Medidas de Fluxo de Informação com Aplicação em Neurociência

Daniel Yasumasa Takahashi

Tese apresentada ao

Programa Interunidades de Pós-graduação

em Bioinformática

da

Universidade de São Paulo

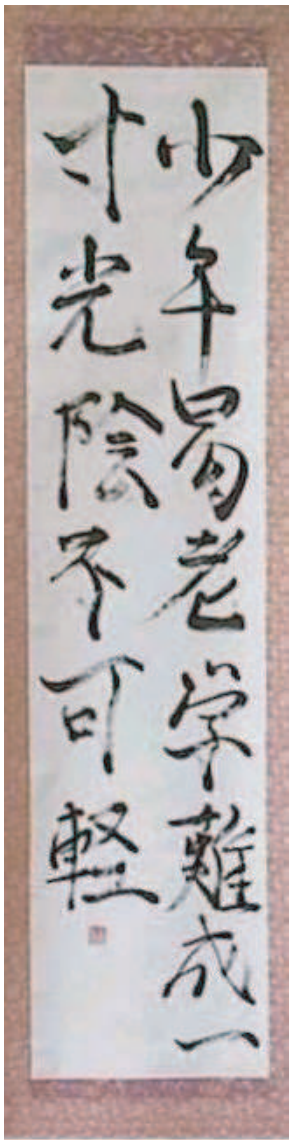
Orientador: Koichi Sameshima

Co-orientador: Luiz Antonio Baccalá

São Paulo 2008

Este trabalho foi realizado com o financiamento da CAPES

(Bolsa de Doutorado).



Frase atribuída a Chu Hsi (1130-1200), porém cuja autoria vem sendo contestada recentemente como sendo de um autor japonês. Corresponde à versão chinesa (ou japonesa) da frase “Ars longa, vita brevis” (Hipócrates). Obra de M. Nishino.

Agradecimentos

Esta tese é fruto de uma vida dedicada ao estudo, ao conhecimento, à pesquisa científica. Portanto, para mim, concluir este trabalho significa vencer um importante desafio. Uma vitória que só foi possível graças a colaboração de pessoas muito especiais:

Professor Koichi Sameshima. Foi ele quem me apresentou a possibilidade de utilizar a matemática no estudo da Neurociência. Mais do que um orientador, foi a pessoa que me guiou no caminho, muitas vezes tortuoso, da pesquisa científica.

Professor Luiz Antonio Baccalá, co-orientador deste trabalho. Sem seu espírito crítico, mas estimulante, questões levantadas na tese correriam o risco de ficar sem solução.

Luiz Henrique Lana, grande amigo com quem mantive longas discussões filosóficas, matemáticas e, principalmente, neurocientíficas, que me ajudaram na produção e finalização deste trabalho.

Professor João Ricardo Sato, amigo, colaborador científico e, acima de tudo, especialista em solucionar problemas estatísticos.

Patrícia Martorelli, competente secretária do programa de pós-graduação em

Bioinformática.

Tenho convicção de que, mais do que minha, essa conquista é principalmente de meus pais. Indiretamente, o processo para a conclusão desta tese revela valores que aprendi com eles, ao lado de meu irmão e irmã: respeito, honestidade, esforço, paciência, curiosidade e, sobretudo, dedicação.

Por fim, não poderia deixar de agradecer a Daiane Tamanaha com quem agora compartilho minha vida.

Prefácio

A proposta inicial do trabalho de tese era estudar a coerência parcial direcionada, medida esta desenvolvida por Koichi Sameshima e Luiz Antonio Baccalá, como medida de dependência direcionada relacionando-a com o conceito de causalidade de Granger e aplicá-la em dados experimentais de neurofisiologia. Durante o desenvolvimento da tese, ficou claro que o entendimento teórico da coerência parcial direcionada só seria possível se inserida num escopo maior de comparação entre medidas de dependência para processos estacionários de segunda ordem, o que modificou ligeiramente a forma da tese, embora mantendo o objetivo inicial. Também ficou claro que a aplicação de qualquer medida de inferência deveria ser amparada em resultados estatísticos assintóticos rigorosos sobre o comportamento dos estimadores, mesmo que estes sejam, no melhor dos casos, apenas aproximações grosseiras do comportamento observado. Esta última parte do trabalho não foi incluída, embora seja importante, porque tornaria a tese pouco concisa. Há três trabalhos publicados Takahashi et al. (2008, 2007); Baccalá et al. (2006), incluídos nos anexos, referentes ao comportamento estatístico de algumas medidas de dependência discutidas nesta tese

do ponto de vista de processos estocásticos.

Além dos objetivos mais específicos apresentados acima, esta tese também é uma tentativa de esclarecer a relação entre algumas medidas de dependência, sobretudo linear, cuja literatura é bastante extensa e com formalismo pouco padronizado, provavelmente pelo fato de seu desenvolvimento envolver áreas do conhecimento distintas como Neurociência, Sociologia, Econometria, Estatística, Física, Matemática e Teoria da Informação. Espera-se que algumas dessas relações entre medidas de dependência tenham se tornado mais explícitas.

SUMÁRIO

LISTA DE FIGURAS

Resumo

Inferência da força de interação nos fenômenos físicos/biológicos é objetivo comum a diversas áreas da ciência. Em particular, nas neurociências tem-se assistido a uma mudança no paradigma experimental em que a atenção tem-se voltado à compreensão da interação entre grupamentos neuronais. Em vista desta demanda surgiram naturalmente diversos métodos estatísticos de medida de dependência entre grupamentos neurais. Alguns foram desenhados para inferência de fluxo de informação, sem contudo precisar o que se entende por fluxo de informação, gerando conseqüentemente controvérsias na literatura.

O principal objetivo deste trabalho é aplicar os conceitos da Teoria da Informação na análise de processos estacionários de segunda ordem para precisar as idéias de fluxo de informação utilizadas na literatura de forma "ad hoc" e obter um melhor entendimento da relação existente entre as diferentes medidas de dependência propostas.

Variáveis aleatórias e processos gaussianos desempenham papel fundamental no desenvolvimento da tese ao permitir estudar quantidades da Teoria da Informação utilizando somente momentos de segunda ordem. Embora, bastante

específico, o modelo gaussiano motiva a introdução de algumas medidas de dependências mais gerais, além de estabelecer limites superiores e inferiores para as medidas de dependência aqui consideradas.

Os desenvolvimentos centrais desta tese são a introdução da definição de variáveis aleatórias inversas associadas a um conjunto de variáveis aleatórias e o estudo de suas propriedades que permitem entender a relação entre a matriz de variância/covariância e sua inversa. Mostra-se que a matriz de variância/covariância das variáveis aleatórias inversas é o inverso da matriz de variância/covariância das variáveis aleatórias associadas. Este fato permite provar a relação entre diferentes medidas de dependência linear propostas na literatura.

Os resultados obtidos para o caso de número finito de variáveis aleatórias são estendidos para séries temporais multivariadas e conduzem a medidas de fluxo de informação. Expressões assintóticas exatas tanto no domínio do tempo como no da freqüência são obtidas para processos estacionários gaussianos.

Por fim, uma aplicação das medidas propostas em dados experimentais é mostrada. Os conjuntos de dados consistem de medidas de potenciais de campo local do hipocampo e córtex pré-frontal registrados durante a execução de tarefa de memória espacial de dois grupos de camundongos: um camundongo controle normal e um hiperdopaminérgico geneticamente modificado.

Summary

The inference of the strength of interaction in physical/biological phenomena is a common objective to many scientific areas. Neuroscience has witnessed a shift of experimental paradigm where the focus is in the understanding of the interaction between groups of neurons. Consequently, new methods were proposed to measure this dependence. Some of them were proposed to infer the information flow alas without defining the precise meaning of these terms, leading to considerable controversy in the literature.

The main aim of this thesis is to use information theoretical ideas for second-order stationary processes to make the idea of information flow precise and thus leading to a better understanding of the relationship between different measures of dependence.

Gaussian random variables and stochastic processes are fundamental to the development of the thesis, allowing the study of information theoretical quantities using only second order moments, though Gaussian models are very special ones, they motivate the definition of gereral measures of dependence and allow bounding the dependence measures studied here.

Inverse random variables associated to a group of random variables and the study of its properties are central do this thesis, for they allow expressing the relationship bewteen the variance/covariance matrix of random variables and its inverse. It is proved that the variance/covariance matrix of the inverse random variables is the inverse of the variance/covariance matrix of the associated random variables.

This last fact is central to explaining the relationship between different measures of linear dependence.

The results obtained for the case of finite number of random variables are extended to multivariate time series and allow defining some measures of information flow. Exact asymptotic expressions, in both time and frequency domains, are obtained for Gaussian stationary processes.

Finally, the proposed measures are illustrated by applying them to data consisting of local field potential from the hippocampus and the pre-frontal cortex during a spatial memory task from two groups of mice: one control and one genetically modified hyperdopaminergic mouse.

CAPÍTULO 1

Introdução

"Clocks tick, bridges and skyscrapers vibrate, neuronal networks oscillate. Are neuronal oscillations an inevitable by-product, similar to bridge vibrations, or an essential part of the brain's design? Mammalian cortical neurons form behavior-dependent oscillating networks of various sizes, which span five orders of magnitude in frequency. These oscillations are phylogenetically preserved, suggesting that they are functionally relevant..." (G. Buzsáki e A. Draguhn, 2004).

A Neurociência tem evoluído a passos rápidos e a década de 1990 ficou conhecida como a Década do Cérebro[1]. Um conceito importante na Neurociência que tem guiado o seu desenvolvimento é o de "áreas neurais funcionais e estruturalmente segregadas[2]". Este se refere a um agrupamento de neurônios espacialmente contíguos juntamente com seu tecido adjacente, cuja atividade apresenta

[1] Com o intuito de chamar a atenção pública e alocar maiores recursos nas áreas envolvendo pesquisa neurocientífica o Congresso Americano denominou a década com início em primeiro de janeiro de 1990 como "Decade of Brain".

[2] A distinção entre os adjetivos "neuronal" e "neural" nem sempre é clara, porém nesta tese o primeiro se refere a neurônios individuais e o último a um grupo de neurônios.

alta correlação com um comportamento animal ou função específica. Diversas técnicas de medidas de atividades neurais têm sido utilizadas para classificar as áreas neurais, desde métodos simples como lesão de uma região específica, observar o seu efeito no animal até métodos sofisticados utilizando imageamento por ressonância magnética e observar a alteração nos sinais de BOLD para tarefas distintas. Há um grande acúmulo de dados relacionados a esses experimentos, e diferentes teorias de funcionamento do sistema nervoso têm sido sugeridas basendo-se neles, porém, parece existir um limite intrínseco nessas abordagens por estudarem as áreas isoladamente no tempo e no espaço, ou seja, em geral tenta-se associar uma função específica para determinadas regiões do sistema nervoso sem se levar em consideração a dinâmica de interação com as outras regiões do sistema nervoso.

A percepção desta limitação naturalmente fez com que na última década houvesse uma mudança de paradigma de investigação, em que o objetivo se tornou caracterizar a relação entre as áreas neurais e reinterpretar as suas funções. A esse estudo da interação dinâmica entre áreas neurais dá-se nome de estudo de *conectividade*.

Há diversos métodos para a inferência de conectividade, incluindo desde aplicação de métodos já estabelecidos na literatura de outras áreas científicas até outros novos motivados nos problemas biológicos. Pode-se dizer que o desenvolvimento de métodos para análise de conectividade se tornou uma importante área de pesquisa em Neurociência. Nota-se, por exemplo, que algumas revistas científicas são especializadas em técnicas de análse como o *Journal of Neuroscience Methods*.

Comum ao desenvolvimento científico em geral, a diversidade de métodos

existentes, se por um lado tem a vantagem de permitir que se utilize o método que melhor se adapta ao problema biológico, é também fonte de controvérsias em que se argumentam os méritos e as desvantagens de determinados métodos baseados em julgamentos filosóficos, biológicos, físicos e matemáticos.

Seria interessante que os métodos pudessem ser classificados de acordo com critérios que envolvessem os diversos aspectos importantes para o uso em neurofisiologia. De fato, na literatura existem alguns esforços neste sentido (Hlaváckova-Schindlera et al., 2007), porém há ainda uma carência de estudos teóricos/ matemáticos que permitam o melhor entendimento das diferenças e semelhanças entre as medidas de conectividade.

Esta tese tem como objetivo principal estudar e elucidar as relações que existem entre algumas medidas de conectividade que têm sido propostas na literatura de Neurociência como sendo relacionadas ao conceito de causalidade de Granger. Neste estudo, a Teoria da Informação desempenha um papel crucial permitindo que se interprete as medidas de conectividade estudadas como sendo de fato medidas de dependência entre determinadas variáveis aleatórias (v.as.) ou séries temporais, o que em muitos casos permite que se entenda o que de fato uma determinada medida de conectividade elucida.

O resultado principal desta tese é a generalização da seguinte proposição[3]:

[3]Optou-se por denominar "Proposição" todos os resultados que foram demonstrados nesta tese, reservando a denominação "Teorema" para resultados conhecidos e provados na literatura.

Proposição 1.0.1. *Sejam X e Y séries univariadas conjuntamente estacionárias e gaussianas. Seja a matriz $f(\lambda)$ de densidade espectral conjunta de X e Y, isto é,*

$$f(\lambda) = \begin{bmatrix} f_{xx}(\lambda) & f_{xy}(\lambda) \\ f_{yx}(\lambda) & f_{yy}(\lambda) \end{bmatrix},$$

em que $\lambda \in [-\pi, \pi)$. Suponha que $c_1 I_n \leq f(\lambda) \leq c_2 I_n$, $c_2 \geq c_1 > 0$, em que, para A, B matrizes $n \times n$, $A - B > 0$ se e somente se $A - B$ for positiva definida.

Seja a representação autorregressiva bivariada

$$\begin{bmatrix} X(t) \\ Y(t) \end{bmatrix} = \sum_{k=1}^{\infty} \begin{bmatrix} A_{xx}(k) & A_{xy}(k) \\ A_{yx}(k) & A_{yy}(k) \end{bmatrix} \begin{bmatrix} X(t-k) \\ Y(t-k) \end{bmatrix} + \begin{bmatrix} \xi_x(t) \\ \xi_y(t) \end{bmatrix}. \tag{1.1}$$

Considere ainda a série dos resíduos de X dado Y, isto é,

$$X(t) = \sum_{k=-\infty}^{\infty} \alpha(k) Y(t-k) + \epsilon_x(t). \tag{1.2}$$

Tome

$$\tilde{A}(\lambda) = I - \sum_{k=1}^{\infty} A(k) e^{-ik\lambda}. \tag{1.3}$$

Tem-se

(1.4)

$$\lim_{j \to \infty} \frac{1}{j+1} E\left(\log \frac{p(\epsilon_x(t), \ldots, \epsilon_x(t-j), \xi_y(t), \ldots, \xi_y(t-j))}{p(\epsilon_x(t), \ldots, \epsilon_x(t-j)) p(\xi_y(t), \ldots, \xi_y(t-j))} \right)$$
$$= -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log \left(1 - \frac{|\tilde{A}_{xy}(\lambda)|^2 \, Var(\xi_x(t))^{-1}}{[\tilde{A}_{xy}(\lambda)^* \; \tilde{A}_{yy}(\lambda)^*] \, Var(\xi_x(t), \xi_y(t))^{-1} [\tilde{A}_{xy}(\lambda) \; \tilde{A}_{yy}(\lambda)]^T} \right) d\lambda,$$

em que a esperança em (1.4) é em relação a todas as v.as. consideradas.

A proposição acima necessita de alguns esclarecimentos. A quantidade

$$\lim_{j\to\infty} \frac{1}{j+1} E\left(\log \frac{p(\epsilon_x(t),\ldots,\epsilon_x(t-j),\xi_y(t),\ldots,\xi_y(t-j))}{p(\epsilon_x(t),\ldots,\epsilon_x(t-j))p(\xi_y(t),\ldots,\xi_y(t-j))}\right) \tag{1.5}$$

é conhecida como taxa de informação mútua entre as séries ϵ_x e ξ_y. Intuitivamente, esta quantidade mede o grau de independência entre as séries. Note que se as séries são independentes, isto é,

$$\begin{aligned} &p(\epsilon_x(t),\ldots,\epsilon_x(t-j),\xi_y(t),\ldots,\xi_y(t-j)) \\ &= p(\epsilon_x(t),\ldots,\epsilon_x(t-j))p(\xi_y(t),\ldots,\xi_y(t-j)) \end{aligned}$$

portanto (1.5) é igual a zero.

De fato, uma possível interpretação para (1.5) é que ela mede o fluxo de informação de Y para X. Esta interpretação se torna aparente uma vez que o lado direito de (1.4) implica que (1.5) é zero se e somente se $\tilde{A}_{xy}(\lambda) = 0$ e que por sua vez implica que $A_{xy}(k) = 0, k \geq 1$. Olhando para a representação autorregressiva (1.1), $A_{xy}(k) = 0,\ k \geq 1$ implica que o passado de Y não influencia $X(t)$ dado que o passado de X é considerado. Em outras palavras se $A_{xy}(k)$ for diferente de zero para algum $k \geq 1$, pode-se concluir que de alguma forma o passado de Y "envia informação" para $X(t)$.

Esta última noção de fluxo de informação é a definição de causalidade de Granger comumente empregada na literatura de Econometria (Lütkepohl, 1993) e, recentemente, também em Neurociência (Sameshima e Baccalá, 1999). É importante salientar que a própria definição de causalidade de Granger é ambígua em muitos casos, e diferentes medidas de causalidade de Granger são equivalentes quando os coeficientes $A_{xy}(k), k \geq 1$ são nulos, porém assumem valores

distintos quando existe causalidade de Granger, o que exige certo cuidado em definir uma medida de caulidade de Granger. É uma questão que se se discute na tese.

Por fim, seguem algumas observações sobre a organização do texto.

Para provar a Proposição 1.0.1 e tornar o método de obtenção de medidas de dependência mais sistemático, foi necessária a introdução de conceitos de Teoria da Informação e medidas de dependência linear assim como a obtenção de alguns resultados matemáticos novos referentes à algumas medidas de dependência. Em alguns casos, a aplicação dos resultados é feita somente no último capítulo. Assim, como tentativa de melhorar a legibilidade, alguns comentários informais sobre os resultados obtidos são feitos no decorrer do texto.

No primeiro capítulo são listadas algumas notações e convenções utilizadas ao longo do texto. Em alguns casos as definições e notações são repetidas quando parecer adequado.

No Capítulo 3 é introduzido o conceito de informação mútua como sendo uma definição geral de medida de dependência entre v.as. Algumas propriedades fundamentais relacionadas à informação mútua são obtidas para se provar resultados em capítulos seguintes. Os principais resultados nesta seção são as expressões de informação mútua para v.as. gaussianas e as identidades e desigualdades envolvendo informação mútua e entropia. Embora, para a obtenção dos resultados tenha-se sempre em mente as v.as. gaussianas e séries temporais estacionárias gaussianas, muitos deles não se restrigem a estas v.as. e os processos. Em particular, existe uma relação entre as chamadas funções de cópulas e a informação mútua, o que permite em muitos casos estender diretamente os resultados de Teoria da Informação obtidos para o caso gaussiano, bastando

para isso simplesmente considerar as v.as. com cópula gaussiana. O conceito de cópula é brevemente introduzida. Um resultado possivelmente inédito que se obtém nessa seção é a parametrização da informação mútua em termos da cópula que caracteriza a distribuição conjunta das v.as. consideradas.

No Capítulo 4 é estudada uma familía de medidas de dependência conhecidas como medidas de dependência linear em que a correlação linear de Pearson é o exemplo mais conhecido. Embora as medidas de dependência linear constituam uma família bastante específica de medidas que em muitos casos não caracteriza totalmente a estrutura de dependência, elas constituem ótimos modelos para o estudo de medidas de dependência em geral. Além do fato que no caso em que as v.as. apresentam distribuição gaussiana conjunta, as medidas de dependência linear caracterizam totalmente a estrutra de dependência entre as v.as. Os principais temas do capítulo são a definição da correlação quadrática total entre duas ou mais v.as. não necessariamente univariadas, a definição da parcialização e inversão de medidas de dependência linear e a relação com a informação mútua no caso gaussiano. Os resultados obtidos nesse capítulo são essenciais para se provar os resutados do Capítulo 7 sobre medidas de dependência entre séries temporais. Em particular, a inversão de medidas de dependência linear tem papel fundamental para a compreensão das medidas de dependência linear em geral e é uma contribuição original desta tese.

No Capítulo 5 são revisados alguns fatos sobre séries temporais estacionárias de segunda ordem das quais as séries estacionárias gaussianas são exemplos importantes. É definida a condição de limitação para séries estacionárias de segunda ordem que garante a validade dos cálculos realizados nesta tese. Um fato importante para estes tipos de séries temporais é a existência da representação

espectral que permite a introdução do conceito de componentes no domínio da freqüência para estes processos.

O Capítulo 6 serve como motivação para se definir algumas medidas de dependência direcionada que são denominadas medidas de causalidade ou fluxo de informação. É importante salientar que o termo causalidade utilizada nesta tese se refere a uma noção particular de relação de preditibilidade entre séries temporais e não ao conceito filosófico de causalidade.

No capítulo 7 são apresentados os principais resultados desta tese. Na Seção 7.1 são provados alguns teoremas assintóticos que são utilizados para se provar os resultados das últimas duas seções. Na seção seguinte, intitulada "medidas de dependência simétrica", algumas medidas de dependência entre séries temporais são definida e algumas propriedades obtidas. As medidas de dependência consideradas nessa seção são simétricas em relação às séries envolvidas e não fornecem a noção de fluxo de informação ou "causalidade". A seção seguinte contém a prova da Proposição 1.0.1.

No Capítulo 8 são apresentados alguns exemplos de aplicação biológica de algumas medidas de dependência entre séries temporais estudadas nesta tese.

No último capítulo são feitas as conclusões gerais e alguns comentários sobre possíveis trabalhos futuros.

CAPÍTULO 2

Notação

Nesta seção, A é uma matriz quadrada $n \times n$ com valores complexos. A matriz B é uma matriz $n \times m$ com valores complexos com elementos $B_{kl},\ k = 1, \ldots, n, l = 1, \ldots, m$, e cujos vetores colunas são denotados por $B_k,\ 1 \leq k \leq m$, isto é, $B = [B_1\ \ \ldots\ \ B_m]$. As matrizes $C_1, \ldots, C_m$ apresentam dimensões finitas e não são necessariamente quadradas. X e Y são variáveis aleatórias (v.as.) n e m-dimensionais complexas. As v.as. complexas $W_1, \ldots, W_k$ e $Z_1, \ldots, Z_l$ são $d_1, \ldots, d_k$ e $c_1, \ldots, c_m$-dimensionais.

- A^T - matriz transposta de A.
- $\overline{A}$ - matriz conjugada complexa de A.
- $A^* = (\overline{A})^T$ - matriz conjugada complexa transposta (hermitiana) de A.
- I_n - matriz identidade de dimensão n. Será denotado simplesmente I caso a dimensão esteja clara pelo contexto.

- $0_{n\times m}$ - matriz nula de dimensão $n \times m$. Será denotado simplesmente 0 se não houver ambigüidade.

- $\mathrm{diag}(C_1, \ldots, C_m)$ - matriz bloco diagonal formada pelas matrizes $C_1, \ldots, C_m$ postas na "diagonal blocada", isto é,

$$\mathrm{diag}(C_1, \ldots, C_m) = \begin{bmatrix} C_1 & 0 & 0 & \ldots & 0 \\ 0 & C_2 & 0 & \ldots & 0 \\ \vdots & \ldots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & C_n \end{bmatrix}.$$

- A_{pq} - elemento da p-ésima linha e q-ésima coluna da matriz A na base canônica.

- $|A|$ - matriz valor absoluto de A termo a termo.

- $\det A$ - determinante de A.

- $\mathrm{tr}A$ - traço de A.

- $\mathrm{vec}B = [B_1^T \ldots B_m^T]^T$ - operador de enfileiramento (*column stacking*).

- $\otimes$ - produto de Kronecker ou produto direto.

- $E_P(X)$ - esperança matemática de X em relação à medida P. Quando a medida estiver clara, esta é omitida.

- $\mathrm{Var}(X)$ - variância de X, isto é, $\mathrm{Var}(X) = E(XX^*) - E(X)E(X)^*$. É uma matriz $n \times n$.

- $\mathrm{Var}(W_1, \ldots, W_k) = \mathrm{Var}((W_1^T, \ldots, W_k^T)^T)$.

- $\mathrm{Cov}(X : Y)$ - covariância entre X e Y, ou seja, $\mathrm{Cov}(X : Y) = E(XY^*) - E(X)E(Y)^*$.

- $\mathrm{Cov}(W_1, \ldots, W_k : Z_1, \ldots, Z_l)$ - covariância entre $W^T = [W_1^T \ \ldots \ W_k^T]$ e $Z^T = [Z_1^T \ \ldots \ Z_l^T]$, ou seja,
  $\mathrm{Cov}(W_1, \ldots, W_k : Z_1, \ldots, Z_l) = \mathrm{Cov}(W : Z)$.

- $E(X/Y)$ - esperança condicional de X dado Y. É uma variável aleatória n-dimensional definida por $E(X/Y) = (E(X_1/Y_1, \ldots, Y_m), \ldots, E(X_n/Y_1, \ldots, Y_m))^T$.

- $R(X/Y) = X - E(X/Y)$ - resíduo da esperança condicional de X dado Y. É uma variável aleatória n-dimensional.

- $\bar{E}(X/Y)$ - projeção ortogonal linear de X, termo a termo, no subespaço de L^2 gerado por Y. É uma variável aleatória n-dimensional.

- $\bar{R}(X/Y) = X - \bar{E}(X/Y)$ - resíduo da projeção ortogonal linear de X no subespaço gerado por Y. É uma variável aleatória n-dimensional.

- $\mathrm{Var}(X/Y) = \mathrm{Var}(\bar{R}(X/Y))$ - variância parcial de X dado Y. É uma matriz constante $n \times n$. Não é uma variância condicional.

- $\mathrm{Cov}(X : Y/Z) = \mathrm{Cov}(\bar{R}(X/Z) : \bar{R}(Y/Z))$ - covariância parcial de X e Y dado Z. É uma matriz constante $n \times m$.

- $\{X_j\}_0^k$ - seqüência de $k+1$ v.as. n-dimensionais.

- $\{X_j\}_0^\infty$ - seqüência unilateral infinita de v.as. n-dimensionais.

2.1 Algumas convenções

As convenções são sempre explicitadas em cada capítulo quando necessárias, porém para facilitar a leitura algumas delas são fixadas nesta seção, com o risco de repetir em outras seções.

Seja $(\Omega, \mathcal{F}, P)$ um espaço de probabilidade. Uma variável aleatória (v.a.) é uma função mensurável de Ω a valores em $\mathbb{R}^n$ ou $\mathbb{C}^n$. Quando $n > 1$ dizemos que a v.a. é multidimensional ou multivariada real (complexa), caso contrário dizemos que é uma v.a. unidimensional ou univariada real (complexa). As v.a. consideradas nesse texto apresentam média (esperança) zero e variância finita a menos que seja explicitado. As matrizes de covariância das v.as. consideradas aqui sempre apresentam posto máximo e portanto são positivas definidas.

Um processo estocástico n-dimensional X é definido como uma família de v.a. $X = \{X(t)\ v.a.\ n - dimensional : t \in J\}$, em que J é o conjunto dos índices. Nesse texto, são considerados os processos estocásticos em tempo discreto, denominados séries temporais, em que $J = \mathbb{Z}$. No caso de a série temporal ser multivariada (n-dimensional com $n > 1$) os k-ésimos componentes univariado da série no tempo t são denotados por $X_k(t), k = 1, \ldots, n$. Em algumas partes do texto o índice subscrito é usado para indicar a k-ésima série não necessariamente univariada e, nesse caso, o significado do índice é explicitado no próprio texto.

Ao se considerar n v.as., n é sempre finito, a menos que seja especificado como infinito.

Utilizou-se alguns termos da Análise Funcional, sobretudo quando os argumentos envolvem número não finito de elementos, embora não seja a linguagem de escolha para o texto em geral. Dado uma família de v.a. $X =$

$\{X(t)\ v.a.\ n-dimensional : t \in J \subset \mathbb{Z}\}$, o espaço gerado por X é o espaço de Hilbert $\mathcal{H} \subset L^2(\Omega, \mathcal{F}, P)$ fechado gerado pelas v.a. $X_k(t),\ t \in J,\ 1 \leq k \leq n$, ou seja, é o subespaço gerado pelos componentes univariados dos elementos da série temporal. O produto escalar de duas v.a. univariadas $X, Y \in \mathcal{H}$ é definido por $\langle X, Y \rangle = E(XY)$. Como as v.as. consideradas nesta tese apresentam média nula $\langle X, Y \rangle = \mathrm{Cov}(X : Y)$.

Duas v.a. unidimensionais X e Y são ortogonais ou não-correlacionadas quando $\langle X, Y \rangle = \mathrm{Cov}(X : Y) = 0$. Se X e Y forem v.as. n e m-dimensionais, diz-se que são ortogonais se todas as combinações lineares de elementos de X e Y da forma $\sum_{k=1}^{n} a_k X_k$ e $\sum_{k=1}^{m} b_k Y_k$, respectivamente, forem ortogonais, ou seja, se $\mathrm{Cov}(X : Y) = 0$.

A convergência de seqüências de v.as. é entendida no sentido de média quadrática, ou seja em $L^2(\Omega, \mathcal{F}, P)$.

Duas séries n-variadas X e Y são iguais se $X_k(t) = Y_k(t)$ em média quadrática para todo $t \in \mathbb{Z}$ e $1 \leq k \leq n$.

O termo regressão estará se referindo à regressão linear com minimização do erro quadrático médio (mínimos quadrados), ou seja, dadas $n+1$ v.a. $Y, X_1, \ldots, X_n$, respectivamente com dimensões $d, d_1, \ldots, d_n$, a regressão ou mais especificamente os coeficientes de regressão de Y em $X_1, \ldots, X_n$ são definidas como sendo as matrizes de coeficientes $A_1, \ldots, A_n$ com dimensões $d \times d_1, \ldots, d \times d_n$, respectivamente, tais que minimizem

$$\mathrm{Tr}\left\{\mathrm{Var}\left(Y - \sum_{k=1}^{n} A_k' X_k\right)\right\}, \tag{2.1}$$

em que $\mathrm{Tr}B, B \in \mathbb{R}^{m \times m}, m \geq 1$, é o traço da matriz B. Eventualmente n pode

ser infinito quando o erro (2.1) estiver bem definido, que é sempre o caso neste texto.

Os resultados já conhecidos e cujas provas estão disponíveis na literatura são apresentados como teoremas e suas demostrações são sempre referenciadas. As proposições nesta tese sempre se referem a resultados (a) novos, (b) que não foram encontrados na literatura sobre o qual se baseou o trabalho ou (c) que embora conhecidos a prova não está disponível de forma simples na literatura. Para as proposições, as demonstrações são feitas na tese.

CAPÍTULO 3

Medidas de dependência - aspectos gerais

"Let ξ and η be random variables on a probability space $(\Omega, \mathcal{A}, P)$, neither of them being constant with probability 1. In almost every field of application of statistics one encounters often the problem that one has to characterize by a numerical value the strength of dependence between ξ and η. (...) With these conventions the following set of postulates for an appropriate measure of dependence, which shall be denoted by $\delta(\xi, \eta)$, seems natural ... "(A.Rényi, 1959)

Comum a praticamente todas as disciplinas que utilizam a Teoria da Probabilidade, a noção de dependência se refere ao vínculo probabilístico entre v.as. ou eventos. Apesar desse papel central, é seguro dizer que inexiste uma definição única que permita aferí-la quantitativamente. Assim, propostas nesse sentido geralmente variam de acordo com especificidades da aplicação em estudo.

Seguramente, a medida de dependência mais amplamente conhecida e usada (por vezes até inapropriadamente), é o *coeficiente de correlação linear* ou

simplemente a *correlação* entre duas v.as. Seu emprego se faz freqüentemente mesmo a despeito de somente indicar independência de modo inequívoco em casos específicos, como quando envolve v.as. conjuntamente gaussianas.

Rényi (1959) propôs sete postulados para explicitar as propriedades de quantidades destinadas a medir dependência que, ainda retendo as propriedades intuitivas da correlação, fossem válidas de forma mais geral. Com base nesta idéia, devidamente generalizada e modificada, Bell (1962) observou que uma quantidade que satisfaz todos os postulados é a *informação mútua*, originariamente introduzida em Teoria da Informação (Shannon e Weaver, 1949; Cover e Thomas, 1991).

Uma segunda abordagem para descrever dependências entre v.as., que é hoje bastante popular na literatura, baseia-se nas funções de cópula, que são distribuições multivariadas cujas marginais univariadas são distribuições uniformes no intervalo $[0, 1]$ (Nelsen, 1999). Pelo celebrado teorema de Sklar (1959), as cópulas permitem representar a distribuição conjunta de v.as. como funções de suas marginais univariadas. Isto permite estudar a dependência entre as v.as. separadamente das propriedades das minúcias relativas às suas marginais univariadas.

Os principais objetivos neste capítulo são (a) introduzir o conceito de entropia e informação mútua e obter algums fórmulas para o caso gaussiano, (b) obter algumas igualdades e desigualdades envolvendo quantidades da Teoria da Informação para serem usadas em capítulos posteriores, (c) definir a função de cópula e (d) relacioná-la com a informação mútua.

Como roteiro do restante deste capítulo, inicia-se pela Seção 3.1 em que se examina o conceito de medida de dependência à luz das idéias de Rènyi e

Bell. Por questão de clareza e ordem histórica, inicialmente são definidas as quantidades da Teoria da Informação para o caso em que as v.as. assumem valores discretos[1], embora este não seja mais utilizado em capítulos subseqüentes. Logo em seguida são definidas as mesmas quantidades para o caso de v.as. que apresentam densidades de probabilidades. Pela sua particular simplicidade e importância, quando envolvem v.as. gaussianas, tanto informação mútua bem como suas generalizações são apresentadas explicitamente.

A seguir, na Seção 3.1.2, examina-se a relação entre a informação mútua e as funções de cópula cujo resultado serve para justificar como a correlação e suas generalizações podem ainda ser úteis para descrever dependência entre variáveis aleatórias gerais.

Na última seção são discutidos os resultados obtidos e como eles se relacionam com as medidas de dependência linear.

Neste capítulo todas as v.as. assumem valores no conjunto dos reais ou num subconjunto deste. O caso em que as v.as. assumem valores complexos pode ser tratado como caso especial, bastando para isto separar as v.as. em partes real e imaginária e então utilizando a teoria desenvolvida para o caso real.

3.1 Dependência

O conceito de dependência entre variáveis aleatórias tem papel crucial no desenvolvimento da teoria dos processos estocásticos assim como na aplicação dos métodos estatísticos. A sua definição exata varia de acordo com a situação, porém para o texto que segue a definição devido a Rényi (1959) parece ade-

[1]Sem perda de generalidade pode-se supor que assumem valores num subconjunto dos números naturais

quada.

Dadas v.as. X e Y definidas num mesmo espaço de probabilidade, Rényi (1959) propôs um conjunto de sete postulados que devem ser satisfeitos por uma medida de dependência $\delta(X, Y)$. Bell (1962) sugere algumas modificações e propõe os seguintes postulados:

1. $\delta(X, Y)$ é definida para quaisquer X e Y definidos no mesmo espaço de probabilidade, tais que nenhum deles seja uma constante com probabilidade 1.

2. $\delta(X, Y) = \delta(Y, X)$.

3. $0 \leq \delta(X, Y) \leq \infty$.

4. $\delta(X, Y) = 0$ se e somente se X e Y forem independentes.

5. $\delta(X, Y)$ assume seu valor máximo, quando finito, se e somente se $X = f(Y)$ e $Y = g(X)$ em que g e f são funções mensuráveis[2].

6. $\delta(X, Y) = \delta(f(X), g(Y))$ se f e g são funções bijetoras da reta real.

7. Se X e Y apresentam distribuição conjunta normal multivariada, $\delta(X, Y)$ é igual ao módulo da correlação linear entre X e Y a menos de uma transformação monotônica estritamente crescente na reta real.

Bell (1962) provou que a informação mútua satisfaz essas condições tornando-a um candidato natural como medida de dependência padrão. Na Teoria da Informação originada nos trabalhos de Shannon e Weaver (1949), a informação mútua apresenta interpretação natural como medida de informação comum entre

[2]Rényi (1959) exigia que o valor máximo fosse um, porém essa exigência não é essencial.

v.as. (Kolmogorov (1957); Dobrushin (1959)) e está intimamente relacionadoa ao conceito de capacidade de canal (Cover e Thomas, 1991).

3.1.1 Informação mútua

A seguinte frase devido a Kolmogorov (1957), embora escrita há mais de meio século, ilustra bem como os conceitos desenvolvidos na Teoria da Informação têm influenciado as ciências experimentais.

> *"Let me note that in my view the applications of the concept of information theory to natural memory devices, to the study of the nervous system and hereditary phenomena, are also very well founded and hold out prospects of being essential in the development of these branches of science." (A. N. Kolmogorov, 1957)*

No caso mais simples em que as v.as. X e Y assumem valores num conjunto $\mathcal{A} \times \mathcal{B}$ tem-se a seguinte definição:

Definição 3.1.1 (Informação mútua)**.** A *informação mútua* entre X e Y IM(X:Y) é definida como

$$\mathrm{IM}(X:Y) = \sum_{k,l} P(X = x_k, Y = y_l) \log \frac{P(X = x_k, Y = y_l)}{P(X = x_k)P(Y = y_l)},$$

em que $(x_k, y_l) \in \mathcal{A} \times \mathcal{B}$. Assume-se $0 \log f/0 = \infty$ para $f > 0$ e $0 \log 0/f = 0$ para $f \geq 0$.

Pode-se mostrar que a informação mútua $\mathrm{IM}(X:Y)$ assume apenas valores não negativos, anulando-se se e somente se X e Y forem independentes (Lloyd, 1962), o que justifica parcialmente o seu uso como medida de dependência entre v.as. Ela assume o valor máximo se e somente se $X = f(Y)$ e $Y = g(X)$ em

que f e g são funções bijetoras mensuráveis (Lloyd, 1962). Neste caso

$$\mathrm{IM}(X : Y) = H(X),$$

em que $H(X)$ é a entropia de X definida a seguir.

Definição 3.1.2 (Entropia)**.** Seja $X = (X_1, \ldots X_n)$ uma v.a. a valores num conjunto enumerável $\mathcal{A}_1 \times \ldots \times \mathcal{A}_n$. A *entropia* $\mathrm{H}(X)$ de X é definida por

$$\mathrm{H}(X_1, \ldots, X_n) = \mathrm{H}(X) = -E_{P(X)}(\log P(X)).$$

A entropia acima definida para v.as. discretas assume somente valores não negativos, o que difere do caso em que as v.a. assumem valores em conjuntos não enumeráveis como na reta real.

Antes de se estudar o caso de v.as. mais geral, considere a definição de informação mútua entre mais de duas v.as. assumindo valores em conjuntos enumeráveis.

Definição 3.1.3 (Informação mútua para mais de duas v.as.)**.** A *informação mútua* $\mathrm{IM}(X_1 : \ldots : X_n)$ entre $X_1, \ldots, X_n$ assumindo valores nos conjuntos enumeráveis $\mathcal{A}_1, \ldots, \mathcal{A}_n$, respectivamente, é definida como

$$\mathrm{IM}(X_1 : \ldots : X_n) = E_{P(X_1,\ldots,X_n)} \left(\log P(X_1, \ldots, X_n) - \log \left\{ \prod_{k=1}^{n} P(X_k) \right\} \right).$$

Pode-se escrever a informação mútua acima em termos de entropias, mais explicitamente,

$$\mathrm{IM}(X_1 : \ldots : X_n) = \sum_{k=1}^{n} H(X_k) - H(X_1, \ldots, X_n),$$

o que permite interpretar a informação mútua como a medida da parte da entropia comum entre as v.as. $X_1, \ldots, X_n$.

A informação mútua entre $X_1, \ldots, X_n$ assume somente valores não negativos[3] e é zero se e somente se

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{k=1}^{n} P(X_k = x_k), \tag{3.1}$$

$$(x_1, \ldots, x_n) \in \mathcal{A}_1 \times \ldots \times \mathcal{A}_1. \tag{3.2}$$

Veja (Lloyd, 1962) para a demonstração.

Observação 3.1.1. A definição 3.1.3 não é a única possível para a informação mútua entre mais de duas variáveis aleatórias. Em alguns casos a versão definida em 3.1.3 é denominada correlação total (Watanabe, 1960). De fato, talvez um nome mais adequado para a informação mútua definida em 3.1.3 seja informação mútua total, pois mede a soma das relações que existem entre as v.as. duas a duas, três a três e assim por diante. Uma definição alternativa para informação mútua para mais de duas v.as. que mede somente o componente comum a todas as v.as. é a seguinte.

Definição 3.1.4 (Informação mútua multipla). Seja J_k o conjunto das partições de $\{1, \ldots, n\}$ com k elementos distintos. A *informação mútua múltipla* (Han, 1980) $\mathrm{I}(X_1 : \ldots : X_n)$ entre $X_1, \ldots, X_n$ assumido valores nos conjuntos enumeráveis $\mathcal{A}_1, \ldots, \mathcal{A}_n$, respectivamente, é definida como

$$\mathrm{I}(X_1 : \ldots : X_n) = \sum_{k=1}^{n} (-1)^{k-1} \sum_{(j_1, \ldots, j_k) \in J_k} H(X_{j_1}, \ldots, X_{j_k}).$$

A definição 3.1.4 é interessante por isolar os componentes das dependências (veja Han (1980) para uma discussão). Difere da informação mútua, definida em 3.1.3, por assumir valores negativos. Além disso, a

[3] A informação mútua para v.as. assumindo valores em conjuntos não enumeráveis também assume valores somente não negativos, o que difere da entropia.

> condição de independência (3.1) é somente suficiente, mas não é necessária para que a informação mútua múltipla seja nula. A condição necessária e suficiente para a nulidade da informação mútua múltipla é denominada condição de semi-independência (Han, 1980), e não é discutida aqui.

Para o caso de v.a. assumindo valores em conjuntos não enumeráveis como o $\mathbb{R}$, a definição geral da informação mútua é mais delicada e pode ser encontrada com detalhes em Masani (1992a,b); Dobrushin (1959); Lloyd (1962). Aqui a definição mais geral é desnecessária e os teoremas a seguir possibilitam calcular explicitamente os valores da informação mútua para os casos de interesse.

Teorema 3.1.1. *Sejam $X_1, \ldots, X_n$ v.as. a valores em $\mathbb{R}^{d_1}, \ldots, \mathbb{R}^{d_n}$ com densidade de probabilidade definidas. Sejam p a densidade de probabilidade conjunta de $X_1, \ldots, X_n$ e $p_1, \ldots, p_n$ as suas densidades de probabilidade marginais, respectivamente. A informação mútua $IM(X_1, \ldots, X_n)$ entre as v.as. $X_1, \ldots, X_n$ pode ser escrita como*

$$IM(X_1 : \ldots : X_n) = \int \cdots \int p(x_1, \ldots, x_n) \log \frac{p(x_1, \ldots, x_n)}{\prod_{k=1}^{n} p_k(x_k)} dx_1 \ldots dx_n, \tag{3.3}$$

se a integral for finita.

Demonstração. Veja Dobrushin (1959) equação (1.2.3). □

O Teorema 3.1.1 possibilita o cálculo da informação mútua em alguns casos importantes, por exemplo, quando as v.a. apresentam distribuição gaussiana de dimensão finita. Na literatura é comum se adotar a fórmula (3.3) como definição de informação mútua (veja por exemplo Cover e Thomas (1991)).

A definição da informação mútua no caso contínuo preserva as propriedades da informação mútua para o caso discreto, isto é, assume apenas valores não

negativos e é zero se e somente se as v.as. são independentes.

Uma propriedade importante da informação mútua é sua invariância em relação às transformações bijetoras, isto é,

Teorema 3.1.2. *Sejam $X_1, \ldots, X_n$ v.as. $d_1, \ldots, d_n$ dimensionais, respectivamente, definidas num mesmo espaço de probabilidade. Tome as funções $f_k : \mathbb{R}^{dk} \to \mathbb{R}^{dk}$ para $k = 1, \ldots, n$, bijetoras mensuráveis com as inversas f_k^{-1} também mensuráveis, então*

$$IM(X_1 : \ldots : X_n) = IM(f_1(X_1) : \ldots : f_n(X_n)). \tag{3.4}$$

Demonstração. Veja Ihara (1964). □

Na prática, o Teorema anterior indica que a informação mútua é invariante quanto à parametrização e portanto, do ponto de vista físico, a forma em que os fenômenos associados às v.as. X e Y são mensuradas não influencia no valor da informação mútua, se for garantido que não ocorra perda de "informação".

Teorema 3.1.3. *Sejam $\{X_k^1\}_0^\infty, \ldots, \{X_k^n\}_0^\infty$ seqüências de v.as. $d_1, \ldots, d_n$ dimensionais. Tem-se*

$$IM(\{X_k^1\}_0^{j_1} : \ldots : \{X_k^n\}_0^{j_n}) \leq IM(\{X_k^1\}_0^{l_1} : \ldots : \{X_k^n\}_0^{l_n}),$$

$$j_k \leq l_k, 1 \leq k \leq n,$$

$$\lim_{j_1, \ldots, j_n \to \infty} IM(\{X_k^1\}_0^{j_1} : \ldots : \{X_k^n\}_0^{j_n}) = IM(\{X_k^1\}_0^{\infty} : \ldots : \{X_k^n\}_0^{\infty}).$$

Demonstração. Veja Lloyd (1962) Teorema 13. □

O Teorema 3.1.3 permite o cálculo da informação mútua entre seqüências de v.as. como um limite de séries de informações mútuas. Em muitos casos o limite não é finito e é útil se definir a taxa de informação mútua.

Definição 3.1.5 (Taxa de informação mútua)**.** Sejam $\{X_k^1\}_0^\infty, \ldots, \{X_k^n\}_0^\infty$ seqüências de v.as. $d_1, \ldots, d_n$ dimensionais. A *taxa de informação mútua* $\mathrm{TIM}(\{X_k^1\}_0^\infty, \ldots, \{X_k^n\}_0^\infty)$ entre seqüências de v.as. é definida como

$$\mathrm{TIM}(\{X_k^1\}_0^\infty : \ldots : \{X_k^n\}_0^\infty) = \lim_{j\to\infty} \frac{1}{j+1}\mathrm{IM}(\{X_k^1\}_0^j : \ldots : \{X_k^n\}_0^j).$$

Nesta tese, um dos objetivos é calcular aproximações para as taxas de informação mútua para as diversas séries de interesse. Os cálculos são feitos no Capítulo 6.

Agora, pode-se calcular a informação mútua para o caso de v.as. com distribuição conjunta gaussiana utilizando o Teorema 3.1.1.

Proposição 3.1.1. *Sejam $X_1, \ldots, X_n$ v.as. conjuntamente gaussianas $d_1, \ldots, d_n$-dimensionais. Assumindo que a matriz de variância/covariância $Var(X_1, \ldots, X_n)$ não seja singular tem-se*

$$IM(X_1 : \ldots : X_n) = -\frac{1}{2}\log\left(\frac{\det Var(X_1, \ldots, X_n)}{\prod_{k=1}^n \det Var(X_k)}\right)$$

Demonstração. Tem-se

$$\begin{aligned}
&\int \cdots \int p(x_1, \ldots, x_n) \log p(x_1, \ldots, x_n) dx_1 \ldots dx_n \\
&= -\frac{1}{2}\log\{(2\pi)^n \det \mathrm{Var}(X_1, \ldots, X_n)\} - \mathrm{Tr}(\mathrm{Var}(X_1, \ldots, X_n)^{-1}\mathrm{Var}(X_1, \ldots, X_n)) \\
&= -\frac{1}{2}\log\{(2\pi)^n \det \mathrm{Var}(X_1, \ldots, X_n)\} - n.
\end{aligned}$$

Usando Teorema 3.1.1 obtém-se o resultado. □

Observação 3.1.2. Embora nesta tese não seja considerado o caso em que a matriz de variância/covariância das v.as. envolvidas seja singular, é possível calcular a informação mútua mesmo nestes casos. Para isto, basta observar

que sempre existe uma matriz $M_{r\times n}$ de dimensão $r \times s$ tal que transforma uma v.a. normal s-dimensional Y com matriz de variância/covariância eventualmente singular numa v.a. normal padrão não singular, isto é,

$$\mathrm{Var}(M_{r\times s}Y) = I_r,$$

em que $r = \mathrm{posto}(\mathrm{Var}(Y))$.

Proposição 3.1.2. *Sejam $X_1, \ldots, X_n$ v.as. conjuntamente gaussianas $d_1, \ldots, d_n$-dimensionais com matriz de variância/covariância Var$(X_1, \ldots, X_n)$ eventualmente singular. Dado $d = \sum_{k=1}^{n} d_k$, tem-se*

$$\begin{aligned} &IM(X_1 : \ldots : X_n) \\ &= -\frac{1}{2}(r - \sum_{k=1}^{n} r_k)(\log(2\pi) + 1) + \frac{1}{2}\log\left(\frac{\det M_{r\times d}M_{r\times d}^T}{\prod_{k=1}^{n}\det M_{r\times d_k}M_{r\times d_k}^T}\right) \end{aligned}$$

Demonstração. Basta padronizar as v.as. $X_1, \ldots X_n$ e $W^T = [X_1^T \ \ldots \ X_n^T]$ e calcular como na demonstração da Proposição 3.1.1 para as v.as. gaussianas padronizadas. □

A Proposição 3.1.3 apresentada a seguir é importante pois permite que se obtenha uma estimativa do erro que se comete ao se considerar apenas as variâncias e covariâncias das v.as. para se calcular a dependência entre as variáveis aleatórias. Na prática, obter a informação completa sobre a distribuição de probabilidade a partir dos dados é uma tarefa difícil e é importante se obter estas estimativas. Antes de enunciar a Proposição 3.1.3, a definição e o teorema a seguir são úteis.

Definição 3.1.6 (Entropia para v.as. contínuas)**.** Sejam $X_1, \ldots, X_n$ v.as. $d_1, \ldots, d_n$-dimensionais com densidades de probabilidade. A entropia $H(X_1, \ldots, X_n)$ das v.as. $X_1, \ldots, X_n$ é definida por

$$\mathrm{H}(X_1, \ldots, X_n) = -\int \cdots \int p(x_1, \ldots, x_n) \log p(x_1, \ldots, x_n) dx_1 \ldots dx_n.$$

A entropia definida acima é muitas vezes denominada entropia diferencial. Embora apresente propriedades semelhantes à entropia definida para o caso discreto (definição 3.1.2) não é o análogo perfeito. A diferença mais notável é a possibilidade de assumir valores negativos. No caso em que as v.as. são contínuas, a distribuição gaussiana apresenta um papel importante como é mostrada pelo seguinte teorema.

Teorema 3.1.4 (Máximo da entropia)**.** *Sejam $Y_1, \ldots, Y_n$ v.as. conjuntamente gaussianas $d_1, \ldots, d_n$ dimensionais e $X_1, \ldots, X_n$ v.as. $d_1, \ldots, d_n$ dimensionais não necessariamente gaussianas. Tome $d = \sum d_k$. Assume-se que a matriz de variância/covariância são iguais, isto é, $Var(Y_1, \ldots, Y_n) = Var(X_1, \ldots, X_n)$. Tem-se*

$$\begin{aligned} & H(X_1, \ldots, X_n) \\ & \leq H(Y_1, \ldots, Y_n) \\ & = \frac{1}{2} \log \left\{ (2\pi e)^d \det Var(Y_1, \ldots, Y_n) \right\} \qquad (3.5) \\ & = \frac{1}{2} \log \left\{ (2\pi e)^d \det Var(X_1, \ldots, X_n) \right\}, \qquad (3.6) \end{aligned}$$

em que $e = \exp(1)$.

Demonstração. Veja (Cover e Thomas, 1991, p. 234, Teorema 9.6.5). □

Pode-se, agora, enunciar e provar a seguinte proposição:

Proposição 3.1.3 (Limitantes para informação mútua)**.** *Sejam $X_1, \ldots, X_n$ v.as. $d_1, \ldots, d_n$ dimensionais não necessariamente gaussianas. Tome $d = \sum d_k$. A seguinte estimativa é válida:*

$$\frac{1}{2}\log\left\{(2\pi e)^d \det Var(X_1, \ldots, X_n)\right\} - H(X_1, \ldots, X_n) \tag{3.7}$$

$$\geq IM(X_1 : \ldots : X_n) - \frac{1}{2}\log\left(\frac{\det Var(X_1, \ldots, X_n)}{\prod_{k=1}^n \det Var(X_k)}\right) \tag{3.8}$$

$$\geq \sum_{k=1}^n H(X_k) - \frac{1}{2}\log\left\{(2\pi e)^d \prod_{k=1}^n \det Var(X_k)\right\}, \tag{3.9}$$

em que (3.7) assume apenas valores não negativos e (3.9) assume apenas valores não positivos. Se as v.as. são conjuntamente gaussianas a igualdade ocorre.

Demonstração. Tem-se a identidade:

$$\mathrm{IM}(X_1 : \ldots : X_n) = \sum_{k=1}^n \mathrm{H}(X_k) - \mathrm{H}(X_1, \ldots, X_n). \tag{3.10}$$

Pelo Teorema 3.1.4

$$\sum_{k=1}^n \mathrm{H}(X_k) \leq \frac{1}{2}\log\left\{(2\pi e)^d \prod_{k=1}^n \det \mathrm{Var}(X_k)\right\}$$

e

$$\mathrm{H}(X_1, \ldots, X_n) \leq \frac{1}{2}\log\left\{(2\pi e)^d \det \mathrm{Var}(X_1, \ldots, X_n)\right\}.$$

Tem-se então

$$\frac{1}{2}\log\left\{(2\pi e)^d \prod_{k=1}^n \det \mathrm{Var}(X_k)\right\} - \mathrm{H}(X_1, \ldots, X_n) \tag{3.11}$$

$$\geq \mathrm{IM}(X_1 : \ldots : X_n) \tag{3.12}$$

$$\geq \sum_{k=1}^n \mathrm{H}(X_k) - \frac{1}{2}\log\left\{(2\pi e)^d \det \mathrm{Var}(X_1, \ldots, X_n)\right\}. \tag{3.13}$$

Subtraindo

$$-\frac{1}{2}\log\left(\frac{\det(2\pi e)^d \mathrm{Var}(X_1,\ldots,X_n)}{(2\pi e)^d \prod_{k=1}^n \det \mathrm{Var}(X_k)}\right)$$

de (3.11), (3.12) e (3.13) obtém-se o resultado. □

Na Seção 3.1.2, sobre cópulas, é mostrado que a igualdade na proposição acima ocorre mesmo quando as v.as. não são gaussianas conjuntamente, bastando que elas apresentem distribuição conjunta com cópula gaussiana.

Corolário 3.1.1. *Sejam $X_1,\ldots,X_n$ como na Proposição 3.1.3. A seguinte estimativa é válida:*

$$\left|IM(X_1:\ldots:X_n) - \frac{1}{2}\log\left(\frac{\det Var(X_1,\ldots,X_n)}{\prod_{k=1}^n \det Var(X_k)}\right)\right| \tag{3.14}$$

$$\leq \frac{1}{2}\log\left\{(2\pi e)^d \prod_{k=1}^n \det Var(X_k)\right\} - H(X_1,\ldots,X_n). \tag{3.15}$$

Demonstração. Como a informação mútua é não negativa tem-se

$$\mathrm{H}(X_1,\ldots,X_n) \leq \sum_{k=1}^n \mathrm{H}(X_k),$$

em particular

$$\frac{1}{2}\log\left\{(2\pi e)^d \det \mathrm{Var}(X_1,\ldots,X_n)\right\} \leq \frac{1}{2}\log\left\{(2\pi e)^d \det \prod_{k=1}^n \mathrm{Var}(X_k)\right\}.$$

Logo, (3.15) é maior que (3.9) e (3.7). □

Este corolário será útil para se obter estimativas para as medidas de dependência linear que são discutidas no Capítulo 4.

Definição 3.1.7 (Informação mútua entre v.a. dada uma outra v.a.)**.** Sejam $X_1, \ldots, X_n, X_{n+1}$ v.a. a valores em $\mathbb{R}^{d_1}, \ldots, \mathbb{R}^{d_n}, \mathbb{R}^{d_{n+1}}$, respectivamente. Sejam p a densidade de probabilidade conjunta de $X_1, \ldots, X_n, X_{n+1}$ e $p(\cdot/x_{n+1}), p_1(\cdot/x_{n+1}), \ldots, p_n(\cdot/x_{n+1})$ as densidades de probabilidade conjunta e marginais de $X_1, \ldots, X_n$ condicionadas em $X_{n+1} = x_{n+1}$, respectivamente. A informação mútua $\mathrm{IM}(X_1, \ldots, X_n/X_{n+1})$ entre as v.as. $X_1, \ldots, X_n$ dado X_{n+1} é definida como

$$\mathrm{IM}(X_1 : \ldots : X_n/X_{n+1}) \tag{3.16}$$
$$= \int \cdots \int p(x_1, \ldots, x_n, x_{n+1}) \log \frac{p(x_1, \ldots, x_n/x_{n+1})}{\prod_{k=1}^{n} p_k(x_k/x_{n+1})} dx_1 \ldots dx_n dx_{n+1},$$

quando a integral existir e será ∞ caso contrário.

A definição 3.1.7, sem ser a mais geral, é suficiente para os objetivos presentes. A definição geral é dada em Wyner (1978) e pode-se mostrar que (3.16) assume somente valores não negativos e é nulo se e somente se $X_1, \ldots, X_n$ forem independentes condicionado em X_{n+1} (cf. Ihara (1964), p.38). Observe que a quantidade acima não é a informação mútua condicional embora na literatura não raramente seja denominada como tal. De fato, embora envolva probabilidades condicionais, a quantidade (3.16) é um número não aleatório, já que se toma a esperança de todas as probabilidades condicionais. Na literatura, muitas vezes define-se a informação mútua condicional que é uma v.a. e define-se a quantidade em 3.1.7 como sendo a esperança desta quantidade. Como a versão condicional da informação mútua não será utilizada nesta tese, optou-se por não definí-la.

Proposição 3.1.4. *Sejam $X_1, \ldots, X_n, X_{n+1}$ v.as. conjuntamente gaussianas $d_1, \ldots, d_n, d_{n+1}$-dimensionais. Assumindo que a matriz de variância/covariância $Var(X_1, \ldots, X_n, X_{n+1})$ não seja singular tem-se*

$$\begin{aligned} &IM(X_1, \ldots, X_n / X_{n+1}) \\ &= -\frac{1}{2} \log \left(\frac{\det Var(X_1, \ldots, X_n, X_{n+1}) \{\det Var(X_{n+1})\}^{n-1}}{\prod_{k=1}^{n} \det Var(X_k, X_{n+1})} \right) \end{aligned} \tag{3.17}$$

Demonstração. Basta verificar que a variância de $X_k, 1 \leq k \leq n$ condicionada em $X_{n+1} = x_{n+1}$ é a variância parcial de X_k dado X_{n+1} (veja Johnson e Wichern (1998)), isto é,

$$\mathrm{Var}(X_k / X_{n+1}) = \mathrm{Var}(X_k) - \mathrm{Cov}(X_k : X_{n+1}) \mathrm{Var}(X_{n+1})^{-1} \mathrm{Cov}(X_{n+1} : X_k),$$

ou seja, a variância de $X_k, 1 \leq k \leq n$ condicionada em $X_{n+1} = x_{n+1}$ não depende do particular valor x_{n+1} sob o qual é condicionado (Johnson e Wichern, 1998).

Agora como

$$\begin{aligned} &\begin{bmatrix} \mathrm{Var}(X_k) & \mathrm{Cov}(X_k : X_{n+1}) \\ \mathrm{Cov}(X_{n+1} : X_k) & \mathrm{Var}(X_{n+1}) \end{bmatrix} \\ &= \begin{bmatrix} I_{d_k} & A \\ 0 & I_{d_{n+1}} \end{bmatrix} \begin{bmatrix} \mathrm{Var}(X_k / X_{n+1}) & 0 \\ 0 & \mathrm{Var}(X_{n+1}) \end{bmatrix} \begin{bmatrix} I_{d_k} & 0 \\ A^T & I_{d_{n+1}} \end{bmatrix}, \end{aligned}$$

em que $A = \mathrm{Cov}(X_k : X_{n+1}) \mathrm{Var}(X_{n+1})^{-1}$, tem-se

$$\det \mathrm{Var}(X_k / X_{n+1}) = \frac{\det \mathrm{Var}(X_k, X_{n+1})}{\det \mathrm{Var}(X_{n+1})}. \tag{3.18}$$

De forma análoga, para a variância de $W^T = [X_1^T \quad \dots \quad X_n^T]$ condicionada em $X_{n+1} = x_{n+1}$ tem-se

$$\mathrm{Var}(X_1, \dots, X_n / X_{n+1}) = \mathrm{Var}(W) - \mathrm{Cov}(W : X_{n+1})\mathrm{Var}(X_{n+1})^{-1}\mathrm{Cov}(X_{n+1} : W)$$

e

$$\det \mathrm{Var}(X_1, \dots, X_n / X_{n+1}) = \frac{\det \mathrm{Var}(X_1, \dots, X_n, X_{n+1})}{\det \mathrm{Var}(X_{n+1})}. \tag{3.19}$$

Agora, pela Definição 3.1.7

$$\begin{aligned} &\mathrm{IM}(X_1, \dots, X_n / X_{n+1}) \\ &= -\frac{1}{2} \log \left(\frac{\det \mathrm{Var}(X_1, \dots, X_n / X_{n+1})}{\prod_{k=1}^{n} \det \mathrm{Var}(X_k / X_{n+1})} \right). \end{aligned} \tag{3.20}$$

Substituindo (3.18) e (3.19) em (3.20), segue o resultado. □

Para uso futuro as seguintes definições e teoremas são úteis.

Definição 3.1.8 (Entropia de v.as. dado outra v.a.)**.** Sejam $X_1, \dots, X_n, X_{n+1}$ v.a. $d_1, \dots, d_n, d_{n+1}$-dimensionais. A entropia $\mathrm{H}(X_1, \dots, X_n / X_{n+1})$ de $X_1, \dots, X_n$ dado X_{n+1} é definida como

$$\begin{aligned} &\mathrm{H}(X_1, \dots, X_n / X_{n+1}) \\ &= -\int \cdots \int p(x_1, \dots, x_n) \log p(x_1, \dots, x_n / x_{n+1}) dx_1 \dots dx_n dx_{n+1}. \end{aligned}$$

Teorema 3.1.5 (Algumas identidades)**.** *Sejam* $X_1, \ldots, X_n, X_{n+1}$ *v.as.* $d_1, \ldots, d_n, d_{n+1}$*-dimensionais com densidades de probabilidade, são válidas as seguintes identidades:*

$$IM(X_1 : \ldots : X_{n+1}) = \sum_{k=1}^{n+1} H(X_k) - H(X_1, \ldots, X_{n+1}); \tag{3.21}$$

$$IM(X_1 : \ldots : X_n / X_{n+1}) = \sum_{k=1}^{n} H(X_k / X_{n+1}) - H(X_1, \ldots, X_n / X_{n+1}); \tag{3.22}$$

$$H(X_1, \ldots, X_{n+1}) = H(X_1) + \sum_{k=1}^{n} H(X_{k+1} / X_1, \ldots, X_k); \tag{3.23}$$

$$H(X_1, \ldots, X_n / X_{n+1}) = H(X_1 / X_{n+1}) + \sum_{k=1}^{n-1} H(X_{k+1} / X_{n+1}, X_1, \ldots, X_k); \tag{3.24}$$

$$IM(X_1, \ldots, X_n : X_{n+1}) = IM(X_1 : X_{n+1}) + \sum_{k=1}^{n-1} IM(X_{k+1} : X_{n+1} / X_1, \ldots, X_k). \tag{3.25}$$

Demonstração. A identidade (3.21) é uma simples conseqüência da Definição 3.1.6 de entropia e da fórmula (3.3) da informação mútua para v.as. contínuas com densidades de probabilidade. (3.22) é conseqüência imediata da Definição 3.1.8 da entropia de v.as. dada outra v.a. e da Definição 3.1.7 de informação mútua entre v.as. dada outra v.a.

As identidades (3.23) e (3.24) são conhecidas como regras da cadeia para entropia e suas demonstrações podem ser encontradas nos teoremas 2.5.1 (caso discreto) e 8.6.2 (caso contínuo) em Cover e Thomas (1991). A identidade 3.25

está provada para o caso em que as v.as. são discretas em Cover e Thomas (1991) e para o caso contínuo a demonstração é idêntica, isto é,

$$
\begin{aligned}
&\mathrm{IM}(X_1,\ldots,X_n : X_{n+1}) \\
&= \mathrm{H}(X_1,\ldots,X_n) + \mathrm{H}(X_{n+1}) - \mathrm{H}(X_1,\ldots,X_n,X_{n+1}) \\
&= \mathrm{H}(X_1,\ldots,X_n) - \mathrm{H}(X_1,\ldots,X_n/X_{n+1}) \\
&= \mathrm{H}(X_1) + \sum_{k=1}^{n-1} \mathrm{H}(X_{k+1}/X_1,\ldots,X_k) - \mathrm{H}(X_1/X_{n+1}) \\
&- \sum_{k=1}^{n-1} \mathrm{H}(X_{k+1}/X_{n+1},X_1,\ldots,X_k) \\
&= \mathrm{H}(X_1) - \mathrm{H}(X_1/X_{n+1}) + \sum_{k=1}^{n-1} (\mathrm{H}(X_{k+1}/X_1,\ldots,X_k) \\
&- \mathrm{H}(X_{k+1}/X_{n+1},X_1,\ldots,X_k)) \\
&= \mathrm{IM}(X_1 : X_{n+1}) + \sum_{k=1}^{n-1} \mathrm{IM}(X_{k+1} : X_{n+1}/X_1,\ldots,X_k).
\end{aligned}
$$

□

Teorema 3.1.6 (Algumas desigualdades)**.** *Sejam* $X_1, \ldots, X_n, X_{n+1}$ *v.as.* $d_1, \ldots, d_n, d_{n+1}$*-dimensionais, são válidas as seguintes desigualdades:*

$$H(X_1, \ldots, X_n) + H(X_{n+1}) \geq H(X_1, \ldots, X_n, X_{n+1}); \tag{3.26}$$

$$H(X_1, \ldots, X_n, X_{n+1}) \geq H(X_1, \ldots, X_n); \tag{3.27}$$

$$H(X_1, \ldots, X_n) \geq H(X_1, \ldots, X_n / X_{n+1}); \tag{3.28}$$

$$\sum_{k=1}^{n+1} H(X_k) \geq H(X_1, \ldots, X_n, X_{n+1}); \tag{3.29}$$

$$IM(X_1 : \ldots : X_n : X_{n+1}) \geq IM(X_1 : \ldots : X_n, X_{n+1}); \tag{3.30}$$

$$IM(X_1 : \ldots : X_{n+1}) \geq IM(X_1 : \ldots : X_n); \tag{3.31}$$

$$IM(X_1 : \ldots : X_n, X_{n+1}) \geq IM(X_1 : \ldots : X_n / X_{n+1});; \tag{3.32}$$

em que as igualdades ocorrem, respectivamente, se e somente se

X_{n+1} *for independente das outras v.as. conjuntamente*;

$X_{n+1} = f(X_1, \ldots, X_n)$*, para alguma função* f *mensurável*;

X_{n+1} *for independente das outras v.as. conjuntamente*;

as v.as. forem independentes;

X_{n+1} *for independente de* X_n;

X_{n+1} *for independente das outras v.as. conjuntamente*;

X_{n+1} *for independente das outras v.as. dois a dois.*

Demonstração. Veja Cover e Thomas (1991, pp.489-493) para a prova das desigualdades (3.26) a (3.29). O restante das desigualdades são conseqüências imediatas das anteriores. □

Definição 3.1.9 (Taxa de entropia de uma seqüência de v.as.)**.** Seja $X_k, k \geq 0$ uma seqüência de v.as. n-dimensionais. A taxa de entropia $\mathrm{h}(X)$ da seqüência X_k é definida como

$$\mathrm{h}(X) = \lim_{j\to\infty} \frac{1}{j}\mathrm{H}(\{X_k\}_0^j). \tag{3.33}$$

Definição 3.1.10 (Entropia de uma v.a. dada uma seqüência de v.as.)**.** A entropia $\mathrm{H}(X/\{Y_k\}_0^\infty)$ de uma v.a. n-dimensional X dado uma seqüência de v.as. m-dimensionais $\{Y_k\}_0^\infty$ é definida como

$$\mathrm{H}(X/\{Y_k\}_0^\infty) = \lim_{j\to\infty} \mathrm{H}(X/\{Y\}_0^j). \tag{3.34}$$

A definição acima é útil na discussão de medidas de dependência para séries temporais.

3.1.2 Cópulas

Para distribuições de probabilidade multivariadas contínuas, as marginais univariadas e a estruturas de dependências podem ser separadas e a relação entre elas é estabelecida por uma família de funções denominada *cópulas*. Fato esse demonstrado por Sklar (cf. Nelsen (1999)). As cópulas têm recebido crescente atenção na literatura estatística por permitir o estudo da estrutura de dependência separadamente das distribuições marginais das v.as., mostrando-se úteis em modelagens e estimações de distribuições multivariadas (Joe, 1997; Nelsen, 1999) e, mais recentemente, tem-se demonstrado sua aplicabilidade na obtenção de resultados assintóticos para séries temporais.

É interessante e natural que se possa estabelecer relações entre cópulas e informação mútua estudada na seção anterior, uma vez que ambas se prestam

para o estudo da dependência entre variáveis aleatórias. Aqui é feita uma breve discussão a esse respeito.

Diz-se que C é uma n-cópula se é uma função de distribuição acumulada conjunta de n v.as. cujas marginais são distribuições uniformes em $[0,1]$. Equivalentemente,

Definição 3.1.11 (Cópula)**.** Uma função $C : [0,1]^n \to [0,1]$ é denominada n-cópula se satisfaz as seguintes condições:

1. $C(u_1, \ldots, u_n)$ é crescente em cada componente u_k.
2. $C(u_1, \ldots, u_{k-1}, 0, u_{k+1}, \ldots, u_n) = 0$.
3. $C(1, \ldots, 1, u_k, 1, \ldots, 1) = u_k$.
4. Para todo $(a_1, \ldots, a_n), (b_1, \ldots, b_n) \in [0,1]^n$ com $a_i \leq b_i$,

$$\sum_{j_1=1}^{2} \cdots \sum_{j_n=1}^{2} (-1)^{j_1+\ldots+j_n} C(x_{1j_1}, \ldots, x_{nj_n}) \geq 0,$$

em que $x_{k1} = a_k$ e $x_{k2} = b_k$ para todo $k \in \{1, \ldots, n\}$.

Em particular uma 1-cópula $C : [0,1] \to [0,1]$ será definida por

$$C(u) = u.$$

Definição 3.1.12 (Cópula absolutamente contínua)**.** Uma n-cópula $C : [0,1]^n \to [0,1]$ é denominada *absolutamente contínua* se, quando considerada como uma função de distribuição acumulada conjunta das n v.a. uniformes em $[0,1]$, ela tem uma densidade $c : [0,1]^n \to \mathbb{R}$ dada por

$$c(u_1, \ldots, u_n) = \frac{\partial^n C}{\partial u_1 \ldots \partial u_n}(u_1, \ldots, u_n).$$

A função c é denominada *densidade de cópula*.

O seguinte teorema é fundamental.

Teorema 3.1.7 (Sklar (1959))**.** *Sejam $X_1, \ldots, X_n$ v.as. a valores reais definidas num mesmo espaço de probabilidade, com distribuições marginais acumuladas $F_k(x_k) = P(X_k \leq x_k)$ e a distribuição acumulada conjunta $F_{1\ldots n}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$, então existe uma n-cópula $C_{1\ldots n}(u_1, \ldots, u_n)$ tal que*

$$F_{1\ldots n}(x_1, \ldots, x_n) = C_{1\ldots n}(F_1(x_1), \ldots, F_n(x_n)), \ \ \forall x_k \in \mathbb{R},\ 1 \leq k \leq n.$$

O reverso também é válido, isto é, dada uma n-cópula $C_{1\ldots n}(u_1, \ldots, u_n)$ e as v.as. $X_1, \ldots, X_n$ com marginais acumuladas $F_k(x_k) = P(X_k \leq x_k)$, a função

$$C_{1\ldots n}(F_1(x_1), \ldots, F_n(x_n)), \ \ \forall x_k \in \mathbb{R},\ 1 \leq k \leq n,$$

define uma função de distribuição acumulada conjunta das variáveis aleatórias.

Definição 3.1.13 (Informação de cópula)**.**

Sejam $X_1 = (X_1^1, \ldots, X_1^{d_1}), \ldots, X_n = (X_n^1, \ldots, X_n^{d_n})$ v.as. $d_1, \ldots, d_n$ dimensionais, respectivamente, definidas num mesmo espaço de probabilidade com distribuições marginais $F_1^1(x_1^1), \ldots, F_1^{d_1}(x_1^{d1}), \ldots, F_n^1(x_n^1), \ldots, F_n^{d_n}(x_n^{d_n})$, respectivamente, e a distribuição acumulada conjunta $F_{1\ldots n}(x_1, \ldots, x_n)$. Tome $d = \sum d_k$.

Seja $C_{1\ldots n}(u_1, \ldots, u_n)$ uma d-cópula associada à $F_{1\ldots n}(x_1, \ldots, x_n)$. A *informação de cópula* $\mathrm{IC}(X_1 : \ldots : X_n)$ entre $X_1, \ldots, X_n$ é definida como sendo a informação mútua $\mathrm{IM}(U_1 : \ldots : U_n)$ entre as v.as. uniformes $U_1, \ldots, U_n$ em $[0,1]^{d_1}, \ldots, [0,1]^{d_n}$ com função de distribuição acumulada $C_{1\ldots n}(u_1, \ldots, u_n)$, isto é,

$$\mathrm{IC}(X_1 : \ldots : X_n) = \mathrm{IM}(U_1 : \ldots : U_n).$$

Em particular, se a cópula for absolutamente contínua com densidade de cópula $c_{1\ldots n}(u_1, \ldots, u_n)$ e definindo-se $c_k(u_k)$ como a densidade asscociada a $C(1, \ldots, 1, u_k, 1, \ldots, 1), 1 \leq k \leq n$, a informação de cópula $\mathrm{IC}(X_1 : \ldots : X_n)$ pode ser escrita como

$$\begin{aligned} &\mathrm{IC}(X_1 : \ldots : X_n) \\ &= \int_{[0,1]^{d_1}} \cdots \int_{[0,1]^{d_n}} c_{1\ldots n}(u_1, \ldots, u_n) \log \frac{c_{1\ldots n}(u_1, \ldots, u_n)}{c_1(u_1) \ldots c_n(u_n)} du_1 \ldots du_n. \end{aligned} \tag{3.35}$$

Agora pode-se enunciar a seguinte proposição que relacionam as cópulas e a informação mútua.

Proposição 3.1.5. *Sejam* $X_1 = (X_1^1, \ldots, X_1^{d_1}), \ldots, X_n = (X_n^1, \ldots, X_n^{d_n})$ *v.as.* $d_1, \ldots, d_n$ *dimensionais definidas num mesmo espaço de probabilidade, com distribuições marginais* $F_1^1(x_1^1), \ldots, F_1^{d_1}(x_1^{d1}), \ldots, F_n^1(x_n^1), \ldots, F_n^{d_n}(x_n^{d_n})$*, respectivamente. Dado* $d = \sum d_k$*. Seja* $C_{1\ldots n}$ *uma d-cópula que define a distribuição acumulada conjunta* $F_{1\ldots n}$*, tal que,*

$$\begin{aligned} &F_{1\ldots n}(x_1^1, \ldots, x_1^{d1}, \ldots, x_n^1, \ldots, x_n^{d_n}) \\ &= C_{1\ldots n}(F_1^1(x_1^1), \ldots, F_1^{d_1}(x_1^{d1}), \ldots, F_n^1(x_n^1), \ldots, F_n^{d_n}(x_n^{d_n})), \\ &\forall x_k \in \mathbb{R},\ 1 \leq k \leq n. \end{aligned}$$

A informação mútua $IM(X_1 : \ldots : X_n)$ *é igual à informação de cópula* $IC(X_1 : \ldots : X_n)$*, isto é,*

$$IM(X_1 : \ldots : X_n) = IC(X_1 : \ldots : X_n). \tag{3.36}$$

Demonstração. Pela definição de informação de cópula, tem-se

$$\begin{aligned} &\mathrm{IC}(X_1 : \ldots : X_n) \\ &\quad = \mathrm{IM}(U_1 : \ldots : U_n) \\ &\quad = \mathrm{IM}((U_1^1, \ldots, U_1^{d_1}) : \ldots : (U_n^1, \ldots, U_n^{d_n})) \\ &\quad = \mathrm{IM}((F_1^1(X_1^1), \ldots, F_1^{d_1}(X_1^{d_1})) : \ldots : (F_n^1(X_n^1), \ldots, F_n^{d_n}(X_n^{d_n}))) \qquad (3.37) \\ &\quad = \mathrm{IM}((X_1^1, \ldots, X_1^{d_1}) : \ldots : (X_n^1, \ldots, X_n^{d_n})) \qquad (3.38) \\ &\quad = \mathrm{IM}(X_1 : \ldots : X_n). \end{aligned}$$

A igualdade entre (3.37) e (3.38) é válida pelo fato de a função

$$F_k : \mathbb{R}^{d_k} \to [0,1]^{d_k}$$

$$(x_k^1, \ldots, x_k^{d_k}) \mapsto (F_k^1(x_k^1), \ldots, F_k^{d_k}(x_k^{d_k})),$$

$1 \leq k \leq n$, ser estritamente crescente e contínua termo a termo, ou seja, é bijetora, permitindo aplicar o Teorema 3.1.2. □

Observação 3.1.3. A proposição 3.1.5 é aparentemente nova. Há resultados para o caso particular de quando as v.as. $X_1, \ldots, X_n$ são univariadas, as distribuições acumuladas são diferenciáveis e a cópula associada é absolutamente contínua. Neste caso, basta fazer uma mudança de variável na integral (3.35) e obtém-se o resultado desejado. Veja Jenison e Reale (2004); Mercierand et al. (2006). Na referência (Mercierand et al., 2006) estuda-se o caso de duas v.as. X_1, X_2 com cópula de Marshall-Olkin definida para $u_1, u_2 \in [0,1]$ por:

$$C(u_1, u_2) = \min(u_1^{1-\theta} u_2, u_1 u_2^{1-\theta}), \quad \theta \in [0,1).$$

Então, a informação mútua entre X_1 e X_2 é dada por

$$\mathrm{IM}(X_1 : X_2) = 2\frac{1-\theta}{2-\theta}\log(1-\theta) - \frac{\theta}{2-\theta} + \frac{\theta^2}{(2-\theta)^2}.$$

A Proposição 3.1.5 permite estudar questões envolvendo a informação mútua utilizando técnicas desenvolvidas para as cópulas e vice-versa. Uma outra conseqüência importante da Proposição 3.1.5 é que a informação mútua não depende das marginais das v.as. envolvidas, mas somente da cópula.

A seguinte definição é útil.

Definição 3.1.14 (Cópula gaussiana)**.** A *n-cópula gaussiana*[a] é definida por

$$C(u_1, \ldots, u_n) = \Phi_\Gamma(\Phi^{-1}_{\Gamma_{11}}(u_1), \ldots, \Phi^{-1}_{\Gamma_{nn}}(u_n)), \tag{3.39}$$

em que Φ_Γ é a função de distribuição acumulada gaussiana n-variada com matriz de variância/covariância Γ e média zero e $\Phi^{-1}_{\Gamma_{kk}}$ são as funções inversas das funções de distribuição acumulada gaussianas univariadas com variância $\Gamma_{kk},\ k = 1, \ldots, n$ e média zero.

[a]Na literatura (Nelsen, 1999) é comum se referir como cópula gaussiana a cópula definida analogamente, porém com distribuições gaussianas com variância um no lugar de distribuições gaussianas com matriz de variâncias/covariâncias quaisquer. As definições são idênticas, bastando normalizar as v.as.

Observe que, se $C(u_1, \ldots, u_k, \ldots, u_n)$ é uma n-cópula gaussiana, $C(u_1, \ldots, 1, \ldots, u_n)$ será uma $(n-1)$-cópula gaussiana, pois,

$$\begin{aligned}
&\Phi_\Gamma(\Phi^{-1}_{\Gamma_{11}}(u_1), \ldots, \Phi^{-1}_{\Gamma_{kk}}(1), \ldots, \Phi^{-1}_{\Gamma_{nn}}(u_n)) \\
&= \Phi_\Gamma(\Phi^{-1}_{\Gamma_{11}}(u_1), \ldots, \infty, \ldots, \Phi^{-1}_{\Gamma_{nn}}(u_n)) \\
&= \Phi_\Gamma(\Phi^{-1}_{\Gamma_{11}}(u_1), \ldots, \Phi^{-1}_{\Gamma_{(k-1)(k-1)}}(u_{k-1}), \Phi^{-1}_{\Gamma_{(k+1)(k+1)}}(u_{k+1}), \ldots, \Phi^{-1}_{\Gamma_{nn}}(u_n)).
\end{aligned}$$

A seguinte definição será útil para enunciar alguns dos resultados desta tese.

Definição 3.1.15 (V.as. com cópula gaussiana)**.** Diz-se que as v.as. $X_1, \ldots, X_n,\ n \geq 2$, univariadas, apresentam *cópula gaussiana com matriz de covariância/variância* Γ, se a função de distribuição acumulada for definida por uma cópula gaussiana com matriz de covariância/variância Γ e as distribuições marginais forem tais que $\mathrm{Var}(X_k) = \Gamma_{kk},\ k = 1, \ldots, n$, ou seja, se as variâncias da cópula e das marginais forem compatíveis.

Diz-se que as v.as. $Y_1, \ldots, Y_n$ não necessariamente univariadas apresentam cópula gaussiana se os seus componentes univariados apresentarem cópula gaussiana.

Sejam $X_1, \ldots, X_n$ v.as. unidimensionais. As v.as. apresentam distribuição gaussiana conjunta n-variada com matriz de variância/covariância Γ se e somente se apresentam cópula gaussiana e marginais gaussianas univariadas com variâncias $\Gamma_{kk},\ k = 1, \ldots, n$.

Pela Proposição 3.1.5, a informação mútua pode ser caracterizada pela função de cópula associada às v.as., independentemente da distribuição de suas marginais, e dessa forma pode-se generalizar a Proposição 3.1.5.

Proposição 3.1.6 (Informação mútua para v.as. com cópula gaussiana)**.** *Sejam $X_1, \ldots, X_n$ v.as. $d_1, \ldots, d_n$-dimensionais com distribuição acumulada conjunta definida por uma cópula gaussiana. Assumindo que a matriz de variância/covariância $Var(X_1, \ldots, X_n)$ não seja singular tem-se*

$$IM(X_1 : \ldots : X_n) = -\frac{1}{2} \log \left(\frac{\det Var(X_1, \ldots, X_n)}{\prod_{k=1}^{n} \det Var(X_k)} \right)$$

Demonstração. Sejam $Y_1, \ldots, Y_n$ v.as. conjuntamente gaussianas $d_1, \ldots, d_n$-dimensionais com $\mathrm{Var}(Y_1, \ldots, Y_n) = \mathrm{Var}(X_1, \ldots, X_n)$. Pela Proposição 3.1.5,

tem-se que

$$\begin{aligned}\mathrm{IM}(Y_1 : \ldots : Y_n) &= -\frac{1}{2}\log\left(\frac{\det \mathrm{Var}(Y_1, \ldots, Y_n)}{\prod_{k=1}^{n} \det \mathrm{Var}(Y_k)}\right) \\ &= -\frac{1}{2}\log\left(\frac{\det \mathrm{Var}(X_1, \ldots, X_n)}{\prod_{k=1}^{n} \det \mathrm{Var}(X_k)}\right).\end{aligned}$$

Agora, pela Proposição 3.1.5 e definição da informação de cópula

$$\begin{aligned}\mathrm{IM}(Y_1 : \ldots : Y_n) &= \mathrm{IC}(Y_1 : \ldots : Y_n) \\ &= \mathrm{IC}(X_1 : \ldots : X_n) \\ &= \mathrm{IM}(X_1 : \ldots : X_n).\end{aligned}$$

□

Com esta proposição obtém-se a mesma fórmula para o caso gaussiano para v.as. com distribuição conjunta não necessariamente gaussiana, mas com cópula gaussiana.

O seguinte corolário que generaliza a equivalência entre a nulidade da correlação e independência para v.as. com cópula gaussiana é importante.

Corolário 3.1.2 (Independência para v.as. com cópula gaussiana)**.** *Sejam $X_1, \ldots, X_n$ v.as. $d_1, \ldots, d_n$-dimensionais, com cópula gaussiana com covariância Γ. As v.as. $X_1, \ldots, X_n$ são independentes se e somente se*

$$\det \Gamma = \prod_{k=1}^{n} \det \mathit{Var}(X_k), \tag{3.40}$$

ou equivalentemente, se e somente se $\mathit{Cov}(X_k : X_l) = 0, 1 \leq k < l \leq n$.

Demonstração. A informação mútua $\mathrm{IM}(X_1, \ldots, X_n)$ é zero se e somente se as v.as. forem independentes. No caso das v.as. com cópula gaussiana, pela

Proposição 3.1.6, $\mathrm{IM}(X_1, \ldots, X_n) = 0$ se e somente se $\det \Gamma = \prod_{k=1}^{n} \det \mathrm{Var}(X_k)$ que por sua vez ocorre se e somente se $\mathrm{Cov}(X_k : X_l) = 0, 1 \leq k < l \leq n$. □

As afirmações envolvendo a informação mútua para v.as. conjuntamente gaussianas se tornam afirmações sobre informação mútua para v.as. com cópula gaussiana pela Proposição 3.1.6. Contudo, deve-se tomar o seguinte cuidado: não é claro que as v.as. com cópula gaussiana sejam fechadas quanto às operações lineares, isto é, se a combinação linear de v.as. com cópula gaussiana resulta numa v.a. com cópula gaussiana. Este último fato limita consideravelmente os resultados que podem ser generalizados, pois operações como parcializações que envolvem a combinação linear de v.as. podem não resultar em v.as. com cópula gaussiana.

3.2 Conclusão

Definiu-se a informação mútua, introduziu-se alguns teoremas e provou-se resultados que são utilizados para se demonstrar os resultados dos capítulos seguintes. No caso em que as v.as. envolvidas são gaussianas, pode-se obter expressões para quantidades da Teoria da Informação baseando-se somente nos momentos de segunda ordem das v.as. Embora as v.as. gaussianas constituam uma família bastante específica de v.as., elas ocupam na Teoria da Informação papel central, permitindo que se obtenha limites superiores e inferiores para a entropia e informação mútua. Algumas igualdades e desiguadades de quantidades da Teoria da Informação foram introduzidas e são utilizadas nos capítulos seguintes.

A relação entre a informação mútua e a cópula estabelecida pela Proposição 3.1.5 permite que alguns resultados sobre a informação mútua sejam obtidos em

termos de cópulas. Em particular, resultados como "se as v.as. forem conjuntamente gaussianas ... a informação mútua ..." pode ser substituída por resultados do tipo "se as v.as. apresentarem cópula gaussiana ... a informação mútua ... ", o que não exige que as marginais sejam gaussianas. A Proposição 3.1.5 justifica denominar a informação mútua como uma medida de dependência e em muitos casos permite que se obtenham fórmulas explícitas para a informação mútua entre v.as. que de outra forma seriam intratáveis. Uma aplicação importante da Proposição 3.1.6 é feita no Capítulo 7 em que é obtida uma expressão exata para a taxa de informação mútua entre séries temporais exigindo somente que as séries sejam estacionárias de segunda ordem e apresentem distribuições conjuntas com cópula gaussiana. É um resultado que, sem o uso da idéia de cópulas, em geral não é simples de obter, pois a taxa de informação mútua não é sequer bem definida, em geral, para processos que não sejam estritamente estacionários.

Desta forma, o trabalho desenvolvido aqui pode ser inserido num contexto mais amplo e com outras possibilidades de generalização além daquelas estudadas especificamente nesta tese.

CAPÍTULO 4

Medidas de dependência linear

"Two organs in the same individual, or in a connected pair of individuals, are said to be correlated when a series of the first organ of a definite size being selected, the mean of the sizes of the correponding second organs is found to be a function of the size of the selected organ. If the mean is independent, the organs are said to be non correlated. Correlation is defined mathematically by any constants, which determine the above function." (Karl Pearson, 1896)

A correlação linear de Pearson, doravante denominada correlação, e suas generalizações baseadas somente nas propriedades do segundo momento das v.as. são denominadas genericamente de medidas de dependência linear. Denominação que provém, provavelmente, do fato que dadas duas v.as. X e Y unidimensionais de média nula, quando o módulo do valor da correlação é um, apresentam uma relação linear exata, isto é, $X = aY,\ a \neq 0$[1]. Por outro

[1]Lembrando que as v.as. nesta tese apresentam média nula.

lado, quando a correlação assume valor no intervalo $(-1, 1)$, pode-se escrever $X = \alpha Y + \xi$ em que $\mathrm{Var}(\xi) \neq 0$ e $\mathrm{Cov}(Y : \xi) = 0$. Ou seja, de certa forma, a correlação mede o grau de linearidade da relação.

As medidas de dependência linear caracterizam a estrutura de dependência das v.as. somente em casos específicos (Nelsen, 1999) como quando a cópula que define a distribuição conjunta é gaussiana (veja Proposição 3.1.6), em particular, quando as v.as. envolvidas apresentam distribuição conjuntamente gaussiana. Este resultado aparentemente restringe a aplicabilidade das medidas linear. Apesar disso, alguns fatos tornam estas medidas bastante atraentes:

1. No caso em que as v.as. são gaussianas, as medidas lineares caracterizam completamente sua estrutura de dependência, ou seja, medidas de dependência gerais, como a informação mútua, reduzem-se a funções das medidas lineares. Isto permite que métodos utilizando medidas lineares sejam generalizados de forma natural.

2. Como discutido no Capítulo 3, é possível estabelecer limites superiores e inferiores para a diferença entre a informação mútua e as medidas de dependência linear.

3. Quando as v.as. são interpretadas como elementos dos espaços de Hilbert $L^2(\Omega, \mathcal{F}, P)$ de todas as funções quadrado integráveis no espaço de probabilidade $(\Omega, \mathcal{F}, P)$, pode-se usar os métodos de Análise Funcional para caracterizar a estrutura de dependência das v.as. não necessariamente gaussianas (Goodman e Johnson (2004); Hannan (1961); Lancaster (1958)) e, notadamente, conceitos como correlação canônica e correlação desempenham papel fundamental (Hannan (1961)).

4. Recentemente, pesquisadores da área de "machine learning"e estatística têm utilizado a teoria dos núcleos dos operadores entre espaços de Hilbert[2] para tratar problemas não lineares por métodos lineares em que conceitos como correlação tem papel central (Cucker e Smale, 2002).

5. As medidas lineares são bastante intuitivas, com interpretação geométrica relacionada ao ângulo e à distância entre subespaços. Por exemplo, a correlação entre duas v.as. X e Y é o cosseno entre os subespaços gerados por X e Y.

6. Seguramente são os métodos mais bem estudados do ponto de vista estatístico e computacional, com estudos de robustez e flutuações estatísticas e de aspectos numéricos computacionais precisos.

7. Embora as medidas de dependência linear apresentem limitações, diferentemente de muitos outros métodos, estes são bem conhecidos.

Neste capítulo, estudam-se em detalhes formas canônicas de construção de medidas de dependência entre v.as. A Proposição 3.1.1 será utilizada sistematicamente para mostrar a interpretação das diferentes medidas lineares à luz da Teoria da Informação.

Inicialmente, na Seção 4.1 é discutida a importante idéia de regressão entre v.as. ou equivalentemente da projeção ortogonal entre subespaços gerados por elas. Sucintamente, dadas duas v.as. X e Y, pode-se escrever uma v.a. X como a soma de uma v.a. não correlacionada e uma outra v.a. com correlação 1 com relação a Y. O procedimento é utilizado sistematicamente para a construção das medidas de dependência linear. Novamente, as v.as. gaussianas desempenham

[2]Na literatura em inglês é conhecida como "reproducing kernel hilbert space theory"

papel central devido a relação que existe entre projeção ortogonal e esperança condicional para esta família de v.as.

Na Seção 4.2.1 discute-se a noção de correlação para duas v.a. unidimensionais. Esta é então generalizada para o caso de duas v.a. multidimensionais e posteriormente para o caso de mais de duas v.a. multidimensionais. A generalização discutida na Seção 4.2.2 é denominada correlação quadrática total e desempenha um papel central nesta tese, apresentando uma relação um para um com a informação mútua entre v.as. com cópula gaussiana.

Na seção 4.2.3 é introduzida a idéia de parcialização das medidas de dependência que consiste em estudar a relação entre duas v.as. X e Y descontando o efeito de uma terceira v.a. Z. A correlação quadrática total parcializada é obtida de forma natural como resultado do procedimento de parcialização e é estabelecida a sua relação com as informações mútuas entre v.as. dado um outro conjunto de v.as.

Na seção 4.2.4 é discutido o conceito de inversão da matriz de covariância /variância. O inverso da matriz de variância/covariância apresenta um papel importante na compreensão das medidas de dependência linear. Embora algumas de suas propriedades tenham apareciado de forma esporádica na literatura, aparentemente não há estudos sistemáticos de suas propriedades e da relação com as medidas de dependência linear. Nesta tese, tentou-se sistematizar o estudo de alguns aspectos da inversão. Em particular, dado um conjunto de v.as. $X_1, \ldots, X_n$, são definidas as v.as. inversas ${}^iX_1, \ldots, {}^iX_n$ que são as v.as. cuja matriz de variância é a matriz inversa da matriz de variância/covariância de $X_1, \ldots, X_n$. A introdução das v.as. inversas permite que se obtenha resultados que de outra forma seriam difíceis de se obter e ao mesmo tempo respondem a

questões como: qual a interpretação para o inverso da matriz de coeficientes da regressão entre duas v.as.?

4.1 Regressão, projeção ortogonal, esperança condicional e v.as. gaussianas

Neste capítulo é estudada a teoria de medidas de dependência linear para v.as. não necessariamente univariadas definidas num espaço de probabilidade $(\Omega, \mathcal{F}, P)$. Para faciliar a discussão são introduzidos nesta seção algumas definições e resultados utilizados neste capítulo. Nesta seção, as v.as. podem ser reais ou complexas, apresentam média nula e a matriz de variância/covariância é positiva definida. As v.as. são ditas apresentarem distribuição gaussiana ou cópula gaussiana se são v.as. reais com distribuição gaussiana multivariada ou com cópula gaussiana, respectivamente.

As v.as. univariadas são entendidas como elementos do espaço de Hilbert separável $L^2(\Omega, \mathcal{F}, P)$ das v.as. univariadas com variância finita definidas num espaço de probabilidade $(\Omega, \mathcal{F}, P)$. O produto escalar $< X, Y >$ entre duas v.as. X e Y unidimensionais é definida como $\langle X, Y \rangle = \text{Cov(X,Y)}$. Por abuso de notação diz-se que uma v.a. n-dimensional $X \in L^2(\Omega, \mathcal{F}, P)$ se $X_k \in L^2(\Omega, \mathcal{F}, P), k = 1, \ldots, n$. A convergência da seqüência de v.as. n-dimensionais $X \in L^2$ é entendida como convergência em L^2 dos seus termos univariados, isto é, $X_k \to X$ para $k \to \infty$ em L^2 se e somente se $X_k^j \to X^j, j = 1, \ldots, n$ para $k \to \infty$ em L^2.

Para o tratamento unificado das v.as. multidimensionais, a seguinte noção de ortogonalidade é útil.

Definição 4.1.1 (Ortogonalidade)**.** Sejam X e Y v.as. n e m-dimensionais. Elas são ditas *ortogonais* ou não correlacionadas se $\mathrm{Cov}(X_k : Y_l) = 0, 1 \leq k \leq n, 1 \leq l \leq m$, isto é, $\mathrm{Cov}(X : Y) = 0$.

Essa definição de ortogonalidade está bem definida mesmo para v.as. de dimensões distintas.

Tem-se a seguinte caracterização das v.as. ortogonais.

Proposição 4.1.1 (Caracterização da ortogonalidade)**.** *As v.as. X e Y n e m-dimensionais são ortogonais se e somente se vale*

$$\det |\mathit{Var}(X, Y)| = \det |\mathit{Var}(X)| \det |\mathit{Var}(Y)|. \tag{4.1}$$

Demonstração. Sejam Z e W v.as. n e m-dimensionais gaussianas tais que $\mathrm{Var}(X, Y) = \mathrm{Var}(Z, W)$. Pela desigualdade (3.26) da Proposição 3.1.6, tem-se que a entropia $H(Z, W) = H(Z) + H(W)$ se e somente se Z e W forem independentes. Duas v.as. conjuntamente gaussianas são independentes se e somente se $\mathrm{Cov}(Z : W) = 0 = \mathrm{Cov}(X : Y)$, ou seja,

$$\det |\mathrm{Var}(X, Y)| = \det |\mathrm{Var}(X)| \det |\mathrm{Var}(Y)|. \tag{4.2}$$

□

Corolário 4.1.1. *As v.as. $X_1, \ldots, X_n$ v.as. $d_1, \ldots, d_n$-dimensionais, respectivamente, são ortogonais duas a duas se e somente se vale*

$$\det |\mathit{Var}(X_1, \ldots, X_n)| = \prod_{k=1}^{n} \det |\mathit{Var}(X_k)|.$$

Demonstração. A prova é por indução. Suponha que vale $\det |\mathrm{Var}(X_1, \ldots, X_{n-1})| = \prod_{k=1}^{n-1} \det |\mathrm{Var}(X_k)|$. Agora, basta observar que X_n e $[X_1 \ \ldots \ X_{n-1}]^T$ são ortogonais, se as v.as. são ortogonais dois a dois. Então aplicando-se a Proposição

4.1.1 obtém-se $\det|\mathrm{Var}(X_1,\ldots,X_n)| = \det|\mathrm{Var}(X_n)|\det|\mathrm{Var}(X_1,\ldots,X_{n-1})|$. O resultado segue pela hipótese de indução e a Proposição 4.1.1. □

Observe que $\mathrm{Cov}(\cdot:\cdot)$ apresenta propriedades muito semelhantes ao produto interno usual, isto é, sejam X,Y e Z v.as. complexas n, m e n-dimensionais, respectivamente, então

1. $\mathrm{Cov}(X:Y) = \mathrm{Cov}(Y:X)^*$.
2. $\mathrm{Cov}(AX:Y) = A\mathrm{Cov}(X:Y)$ em que A é uma matriz complexa $n\times n$.
3. $\mathrm{Cov}(X:BY) = \mathrm{Cov}(X:Y)B^*$ em que B é uma matriz complexa $m\times m$.
4. $\mathrm{Cov}(X:X)$ é positiva semidefinida.
5. $\mathrm{Cov}(X:X) = 0$ se e somente se $X = 0$.
6. $\mathrm{Cov}(X+Z:Y) = \mathrm{Cov}(X:Y) + \mathrm{Cov}(Z:Y)$.

Não se trata de um produto interno (escalar) usual, pois em geral os valores deste produto interno são matrizes. De fato, o espaço das v.as. n-dimensionais com variância finita juntamente com esse "produto interno" foram estudadas por Wiener e Masani (1957) que desenvolveram a teoria dos processos estacionários n-dimensionais utilizando esse espaço, porém para o estudo desenvolvido nesta tese não é necessário o uso dessa teoria, exceto algumas propriedades que são introduzidas como definições e teoremas no que segue.

Para o caso de v.as. unidmensionais a noção de *subespaço* é a usual, ou seja, é um subconjunto $M\subset L^2$ não vazio tal que, se $X,Y\in M$, então $aX + bY\in M$ para todo a,b reais (complexas) se as v.as. forem reais (complexas) e é fechado na topologia da norma da variância. O *subespaço gerado* pelas

v.as. unidimensionsionais $X_1, \ldots, X_n$, n eventualmente infinto, é o conjunto $M = \overline{\text{span}}\{X_1, \ldots, X_n\}$.

Para as v.as. multidimensionais é definida a seguinte noção de subespaço gerado.

Definição 4.1.2 (Subespaço gerado)**.** O subespaço gerado de L^2 pelas v.as. $X_1, \ldots, X_n$ $d_1, \ldots, d_n$-dimensionais com n eventualmente infinito, com $\sup_{k \geq 1} d_k < \infty$, é o subespaço de L^2 gerado pelas v.as. unidimensionais $X_1^1, \ldots, X_1^{d_1}, \ldots, X_n^1, \ldots, X_n^{d_n}$.

Doravante, o termo subespaço indica o subespaço de L^2.

O seguinte teorema é fundamental.

Teorema 4.1.1 (Projeção ortogonal)**.** *Seja $X \in L^2$ uma v.a. unidimensional e $M \subset L^2$ um subespaço, então existe uma única v.a. unidimensional $Y \in M$ e uma única v.a. unidimensional ξ tal que,*

$$X = Y + \xi,$$

$$Cov(Z : \xi) = 0, \forall Z \in M.$$

A v.a. Y é o único elemento de M que satisfaz

$$Var(X - Y) = \inf_{Y' \in M} Var(X - Y').$$

A v.a. Y é denominada projeção ortogonal de X em M e é denotada $\bar{E}(X/M)$ ou $\bar{E}(X/X_1, \ldots, X_n)$ quando M for gerado pelas v.as. unidimensionais $X_1, \ldots, X_n$. ξ é denominada resíduo da projeção ortogonal de X em M e é denotada por $\bar{R}(X/M)$ ou $\bar{R}(X/X_1, \ldots, X_n)$.

No caso das v.as. multidimensionais define-se a projeção ortogonal da seguinte forma.

Definição 4.1.3 (Projeção ortogonal para v.as. multidimensionais)**.** Sejam $X, X_1, \ldots, X_n$ v.as. $d, d_1, \ldots, d_n$-dimensionais, respectivamente, com n eventualmente infinito e $\sup_{k\geq 1} d_k < \infty$. A projeção ortogonal $\bar{E}(X/M)$ da v.a. X no subespaço M gerado pelas v.as. $X_1, \ldots, X_n$ é a única v.a. d-dimensional $\bar{E}(X/M) = [\bar{E}(X_1/M) \ \ldots \ \bar{E}(X_d/M)]^T$, ou seja, a projeção de X em M é a v.a. formada pelas projeções ortogonais de seus componentes univariados. A notação $\bar{E}(X/X_1, \ldots, X_n)$ também indica a projeção de X em M.

O seguinte teorema é uma conseqüência imediata do Teorema 4.1.1.

Teorema 4.1.2. *Seja $\mathcal{M}$ um conjunto não vazio de v.as. reais (complexas) d-dimensionais em L^2 tal que (a) se $f, g \in \mathcal{M}$, $Af + Bg \in \mathcal{M}$ para todas as matrizes $d \times d$ reais (complexas) e (b) fechada. A projeção $E(X/M)$ de X no subespaço M gerado pelas v.as em $\mathcal{M}$ é o único elemento de $\mathcal{M}$ que satisfaz*

$$TrVar(X - E(X/M)) = \inf_{Y' \in \mathcal{M}} TrVar(X - Y').$$

Demonstração. Veja Wiener e Masani (1957, p.131, lema 5.8) □

Um conceito relacionado a projeção é o de *regressão*, termo utilizado neste texto para indicar a regressão linear com minimização de erro quadrático médio, isto é,

Definição 4.1.4 (Regressão linear quadrática)**.** Sejam as v.as. $Y, X_1, \ldots, X_n \in L^2$, n eventualmente infinito, com dimensões $d, d_1, \ldots, d_n$, respectivamente e $\sup_{k\geq 1} d_k < \infty$. Os *coeficientes de regressão* de Y nas v.as. $X_1, \ldots, X_n$ são definidos como sendo as matrizes de coeficientes $A_1, \ldots, A_n$ com dimensões $d \times d_1, \ldots, d \times d_n$, respectivamente, tais que minimizem o *erro quadrático médio*

$$\text{Tr}\left\{\text{Var}\left(Y - \sum_{k=1}^{n} A_k' X_k\right)\right\}, \tag{4.3}$$

em que $\text{Tr}B, B \in \mathbb{R}^{m\times m}, m \geq 1$, é o traço da matriz B. A v.a. sobre a qual se calcula a variância é denominada *resíduo* da regressão. A v.a. $\sum_{k=1}^{n} A_k X_k$ é denominada projeção ortogonal de Y no subespaço gerado por $X_1, \ldots, X_n$.

Pelo Teorema 4.1.2 é claro que $\sum_{k=1}^{n} A_k X_k = E(Y/X_1, \ldots, X_n)$ e portanto a projeção sempre existe e é única. Isto implica, em particular, que os coeficientes $A_1, \ldots, A_n$ existem e são únicos. No problema de regressão, um dos principais objetivos é recuperar os coeficientes de regressão utilizando somente a matriz de covariância/variância $\text{Var}(Y, X_1, \ldots, X_n)$, o que é sempre possível no caso em que n é finito e a matriz $\text{Var}(X_1, \ldots, X_n)$ é positiva definida.

No caso em que n é infinito, o problema é mais delicado, pois depende do procedimento específico utilizado para recuperar os coeficientes. Esta questão se torna importante principalmente no caso de séries temporais como é discutida mais adiante.

Outro conceito relacionado à projeção é a esperança condicional. No caso das v.as. com variância finita tem-se uma equivalência, num certo sentido, entre a projeção e a esperança condicional.

Teorema 4.1.3. *Seja $L^2(\Omega, \mathcal{F}, P)$ o espaço das v.as. unidimensionais com variância finita definidas num espaço de probabilidade $(\Omega, \mathcal{F}, P)$. Seja $X \in L^2(\Omega, \mathcal{F}, P)$ e a σ-algebra $\mathcal{G} \subset \mathcal{F}$, a esperança condicional $E(X/\mathcal{G})$ de X dado $\mathcal{G}$ é a projeção $E(X/\mathcal{H})$ de X sobre o subespaço $\mathcal{H} \subset L^2(\Omega, \mathcal{F}, P)$ das funcões $\mathcal{G}$-mensuráveis. Em particular, seja $X \in L^2(\Omega, \mathcal{F}, P)$ e $\mathcal{G}$ a σ-algebra gerada pelas v.as. $Y_k \in L^2(\Omega, \mathcal{F}, P), k = 1, \ldots, n$, então a esperança condicional $E(X/\mathcal{G})$ é a v.a. que satisfaz*

$$Var(X - E(X/\mathcal{G})) \leq Var(X - g),$$

em que g é $\mathcal{G}$-mensurável, ou seja, g pode ser escrito como $g(Y_1, \ldots, Y_n)$.

Demonstração. Veja Loéve (1994, p. 128). □

O caso multivariado do Teorema acima é uma simples conseqüência do caso univariado, pois, dada a v.a. n-dimensional $Y \in L^2(\Omega, \mathcal{F}, P)$, $E(Y/\mathcal{G}) = [E(Y_1/\mathcal{G}) \ \ldots \ E(Y_n/\mathcal{G})]^T$, ou seja, basta considerar os termos univariados separadamente.

Em geral, calcular a esperança condicional não é simples, embora para v.as. gaussianas valha:

Teorema 4.1.4. *Sejam $Y, X_1, \ldots, X_n$ v.as. conjuntamente gaussianas $d, d_1, \ldots, d_n$-dimensionais, n eventualmente infinito e $\sup_{k\geq 1} d_k < \infty$. A esperança condicional de Y dado $X_1, \ldots, X_n$ é igual à projeção ortogonal de Y no subespaço gerado por $X_1, \ldots, X_n$, isto é,*

$$E(Y/X_1, \ldots, X_n) = \bar{E}(Y/X_1, \ldots, X_n).$$

Em particular,

$$TrVar(Y - \bar{E}(Y/X_1, \ldots, X_n)) \leq TrVar(Y - \bar{E}(Y/g(X_1, \ldots, X_n))),$$

Em que $g : \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_n} \to \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_n}$ e $TrVar(g(x_1, \ldots, X_n)) < \infty$.

Demonstração. Veja Loéve (1994, p. 128). □

O teorema acima mostra que, em geral, a troca do operador de projeção ortogonal $\bar{E}(\cdot/\cdot)$ pela esperança condicional $E(\cdot/\cdot)$ permite uma generalização ipsis literis de muitos conceitos desenvolvidos nesta tese.

Por fim, têm-se os seguintes resultados que ilustram o fato de as v.as. gaussianas estarem intimamente relacionadas com as medidas de dependência linear, a ponto de muitos autores misturarem o conceito de gaussianidade com linearidade.

Teorema 4.1.5. *Sejam $X_1, \ldots, X_n$ v.as. não necessariamente gaussianas $d_1, \ldots, d_n$-dimensionais com n eventualmente infinito, $\sup_{k\geq 1} d_k < \infty$. Existem v.as. $Y_1, \ldots, Y_n$ gaussianas $d_1, \ldots, d_n$-dimensionais tais que $Var(X_k, X_l) = Var(Y_k, Y_l), k, l = 1, \ldots, n$.*

Demonstração. Veja Loéve (1994, p. 133). □

Teorema 4.1.6. *O espaço das v.as gaussianas unidimensionais é fechado quanto à combinação linear e a convergência na norma da variância.*

Demonstração. Veja o teorema do fecho (B) na p. 134 e a observação 37.6 na p. 151 em Loéve (1994). Também veja Ibragimov e Rozanov (1978, pp. 5-6) para uma discussão sobre a convergência de séries de v.as. gaussianas. □

4.2 Medidas de dependência entre v.as.

4.2.1 Correlação

A correlação entre duas variáveis aleatórias talvez seja uma das medidas de dependência mais clássicas e bem estabelecidas na literaratura científica. Uma revisão interessante sobre aspectos históricos da correlação linear pode ser encontrada em Rodgers e Nicewander (1988).

Definição 4.2.1 (Correlação linear)**.** Sejam X e Y duas v.a. unidimensionais. A *correlação linear* ou simplesmente correlação $\rho(X : Y)$ entre X e Y é definida por

$$\rho(X : Y) = \frac{\mathrm{Cov}(X : Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

É imediato pela desigualdade de Cauchy-Schwartz que $|\rho(X : Y)| \leq 1$. Se X e Y forem não correlacionadas $\rho(X : Y) = 0$ e $|\rho(X : Y)| = 1$ se e somente se $Y = aX$, $a \neq 0$.

A correlação é invariante a mudança de escala (a menos do sinal) e translação, ou seja,

$$\rho(X : Y) = \frac{ab}{|ab|}\rho(aX - c : bY - d), \quad a, b \neq 0.$$

Pode-se associar o significado geométrico de ângulo entre X e Y a $\rho(X : Y)$, pois a correlação pode ser escrita como

$$\rho(X : Y) = \frac{< X, Y >}{\sqrt{\|X\|\|Y\|}},$$

em que o produto escalar $< X, Y >= \mathrm{Cov}(X : Y)$.

Quando as v.as. X e Y apresentam cópula gaussiana, existe uma relação simples entre a correlação e a informação mútua entre X e Y.

Proposição 4.2.1 (Correlação e informação mútua)**.** *Sejam X e Y v.as. unidimensionais com cópula gaussiana. A informação mútua $IM(X : Y)$ pode ser escrita como*

$$IM(X : Y) = -\frac{1}{2} \log \left(\frac{\det Var(X, Y)}{Var(X)\, Var(Y)} \right) \tag{4.4}$$

$$= -\frac{1}{2} \log \left(1 - \rho(X : Y)^2\right). \tag{4.5}$$

Demonstração. Basta verificar que

$$\begin{aligned} \frac{\det \mathrm{Var}(X, Y)}{\mathrm{Var}(X)\mathrm{Var}(Y)} &= \frac{\mathrm{Var}(Y)\mathrm{Var}(X) - |\mathrm{Cov}(X : Y)|^2}{\mathrm{Var}(X)\mathrm{Var}(Y)} \\ &= 1 - \frac{|\mathrm{Cov}(X : Y)|^2}{\mathrm{Var}(X)\mathrm{Var}(Y)}. \end{aligned}$$

□

Para o caso de v.as. multivariadas a seguinte definição de matriz de correlação permite caracterizar a inter-relação entre as v.as. duas a duas.

Definição 4.2.2. Sejam $X_1, \ldots, X_n$ v.as. não necessariamente unidimensionais. A matriz de correlação $\mathrm{Corr}(X_1 : \ldots : X_n)$ entre as v.as. $X_1, \ldots, X_n$ é definida por

$$\mathrm{Corr}(X_1 : \ldots : X_n)$$
$$= \begin{bmatrix} I & \cdots & \mathrm{Var}(X_1)^{-1/2}\mathrm{Cov}(X_1 : X_n)\mathrm{Var}(X_n)^{-1/2} \\ \vdots & \ddots & \vdots \\ \mathrm{Var}(X_n)^{-1/2}\mathrm{Cov}(X_n : X_1)\mathrm{Var}(X_1)^{-1/2} & \cdots & I \end{bmatrix},$$

ou, equivalentemente,

$$\mathrm{Corr}(X_1 : \ldots : X_n)$$
$$= \mathrm{diag}(\mathrm{Var}(X_1), \ldots, \mathrm{Var}(X_n))^{-1/2}\mathrm{Var}(X_1, \ldots, X_n)\mathrm{diag}(\mathrm{Var}(X_1), \ldots, \mathrm{Var}(X_n))^{-1/2},$$

em que, dadas as matrizes $B_1, \ldots, B_n$, $\mathrm{diag}(B_1, \ldots, B_n)$ indica a matriz bloco diagonal com as matrizes $B_1, \ldots, B_n$ dispostas nas diagonais em blocos (veja a notação no capítulo 2).

No caso em que $X_1, \ldots, X_n$ são v.as. unidimensionais, a matriz de correlação é simplesmente a matriz cujos elementos são as correlações entre as v.as., isto é,

$$\mathrm{Corr}(X_1 : \ldots : X_n) = \begin{bmatrix} \rho(X_1 : X_1) & \cdots & \rho(X_1 : X_n) \\ \vdots & \ddots & \vdots \\ \rho(X_n : X_1) & \cdots & \rho(X_n : X_n) \end{bmatrix},$$

em que $\rho(X_k : X_k) = 1,\ k = 1, \ldots, n.$

A matriz de correlação tem a vantagem de exibir a estrutura de dependência dois a dois isoladamente, porém apresenta a desvantagem de não ser uma medida de dependência que resume a estrutura de dependência em um único número, como será feito na próxima seção.

4.2.2 Correlação quadrática total

Há diferentes generalizações multivariadas do conceito de coeficiente correlação ou para o módulo quadrado do coeficiente de correlação, porém para o propósito desta tese a definição abaixo é a mais adequada.

Definição 4.2.3. Sejam X e Y v.as. n e m-dimensionais, respectivamente, a *correlação quadrática total* $\mathrm{CQT}(X:Y)$ entre X e Y é definida como:

$$\mathrm{CQT}(X:Y) = 1 - \frac{\det \mathrm{Var}(X,Y)}{\det \mathrm{Var}(X) \det \mathrm{Var}(Y)} \tag{4.6}$$

$$= 1 - \det \mathrm{Corr}(X:Y). \tag{4.7}$$

No caso de X e Y serem unidimensionais $\mathrm{CQT}(X:Y) = \rho(X:Y)^2$.

> *Observação* 4.2.1. A correlação quadrática total (CQT) foi denominada correlação generalizada por Kotz et al. (2000), porém pelo fato de a CQT ser uma generalização do módulo quadrático da correlação e não da correlação, optou-se pela primeira nomenclatura que exprime melhor o conceito.

A definição acima pode ser facilmente generalizada para mais de duas v.as.

Definição 4.2.4. Sejam $X_1, \ldots, X_n$ v.as. cada uma não necessariamente univariada, a correlação quadrática total $\mathrm{CQT}(X_1 : \ldots : X_n)$ entre as v.as. é dada por

$$\mathrm{CQT}(X_1 : \ldots : X_n) = 1 - \frac{\det \mathrm{Var}(X_1, \ldots, X_n)}{\det \mathrm{Var}(X_1) \ldots \det \mathrm{Var}(X_n)} \tag{4.8}$$

$$= 1 - \det \mathrm{Corr}(X_1 : \ldots : X_n). \tag{4.9}$$

.

No caso em que as v.as. apresentam cópula gaussiana obtém-se a seguinte relação entre a CQT e a informação mútua.

Proposição 4.2.2. *Sejam $X_1, \ldots, X_n$ v.as. com cópula gaussiana cada uma não necessariamente univariada, a informação mútua $IM(X_1 : \ldots : X_n)$ é uma função monotônica crescente da correlação quadrática total $CQT(X_1 : \ldots : X_n)$, isto é,*

$$IM(X_1 : \ldots : X_n) = -\frac{1}{2}\log(1 - CQT(X_1 : \ldots : X_n)). \tag{4.10}$$

Demonstração. É uma conseqüência imediata da Proposição 3.1.6 e da definição de CQT. □

A proposição acima é verdadeira mesmo no caso singular se a proposição 3.1.2 for utilizada e a definição da CQT for modificada de acordo, porém por simplicidade, como já foi destacada anteriormente, as v.as consideradas neste texto apresentam matrizes de variância/covariância não singulares.

O seguinte corolário pode ser obtido.

Corolário 4.2.1. *Sejam $X_1, \ldots, X_n$ v.as. cada uma não necessariamente univariadas, a $CQT(X_1 : \ldots : X_n)$ assume valor no intervalo $[0, 1)$. Em particular, $CQT(X_1 : \ldots : X_n) = 0$ se e somente se as v.as. forem ortogonais duas a duas.*

Demonstração. Pelo corolário 4.1.1, tem-se $\mathrm{CQT}(X_1 : \ldots : X_n) = 0$ se e somente se $X_1, \ldots, X_n$ forem ortogonais. Sejam $Y_1, \ldots, Y_n$ v.as. gaussianas tais que $\mathrm{Var}(X_1, \ldots, X_n) = \mathrm{Var}(Y_1, \ldots, Y_n)$ que existe pelo Teorema 4.1.5. Agora, pela Proposição 4.2.2 e o fato da informação mútua assumir somente valores não negativos, segue que $0 \leq \mathrm{CQT}(X_1 : \ldots : X_n) = \mathrm{CQT}(Y_1 : \ldots : Y_n) < 1$. Pelo fato de as v.as. consideradas neste capítulo apresentarem matriz $\mathrm{Var}(X_1, \ldots, X_n)$ positiva definida, o caso em que $\mathrm{CQT}(X_1 : \ldots : X_n) = 0$ é excluído. □

4.2.3 Parcialização

Aqui é discutido o procedimento de parcialização das medidas de dependência, uma forma de estudar a dependência linear entre um conjunto de v.as. "descontando" parte da relação devido a um outro grupo de v.as. Mais especificamente, tome três v.as. univariadas X, Y e Z. Uma possível questão é como medir a correlação entre X e Y subtraindo aquela parte da relação linear devido a Z. Para tanto, calcula-se o resíduo ξ_x da regressão de X em Z e o resíduo ξ_y da regressão de Y em Z. Agora, ξ_x e ξ_y são ortogonais a Z e portanto a correlação $\rho(\xi_x : \xi_y)$ é a correlação entre os componentes de X e Y que não apresentam dependência linear com Z. A correlação $\rho(\xi_x : \xi_y)$ é conhecida como correlação parcial e indicada por $\rho(X : Y/Z)$.

> *Observação* 4.2.2. Note que, em geral, a correlação parcial ou parcializada não é a correlação condicionada, embora na literatura exista um certo grau de confusão sobre estes dois conceitos. Isto se deve à existência de diferentes definições de correlação condicionada e também de casos para o qual a correlação parcializada e a condicional são equivalentes. Notadamente, no caso gaussiano, em que a correlação parcial nula indica independência condicional, obtém-se os mesmos valores para as duas correlações. Uma discussão bastante interessante sobre as diferenças e condições de equivalência entre os dois conceitos pode ser encontrada em (Baba et al., 2004). Para finalizar esta pequena consideração, em geral, a palavra "condicional" ou "condicionada" é reservada para quantidades em que elas mesmas são v.as., o que não é o caso da correlação parcial que é sempre um valor não aleatório.

Definem-se as medidas de dependência linear parcializadas da seguinte forma:

Definição 4.2.5 (Medidas parcializadas)**.** Sejam $X_1, \ldots, X_n$ v.as. com dimensões $d_1, \ldots, d_n$, respectivamente, e $M \subset L^2$ um subespaço tal que o subespaço gerado por X_k, $k = 1, \ldots, n$ não está contido em M, ou seja, $\text{span}\{X_k^1, \ldots, X_k^{d_k}\} \nsubseteq M$, $k = 1, \ldots, n$[a]. Sejam $\xi_1, \ldots, \xi_n$ os resíduos de projeção ortogonal de $X_1, \ldots, X_n$ em M, isto é, $\xi_k = \bar{R}(X_k/M)$, $k = 1, \ldots, n$. A CQT parcializada $\text{CQT}(X_1 : \ldots : X_n/M)$ e a matriz de correlação parcializada $\text{Corr}(X_1 : \ldots : X_n/M)$ de $X_1, \ldots, X_n$ dado M são definidas como

$$\text{CQT}(X_1 : \ldots : X_n/M) = \text{CQT}(\xi_1 : \ldots : \xi_n), \tag{4.11}$$

$$\text{Corr}(X_1 : \ldots : X_n/M) = \text{Corr}(\xi_1 : \ldots : \xi_n). \tag{4.12}$$

Se o subespaço M for gerado pelas v.as. $Z_1, \ldots, Z_m$, m eventualmente infinito, pode se denotá-las, respectivamente, por $\text{CQT}(X_1 : \ldots : X_n/Z_1, \ldots, Z_m)$ e $\text{Corr}(X_1 : \ldots : X_n/Z_1, \ldots, Z_m)$.

[a] Esta última restrição é somente para garantir que a matriz de variância/covariância dos resíduos não seja singular.

Por esta definição, é claro que a CQT e a matriz de correlação parcializadas apresentam as mesmas propriedades da respectivas medidas não parcializadas, em particular,

Proposição 4.2.3. *Sejam $X_1, \ldots, X_n$ e $M \subset L^2$ como na Definição 4.2.5, a CQT$(X_1 : \ldots : X_n/M)$ assume valor no intervalo $[0, 1)$. Em particular, CQT$(X_1 : \ldots : X_n/M) = 0$ se e somente se as v.as. forem ortogonais duas a duas dado M.*

Demonstração. Como $\text{CQT}(X_1 : \ldots : X_n/M) = \text{CQT}(\xi_1 : \ldots : \xi_n)$, em que $\xi_1, \ldots, \xi_n$ são os resíduos da projeção ortogonal de $X_1, \ldots, X_n$ em M, basta aplicar o Corolário 4.2.1 em $\text{CQT}(\xi_1 : \ldots : \xi_n)$. □

Intimamente relacionado à Definição 4.2.5 de CQT parcializada é a definição

de *variância parcial* de $X_1, \ldots, X_n$ dado M que é simplesmente $\mathrm{Var}(X_1, \ldots, X_n/M)$ $= \mathrm{Var}(\xi_1, \ldots, \xi_n)$. Quando M é gerado pelas v.as. $Z_1, \ldots, Z_m$, m eventualmente infinito, pode se denotá-lo por $\mathrm{Var}(X_1, \ldots, X_n/Z_1, \ldots, Z_m)$. A seguinte proposição é útil.

Proposição 4.2.4. *Sejam $X_1, \ldots, X_n$ como na Definição 4.2.5 e $Z_1, \ldots, Z_m$ v.as. não necessariamente unidimensionais com n e m finitos. A variância parcializada $Var(X_1, \ldots, X_n/Z_1, \ldots, Z_m)$ de $X_1, \ldots, X_n$ dado $Z_1, \ldots, Z_m$ é dada por*

(4.13)

$$Var(X_1, \ldots, X_n/Z_1, \ldots, Z_m) = Var(X_1, \ldots, X_n)$$
$$- Cov(X_1, \ldots, X_n : Z_1, \ldots, Z_m)\, Var(Z_1, \ldots, Z_m)^{-1}\, Cov(Z_1, \ldots, Z_m : X_1, \ldots, X_n).$$

Demonstração. Tomando $X^T = [X_1^T \ \ldots \ X_n^T]$ e
$Z^T = [Z_1^T \ \ldots \ Z_n^T]$, tem-se, por definição,

$$\begin{aligned}\mathrm{Var}(X_1, \ldots, X_n/Z_1, \ldots, Z_m) &= \mathrm{Var}(X/Z) \\ &= \mathrm{Var}(\xi),\end{aligned}$$

em que ξ é o resíduo da regressão de X em Z.

Também por definição

$$X = AZ + \xi, \tag{4.14}$$

em que $\mathrm{Cov}(\xi : Z) = 0$. Logo,

$$\mathrm{Cov}(X : Z) = A\mathrm{Var}(Z),$$

ou seja, $A = \text{Cov}(X : Z)\text{Var}(Z)^{-1}$. Agora substituindo em (4.14) tem-se

$$\xi = X - \text{Cov}(X : Z)\text{Var}(Z)^{-1}Z.$$

Então,

$$\begin{aligned}
\text{Var}(\xi) &= \text{Cov}(\xi : \xi) \\
&= \text{Cov}(\xi : X) \\
&= E(X - \text{Cov}(X : Z)\text{Var}(Z)^{-1}Z, X) \\
&= \text{Var}(X) - \text{Cov}(X : Z)\text{Var}(Z)^{-1}\text{Cov}(Z : X).
\end{aligned}$$

□

A *covariância parcializada* é definida analogamente à variância parcializada. Sejam X e Y v.as. não necessariamente univariadas e $M \subset L^2$ um subespaço. A covariância de X e Y dado M é simplesmente $\text{Cov}(X : Y/M) = \text{Cov}(\xi_x, \xi_y)$, em que ξ_x e ξ_y são, respectivamente, os resíduos da projeção ortogonal de X e Y em M. Quando M é gerado pelas v.as. $Z_1, \ldots, Z_m$, m eventualmente infinito, pode-se denotá-lo por $\text{Cov}(X : Y/Z_1, \ldots, Z_m)$.

Corolário 4.2.2. *Sejam X, Y e Z v.as. n, m e d-dimensionais, respectivamente. A covariância parcializada de X e Y dado Z é*

$$Cov(X : Y/Z) = Cov(X : Y) - Cov(X : Z)\, Var(Z)^{-1}\, Cov(Z : Y). \tag{4.15}$$

Demonstração. Pela Proposição 4.2.4

$$\begin{aligned}
&\mathrm{Var}(X, Y/Z) = \mathrm{Var}(X, Y) - \mathrm{Cov}(X, Y : Z)\mathrm{Var}(Z)^{-1}\mathrm{Cov}(Z : X, Y) \\
&= \begin{bmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X : Y) \\ \mathrm{Cov}(Y : X) & \mathrm{Var}(Y) \end{bmatrix} \\
&- \begin{bmatrix} \mathrm{Cov}(X : Z)\mathrm{Var}(Z)^{-1}\mathrm{Cov}(Z : X) & \mathrm{Cov}(X : Z)\mathrm{Var}(Z)^{-1}\mathrm{Cov}(Z : Y) \\ \mathrm{Cov}(Y : Z)\mathrm{Var}(Z)^{-1}\mathrm{Cov}(Z : X) & \mathrm{Cov}(Y : Z)\mathrm{Var}(Z)^{-1}\mathrm{Cov}(Z : Y) \end{bmatrix}.
\end{aligned}$$

Comparando os elementos desta última matriz com $\mathrm{Var}(X, Y/Z)$, obtém-se o resultado. □

O seguinte resultado é utilizado repetidas vezes nesta tese e é útil enunciá-lo como proposição para facilitar os desenvolvimentos que se seguem.

Proposição 4.2.5. *Sejam $X_1, \ldots, X_n, X_{n+1}$ v.as. $d_1, \ldots, d_n, d_{n+1}$-dimensionais. A seguinte decomposição do determinante da variância é válida:*

$$\begin{aligned}
&\det \mathit{Var}(X_1, \ldots, X_n, X_{n+1}) \\
&= \det \mathit{Var}(X_{n+1}) \det \mathit{Var}(X_1, \ldots, X_n/X_{n+1}) \qquad (4.16) \\
&= \det \mathit{Var}(X_{n+1}) \prod_{k=1}^{n} \det \mathit{Var}(X_{n+1-k}/X_{n+1}, \ldots, X_{n+2-k}). \qquad (4.17)
\end{aligned}$$

Demonstração. Sejam $Y_1, \ldots, Y_n, Y_{n+1}$ v.as. conjuntamente gaussianas $d_1, \ldots, d_n$, d_{n+1}-dimensionais tais que $\mathrm{Var}(Y_1, \ldots, Y_{n+1}) = \mathrm{Var}(X_1, \ldots, X_{n+1})$. Segue da identidade (3.23) do Teorema 3.1.5 que

$$\begin{aligned}
&\mathrm{H}(Y_1, \ldots, Y_n, Y_{n+1}) = \mathrm{H}(Y_n) + \mathrm{H}(Y_1, \ldots, Y_n/Y_{n+1}) \qquad (4.18) \\
&= \mathrm{H}(Y_{n+1}) + \sum_{k=1}^{n} \mathrm{H}(Y_{n+1-k}/Y_{n+1}, \ldots, Y_{n+2-k}). \qquad (4.19)
\end{aligned}$$

Pelas Proposições 3.1.1 3.1.4, e (4.18) tem-se

$$\begin{aligned}
&\det \mathrm{Var}(X_1, \ldots, X_n, X_{n+1}) \\
&= \det \mathrm{Var}(Y_1, \ldots, Y_n, Y_{n+1}) \\
&= \det \mathrm{Var}(Y_{n+1}) \det \mathrm{Var}(Y_1, \ldots, Y_n / Y_{n+1}) \\
&= \det \mathrm{Var}(X_{n+1}) \det \mathrm{Var}(X_1, \ldots, X_n / X_{n+1}).
\end{aligned}$$

Por sua vez (4.19) implica

$$\begin{aligned}
&\det \mathrm{Var}(X_1, \ldots, X_n, X_{n+1}) \\
&= \det \mathrm{Var}(Y_1, \ldots, Y_n, Y_{n+1}) \\
&= \det \mathrm{Var}(Y_{n+1}) \prod_{k=1}^{n} \det \mathrm{Var}(Y_{n+1-k} / Y_{n+1}, \ldots, Y_{n+2-k}) \\
&= \det \mathrm{Var}(X_{n+1}) \prod_{k=1}^{n} \det \mathrm{Var}(X_{n+1-k} / X_{n+1}, \ldots, X_{n+2-k}).
\end{aligned}$$

□

Proposição 4.2.6. *Sejam $X_1, \ldots, X_n, X_{n+1}$ v.as. $d_1, \ldots, d_n, d_{n+1}$-dimensionais. A CQT parcializada $CQT(X_1 : \ldots : X_n / X_{n+1})$ entre $X_1, \ldots, X_n$ dado X_{n+1} é expressa por*

(4.20)
$$CQT(X_1 : \ldots : X_n / X_{n+1}) = \frac{\det Var(X_1, \ldots, X_n / X_{n+1})}{\prod_{k=1}^{n} \det Var(X_k / X_{n+1})}$$

(4.21)
$$= \frac{\det Var(X_1, \ldots, X_n, X_{n+1}) \det Var(X_{n+1})^{n-1}}{\prod_{k=1}^{n} \det Var(X_k, X_{n+1})}.$$

Demonstração. A identidade (4.20) segue da definição de CQT condicional e da

variância condicional. A equação (4.21) segue do fato que, pelo Corolário 4.2.5,

$$\det \mathrm{Var}(X_1, \ldots, X_n / X_{n+1}) = \det \mathrm{Var}(X_1, \ldots, X_n, X_{n+1}) \det \mathrm{Var}(X_{n+1})^{-1}$$

e

$$\det \mathrm{Var}(X_k / X_{n+1}) = \det \mathrm{Var}(X_k, X_{n+1}) \det \mathrm{Var}(X_{n+1})^{-1},$$

para $k = 1, \ldots, n$. □

No caso gaussiano tem-se a seguinte relação:

Proposição 4.2.7. *Sejam $X_1, \ldots, X_n, X_{n+1}$ v.as. conjuntamente gaussianas $d_1, \ldots, d_n, d_{n+1}$-dimensionais. Assumindo que a matriz de variância/covariância $Var(X_1, \ldots, X_n, X_{n+1})$ não seja singular tem-se*

$$IM(X_1 : \ldots : X_n / X_{n+1}) = -\frac{1}{2} \log \left(1 - CQT(X_1 : \ldots : X_n / X_{n+1})\right). \tag{4.22}$$

Em particular, $CQT(X_1 : \ldots : X_n / X_{n+1}) = 0$ se e somente se $X_1, \ldots, X_n$ são condicionalmente independentes dado X_{n+1}.

Demonstração. Pelas Proposições 3.1.4 e 4.2.6, a identidade (4.22) é imediata. A independência condicional segue do fato de que a informação mútua condicional é zero se e somente se as v.as. são condicionalmente independentes (veja observação logo abaixo da Proposição 3.1.7). □

4.2.4 Inversão

O inverso ou uma versão adequadamente normalizada do inverso da matriz de correlação/covariância é comumente empregada em problemas de regressão.

Uma revisão sobre o uso da inversa da matriz de correlação em estatística multivariada é feita em Raveh (1985) e para uma discussão sobre a relação entre inverso da matriz de covariância e a interpolação de séries temporais estacionárias vide Bhansali (1990).

Nesta seção, o objetivo é explicitar o papel da inversa da correlação/covariância na interpretação das medidas de dependência linear entre as v.as. Inicialmente, alguns resultados gerais sobre inversas de matrizes são obtidas e então utilizadas para se reinterpretar algumas medidas de dependências linear.

Lema 4.2.1 (Decomposição de Frobenius-Schur)**.** *Sejam A_{11}, A_{12}, A_{21} e A_{22} matrizes de dimensões $n \times n$, $n \times m$, $m \times n$ e $m \times m$. Suponha que A_{11} e A_{22} sejam não singulares. Considere a matriz particionada*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}. \tag{4.23}$$

As seguintes identidades são válidas:

$$A = \begin{bmatrix} I & 0 \\ A_{21}A_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix} \begin{bmatrix} I & A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix}; \tag{4.24}$$

e

$$A = \begin{bmatrix} I & A_{21}A_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{11} - A_{12}A_{22}^{-1}A_{22} & 0 \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ A_{22}^{-1}A_{21} & I \end{bmatrix}; \tag{4.25}$$

Demonstração. Como $A_{22} - A_{21}A_{11}^{-1}A_{12}$ e $A_{11} - A_{12}A_{22}^{-1}A_{22}$ existem, basta multiplicar as matrizes e verificar que o produto coincide com a matriz A. □

O lema acima já foi utilizado em algumas provas nas seções anteriores.

O seguinte lema bem conhecido é importante.

Lema 4.2.2 (Inversa da matriz particionada)**.** *Sejam $A_{11}, A_{12}, A_{21}, A_{22}$ e A como no Lema 4.2.1. Suponha ainda que A é não singular. Defina as matrizes $D = (A_{11} - A_{12}A_{22}^{-1}A_{22})^{-1}$ e $G = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$. A inversa da matriz A pode ser escrita como*

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}GA_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}G \\ -GA_{21}A_{11}^{-1} & G \end{bmatrix} \tag{4.26}$$

$$= \begin{bmatrix} D & -DA_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}D & A_{22}^{-1} + A_{22}^{-1}A_{21}DA_{12}A_{22}^{-1} \end{bmatrix}. \tag{4.27}$$

Demonstração. Usando a identidade (4.23) do Lema 4.2.1 tem-se

$$A^{-1} = \begin{bmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & G \end{bmatrix} \begin{bmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{bmatrix}. \tag{4.28}$$

Agora basta multiplicar as matrizes e obtém-se (4.26).

Para obter (4.27) basta utilizar (4.24) e proceder analogamente. □

Agora, seguem alguns corolários para os lemas.

Corolário 4.2.3. *Sejam X e Y v.as. n e m-dimensionais. Seja $X = BY + \xi_x$ e $Y = CX + \xi_y$ as equações de regressão de X em Y e vice-versa. A inversa da matriz de variância/covariância $Var(X, Y)$ pode-ser escrita da seguinte forma*

$$Var(X,Y)^{-1} = \begin{bmatrix} Var(X/Y)^{-1} & -Var(X/Y)^{-1}B \\ -Var(Y/X)^{-1}C & Var(Y/X)^{-1} \end{bmatrix} \tag{4.29}$$

$$= \begin{bmatrix} Var(X/Y)^{-1} & -C^T Var(Y/X)^{-1} \\ -B^T Var(X/Y)^{-1} & Var(Y/X)^{-1} \end{bmatrix}. \tag{4.30}$$

Demonstração. Pela Proposição 4.2.4,

$$\text{Var}(X/Y) = \text{Var}(X) - \text{Cov}(X : Y)\text{Var}(Y)^{-1}\text{Cov}(Y : X), \tag{4.31}$$

$$\text{Var}(Y/X) = \text{Var}(Y) - \text{Cov}(Y : X)\text{Var}(X)^{-1}\text{Cov}(X : Y). \tag{4.32}$$

Pela equação de regressão

$$\text{Cov}(X : Y) = B\text{Var}(Y), \tag{4.33}$$

$$\text{Cov}(Y : X) = C\text{Var}(X), \tag{4.34}$$

de onde segue que $B = \text{Cov}(X : Y)\text{Var}(Y)^{-1}$ e $C = \text{Cov}(Y : X)\text{Var}(X)^{-1}$.

Usando o Lema 4.2.2 segue o resultado. □

Usando os resultados anteriores pode-se provar a seguinte frase que faz parte do folclore da Estatística: a inversa da matriz de correlação é a matriz das correlações parciais. Antes, é útil introduzir uma notação para as *submatrizes* e definir a *matriz de correlação inversa*.

Definição 4.2.6 (Submatriz)**.** Seja A uma matriz $n \times m$. A submatriz $[A]_L^K$ de A é uma matriz $\#L \times \#K$, em que $\#L$ indica o número de elementos do conjunto, formada pelos (l, k)-ésimos elementos de A em que $l \in L \subset \{1, \ldots, n\}$ e $k \in K \subset \{1, \ldots, m\}$. Ou seja, $[A]_L^K$ é a submatriz formada pelos elementos de A cujos índices das linhas são elementos de L e da coluna são elementos de K.

Definição 4.2.7 (Matriz de correlação inversa)**.** Seja $X_1, \ldots, X_n$ v.as. $d_1, \ldots, d_n$-dimensionais. Tome $d = \sum_{k=1}^{n} d_k$.

Defina $D_k = \{\sum_{l=0}^{k-1} d_l + 1, \ldots, \sum_{l=0}^{k} d_l\}, k = 1, \ldots, n$, em que $d_0 = 0$. Seja Λ a matriz

$$\Lambda = \text{diag}([\text{Corr}(X_1, \ldots, X_n)^{-1}]_{D_1}^{D_1}, \ldots, [\text{Corr}(X_1, \ldots, X_n)^{-1}]_{D_n}^{D_n}).$$

A matriz de correlação inversa $\text{iCorr}(X_1 : \ldots : X_n)$ é definida por

$$\text{iCorr}(X_1, \ldots, X_n) = \Lambda^{-1/2}\text{Corr}(X_1, \ldots, X_n)^{-1}\Lambda^{-1/2}. \tag{4.35}$$

Ou seja, a matriz de correlação inversa é a inversa da matriz de correlação normalizada pelas diagonais blocadas. No caso em que as v.as. são unidimensionais, a matriz de correlação inversa é a inversa da matriz de correlação normalizada pela diagonal principal.

Proposição 4.2.8 (Correlação inversa e correlação parcial)**.** *Seja* $X = [X_1 \ \ldots \ X_n]^T$ *uma v.a. n-dimensional. O módulo do* (l, k)*-ésimo elemento da matriz de correlação inversa* $iCorr(X_1 : \ldots : X_n)$*,* $[iCorr(X_1 : \ldots : X_n)]_l^k$*, é o módulo da correlação parcial* $|\rho(X_l, X_k/X^{l,k})|$ *de* X_l *e* X_k *dado o restante dos* $(n-2)$ *componentes de* X *denotado por* $X^{l,k}$*.*

Demonstração. Sem perda de generalidade, assuma que $l = 1$ e $k = 2$. Se não for o caso, basta permutar as linhas e as colunas e verificar a alteração do sinal nos determinantes devido à permutação. Pelo Corolário 4.2.3,

$$[\text{Var}(X)^{-1}]_{\{1,2\}}^{\{1,2\}} = \text{Var}(X_1, X_2/X^{1,2})^{-1}.$$

Como $\text{Var}(X_1, X_2/X^{1,2})^{-1}$ é uma matriz 2×2, apresenta a seguinte forma

simples

$$\begin{aligned}&\mathrm{Var}(X_1, X_2/X^{1,2})^{-1}\\&= \frac{1}{\det \mathrm{Var}(X_1, X_2/X^{1,2})} \begin{bmatrix} \mathrm{Var}(X_2/X^{1,2}) & -\mathrm{Cov}(X_2 : X_1/X^{1,2}) \\ -\mathrm{Cov}(X_1 : X_2/X^{1,2}) & \mathrm{Var}(X_1/X^{1,2}) \end{bmatrix}.\end{aligned}$$

Assim

$$\begin{aligned}&|[\mathrm{iCorr}(X_1 : \ldots : X_n)]_1^2|\\&= \frac{|\mathrm{Cov}(X_2 : X_1/X^{1,2}) \det \mathrm{Var}(X_1, X_2/X^{1,2})^{-1}|}{\{|\mathrm{Var}(X_2/X^{1,2})\mathrm{Var}(X_1/X^{1,2}) \det \mathrm{Var}(X_1, X_2/X^{1,2})^{-2}|\}^{1/2}}\\&= \frac{|\mathrm{Cov}(X_2 : X_1/X^{1,2})|}{\{|\mathrm{Var}(X_2/X^{1,2})\mathrm{Var}(X_1/X^{1,2})|\}^{1/2}}\\&= |\rho(X_2, X_1/X^{1,2})|\\&= |\rho(X_1, X_2/X^{1,2})|.\end{aligned}$$

□

Esta proposição não se generaliza naturalmente para o caso geral em que as v.as. são multivariadas, ou seja, assumindo as mesmas hipóteses da Definição 4.2.7 e da Proposição 4.2.8 não é verdade, em geral, que

$$[\mathrm{iCorr}(X_1 : \ldots : X_n)]_{D_l}^{D_k} = Q,$$

em que $Q = \mathrm{Var}(X_k/X^{kl})^{-1/2}\mathrm{Cov}(X_k : X_l/X^{kl})\mathrm{Var}(X_l/X^{kl})^{-1/2}$.

De fato, após um cálculo trabalhoso, obtém-se que

$$[\mathrm{iCorr}(X_1 : \ldots : X_n)]_{D_l}^{D_k}$$
$$= \mathrm{Var}(X_k/X^k)^{-1/2}\mathrm{Var}(X_k/X^{kl})^{1/2}Q\mathrm{Var}(X_l/X^{kl})^{-1/2}\mathrm{Cov}(X_l/X^l)^{1/2}.$$

No caso em que as v.as. são univariadas

$$\mathrm{Var}(X_k/X^k)^{-1/2}\mathrm{Var}(X_k/X^{kl})^{1/2} = (\mathrm{Var}(X_l/X^{kl})^{-1/2}\mathrm{Cov}(X_l/X^l)^{1/2})^{-1},$$

e, portanto, segue a validade da Proposição 4.2.8.

Como este fato não será mais utilizado nesta tese e se trata apenas de um cálculo tedioso, os detalhes da prova não são apresentados. O fato importante é que, embora a Proposição 4.2.8 não se generalize naturalmente, pode-se mostrar que existe uma importante relação entre as correlações inversas e as correlações parciais. Os resultados que se seguem são importantes neste contexto.

A seguir são definidas as v.as. inversas e é demonstrado que as v.as. inversas apresentam a estrutura de dependência linear determinada pela matriz de correlação inversa.

Definição 4.2.8 (V.as. inversas)**.** Sejam $X_1, \ldots, X_n$ v.as. $d_1, \ldots, d_n$-dimensionais , n eventualmente infinito, $\sup_{k\geq 1} d_k < \infty$. As v.as. inversas ${}^iX_k$, para $k = 1, \ldots, n$, são, respectivamente, os resíduos das projeções ortogonais de X_k em X^k que é o subespaço gerado pelo restante das v.as. $X_l, l \neq k$, normalizadas pelo inverso das suas variâncias. Mais especificamente,

$$ {}^iX_k = \mathrm{Var}(X_k/X^k)^{-1}\bar{R}(X_k/X^k), \tag{4.36}$$

e portanto $\mathrm{Var}({}^iX_k) = \mathrm{Var}(X_k/X^k)^{-1}$.

É claro pela definição que vale $\mathrm{Cov}(X_k, {}^i X_k) = I, k = 1, \ldots, n$ e $\mathrm{Cov}(X_k, {}^i X_l) = 0, k \neq l$, ou seja, as v.as. e suas inversas são bi-ortonormais.

A seguinte proposição justifica a introdução da definição de v.as. inversas.

Proposição 4.2.9. *Sejam $X_1, \ldots, X_n$ v.as. não necessariamente unidimensionais e ${}^iX_1, \ldots, {}^i X_n$ as suas respectivas v.as. inversas. Tem-se*

$$ \mathit{Var}(X_1, \ldots, X_n)^{-1} = \mathit{Var}({}^iX_1, \ldots, {}^i X_n) \tag{4.37}$$

$$ \mathit{iCorr}(X_1 : \ldots : X_n) = \mathit{Corr}({}^iX_1 : \ldots : {}^i X_n). \tag{4.38}$$

Demonstração. Pela definição de v.a. inversa tem-se a seguinte equação de regressão:

$$\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} A_{11} & \ldots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{n1} & \ldots & A_{nn} \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} + \begin{bmatrix} \mathrm{Var}(\bar{R}(X_1/X^1))({}^iX_1) \\ \vdots \\ \mathrm{Var}(\bar{R}(X_n/X^n))({}^iX_n) \end{bmatrix},$$

em que $A_{kk} = 0, k = 1, \ldots, n$. Equivalentemente,

$$\begin{bmatrix} I - A_{11} & \ldots & -A_{1n} \\ \vdots & \ddots & \vdots \\ -A_{n1} & \ldots & I - A_{nn} \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} \mathrm{Var}(\bar{R}(X_1/X^1))({}^iX_1) \\ \vdots \\ \mathrm{Var}(\bar{R}(X_n/X^n))({}^iX_n) \end{bmatrix}. \tag{4.39}$$

Denomine A a matriz no lado esquerdo em (4.39) e

$$V = \mathrm{diag}(\mathrm{Var}(\bar{R}(X_1/X^1)), \ldots, \mathrm{Var}(\bar{R}(X_n/X^n))).$$

Multiplicando (4.39) por $[{}^iX_1^T \ \ldots; {}^iX_n^T]$, obtém-se

$$A\mathrm{Cov}(X_1, \ldots, X_n :{}^i X_1, \ldots, {}^i X_n) = V\mathrm{Var}({}^iX_1, \ldots, {}^i X_n).$$

Como $\mathrm{Cov}(X_k :{}^i X_l) = 0,\ k \neq l$ e $\mathrm{Cov}(X_k, {}^i X_k) = \mathrm{Var}\bar{R}(X_k/X^k)\mathrm{Var}({}^iX_k) = I$, tem-se

$$A = V\mathrm{Var}({}^iX_1, \ldots, {}^i X_n). \tag{4.40}$$

Agora substituindo (4.40) em (4.39) e multiplicando por $[X_1^T \ \ldots; X_n^T]$,

$$\begin{aligned} V\mathrm{Var}({}^iX_1, \ldots, {}^i X_n)\mathrm{Var}(X_1, \ldots, X_n) &= V\mathrm{Cov}({}^iX_1, \ldots, {}^i X_n : X_1, \ldots, X_n) \\ &= V. \end{aligned}$$

Portanto,

$$\mathrm{Var}({}^iX_1, \ldots, {}^i X_n) = \mathrm{Var}(X_1, \ldots, X_n)^{-1},$$

o que conclui (4.37). Finalmente, usando a definição de matriz de correlação inversa conclui-se (4.38). □

Corolário 4.2.4 (Reflexividade)**.** *Sejam $X_1, \ldots, X_n$ v.as. não necessariamente univariadas e ${}^{i}X_1, \ldots, {}^{i}X_n$ as suas respectivas v.as. inversas. Sejam ${}^{ii}X_1, \ldots, {}^{ii}X_n$ as v.as. inversas das v.as. inversas. Tem-se*

$$Var(X_1, \ldots, X_n) = Var({}^{ii}X_1, \ldots, {}^{ii}X_n) \tag{4.41}$$

$$Corr(X_1 : \ldots : X_n) = Corr({}^{ii}X_1 : \ldots : {}^{ii}X_n). \tag{4.42}$$

Demonstração. Conseqüência imediata da definição de matriz de correlação inversa e da Proposição 4.2.9. □

Agora é apresentada uma relação importante que existe entre as matrizes de coeficientes de regressão da v.as. e das suas inversas.

O ponto importante da proposição a seguir é o fato que as v.as. inversas "desparcializam" os coeficientes de regressão. Mais especificamente, sejam X, Y e Z v.as. unidimensionais e ${}^{i}X, {}^{i}Y, {}^{i}Z$ as respectivas inversas. Considere a regressão de X nas outras v.as., isto é,

$$X = aY + bZ + c({}^{i}X).$$

O coeficiente a é proporcional a $\mathrm{Cov}(X : Y/Z)$. Agora considere

$${}^{i}X = \alpha({}^{i}Y) + \beta({}^{i}Z) + \gamma X.$$

O coeficiente α é proporcional a $\mathrm{Cov}(X : Y)$. Ou seja, os coeficientes da regressão entre as v.as. fornece essencialmente a estrutura de dependência parcializada enquanto os coeficientes de regressão das v.as. inversas fornecem a

estrutura de dependência não parcializada. Neste sentido, as v.as. e suas inversas são duais uma em relação a outra.

Proposição 4.2.10. *Sejam $X_1, \ldots, X_n$ v.as. $d_1, \ldots, d_n$-dimensionais e ${}^iX_1, \ldots, {}^iX_n$ as suas respectivas v.as. inversas definidas em (4.36). Considere as equações de regressão*

$$X_1 = A_2X_2 + \sum_{k=3}^{n} A_kX_k + Var(X_1/X^1)({}^iX_1) \tag{4.43}$$

e

$${}^iX_1 = G_2({}^iX_2) + \sum_{k=3}^{n} G_k({}^iX_k) + Var(X_1)^{-1}(X_1). \tag{4.44}$$

em que a v.a. com índice sobrescrito $X^k, k = 1, \ldots, n$ é o vetor formado por todos as v.as. de $X_1, \ldots, X_n$ exceto X_k. Analogamente para ${}^iX^k$ e ${}^iX^{k,l}$.

Tem-se

$$A_2 = Cov(X_1 : X_2/X^{1,2})\, Var(X_2/X^{1,2})^{-1} \tag{4.45}$$

e

$$G_2 = Var(X_1)^{-1}\, Cov(X_1 : X_2). \tag{4.46}$$

Demonstração. As equações (4.43) e (4.44) são de fato equações de regressão pela bi-ortogonalidade das v.as. e suas inversas, isto é, $\mathrm{Cov}(X_k, {}^iX_l), k \neq l$ e portanto $\mathrm{Var}(X_1/X^1)({}^iX_1)$ é de fato o resíduo em (4.43) e $\mathrm{Var}(X_1)^{-1}(X_1)$ é o resíduo em (4.44).

Pode-se calcular diretamente o coeficiente, porém o seguinte método é mais elucidativo. Tome $d = \sum_{k=1}^{n} d_k$. Defina $D_k = \{\sum_{l=0}^{k-1} d_l + 1, \ldots, \sum_{l=0}^{k} d_l\}, k =$

$1, \ldots, n$, em que $d_0 = 0$. Tem-se, aplicando duas vezes a relação da inversa da matriz particionada obtida no Corolário 4.2.3

$$\begin{aligned}
&[\mathrm{Var}(X_1, \ldots, X_n)^{-1}]_{D_1 \cup D_2}^{D_1 \cup D_2} \\
&= \begin{bmatrix} \mathrm{Var}(X_1/X^{12}) & \mathrm{Cov}(X_1 : X_2/X^{12}) \\ \mathrm{Cov}(X_2 : X_1/X^{12}) & \mathrm{Var}(X_2/X^{12}) \end{bmatrix}^{-1} \\
&= \begin{bmatrix} \mathrm{Var}(X_1/X^1)^{-1} & -\mathrm{Var}(X_1/X^1)^{-1}B \\ \mathrm{Var}(X_2/X^2)^{-1}C & \mathrm{Var}(X_2/X^2)^{-1} \end{bmatrix},
\end{aligned}$$

em que C é o coeficiente de regressão em

$$\bar{R}(X_2/X^{12}) = C\bar{R}(X_1/X^{12}) + \bar{R}(X_2/X^2),$$
$$\bar{R}(X_1/X^{12}) = B\bar{R}(X_2/X^{12}) + \bar{R}(X_1/X^1).$$

Agora pode-se ver que

$$C = \mathrm{Cov}(X_2 : X_1/X^{12})\mathrm{Var}(X_1/X^{12})^{-1}$$
$$B = \mathrm{Cov}(X_1 : X_2/X^{12})\mathrm{Var}(X_2/X^{12})^{-1}.$$

Defina

$$V = \mathrm{diag}(\mathrm{Var}(X_1/X^1), \ldots, \mathrm{Var}(X_n/X^n)).$$

Agora, pela Proposição 4.2.9 e substituindo (4.40) na equação (4.39) tem-se

$$V\mathrm{Var}(X_1, \ldots, X_n)^{-1}X = V({}^iX),$$

em que $X^T = [X_1^T \ \dots \ X_n^T]$ e ${}^iX^T = [{}^iX_1^T \ \dots \ {}^iX_n^T]$. Assim comparando os coeficientes

$$\begin{aligned} A_2 &= -[V\mathrm{Var}(X_1,\dots,X_n)^{-1}]_{D_1}^{D_2} \\ &= \mathrm{Var}(X_1/X^1)\mathrm{Var}(X_1/X^1)^{-1}B \\ &= B. \end{aligned}$$

Agora defina

$$U = \mathrm{diag}(\mathrm{Var}(X_1),\dots,\mathrm{Var}(X_n)).$$

Novamente pela Proposição 4.2.9 e substituindo (4.40) na equação (4.39) tem-se

$$\mathrm{Var}(X_1,\dots,X_n)^{-1}UU^{-1}X = ({}^iX)$$

e portanto

$$U^{-1}X = U^{-1}\mathrm{Var}(X_1,\dots,X_n)({}^iX).$$

Comparando os coeficientes

$$\begin{aligned} G_2 &= -[U^{-1}\mathrm{Var}(X_1,\dots,X_n)]_{D_1}^{D_2} \\ &= \mathrm{Var}(X_1)^{-1}\mathrm{Cov}(X_1 : X_2) \end{aligned}$$

□

As Proposições 4.2.9 e 4.2.10 acima, mostram que as v.as. inversas são v.as cujas dependências internas (entre os componentes univariadas da mesma v.as. multidimensional) foram parcializadas e as dependências externas (entre as v.as.

multidimensionais) foram desparcializadas.

Uma aplicação do conceito desenvolvido nesta seção é a cosntrução de medidas de dependência linear simplesmente substituindo as v.as pelas suas inversas. Por exemplo, pode-se construir a *CQT inversa* denotada iCQT da seguinte forma.

Definição 4.2.9. Sejam $X_1, \ldots, X_n$ v.as. não necessariamente unidimensionais e ${}^iX_1, \ldots, {}^iX_n$ as v.as. inversas. A CQT inversa $\mathrm{iCQT}(X_1 : \ldots : X_n)$ é definida como

$$\mathrm{iCQT}(X_1 : \ldots : X_n) = \mathrm{CQT}({}^iX_1 : \ldots : {}^iX_n). \tag{4.47}$$

No caso de duas v.as. X e Y com inversas ${}^iX$ e ${}^iY$,

$$\mathrm{iCQT}(X : Y) = \mathrm{CQT}({}^iX : {}^iY) = \mathrm{CQT}(X : Y), \tag{4.48}$$

pois ,

$$\begin{aligned}
&\frac{\det \mathrm{Var}({}^iX, {}^iY)^{-1}}{\det \mathrm{Var}({}^iX) \det \mathrm{Var}({}^iY)} \\
&= \frac{\det \mathrm{Var}(X, Y)^{-1}}{\det \mathrm{Var}(X/Y)^{-1} \det \mathrm{Var}(Y/X)^{-1}} \\
&= \frac{\det \mathrm{Var}(X, Y)}{\det \mathrm{Var}(X) \det \mathrm{Var}(Y)}.
\end{aligned}$$

Entretanto, no caso de mais de duas v.as. as medidas são, em geral, diferentes.

As v.as. inversas apresentam propriedades adicionais interessantes que possibilitam melhor interpretação.

Proposição 4.2.11. *Sejam X, Y, Z v.as. não necessariamente univariadas e ${}^iX, {}^iY, {}^iZ$ as suas respectivas v.as. inversas. Tem-se*

$$CQT({}^iX :^i Y) = CQT(X : Y/Z) \tag{4.49}$$

$$CQT(X : Y) = CQT({}^iX :^i Y/{}^iZ). \tag{4.50}$$

Demonstração. Pela Proposição 4.2.9

$$\mathrm{Var}({}^iX, {}^iY, {}^iZ) = \mathrm{Var}(X, Y, Z)^{-1}.$$

Então, pelo Corolário 4.2.3

$$\mathrm{Var}({}^iX, {}^iY) = \mathrm{Var}(X, Y/Z)^{-1}.$$

Agora pela identidade (4.48)

$$\mathrm{CQT}({}^iX :^i Y) = \mathrm{CQT}(X : Y/Z).$$

A identidade (4.50) segue (4.49) e da reflexividade (Corolário 4.2.4). □

As seguintes proposições são úteis para se obter as inversas de algumas medidas de dependências para séries temporais. Os resultados a seguir se distingüem dos anteriores por considerar a dependência entre elementos de dois grupos de v.as., cada uma com as suas v.as. inversas.

Proposição 4.2.12. *Sejam X_1, X_2 v.as. n_1 e n_2 dimensionais e ${}^iX_1, {}^iX_2$ suas inversas, respectivamente. Sejam Y_1, Y_2 v.as. m_1 e m_2 dimensionais e ${}^iY_1, {}^iY_2$ suas inversas, respectivamente. Tem-se*

$$CQT({}^iX_1 :^i Y_1/{}^iY_2) = CQT(X_1 : Y_1/X_2). \tag{4.51}$$

Demonstração. Defina $X^T = [X_1^T \ X_2^T]$, $Y^T = [Y_1^T \ Y_2^T]$ e a matriz

$$H = \mathrm{Cov}(Y : X)\mathrm{Var}(X)$$

tal que

$$HX = Y. \tag{4.52}$$

Defina $W^T = [{}^iX_1^T \ {}^iX_2^T]$ e $Z^T = [{}^iY_1^T \ {}^iY_2^T]$. Observe que ${}^iX$ e ${}^iY$ não são, em geral, iguais a W e Z, respectivamente. Agora, seja A tal que

$$W = AZ.$$

Então,

$$\mathrm{Cov}(HX : W) = \mathrm{Cov}(Y : AZ),$$

o que implica

$$A^* = H. \tag{4.53}$$

Calculando, tem-se

$$A^* = \begin{bmatrix} \mathrm{Cov}(Y_1 : X_1/X_2)\mathrm{Var}(X_1/X_2)^{-1} & \mathrm{Cov}(Y_1 : X_2/X_1)\mathrm{Var}(X_2/X_1)^{-1} \\ \mathrm{Cov}(Y_2 : X_1/X_2)\mathrm{Var}(X_1/X_2)^{-1} & \mathrm{Cov}(Y_2 : X_2/X_1)\mathrm{Var}(X_2/X_1)^{-1} \end{bmatrix}.$$

Comparando os termos

(4.54)

$$\mathrm{Cov}({}^iX_1 :^i Y_1/{}^iY_2)\mathrm{Var}({}^iY_1/{}^iY_2)^{-1} = \left(\mathrm{Cov}(Y_1 : X_1/X_2)\mathrm{Var}(X_1/X_2)^{-1}\right)^*.$$

De forma análoga, defina $P^T = [X_1^T \ Y_2^T]$, $Q^T = [Y_1^T \ X_2^T]$,$U^T = [{}^iX_1^T \ {}^iY_2^T]$, $V^T = [{}^iY_1^T \ {}^iX_2^T]$.Tem-se

$$P = JQ$$

e

$$MU = V.$$

Analogamente a (4.53), tem-se

$$M = J^*,$$

de onde se conclui que

(4.55)
$$\begin{aligned} &\mathrm{Cov}({}^iY_1 :^i X_1/{}^iY_2)\mathrm{Var}({}^iX_1/{}^iY_2)^{-1} \\ &= \left(\mathrm{Cov}(X_1 : Y_1/X_2)\mathrm{Var}(Y_1/X_2)^{-1}\right)^*. \end{aligned}$$

Agora,

(4.56)

$$1 - \mathrm{CQT}({}^{i}X_1 :^{i} Y_1/{}^{i}Y_2)$$

$$= \det(I - \mathrm{Cov}({}^{i}X_1 :^{i} Y_1/{}^{i}Y_2)\mathrm{Var}({}^{i}Y_1/{}^{i}Y_2)^{-1}\mathrm{Cov}({}^{i}Y_1 :^{i} X_1/{}^{i}Y_2)\mathrm{Var}({}^{i}X_1/{}^{i}Y_2)^{-1}).$$

Usando (4.54) e (4.55) e substituindo em (4.56) obtém-se

(4.57)

$$1 - \mathrm{CQT}({}^{i}X_1 :^{i} Y_1/{}^{i}Y_2)$$

$$= \det(I - \mathrm{Var}(X_1/X_2)^{-1}\mathrm{Cov}(X_1 : Y_1/X_2)\mathrm{Var}(Y_1/X_2)^{-1}\mathrm{Cov}(Y_1 : X_1/X_2)).$$

Finalmente,

$$1 - \mathrm{CQT}({}^{i}X_1 :^{i} Y_1/{}^{i}Y_2) = 1 - \mathrm{CQT}(X_1 : Y_1/{}^{i}X_2),$$

o que finaliza a prova. □

Corolário 4.2.5. *Sejam $X_1, \ldots, X_n$ v.as. não necessariamente unidimensionais e ${}^{i}X_1, \ldots, {}^{i}X_n$ suas v.as. inversas, respectivamente. Considere também $Y_1, \ldots, Y_m$ outras v.as. não necessariamente unidimensionais e ${}^{i}Y_1, \ldots, {}^{i}Y_m$ suas respectivas v.as. inversas. Tem-se*

(4.58) $$CQT({}^{i}X_p :^{i} Y_q/{}^{i}Y^q) = CQT(X_p : Y_q/X^p),$$

$p = 1, \ldots, n$ e $q = 1, \ldots, m$, em que ${}^{i}Y^q$ é a v.a. formada por ${}^{i}Y_k$, $k \neq q$. X^p é definido analogamente.

Demonstração. Imediato pela Proposição 4.2.12. □

Uma outra proposição que é necessária para se provar as proposições da próxima seção é bastate semelhante ao anterior.

Proposição 4.2.13. *Sejam X_1, X_2 v.as. n_1 e n_2 dimensionais e ${}^iX_1, {}^iX_2$ suas inversas, respectivamente. Sejam Y_1, Y_2 v.as. m_1 e m_2 dimensionais e ${}^iY_1, {}^iY_2$ suas inversas, respectivamente. Tem-se*

$$CQT({}^iX_1 : \bar{R}({}^iY_1/{}^iY_2)) = CQT(\bar{R}(X_1/X_2) : Y_1). \tag{4.59}$$

Demonstração. A seguinte observação é importante. Sejam X, Y e Z v.as. não necessariamente unidimensionais.

$$\mathrm{Cov}(X : Y/Z) = \mathrm{Cov}(X : \bar{R}(Y/Z)). \tag{4.60}$$

Basta notar que $X = AZ + \bar{R}(X/Z)$ em que $\mathrm{Cov}(Z : \bar{R}(X/Z)) = 0 = \mathrm{Cov}(Z : \bar{R}(Y/Z))$. Logo,

$$\begin{aligned}\mathrm{Cov}(X : Y/Z) &= \mathrm{Cov}(\bar{R}(X/Z) : \bar{R}(Y/Z)) \\ &= \mathrm{Cov}(AZ + \bar{R}(X/Z) : \bar{R}(Y/Z)).\end{aligned}$$

Agora, por (4.54)

(4.61)

$$\mathrm{Cov}({}^iX_1 :^i Y_1/{}^iY_2)\mathrm{Var}({}^iY_1/{}^iY_2)^{-1} = \left(\mathrm{Cov}(Y_1 : X_1/X_2)\mathrm{Var}(X_1/X_2)^{-1}\right)^*.$$

Usando (4.60) e substituindo em (4.61) tem-se

(4.62)

$$\mathrm{Cov}({}^iX_1 : \bar{R}({}^iY_1/{}^iY_2))\mathrm{Var}({}^iY_1/{}^iY_2)^{-1} = \left(\mathrm{Cov}(Y_1 : \bar{R}(X_1/X_2))\mathrm{Var}(X_1/X_2)^{-1}\right)^*.$$

Agora, observe que

$$
\begin{aligned}
&1 - \mathrm{CQT}({}^iX_1 : \bar{R}({}^iY_1/{}^iY_2)) \\
&= \det(I - \mathrm{Cov}({}^iX_1 : \bar{R}({}^iY_1/{}^iY_2))\mathrm{Var}({}^iY_1/{}^iY_2)^{-1}\mathrm{Cov}(\bar{R}({}^iY_1/{}^iY_2) :^i X_1)\mathrm{Var}({}^iX_1)^{-1}) \\
&= \det(I - \mathrm{Cov}({}^iX_1 : \bar{R}({}^iY_1/{}^iY_2))\mathrm{Var}({}^iY_1/{}^iY_2)^{-1} \\
&\mathrm{Var}({}^iY_1/{}^iY_2)\mathrm{Var}({}^iY_1/{}^iY_2)^{-1}\mathrm{Cov}(\bar{R}({}^iY_1/{}^iY_2) :^i X_1)\mathrm{Var}({}^iX_1)^{-1}) \\
&= \det(I - \mathrm{Cov}({}^iX_1 : \bar{R}({}^iY_1/{}^iY_2))\mathrm{Var}({}^iY_1/{}^iY_2)^{-1} \\
&\mathrm{Var}({}^iY_1/{}^iY_2)\mathrm{Var}({}^iY_1/{}^iY_2)^{-1}\mathrm{Cov}(\bar{R}({}^iY_1/{}^iY_2) :^i X_1)\mathrm{Var}(X_1/X_2)).
\end{aligned}
$$

Usando a identidade (4.62) e pelas Proposição 4.2.9 e Corolário 4.2.3, obtém-se

$$
\begin{aligned}
&1 - \mathrm{CQT}({}^iX_1 : \bar{R}({}^iY_1/{}^iY_2)) && (4.63)\\
&= \det(I - \mathrm{Var}(X_1/X_2)^{-1}\mathrm{Cov}(\bar{R}(X_1/X_2) : Y_1)\mathrm{Var}(Y_1)^{-1} && (4.64)\\
&\quad \mathrm{Cov}(Y_1 : \bar{R}(X_1/X_2))\mathrm{Var}(X_1/X_2)^{-1}\mathrm{Var}(X_1/X_2)) \\
&= 1 - \mathrm{CQT}(\bar{R}(X_1/X_2) : Y_1),
\end{aligned}
$$

obtendo-se o resultado desejado. □

Definir as v.as. inversas para infinitas v.as. é um problema bem mais difícil e é estudada no próximo capítulo para o caso de séries estacionárias em que os resultados desta seção se generalizam ipsis literis para os casos em que se pode associar uma representação espectral adequada.

4.3 Conclusão

Neste capítulo foram introduzidas algumas medidas de dependência linear que naturalmente se relacionam com a informação mútua no caso em que as v.as. são gaussianas e em alguns casos relacionados também a v.as. não necessariamente gaussianas, mas com cópulas gaussianas. A principal medida introduzida foi a CQT que permite inferir a relação linear de mais de duas v.as. multivariadas. Foram estudados dois métodos canônicos de construção de medidas de dependência a partir de outras medidas de dependência linear: (a) parcialização e (b) inversão. Ambas as construções apresentam uma relação intrínseca de tal sorte que no caso em que se deseja estudar a relação entre as v.as. unidimensionais a parcialização pode ser obtida da inversão e vice-versa. No caso geral, quando se deseja estudar a dependência entre mais de duas v.as. a relação entre os métodos é menos simples, porém ainda mantém uma relação importante.

Uma vez que (a) no Capítulo 3 foram revisados e desenvolvidos alguns conceitos de Teoria da Informação e dependências em geral e (b) neste capítulo desenvolveu-se conceitos e métodos de construção de medidas a partir de outras, o capítulo seguinte se concentra nas suas aplicações para séries temporais com o objetivo final de reinterpretar e generalizar algumas medidas existentes na literatura, em particular a coerência parcial direcionada.

CAPÍTULO 5

Séries temporais - um resumo

"Absolute, true, and mathematical time, in and of itself and of its own nature, without reference to anything external, flows uniformly and by another name is called duration. Relative, apparent, and common time is any sensible and external measure (precise or imprecise) of duration by means of motion; such a measure - for example, an hour, a day, a month, a year - is commonly used instead of true time." (Isaac Newton, Principia, 1726)

Neste capítulo faz-se um sumário de alguns fatos sobre séries temporais estacionárias de segunda ordem que são utilizados para o desenvolvimento dos próximos capítulos. Os resultados não são demonstrados pois são bem cohecidas na literatura. As referências padrão para este capítulo são Rozanov (1967); Hannan (1970); Hannan e Deistler (1988); Brillinger (1981); Lütkepohl (1993) em ordem de maior para menor sofisticação matemática e menor para maior ênfase em aplicação em dados. As referências para algumas questões de análise harmônica

para séries temporais multivariadas são Wiener e Masani (1957, 1958); Masani (1960); Helson e Lowdenslager (1958, 1962).

As séries temporais n-variadas X consideradas nesta tese são seqüências de v.as. reais n-variadas $\{\ldots, X(-1), X(0), X(1), \ldots\}$ infinitas bilaterais com índices no conjunto dos inteiros. Os índices entre parênteses são denominados tempo. Se Y é uma outra série m-variada, a série $(n+m)$-variada W tal que $W_t^T = [X(t)^T\ Y(t)^T], \forall t \in \mathcal{Z}$ pode ser indicada como $W^T = [X^T\ Y^T]$.

Seja X uma série n-variada, a seguinte notação é util

$$\{X\}_s^t = [X(s)^T\ \ldots\ X(t)^T]^T,\ \ s \leq t.$$

O espaço gerado pela série X é o subespaço de L^2 gerado pelos elementos da série conforme a Definição 4.1.2.

$$\Gamma_x(t,\tau) = \mathrm{Cov}(X(t), X(t-\tau)),\ \forall t, \tau \in \mathbb{Z},$$

em que t é a variável de tempo global e τ é de tempo local ou de atraso.

Uma série temporal X é dita estacionária em senso amplo ou de segunda ordem se a função de autocovariância for finita para todos os valores e depender somente de atraso, ou seja, $\Gamma_x(t,\tau) = \Gamma_x(0,\tau),\ \forall t, \tau \in \mathbb{Z}$. A função de autocovariância nesse caso será escrita simplesmente como $\Gamma_x(\tau)$.

Afirmar que n-séries $X_1, \ldots, X_n$, cada uma não necessariamente univariada, são conjuntamente estacionárias, equivale a dizer que a série $W^T = [X_1^T\ \ldots\ X_n^T]$ é estacionária.

Pode-se demonstrar que um processo é estacionário se e somente se pode ser escrito como a transformada de Fourier-Stieltjes de um processo aleatório com

incrementos ortogonais, ou seja,

$$X(t) = \int_{-\pi}^{\pi} e^{it\lambda} dZ_x(\lambda), \tag{5.1}$$

em que, dados Λ e $\Lambda' \subset [-\pi, \pi)$, $\Lambda \cap \Lambda' = \emptyset$, tem-se $\mathrm{Cov}(Z_x(\Lambda), Z_x(\Lambda')) = 0$. A representação integral (5.1) é denominada representação espectral do processo X e a igualdade é válida em média quadrática.

É interessante ressaltar que uma grande família de processos denominados processos harmonizáveis, que não são necessariamente estacionários, pode ser representada como a transformada de Fourier-Stieltjes de processos aleatórios, ou, de forma mais geral, de medidas aleatórias (medidas a valores num espaço de Hilbert), ou seja, apresentam a representação integral (5.1) em que no caso geral os incrementos não são necessariamente ortogonais. Embora não explorada nesta tese, alguns dos resultados obtidos para o caso estacionário se generalizam para esta família de processos pelo menos formalmente.

Uma conseqüência importante da existência da representação espectral (5.1) para as séries estacionárias é a possibilidade de se escrever qualquer elemento ξ do espaço gerado pela série n-dimensional X como

$$\xi = \int_{-\pi}^{\pi} \phi(\lambda) dZ_x(\lambda),$$

em que ϕ é uma função matricial de posto completo com dimensão $m \times n$ em que $m \leq n$.

A função ϕ é denominada filtro que gera ξ a partir de X. Em geral, quando X é estacionária, associa-se, não somente uma v.a., mas uma série ξ denominada

série filtrada de X com filtro ϕ em que

$$\xi(t) = \int_{-\pi}^{\pi} e^{it\lambda}\phi(\lambda)dZ_x(\lambda).$$

Dado um processo estacionário X, denomina-se função de distribuição espectral de X a função $F_x(\Lambda) = \mathrm{Var}(Z_x(\Lambda)),\ \Lambda \subset [-\pi, \pi)$. Tem-se:

$$\mathrm{Var}(X(t)) = \mathrm{Var}\left(\int_{-\pi}^{\pi} e^{it\lambda}dZ_x(\lambda)\right) = \int_{-\pi}^{\pi} dF_x(\lambda).$$

De forma mais geral,

$$\begin{aligned}\mathrm{Cov}(X(t), X(s)) &= \mathrm{Cov}\left(\int_{-\pi}^{\pi} e^{it\lambda}dZ_x(\lambda), \int_{-\pi}^{\pi} e^{is\lambda}dZ_x(\lambda)\right)\\ &= \int_{-\pi}^{\pi} e^{i(t-s)\lambda}dF_x(\lambda),\end{aligned}$$

ou seja,

$$\Gamma_x(\tau) = \int_{-\pi}^{\pi} e^{i\tau\lambda}dF_x(\lambda),$$

o que justifica denotar $\mathrm{Var}(dZ_x(\lambda)) = dF_x(\lambda)$.

Diversas propriedades do processo X podem ser descritas pelas condições sobre F. Neste texto, serão considerados os processos estacionários n-dimensionais X tais que as funções de distribuição espectral F_x sejam absolutamente contínuas em relação à medida de Lebesgue no intervalo $[-\pi, \pi)$. Nesse caso, existe, pelo teorema de Radon-Nikodym, uma função densidade espectral $f_x(\lambda) = \frac{dF(\lambda)}{d\lambda}$ de $[-\pi, \pi)$ em $\mathbb{R}^{n\times n}$. No caso em que o processo é real, que é o caso dos processos desta tese, $f_x(\lambda) = f_x(\lambda)^*$ e $f_x(-\lambda) = \overline{f_x(\lambda)}$.

Sejam as séries estacionárias n e m-variadas X e Y e a série $W^T = [X^T\ Y^T]$. A densidade espectral de W, f_w, denotada também por $f_{(xy)}$ pode ser particionada da seguinte forma:

$$f_{(xy)}(\lambda) = \begin{bmatrix} f_x(\lambda) & f_{xy}(\lambda) \\ f_{yx}(\lambda) & f_y(\lambda) \end{bmatrix}, \ \forall \lambda \in [-\pi, \pi),$$

em que f_x e f_y são as densidades espectrais de X e Y, respectivamente, e f_{xy} é a densidade espectral cruzada de X e Y. Uma vez que se assume que as séries sejam reais, vale a relação $f_{yx}(\lambda) = f_{xy}(\lambda)^*$.

Será assumida ainda ao longo do capítulo que cada função densidade f tenha uma inversa f^{-1} e que ambas apresentem autovalores limitados para todas as freqüências, ou seja,

Condição 5.0.1 (Condição de limitação)**.** *Seja X um processo estacionário n-variado com matriz de densidade espectral f_x. Diz-se que X satisfaz a condição de limitação se*

$$c_1 I_n \leq f_x(\lambda) \leq c_2 I_n, \ \ 0 < c_1 \leq c_2 < \infty, \tag{5.2}$$

para quase todo $\lambda \in [-\pi, \pi)$.

Essa última condição é suficiente para a validade, na maioria dos casos, dos cálculos formais que serão apresentados no decorrer do texto. Caso não haja ambigüidade, essa condição será sempre assumida. Quando forem necessárias outras condições, estas serão apresentadas explicitamente. Em geral, a condição de limitação não é necessária, porém é suficientemente geral para os propósitos desta tese. De fato, a condição de limitação é uma exigência natural para a validade de muitos resultados estatísticos (Taniguchi e Kakizawa, 2000; Cheng

e Pourahmadi, 1992).

O passado de $X(t)$ denotado X^{t-} é o subespaço gerado pelas v.as. $X(s),\ s < t$. O futuro de $X(t)$ denotado X^{t+} é o subespaço garado pelas v.a. $X(s),\ s > t$.

Define-se a inovação ou processo fundamental de uma série n-variada X no tempo t como sendo os resíduos $\xi(t)$ da projeção ortogonal de $X(t)$ termo a termo em seu passado X^{t-}. Dessa forma, a inovação $\xi(t)$ no tempo t é uma v.a. ortogonal ao passado X^{t-} de tal forma que tem-se a decomposição única $X(t) = \bar{E}(X(t)/X^{t-}) + \xi(t)$. Intuitivamente, $\xi(t) = [\xi_1(t), \ldots, \xi_n(t)]^T$ é a parte de $X(t)$ que não pôde ser explicada pelo seu passado. Tem-se que as inovações são mutualmente ortogonais para tempos distintos, ou seja, $\mathrm{Cov}(\xi(t), \xi(s)) = 0, \forall t, s \in \mathbb{Z}$ e $t \neq s$. A condição de limitação (5.2) é suficiente para garantir que as inovações geram o mesmo espaço que o original, isto é, $X^{t-} = \xi^{t-},\ \forall t \in \mathcal{Z}$.

Uma série n-variada X satisfazendo a condição de limitação (5.2) apresenta a seguinte representação média móvel (MM) causal em termos da sua inovação ξ:

$$X(t) = \sum_{k=0}^{\infty} H(k)\xi(t-k), \tag{5.3}$$

em que $H(k)$ para todo $k \geq 0$ é uma matriz de dimensões $n \times n$ e $H(0) = I$. Caso $H(k) = 0$ para $k > q$ e $H(k) \neq 0$ para $k = 0$ e $k = q$, a representação MM (5.3) é dita de ordem q. O processo X apresenta também uma representação MM anti-causal

$$X(t) = \sum_{k=0}^{\infty} G(k)\upsilon(t+k), \tag{5.4}$$

em que $\upsilon(t)$ é o resíduo da projeção ortogonal de $X(t)$ em X^{t+} e $\mathrm{Cov}(\upsilon(t), \upsilon(s)) = 0, \forall t, s \in \mathbb{Z}$. A relação entre as representações causal e anti-causal no caso multivariado não é simples e foi caracterizada por Soltani e Mohammadpour (Soltani e Mohammadpour, 2006).

Dada uma representação MM para um processo estacionário X, o símbolo MM $\tilde{H}(\lambda)$ de X é definida como sendo uma função matricial definida em $[-\pi, \pi)$ tal que

$$\tilde{H}(\lambda) = \sum_{k=0}^{\infty} H(k) e^{-ik\lambda},$$

em que $H(k), k \geq 0$ são as matrizes de coeficientes da representação AR de X. Na literatura, principalmente de engenharia, $\tilde{H}$ é denominado também função de transferência de X.

Dado um processo estacionário n-variado X satisfazendo a condição de limitação pode-se representá-lo na forma

$$X(t) = \sum_{k=0}^{\infty} E(k)\eta(t-k), \tag{5.5}$$

em que η é um processo não correlacionado que não seja o processo de inovação ξ e nem um produto de ξ com uma matriz unitária. Nesse caso tem-se a seguinte importante propriedade (Rozanov (1967)):

$$\mathrm{Var}(\xi(t)) > \mathrm{Var}(\eta(t)),$$

que é a notação para indicar que a diferença $\mathrm{Var}(\xi(t)) - \mathrm{Var}(\eta(t))$ é positiva definida.

Teorema 5.0.1 (Szegö)**.** *Sob a condição de limitação, a seguinte identidade é válida para uma série estacionária n-dimensional X com função densidade espectral f_x e inovação ξ:*

$$\det\ Var(\xi(t)) = (2\pi)^n \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\det f_x(\lambda)d\lambda\right\}, \tag{5.6}$$

para todo $t \in \mathbb{Z}$ e tem-se $\det\ Var(\xi(t)) > 0$.

Doravante a (5.6) será denominada identidade de Szegö. Este resultado é crucial para relacionar as medidas nos domínios do tempo e da freqüência.

Pode-se mostrar, sob a condição de limitação, a existência da seguinte representação autorregressiva (AR) convergente em média quadrática do processo X:

$$X(t) = \sum_{k=1}^{\infty} A(k)X(t-k) + \xi(t), \tag{5.7}$$

em que ξ é o processo de inovação e $A(k), k \geq 1$ são matrizes de dimensões $n \times n$. Caso $A(k) = 0$ para $k > p$ e $A(k) \neq 0$ para $k = p$ a representação AR (5.7) é dita de ordem p. Freqüentemente a representação (5.7) é denominada representação AR reduzida. Há uma segunda forma AR freqüentemente utilizada denominada forma AR primária escrita como:

$$X(t) = \sum_{k=0}^{\infty} A^{\dagger}(k)X(t-k) + \xi^{\dagger}(t), \tag{5.8}$$

em que $A^{\dagger}(0)$ é triangular inferior, $A^{\dagger}_{pp}(0) = 0$ para $1 \leq p \leq n$ e $\xi^{\dagger}$ é o processo de inovação ortogonalizado multiplicando-se a inversa do fator de Cholesky normalizado na diagonal principal, ou seja, $\mathrm{Cov}(\xi(t)) = LL^T$ em que $M =$

$L(\text{diag}(L))^{-1}$ é triagular inferior e $\xi^{\dagger} = M^{-1}\xi$. Tem-se $\text{Cov}(\xi^{\dagger}(t)) = \text{diag}(L)^2$. Um cálculo simples mostra que $A^{\dagger}(0) = I - M^{-1}$.

Dada uma representação AR para um processo estacionário X, o símbolo AR $\tilde{A}(\lambda)$ de X é definida como sendo uma função matricial definida em $[-\pi, \pi)$ tal que

$$\tilde{A}(\lambda) = I - \sum_{k=1}^{\infty} A(k)e^{-ik\lambda},$$

em que $A(k), k \geq 1$ são as matrizes de coeficientes da representação AR de X.

Uma outra propriedade importante garantida também pela condição de limitação é a existência do processo de interpolação ou inovação bilateral W definida como a família de v.a. formada pelos resíduos da projeção ortogonal de $X(t)$ em seu passado X^{t-} e seu futuro X^{t+}, ou seja,

$$X(t) = \sum_{k=1}^{\infty} B(k)X(t-k) + \sum_{k=1}^{\infty} F(l)X(t+l) + W(t),$$

em que $W(t)$ é ortogonal ao subespaço expandido pelos elementos de X^{t-} e X^{t+}.

Definição 5.0.1. O processo ${}^{i}X$ obtido pela nomalização de $W(t)$ por sua variância $\text{Var}(W(0))$, tal que, ${}^{i}X(t) = \text{Var}(W(0))^{-1}W(t)$ é denominado processo inverso associado a X.

A propriedade fundamental do processo inverso é que a matriz de densidade espectral do processo inverso é exatamente o inverso da densidade espectral do processo original, isto é,

$$f_w(\lambda) = f_x(\lambda)^{-1},\ \lambda \in [-\pi, \pi).$$

Uma conseqüência imediata é a seguinte propriedade: dado o processo esta-

cionário X satisfazendo a condição (5.2) com representação AR (5.7) o processo inverso ${}^{i}X$ associado tem representação MM dada por

$$ {}^{i}X(t) = \sum_{k=0}^{\infty} A(k)^{T} \eta(t+k), $$

em que o processo η está relacionado à inovação de X por $\eta(t) = \mathrm{Var}(\xi(0))^{-1}\xi(t)$. Assim, existe uma relação direta entre a representação AR de X e a representação MM de ${}^{i}X$. Em particular, sabe-se que se o processo X apresenta representação AR de ordem p, então o processo inverso ${}^{i}X$ apresenta representação MM de ordem p anti-causal cujas matrizes de coeficientes são as transpostas daquelas da representação AR de X.

A série temporal n-dimensional X será denominada gaussiana se a distribuição conjunda de um número finito de elementos de X for gaussiana. Mais explicitamente, X é uma série temporal gaussiana se dado $p \in \mathbb{N}_{+}$ e $t_k \in \mathbb{Z}, 1 \leq k \leq p$, a probabilidade conjunta de $\{X(t_1), \ldots, X(t_p)\}$ apresentar distribuição gaussiana multivariada, eventualmente degenerada, isto é, $\det \mathrm{Var}(X(t_1), \ldots, X(t_p)) = 0$. Neste texto não serão considerados os casos degenerados.

As séries temporais gaussianas são o protótipo para se estudar séries estacionárias em senso amplo, uma vez que as séries gaussianas estacionárias em senso estrito são séries estacionárias em senso amplo. Nota-se também que dada uma série estacionária em senso amplo, pelo Teorema 4.1.5, é sempre possível associar uma série gaussiana com a mesma estrutura de variância/covariância.

CAPÍTULO 6

Fluxo de informação ou causalidade - observações

"It is true that the law of causality cannot be demonstrated any more than it can be logically refuted: it is neither correct nor incorrect; it is a heuristic principle; it points the way, and in my opinion it is the most valuable pointer that we posses in order to find a path through the confucions of events, and in order to know in what direction scientific investigation must proceed so that it shall reach useful results." (Max Plank, 1936)

"The concept cause, as it occurs in the works of most philosophers, is one which is apparently not used in any advanced science. But the concepts that are used have been developed from primitive concept (which is that prevalent among philosophers), and the primitive concept, as I shall try to show, still has importance as the source of approximate generalisations and pre-scientific inductions, and as a concept which is valid when suitably limited."(Bertrand Russel,

1948)

A causalidade sempre foi intimamente relacionada à prática e filosofia científica e sua discussão envolve controvérsias.

Não é o objetivo desta tese discutir aspectos filosóficos profundos sobre o conceito de causalidade, porém é inevitável que se faça algumas considerações simples, com certo grau de subjetividade, que motivem a introdução de definições de medidas de dependência para o qual é possível associar a idéia de direção no tempo, inspirando-se nas idéias de causalidade ou fluxo de informação. Aqui são apresentados dois exemplos utilizando modelos lineares gaussianos em que o conceitos de dependência direcionada é discutida. A apresentação nesta seção é informal e as demonstrações e definições precisas são feitas nas Seções 7.1, 7.3 e 7.3 deste capítulo.

A literatura sobre métodos de inferência de causalidade e/ou fluxo de informação é bastante ampla (veja Pearl (2000) para uma discussão da literatura) e é estudada sobre diferentes nomes: redes bayesianas, modelos gráficos, dependências multivariadas, modelos de intervenção e outros. Em geral, é difícil classificar as diferentes propostas por envolverem considerações filosóficas e aspectos técnicos/metodológicos díspares. Em particular, é comum, principalmente na literatura estatística, biológica e de inteligência artificial discutir as definições de causalidade e fluxo de informação sem se considerar explicitamente o papel do tempo, enquanto na literatura física e econométrica o parâmetro tempo freqüentemente tem um papel mais explícito, aparentemente.

Para o objetivo desta tese, o parâmetro tempo é importante e permite que se dividam os processos estocásticos em passado, presente e futuro, uma vez que se fixe um tempo t de referência, sendo que o futuro é indicado pelos índices $s > t$,

o passado pelos índices $s < t$ e o presente por $s = t$, por convenção. Também assume-se que o sentido do tempo seja do passado para o futuro. A palavra causalidade é associada a quantidades que relacionam o passado ao presente ou futuro, sendo que a "causa" ou a origem é sempre associado ao passado e o "efeito" ou chegada é sempre associado ao presente ou futuro.

6.0.1 Modelo 1

O Modelo 1 apresentado abaixo serve para motivar e discutir algumas definições de causalidade de Granger.

Exemplo 6.0.1 (Modelo 1)**.** *Sejam X e Y séries univariadas conjuntamente estacionárias e gaussianas que satisfazem a condição de limitação com representação AR*

$$\begin{bmatrix} X(t) \\ Y(t) \end{bmatrix} = \sum_{k=1}^{\infty} \begin{bmatrix} A_{xx}(k) & A_{xy}(k) \\ A_{yx}(k) & A_{yy}(k) \end{bmatrix} \begin{bmatrix} X(t-k) \\ Y(t-k) \end{bmatrix} + \begin{bmatrix} \xi_x(t) \\ \xi_y(t) \end{bmatrix} \tag{6.1}$$

e representação MM associada

$$\begin{bmatrix} X(t) \\ Y(t) \end{bmatrix} = \sum_{k=0}^{\infty} \begin{bmatrix} H_{xx}(k) & H_{xy}(k) \\ H_{yx}(k) & H_{yy}(k) \end{bmatrix} \begin{bmatrix} \xi_x(t) \\ \xi_y(t) \end{bmatrix}. \tag{6.2}$$

Considere representação AR univariada

$$X(t) = \sum_{k=1}^{\infty} \alpha(k) X(t-k) + \eta_x(t). \tag{6.3}$$

Pode-se ainda projetar $X(t)$ em Y^{t-} e obter

$$X(t) = \sum_{k=1}^{\infty} \beta(k) Y(t-k) + \epsilon_x(t). \tag{6.4}$$

Note que, uma vez que X e Y satisfazem a condição de limitação, X isoladamente também o satisfaz.

Suponha que se observa um processo físico em que é realizada a mensuração em tempo discreto e que tal mensuração seja suficiente para descrever todo o sistema de interesse. Em particular, suponha que o sistema esteja isolado. Assuma ainda que o processo abservado possa ser representado perfeitamente pelo Modelo 1. A questão é: o que é uma definição razoável para causalidade e medida de causalidade? A distinção entre definição de causalidade e de me-

dida de causalidade é necessária, pois a causalidade refere-se à existência ou não de determinadas condições e não é necessário que se defina o grau de causalidade, este último que se refere às medidas de causalidade. É interessante notar que diferentes medidas de causalidade assumem mesmo valor quando não há causalidade como é discutido a seguir.

Para se medir a causalidade, parece razoável perguntar se a probabilidade de observar $X(t) \in A \subset \mathbb{R}$ condicionado em X^{t-} é maior ou menor que a probabilidade de $X(t) \in A$ condicionada em X^{t-} e Y^{t-}, em outras palavras, se o passado de Y ajudar na predição de $X(t)$ espera-se que $p(X(t) = x(t)/X^{t-}, Y^{t-})$ seja maior que $p(X(t) = x(t)/X^{t-})$ em algum sentido. Uma primeira tentativa seria estudar a quantidade

$$\lim_{j\to\infty} E(p(X(t)/X(t-1), Y(t-1), \ldots, X(t-j), Y(t-j)) - p(X(t)/X(t-1), \ldots, X(t-j))),$$

em que a esperança é em relação a todas as v.as. envolvidas. No entanto, a quantidade acima é sempre nula, pois a probabilidade soma um em cada um dos termos dentro da esperança. Assim, a média da diferença das densidades de probabilidades condicionais não pode ser utilizada como critério para decidir se existe ou não causalidade. Do ponto de vista da Teoria da Informação, o problema está na escala que deve ser logarítmica, ou seja, deve se utilizar como medida de causalidade a quantidade

$$\lim_{j\to\infty} E(\log p(X(t)/X(t-1), Y(t-1), \ldots, X(t-j), Y(t-j)) - \log p(X(t)/X(t-1), \ldots, X(t-j))),$$

ou equivalentemente

$$\lim_{j\to\infty}(\mathrm{H}(X(t)/X(t{-}1),\ldots,X(t{-}j)){-}\mathrm{H}(X(t)/X(t{-}1),Y(t{-}1),\ldots,X(t{-}j),Y(t{-}j))).$$

Note que se manteve a idéia inicial de comparar as probabilidades condicionadas, porém agora utilizando uma nova escala.

Usando a identidade (3.24), esta nova quantidade pode ser escrita como

$$\begin{aligned}\lim_{j\to\infty}&(\mathrm{H}(X(t)/X(t-1),\ldots,X(t-j))\\ &-\mathrm{H}(X(t)/X(t-1),Y(t-1),\ldots,X(t-j),Y(t-j)))\\ &=\lim_{j\to\infty}\mathrm{IM}(X(t):Y(t-1),\ldots,Y(t-j)/X(t-1),\ldots,X(t-j)),\end{aligned}$$

e, portanto, assume apenas valores não negativos e é zero se e somente se $X(t)$ e Y^{t-} forem condicionalmente independentes dado X^{t-}, o que é razoável para uma definição de medida de causalidade. Esta quantidade é denominada *medida de causalidade de Granger*.

Utilizando argumentos análogos, uma outra quantidade que pode ser proposta é

$$\begin{aligned}\lim_{j\to\infty}&(\mathrm{H}(X(t))-\mathrm{H}(X(t)/Y(t-1),\ldots,Y(t-j)))\\ &=\lim_{j\to\infty}\mathrm{IM}(X(t):Y(t-1),\ldots,Y(t-j)).\end{aligned}$$

Esta última quantidade essencialmente mede a dependência de Y^{t-} e $X(t)$ sem se importar com X^{t-}.

Dada estas considerações, as seguintes condições podem ser utilizadas para verificar a existência de fluxo de informação de Y para X:

1. $A_{xy}(k) \neq 0$ para algum $k \geq 1$.

2. $\mathrm{Var}(\xi_x(t)) \neq \mathrm{Var}(\eta_x(t))$.

3. $\lim_{n\to\infty} \mathrm{IM}(X(t) : \{Y\}_{(t-1)}^{(t-n)} / \{X\}_{(t-1)}^{(t-n)}) = \mathrm{IM}(X(t) : Y^{t-} / X^{t-}) \neq 0$.

4. $\beta(k) \neq 0$ para algum $k \geq 1$.

5. $\mathrm{Var}(\epsilon_x(t)) \neq \mathrm{Var}(X(t))$.

6. $\lim_{n\to\infty} \mathrm{IM}(X(t) : \{Y\}_{(t-1)}^{(t-n)}) = \mathrm{IM}(X(t) : Y^{t-}) \neq 0$.

A Condição 1 é uma escolha natural se o modelo (6.3) for interpretado em termos de regressão e é prática comum na comunidade estatística. Na comunidade de séries temporais, a Condição 1 é conhecida como condição de existência de *causalidade de Granger* de Y para X (Lütkepohl, 1993).

A Condição 2 é baseada na seguinte interpretação. Se Y de fato envia informação nova para X que não esteja presente no passado de X, o erro que se comete em se prever $X(t)$ usando X^{t-} e Y^{t-} deveria ser menor que o erro que se comete quando se utiliza somente X^{t-} para se prever $X(t)$. Esta condição também é conhecida como condição de existência de causalidade de Granger de Y para X na literatura de econometria (Lütkepohl, 1993) e sabe-se que as Condições 1 e 2 são equivalentes no caso bivariado e mais geralmente a equivalência é válida mesmo para o caso de X e Y não serem univariados (Lütkepohl, 1993).

A idéia da Condição 3 é essencialmente a mesma da Condição 2, porém, em vez de utilizar a noção de erro de predição foi utilizada a noção de informação

em comum, ou seja, se Y envia informação para X distinta daquela que já estava contida no passado de X a informação mútua de $X(t)$ e Y^{t-} dado X^{t-} deve ser diferente de zero. Em outras palavras, se $X(t)$ e Y^{t-} não forem condicionalmente independentes dado X^{t-}, Y^{t-} está enviando nova informação para $X(t)$. No caso estacionário gaussiano, que é o caso considerado, o limite existe e é dado por (Proposição 7.2.2):

$$\mathrm{IM}(X(t) : Y^{t-}/X^{t-}) = -\frac{1}{2}\log\frac{\mathrm{Var}(\xi_x(t))}{\mathrm{Var}(\eta(t))},$$

de onde se conclui que as condições (2) e (3) são equivalentes. A equivalência da condição (1) sai como corolário do fato de a representação AR ser a única que minimiza o erro quadrático de predição. Assim, se $\mathrm{Var}(\xi_x(t)) = \mathrm{Var}(\eta_x(t))$, a primeira linha de 6.1 é igual a 6.3 e portanto $A_{xy}(k) = 0,\ k \geq 1$.

A condição 4 é diferente das anteriores, pois considera-se que existe informação em comum entre o passado de Y e o presente de X se existir alguma correlação entre $X(t)$ e Y^{t-} mesmo que a origem da correlação seja a parte de X^{t-} que foi transmitida para Y^{t-}.

As condições 5 e 6 são equivalentes à condição 4, porém parafraseando em termos da variância do resíduo de predição e em termos de informação mútua. Pode-se mostrar, usando a Proposição 7.1.1, que

$$\mathrm{IM}(X(t) : Y^{t-}) = -\frac{1}{2}\log\frac{\mathrm{Var}(\epsilon_x(t))}{\mathrm{Var}(X(t))}. \tag{6.5}$$

Para entender intuitivamente a diferença entre as condições é interessante se fazer alguns cálculos formais.

Formalmente,

$$A_{xy}(1) = \frac{\mathrm{Cov}(X(t) : Y(t-1)/X^{t-}, Y^{(t-1)-})}{\mathrm{Var}(Y(t-1)/X^{t-}, Y^{(t-1)-})}.$$

Usando a Proposição 4.2.4 e sem se preocupar com o fato de envolver matrizes de tamanho infinito[1] tem-se

$$\begin{aligned}
&\mathrm{Cov}(X(t) : Y(t-1)/X^{t-}, Y^{(t-1)-}) \\
&= \mathrm{Cov}(X(t) : Y(t-1)/Y^{(t-1)-}) \\
&- \mathrm{Cov}(X(t) : X^{t-}/Y^{(t-1)-})\mathrm{Var}(X^{t-}/Y^{(t-1)-})^{-1}\mathrm{Cov}(X^{t-} : Y(t-1)/Y^{(t-1)-}).
\end{aligned}$$

Por outro lado

$$\beta(1) = \frac{\mathrm{Cov}(X(t) : Y(t-1)/Y^{(t-1)-})}{\mathrm{Var}(Y(t-1)/Y^{(t-1)-})}.$$

Agora, supondo que os cálculos formais sejam válidos, pode-se observar que $\beta(1) = 0$ não implica em geral que $A_{xy}(1) = 0$ e vice-versa. A razão disto é que eventualmente tudo que o passado de Y tem em comum com $X(t)$ já pode estar contido no passado do próprio X. Posto desta forma, fica claro que a condição (6) não é adequada como medida de causalidade ou fluxo de informação. Em outras palavras, suponha que de fato existe fluxo de informação de Y para X, porém não há fluxo de informação de X para Y. Neste caso, a quantidade

$$\mathrm{IM}(Y(t) : X^{t-}) \tag{6.6}$$

[1] Pode-se justificar rigorosamente os cálculos utilizando a representação espectral dos processos, porém isso acrescentaria a introdução de aspectos que não são necessários para o objetivo da tese e assim foi evitado

não é nulo, pois, de fato, X^{t-} apresenta informação em comum com Y^{t-} que por sua vez pode apresentar informação em comum com $Y(t)$ e, neste caso, $Y(t)$ e X^{t-} não são independentes e portanto (6.6) não é nulo, o que não é razoável para uma medida de fluxo de infomação. Já a quantidade

$$\mathrm{IM}(Y(t) : X^{t-}/Y^{t-}) \tag{6.7}$$

é nula, pois a fonte de informação em comum entre $Y(t)$ e X^{t-} neste caso é somente Y^{t-} cuja contribuição é totalmente subtraída. Assim, a quantidade (6.7) parece ser mais adequada e explicita a importância da representação AR (6.1) para o Modelo 1 e justifica a prática na comunidade de séries temporais de se testar a nulidade dos coeficientes que relacionam as diferentes séries no modelo autorregressivo multivariado. É interessante notar que na literatura a quantidade (6.7) tem surgido e ressurgido em casos específicos com diferentes nomes em diferentes disciplinas como em Física (Schreiber, 2000; Matsumoto e Tsuda, 1988), Estatística (Geweke, 1982, 1984), Engenharia (Kamitake et al., 2008; Caines e Chan, 1975) e Teoria da Informação Massey e Massey (2005); Marko (1973). As condições (4), (5) e (6), de maneira geral, são associadas à falácia:

"Post hoc ergo propter hoc" (autor desconhecido)

Depois disto, portanto devido a isto

Vale ressaltar que o Modelo 1 considerado é bivariado e embora seja o modelo padrão para se estudar definições de causalidade e fluxo de informação, é um modelo bastante específico que não apresenta dificuldades que podem aparecer no caso multivariado geral. Uma destas dificuldades é estudada no Modelo 2.

Contudo, antes de verifcar o caso multivariado, uma outra medida de fluxo de informação é introduzida.

A medida proposta a seguir é baseada em idéias de identificação de sistemas em que o estudo de sistemas com retroalimentação faz parte da teoria. O ponto principal é a interpretação da seguinte equação:

$$Y(t) = \sum_{k=1}^{\infty} B_{yx}(k)Y(t-k) + \sum_{k=0}^{\infty} B_{yy}(k)X(t-k) + \zeta_y(k).$$

Observe que $\zeta_y(t)$ na equação acima é o resíduo de regressão de $Y(t)$ no passado de Y e no presente e passado de X. Ou seja, é a parte de $Y(t)$ que é realmente nova e que não é devido o passado de Y e nem do presente e passado de X. A série ζ é conhecida como inovação ortogonalizada e pode-se mostrar que $\zeta_y(t) = \bar{R}(\xi_y(t)/\xi_x(t))$. Agora, pode-se perguntar se para medir o fluxo de informação de Y para X não seria mais adequado que se medisse quanta informação o passado de ζ_y tem em comum com $X(t)$ que já não esteja contida no passado de X, uma vez que ζ_y representa a parte de Y que é realmente dele. Assim, a seguinte medida pode ser introduzida:

$$\mathrm{IM}(X(t) : \zeta_y^{t-}/X^{t-}). \tag{6.8}$$

No contexto de processos estacionários de segunda ordem, sem utilizar a terminologia da Teoria da Informação, a quantidade (6.8) foi introduzida por Hosoya (1991) e é denominada aqui medida de causalidade de Hosoya. De fato, Hosoya definiu (6.8) utilizando a variância dos resíduos da regressão, cujo correspondente para o caso gaussiano em termos da Teoria da Informação é a expressão (6.8). A medida proposta por Hosoya apresenta uma série de pro-

priedades interessantes, sendo uma delas o fato de se anular se e somente se a medida de causalidade de Granger é nula, sendo assim, a medida de Hosoya é uma outra candidata possível para se verificar a existência ou não de causalidade de Granger.

É interessante que a importância da inovação para se verificar a causalidade de Granger já havia sido sugerida nos trabalhos de Sims (1972) e Caines e Chan (1975) em que é demonstrado que, num sistema com duas séries estacionárias de segunda ordem, a causalidade de Granger pode ser verificada tanto pela nulidade dos coeficientes da representação AR ou MM. Mais especificamente, no Modelo 1, $\mathrm{IM}(X(t) : Y^{t-}/X^{t-}) = 0$ se e somente se $A_{xy}(k) = 0, k \geq 1$ que equivale no Modelo 1 à condição $H_{xy}(k) = 0, k \geq 0$. Esta última condição de fato motiva a definição da quantidade $\mathrm{IM}(X(t) : \zeta_y^{t-1}/X^{t-})$. Um cuidado que se deve tomar é que, embora no caso de duas séries não necessariamente univariadas, a equivalência entre a nulidade dos respectivos coeficientes da representação AR e MM seja válida, no caso mais geral, em que se envolvem três ou mais séries não necessariamente univariadas a nulidade de uma, em geral, não implica a nulidade da outra e portanto o siginificado das medidas de causalidade de Granger de Hosoya são distintas.

Diferentemente da medida de causalidade de Granger, a medida de Hosoya, aparentemente, não tem correspodentes imediatos, o que pode ser justificado pelo fato de apresentar uma interpretação menos imediata que a medida de causalidade de Granger e pela dificuldade de definir em termos probabilísticos a inovação ortogonalizada para processos estacionários não necessariamente gaussianos. Veja Ronsenblatt (1971) para uma discussão sobre representações de processos estacionários em geral como funções de processos i.i.d., que no caso

gaussiano corresponde à representação MM.

Pode-se mostrar (veja Seção 7.3) que no caso gaussiano

$$\mathrm{IM}(X(t) : Y^{t-}/X^{t-}) \geq \mathrm{IM}(X(t) : \zeta_y^{t-}/X^{t-}),$$

e que

$$\mathrm{IM}(X(t) : \zeta_y^{t-}/X^{t-}) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \log(1 - \mathrm{CQT}(dZ_x(\lambda) : dZ_{\xi_y}(\lambda)/dZ_{\xi_x}(\lambda)))d\lambda.$$

Um fato importante é que se pode parametrizar $\mathrm{CQT}(dZ_x(\lambda) : dZ_{\xi_y}(\lambda)/dZ_{\xi_x}(\lambda))$ pelos coeficientes da representação AR (6.1) e MM (6.2) da seguinte forma. Defina

$$\tilde{A}_{zw}(\lambda) = \delta_{zw} - \sum_{k=1}^{\infty} A_{zw}(k)e^{-i\lambda k},$$
$$\tilde{H}_{zw}(\lambda) = \sum_{k=0}^{\infty} H_{zw}(k)e^{-i\lambda k},$$

em que $z = x$ ou y e $z = x$ ou y. Ainda, $\delta_{zw} = 1$ se $z = w$ e $\delta_{zw} = 0$ caso contrário. Assim,

$$\begin{aligned} &\mathrm{CQT}(dZ_x(\lambda) : dZ_{\xi_y}(\lambda)/dZ_{\xi_x}(\lambda)) \\ &= \frac{|\tilde{H}_{xy}(\lambda)|^2 \mathrm{Var}(\xi_y(t)/\xi_x(t))}{[\tilde{H}_{xx}(\lambda)\ \tilde{H}_{xy}(\lambda)]\mathrm{Var}(\xi_x(t), \xi_y(t))[\tilde{H}_{xx}(\lambda)\ \tilde{H}_{xy}(\lambda)]^*} \qquad (6.9) \\ &= \frac{|\tilde{A}_{xy}(\lambda)|^2 \mathrm{Var}(\xi_x(t))^{-1}}{[\tilde{A}_{xy}(\lambda)^*\ \tilde{A}_{yy}(\lambda)^*]\mathrm{Var}(\xi_x(t), \xi_y(t))^{-1}[\tilde{A}_{xy}(\lambda)\ \tilde{A}_{yy}(\lambda)]^T}. \qquad (6.10) \end{aligned}$$

As expressões (6.9) e (6.10) explicitam a relação entre a medida de Hosoya e os coeficientes da represetação AR e MM. De certa forma é surpreendente

que se possa interpretar a medida de Hosoya tanto em termos dos coeficientes da representação MM assim como pelos coeficientes da representação AR, pois ambos apresentam significado bastante distintos. Mais explicitamente, tem-se

$$H_{xy}(k) = \frac{\mathrm{Cov}(X(t) : Y(t-k)/X(t-k), X^{(t-k)-}, Y^{(t-k)-})}{\mathrm{Var}(\xi_y(t-k)/\xi_x(t-k))} \tag{6.11}$$

e

$$A_{xy}(k) = \frac{\mathrm{Cov}(X(t) : Y(t-k)/X(t-1), Y(t-1), \ldots, X(t-k), X^{(t-k)-}, Y^{(t-k)-})}{\mathrm{Var}(Y(t-k)/X(t-1), Y(t-1), \ldots, X(t-k), X^{(t-k)-}, Y^{(t-k)-})}. \tag{6.12}$$

Ou seja, $H_{xy}(k)$ representa a relação entre $X(t)$ e $Y(t-k)$ que não é devido a $X(t-k), X^{(t-k)-}$ e $Y^{(t-k)-}$, em outras palavras, está relacionada à informação de $Y(t-k)$ que "alcança" $X(t)$, não se importando com o caminho que esse percorre de $Y(t-k)$ até chegar a $X(t)$. Por outro lado, $A_{xy}(k)$ está relacionada à informação que sai de $Y(t-k)$ e chega a $X(t)$ "diretamente" sem que passe por $X(t-1), Y(t-1), \ldots, X(t-k+1), Y(t-k+1)$, ou seja, os coeficientes da representação MM estão relacionados com a noção de *alcance da informação* enquanto os coeficientes da representação AR estão relacionados com a noção de *informação direta*. Desta forma, não é imediato que no caso bivariado ambas as noções resultem numa mesma medida que é a medida de causalidade de Hosoya. Este fato é discutido com mais detalhes no Capítulo 7.

Como última observação, pode-se dizer que a igualdade entre (6.9) e (6.10), é um confusor na literatura. A existência da igualdade foi indicada inicialmente em Sameshima e Baccalá (1999), no caso particular de séres estacionárias de segunda ordem com matriz de variância/covariância dos resíduos igual a identidade, e o caso geral foi provado em Takahashi et al. (2006). O fato da causalidade

de Granger estar definida explicitamente para o caso de duas séries e poder ser enunciada utilizando tanto a representação AR como MM, tem feito com que diferentes trabalhos definam a causalidade de Granger e generalizem-na de formas distintas. Neste sentido, um dos objetivos do Capítulo 7 é tentar elucidar melhor a diferença que existe entre as medidas de causalidade de Granger e Hosoya, ou dita de outra forma, entre a causalidade baseada na representação AR e MM.

6.0.2 Modelo 2

O Modelo 2 abaixo serve para motivar a existência de uma dualidade entre os conceitos de causalidade e serve para analisar melhor a diferença de interpretação que existe entre as representações AR e MM. O modelo foi sugerido por Hosoya (2001).

Exemplo 6.0.2 (Modelo 2)**.** *Sejam X, Y e Z séries univariadas conjuntamente estacionárias e gaussianas com representação AR*

$$X(t) = -0.25Y(t-2) + 0.5Z(t-1) + \epsilon(t) \tag{6.13}$$

$$Y(t) = \xi(t) \tag{6.14}$$

$$Z(t) = 0.5Y(t-1) + \eta(t) \tag{6.15}$$

com $Var(\epsilon(t), \xi(t), \eta(t)) = I$ *e representação MM*

$$X(t) = \epsilon(t) + 0.5\eta(t-1) \tag{6.16}$$

$$Y(t) = \xi(t) \tag{6.17}$$

$$Z(t) = \eta(t) + 0.5\xi(t-1). \tag{6.18}$$

Como a representação MM é inversível a representação AR acima é de fato a representação AR estável (Lütkepohl, 1993).

Hosoya (2001) supôs que a série tenha sido gerada utilizando a representação MM e então concluiu que Y não causa X pois são independentes, porém quando se analisa a representação AR observa-se que $Y(t-2)$ aparece na equação de regressão de $X(t)$ e então acaba se concluindo "erroneamente", segudo Hosoya, que Y causa X ou manda informação para X. Em seu trabalho, Hosoya (2001) sugere uma medida de causalidade que não sofre deste "poblema".

Embora interessante, o argumento de Hosoya não é totalmente convincente, pois se a série é gerada utilizando a representação AR parece ser razoável assumir que Y causa X. A questão natural que surge é: qual a razão desta diferença de interpretação dependendo da representação AR ou MM que escolhe? Do ponto de vista interpretativo não é satisfatório que a interpretação dependa da representação que se assume ser a geradora do processo (AR ou MM). Analisando este exemplo, fica claro que a diferença de interpretação é devido à diferença entre as formas de independência condicionada consideradas. Mais especificamente, na interpretação de Hosoya, a causalidade entre as séries deve ser interpretada sem condicionamento, ou seja, parafraseando-se as considerações de Hosoya em termos da Teoria da Informação, considera-se que não há causalidade se $\mathrm{IM}(X(t) : \xi^{t-}/X^{t-}, \epsilon^{t-}, \eta^{t-}) = 0$ que no Modelo 2 correponde à condição $\mathrm{IM}(X(t) : Y^{t-}/\epsilon^{t-}, \eta^{t-})$. Já na interpretação utilizando a representação AR, somente considera-se que não há causalidade se $\mathrm{IM}(X(t) : Y^{t-}/X^{t-}, Z^{t-}) = 0$. Pode-se ver que $\mathrm{IM}(X(t) : Y^{t-}/\epsilon^{t-}, \eta^{t-})$ é nulo e portanto não há causalidade segundo Hosoya, porém $\mathrm{IM}(X(t) : Y^{t-}/X^{t-}, Z^{t-}) \neq 0$, pois na representação AR (6.13) vê-se um coeficiente não nulo entre $X(t)$ e $Y(t-2)$. A razão para

isto está no fato de X^{t-} e Y^{t-} serem independentes, porém dependentes condicionalmente em Z^{t-}.

De fato há argumentos prós e contras às duas condições para não causalidade e dependendo da situação uma é mais adquada do que a outra. No capítulo seguinte é desenvolvida uma forma sistemática de se estudar medidas de causalidade e generalizar se for o caso.

CAPÍTULO 7

Medidas de dependência entre séries temporais

"Or again, in the study of brain waves we may be able to abtain electroencephalograms more or less corresponding to electrical activity in different parts of the brain. Here the study of the coeffcients of causality running both ways and of their analogue for sets of more than two functions f may be useful in determining what part of the brain is driving what other part of the brain in its normal activity."
(Nobert Wiener, 1959)

Neste capítulo, os conceitos desenvolvidos nos Capítulos 3 e 4 são utilizadas para um estudo sistemático da dependência entre séries temporais motivadas no Capítulo 6. O objetivo principal neste capítulo é estudar, à luz dos conceitos de medidas de dependência estudadas nos capítulos anteriores, a idéia de *causalidade de Granger* introduzida por Granger (1969) e relacioná-la com uma medida de "fluxo de informação" no domínio da freqüência denominada *coerência parcial direcionada* introduzida em Sameshima e Baccalá (1999); Baccalá e Sameshima

(2001). Para atingir o objetivo e tornar as idéias mais naturais, alguns preparativos são feitos até que se obtenha o resultado final.

É interessante observar que a coerência parcial direcionada e outras medidas "fluxo de informação" sempre foram implicitamente consideradas relacionadas ao conceito de Granger. Este capítulo tem o objetivo de explicitar essa relação usando como conceito-chave a noção de informação desenvolvida na Teoria de Informação (Shannon e Weaver, 1949).

Para as questões deste capítulo, existem duas diferenças principais em relação às considerações feitas nos Capítulos 3 e 4. A primeira refere-se ao fato de o estudo de dependências entre séries temporais envolverem necessariamente o estudo de dependência entre infinitas v.as., o que exige maiores cuidados para se verificar a validade matemática das medidas de dependência sugeridas matematicamente. A postura neste capítulo é sempre interpretar as medidas como sendo limites de uma seqüência de medidas definidas para um número finito de v.as.

A segunda diferença refere-se à interpretação dada à assimetria no tempo. Tipicamente divide-se as séries temporais em passado, presente e futuro, e as medidas de dependência entre séries temporais devem ser consistentes com a interpretação no tempo. Isso introduz novas dificuldades para o estudo de medidas de dependência entre séries temporais.

Como roteiro deste capítulo, na Seção 7.1 são obtidos resultados sobre o comportamento assintótico de algumas quantidades da Teoria da Informação que são utilizadas para obter os resultados principais desta tese. A expressão exata para as taxas de entropia para processos gaussianos estacionários é obtida.

Na Seção 7.2 as medidas de dependência linear entre séries temporais denominadas simétricas são discutidas. A simetria diz respeito ao fato de as me-

didas definidas nesta seção não introduzirem assimetria de dependência entre as séries envolvidas. É discutida nesta seção o papel da representação espectral que, dentre as diversas propriedades que apresenta, permite o estudo da dependência entre séries temporais utilizando ipsis literis os métodos desenvolvidos para v.as. no Capítulo 4.

Na Seção 7.3 é apresentado o resultado principal desta tese que é a relação entre a coerência parcial direcionada e a causalidade de Granger. Outras medidas de fluxo de informação são discutidas e comparadas.

7.1 Alguns teoremas assintóticos para séries temporais estacionárias gaussianas

Nesta seção são apresentados alguns resultados sobre o comportamento assintótico de séries temporais estacionárias gaussianas que satisfazem a condição de limitação (5.2). A maioria dos resultados é bem conhecida e podem ser apresentadas utilizando somente a Teoria da Informação ou somente a teoria dos processos gaussianos estacionários. Aqui é feita uma ponte entre os dois que, embora seja clara, não aparece explicitamente na literatura, com a exceção de Pinsker (1964), Ihara (1964).

A seguinte proposição é útil.

Proposição 7.1.1. *Seja X uma v.a. n-dimensional e Z um processo m-dimensional, não necessariamente estacionário, conjuntamente gaussianos tais que* det *Var*$(\bar{R}(X/Z^{t-})) > 0$. *Tem-se*

$$\lim_{j\to\infty} H(X/Z(t-1),\ldots,Z(t-j)) \tag{7.1}$$

$$= \frac{n}{2}\log(2\pi e) + \frac{1}{2}\log\det Var(X/Z^{t-}). \tag{7.2}$$

Demonstração. A prova é uma conseqüência imediata da convergência de seqüências monotonicamente não decrescentes de operadores de projeção ortogonal (veja, por exemplo, p.68, Akhiezer e Glazman (1993)). Seja a seqüência $\xi_x^{(j)} = \bar{R}(X/Z(t-1),\ldots,Z(t-j))$, $j \geq 1$, de projeções em subespaços monotonicamente crescentes e defina $\xi_x = \bar{R}(X/Z^{t-})$. Pela convergência de seqüências montônicas de projeções ortogonais, $\xi_x^{(j)} \to \xi_x$ em L^2 e portanto $\mathrm{Var}(\xi_x^{(j)}) \to \mathrm{Var}(\xi_x)$. Logo

$$\frac{n}{2}\log(2\pi e) + \frac{1}{2}\log\det\mathrm{Var}(\xi_x^{(j)}) \to \frac{n}{2}\log(2\pi e) + \frac{1}{2}\log\det\mathrm{Var}(\xi_x)$$

o que conclui a demonstração. □

Um corolário imediato é o seguite.

Corolário 7.1.1. *Sejam X e Y processos n e m-dimensionais conjuntamente estacionários e gaussianos que satisfazem a condição de limitação conjuntamente. Considere a representação AR*

$$\begin{bmatrix} X(t) \\ Y(t) \end{bmatrix} = \sum_{k=1}^{\infty} \begin{bmatrix} A_{xx}(k) & A_{xy}(k) \\ A_{yx}(k) & A_{yy}(k) \end{bmatrix} \begin{bmatrix} X(t-k) \\ Y(t-k) \end{bmatrix} + \begin{bmatrix} \xi_x(t) \\ \xi_y(t) \end{bmatrix}. \tag{7.3}$$

Tem-se

$$\lim_{j\to\infty} H(X(t)/X(t-1), Y(t-1), \ldots, X(t-j), Y(t-j)) \tag{7.4}$$

$$= \frac{n}{2}\log(2\pi e) + \frac{1}{2}\log\det Var(\xi_x(t)). \tag{7.5}$$

Demonstração. Aplicação direta da Proposição 7.1.1, tomando $X(t)$ como v.a. e escolhendo como processo a série Z definida por $Z(t)^T = [X(t)^T\ Y(t)^T], t \in \mathbb{Z}$. □

As seguintes identidades são fundamentais e dizem respeito à taxa de entropia na Definição (3.1.9).

Teorema 7.1.1. *Seja X uma série estacionária ergódica com densidade de probabilidade p_x tal que $H(X(t)/X(t-1), \ldots, X(t-j)) > -\infty$ para todo j maior que algum $m > 0$. Tem-se*

$$h(X) = \lim_{j\to\infty} \frac{1}{j+1} H(X(t), \ldots, X(t-j)) \tag{7.6}$$

$$= \lim_{j\to\infty} H(X(t)/X(t-1), \ldots, X(t-j)). \tag{7.7}$$

Demonstração. A igualdade entre (7.6) e (7.7) é bem conhecida e é válida para processos estritamente estacionários em geral (não necessariamente ergódicas) com densidades tais que $\mathrm{H}(X(t)/X(t-1), \ldots, X(t-j)) > -\infty$ para todo j maior que algum m. A prova pode ser encontrada em Ihara (1964, p. 60)

Teorema 2.1.1. □

O seguinte corolário é utilizado repetidas vezes nas seções seguintes.

Corolário 7.1.2. *Seja X uma série temporal gaussiana estacionária n-dimensional com matriz de densidade espectral f_x e que satisfaz a condição de limitação. Seja η_x o seu processo de inovação. Tem-se*

$$h(X) = \frac{n}{2}\log(2\pi e) + \frac{1}{2}\log\det Var(\eta_x(t)) \tag{7.8}$$

$$= \frac{n}{2} + n\log(2\pi) + \frac{1}{4\pi}\int_{-\pi}^{\pi}\log\det f_x(\lambda)d\lambda. \tag{7.9}$$

Demonstração. Tem-se $\mathrm{H}(X(t)/X(t-1),\ldots,X(t-j)) > -\infty$ para todo $j \geq 1$, pois $\mathrm{Var}(\bar{R}(X(t)/X(t-1),\ldots,X(t-j))) \geq \mathrm{Var}(\bar{R}(X(t)/X^{t-}))$ e portanto

$$\mathrm{H}(X(t)/X(t-1),\ldots,X(t-j)) \geq \frac{n}{2}\log(2\pi e) + \frac{1}{2}\log\det\mathrm{Var}(\eta_x(t)) > -\infty,$$

para todo $j \geq 1$, em que para a última desigualdade usou-se o Teorema 5.0.1 que garante $\mathrm{Var}(\eta_x(t)) > 0$. Assim, igualdade entre (7.7) e (7.8) é conseqüência da Proposição 7.1.1 em que se tomou como série Z da proposição o próprio passado de X. A igualdade entre (7.8) e (7.9) é conseqüência da identidade de Szegö (Teorema 5.0.1). □

7.2 Medidas simétricas

Dadas duas séries, é uma questão natural se perguntar pela informação em comum que elas apresentam. Tipicamente, a informação em comum entre dois processos é infinito, fornecendo pouca informação sobre as dependências, porém a taxa com que as medidas de dependência crescem é bem comportada e fornece informações mais interessantes.

A seguir é definida uma taxa de informação mútua entre perocessos que é natural e utiliza a definição de taxa de informação mútua entre seqüências da Definição 3.1.5.

Definição 7.2.1. Sejam $X_1, \ldots, X_n$ séries não necessariamente univariadas conjuntamente estacionárias com densidades. A taxa de informação mútua $\mathrm{TIM}(X_1 : \ldots : X_n)$ entre $X_1, \ldots, X_n$ é definida como

$$\mathrm{TIM}(X_1 : \ldots : X_n) = \lim_{j\to\infty} \frac{1}{j+1}\mathrm{IM}(\{X_1\}_t^{t-j} : \ldots : \{X_n\}_t^{t-j}),$$

quando o limite existir.

Proposição 7.2.1. *Sejam $X_1, \ldots, X_n$ séries não necessariamente univariadas conjuntamente estacionárias com densidades, não necessariamente gaussianas, tais que $H(X_1(t), \ldots, X_n(t)/\{X_1\}_{t-1}^{t-j}, \ldots, \{X_n\}_{t-1}^{t-j}) > -\infty$ para todo j maior que algum $m > 0$. A taxa de informação mútua $TIM(X_1 : \ldots : X_n)$ é dada por*

$$TIM(X_1 : \ldots : X_n) = \sum_{k=1}^{n} h(X_k) - h(X_1, \ldots, X_n). \tag{7.10}$$

Demonstração. Pela Definição 7.2.1, basta calcular

$$\begin{aligned}
&\lim_{j\to\infty} \frac{1}{j+1}\mathrm{IM}(\{X_1\}_t^{t-j} : \ldots : \{X_n\}_t^{t-j}) \\
&= \lim_{j\to\infty} \frac{1}{j+1}\Big(\sum_{k=1}^{n} \mathrm{H}(\{X_k\}_t^{t-j}) - \mathrm{H}(\{X_1\}_t^{t-j}, \ldots, \{X_n\}_t^{t-j})\Big) \\
&= \sum_{k=1}^{n} \mathrm{h}(X_k) - h(X_1, \ldots, X_n),
\end{aligned}$$

em que a última igualdade segue do fato de

$$\sum_{k=1}^{n} \mathrm{H}(X_k(t)/\{X_k\}_{t-1}^{t-j}) \geq \mathrm{H}(X_1(t), \ldots, X_n(t)/\{X_1\}_{t-1}^{t-j}, \ldots, \{X_n\}_{t-1}^{t-j}),$$

e pela Proposição 7.1.1. □

Agora, é obtida a seguinte identidade:

Proposição 7.2.2. *Sejam X e Y séries n e m-dimensionais conjuntamente estacionárias e gaussianas com densidades espectrais $f_{(xy)}$ conjunta e f_x, f_y individuais que satisfazem a condição de limitação. Sejam $\int e^{i\lambda t} dZ_x(\lambda)$ e $\int e^{i\lambda t} dZ_y(\lambda)$ as suas respectivas representações espectrais. A taxa de informação mútua $TIM(X:Y)$ entre X e Y pode ser escrita como*

$$
\begin{aligned}
& TIM(X:Y) \\
(7.11) \qquad & = -\frac{1}{2}\log \frac{\det Var(X(t), Y(t)/X^{t-}, Y^{t-})}{\det Var(X(t)/X^{t-}) \det Var(Y(t)/Y^{t-})} \\
(7.12) \qquad & = -\frac{1}{4\pi}\int_{-\pi}^{\pi} \log\left(\frac{\det f_{(xy)}(\lambda)}{\det f_x(\lambda) \det f_y(\lambda)}\right) d\lambda \\
(7.13) \qquad & = -\frac{1}{4\pi}\int_{-\pi}^{\pi} \log(1 - CQT(dZ_x(\lambda) : dZ_y(\lambda)))d\lambda.
\end{aligned}
$$

Demonstração. Como as séries satisfazem a condição de limitação, elas satisfazem as condições da Proposição 7.2.1 e portanto

$$\mathrm{TIM}(X:Y) = \mathrm{h}(X) + \mathrm{h}(Y) - \mathrm{h}(X,Y).$$

Pelo Corolário 7.1.2

$$
\begin{aligned}
& \mathrm{h}(X) + \mathrm{h}(Y) - \mathrm{h}(X,Y) \\
& = -\frac{1}{2}\log \frac{\det \mathrm{Var}(X_(t), Y(t)/X^{t-}, Y^{t-})}{\det \mathrm{Var}(X(t)/X^{t-}) \det \mathrm{Var}(Y_(t)/Y^{t-})} \\
& = -\frac{1}{4\pi}\int_{-\pi}^{\pi} \log\left(\frac{\det f_{(xy)}(\lambda)}{\det f_x(\lambda) \det f_y(\lambda)}\right) d\lambda.
\end{aligned}
$$

A identidade (7.13) segue da definição de CQT, isto é,

$$\begin{aligned} 1 - \mathrm{CQT}(dZ_x(\lambda) : dZ_y(\lambda)) &= \frac{\det \mathrm{Var}(dZ_x(\lambda), dZ_y(\lambda))}{\det \mathrm{Var}(dZ_x(\lambda)) \det \mathrm{Var}(dZ_y(\lambda))} \\ &= \frac{\det f_{(xy)}(\lambda)}{\det f_x(\lambda) \det f_y(\lambda)}. \end{aligned}$$

□

No caso em que X e Y são unidimensionais $\mathrm{CQT}(dZ_x(\lambda) : dZ_y(\lambda))$ é exatamente o módulo quadrático da coerência entre X e Y, isto é,

$$\begin{aligned} \mathrm{CQT}(dZ_x(\lambda) : dZ_y(\lambda)) &= 1 - \frac{f_x(\lambda) f_y(\lambda) - |f_{xy}(\lambda)|^2}{f_x(\lambda) f_y(\lambda)} \\ &= \frac{|f_{xy}(\lambda)|^2}{f_x(\lambda) f_y(\lambda)}, \end{aligned}$$

em que f_{xy} é o espectro cruzado entre X e Y. Neste caso,

$$\begin{aligned} &\mathrm{TIM}(X : Y) \\ &= -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log \left(1 - \frac{\det f_{(xy)}(\lambda)}{\det f_x(\lambda) \det f_y(\lambda)} \right) d\lambda, \end{aligned}$$

o que mostra que $\mathrm{TIM}(X : Y) = 0$ se e somente se a coerência entre X e Y é zero para $\lambda \in [-\pi, \pi)$ quase certamente, o que é coerente.

Observação 7.2.1. A Proposição 7.2.2 foi provada para o caso de processos gaussianos contínuos e univariados por Gelfand e Yaglom (1959). Para o caso de processos gaussianos multivariados discretos a Proposição 7.2.2 foi demonstrada por Pinsker (1964) com condições menos restritivas que a condição de limitação assumida nesta tese. Ambos os trabalhos chamam a taxa de informação mútua como informação mútua média, que não é a denominação usual na literatura

de Teoria da Informação, além de não representar a idéia da quantidade que é de fato uma taxa e não uma média. A demonstração feita nesta tese é distinta daquela usada em Pinsker (1964). Aqui é utilizada diretamente a identidade de Szegö, o que facilita consideravelmente a demonstração. É interessante que, na literatura, muitas vezes o artigo de Gelfand e Yaglom (1959) é citado como fonte do resultado da Proposição 7.2.2, embora não seja o caso.

A generalização da Proposição 7.2.2 para o caso de mais de duas v.as. é imediata.

Proposição 7.2.3. *Sejam $X_1, \ldots, X_n$ séries não necessariamente univariadas conjuntamente estacionárias e gaussianas com densidades espectrais $f_{(x_1 \ldots x_n)}$ conjunta e $f_{x_1}, \ldots, f_{x_n}$ individuais que satisfazem a condição de limitação conjuntamente. Sejam $\int e^{i\lambda t} dZ_{x_k}(\lambda)$, $k = 1, \ldots, n$, as suas representações espectrais, respectivamente. A taxa de informação mútua $TIM(X_1 : \ldots : X_n)$ entre $X_1, \ldots, X_n$ pode ser escrita como*

$$
\begin{aligned}
& TIM(X_1 : \ldots : X_n) \\
(7.14) \qquad & = -\frac{1}{2} \log \frac{\det Var(X_1(t), \ldots, X_n(t)/X_1^{t-}, \ldots, X_n^{t-})}{\det Var(X_1(t)/X_1^{t-}) \ldots \det Var(X_n(t)/X_n^{t-})} \\
(7.15) \qquad & = -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log \left(\frac{\det f_{(x_1 \ldots x_n)}(\lambda)}{\det f_{x_1}(\lambda) \ldots \det f_{x_n}(\lambda)} \right) d\lambda \\
(7.16) \qquad & = -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log(1 - CQT(dZ_{x_1}(\lambda) : \ldots : dZ_{x_n}(\lambda))) d\lambda.
\end{aligned}
$$

Demonstração. A prova é idêntica à demonstração da Proposição 7.2.3. □

Alguns resultados obtidos para o caso finito agora podem ser generalizados utilizando-se os resultados anteriores.

Proposição 7.2.4. *Sejam $X_1, \ldots, X_n$ séries $d_1, \ldots, d_n$ dimensionais estacionárias de segunda ordem não necessariamente gaussianas com densidades espectrais conjunta $f_{(x_1 \ldots x_n)}$ e marginais f_{x_k}, $k = 1, \ldots, n$. Suponha $H(X_1(t), \ldots, X_n(t)/\{X_1\}_{t-1}^{t-j}, \ldots, \{X_n\}_{t-1}^{t-j}) > -\infty$ para todo j maior que algum $m > 0$. Tome $d = \sum d_k$. Os seguintes limites são válidos:*

$$\frac{d}{2} + d\log(2\pi) + \frac{1}{4\pi}\int_{-\pi}^{\pi} \log\det f_{(x_1 \ldots x_n)}(\lambda)d\lambda - h(X_1, \ldots, X_n) \tag{7.17}$$

$$\geq TIM(X_1 : \ldots : X_n) - \frac{1}{4\pi}\int_{-\pi}^{\pi} \log(1 - CQT(dZ_{x_1}(\lambda) : \ldots : dZ_{x_n}(\lambda)))^{-1}d\lambda \tag{7.18}$$

$$\geq \sum_{k=1}^{n} h(X_k) - \frac{d}{2} - d\log(2\pi) - \frac{1}{4\pi}\sum_{k=1}^{n}\int_{-\pi}^{\pi} \log\det f_{x_k}(\lambda)d\lambda, \tag{7.19}$$

em que (7.17) assume apenas valores não negativos e (7.19) assume apenas valores não positivos. Se as séries forem conjuntamente gaussianas tem-se a igualdade.

Demonstração. A prova se obtem pela aplicação direta das desigualdades da Proposição 3.1.3 para as séries e tomando os limites adequadamente. As identidades do Corolário 7.1.2 (equação (7.16))e Proposição 7.2.3 (equação (7.9)) concluem a demonstração. □

Outra conseqüência da Proposição 7.2.2 é a possibilidade de se calcular a taxa de informação mútua para processos não necessariamente gaussianos, mas que apresentem cópula gaussiana. A idéia é simplesmente usar o fato que no caso de um número finito de v.as. pode-se associar v.as. gaussianas com a mesma estrutura de variância/covariância. Para esse conjunto de v.as. gaussianas pode-se calcular explicitamente a informação mútua. Agora, usando o fato que a informação mútua depende somente da cópula associada à distribuição conjunta das v.as. chega-se a conclusão que pode-se obter a mesma fórmula do caso gaussiano para todas as v.as. com cópulas gaussianas com a mesma estrutura

de variância/covariância. Este o conteúdo do Corolário 3.1.6.

No caso de séries temporais, a taxa de informação mútua é simplesmente o limite da informação mútua adequadamente normalizada. Assim, tomando processos conjuntamente estacionários de segunda ordem cujas distribuições conjuntas finitas apresentam cópula gaussiana, basta associar processos gaussianos conjuntamente estacionários com a mesma função de autocovariância conjunta para o qual se pode calcular a taxa de informação mútua. Agora é imediato que a taxa de informação mútua para os processos originais apresentam o mesmo valor daquela obtida para os processos gaussianos associados.

A única dúvida que resta é a existência de tais processos. Porém, é claro que os processos gaussianos são exemplos de processos com cópula gaussiana e ainda outros exemplos podem ser construídos utilizando o Teorema de Existência de Kolmogorov (vide Billingsley (1995, Teorema 36.2, p.486)). Um estudo sobre processos definidos por cópulas é feita em Schmitz (2003).

A proposição a seguir resume estas observações.

Proposição 7.2.5. *Sejam $X_1, \ldots, X_n$ processos não necessariamente univariados conjuntamente estacionários de segunda ordem e cuja distribuições conjuntas finitas apresentam cópula gaussiana. Sejam $f_{(x_1 \ldots x_n)}$ as densidades espectrais conjuntas e $f_{x_1}, \ldots, f_{x_n}$ as densidades espectrais individuais que satisfazem a condição de limitação conjuntamente. Sejam $\int e^{i\lambda t} dZ_{x_k}(\lambda)$, $k = 1, \ldots, n$ as representações espectrais de X_k. A taxa de informação mútua entre $X_1, \ldots, X_n$ pode ser escrita como*

$$
\begin{aligned}
& TIM(X_1 : \ldots : X_n) \\
(7.20) \qquad & = -\frac{1}{2} \log \frac{\det Var(X_1(t), \ldots, X_n(t)/X_1^{t-}, \ldots, X_n^{t-})}{\det Var(X_1(t)/X_1^{t-}) \ldots \det Var(X_n(t)/X_n^{t-})} \\
(7.21) \qquad & = -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log \left(\frac{\det f_{(x_1 \ldots x_n)}(\lambda)}{\det f_{x_1}(\lambda) \ldots \det f_{x_n}(\lambda)} \right) d\lambda \\
(7.22) \qquad & = -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log(1 - CQT(dZ_{x_1}(\lambda) : \ldots : dZ_{x_n}(\lambda))) d\lambda.
\end{aligned}
$$

Demonstração. Sejam $Y_1, \ldots, Y_n$ processos gaussianos tais que $\mathrm{Cov}(Y_k(t), Y_l(s)) = \mathrm{Cov}(X_k(t), X_l(s)), \forall t, s \in \mathbb{Z}, k, l = 1, \ldots, n$. Pelo Corolário 3.1.6

$$
\mathrm{IM}(\{Y_1\}_t^{t-j} : \ldots : \{Y_n\}_t^{t-j}) = \mathrm{IM}(\{X_1\}_t^{t-j} : \ldots : \{X_n\}_t^{t-j}),
$$

para todo $j \geq 0$. Logo

$$
\lim_{j \to \infty} \frac{1}{j+1} \mathrm{IM}(\{Y_1\}_t^{t-j} : \ldots : \{Y_n\}_t^{t-j}) = \lim_{j \to \infty} \frac{1}{j+1} \mathrm{IM}(\{X_1\}_t^{t-j} : \ldots : \{X_n\}_t^{t-j}).
$$

Assim, pela Proposição 7.2.2, segue o resultado. □

A proposição acima, aparentemente simples, é interessante no sentido de permitir a construção de processos que não são estacionários em senso estrito cujas informações mútuas podem ser calculadas explicitamente, o que é em geral

um problema difícil.

As versões parcializadas das definições e teoremas acimas podem ser obtidos com modificações adequadas. Uma possibilidade natural para uma medida de dependência parcializada é dada pela seguinte definição.

Definição 7.2.2 (Taxa de informação mútua dada uma outra série)**.** Sejam $X_1, \ldots, X_n$ e Y séries não necessariamente univariadas conjuntamente estacionárias com densidades de probabilidade. A *taxa de informação mútua* $\mathrm{TIM}(X_1 : \ldots : X_n/Y)$ entre $X_1, \ldots, X_n$ dado Y é definida, quando os limites existirem, como

$$\mathrm{TIM}(X_1 : \ldots : X_n/Y) = \lim_{j\to\infty} \frac{1}{j+1} \mathrm{IM}(\{X_1\}_t^{t-j} : \ldots : \{X_n\}_t^{t-j}/Y),$$

em que

$$\mathrm{IM}(\{X_1\}_t^{t-j} : \ldots : \{X_n\}_t^{t-j}/Y) = \lim_{p\to\infty} \mathrm{IM}(\{X_1\}_t^{t-j} : \ldots : \{X_n\}_t^{t-j}/\{Y\}_{-p}^{p}).$$

No caso gaussiano pode-se provar sua existência e obter a expressão explícita da versão parcializada da taxa de informação mútua entre processos.

Proposição 7.2.6. *Sejam $X_1, \ldots, X_n$ e Y séries não necessariamente univariadas conjuntamente estacionárias e gaussianas com densidades espectrais $f_{(x_1 \ldots x_n y)}$ conjunta e $f_{x_1 y}, \ldots, f_{x_n y}, f_y$ densidades espectrais das séries dos respectivos índices. Suponha que as séries satisfaçam a condição de limitação conjuntamente. Sejam $\int e^{i\lambda t} dZ_{x_k}(\lambda)$, $k = 1, \ldots, n$ e $\int e^{i\lambda t} dZ_y(\lambda)$ as suas representações espectrais, respectivamente. A taxa de informação mútua $TIM(X_1 : \ldots : X_n/Y)$ entre $X_1, \ldots, X_n$ dado Y pode ser escrita como*

$$
\begin{aligned}
& TIM(X_1 : \ldots : X_n/Y) \\
(7.23) \qquad & = -\frac{1}{2} \log \frac{\det Var(X_1(t), \ldots, X_n(t)/X_1^{t-}, \ldots, X_n^{t-}, Y)}{\det Var(X_1(t)/X_1^{t-}, Y) \ldots \det Var(X_n(t)/X_n^{t-}, Y)} \\
(7.24) \qquad & = -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log \left(\frac{\det f_{(x_1 \ldots x_n y)}(\lambda) \{\det f_y(\lambda)\}^{n-1}}{\det f_{x_1 y}(\lambda) \ldots \det f_{x_n y}(\lambda)} \right) d\lambda \\
(7.25) \qquad & = -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log(1 - CQT(dZ_{x_1}(\lambda) : \ldots : dZ_{x_n}(\lambda)/dZ_y(\lambda))) d\lambda.
\end{aligned}
$$

Demonstração. Basta observar que no caso gaussiano, assim como no caso finito, $\mathrm{TIM}(X_1 : \ldots : X_n/Y) = \mathrm{TIM}(\xi_1 : \ldots : xi_n)$, em que $\xi_k(t) = \bar{R}(X(t)/Y)$, $t \in \mathbb{Z}$ $k = 1, \ldots, n$. Aplicando a identidade (7.14) obtém-se (7.23). Agora, sabe-se que (veja por exemplo Brillinger (1981, p.296, equação (8.3.8)))

$$f_{\xi_k}(\lambda) = f_{x_k}(\lambda) - f_{x_k y}(\lambda) f_y(\lambda)^{-1} f_{y x_k}(\lambda),$$

para $k = 1, \ldots, n$, $\lambda \in [-\pi, \pi)$. Lembre que $f_{x_k y}$ em que o índice não apresenta parênteses é o espectro cruzado. Logo,

$$\det f_{\xi_k}(\lambda) = \frac{\det f_{(x_k y)}(\lambda)}{\det f_y(\lambda)}.$$

Analogamente

$$\det f_{\xi_1 \ldots \xi_n}(\lambda) = \frac{\det f_{(x_1 \ldots x_n y)}(\lambda)}{\det f_y(\lambda)}.$$

Utilizando (7.15) obtém-se (7.24). A identidade (7.25) é obtida simplesmente pela aplicação da definição de CQT parcializada (Definição 4.2.5). □

Observe que no caso em que X_1, X_2 e Y são processos univariados conjuntamente estacionários e gaussianos, a CQT$(dZ_{x_1}(\lambda) : dZ_{x_2}(\lambda)/dZ_y)$ é simplesmente a coerência parcial entre X_1 e X_2 dado Y.

Para finalizar a analogia com o caso de v.as. finitas, pode-se definir a taxa de informação mútua inversa que simplesmente consiste em calcular as mesmas medidas de dependência entre séries definidas anteriormente para as séries inversas (Definição 5.0.1). Definir os processos inversos para processos estritamente estacionários em geral não parece simples.

No capítulo de séries temporais foram definidas as séries inversas de uma série n-dimensional. É útil definir o siginificado de séries inversas para um conjunto de séries temporais estacionárias.

Definição 7.2.3 (Processos inversos para um conjunto finito de v.as.)**.** Sejam $X_1, \ldots, X_n$ séries não necessariamente unidimensionais. As suas respectivas séries inversas ${}^iX_1, \ldots, {}^iX_n$ são definidas como sendo os respectivos componentes da série inversa de $X^T = [X_1^T \ \ldots \ X_n^T]$, isto é, ${}^iX^T = [{}^iX_1^T \ \ldots \ {}^iX_n^T]$.

Definição 7.2.4 (Taxa de informação mútua inversa)**.** Sejam $X_1, \ldots, X_n$ séries não necessariamente univariadas conjuntamente estacionárias e gaussianas. Sejam ${}^iX_1, \ldots, {}^iX_n$ as respectivas séries inversas. A *taxa de informação mútua inversa* $\mathrm{iTIM}(X_1 : \ldots : X_n)$ entre $X_1, \ldots, X_n$ é definida como

$$\mathrm{iTIM}(X_1 : \ldots : X_n) = \lim_{j \to \infty} \frac{1}{j+1} \mathrm{IM}(\{{}^iX_1\}_t^{t-j} : \ldots : \{{}^iX_n\}_t^{t-j}),$$

quando o limite existir.

Os processos inversos para processos estacionários exercem exatamente o mesmo papel que as v.as. inversas tal que assim como as v.as. inversas são as v.as. cuja matriz de variância/covariância é o inverso da matriz das v.as. originais; o processo inverso é o processo cuja matriz de densidade espectral é o inverso da matriz espectral do processo original. Logo, os resultados esperados ao se considerar os processos inversos são semelhantes daqueles obtidos para as v.as. inversas e espera-se que exista uma forte relação com as medidas parcializadas que é de fato o caso.

O resultado que se obtém para a taxa de informação mútua inversa é o seguinte.

Proposição 7.2.7. *Sejam $X_1, \ldots, X_n$ séries não necessariamente univariadas conjuntamente estacionárias e gaussianas com densidades espectrais $f_{(x_1 \ldots x_n)}$ conjunta e $f_{x_1}, \ldots, f_{x_n}$ as densidades espectrais das séries dos respectivos índices. Suponha que as séries satisfaçam a condição de limitação conjuntamente. Defina $(X^k)^T = [X_1^T \ldots X_{k-1}^T \; X_{k+1}^T \ldots X_n^T], k = 1, \ldots, n$, ou seja, é a série formada por todas as séries exceto X_k. Sejam $\int e^{i\lambda t} dZ_{x_k}(\lambda),\ k = 1, \ldots, n$, as suas representações espectrais, $\int e^{i\lambda t} d^i Z_{x_k}(\lambda),\ k = 1, \ldots, n$, as representações espectrais das v.as. inversas ${}^i X_k$ e $\int e^{i\lambda t} dZ_{x^k}(\lambda),\ k = 1, \ldots, n$ a representação espectral de X^k. A taxa de informação mútua inversa $iTIM(X_1 : \ldots : X_n)$ entre $X_1, \ldots, X_n$ pode ser escrita como*

$$
\begin{aligned}
& TIM(X_1 : \ldots : X_n) \\
(7.26) \quad & = -\frac{1}{2} \log \frac{\det Var(X_1(t), \ldots, X_n(t)/X_1^{t-}, \ldots, X_n^{t-})^{-1}}{\det Var(X_1(t)/X_1^{t-}, X^1)^{-1} \ldots \det Var(X_n(t)/X_n^{t-}, X^n)^{-1}} \\
(7.27) \quad & = -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log \left(\frac{\det f_{(x_1 \ldots x_n)}(\lambda)^{-1}}{\det f_{x_1/x^1}(\lambda)^{-1} \ldots \det f_{x_n/x^n}(\lambda)^{-1}} \right) d\lambda \\
(7.28) \quad & = -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log(1 - CQT(d^i Z_{x_1}(\lambda) : \ldots : d^i Z_{x_n}(\lambda)) d\lambda,
\end{aligned}
$$

em que $f_{x_k/x^k}, k = 1, \ldots, n$ é o espectro parcializado de X_k dado o resto dos processos, ou seja,

$$
(7.29) \qquad f_{x_k/x^k}(\lambda) = f_{x_k}(\lambda) - f_{x_k x^k}(\lambda) f_{x^k}(\lambda)^{-1} f_{x^k x_k}(\lambda).
$$

Demonstração. A identidade (7.28) é imediato por (7.22). O restante das identidades são conseqüências do fato da matriz espectral do processo inverso ser dada por $f_{(x_1 \ldots x_n)}(\lambda)^{-1},\ \lambda \in [-\pi, \pi)$, ou seja, pelo inverso da matriz de densidade espectral conjunta dos processos. □

No caso de duas e três séries, resultados análogos ao caso de v.as. são obtidos.

Mais especificamente:

Proposição 7.2.8. *Sejam X e Y duas séries não necessariamente univariadas conjuntamente estacionárias e gaussianas que satisfazem a condição de limitação e ${}^iX$ e ${}^iY$ as suas séries inversas respectivamente. Tem-se*

$$TIM({}^iX :^i Y) = TIM(X : Y). \tag{7.30}$$

Demonstração. É uma conseqüência imediada das Proposições 7.2.3 e 7.2.7 e de (4.48). □

Proposição 7.2.9. *Sejam X, Y e Z séries não necessariamente univariadas conjuntamente estacionárias e gaussianas que satisfazem a condição de limitação e ${}^iX$,${}^iY$ e ${}^iZ$ as suas séries inversas respectivamente. Tem-se*

$$TIM({}^iX :^i Y) = TIM(X : Y/Z). \tag{7.31}$$

Demonstração. É uma conseqüência das Proposições 4.2.11 e 7.2.7. □

Com esses resultados pode-se concluir que, no caso de processos estacionários e gaussianos, pode-se obter diferentes medidas de dependências que são análogas completas das medidas de dependência linear entre v.as.

7.3 Medidas de dependências assimétricas

Nesta seção são finalmente obtidas expressões para algumas medidas de fluxo de informação entre séries temporais.

A primeira medida de fluxo de informação é a representação em termos da Teoria da Informação do conceito da causalidade de Granger foi proposta inicialmente por Geweke (1982) para processos estacionários gaussianos.

Definição 7.3.1. Sejam X e Y séries conjuntamente estacionárias em senso estrito. A *medida de causalidade de Granger* de Y para X é definida como

$$\mathrm{IM}(X(t) : Y^{t-}/X^{t-}) = \lim_{j\to\infty} \mathrm{IM}(X(t) : \{Y\}_{t-1}^{t-j}/X^{t-}), \tag{7.32}$$

quando o limite existir.

Proposição 7.3.1. *Sejam X e Y séries conjuntamente estacionárias e gaussianas que satisfazem a condição de limitação. A medida de causalidade de Granger é dada por*

$$IM(X(t) : Y^{t-}/X^{t-}) = -\frac{1}{2}\log\frac{Var(X(t)/X^{t-}, Y^{t-})}{Var(X(t)/X^{t-})}. \tag{7.33}$$

Demonstração. Como as séries satisfazem a condição de limitação. Pode-se escrever

$$\mathrm{IM}(X(t) : Y^{t-}/X^{t-}) = \mathrm{H}(X(t)/X^{t-}) - \mathrm{H}(X(t)/X^{t-}, Y^{t-}).$$

Aplicando-se a Proposição 7.1.1 segue o resultado. □

Infelizmente não é claro como representar a medida de causalidade de Granger como uma taxa de informação mútua, o que possibilitaria se obter uma expressão baseada na representação espectral. No artigo de Geweke (1982), pode-se verificar uma expressão baseada nas densidades espectrais que limita inferiormente a medida de causalidade de Granger e que é denominada medida de retroalimentação no domínio da freqüência por Geweke. Geweke argumenta no seu artigo que na maioria dos casos ocorre a igualdade. Este argumento é motivo de controvérsia e confusão na literatura. De fato, Hosoya (1991) resolveu parcialmente a controvérsia sobre o significado da medida de retroalimentação no

domínio da freqüência, identificando as condições necessárias e suficientes em que a afirmação de Geweke é válida, embora as condições obtidas não sejam passíveis de checagem em geral utilizando as representações MM e AR do processo. O trabalho de Hosoya (1991), embora baseado no trabalho de Geweke (1982), procurou obter os resultados em condições mais gerais que aquelas consideradas por Geweke, em particular, não foi considerado que as séries fossem gaussianas. Dessa forma, embora interessantes, os resultados de Hosoya não dizem respeito às quantidades da Teoria da Informação e não é claro pela apresentação de Hosoya como obter as medidas de dependência propostas no trabalho no contexto da Teoria da Informação.

O que segue são resultados que clarificam a relação entre as medidas propostas por Geweke (1982) e Hosoya (1991) com a Teoria da Informação. Logo em seguida são obtidas generalizações para os processos inversos a partir do que se deriva a interpretação para a coerência parcial direcionada no contexto da Teoria da Informação, isto é, a coerência parcial direcionada é a medida de dependência de fluxo de informação obtida ao se substituir os processos pelos seus processos inversos. As generalizações destas medidas são obtidas assim como as suas interpretações.

Definição 7.3.2 (Medida de fluxo de informação de Hosoya)**.** Sejam X e Y séries conjuntamente estacionárias em senso estrito não necessariamente unidimensionais. Sejam η_x e η_y séries estacionárias tais que $\eta_x(t) = R(X(t)/X^{t-}, Y(t), Y^{t-})$ e $\eta_y(t) = R(Y(t)/X(t), X^{t-}, Y^{t-})$. A *medida de fluxo de informação* de Y para X é definida como

$$\text{TIM}(X : \eta_y), \tag{7.34}$$

quando o limite existir.

Observe que, na definição acima, η_x, η_y são os resíduos da esperança condicional e não da projeção ortogonal linear. Também note que η_x, η_y são de fato processos conjuntamente estacionários. No caso gaussiano pode-se obter expressões para a medida de fluxo de informação parametrizadas pelos coeficientes da representação MM e AR.

Observação 7.3.1. A definição acima de medida de fluxo de informação de Hosoya é motivada pelas seguintes observações.

A v.a. $\eta_y(t)$ é ortogonal em relação ao passado de X, isto é, $\text{Cov}(\eta_y(t) : X(s)) = 0$ para todo $s \leq t$. Logo, $\text{TIM}(X : \eta_y)$ é a medida da taxa de informação entre o processo X e o passado do processo η_y. O processo η_y é a parte de Y livre de retroalimentação, pois é a parte de Y que não é devido ao passado de Y e nem ao passado e presente de X.

Ainda, como discutido no Capítulo 6, Hosoya (1991) definiu a seguinte medida de fluxo de informação, denominada por ele "measure of one-way effect", para processos estacionários de segunda ordem não necessariamente

univariados e não necessariamente gaussianos X e Y:

$$\log \frac{\mathrm{Var}(X(t)/X^{t-})}{\mathrm{Var}(X(t)/\eta_y^{t-})},$$

em que $\eta_y(t) = \bar{R}(Y(t)/X(t), X^{t-}, Y^{t-})$. Observe que $\eta_y(t)$ definido aqui é o resíduo da projeção ortogonal linear de $Y(t)$ sobre o presente e passado de X e passado de Y. Pela Proposição 7.1.1, pode-se observar que quando X e Y são processos gaussianos estacionários

$$\begin{aligned}\log \frac{\mathrm{Var}(X(t)/X^{t-})}{\mathrm{Var}(X(t)/X^{t-}, \eta_y^{t-})} &= 2(H(X(t)/X^{t-}) - H(X(t/X^{t-}, \eta_y^{t-}))) \\ &= 2\mathrm{IM}(X(t) : \eta_y^{t-}/X^{t-}),\end{aligned}$$

que é a quantidade (6.8) discutida no Capítulo 6. Note que Hosoya (1991) não estabelece a relação da medida proposta por ele e a informação mútua, pois ele não supõe que o processo seja gaussiano.

Um resultado importante em Hosoya (1991) é a demonstração da seguinte identidade para processos estacionários de segunda ordem:

$$\log \frac{\mathrm{Var}(X(t)/X^{t-})}{\mathrm{Var}(X(t)/\eta_y^{t-})} = -\frac{1}{2\pi}\int_{-\pi}^{\pi} \log \frac{\det \mathrm{Var}(dZ_\zeta(\lambda))}{\det \mathrm{Var}(dZx(\lambda))} d\lambda,$$

em que $\zeta(t) = \bar{R}(X(t)/\eta_y)$, ou seja, é o resíduo de $X(t)$ projetado sobre todo processo η_y.

Agora, pela definição de espectro parcial, tem-se que

$$\mathrm{Var}(dZ_\zeta(\lambda)) = \mathrm{Var}(dZ_x(\lambda)/dZ_{\eta_y}(\lambda))$$

e portanto

$$\begin{aligned}
&\det \mathrm{Var}(dZ_\zeta(\lambda)) \det \mathrm{Var}(dZ_{\eta_y}(\lambda)) \\
&= \det \mathrm{Var}(dZ_x(\lambda)/dZ_{\eta_y}(\lambda)) \det \mathrm{Var}(dZ_{\eta_y}(\lambda)) \\
&= \det \mathrm{Var}(dZ_x(\lambda), dZ_{\eta_y}(\lambda)).
\end{aligned}$$

Assim

$$\log \frac{\mathrm{Var}(X(t)/X^{t-})}{\mathrm{Var}(X(t)/\eta_y^{t-})} = -\frac{1}{2\pi}\int_{-\pi}^{\pi} \log \frac{\det \mathrm{Var}(dZ_x(\lambda), dZ_{\eta_y}(\lambda))}{\det \mathrm{Var}(dZ_x(\lambda)) \det \mathrm{Var}(dZ_{\eta_y}(\lambda))} d\lambda.$$

Finalmente, assumindo que X e Y sejam processos gaussianos estacionários, η_y também será um processo gaussiano estacionário e, portanto, utilizando a Proposição 7.2.2, tem-se

$$\begin{aligned}
&2\mathrm{IM}(X(t) : \eta_y^{t-}/X^{t-}) \\
&= \log \frac{\mathrm{Var}(X(t)/X^{t-})}{\mathrm{Var}(X(t)/\eta_y^{t-})} \\
&= -\frac{1}{2\pi}\int_{-\pi}^{\pi} \log \frac{\det \mathrm{Var}(dZ_x(\lambda), dZ_{\eta_y}(\lambda))}{\det \mathrm{Var}(dZ_x(\lambda)) \det \mathrm{Var}(dZ_{\eta_y}(\lambda))} d\lambda \\
&= 2\mathrm{TIM}(X(t) : \eta_y).
\end{aligned}$$

Esta última identidade juntamente com o argumento intuitivo do começo desta observação justificam a introdução da Definição 7.3.2.

Proposição 7.3.2. *Sejam X e Y séries conjuntamente estacionárias e gaussianas não necessariamente unidimensionais que satisfazem a condição de limitação conjuntamente. Seja a representação MM*

$$\begin{bmatrix} X(t) \\ Y(t) \end{bmatrix} = \sum_{k=0}^{\infty} \begin{bmatrix} H_{xx}(k) & H_{xy}(k) \\ H_{yx}(k) & H_{yy}(k) \end{bmatrix} \begin{bmatrix} \xi_x(t-k) \\ \xi_y(t-k) \end{bmatrix}. \tag{7.35}$$

Seja $\tilde{H}$ o símbolo MM de X. Sejam η_x e η_y séries estacionárias gaussianas tais que $\eta_x(t) = \bar{R}(X(t)/X^{t-}, Y(t), Y^{t-}) = \bar{R}(\xi_x(t)/\xi_y(t))$ e $\eta_y(t) = \bar{R}(Y(t)/X(t), X^{t-}, Y^{t-}) = \bar{R}(\xi_y(t)/\xi_x(t))$. A medida de fluxo de informação $TIM(X : \eta_y(t))$ de Y para X pode ser calculada como

$$TIM(X : \eta_y) = -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log(1 - CQT(dZ_x(\lambda) : dZ_{\eta_y}(\lambda)))d\lambda \tag{7.36}$$

$$= -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log\det(I - f_x^{-1/2}(\lambda)\tilde{H}_{xy}(\lambda)\, Var(\xi_y(t)/\xi_x(t))\tilde{H}_{xy}(\lambda)^* f_x^{-1/2}(\lambda))d\lambda. \tag{7.37}$$

Demonstração. A identidade (7.36) é imediata pela Proposição 7.2.3. A identidade (7.37) é obtida calculando-se

$$\begin{aligned} &1 - \mathrm{CQT}(dZ_x(\lambda) : dZ_{\eta_y}(\lambda)) \\ &= \det(I - f_x(\lambda)^{-1/2} f_{x\eta_y}(\lambda) f_{\eta_y}^{-1} f_{\eta_y} f_{\eta_y}^{-1} f_{\eta_y x}(\lambda) f_x(\lambda)^{-1/2}). \end{aligned}$$

Agora, como

$$H_{xy}(\lambda) = f_{x\eta_y}(\lambda) f_{\eta_y}^{-1},$$

segue o resultado. □

A generalização para o caso de mais de duas séries é imediata.

Definição 7.3.3 (Generalização da medida de fluxo de informação de Hosoya para mais de duas séries)**.** Sejam $X_1, \ldots, X_n$ séries conjuntamente estacionárias em senso estrito não necessariamente unidimensionais. Sejam $\eta_1, \ldots, \eta_n$ séries estacionárias tais que $\eta_k(t) = R(X_k(t)/X_k^{t-}, X^k(t), (X^k)^{t-}),\ k = 1, \ldots, n$, em que X^k é a série formada pelas séries $X_l, l \neq k$. A *medida de fluxo de informação* de X_q para X_p é definida como

$$\mathrm{TIM}(X_p : \eta_q), \tag{7.38}$$

quando o limite existir.

Note novamente que $\eta_k,\ k = 1, \ldots, n$, são so resíduos da esperança condicional e não da projeção linear ortogonal.

Proposição 7.3.3. *Sejam $X_1, \ldots, X_n$ séries conjuntamente estacionárias e gaussianas não necessariamente unidimensionais que satisfazem a condição de limitação cojuntamente. Seja a representação MM*

$$\begin{bmatrix} X_1(t) \\ \vdots \\ X_n(t) \end{bmatrix} = \sum_{k=0}^{\infty} \begin{bmatrix} H_{11}(k) & \ldots & H_{1n}(k) \\ \vdots & \ddots & \vdots \\ H_{n1}(k) & \ldots & H_{nn}(k) \end{bmatrix} \begin{bmatrix} \xi_1(t-k) \\ \vdots \\ \xi_n(t-k) \end{bmatrix} \tag{7.39}$$

e $\tilde{H}$ o símbolo MM de $X^T = [X_1^T \ldots X_n^T]$. Sejam $\eta_1, \ldots, \eta_n$ séries estacionárias gaussinas tais que $\eta_k(t) = R(X_k(t)/X_k^{t-}, X^k(t), (X^k)^{t-}) = \bar{R}(\xi_k(t)/\xi^k(t)), k = 1, \ldots, n$, em que X^k é a série formada pelas séries $X_l, l \neq k$ e ξ^k é a série formada por $\xi_l, l \neq k$. A medida de fluxo de informação $TIM(X_p : \eta_q)$ de X_q para X_p pode ser calculada como

$$TIM(X_p : \eta_q) = -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log(1 - CQT(dZ_{x_p}(\lambda) : dZ_{\eta_q}(\lambda)))d\lambda \tag{7.40}$$

$$= -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log \det(I - f_{x_p}^{-1/2}(\lambda)\tilde{H}_{pq}(\lambda)\, Var(\xi_q(t)/\xi^q(t))\tilde{H}_{pq}(\lambda)^* f_{x_p}^{-1/2}(\lambda))d\lambda. \tag{7.41}$$

Demonstração. A identidade (7.40) é imediata pela Proposição 7.2.3. A identidade (7.41) é obtida calculando-se

$$\begin{aligned} &1 - \mathrm{CQT}(dZ_p(\lambda) : dZ_{\eta_q}(\lambda)) \\ &= \det(I - f_{x_p}(\lambda)^{-1/2} f_{x_p\eta_q}(\lambda) f_{\eta_q}^{-1} f_{\eta_q} f_{\eta_q}^{-1} f_{\eta_q x_p}(\lambda) f_{x_p}(\lambda)^{-1/2}). \end{aligned}$$

Agora, como

$$H_{pq}(\lambda) = f_{x_p\eta_q}(\lambda) f_{\eta_q}^{-1},$$

segue o resultado. □

Esta última proposição pode ser interpretada como uma generalização de algumas medidas de dependência entre séries temporais estacionárias propostas na literatura. Para isto, note que, no caso em que as séries $X_1, \ldots, X_n$ são univariadas, a identidade (7.41) assume uma expressão simplificada, pois

$$\mathrm{CQT}(dZ_p(\lambda) : dZ_{\eta_q}(\lambda)) = \frac{|H_{pq}(\lambda)|^2 \mathrm{Var}(\xi_q(t)/\xi^q(t))}{f_{x_p}(\lambda)}.$$

Nestas condições, $|H_{pq}(\lambda)|^2 \mathrm{Var}(\xi_q(t)/\xi^q(t)) f_x(\lambda)^{-1}$ é o módulo quadrático *coerência direcionada* de X_q para X_p na freqüência $\lambda \in [-\pi, \pi)$ quando a matriz de variância/covariância dos resíduos $\mathrm{Var}(\xi_1(t)), \ldots, \xi_n(t)$ é diagonal (Baccalá et al., 1999). Ainda, quando $\mathrm{Var}(\xi_1(t)), \ldots, \xi_n(t)$ é a matriz identidade, a quantidade $|H_{pq}(\lambda)|^2 \mathrm{Var}(\xi_q(t)/\xi^q(t)) f_{x_p}(\lambda)^{-1}$ é o módulo quadrático da *função de transferência direcionada* introduzida em Kaminski e Blinowska (1991). É interessante que as expressões para medidas de fluxo de informação, embora motivadas de formas distintas, apresentem relações explícitas entre elas.

Esta última observação motiva a introdução da seguinte medida de fluxo de informação no domínio da freqüência:

Definição 7.3.4 (Coerência direcionada quadrática)**.** Sejam $X_1, \ldots, X_n$ séries estacionárias de segunda ordem que satisfazem a condição de limitação conjuntamente. Sejam $\eta_1, \ldots, \eta_n$ séries estacionárias de segunda ordem tais que $\eta_k(t) = R(X_k(t)/X_k^{t-}, X^k(t), (X^k)^{t-}), k = 1, \ldots, n$, em que X^k é a série formada pelas séries $X_l, l \neq k$. A coerência direcionada quadrática baseada na Teoria da Informação CDQ^{TI} de X_q para X_p na freqüência $\lambda \in [-\pi, \pi)$ é definida por[a]

$$\mathrm{CDQ}_{pq}^{TI}(\lambda) = \mathrm{CQT}(dZ_{x_p}(\lambda) : dZ_{\eta_q}(\lambda)).$$

[a]O sobrescrito TI indica que é a versão relacionada à Teoria da Informação para diferenciar da coerência direcionada (CD) definida em (Baccalá et al., 1999).

É importante salientar que tanto Geweke (1982) como Hosoya (1991) introduziram as medidas de dependência apenas para o caso de duas séries não necessariamente univariadas baseando-se na representação espectral dos processos, sem explicitar a relação com quantidades da Teoria da Informação. Também é importante notar que ambos os autores generalizaram as medidas de fluxo de informação propostas para o caso de mais de duas séries, porém as generalizações obtidas são distintas da Definição 7.3.3, mesmo no caso gaussiano.

De fato, Geweke (1984) e Hosoya (2001) generalizaram as medidas de fluxo de informação entre duas séries para o caso de mais de duas séries, obtendo generalizações distintas. As generalizações sugeridas são apresentadas no apêndice por não ser o foco principal da tese.

Uma questão importante é definir uma medida de fluxo de informação para as séries inversas, isto é,

Definição 7.3.5. Sejam $X_1, \ldots, X_n$ séries conjuntamente estacionárias em senso estrito não necessariamente unidimensionais que satisfazem a condição de limitação e ${}^iX_1, \ldots, {}^iX_n$ as suas séries inversas. Sejam $\eta_1, \ldots, \eta_n$ séries estacionárias tais que $\eta_k(t) = R(X_k(t)/X_k^{t-}, X^k(t), (X^k)^{t-}), k = 1, \ldots, n$, em que X^k é a série formada pelas séries $X_l, l \neq k$. As séries ${}^i\eta_1, \ldots, {}^i\eta_n$ são as suas inversas. A medida inversa de fluxo de informação de X_q para X_p é definida como

$$\mathrm{TIM}({}^iX_p :^i \eta_q), \tag{7.42}$$

quando o limite existir.

Proposição 7.3.4. *Sejam $X_1, \ldots, X_n$ séries conjuntamente estacionárias e gaussianas não necessariamente unidimensionais que satisfazem a condição de limitação conjuntamente. Seja a representação AR*

$$\begin{bmatrix} X_1(t) \\ \vdots \\ X_n(t) \end{bmatrix} = \sum_{k=1}^{\infty} \begin{bmatrix} A_{11}(k) & \ldots & A_{1n}(k) \\ \vdots & \ddots & \vdots \\ A_{n1}(k) & \ldots & A_{nn}(k) \end{bmatrix} \begin{bmatrix} X_1(t-k) \\ \vdots \\ X_n(t-k) \end{bmatrix} + \begin{bmatrix} \xi_1(t) \\ \vdots \\ \xi_n(t) \end{bmatrix} \tag{7.43}$$

e $\tilde{A}$ o símbolo AR de $X^T = [X_1^T \ldots X_n^T]$. Sejam $\eta_1, \ldots, \eta_n$ séries estacionárias gaussinas tais que $\eta_k(t) = R(X_k(t)/X_k^{t-}, X^k(t), (X^k)^{t-} = \bar{R}(\xi_k(t)/\xi^k(t)), k = 1, \ldots, n$, em que X^k é a série formada pelas séries X_l, $l \neq k$ e ξ^k é a série formada por ξ_l, $l \neq k$. Ainda, defina $\epsilon_k(t) = \bar{R}(X_k(t)/X^k)$, $k = 1, \ldots, n$. A medida inversa de fluxo de informação $TIM({}^iX_p :^i \eta_q)$ de X_q para X_p pode ser calculada como

$$TIM({}^iX_p :^i \eta_q) = TIM(\epsilon_p : \xi_q) \tag{7.44}$$

$$= -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log(1 - CQT(\bar{R}(dZ_{x_p}(\lambda)/dZ_{x^p}(\lambda)) : dZ_{\xi_q}(\lambda)))d\lambda \tag{7.45}$$

$$= -\frac{1}{4\pi} \int_{-\pi}^{\pi} \log\det(I - f_{x_p/x^p}^{1/2}(\lambda)\tilde{A}_{pq}(\lambda)^* Var(\xi_p(t))^{-1}\tilde{A}_{pq}(\lambda)f_{x_p/x^p}^{1/2}(\lambda))d\lambda, \tag{7.46}$$

em que f_{x_p/x^p} é a função de densidade espectral de X_p parcializada por X^p, isto é,

$$f_{x_p/x^p}(\lambda) = Var(dZ_{x_p}(\lambda)/dZ_{x^p}(\lambda)) \tag{7.47}$$

$$= f_{x_p}(\lambda) - f_{x_p x^p}(\lambda)f_{x^p}(\lambda)^{-1}f_{x^p x_p}(\lambda). \tag{7.48}$$

Demonstração. A identidade (7.45) segue das Proposições 4.2.13 e 7.2.3, pois

$$\text{CQT}(d^i Z_{x_p}(\lambda) : \bar{R}(d^i Z_{\xi_q}(\lambda)/d^i Z_{\xi^q}(\lambda))) \tag{7.49}$$

$$= \text{CQT}(\bar{R}(dZ_{x_p}(\lambda)/dZ_{x^p}(\lambda)) : dZ_{\xi_q}(\lambda)). \tag{7.50}$$

A identidade (7.44) segue da identidade (7.45).

Para obter (7.46) observe que pela equação (4.63)

$$\begin{aligned}
&1 - \text{CQT}(d^i Z_{x_p}(\lambda) : \bar{R}(d^i Z_{\xi_q}(\lambda)/d^i Z_{\xi^q}(\lambda)) \\
&= \det(I - \text{Var}(dZ_{x_p}/dZ_{x^p})^{-1}\text{Cov}(\bar{R}(dZ_{x_p}/dZ_{x^p}) : dZ_{\xi_q})\text{Var}(Z_{\xi_q})^{-1} \\
&\text{Cov}(dZ_{\xi_q} : \bar{R}(dZ_{x_p}/dZ_{x^p}))\text{Var}(dZ_{x_p}/dZ_{x^p})^{-1}\text{Var}(dZ_{x_p}/dZ_{x^p})) \\
&= \det(I - f^{1/2}_{x_p/x^p}(\lambda)\tilde{A}_{pq}(\lambda)^*\text{Var}(\xi_p(t))^{-1}\tilde{A}_{pq}(\lambda)f^{1/2}_{x_p/x^p}(\lambda)),
\end{aligned}$$

em que a útima igualdade é devido ao fato de

$$\begin{aligned}
\tilde{A}_{pq}(\lambda) &= \text{Var}(dZ_{x_p}/dZ_{x^p})^{-1}\text{Cov}(dZ_{x_p} : dZ_{\xi_q}/dZ_{x^p}) \\
&= \text{Var}(dZ_{x_p}/dZ_{x^p})^{-1}\text{Cov}(\bar{R}(dZ_{x_p}/dZ_{x^p}) : dZ_{\xi_q}).
\end{aligned}$$

□

No caso em que as séries $X_1, \ldots, X_n$ são séries univariadas conjuntamente estacionárias de segunda ordem

$$\text{CQT}(\bar{R}(dZ_{x_p}(\lambda)/dZ_{x^p}(\lambda)) : dZ_{\xi_q}(\lambda)) = \frac{|A_{pq}(\lambda)|^2\text{Var}(\xi_q(t))^{-1}}{f_{x_p/x^p}(\lambda)},$$

e, neste caso, assim como para a medida de fluxo de informação (Definição

7.3.3), o caso em que a matriz de variância/covariância das inovações é diagonal é equivalente ao modulo quadrático da coerência parcial direcionada generalizada (Baccalá et al., 2007). No caso em que a matriz de variância/covariânca das invações é a matriz identidade obtém-se o módulo quadrático da coerência parcial direcionada introduzida em Baccalá e Sameshima (2001). Para uso futuro, esta última quantidade é definida a seguir.

Definição 7.3.6 (Coerência parcial direcionada de Baccalá e Sameshima (2001))**.** Sejam $X_1, \ldots, X_n$ séries univariadas estacionárias de segunda ordem que satisfazem a condição de limitação conjuntamente. Seja a representação AR

$$\begin{bmatrix} X_1(t) \\ \vdots \\ X_n(t) \end{bmatrix} = \sum_{k=1}^{\infty} \begin{bmatrix} A_{11}(k) & \ldots & A_{1n}(k) \\ \vdots & \ddots & \vdots \\ A_{n1}(k) & \ldots & A_{nn}(k) \end{bmatrix} \begin{bmatrix} X_1(t-k) \\ \vdots \\ X_n(t-k) \end{bmatrix} + \begin{bmatrix} \xi_1(t) \\ \vdots \\ \xi_n(t) \end{bmatrix} \tag{7.51}$$

e $\tilde{A}$ o símbolo AR de $X^T = [X_1^T \ldots X_n^T]$. A coerência parcial direcionada $\mathrm{CPD}_{pq}(\lambda)$ de X_q para X_p é definida como:

$$\mathrm{CPD}_{pq}(\lambda) = \frac{\tilde{A}_{pq}(\lambda)}{\sqrt{\sum_{k=1}^{n} |\tilde{A}_{kq}|^2}}.$$

Analogamente ao caso da medida de fluxo de informação, pode-se introduzir a seguinte medida inversa de fluxo de informação no domínio da freqüência:

Definição 7.3.7 (Coerência parcial direcionada quadrática)**.** Sejam $X_1, \ldots, X_n$ séries estacionárias de segunda ordem que satisfazem a condição de limitação conjuntamente. Sejam $\xi_1, \ldots, \xi_n$ as inovações, isto é, séries estacionárias de segunda ordem tais que $\xi_k(t) = R(X_k(t)/X^{t-}), k = 1, \ldots, n$, em que X é a série formada por todas as séries. A coerência parcial direcionada baseado na Teoria da Informação de X_q para X_p na freqüência $\lambda \in [-\pi, \pi)$ é definida por[a]

$$\mathrm{CPD}_{pq}^{TI}(\lambda) = \mathrm{CQT}(\bar{R}(dZ_{x_p}(\lambda)/dZ_{x^p}(\lambda)) : dZ_{\xi_q}(\lambda)).$$

[a]O sobrescrito TI indica que é a versão relacionada à Teoria da Informação para diferenciar da coerência parcial direcionada (CPD) definida em (Baccalá e Sameshima, 2001).

Observe que as Definições 7.3.4 e 7.3.7 assumem apenas que as séries sejam estacionárias de segunda ordem e que satisfaçam a condição de limitação. De fato, a definição faz sentido exigindo apenas que a matriz de densidades espectrais conjunta dos processos seja inversível e sua inversa seja absolutamente integrável (vide Rozanov (1967) para o significado desta condição). No caso de dados neurofisiológicos, diferentes faixas de freqüências estão associadas a diferentes fenômenos biológicos, portanto medidas de dependência no domínio da freqüência desempenham papel importante na interpretação dos resultados de análise de dependência entre diferetens áreas neurais.

7.4 Conclusão

Os resultados obtidos para v.as. de dimensões finitas no Capítulo 4 foram generalizados para o caso de séries estacionárias de segunda ordem multivariada. Quando as séries são gaussianas as medidas propostas apresentam naturalmente interpretação como taxa de informação mútua entre séries estritamente estacionárias.

No caso de processos estacionários de segunda ordem, as medidas propostas neste capítulo podem ser parametrizadas pelos coeficientes da representação AR e/ou MM. No caso das medidas de fluxo de informação e das medidas inversas associadas, demonstrou-se que a parametrização está relacionada às medidas de fluxo de informação previamente introduzidas na literatura, relacionando estas medidas sob o mesmo formalismo matemático.

A técnica desenvolvida neste capítulo permite a sistematização do estudo de algumas medidas de dependência linear para processos estacionários de segunda ordem, em particular para processos estacionários gaussianos. Uma vantagem da técnica deste capítulo é a possibilidade de se obter uma representação no domínio da freqüência de medidas definidas no domínio do tempo e vice-versa, o que permite a extensão quase que imediata de métodos multivariados desenvolvidos para v.as. para o caso de séries temporais estacionárias de segunda ordem.

CAPÍTULO 8

Exemplos

Neste capítulo são apresentadas aplicações das medidas de dependência para séries temporais discutidas no Capítulo 7 para dados simulados e empíricos. A ênfase é dada à coerência parcial direcionada quadrática para ilustrar algumas de suas propriedades. Nos dois primeiros exemplos são apresentados dois modelos que salientam as diferenças entre as medidas de fluxo de informação (Definição 7.3.3) e sua inversa (Definição 7.3.4), ou equivalentemente, entre a coerência direcionada quadrática (Definição 7.3.4) e coerência parcial direcionada quadrática (Definição 7.3.7). O terceiro exemplo é uma aplicação da coerência parcial direcionada quadrática (Definição 7.3.7) em dados obtidos de um camundongo normal e um com hiperdopaminergia. Este último exemplo ilustra algumas conclusões que se pode obter aplicando-se as medidas discutidas no Capítulo 7 em dados neurofisológicos.

8.1 Uma modificação do Modelo 2 da subseção 6.0.2

O modelo considerado aqui é uma modificação do Modelo 2 introduzido na Subseção 6.0.2 e é definido a seguir.

Exemplo 8.1.1 (modificação do Modelo 2 da Subseção 6.0.2)**.** *Sejam X, Y e Z séries univariadas conjuntamente estacionárias e gaussianas com representação AR*

$$X(t) = -0.64Y(t-2) + 0.8Z(t-1) + \epsilon(t) \tag{8.1}$$

$$Y(t) = \xi(t) \tag{8.2}$$

$$Z(t) = 0.8Y(t-1) + \eta(t) \tag{8.3}$$

com $Var(\epsilon(t), \xi(t), \eta(t)) = I$ *e representação MM*

$$X(t) = \epsilon(t) + 0.8\eta(t-1) \tag{8.4}$$

$$Y(t) = \xi(t) \tag{8.5}$$

$$Z(t) = 0.8\xi(t-1) + \eta(t). \tag{8.6}$$

Pela Proposição 7.3.3, que relaciona os coeficentes da representação MM e a medida de fluxo de informação de Hosoya, fica claro que para o processo gerado pelo modelo acima, tem-se

$$\mathrm{TIM}(X : \zeta_y) = 0, \tag{8.7}$$

em que $\zeta_y(t) = \bar{R}(\xi(t)/\epsilon(t), \eta(t)) = \xi(t)$, pois os coeficientes da representação MM que relacionam ξ^{t-} e $X(t)$ são nulos.

Uma realização do modelo acima com 200 pontos para cada série foi gerada para ilustrar a afirmação acima. A partir dos dados foi estimado um modelo AR utilizando o algoritmo Nuttall-Strand (Schlögl, 2006). Os parâmetros do modelo AR estimados foram então utilizados para calcular as estimativas das coerências direcionadas quadráticas CDQ^{TI} entre as séries. Na Figura 8.1 está apresentado o resultado da estimação.

Agora, por outro lado, pela Proposição 7.3.4, que relaciona os coeficentes da representação AR e a medida inversa de fluxo de informação, tem-se

$$\text{TIM}(\epsilon_x : \xi_y) \neq 0, \tag{8.8}$$

em que $\epsilon_x(t) = \bar{R}(X(t)/Y, Z)$, pois o coeficiente que relaciona $Y(t-1)$ a $X(t)$ não é nulo.

A Figura 8.2 apresenta o resultado da estimação das coerências parciais direcionadas quadráticas CPDQ^{TI} entre as séries utilizando uma outra realização de 200 pontos para cada série do modelo 8.1.1.

É interessante notar que o teste de causalidade de Granger de Y para X consiste em verificar a nulidade dos coeficientes que relacionam $X(t)$ e Y^{t-} na representação AR (Lütkepohl, 1993), e, portanto, coincide com o resultado da medida inversa de fluxo de informação (8.8), mas não com o resultado da medida de fluxo de informação (8.7).

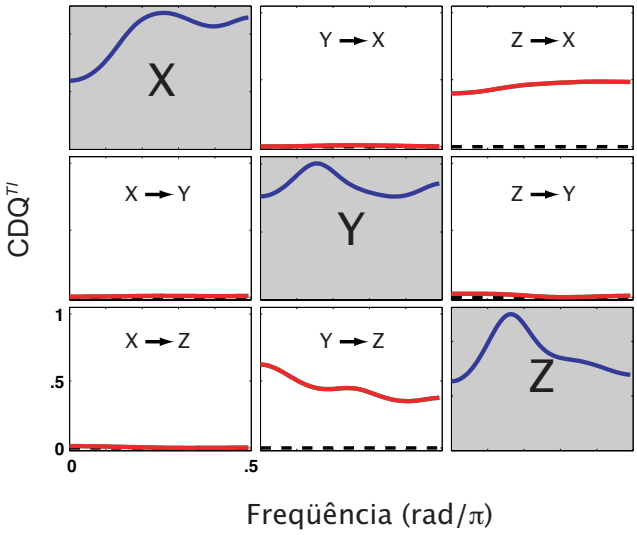


Figura 8.1: Coerência direcionada quadrática estimada para uma realização do modelo 8.1.1. Os quadros da diagonal principal são as densidades espectrais de X, Y e Z estimadas utilizando o modelo AR estimado, nesta ordem de cima para baixo. A linha tracejada preta representa o valor nulo. A linha contínua vermelha representa o valor da coerência direcionada quadrática estimada em cada freqüência.

8.2 O modelo “inverso” do modelo do Exemplo 8.1.1

No exemplo anterior foi apresentado um modelo em que ocorre a nulidade da coerência direcionada de Y para X para todas as freqüências, mas a coerência parcial direcionada não é nula em todas as freqüências. Aqui, é apresentado um modelo em que ocorre o inverso, isto é, a coerência direcionada de Y para X não é nula em todas as freqüências, mas a coerência parcial direcionada é nula em todas as freqüências.

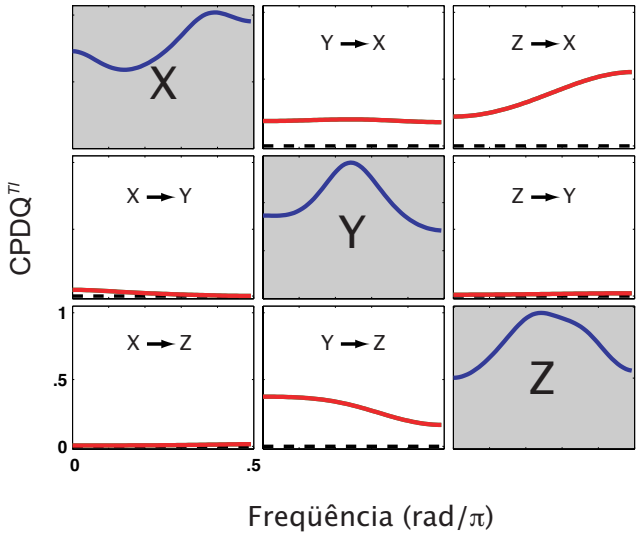


Figura 8.2: Coerência parcial direcionada quadrática estimada para uma realização do modelo 8.1.1. Os quadros da diagonal principal são as densidades espectrais de X, Y e Z estimadas utilizando o modelo AR estimado, nesta ordem de cima para baixo. A linha tracejada preta representa o valor nulo. A linha contínua vermelha representa o valor da coerência parcial direcionada quadrática estimada em cada freqüência.

Exemplo 8.2.1 (“Inverso” do modelo 2 do Exemplo 8.1.1)**.** *Sejam X, Y e Z séries univariadas conjuntamente estacionárias e gaussianas com representação AR*

$$X(t) = -0.8Z(t-1) + \epsilon(t) \tag{8.9}$$

$$Y(t) = \xi(t) \tag{8.10}$$

$$Z(t) = -0.8Y(t-1) + \eta(t) \tag{8.11}$$

com $Var(\epsilon(t), \xi(t), \eta(t)) = I$ e representação MM

$$X(t) = \epsilon(t) + 0.64\xi(t-2) - 0.8\eta(t-1) \tag{8.12}$$

$$Y(t) = \xi(t) \tag{8.13}$$

$$Z(t) = -0.8\xi(t-1) + \eta(t). \tag{8.14}$$

Como o modelo MM considerado acima é inversível, a representação AR acima é de fato estável (Lütkepohl, 1993).

A semelhança do modelo acima (Exemplo 8.2.1) e o modelo do Exemplo 8.1.1 é devido ao fato de uma ser obtida invertendo os coeficientes AR e MM do outro. Pode-se observar que

$$\mathrm{TIM}(\epsilon_x : \xi_y) = 0,$$

em que $\epsilon_x(t) = \bar{R}(X(t)/Y, Z)$, pois os coeficientes que relacionam $X(t)$ e Y^{t-} na representação AR são todos nulos.

Por outro lado,

$$\mathrm{TIM}(X : \zeta_y) \neq 0,$$

em que $\zeta_y(t) = \bar{R}(\xi(t)/\epsilon(t), \eta(t)) = \xi(t)$, pois o coeficiente que relaciona $X(t)$ e $\xi(t-2)$ é não nulo.

Uma realização do modelo 8.2.1 com 200 pontos para cada série foi gerada e os resultados das estimações das coerências direcionadas quadráticas são apresentadas na Figura 8.3. Observe a presença de fluxo de informação de Y para X.

Uma outra realização do modelo 8.2.1 com 200 pontos para cada série foi gerada e os resultados das estimações das coerencias parciais direcionadas são apresentados na Figura 8.4. Observe a ausência de fluxo de informação de Y para X.

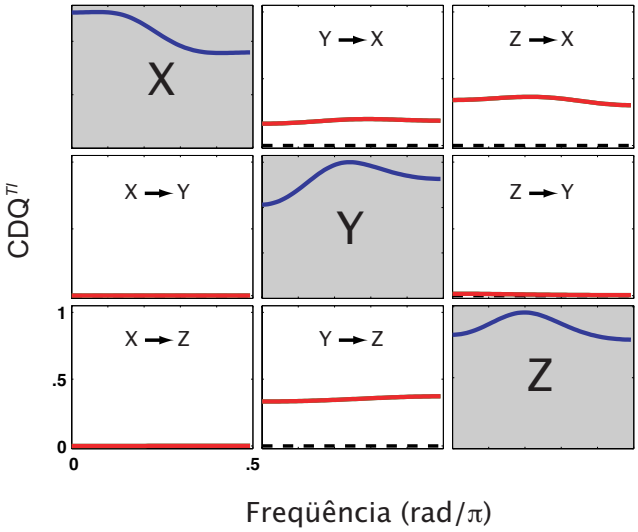


Figura 8.3: Coerência direcionada quadrática estimada para uma realização do modelo 8.2.1. Vide legenda da Figura 8.1.

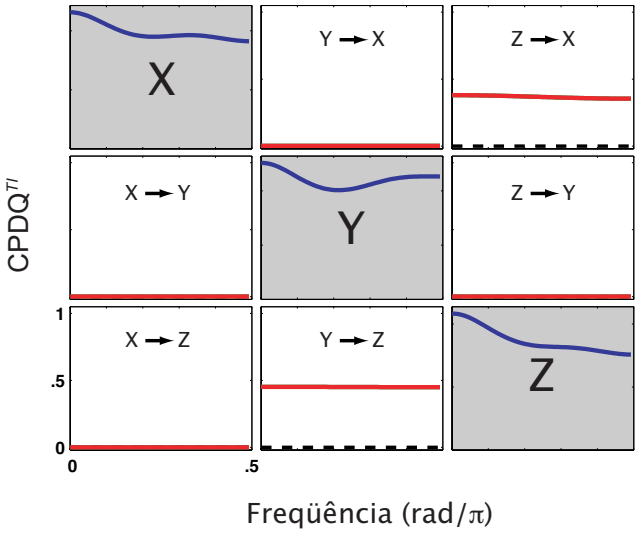


Figura 8.4: Coerência parcial direcionada quadrática estimada para uma realização do modelo 8.2.1. Vide legenda da Figura 8.2.

8.3 Camundongos hiperdopaminérgicos

A via de comunicação entre hipocampo e área pré-frontal desempenha um papel importante nas funções cognitivas de alto nível. A via dopaminérgica exerce influência crítica sobre as atividades no circuito hipocampo e córtex pré-frontal e as alterações dopaminérgicas tem sido apontadas como mediadoras da patogênesis de diversas doenças psiquiátricas como esquizofrenia e transtorno do déficit de atenção com hiperatividade. Utilizando um camundongo geneticamente modificado para apresentar hiperdopaminergia[1], e comparando-o a um camundongo controle normal, verifca-se alterações na dinâmica de interação entre hipocampo e área pré-frontal representadas pelas diferenças nas coerências parciais direcionadas entre os sinais de potencial de campo local registrados no hipocampo e córtex pré-frontal. Os mesmos dados são utilizados para calcular a coerência entre as áreas neurais.

Os dados utilizados nesta seção fazem parte de um conjunto de dados utilizados num trabalho realizado em colaboração com Kafui Dzirasa[2] que gerou o artigo (Dzirasa et al., 2008) submetido a uma revista internacional. A utilização dos resultados obtidos aqui foi realizada com o consentimento do autor principal do trabalho.

As Figuras 8.5 e 8.6 apresentam os resultados das estimações da coerência parcial direcionada quadrática baseada na Teoria da Informação (CPD^{TI}) e dos módulos quadráticos da coerência e da coerência parcial direcionada (CPD), definida em (Baccalá e Sameshima, 2001), entre os sinais de potencial de campo local registrados no hipocampo e córtex pré-frontal em um camundongo nor-

[1] Os camundongos hiperdopaminérgicos apresentam aumento persistente do nível de dopamina extracelular no cérebro.

[2] Department of Neurobiology, Duke University.

mal (Figura 8.5) e hiperdopaminérgico (Figura 8.6) realizando uma tarefa de memória espacial que se inicia após 60 segundos do início do registro do potencial de campo local.

Na Figura 8.5, observa-se que o módulo quadrático da coerência mostra uma interação entre as áreas aproximadamente constante ao longo do tempo na faixa de freqüência próxima a 8Hz, que é conhecida como banda de freqüência teta na literatura (Buzsáki, 2005) e tem sido correlacionado às tarefas que exigem memória espacial.

Os resultados obtidos na análise de coerência, embora sejam interessantes, não permitem inferir o sentido da interação, isto é, qual das estruturas está enviando informação. Já a coerência parcial direcionada quadrática mostra que há fluxo de informação tanto do hipocampo para o córtex pré-frontal como do córtex pré-frontal para o hipocampo, ou seja, existe retroalimentação, porém em freqüências distintas. A coerência parcial direcionada quadrática do hipocampo para o córtex é mais nítida na faixa de freqüência próxima a 8Hz, em acordo com o resultado observado pela coerência, já a coerência parcial direcionada quadrática do córtex para o hipocampo é mais nítida numa faixa de freqüência em torno de 4Hz, diferenciando da ausência de fluxo de informação do córtex para o hipocampo observado no animal normal controle.

O módulo quadrático da coerência parcial direcionada (Definição 7.3.6) foi calculada para comparação. Observa-se que a coerência parcial direcionada quadrática CPDQ^{TI} permite uma melhor apreciação do fato de não haver fluxo de informação do córtex pré-frontal para o hipocampo se comparada ao módulo quadrado da coerência parcial direcionada $|\mathrm{CPD}|^2$. Este último fato se deve essencialmente a não invariância quanto a escala dos sinais da coerência par-

cial direcionada definida em Baccalá e Sameshima (2001). Vide Baccalá et al. (2007) para uma discussão e solução deste fato. Note que a coerência parcial direcionada quadrática baseada na Teoria da Informação é invariante quanto a escala.

Na Figura 8.6, observa-se que a coerência mostra uma interação ao longo do tempo na faixa de freqüência próxima a 8Hz que se torna mais intensa a partir dos 60 segundos quando ocorre o início da tarefa de memória motora. Em comparação ao animal controle, observa-se que, no animal com hiperdopaminergia, o alto valor do módulo quadrático da coerência na faixa de freqüência próxima a 8Hz é mais sustentado ao longo do tempo. A coerência parcial direcionada quadrática baseada na Teoria da Informação torna claro que existe uma direcionalidade do hipocampo para o córtex pré-frontal na faixa de freqüência próxima a 8Hz, mas não no sentido oposto.

É interessante obsevar que o módulo quadrático da coerência parcial direcionada não torna esta última observação visual tão clara quanto a coerência parcial direcionada quadrática.

Um resultado interessante desta análise é o fato de a coerência parcial direcionada quadrática diferenciar claramente a dinâmica de interação entre o hipocampo e o córtex pré-motor de um camundongo normal e com hiperdopaminergia, enquanto que a coerência é elevada na faixa de freqüência em torno de 8Hz em ambos os camundongos, não permitindo uma diferenciação qualitativa tão nítida. Assim, fica claro, neste exemplo, que a inferência da interação entre áreas neurais associada a um conceito de fluxo de informação desempenha papel importante no entendimento da dinâmica do sistema nervoso.

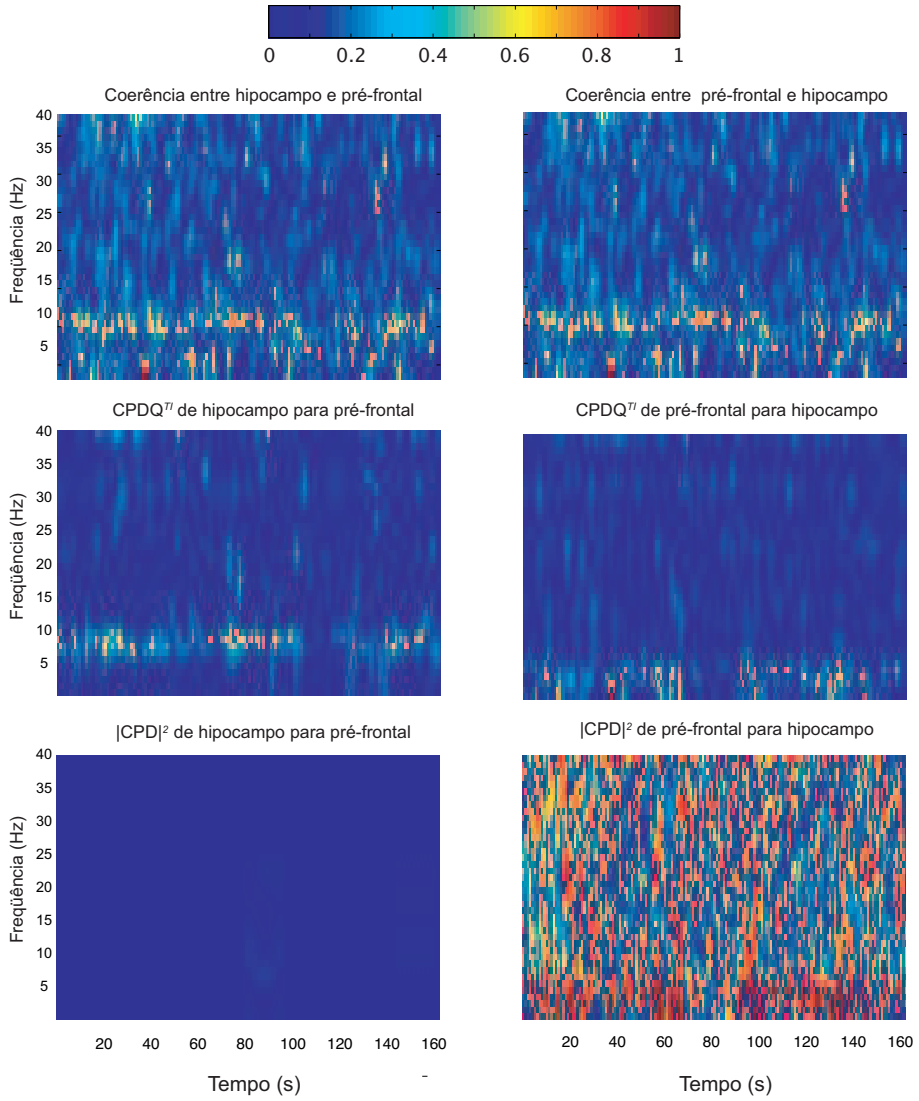


Figura 8.5: Resultado da análise de dados de camundongo normal controle. Cada quadro apresenta as estimativas do módulo quadrático da coerência, da coerência parcial direcionada quadrática e do módulo quadrado da coerência parcial direcionada (Definição 7.3.6), nesta ordem de cima para baixo. As cores representam os valores das estimativas num determinado tempo e freqüência.

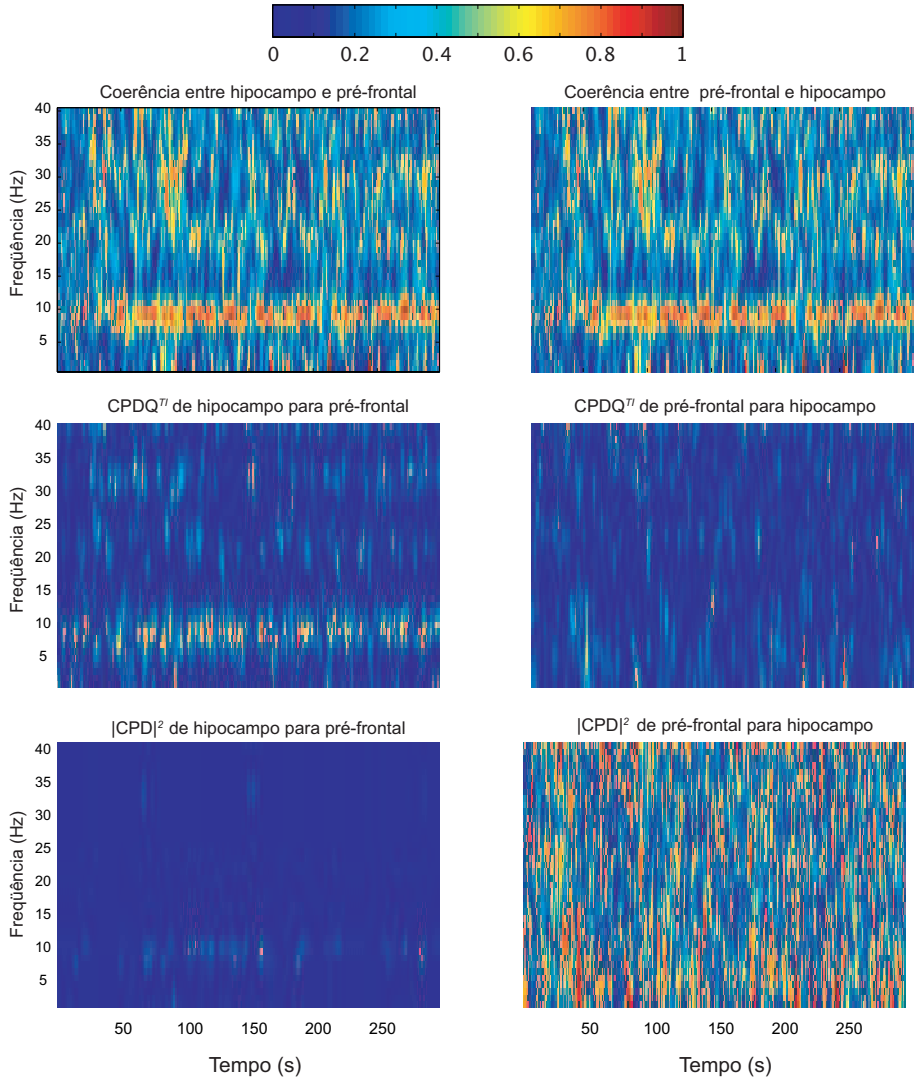


Figura 8.6: Resultado da análise de dados de camundongo hiperdopaminérgico. Vide legenda da Figura 8.5

8.4 Conclusão

A coerência direcionada quadrática e a coerência parcial direcionada quadrática apresentam propriedades distintas assim como a medida de fluxo de informação e a sua inversa apresentam interpretações distintas.

A interpretação da causalidade de Granger comumente empregada em Econometria (Lütkepohl, 1993), que consiste em verificar a nulidade dos coeficientes da representação AR, é compatível com a definição de medida inversa de fluxo de informação, mas não com a definição de medida de fluxo de informação de Hosoya. Este fato é conseqüência das Proposições 7.3.3 e 7.3.4 e dos Exemplos 8.1.1 e 8.2.1 apresentados. Exemplos de estimações para realizações geradas por cada modelo foram apresentadas.

Foi realizada uma aplicação da coerência parcial direcionada quadrática em dados experimentais comparando a dinâmica de interação entre o hipocampo e córtex pré-frontal de um camundongo normal controle e um hiperdopaminérgico. A aplicação ilustra possíves interpretações dos resultados das medidas de fluxo de informação discutidas no Capítulo 7 e é um exemplo de análise de dados dentro do novo paradigma da Neurociência que consiste no entendimento da interação de diferentes áreas neurais.

CAPÍTULO 9

Conclusão

Nesta tese foram explorados alguns conceitos e resultados da Teoria da Informação e processos gaussianos estacionários para se obter medidas de dependência entre séries temporais que, se forem adequadamente interpretadas, podem ser entendidas como medidas de fluxo de informação.

Os conceitos de v.as. e processos inversos desempenham papel fundamental na sistematização da construção de medidas de dependências. Assim, dada uma medida de dependência linear, sempre é possível obter o seu inverso que é simplesmente definido como sendo a mesma medida de dependência calculada sobre as v.as. ou processos inversos. Esta medida inversa, é, num certo sentido, a versão parcializada da medida original. Este fato, aparentemente simples, permite que se obtenham resultados sobre as relações entre diferentes medidas de dependência de forma sistematizada. Em particular, demonstrou-se o seguinte quadro de relações:

medida de dependência		inversa
correlação	$\longleftrightarrow$	correlação parcial
CQT	$\longleftrightarrow$	CQT parcial
matriz de correlação	$\longleftrightarrow$	matriz de correlação parcial
coerência	$\longleftrightarrow$	coerência parcial
função de transferência direcionada	$\longleftrightarrow$	coerência parcial direcionada
coerência direcionada	$\longleftrightarrow$	coerência parcial direcionada generalizada
medida de fluxo de informação	$\longleftrightarrow$	medida inversa de fluxo de informação
CDQ^{TI}	$\longleftrightarrow$	CPDQ^{TI}

Um aspecto importante do quadro de relações acima é a sua reflexividade, ou seja, dado uma medida de dependência, pode-se obter a sua inversa que por sua vez tem como inversa a medida de dependência inicial, ou seja, neste sentido, uma medida de dependência e sua inversa são duais.

Os métodos desenvolvidos nesta tese sugerem a sua aplicabilidade no estudo de outras medidas de dependência, além daquelas estudadas nesta tese e será tópico de estudos futuros.

Há pelo menos dois caminhos para a generalização dos resultados obtidos. O primeiro consiste na obtenção de resultados análogos aos obtidos nesta tese para processos estacionários não necessariamente gaussianos e, o segundo, consiste na generalização dos resultados para processos gaussianos não estacionários. Para o primeiro, a abordagem natural parece ser o estudo da Teoria da Informação e

a obtenção de um processo análogo ao processo inverso para séries gaussianas. Para a segunda generalização, o estudo dos processos harmonizáveis (Rao, 1984) parece ser uma alternativa promissora para se construir uma teoria de medidas lineares entre processos no domínio tempo-freqüência. Em ambos os casos a teoria existente ainda é incompleta e parece existir espaço para muito trabalho.

O fato de as Definições 7.3.4 e 7.3.7 se basearem nas representações espectrais dos processos permite que se generalize as medidas de diferentes formas. Em particular, nesta tese somente foram exploradas com certa generalidade as medidas de dependência linear entre séries estacionárias de segunda ordem, o que se reduz em muitos casos ao estudo das medidas de dependência linear entre os componentes espectrais dos processos numa mesma freqüência. As generalizações dos resultados obtidos aqui conduzem imediatamente ao estudo das medidas de dependência entre os componentes espectrais em freqüências distintas, que constituem tópicos a serem explorados com grande potencial de aplicabilidade.

Foi apresentado um exemplo de aplicação de algumas das medidas de dependência propostas nesta tese em dados neurofisiológicos. Foi mostrado, no exemplo, que os conceitos estudados aqui permitem que se obtenham novas interpretações para as relações de dependência entre diferentes áreas neurais, permitindo um melhor entendimento da dinâmica de interação no sistema nervoso.

Por fim, para a aplicação de dados empíricos é importante e, em muitos casos, necessário que se obtenha resultados estatísticos que garantam a aplicabilidade do método, o que se traduz na maioria dos casos na demonstração da consistência assintótica dos estimadores, além da obtenção das suas distribuições

assintóticas. Como observado no prefácio, foram obtidos alguns resultados neste sentido e parte deles já foram publicados em forma de artigos (Takahashi et al., 2008, 2007) e capítulo de livro (Baccalá et al., 2006). Estes resultados não foram discutidos aqui, porém, constituem uma parte importante do trabalho realizado e também a ser realizado.

REFERÊNCIAS BIBLIOGRÁFICAS

N.I. Akhiezer e I.M Glazman. *Theory of Linear Operator in Hilbert Space, Two Volumes Bound as One*. Dover: New York, 1993.

K. Baba, R. Shibata, e M. Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australlian & New Zealand Journal of Statistics*, 46: 657–664, 2004.

L. A. Baccalá e K. Sameshima. Partial directed coherence: A new concept in neural structure determination. *Biological Cybernetics*, 84: 463–474, 2001.

L. A. Baccalá, K Sameshima, G. Ballester, A. C. Valle, e C. Timo-Iaria. Studying the interaction between brain structures via directed coherence and Granger causality. *Applied Signal Processing*, 5: 40–48, 1999.

L. A. Baccalá, D. Y. Takahashi, e K. Sameshima. Generalized partial directed coherence. In *Cardiff Proceedings of the 2007 15th International Conference on Digital Signal Processing (DSP2007)*, pages 162–166, 2007.

L.A. Baccalá, D. Y. Takahashi, e K. Sameshima. Computer intensive testing

for the influence between time-series. in:*Handbook of Time Series Analysis*, ed: Bjorn Shelter, Jens Timmer and Matthias Winterhalder. pages 411–435. Wiley-VCH, 2006.

C. B. Bell. Mutual information and maximal correlation measures of dependence. *Annals of Mathematical Statistics*, 33: 587–595, 1962.

R. J. Bhansali. On a relationship between the inverse of a stationary covariance matrix and the linear interpolator. *Journal of Applied Probability*, 27: 156–170, 1990.

P. Billingsley. *Probability and Measure, 3ed.* John-Wiley & Sons: New York, 1995.

D. R. Brillinger. *Time Series: Data Analysis and Theory, Expanded Edition.* Holden-Day: San Francisco, 1981.

G. Buzsáki. Theta rythm of navigation: link between path integration and landmark navigation, episodic and semantic memory. *Hippocampus*, 15: 827–840, 2005.

P. Caines e C. Chan. Feedback between stationary stochastic processes. *IEEE Transactions on Automatic Control*, 20: 498–508, 1975.

R. Cheng e M. Pourahmadi. The mixing rate of a stationary multivariate process. *Journal of Theoretical Probability*, 6: 603–617, 1993.

T.M Cover e J.A. Thomas. *Information Theory.* Wiley: New Jersey, 1991.

F. Cucker e S. Smale. On the mathematical foundation of learning. *Bulletin of American Mathematical Society*, 39: 1–49, 2002.

K. Dzirasa, D. Y. Takahashi, J. Staplenton, R.R. Gainetdinov, M. Lavine, K. Sameshima, M. G. Caron, M. A. L. Nicolelis. Persistent hyperdopaminergia alters activity across the hippocampal-prefrontal pathway. *Submetido*, 2008.

R. L. Dobrushin. General formulation of Shannon's main theorem of information theory. *Usp. Mat. Nauk (in Russian). Translated in Amer. Math. Soc. Trans., vol. 33, pp. 323–438.*, 14: 3–104, 1959.

I.M. Gelfand e A.M. Yaglom. Calculation of amount of information about a random function contained in another such function. *American Mathematical Society Translation Series*, 2: 3–52, 1959.

J. F. Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association,*, 77: 304–313, 1982.

J. F. Geweke. Measures of conditional linear dependence and feedback between time series. *Journal of the American Statistical Association*, 79:907–915, 1984.

I. N. Goodman e D. H. Johnson. Orthogonal decomposition of multivariate statistical dependence measure. *Procceding of ICASSP*, pages 1017–1020, 2004.

C. W. J. Granger. Investigating causal relation by econometric models and cross-spectral methods. *Econometrica*, 37: 424–438, 1969.

T. S. Han. Multiple mutual informations and multiple interactions in frequency data. *Information and Control*, 46: 26–45, 1980.

E. Hannan. The general theory of canonical corrleation and its relation to

functional analysis. *Journal of Australian Mathematical Society*, 2: 229–242, 1961.

E. J. Hannan. *Multiple Time Series*. John Wiley & Sons Inc.: New York, 1970.

E. J. Hannan e M. Deistler. *The Statistical Theory of Linear Systems*. Wiley: New York, 1988.

H. Helson e D. Lowdenslager. Prediction theory and fourier series in several variables, Part I. *Acta Mathematica*, 99:165–202, 1958.

H. Helson e D. Lowdenslager. Prediction theory and fourier series in several variables, Part II. *Acta Mathematica*, 106:175–213, 1962.

K. Hlavácková-Schindlera, M. V. M. Palusb e J. Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441:1–46, 2007.

Y. Hosoya. Elimination of third-series effect and defining partial measures of causality. *Journal of Time Series Analysis*, 22:537–554, 2001.

Y. Hosoya. The decomposition and measurement of the interdependency between second-order stationary processes. *Probability Theory and Related Fields*, 88:429–444, 1991.

I. A. Ibragimov e Y. A. Rozanov. *Gaussian Random Processes*. Springer, 1978.

S. Ihara. *Information Theory for Continuous System*. World Scientific Publishing: Singapura, 1964.

R. L. Jenison e R. A. Reale. The shape of neural dependence. *Neural Computation*, 16:665–672, 2004.

G. Mercierand, S. Derrodeand, W. Pieczynskiand, J Nicolasand, A. Joannic-Chardin e J. Inglada. Copula-based stochastic kernels for abrupt change detection. Proceedings of IGARSS 06, pages 665–672, 2006.

H. Joe. *Multivariate Models and Dependence Concepts*. Chapman & Hall: London, 1997.

R. A. Johnson e D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 4 edition, 1998.

M.J. Kaminski e K.J. Blinowska. A new method of the description of the information flow in the brain structures. *Biological Cynernetics*, 65:203–210, 1991.

T. Kamitake, H. Harashima, e H. Miyakawa. A time-series analysis method based on the directed transinformation. *Electronics and Communications in Japan (Part I: Communications)*, 67:1–9, 2008.

A. N. Kolmogorov. Theory of transmission of information. *Session on Scientific Problems of Automatization in Industry, Plenary Talks, Izdat. Akad. Nauk SSSR, Moscow, English transl.*, 1:66–99, 1957.

S. Kotz, N. Balakrishnan, e N. L. Johnson. *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley, New York, 2000.

H. O. Lancaster. The structure of bivariate distribution. *Annals of Mathematical Statistics*, 29:719–736, 1958.

S. P. Lloyd. On measure of stochastic dependence. *Theory of Probability and its Applications*, 7:301–312, 1962.

M. Loéve. *Probability Theory II*. Springer-Verlag: New York, fourth edition, 1994.

H. Lütkepohl. *Introduction to Multiple Time Series Analysis*. Springer-Verlag: Berlin, 1993.

H. Marko. The bidirectional communication theory–a generalization of information theory. *IEEE Transaction on Communication*, 21: 1345–1351, Dec 1973.

P. R. Masani. The prediction theory of multivariate stochastic process, Part III. *Acta Mathematica*, 104:141–162, 1960.

P. R. Masani. The measure theoretic aspects of entropy, Part I. *Journal of Computational and Applied Mathematics*, 40:215–232, 1992a.

P. R. Masani. The measure theoretic aspects of entropy, Part II. *Journal of Computational and Applied Mathematics*, 44:245–260, 1992b.

J. L. Massey e P. C. Massey. Conservation of mutual and directed information. In *Proceedings International Symposium on Information Theory ISIT 2005*, 157–158, 2005. doi: 10.1109/ISIT.2005.1523313.

K. Matsumoto e I. Tsuda. Calculation of information flow rate from mutual information. *Journal of Physics A: Mathematical and General*, 21:1405–1414, 1988.

R. Nelsen. *An Introduction to Copulas*. Springer: New York, 1999.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press: Cambridge, 2000.

M. S. Pinsker. *Information and Information Stability of Random Variables and Processes*. Holden-Day: San Francisco, 1964.

M. M. Rao. The spectral domain of multivariate harmonizable processes. *Proceedings of the National Academy of Sciences of the United States of America*, 81:4611–4612, 1984.

A. Raveh. On the use of the inverse of the correlation matrix in multivariate data analysis. *The American Statistian*, 39: 39–42, 1985.

A. Rényi. On measures of dependence. *Acta Mathematica Hungarica*, 10: 441–451, 1959.

J. L. Rodgers e W. A. Nicewander. Thirteen ways to look at correlation coefficients. *The American Statisticians*, 42: 59–66, 1988.

M. Ronsenblatt. *Markov Process: Structure and Asymptotic Behavior*. Springer: Berlin, 1971.

Y. Rozanov. *Stationary Random Process*. Holden-Day: San Francisco, 1967.

K. Sameshima e L. A. Baccalá. Using partial directed coherence to describe neuronal ensemble interactions. *Journal of Neuroscience Methods*, 94:93–103, 1999.

A. Schlögl A comparison of multivariate autoregressive estimators. *Signal Processing*, 86:2426–2429, 2006.

V. Schmitz. *Copulas and Stochastic Processes*. PhD thesis, Institute of Statistics of Aachen University, 2003.

T. Schreiber. Measuring information transfer. *Physical Review Letter*, 85:461–464, 2000.

C. E. Shannon e W. Weaver. *The Mathematical Theory of Communication*. The Univeristy Of Illinois Press: Illinois, 1949.

C. A. Sims. Money, income, and causality. *The American Economic Review*, 62:540–552, 1972.

A. Sklar. Fonctions de repartition n dimensions et leurs marges. *Publ Inst Statist Univ Paris*, 8:229–231, 1959.

A. R. Soltani e M. Mohammadpour. Moving average representations for multivariate stationary processes. *Journal of Time Series Analysis*, 27:831–841, 2006.

D. Y. Takahashi, L.A. Baccalá e K. Sameshima. Connectivity inference via partial directed coherebce: asymptotic results. *Journal of Applied Statistics*, 34:1259–1273, 2007.

D. Y. Takahashi, L.A. Baccalá e K. Sameshima. Partial directed coherence asymptotics for VAR processes of infinite order. *International Journal of Bioelectromagnetism*, 10:31–36, 2008.

D. Y. Takahashi, L. Baccalá e K. Sameshima. On Granger causality e mutual information. In *Poster apresentado na 11 Escola Brasileira de Pobabilidade*, 2006.

M. Taniguchi e Y. Kakizawa. *Asymptotic Theory of Statistical Inference for Time Series*. Springer-Verlag: New York, 2000.

S. Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4:66–82, 1960.

N. Wiener e P. Masani. Prediction theory of multivariate stochastics processes, part I. *Acta Mathematica*, 98:111–150, 1957.

N. Wiener e P. Masani. Prediction theory of multivariate stochastics processes. part II. *Acta Mathematica*, 99:93–137, 1958.

A. D. Wyner. A definition of conditional mutual information for arbitrary ensembles. *Information and Control*, 38:51–59, 1978.

Anexos