ e $v^{pca}_{y|x}$, respectivamente. Resultados gerados a partir da *trial 2*

Tabela 6.2: Matriz de correlação comparando a representatividade da primeira componente principal de cada grupo com as medidas de associação baseadas em variância compartilhada. Resultados extraídos a partir da *trial 2*

	1ª PC x	1ª PC y	$c_{pca}^{x\|y}$	$c_{pca}^{y\|x}$	$c_{cca}^{x\|y}$	$c_{cca}^{y\|x}$
1ª PC x	1	0,0451	0,2141	0,1588	0,1289	0,1200
1ª PC y	0,0451	1	0,4795	0,6381	0,4956	0,5815
$c_{pca}^{x\|y}$	0,2141	0,4795	1	0,9169	0,9873	0,9072
$c_{pca}^{y\|x}$	0,1588	0,6381	0,9169	1	0,9348	0,9830
$c_{cca}^{y\|x}$	0,1289	0,4956	0,9873	0,9348	1	0,9294
$c_{cca}^{y\|x}$	0,1200	0,5815	0,9072	0,9830	0,9294	1

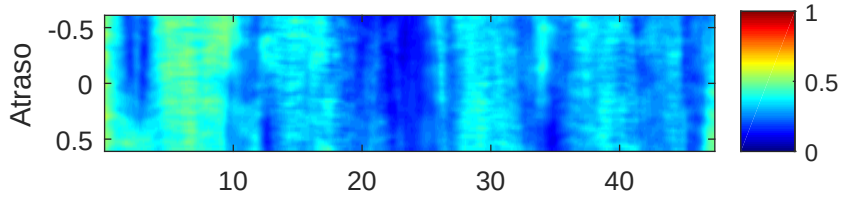

Figura 6.5: Mapas de correlação realizados para a *Trial 1* dos experimentos descritos em [1]. A medida de associação utilizada foi h.

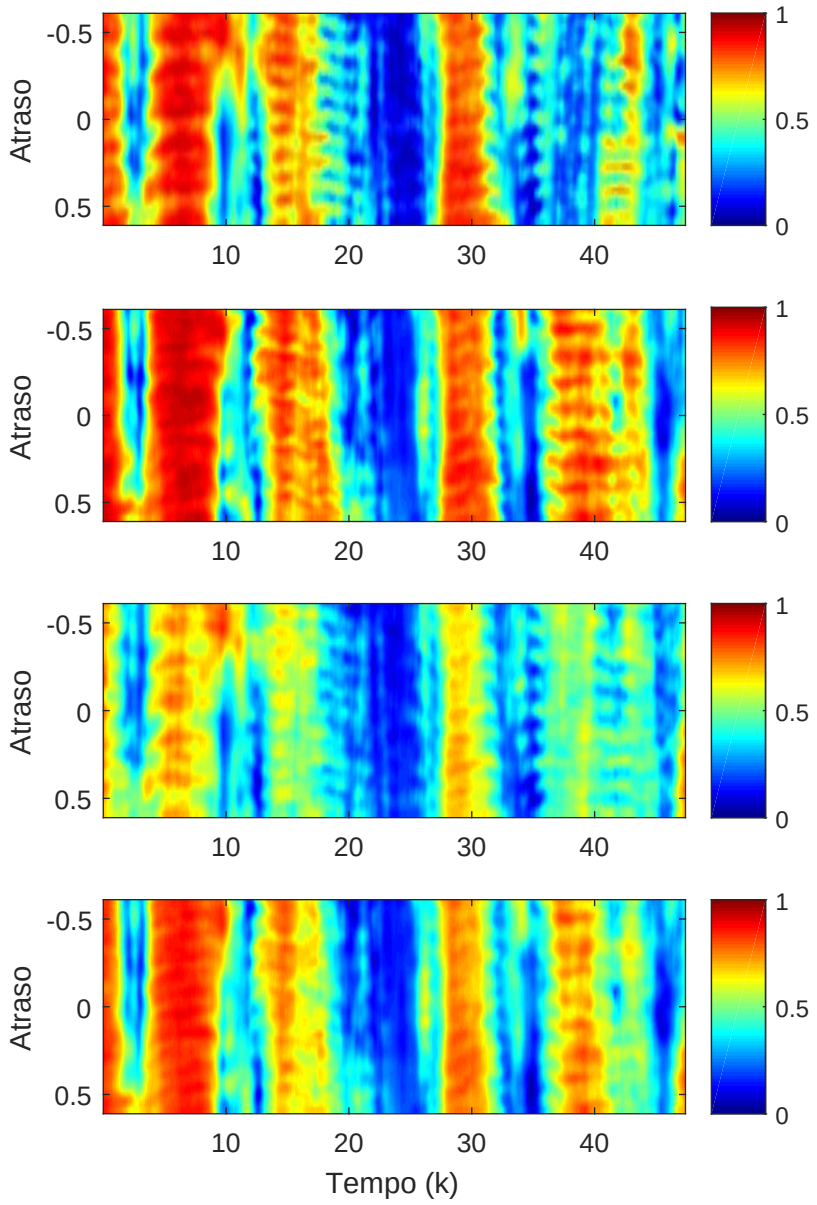

Figura 6.6: Mapas de correlação realizados para a *Trial 1* dos experimentos descritos em [1]. As medidas de associação apresentadas são $v^{pca}_{x|y}$, $v^{pca}_{y|x}$, $v^{cca}_{x|y}$ e $v^{cca}_{y|x}$

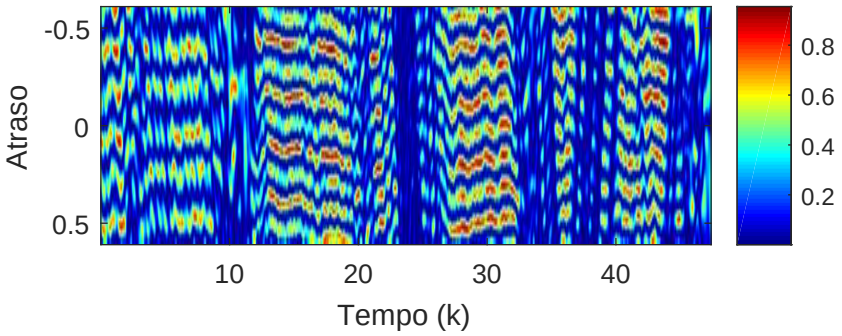

Figura 6.7: Mapa de associação bivariada para onde as variáveis consideradas foram o sensor *TT* para *EVB* e *TR* para *CTB* (*Trial 1*)

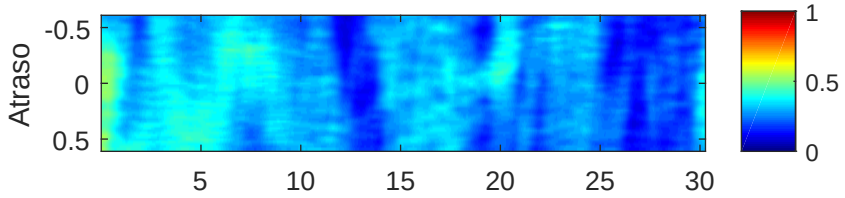

Figura 6.8: Mapas de correlação realizados para a *Trial 2* dos experimentos descritos em [1]. A medida de associação utilizada foi h.

variáveis presentes na matriz $\boldsymbol{A_2}$.

Os testes foram realizados para duas sentenças, ambas pronunciadas pelo locutor *EVB*: *Sam sat on top of the potato cooker and Tommy cut up a bag of tiny potatoes and popped the beet tips into the pot.* (sentença 1) e *When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow.* (sentença 2).

Após uma comparação entre as medidas de associação estáticas entre os dois trabalhos, são apresentados os valores da associações instantâneas para as medidas v^{pca}, v^{cca} e h seguidas pelos mapas de associação. Por fim, é analisada a efetividade do modelo média móvel exponencial e dos vetores autoregressivos para remover a dinâmica dos sistemas, estimando os parâmetros dos modelos e comparando os resultados da nova abordagem com os gerados anteriormente.

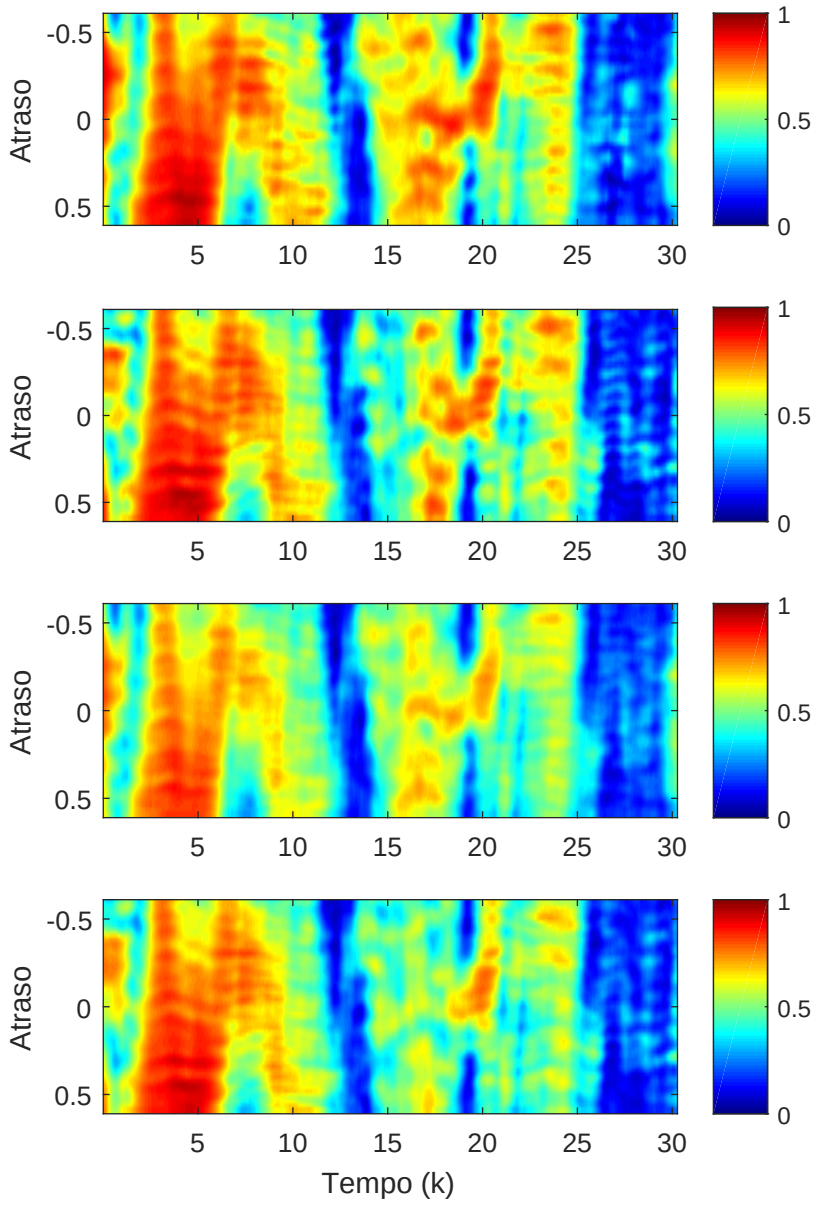

Figura 6.9: Mapas de correlação realizados para a *Trial 2* dos experimentos descritos em [1]. As medidas de associação apresentadas são $v^{pca}_{x|y}$, $v^{pca}_{y|x}$, $v^{cca}_{x|y}$ e $v^{cca}_{y|x}$

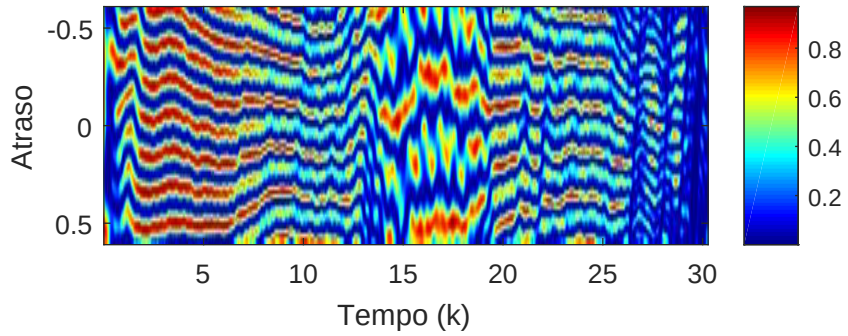

Figura 6.10: Mapa de associação bivariada para onde as variáveis consideradas foram o sensor *TT* para *CTB* e o sensor *TR* para *EVB* (*Trial 2*)

6.2.1 Resultados para associação estática entre grupos

Em [8], os autores apresentam uma metodologia para quantificar a relação entre os movimentos da face e do trato vocal, com o objetivo de estabelecer um *baseline* para servir como referência em trabalhos posteriores. Isso ocorre, pois dada a simplicidade do método utilizado, qualquer outro que utilize um ferramental mais complexo deve apresentar maior acurácia. Apesar das medidas de associação propostas serem baseadas em modelos lineares, assim como em [8], elas possuem algumas diferenças e estas refletem significativamente no valor final do coeficiente de associação e, por este motivo, nesta seção são apresentadas as medidas de associação listadas no Capítulo 3 calculadas de forma estática.

O procedimento utilizado em [8] pode ser dividido em duas partes. Primeiramente, os autores estimam o valor de cada uma das variáveis presentes em $\boldsymbol{x}$ a partir do conjunto de variáveis em $\boldsymbol{y}$ por meio de uma transformação afim [4], e vice e versa. Em um segundo momento, para se quantificar a associação, é utilizado o *Pearson product-moment correlation coefficient*, definido matematicamente como

$$\boldsymbol{R}_{x\hat{x}} = \frac{\sigma^2_{x\hat{x}}}{\sigma_x \sigma_{\hat{x}}} = \frac{tr(\boldsymbol{C}_{x\hat{x}})}{\sqrt{tr(\boldsymbol{C}_{\hat{x}\hat{x}})tr(\boldsymbol{C}_{xx})}} \tag{6.1}$$

onde $\boldsymbol{C}_{xx}$ é a matriz de covariância do vetor aleatório $\boldsymbol{x}$, $\boldsymbol{C}_{\hat{x}\hat{x}}$ a matriz de covariância do valor estimado de $\boldsymbol{x}$ a partir de $\boldsymbol{y}$, $\hat{\boldsymbol{x}}$, e $\boldsymbol{C}_{x\hat{x}}$ a matriz de

covariância cruzada.

Para garantir que os procedimentos sejam próximos, as matrizes de covariância serão estimadas como em [8]. Desta forma, o valor médio esperado das variáveis foi definido como

$$\hat{\boldsymbol{\mu}}_x = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{x}_k \tag{6.2}$$

$$\hat{\boldsymbol{\mu}}_y = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{y}_k \tag{6.3}$$

e as matrizes de covariância definidas como

$$\hat{\boldsymbol{C}}_{xx} = \frac{1}{N-1}\sum_{k=1}^{N} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_x)(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_x)^T \tag{6.4}$$

e

$$\hat{\boldsymbol{C}}_{yy} = \frac{1}{N-1}\sum_{k=1}^{N} (\boldsymbol{y} - \hat{\boldsymbol{\mu}}_y)(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_y)^T \tag{6.5}$$

As estimações das medidas de associação foram realizadas por meio de validação cruzada, para garantir que os valores encontrados são independentes do conjunto de amostras utilizados. Os valores médios de treino e teste, bem como os desvios padrões, para cada uma das medidas são apresentados na Tabela 6.3. Nesta tabela também foram adicionados os valores calculados para o método utilizado em [8], referenciados como $\gamma_{x|y}$ e $\gamma_{y|x}$.

Os valores encontrados para as $v^{pca}_{x|y}$ e $v^{pca}_{y|x}$ estão abaixo do valores do *Pearson product-moment correlation*. Isso ocorre pelo fato de a medida calculada em [8] ser uma medida de correlação e, caso ela for elevada ao quadrado, o resultado será exatamente o encontrado neste trabalho. O fato do valor médio de $v^{cca}_{x|y}$ e $v^{cca}_{y|x}$ estar abaixo do método baseado na PCA indica que sensores com alta variância possuem maior representatividade na associação que sensores com baixa variância.

Tabela 6.3: Médias e desvios padrões das medidas de associação estáticas

	μ **treino**	μ **teste**	σ **treino**	σ **teste**
$\gamma_{x\|y}$	0,93	0,92	$4,7 \times 10^{-3}$	$4,7 \times 10^{-3}$
$\gamma_{y\|x}$	0,79	0,77	$7,7 \times 10^{-3}$	$8,9 \times 10^{-3}$
h	0,45	0,41	$6,2 \times 10^{-3}$	$6,6 \times 10^{-3}$
$v^{pca}_{x\|y}$	0,86	0,86	$8,3 \times 10^{-3}$	$2,5 \times 10^{-2}$
$v^{pca}_{y\|x}$	0,63	0,63	$1,3 \times 10^{-2}$	$2,5 \times 10^{-2}$
$v^{cca}_{x\|y}$	0,71	0,71	$7,8 \times 10^{-3}$	$5,5 \times 10^{-2}$
$v^{cca}_{y\|x}$	0,62	0,61	1×10^{-2}	4×10^{-2}

6.2.2 Medidas de associação variantes no tempo

Nesta seção são estimadas as associações variantes no tempo com atraso nulo entre as séries temporais, a exemplo de como foi feito com a base 1. Para realizar estas simulações, foram concatenadas as repetições da sentença 1 e da sentença 2 como uma série temporal única, para facilitar o processo de geração dos resultados, como descrito o Capítulo 2.

Os resultados apresentados nas figuras 6.11, 6.12 e 6.13 mostram que a coordenação entre os movimentos possui valor condicionado de acordo com a sentença dita pelo locutor, pois os valores para a primeira metade dos gráficos (Sentença 1) são inferiores aos da segunda metade (Sentença 2). Desta forma, pode-se inferir que os movimentos da face, do trato vocal e os coeficientes LSP possuem acoplamento condicionado ao que está sendo mencionado pelo locutor e tal resultado realça a necessidade de que em aplicações de fala, talvez seja mais adequado realizar o mapeamento de um grupo de variáveis a partir do outro de forma dinâmica.

Outro resultado expressivo é a similaridade existente entre as medidas de associação instantâneas e os valores quadráticos médios do sinal de voz (calculados por quadro, a exemplo dos coeficientes LSP) apresentados na Figura 6.14, sobre os quais foram aplicados o filtro média móvel com decaimento exponencial, atenuando-se as componentes de alta frequência. Desta forma, pode-se inferir que os movimentos são mais coordenados na segunda sentença, onde o valor médio quadrático do sinal

de fala é maior. Uma análise detalhada para determinar as causas de tal fenômeno é deixada para estudos futuros.

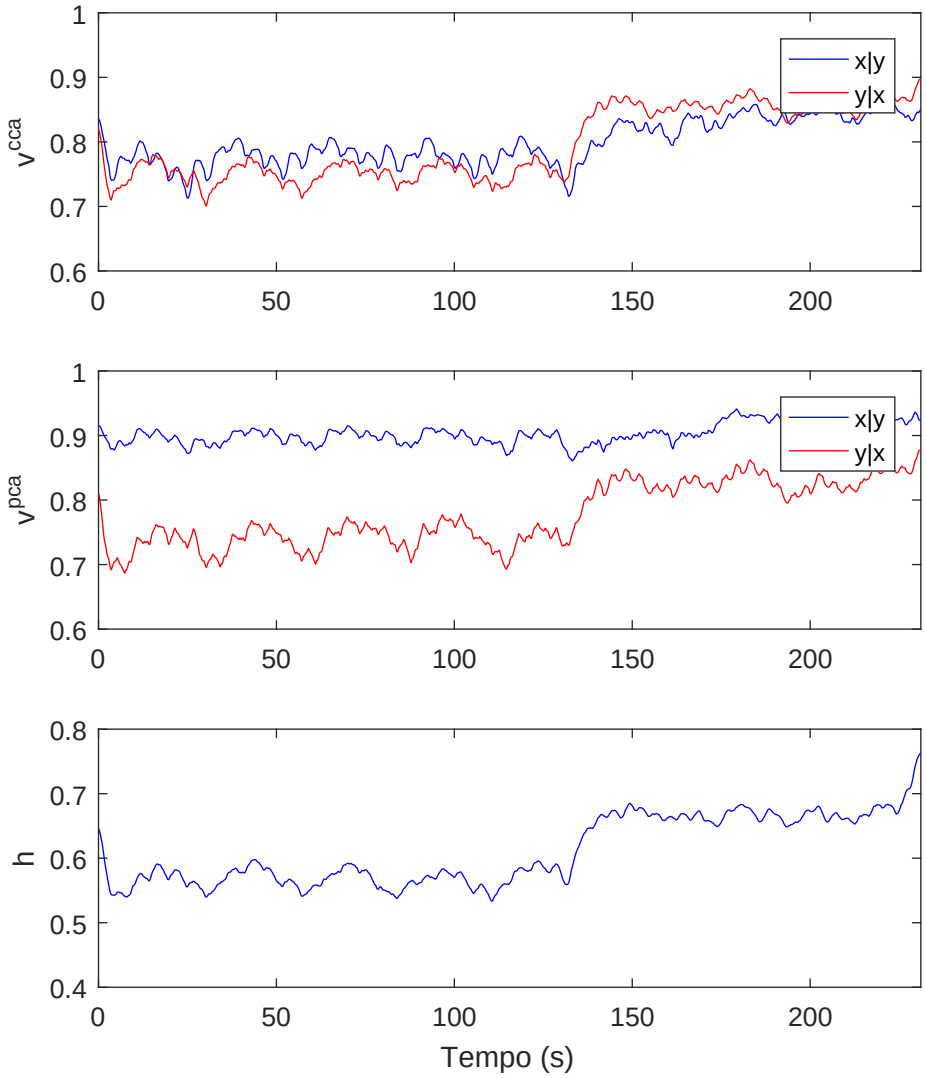

Figura 6.11: Coeficientes de associação $v^{cca}_{x|y}$ e $v^{cca}_{y|x}$, $v^{pca}_{x|y}$ e $v^{pca}_{y|x}$ e h para o caso 1D. Os pequenos vales que aparecem no gráfico são ocasionados pela concatenação entre as elocuções e por isso devem ser desconsiderados.

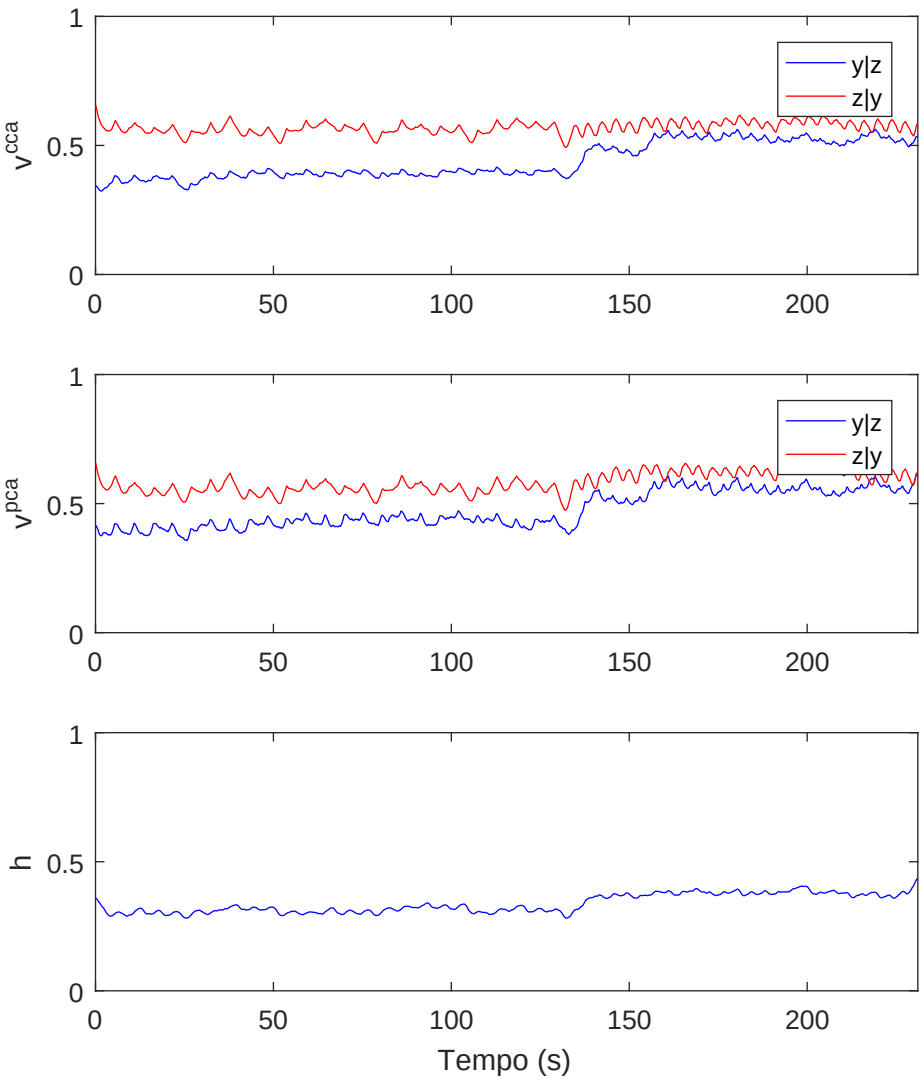

Figura 6.12: Coeficiente de associação variantes no tempo para o caso 1D para os movimentos do trato vocal e acústica da fala. Os pequenos vales que aparecem no gráfico são ocasionados pela concatenação entre as elocuções e por isso devem ser desconsiderados.

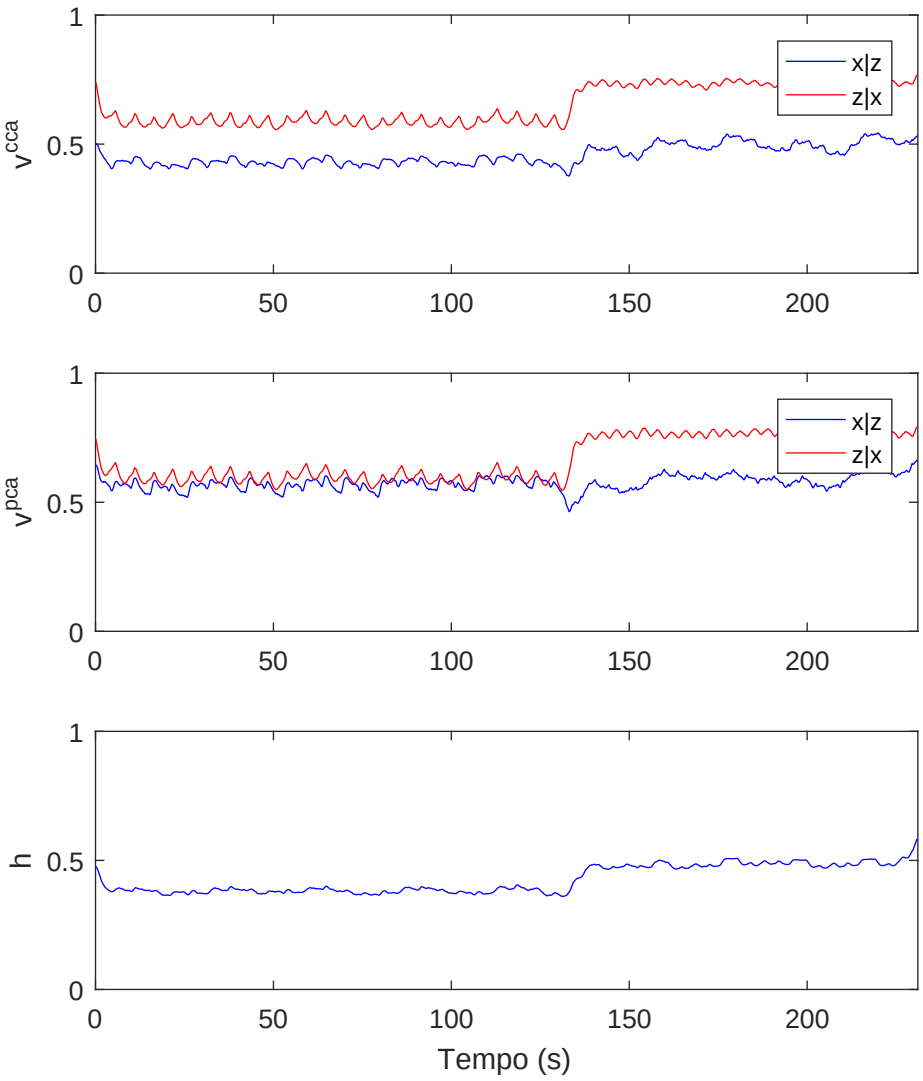

Figura 6.13: Coeficiente de associação variantes no tempo para o caso 1D para os movimentos da face e acústica da fala. Os pequenos vales que aparecem no gráfico são ocasionados pela concatenação entre as elocuções e por isso devem ser desconsiderados.

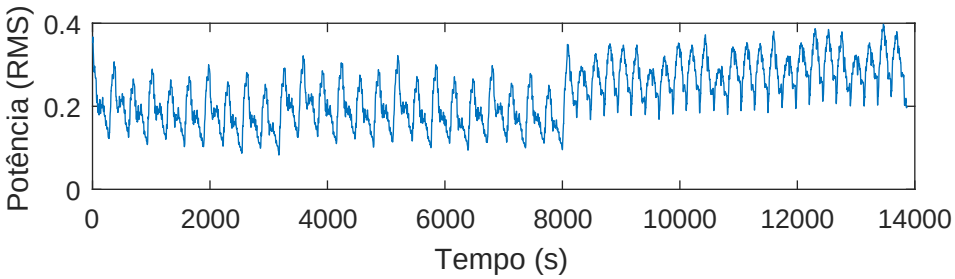

Figura 6.14: Valor da potência extraída do sinal de voz após aplicação do filtro de média móvel. Os pequenos vales que aparecem no gráfico são ocasionados pela concatenação entre as elocuções e por isso devem ser desconsiderados.

A exemplo da seção anterior, foram gerados mapas de associação a partir de uma elocução da sentença 1 e outra da sentença 2. Os seguintes mapas foram gerados para cada uma das sentenças foram:

1. Mapa de associação calculado a partir de h entre os movimentos do trato vocal ($\boldsymbol{y}$) e da face ($\boldsymbol{x}$) (Sentença 1: Figura 6.15 e Sentença 2: Figura 6.21).

2. Mapa de associação calculado a partir de h entre os movimentos da face ($\boldsymbol{x}$) e acústica da fala ($\boldsymbol{z}$) (Sentença 1: Figura 6.16 e Sentença 2: Figura 6.22).

3. Mapa de associação calculado com a medida h entre os movimentos do trato vocal ($\boldsymbol{y}$) e acústica da fala ($\boldsymbol{z}$) (Sentença 1: Figura 6.17 e Sentença 2: Figura 6.23).

4. Mapas de associação calculado com as medidas v^{pca} e v^{cca} entre os movimentos do trato vocal ($\boldsymbol{y}$) e da face ($\boldsymbol{x}$) (Sentença 1: Figura 6.18 e Sentença 2: Figura 6.24).

5. Mapas de associação calculado com as medidas v^{pca} e v^{cca} entre os movimentos da face ($\boldsymbol{x}$) e acústica da fala ($\boldsymbol{z}$) (Sentença 1: Figura 6.19 e Sentença 2: Figura 6.25).

6. Mapas de associação calculado com as medidas v^{pca} e v^{cca} entre os movimentos do trato vocal ($\boldsymbol{y}$) e acústica da fala ($\boldsymbol{z}$) (Sentença 1: Figura 6.20 e Sentença 2: Figura 6.26).

Para todas as medidas, os mapas apresentam uma linha avermelhada ao centro, indicando que os grupos estão em fase para ambas as sentenças pronunciadas. Uma outra hipótese para justificar tal resultado seria uma não detecção de atrasos instantâneos por conta da atenuação sobre as variações de alta frequência ocasionadas pelo filtro média móvel.

Nos gráficos onde foram apresentados os resultados para as medidas de associação baseadas em variância compartilhada consegue-se notar um outro resultado interessante. Em todas as sentenças, a variância da face estimada a partir do trato vocal apresenta uma grande discrepância de valores entre o valor no atraso nulo e os valores de outros atrasos, sendo o primeiro muito mais elevado que o último. Já no caso da variância do trato vocal explicada a partir da face o gráfico se torna homogêneo com associações acima de 0.6 distribuídas ao longo do gráfico. Isso confirma uma hipótese que os movimentos da face são uma função do trato vocal e não o oposto. Em outras palavras, para cada posição do trato vocal existe uma e somente uma configuração da face, mas para uma configuração da face, podem existir diferentes funções do trato vocal.

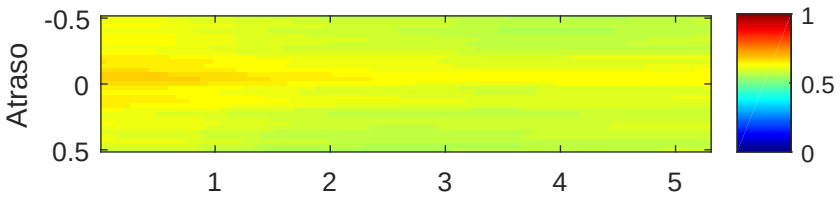

Figura 6.15: Mapa de associação para o coeficiente de associação h, representando a relação entre o movimento da face e o movimento do trato vocal para a sentença 1.

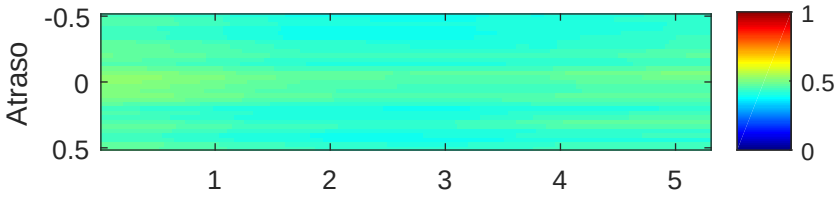

Figura 6.16: Mapa de associação para o coeficiente de associação h, representando a relação entre da acustica da fala e o movimento da face para a sentença 1.

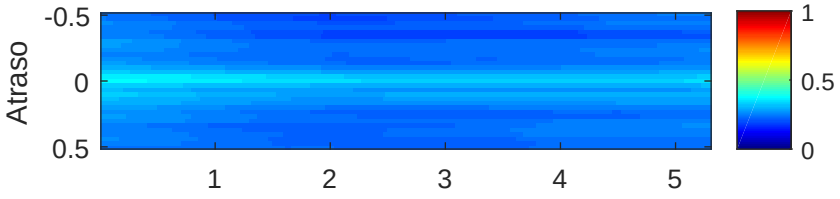

Figura 6.17: Mapa de associação para o coeficiente de associação h, representando a relação entre da acustica da fala e o movimento do trato vocal para a sentença 1.

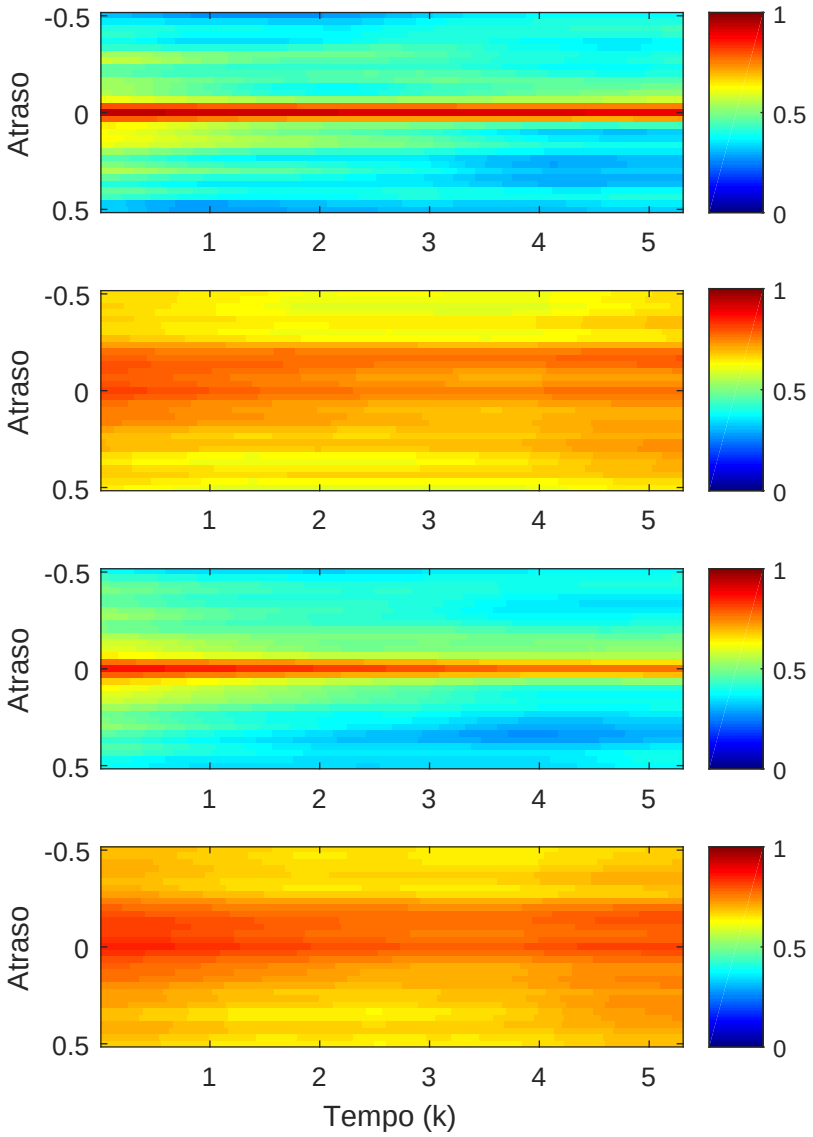

Figura 6.18: Mapas de associação gerados para a sentença 1 a partir do modelo média móvel. Os mapas correspondem aos coeficientes $v^{pca}_{x|y}$, $v^{pca}_{y|x}$, $v^{cca}_{x|y}$ e $v^{cca}_{y|x}$.

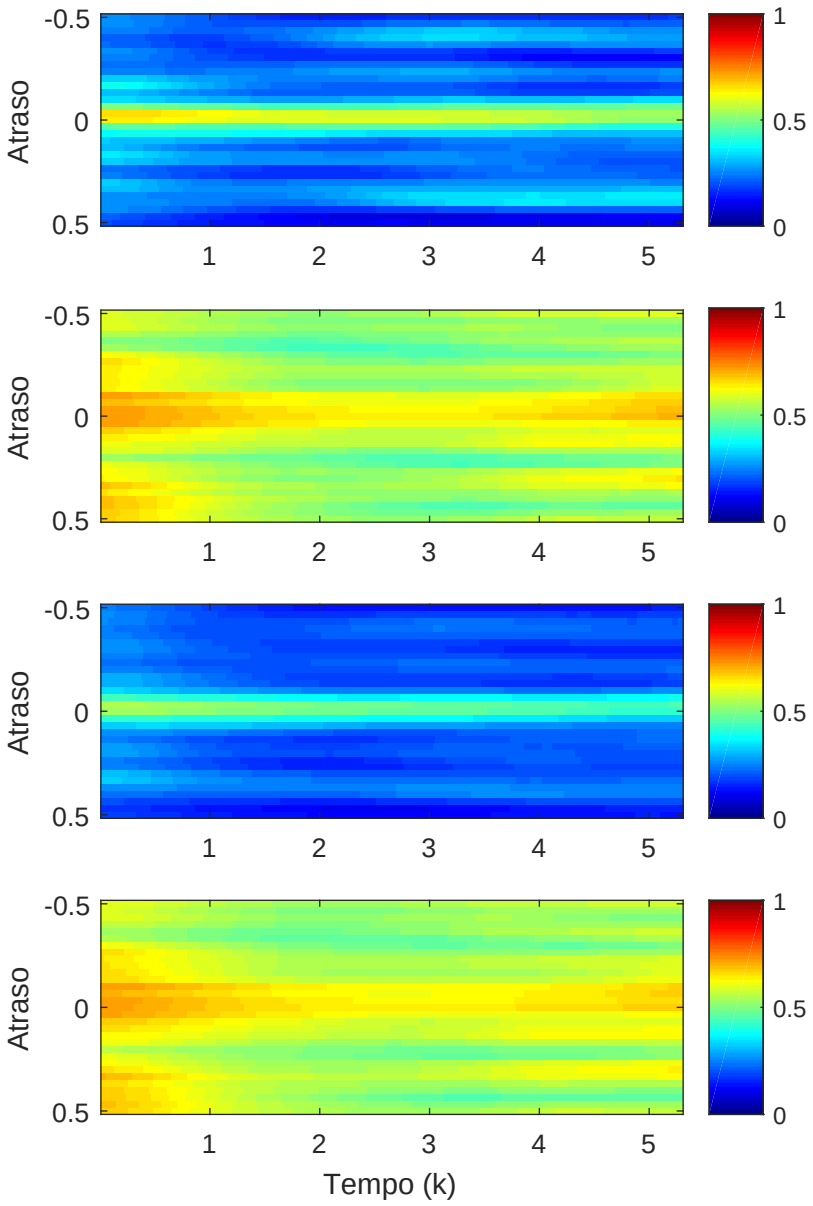

Figura 6.19: Mapas de associação gerados para a sentença 1 a partir do modelo média móvel. Os mapas correspondem aos coeficientes $v^{pca}_{x|z}$, $v^{pca}_{z|x}$, $v^{cca}_{x|z}$ e $v^{cca}_{z|x}$.

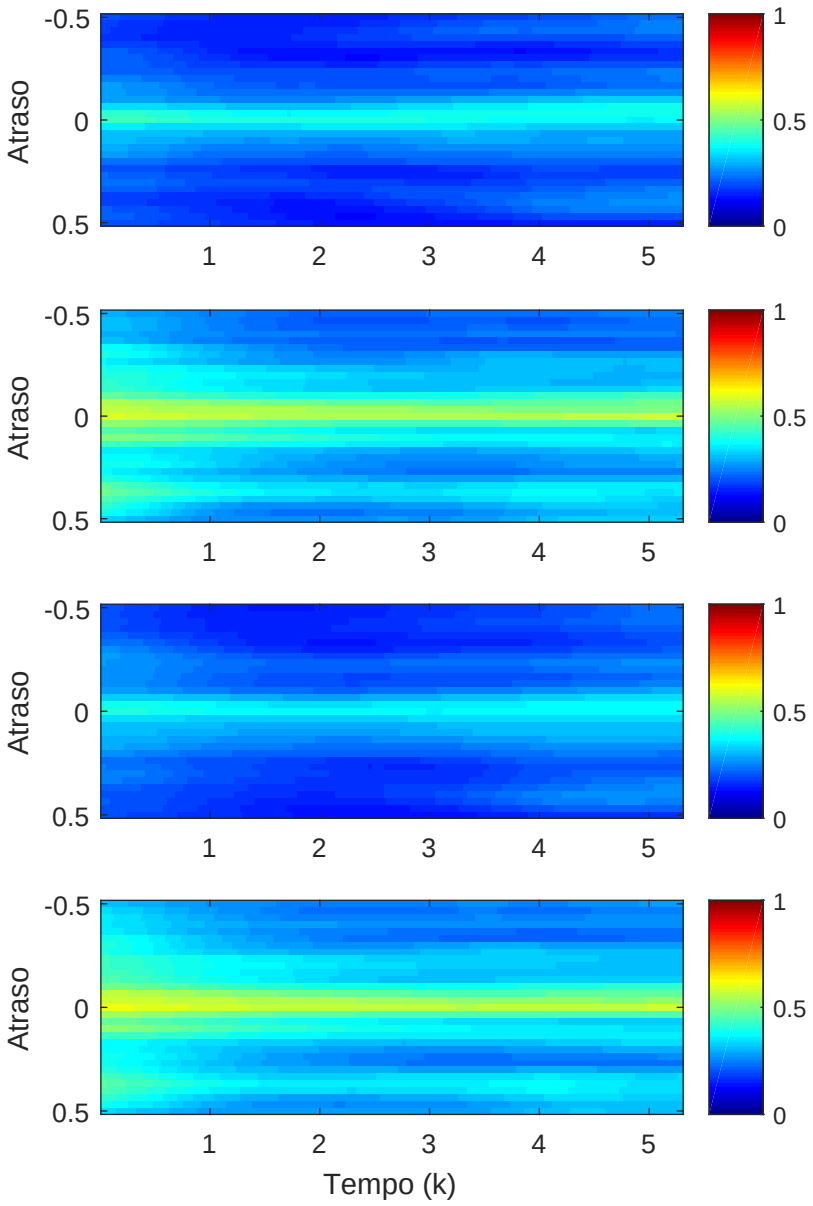

Figura 6.20: Mapas de associação gerados para a sentença 1 a partir do modelo média móvel. Os mapas correspondem aos coeficientes $v^{pca}_{y|z}$, $v^{pca}_{z|y}$, $v^{cca}_{y|z}$ e $v^{cca}_{z|y}$.

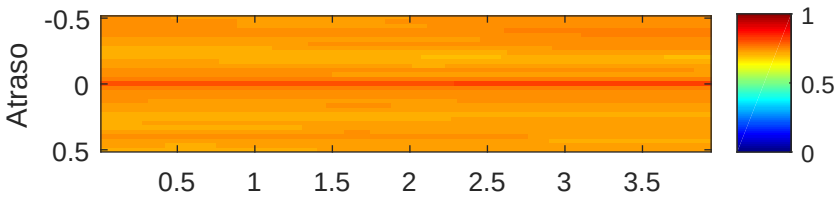

Figura 6.21: Mapa de associação para o coeficiente de associação *h*, representando a relação entre o movimento da face e o movimento do trato vocal para a sentença 2.

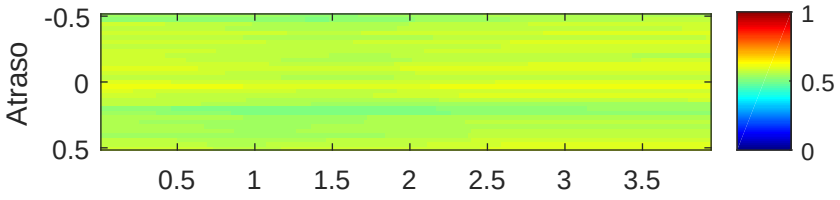

Figura 6.22: Mapa de associação para o coeficiente de associação *h*, representando a relação entre da acustica da fala e o movimento da face para a sentença 2.

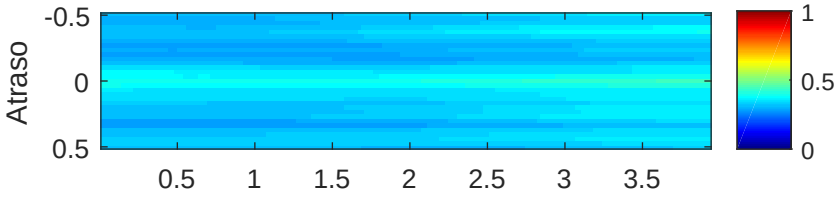

Figura 6.23: Mapa de associação para o coeficiente de associação *h*, representando a relação entre da acustica da fala e o movimento do trato vocal para a sentença 2.

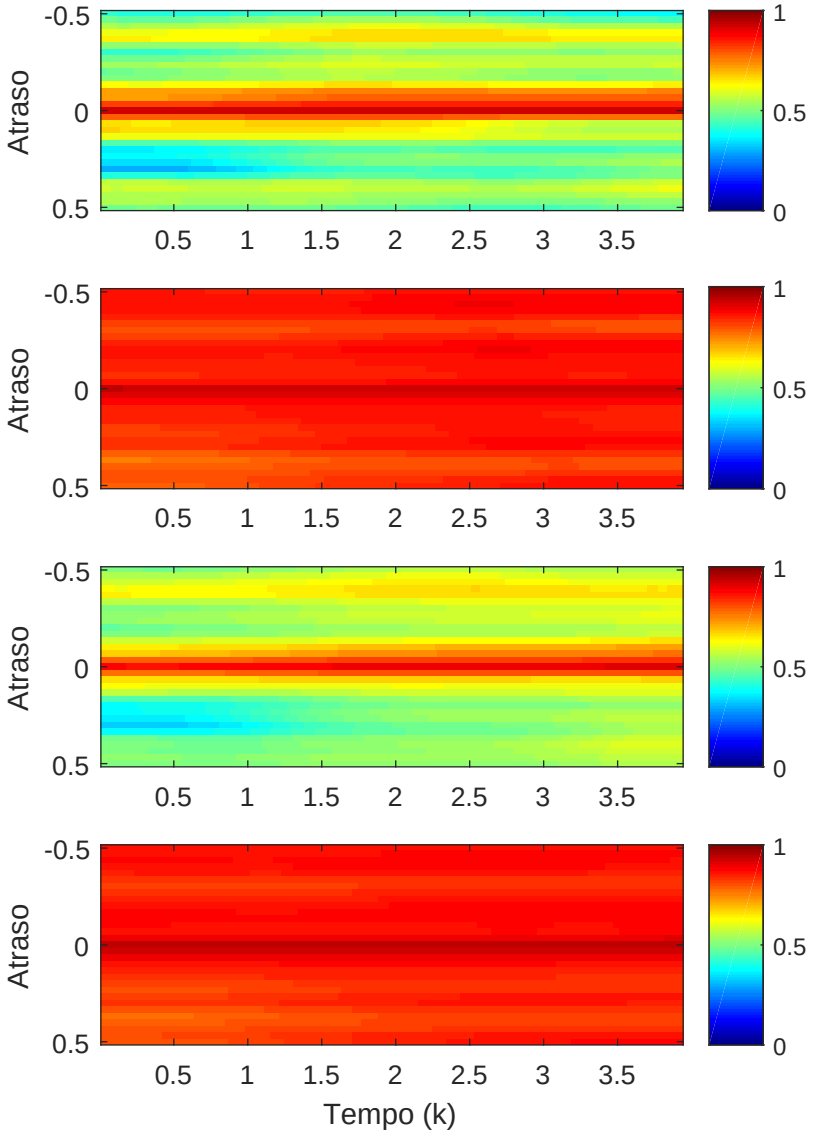

Figura 6.24: Mapas de associação gerados para a sentença 2 a partir do modelo média móvel. Os mapas correspondem aos coeficientes $v^{pca}_{x|y}$, $v^{pca}_{y|x}$, $v^{cca}_{x|y}$ e $v^{cca}_{y|x}$.

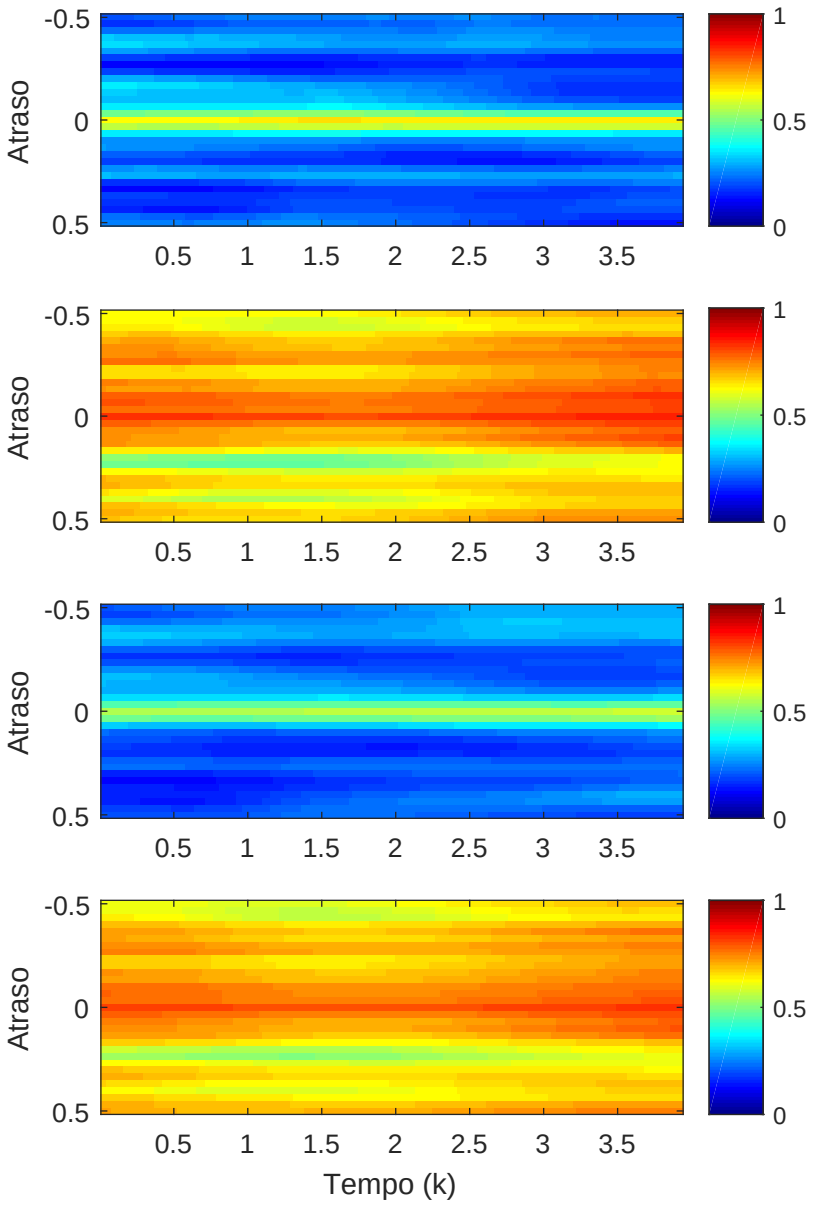

Figura 6.25: Mapas de associação gerados para a sentença 2 a partir do modelo média móvel. Os mapas correspondem aos coeficientes $v^{pca}_{x|z}$, $v^{pca}_{z|x}$, $v^{cca}_{x|z}$ e $v^{cca}_{z|x}$.

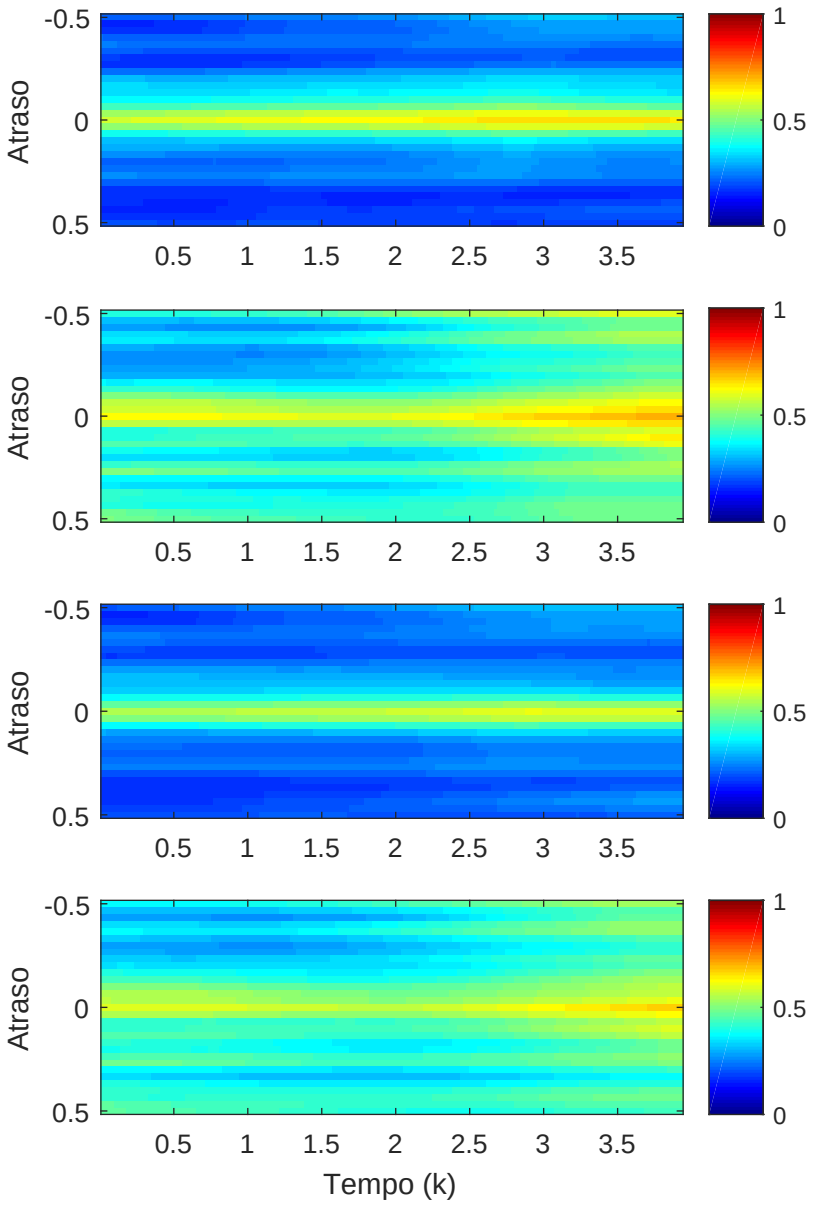

Figura 6.26: Mapas de associação gerados para a sentença 2 a partir do modelo média móvel. Os mapas correspondem aos coeficientes $v^{pca}1_{y|z}$, $v^{pca}_{z|y}$, $v^{cca}_{y|z}$ e $v^{cca}_{z|y}$.

6.2.3 Resultado da remoção de correlação entre as amostras: a utilização dos vetores autoregressivos

Conforme descrito brevemente no capítulo 4, os métodos da família da PCA e da CCA são capazes de representar totalmente a relação entre as variáveis quando os dados apresentam comportamento gaussiano, pelo fato de uma distribuição normal ser representada por seus momentos de primeria e segunda ordem. Este normalmente é encontrado quando as amostras estão descorrelacionadas, ou em outras palavras, quando elas apresentam o comportamento de um ruído branco [37]. Para instigar a discussão do tema, na Figura 6.27 é apresentado o histograma dos desvios em relação a média estática de uma variável dentro do grupo que representa o movimento da face, $x_1(k)$. Pode-se notar por meio de uma simples inspeção visual que os desvios não aparentam ser originados por uma distribuição normal. Por este motivo esta subseção se propõe a realizar uma discussão sobre as consequências de remover as informações redundantes sobre os resultados finais com o auxílio dos vetores autoregressivos.

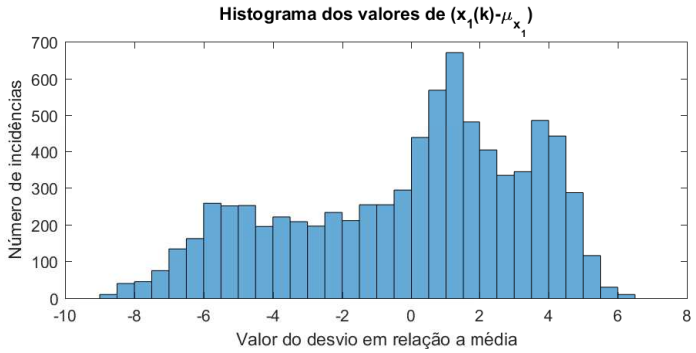

Figura 6.27: Histograma dos desvios em relação a média da primeira variável dos sensores ópticos que captam o movimento facial, $x_1(k)$

Para isso, o comportamento dinâmico do sistema foi estimado por um modelo de vetores autoregressivos de primeira ordem ($L = 1$). Para testar a estabilidade do modelo,foi utilizado validação cruzada. Para quantificar a porcentagem dos movimentos que não pode ser estimada a partir de

dados dos instantes anteriores, foi calculada a razão entre a variância dos resíduos e a variância total do grupo definida como

$$\%_{\boldsymbol{x}(k)-\hat{\boldsymbol{x}}(k)} = \frac{tr(\boldsymbol{C}_{\epsilon_x \epsilon_x})}{\sum_{i=1}^{N_x} \lambda_{x,i}} \tag{6.6}$$

onde $\hat{\boldsymbol{x}}(k)$ são os valores estimados do vetor aleatório $\boldsymbol{x}(k)$ pelo método dos vetores autoregressivos a partir de amostras dos instantes anteriores e $\boldsymbol{\epsilon_x}(k)$ é um vetor definido como

$$\boldsymbol{\epsilon_x}(k) = \boldsymbol{x}(k) - \hat{\boldsymbol{x}}(k) \tag{6.7}$$

Para o vetor aleatório $\boldsymbol{y}(k)$ e sua estimação $\hat{\boldsymbol{y}}(k)$ tal medida pode ser estimada seguindo o mesmo raciocínio. Na Tabela 6.4 são apresentados valores médios e desvios padrões encontrados nas etapas de teste e treino para ambos os grupos de variáveis, todos apresentados no formato de porcentagem. Na Figura 6.28 são apresentados os valores reais e estimados pelo modelo para a primeira série temporal do grupo $\boldsymbol{x}$, $x_1(k)$. Na Figura 6.29 é apresentada uma comparação entre o histograma apresentado no começo desta seção (dos desvios em relação a um valor médio estático) e as inovações obtidas a partir dos vetores autoregressivos. Como pode-se observar, o comportamento é bem mais próximo de uma gaussiana. Tal resultado é confirmado com a execução do teste de *Jarque-Bera* sobre as inovações. Não foi possível rejeitar a hipótese nula com o nível de significância $\alpha = 0.05$, e o parâmetro do teste apresentou um valor muito superior ao das simulações anteriores, indicando um melhor desempenho do VAR para remover informação redundante entre amostras. Por fim, conforme descrito em [37], de acordo com o teorema do limite central, caso as amostras estejam descorrelacionadas, a distribuição de probabilidade das mesmas tende a ser uma normal. Assim, o fato do VAR atingir um alto valor no teste de *Jarque-Bera* (descrito no capítulo de anexos) está intimamente ligado ao fato de remover redundância entre as amostras. Isso pode ser comprovado pelas funções de autocorrelação da série $x_1(k)$ e das inovações obtidas após aplicação do VAR, $\epsilon_{x_1}(k)$

Tabela 6.4: Resultado da porcentagem de variância que não pode ser estimada pelos instantes anteriores. Assim como nas simulações das medidas de associação, foi implementado validação cruzada.

	μ **Treino**	σ **Treino**	μ **Teste**	σ **Teste**
$\boldsymbol{x}$	$4,09\%$	0,4098	4,56	$0,4530\%$
$\boldsymbol{y}$	$7,1\%3$	0,5795	$7,42\%$	0,6324

apresentadas na Figura 6.30. Estes gráficos explicitam o fato da função de autocorrelação das inovações ser muito mais próxima de um impulso, como desejado, que a função de autocorrelação do dado bruto.

Os resultados para estimar a associação variante no tempo realizados na subseção anterior foram repetidos, todavia com o VAR sendo aplicado previamente sobre os grupos de sinais. Na Figura 6.31 são apresentados os resultados para as medidas de associação instantânea enquanto nas figuras 6.32, 6.34 e 6.33, 6.35 são apresentados os mapas de associação para as sentenças 1 e 2 respectivamente. O valor de η considerado foi 0.01.

Dois pontos devem ser destacados. O primeiro deles é uma queda significativa no valor da associação instantânea em comparação com os resultados anteriores, ocasionado pela redução na redundância entre as amostras. O segundo ponto é que no caso da medida de associação baseada na variância compartilhada com auxílio da PCA, houve uma inversão dos resultados: na segunda sentença é possível extrair mais variância do trato vocal a partir da face, do que o oposto, contrariando os princípios de produção de fala. O erro pode ter sido originada por um efeito adverso da utilização dos vetores autoregressivos: a queda da relação sinal-ruído. Como o VAR remove a maior parte das componentes de baixa frequência, o resultado de sua implementação pode ser facilmente comparado a de um filtro passa altas. Como a variância dos resíduos é baixa em comparação com a do sinal original, pode-se dizer que uma grande porcentagem destas inovações na realidade é ruído originado pelo processo de medição.

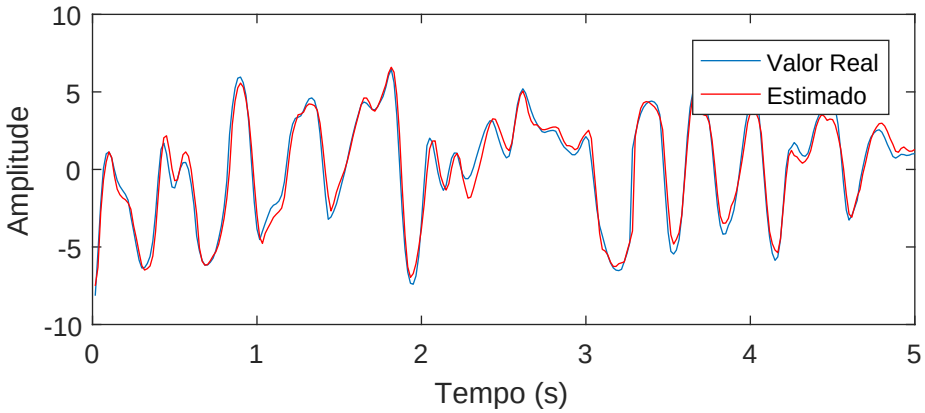

Figura 6.28: Comparativo entre os valores reais das séries temporais e os valores estimados pelo modelo de vetores autoregressivos. Os resultados em questão foram gerados para a série temporal $x_1(k)$.

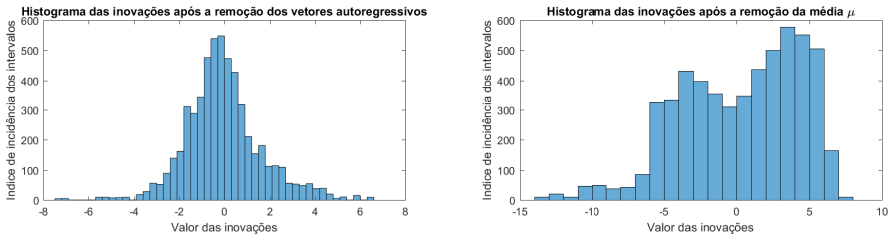

Figura 6.29: Histograma dos desvios em relação à predição feita pelo vetor autoregressivo de primeira ordem, a esquerda, e dos desvios em relação a média estática da população, a direita. Todos os resultados foram extraídos da série temporal $x_1(k)$

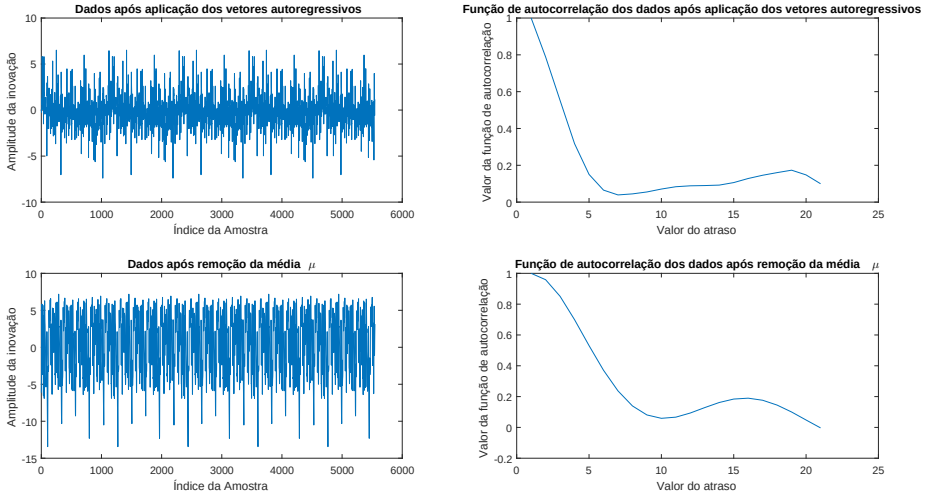

Figura 6.30: Desvios e função de autocorrelação dos mesmos em relação ao vetor autoregressivo e a média estática da população.

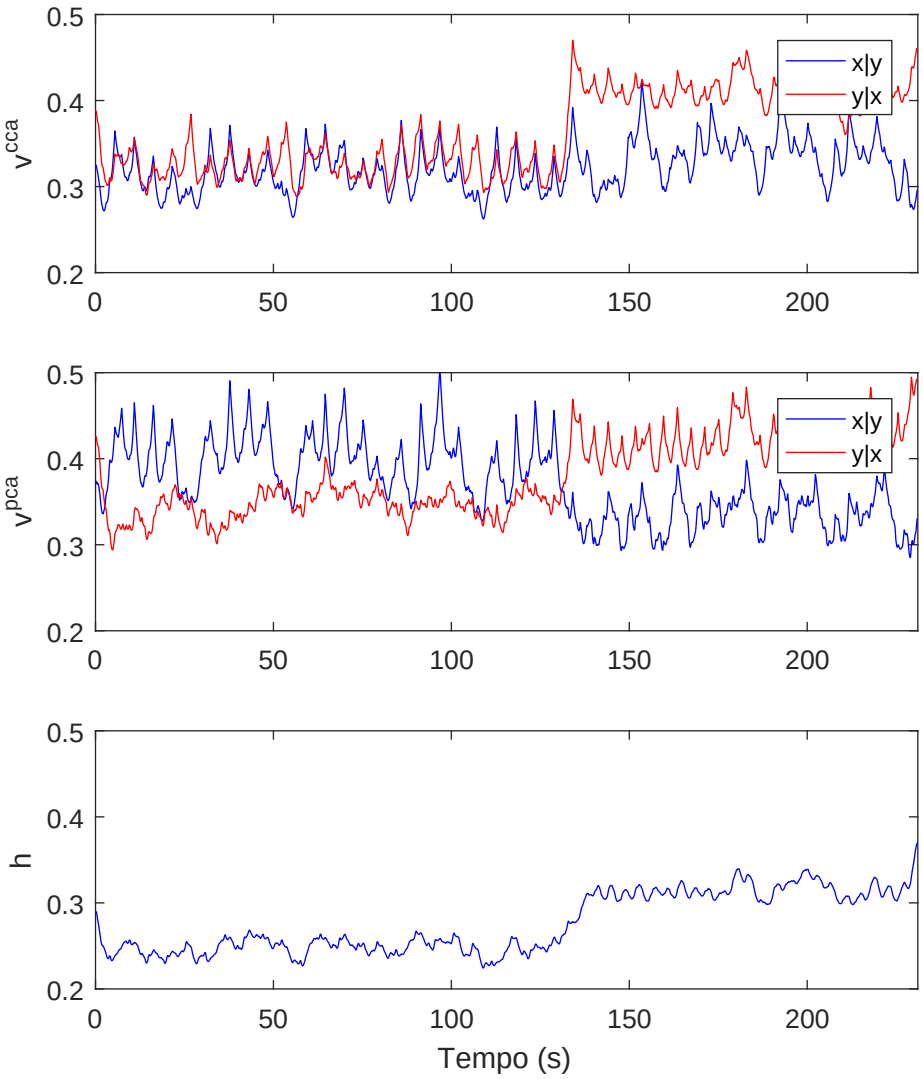

Figura 6.31: Coeficientes de associação $v^{cca}_{x|y}$ e $v^{cca}_{y|x}$, $v^{pca}_{x|y}$ e $v^{pca}_{y|x}$ e h para o caso 1D (com VAR)

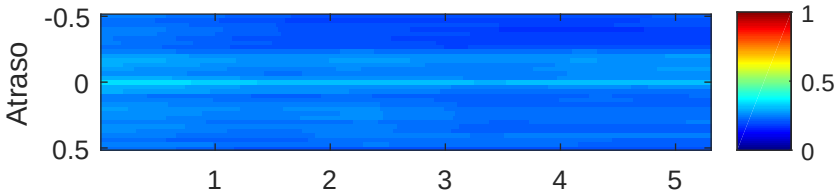

Figura 6.32: Mapa de associação gerado a partir do filtro média móvel para a sentença 1, com base no coeficiente h (Com VAR

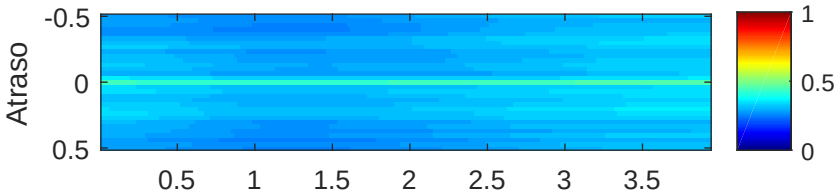

Figura 6.33: Mapas de associação gerado a partir do filtro média móvel para a sentença 2, com base no coeficiente h (Com VAR).

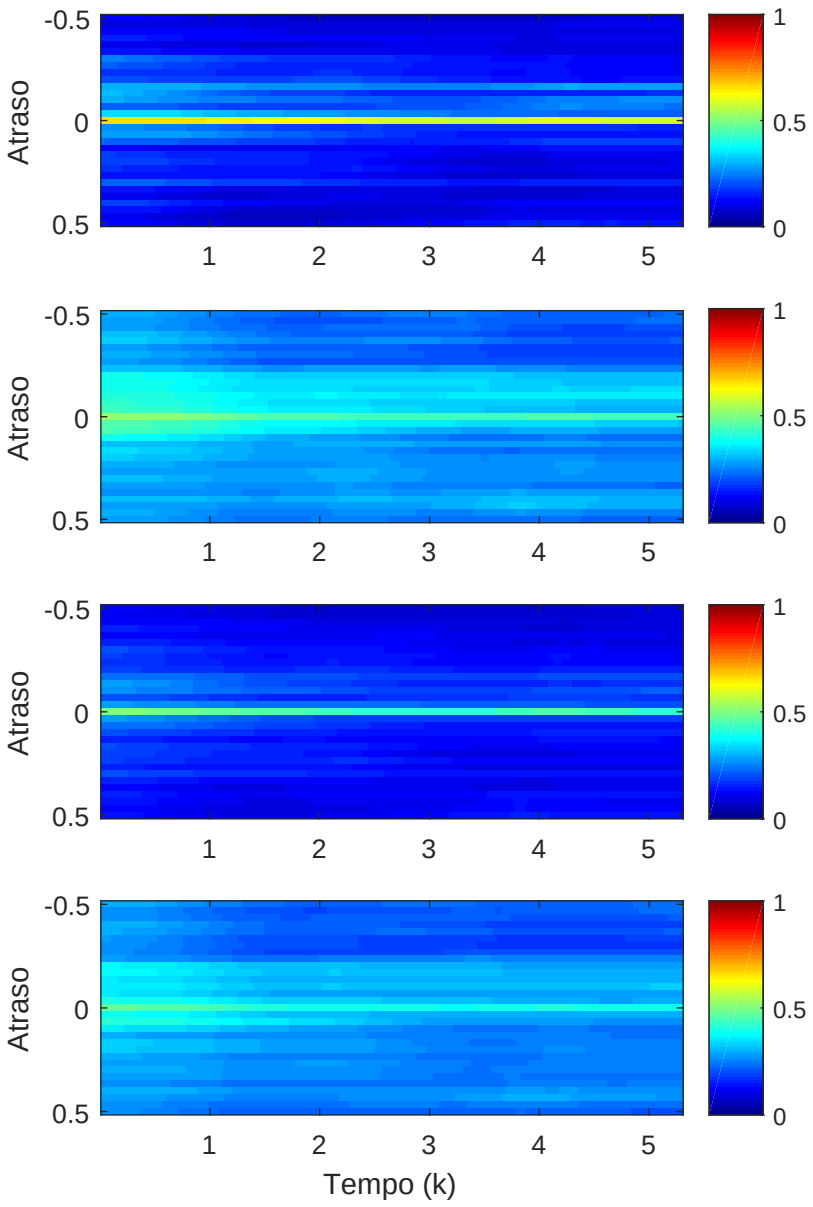

Figura 6.34: Mapas de associação gerados para a sentença 1 a partir do modelo média móvel. Os mapas correspondem aos coeficientes $v^{pca}_{x|y}$, $v^{pca}_{y|x}$, $v^{cca}_{x|y}$ e $v^{cca}_{y|x}$ (Com VAR)

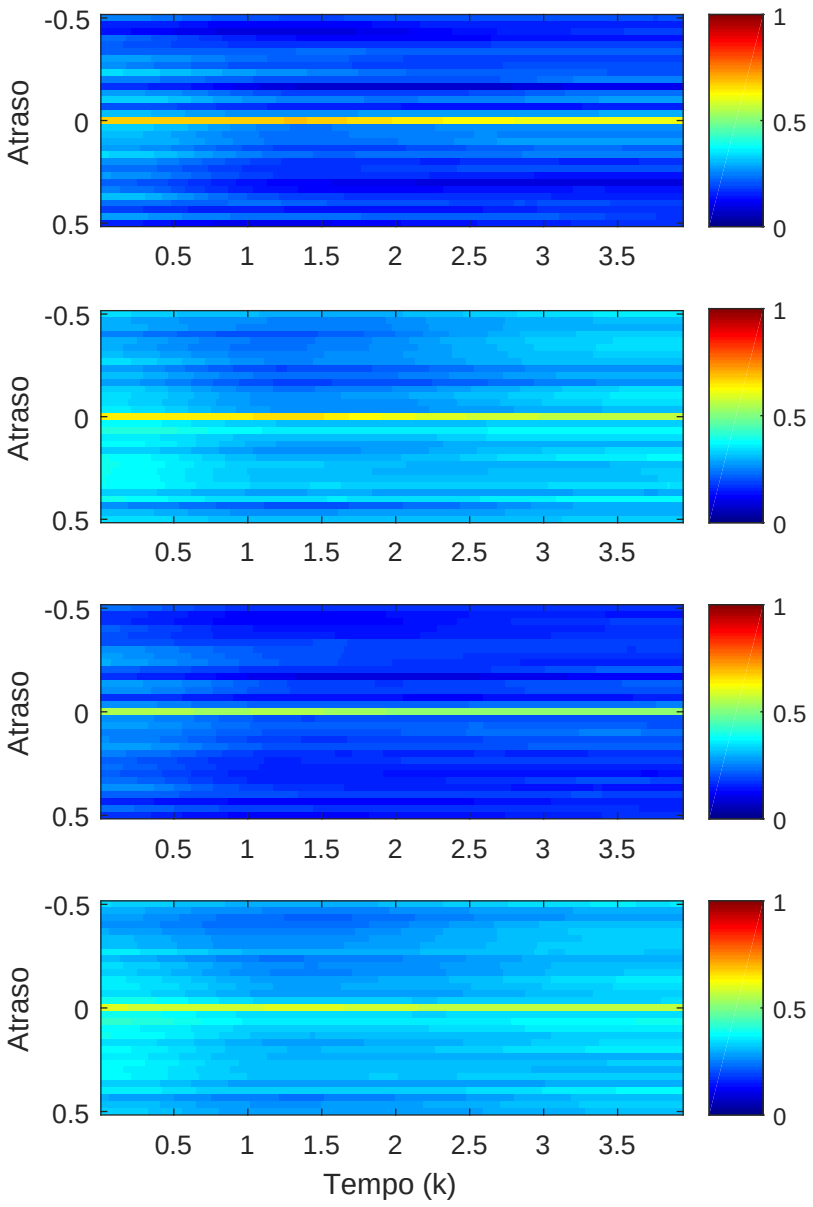

Figura 6.35: Mapas de associação gerados para a sentença 2 a partir do modelo média móvel. Os mapas correspondem aos coeficientes $v^{pca}_{x|y}$, $v^{pca}_{y|x}$, $v^{cca}_{x|y}$ e $v^{cca}_{y|x}$.

6.3 Base de Dados 3

Como mencionado no capítulo de associação entre os grupos de variáveis, uma forma de avaliar a associação entre dois grupos pode ser definida como a máxima correlação existente entre os domínios [5]. Tal medida *c* pode ser aplicável em arbitragem estatística, onde deseja-se encontrar um estimador de mínima variância entre pares/grupos de ativos [11].

No caso bivariado, máxima correlação significa mínima variância e, assim, a CCA, a regressão linear e a teoria do portfólio moderno são equivalentes, pois apresentam o mesmo resultado. Contudo, esta premissa não pode ser assumida na análise entre dois grupos de variáveis, pois caso cada grupo possua duas ou mais variáveis, é impossível realizar uma regressão linear. Neste caso, existem duas alternativas: estimar a combinação linear de variáveis que leve a uma máxima correlação (CCA) ou a uma mínima variância (*Markowitz* e/ou C-VaR).

Nesta seção será realizada, inicialmente, uma análise entre um par de variáveis, com o objetivo de avaliar as diferenças ocorridas no uso das variações percentuais ou das séries temporais puras como entrada do modelo. Também serão comparados os resultados apresentados pela medida de associação *c* e pelo *C-VaR*.

Em um segundo momento, o problema será expandido para o cenário multivariado, onde serão utilizadas as séries temporais de preços dos 10 ativos listados no capítulo 2 e armazenadas na tabela ***MPD***. Como no cenário multivariado, *c* e o portfólio neutro de *Markowitz* não apresentam o mesmo resultado, os métodos foram comparados, para o caso estático e variante no tempo. Nesta parte não foram apresentados resultados para o *C-VaR*, pois este último apresentou resultados semelhantes ao portfólio de mínima variância.

6.3.1 Pairs Trading e a associação entre um par de variáveis

Para as simulações desta seção serão utilizadas séries temporais de preços dos papéis VALE3 e VALE5 coletados com duas taxas de amostragem

diferentes.

Estas ações foram escolhidas, porque possuem alta liquidez dentre as pertencentes ao índice Bovespa, bem como porque são emitidas pela mesma empresa (siderurgica VALE S/A), o que demonstra alta chance de cointegração, já que os valores são influenciados pelos mesmos fatores.

Nas Tabelas 6.5 e 6.7 são apresentados os valores estimados e as precisões dos pesos do portfólio nas etapas de teste e treino. Nas Tabelas 6.6 e 6.8 são apresentados as médias e desvios padrões do risco atribuído aos métodos.

Uma vez que as medidas de risco possuem valores diferentes, uma boa prática é avaliar o método com base na incerteza das estimações dos pesos do portfólio. Como é possível notar, quando o retorno geométrico é utilizado, o nível de precisão aumenta significativamente dada a melhoria da relação sinal-ruído em relação ao uso dos retornos percentuais. Tal resultado é corroborado ao analisar as figuras 6.36 e 6.37 onde são apresentados os valores dos *spreads* resultantes para cada um dos portfólios neutros. Os resultados para o *VaR* não foram apresentados pela semelhança com os resultantes do *C-VaR*.

Até o presente momento, nenhuma análise foi realizada entre versões defasadas das séries temporais dos preços dos ativos, considerando que os ativos estão em fase. Para validar tal hipótese, os mapas de associação foram calculados sobre as séries temporais de preços de ativos. Nas figuras 6.38 e 6.39 são apresentados os mapas de associação gerados com os retornos geométricos para preços amostrados diariamente e a cada minuto respectivamente. Como é possível observar, para o caso com os dados com granularidade reduzida, os preços dos ativos podem não estar em fase, indicando que uma ação está respondendo mais rapidamente ao mercado do que a outra. Tal propriedade pode ser explorada no desenvolvimento de um algoritmo de *trading*.

Tabela 6.5: Valores médios de risco obtidos a partir das medidas *C-VaR* e de *Markowitz*, tendo como entrada retornos geométricos e percentuais. Os dados foram coletados com granularidade diária.

	$\bar{w}_x$	$\bar{w}_y$	σ_{w_x}	σ_{w_y}
Markowitz	0.8051	-0.3414	0.1387	0.1459
β **- CVAR**	0.8080	-0.3433	0.1353	0.1384
Markowitz - Retornos Geométricos	0.5369	-0.4986	0.0070	0.0005

Tabela 6.6: Parâmetros dos pesos estimados para uma estratégia de arbitragem estatística entre dois ativos. Os dados foram extraídos com granularidade diária.

	μ_{risco} **treino**	σ_{risco} **treino**	μ_{risco} **teste**	σ_{risco} **teste**
Markowitz	3.4063e-04	1.8485e-04	3.8980e-04	2.2196e-04
β **- CVAR**	0.0393	0.0088	0.0330	0.0079
Markowitz - Geom.	1.4027e-06	2.4295e-05	6.0017e-07	1.0395e-05

Tabela 6.7: Valores médios de risco obtidos a partir das medidas *C-VaR* e de *Markowitz*, tendo como entrada retornos geométricos e percentuais. Os dados foram coletados a cada minuto.

	$\bar{w}_x$	$\bar{w}_y$	σ_{w_x}	σ_{w_y}
Markowitz	0.9716	-0.0955	0.0600	0.1222
β **- CVAR**	0.9531	-0.1250	0.0836	0.1489
Markowitz - Retornos Geométricos	0.6938	-0.4607	0.0138	0.0058

Tabela 6.8: Parâmetros dos pesos estimados para uma estratégia de arbitragem estatística entre dois ativos. Os dados foram extraídos a cada minuto.

	μ_{risco} **treino**	σ_{risco} **treino**	μ_{risco} **teste**	σ_{risco} **teste**
Markowitz	1.7915e-06	7.0644e-07	1.9520e-06	7.8833e-07
β **- CVAR**	0.0030	7.9641e-04	0.0023	3.8063e-04
Markowitz - Geom.	8.9093e-05	1.3831e-05	1.5243e-06	1.0368e-06

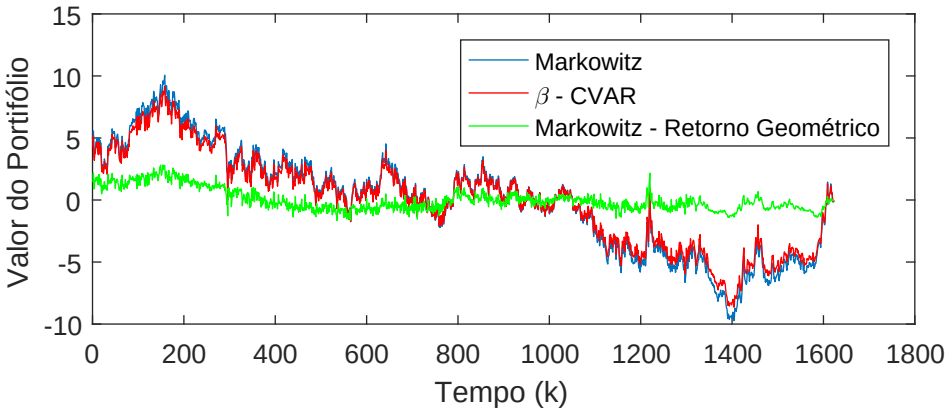

Figura 6.36: Resultados dos valores do *spread*, dados com granularidade diária.

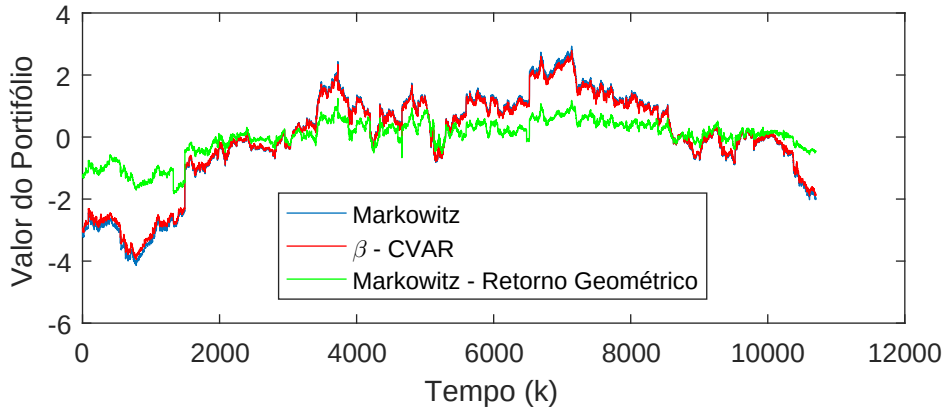

Figura 6.37: Resultados dos valores do *spread*, dados com granularidade de um minuto.

6.3.2 Arbitragem estatística no caso multivariado e associação entre grupos de variáveis

Nesta parte do capítulo os testes realizados na seção anterior serão expandidos para o cenário multivariado. As ações selecionadas para o estudo são apresentadas na Figura 6.40. No gráfico, estão representadas três grandezas que descrevem as principais características dos papéis. O eixo das abcissas representa o valor médio de risco do ativo, sendo definido aqui como o desvio padrão dos retornos percentuais. No eixo das or-

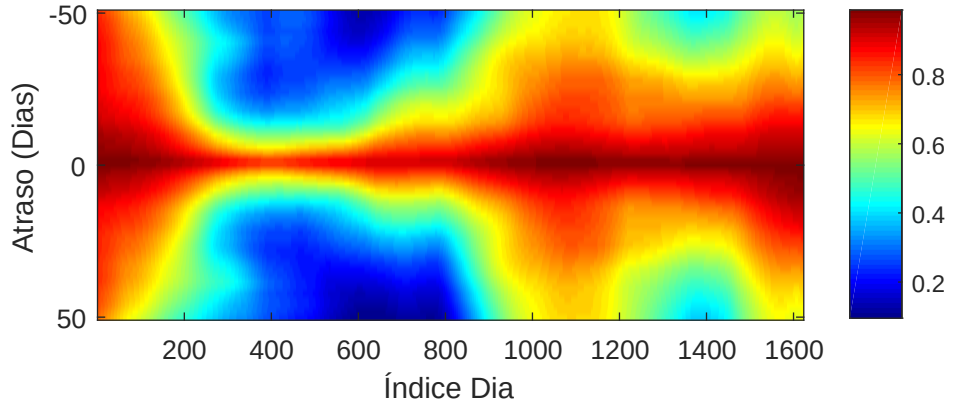

Figura 6.38: Mapas de associação para os valores dos retornos geométricos das ações *VALE3* e *VALE5* com amostragem diária.

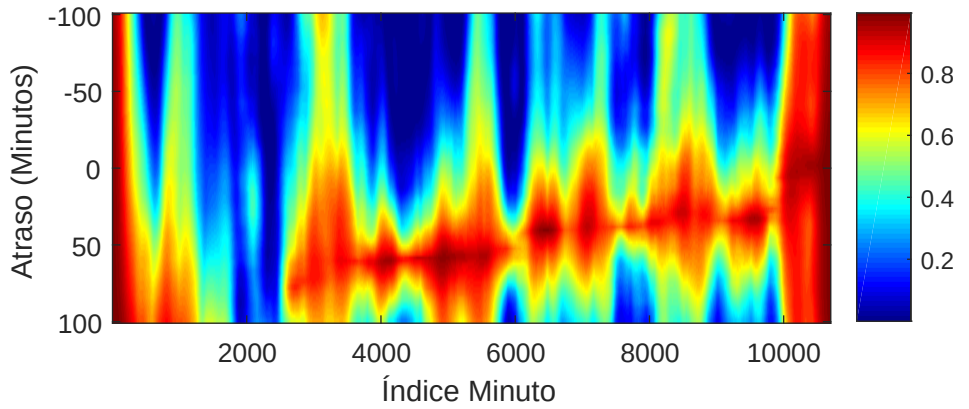

Figura 6.39: Mapas de associação para os valores dos retornos geométricos das ações *VALE3* e *VALE5* com dados coletados a cada minuto.

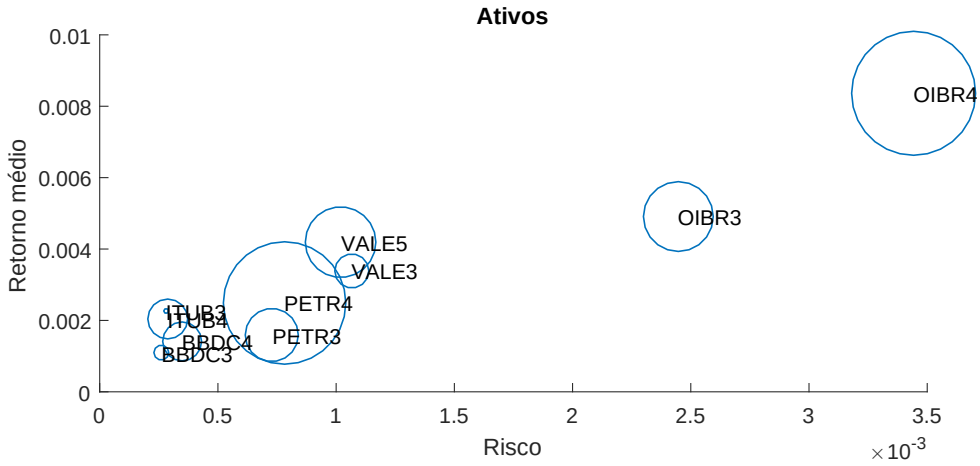

Figura 6.40: Matriz Risco x Retorno Médio x Liquidez

denadas é apresentado o valor médio dos retornos percentuais de cada um dos ativos. O tamanho da bolha associada a cada ativo descreve uma medida de liquidez: o número médio de lotes daquele papel negociados diariamente.

Em um primeiro momento será realizada a avaliação do portfólio para múltiplos ativos e a divisão destes em dois subgrupos. No experimento foi utilizada validação cruzada, em que aplicou-se 100 amostras para cada etapa de teste e treino e o procedimento foi repetido 100 vezes. Em todas as 100 etapas de treino realizadas, o peso do portfólio foi multiplicado pelo sinal do peso referente a PETR4, de forma que este último fosse sempre positivo. Isso garante que não exista etapas de treino onde dois grupos possuam a mesma correlação canônica com os sinais dos pesos invertidos. Os histogramas dos pesos são apresentados na Figura 6.41 para os retornos percentuais e 6.42 para os retornos baseados no retorno geométrico. Ao contrário, no caso onde existem duas variáveis, não existe uma variação muito grande nos resultados, o que indica que quando o número de ativos aumenta o impacto da relação sinal ruído é reduzido.

Todavia, deve-se ressaltar que os histogramas dos pesos estimados para as diferentes bases de treino indicam que as estimações se comportam como uma mistura de gaussianas. Isso ocorre pelo fato das ações terem sido escolhidas em pares (emitidas pela mesma companhia), o que

gera certa instabilidade aos métodos. por possuírem comportamento semelhante [7].

A solução proposta para transformar as misturas de gaussianas em uma única gaussiana foi definir, a priori, qual o sinal do peso de cada ativo. Este foi definido como o sinal do valor médio das estimações encontradas para as diversas etapas de treino e é mostrado na tabela 6.9. Como descrito no capítulo que descreve a teoria de arbitragem estatística, os ativos são divididos em grupos de acordo com o sinal do peso associado a este. Como os valores dos resultados para as etapas de treino foram muito próximos tanto para os retornos geométricos quanto para os log-retornos, foi apresentada somente uma tabela.

Definidos os ativos em dois grupos, a mesma simulação foi realizada restringindo o valor do coeficiente dentro do intervalo $[0, 1]$ para os ativos com sinal positivo e $[-1, 0]$ para os ativos com sinal negativo. O resultado do procedimento para quando os dados de entrada foram os retornos percentuais é apresentado na Figura 6.43 e para quando os dados de entrada foram os retornos geométricos na Figura 6.44. Esta simulação apresenta um resultado importante e contundente para este trabalho: estimar os pesos buscando encontrar a mínima variância é mais eficiente que estimar os mesmos buscando encontrar máxima máxima correlação entre os grupos, pois a precisão do estimador se torna muito maior.

Nas figuras 6.45 e 6.46 são apresentados histogramas dos valores de treino e de teste das correlações entre os grupos e da variância entre eles para as diferentes etapas de treino e de teste. Os resultados estão muito próximos, demonstrando que mesmo no caso multivariado, máxima correlação e mínima variância caminham juntas.

Até o momento, foi realizada uma avaliação estática dos pesos do portfólio e, por este motivo, não foi considerada uma relação de como a volatilidade e os pesos do portfólio variam em um sistema tão dinâmico. Assim, pode ser que a mistura de gaussianas presente nas figuras 6.41 e 6.42 surge pelo fato de que os ativos podem estar mudando de grupo ao longo do tempo. Por este motivo, o filtro proposto por [2] foi utilizado para se avaliar matrizes de correlação instantâneas.

Tabela 6.9: Valor médio e desvio padrão da simulação feita para os grupos quando os sinais dos pesos ainda não foram definidos.

	$\bar{w}$ **CCA**	σ_w **CCA**	$\bar{w}$ **MKW**	σ_w **MKW**
PETR4	0.1234	0.0370	0.3351	0.0473
PETR3	-0.1212	0.0326	-0.2944	0.0420
ITUB3	0.0188	0.1005	0.0190	0.0415
ITUB4	-0.0451	0.1330	-0.0629	0.0550
VALE3	0.0844	0.1098	0.0670	0.0386
VALE5	-0.0867	0.1196	-0.0654	0.0425
OIBR3	0.0022	0.0138	0.0013	0.0114
OIBR4	-0.0027	0.0121	-0.0033	0.0089
BBDC3	-0.0214	0.1187	-0.0108	0.0516
BBDC4	0.0443	0.1292	0.0146	0.0714

Na Figura 6.47 é apresentado um mapa de calor que descreve o valor instantâneo dos coeficientes relacionados a cada um dos ativos selecionados e em 6.48 o mesmo resultado é gerado tendo como entrada os retornos geométricos. As figuras indicam um número excessivo de mudança de grupo dos ativos ao longo do tempo, indicando que mesmo introduzindo uma adaptatividade a variações na matriz de covariância, uma definição prévia dos sinais dos pesos atribuídos aos ativos se mostra necessária. Na Figura 6.49 são apresentados os resultados dos valores dos pesos para o estimador de mínima variância quando os grupos são previamente definidos. Neste caso, é possível observar que os pesos variam ao longo do tempo, todavia de forma mais suave, que é identificado pelas suaves transições entre vermelhos e azuis mais claros e mais escuros.

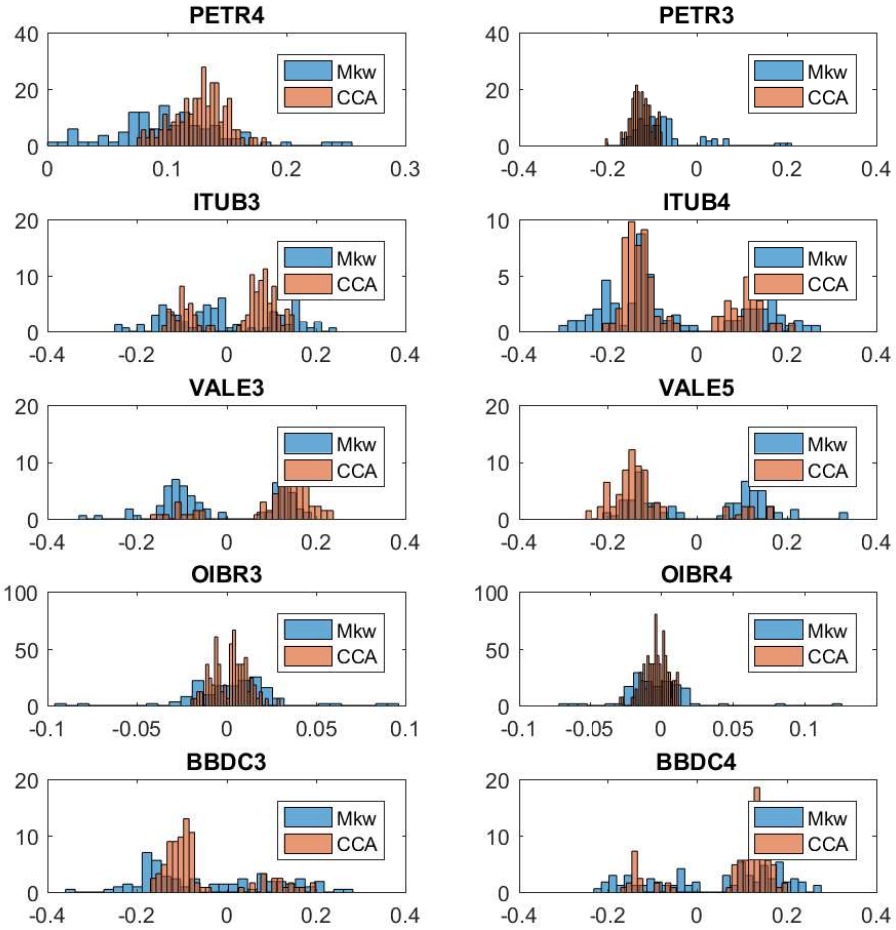

Figura 6.41: Histograma dos pesos do portfólio neutro em relação ao mercado obtidos com o critério de mínima variância de Markowitz e de máxima correlação obtido pela CCA. Nesta simulação, as entradas do modelo foram os retornos percentuais e os grupos não foram definidos a priori.

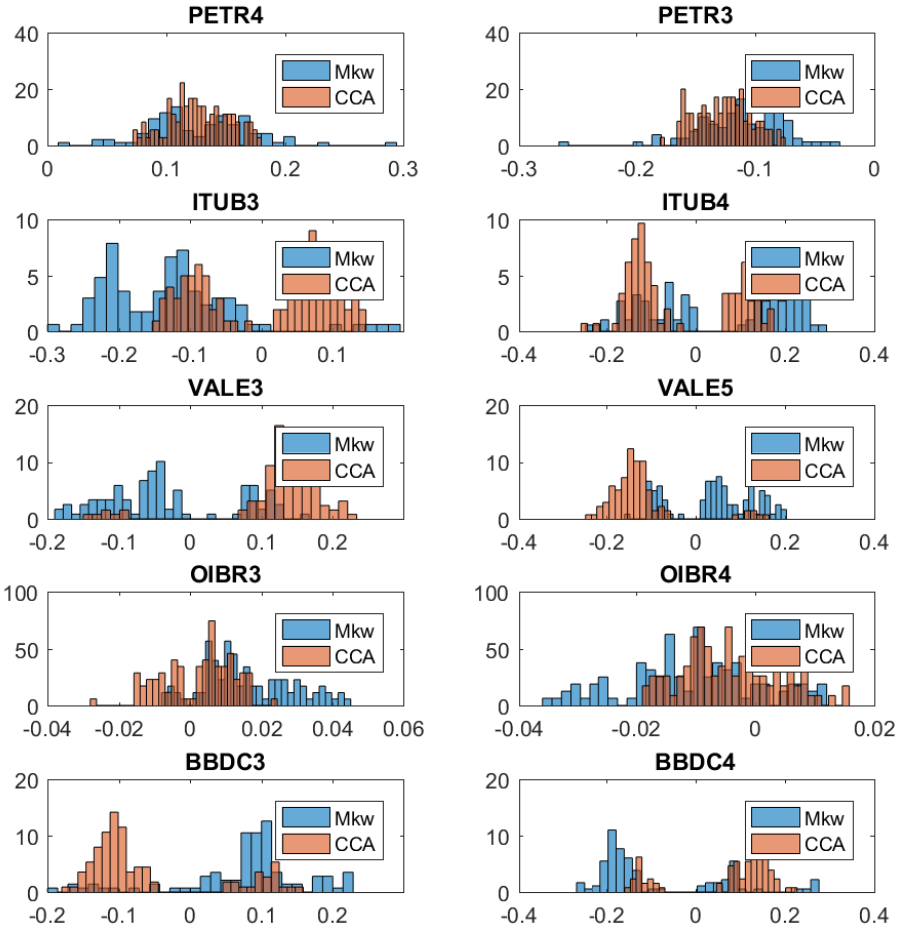

Figura 6.42: Histograma dos pesos do portfólio neutro em relação ao mercado obtidos com o critério de mínima variância de Markowitz e de máxima correlação obtido pela CCA, tendo como base os retornos geométricos para o cálculo da matriz de covariância ao invés das variações percentuais. Os grupos não foram definidos.

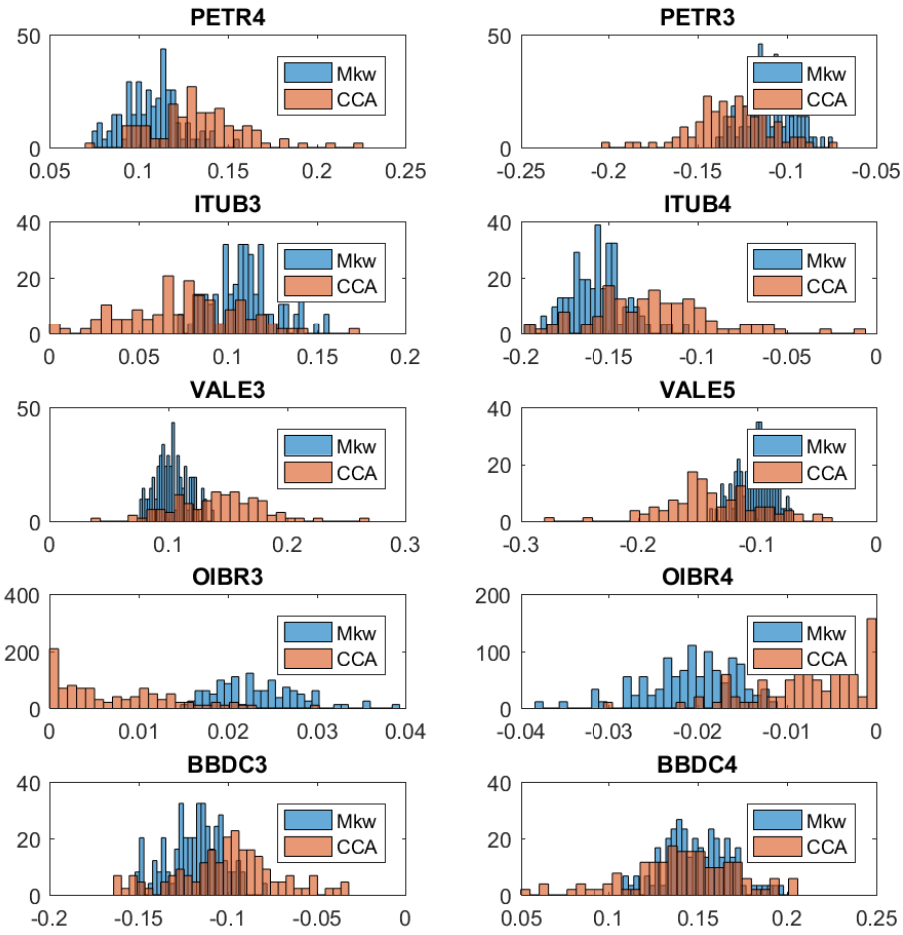

Figura 6.43: Histograma dos pesos do portfólio neutro em relação ao mercado obtidos com o critério de mínima variância de Markowitz e de máxima correlação obtido pela CCA. Nesta simulação os grupos foram definidos e os dados de entrada do modelo foram os retornos percentuais.

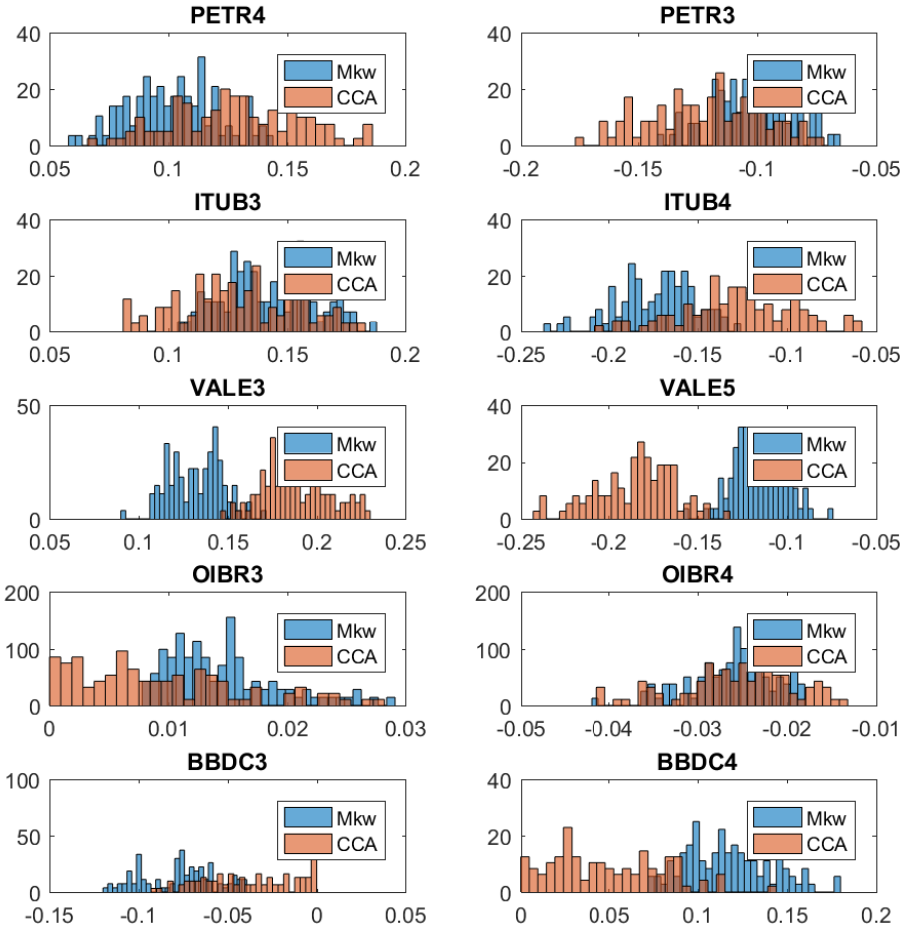

Figura 6.44: Histograma dos pesos do portfólio neutro em relação ao mercado obtidos com o critério de mínima variância de Markowitz e de máxima correlação obtido pela CCA. Nesta simulação os grupos foram definidos e os dados de entrada do modelo foram os retornos geométricos.

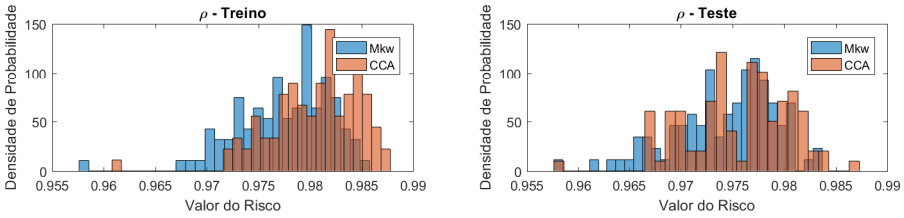

Figura 6.45: Valor da correlação dos dois portfólios estabelecidos (Mkw e CCA).

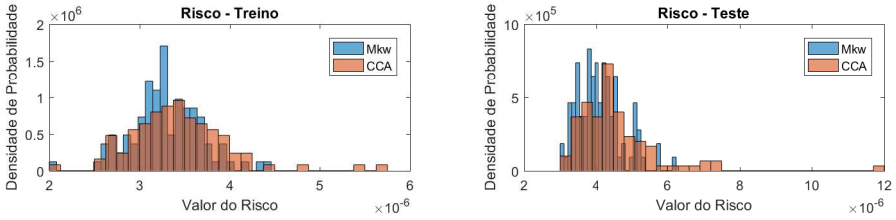

Figura 6.46: Valor da variância (ou risco) dos dois portfólios estabelecidos (Mkw e CCA).

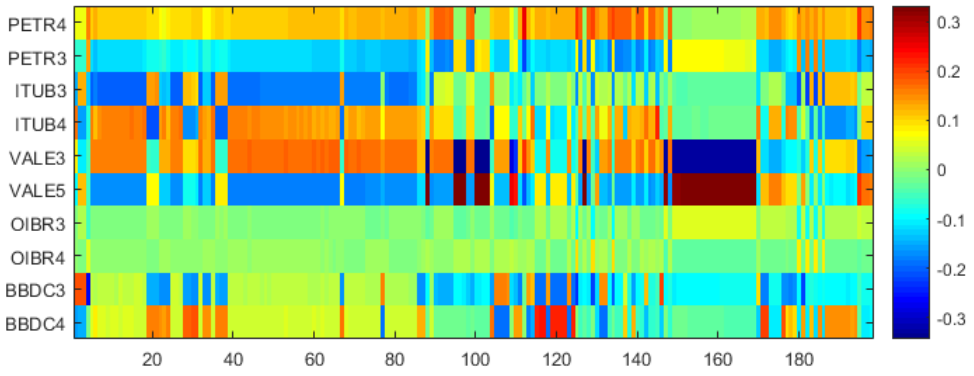

Figura 6.47: Mapa de calor que apresenta como os valores dos pesos do portfólio variam ao longo do tempo para cada um dos ativos selecionados. A granularidade dos dados é diária e foram utilizados os log-retornos.

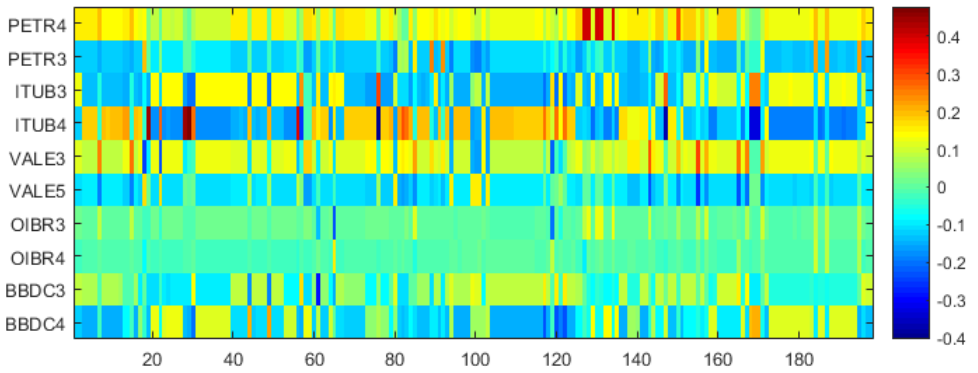

Figura 6.48: Mapa de calor que apresenta como os valores dos pesos do portfólio variam ao longo do tempo para cada um dos ativos selecionados. A granularidade dos dados é diária e foram utilizados os retornos geométricos.

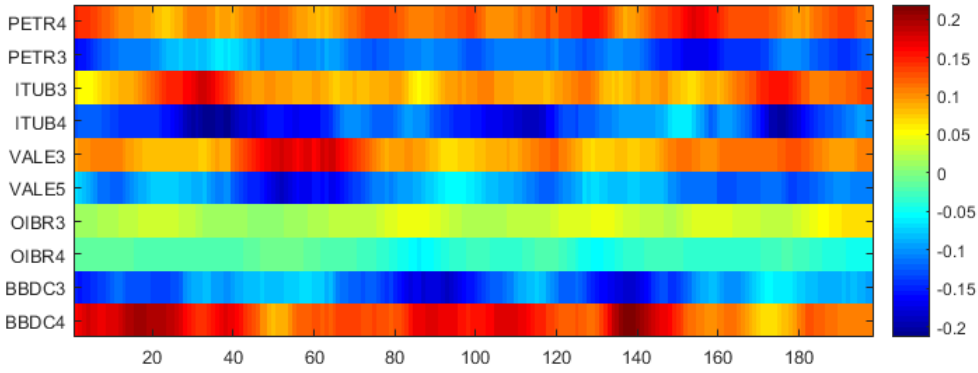

Figura 6.49: Mapa de calor que apresenta como os valores dos pesos do portfólio variam ao longo do tempo para cada um dos ativos selecionados. Nesta simulação os grupos foram definidos, restringindo assim o sinal dos pesos. A granularidade dos dados é diária.

Comentários Finais

Neste capítulo foram apresentados os resultados do método desenvolvido sobre cada uma das bases de dados selecionadas. No caso da primeira base de dados, os resultados permitiram realizar uma comparação entre a associação instantânea no caso bivariado e multivariado. Dentre os principais resultados, o aumento no número de variáveis para representar cada domínio auxiliou na detecção do momentos de maior associação, todavia, foi apresentada uma perda de precisão na detecção de um caminho de máxima associação sobre o mapa. Um outro resultado a ser destacado foi a provável relação entre o valor da associação instantânea e a representatividade da primeira componente principal dentro de cada grupo, indicando um tema para ser melhor investigado em trabalhos futuros.

No caso da *base 2*, quando o método foi utilizado para encontrar a associação entre o movimento do trato vocal e da face, o principal resultado encontrado foi que existe uma relação entre estes domínios e a potência do sinal de voz encontrado. Os mapas de associação indicaram que os domínios estão em fase.

Para a *base 3*, no caso bivariado foi encontrado um atraso entre os preços dos ativos *VALE3* e *VALE5* quando o valor dos preços destes foram amostrados a cada minuto, indicando a possibilidade de se utilizar o mapa de associação em algoritmos do tipo *daytrade*, onde operações são realizadas ao longo do dia. Além disso, o método provou ser mais efetivo quando alimentado pelos valores dos retornos geométricos ao invés dos retornos percentuais (ou log-retornos). No caso multivariado, foi apresentado que os grupos de ativos devem ser definidos a priori antes de se utilizar a ferramenta. Além disso, foi concluído que é mais robusto neste tipo de aplicação buscar encontrar a mínima variância entre os grupos ao invés da máxima correlação, dada a incerteza sobre os valores dos pesos.

A escolha da medida de associação está relacionada com o objetivo de interesse. Caso o objetivo seja mapear um domínio a partir do outro e descrever como esta relação varia ao longo do tempo [8], as medidas de associação que tomam como base variância compartilhada (v) são mais

indicadas. Caso o objetivo seja quantificar a coordenação entre grupos de variáveis, talvez a medida de associação baseada na probabilidade de descorrelação (h) seja mais útil, pois apresenta uma medida única de coordenação, como no caso bivariado [2]. No caso do mercado financeiro, como o objetivo é encontrar a máxima correlação entre os grupos (c), não faz sentido o uso de outras medidas.

Capítulo 7

Conclusão

A aplicação do método implementado para o cálculo da associação entre grupos de variáveis sobre as três bases de dados apresentadas possibilitou realizar uma análise do desempenho deste em três cenários diferentes.

Quando o método estimou a associação instantânea entre os movimentos dos tratos vocais de dois locutores, três resultados merecem destaque: o maior destaque os instantes onde a associação é mais elevada, representados por colocações mais avermelhadas sobre os mapas; a relação entre a razão da variância da primeira componente principal pela variância total do grupo e o valor calculado da associação instantânea entre grupos; e a detecção da representatividade instantânea de cada variável na associação entre os domínios.

No caso da associação entre os movimentos do trato vocal e da face, foi detectado que o volume de informação redundante entre os grupos pode estar relacionado com a potência do sinal de voz emitido pelo locutor. Esta informação pode auxiliar a criar animações de face mais realistas e, consequentemente, auxiliar em sistemas de codificação audiovisual da fala.

No caso da aplicação em arbitragem, ambos os métodos (portfólio de mínima variância e máxima correlação) forneceram resultados equivalentes em relação ao risco. Todavia, os pesos atribuídos a cada um dos ativos se mostraram mais estáveis com estimador de mínima variância que do

que com o de máxima correlação. Outro resultado da simulação está condicionado a escolha entre remover ou não remover a redundância entre as amostras das séries temporais. No caso bivariado, eliminar a correlação entre amostras reduziu a relação sinal-ruído a ponto do estimador não conseguir encontrar o real valor da associação e os respectivos pesos. Entretanto, no caso multivariado, tal tipo de erro é reduzido pelo aumento do número de ativos sendo a solução encontrada idêntica para ambos os dados de entrada.

O método da forma como foi implementado já pode auxiliar a detectar padrões em uma série de estudos futuros, porém existem melhorias a serem realizadas sobre o algoritmo que não foram desenvolvidas neste trabalho. O seu desempenho pode ser melhorado eliminando etapas que aumentam o custo computacional como, por exemplo, as operações de raíz quadrada sobre matriz necessárias para calcular as componentes canônicas. Com uma eventual melhoria da eficiência do método, seria possível utilizar outros algoritmos para se avaliar a matriz de covariância que fossem mais robustos que o filtro média móvel exponencial e encontrar resultados superiores aos descritos neste trabalho.

Outra melhoria possível no algoritmo seria adaptá-lo para considerar valores de atrasos diferentes para variáveis dentro do mesmo grupo, pois em seu formato atual, o método considera que todas as variáveis dentro de cada grupo estão em fase. Caso tal análise seja realizada, o caminho de máxima associação pode ficar mais nítido, todavia, será necessária uma outra estrutura de visualização de dados que suporte todas as combinações de atrasos entre variáveis possíveis.

Este trabalho tem como principal contribuição a expansão dos mapas de correlação do caso bivariado para o multivariado, aqui chamados de mapas de associação, possibilitando, assim, que variáveis inicialmente descartadas possam ser consideradas em novos estudos. A utilização do método em processamento audiovisual da fala e finanças quantitativas ilustra como este transcende as áreas do conhecimento, e assim, espera-se que o mesmo possa auxiliar pesquisadores, independentemente do ramo de estudo.

Capítulo 8

Anexos

8.1 Teste de *Jarque-Bera*

O teste de Jarque-Beta calcula um coeficiente que estima a probabilidade de que amostras tenham sido encontradas por meio de uma distribuição normal, encontrado matematicamente via

$$JB = \frac{n}{6}(S^2 + \frac{(K-3)^2}{4}) \tag{8.1}$$

onde S é o valor do terceiro momento ponderado pela variância

$$S = \frac{\hat{\mu}_3}{\hat{\mu}_2^{3/2}} \tag{8.2}$$

e K é o momento de quarta ordem ponderado pela variância

$$K = \frac{\hat{\mu}_4}{\hat{\mu}_2^2} \tag{8.3}$$

onde μ_i é a estimativa do momento central de ordem i que é dado por

$$\hat{\mu}_i = \frac{1}{N}\sum_{k=1}^{N}(x_k - \hat{x})^i \tag{8.4}$$

Bibliografia

[1] E. Vatikiotis-Bateson, A. V. Barbosa, and C. T. Best, "Articulatory coordination of two vocal tracts," *Journal of Phonetics*, vol. 44, pp. 167–181, May 2014.

[2] A. V. Barbosa, R.-M. Déchaine, E. Vatikiotis-Bateson, and H. C. Yehia, "Quantifying time-varying coordination of multimodal speech signals using correlation map analysis," *The Journal of the Acoustical Society of America*, vol. 131, pp. 2162–2172, Mar. 2012.

[3] M. M. Mukaka, "A guide to appropriate use of Correlation coefficient in medical research," *Malawi Medical Journal*, vol. 24, pp. 69–71, Jan. 2012.

[4] R. J. Harris, *A Primer of Multivariate Statistics*. Psychology Press, May 2001. Google-Books-ID: fRRWBQAAQBAJ.

[5] N. H. Timm, *Applied Multivariate Analysis*. Springer Science & Business Media, June 2002. Google-Books-ID: PyLMNcpuoEwC.

[6] M. Borga, "Canonical correlation: a tutorial," *On line tutorial http://people. imt. liu. se/magnus/cca*, vol. 4, p. 5, 2001.

[7] M. I. Alpert and R. A. Peterson, "On the Interpretation of Canonical Analysis," *Journal of Marketing Research (JMR)*, vol. 9, pp. 187–192, May 1972.

[8] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative Association of Vocal-tract and Facial Behavior," *Speech Commun.*, vol. 26, pp. 23–43, Oct. 1998.

[9] B. H. Repp and Y.-H. Su, "Sensorimotor synchronization: a review of recent research (2006-2012)," *Psychonomic Bulletin & Review*, vol. 20, pp. 403–452, June 2013.

[10] M. Baxter and R. G. King, "Measuring Business Cycles Approximate Band-Pass Filters for Economic Time Series," Working Paper 5022, National Bureau of Economic Research, Feb. 1995.

[11] A. Pole, *Statistical Arbitrage: Algorithmic Trading Insights and Techniques*. John Wiley & Sons, July 2011. Google-Books-ID: xSjXTnKqIKoC.

[12] E. Gatev, W. N. Goetzmann, and K. G. Rouwenhorst, "Pairs Trading: Performance of a Relative-Value Arbitrage Rule," *Review of Financial Studies*, vol. 19, pp. 797–827, Sept. 2006.

[13] G. Vidyamurthy, *Pairs Trading: Quantitative Methods and Analysis*. John Wiley & Sons, Feb. 2011.

[14] S. Salvador and P. Chan, "Toward Accurate Dynamic Time Warping in Linear Time and Space," *Intell. Data Anal.*, vol. 11, pp. 561–580, Oct. 2007.

[15] T. Giorgino, "Computing and visualizing dynamic time warping alignments in R: the dtw package," *Journal of statistical Software*, vol. 31, no. 7, pp. 1–24, 2009.

[16] P. D. J. Benesty, J. Chen, Y. Huang, and P. I. Cohen, "Pearson Correlation Coefficient," in *Noise Reduction in Speech Processing*, no. 2 in Springer Topics in Signal Processing, pp. 1–4, Springer Berlin Heidelberg, 2009.

[17] L. Aguirre, *Introdução à Identificação de Sistemas – Técnicas Lineares e Não-Lineares Aplicadas a Sistemas Reais*. Editora UFMG.

[18] D. Stewart and W. Love, "A general canonical correlation index," *Psychological Bulletin*, vol. 70, pp. 160–163, Sept. 1968.

[19] S. S. Wilks, "Certain Generalizations in the Analysis of Variance," *Biometrika*, vol. 24, no. 3/4, pp. 471–494, 1932.

[20] C. R. Rao and H. Yanai, "General definition and decomposition of projectors and some applications to statistical problems," *Journal of Statistical Planning and Inference*, vol. 3, pp. 1–17, Jan. 1979.

[21] S. Srkk, *Bayesian Filtering and Smoothing*. New York, NY, USA: Cambridge University Press, 2013.

[22] T. Bollerslev, R. F. Engle, and D. B. Nelson, "Chapter 49 Arch models," vol. 4, pp. 2959–3038, Elsevier, 1994.

[23] N. Sugamura and F. Itakura, "Speech analysis and synthesis methods developed at ECL in NTT — From LPC to LSP —," *Speech Communication*, vol. 5, pp. 199–215, June 1986.

[24] L. R. Rabiner and B. Gold, *Theory and application of digital signal processing*. Prentice-Hall, 1975. Google-Books-ID: iAxTAAAAMAAJ.

[25] S. Furui, *Digital Speech Processing: Synthesis, and Recognition, Second Edition,*. Taylor & Francis, Nov. 2000.

[26] Oppenheim, *Discrete-Time Signal Processing*. Pearson Education, 1999.

[27] L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition*. Englewood Cliffs, N.J.: PTR Prentice Hall, 1993. OCLC: 26674087.

[28] J. D. Hamilton, *Time Series Analysis*. Princeton University Press, Jan. 1994.

[29] B. G. Malkiel, "Efficient Market Hypothesis," in *The World of Economics* (J. Eatwell, M. Milgate, and P. Newman, eds.), The New Palgrave, pp. 211–218, Palgrave Macmillan UK, 1991.

[30] A. Timmermann and C. W. J. Granger, "Efficient market hypothesis and forecasting," *International Journal of Forecasting*, vol. 20, pp. 15–27, Jan. 2004.

[31] G. Appel, *Technical Analysis: Power Tools for Active Investors*. FT Press, first ed., 2005.

[32] B. G. Malkiel and E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work*," *The Journal of Finance*, vol. 25, pp. 383–417, May 1970.

[33] R. J. Elliott, J. Van Der Hoek *, and W. P. Malcolm, "Pairs trading," *Quantitative Finance*, vol. 5, pp. 271–276, June 2005.

[34] R. F. Engle and C. W. J. Granger, "Co-Integration and Error Correction: Representation, Estimation, and Testing," *Econometrica*, vol. 55, no. 2, pp. 251–276, 1987.

[35] J. H. Stock and M. W. Watson, "A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems," *Econometrica*, vol. 61, no. 4, pp. 783–820, 1993.

[36] R. T. Rockafellar and S. Uryasev, "Optimization of Conditional Value-at-Risk," *Journal of Risk*, vol. 2, pp. 21–41, 2000.

[37] S. M. Ross, *A First Course in Probability*. Pearson Prentice Hall, 2010. Google-Books-ID: Bc1FAQAAIAAJ.