Murilo Diniz Rocha Filho

Elementos de Probabilidade, Econometria e Aprendizagem de Máquinas com Aplicações em Séries Temporais Econômicas e Financeiras

Brasília

Julho de 2018

Murilo Diniz Rocha Filho

Elementos de Probabilidade, Econometria e Aprendizagem de Máquinas com Aplicações em Séries Temporais Econômicas e Financeiras

Monografia apresentada ao Departamento de Economia da Universidade de Brasília como requisito parcial para a obtenção do título de Bacharel em Economia.

Universidade de Brasília – UnB

Faculdade de Economia, Administração e Contabilidade

Departamento de Economia

Orientador: Dr. Daniel Oliveira Cajueiro

Brasília

Julho de 2018

Resumo

O presente trabalho tem por objetivo complementar a literatura sobre o vínculo entre os elementos de probabilidade, econometria e aprendizagem de máquinas que podem ajudar a responder questões referentes à previsão em series temporais econômicas e financeiras, sugerindo que uma abordagem robusta envolve a utilização complementar de modelos e metodologias distintas. Para tanto, foi realizado um estudo bibliográfico sobre a evolução da literatura que analisa os fundamentos matemáticos e probabilísticos de alguns dos principais modelos utilizados em economia e finanças. Mais especificamente, são apresentadas algumas das principais abordagens econométricas modernas, como vetores autoregressivos e modelos de heterocedasticidade condicional, bem como modelos lineares dinâmicos e a representação em espaço de estado.

Palavras-chave: Séries Temporais, Processos Estocásticos, Métodos Quantitativos.

Abstract

The present paper aims to complement the literature on the link between the elements of probability, econometrics and machine learning that can help answer questions related to economic and financial time series forecasting, suggesting that a robust approach involves the complementary use of models and different approaches. For this, a bibliographic study was carried out on the evolution of the literature that analyzes the mathematical and probabilistic fundamentals of some novel models with application in economics and finance. More specifically, some of the main modern econometric approaches are presented, such as vector autoregressions and conditional heteroscedasticity models, as well as dynamic linear models and the state space representation..

Keywords: Time Series, Stochastic Processes, Quantitative Methods.

Lista de ilustrações

Sumário

Introdução

De acordo com Tsay (2005), a análise de séries temporais financeiras preocupa-se com a teoria e prática da avaliação de ativos ao longo do tempo, sendo uma disciplina altamente empírica, cuja teoria forma as bases para a realização de inferências. Há no entanto uma diferença relevante entre a análise de séries temporais financeiras e outros tipos de análises de séries temporais: tanto a teoria financeira quanto as séries temporais empíricas contém em si um forte elemento de incerteza. Como resultado da incerteza, a teoria e métodos estatísticos desempenham um papel importante na análise de séries temporais financeiras (TSAY, 2005).

Campbell et al. (1997) afirmam que uma das questões mais duradouras no campo da econometria financeira é se os preços de ativos são previsíveis. Os autores afirmam que o campo de economia financeira moderna está fortemente enraizado nas tentativas de "bater o mercado", uma tarefa que ainda é de grande interesse, e alguns afirmam não ser possível.

A possibilidade de previsão dos preços dos ativos está diretamente relacionada ao conceito de "eficiência"de mercado. De acordo com Malkiel e Fama (1970), um mercado é dito eficiente com respeito a um conjunto de informações se os preços refletem completamente tal conjunto de informação. A eficiência de mercados é um tema controverso, tanto dentro da academia quanto dentre praticantes do mercado financeiro. De fato, a história sugere que "bater o mercado"é uma tarefa um tanto quanto desafiadora.

Apesar das críticas relativas à hipótese de mercados eficientes, é válido lembrar que a ciência está preocupada com a procura da melhor hipótese, e até que uma hipótese falha seja substituída por uma hipótese mais robusta, as críticas possuem valor limitado Sewell (2011). No presente trabalho são apresentados elementos de Probabilidade, Econometria e Aprendizagem de Máquinas que podem ajudar a responder as questões de previsibilidade dos preços de ativos e consequentemente de eficiência de mercado.

Mais especificamente, no primeiro capítulo são apresentadas as fundações conceituais e matemáticas da probabilidade, tratamos de alguns teoremas importantes da teoria da probabilidade e são introduzidos alguns conceitos relacionados à distribuições de probabilidade. No segundo capítulo, são introduzidos os conceitos de passeio aleatório e cadeias de Markov. No terceiro capítulo são apresentadas algumas das principais abordagens econométricas de séries temporais, detalhando os processos autoregressivos de média móvel, modelos de heterocedasticidade condicional, vetores autoregressivos e o modelo de correção de erros. No quarto capítulo são introduzidas a abordagem Bayesiana e modelos lineares dinâmicos, e no quinto e último capítulo são detalhados os métodos de seleção de modelos, regularização e árvores de classificação e regressão. O objetivo deste trabalho é detalhar algumas das definições e conceitos relevantes

para o entendimento das possibilidades de aplicação de métodos quantitativos em economia e finanças. Com o objetivo de motivar a discussão, abaixo são descritos de forma concisa alguns dos principais conceitos e possibilidades de aplicação das metodologias apresentadas no trabalho.

De modo mais geral, podemos definir uma série temporal como uma quantidade que é medida sequencialmente em um determinado intervalo de tempo. Sob uma perspectiva estatística, assumimos que as séries temporais são realizações de sequências de variáveis aleatórias e que há um processo gerador subjacente para a série, baseado em uma ou mais distribuições estatísticas contendo as variáveis extraídas.

Segundo Enders (2008), a tarefa de um econometrista moderno é a de desenvolver modelos simples, capazes de prever, interpretar e testar hipóteses relativas a dados econômicos e financeiros. O autor sugere que uma série temporal pode ser decomposta em componentes de tendência, sazonalidade, cíclico e irregular. Ademais, é comum encontrar elementos estocásticos nos componentes de tendência, sazonalidade e irregular. Como exemplo, consideremos a série do índice Bovespa, no período 01/2007 à 07/2018:

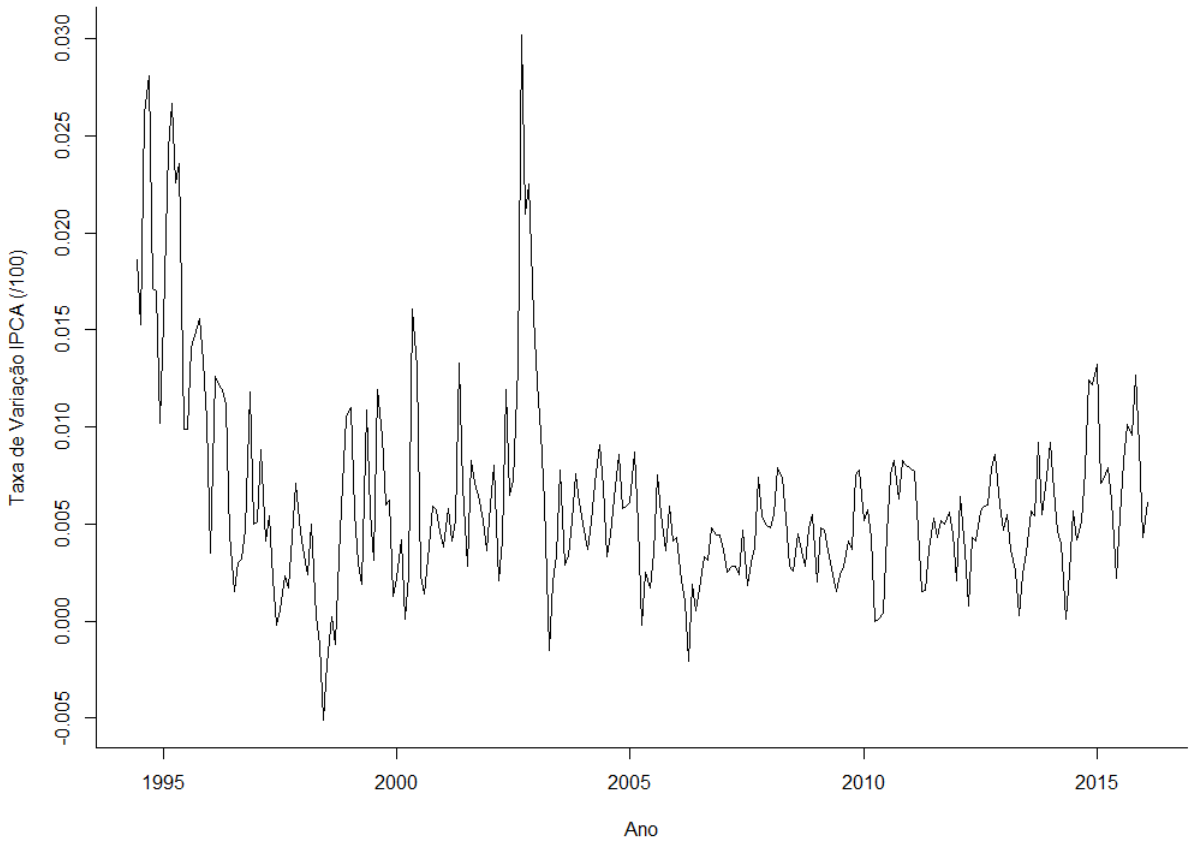

Figura 1 – Variação Mensal IPCA

Fonte: Banco Central do Brasil

Com o objetivo de compreender o processo gerador da série temporal, seguindo a

sugestão de Enders (2008), podemos realizar a decomposição da série em componentes, como

$$y_t = TD_t + sz_t + \varepsilon_t, \tag{1}$$

onde TD_t representa a tendência, sz_t é um efeito sazonal e ε_t é um termo de erro.

Para o caso da série do IPCA mensal, com base em Kendall, Stuart e Ord (1983) e Metcalfe e Cowpertwait (2009), obtemos a serie decomposta

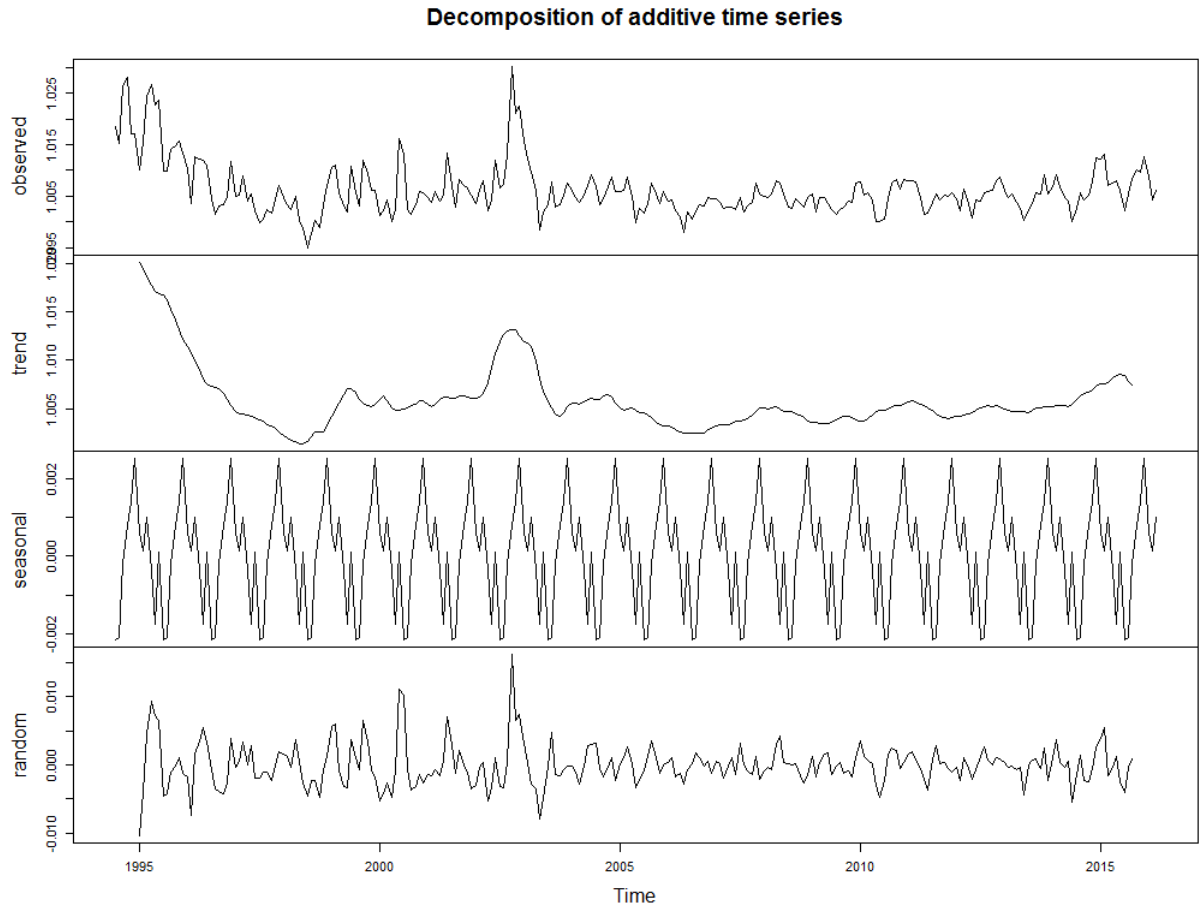

Figura 2 – Decomposição do IPCA

Fonte: Elaboração do Autor

De acordo com Enders (2008), cada um dos componentes da série pode ser representado a partir de uma equação de diferenças, de modo que cada uma das equações expresse o valor da variável como função de suas próprias defasagens e de outras variáveis. Em geral, ao lidar com séries temporais em economia e finanças, estaremos tratando da estimação de equações de diferenças que contêm componentes estocásticos. No entanto, em algumas séries o componente estocástico, é mais acentuado. Nestes casos, não é possível realizar uma decomposição como a realizada para o IPCA.

A série do preço de um ativo, é um exemplo de série com forte componente estocástico, ou aleatório. Faz-se necessário então o entendimento de alguns conceitos importantes, que utilizaremos no contexto da modelagem de séries temporais.

Consideremos a série do Índice Bovespa, por exemplo

Figura 3 – Índice Bovespa (BVSP)

Fonte: Yahoo Finance

A partir de uma inspeção visual, não fica evidente a existência de uma tendência, ou sazonalidade. O que pode ser comprovado a partir da tentativa de decomposição de uma série como esta em um software estatístico. Nestes casos, estamos lidando com um componente aleatório e, em série econômicas e financeiras, os elementos consecutivos de um componente estocástico podem possuir algum grau de correlação. Isto significa que o comportamento de pontos sequenciais em uma série podem afeta uns aos outros, tornando-os dependentes.

Um dos objetivos de um econometrista ou analista quantitativo é identificar a estrutura destas correlações, de forma a obter melhores previsões. Quando as observações sequenciais de uma série são correlacionadas na forma descrita acima, dizemos que a série apresenta correlação serial. Para a definirmos a correlação serial, é importante que façamos a definição de alguns

conceitos simples, porém importantes.

De forma simplificada, o valor esperado, ou esperança, $E(x)$ de uma variável aleatória x é dado pelo valor médio da população μ, e obtemos $E(x) = \mu$. A variância caracteriza a dispersão de uma variável aleatória, sendo definida como a esperança dos desvios quadrados da média e denotada por $\sigma^2(x) = E[(x-\mu)^2]$. O desvio padrão é definido como a raiz quadrada da variância da variável aleatória x. A covariância diz respeito a relação linear entre duas variáveis. Considerando as variáveis aleatórias x e y, com esperança μ_x e μ_y, respectivamente, a covariância é dada por $\sigma(x,y) = E[(x-\mu_x)(y-\mu_y)]$. A correlação é uma medida adimensional do comovimento de duas variáveis. Em essência, a correlação é igual à covariância de duas variáveis aleatórias normalizada por seus respectivos desvios padrão e denotada por

$$\rho(x,y) = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x \sigma_y}.$$

Como exemplo, consideremos duas séries artificiais com 200 períodos de extensão, $x(t)$ e $y(t)$ O processo gerador das mesmas é definido da soma de um componente de tendência, $T = 1,2,\cdots,200$ adicionado a um componente estocástico gerado aleatoriamente, com base em Wichmann e Hill (1982) e Johnson, Kotz e Balakrishnan (1995), o qual segue uma distribuição normal padrão multiplicada por um escalar $c = 15$. Obtemos que a covariância e correlação amostrais são dadas por $Cov = 3507,54$ e $Corr = 0,93$. É importante notar que a principal diferença entre a covariância e a correlação está na escala

A figura 4 representa as séries $x(t)$ e $y(t)$, bem como o gráfico de correlação das séries.

Dadas as definições acima, podemos tratar agora de suas aplicações a dados de séries temporais. A média de uma série temporal x_t , é definida como a esperança da média μ_t , onde μ_t é função do tempo. Uma série temporal é considerada estacionária na média se $\mu(t) = \mu$ é uma constante. Assumindo que uma série temporal é estacionária na média, definimos sua variância $\sigma^2(t)$ como

$$\sigma^2(t) = E[(x_t - \mu)^2],$$

onde $\sigma^2(t)$ é também função do tempo. Assumindo que a variância da população σ^2 é constante, podemos definir a variância amostral como

$$Var(x) = \frac{\sum(x_t - \bar{x})^2}{n-1}.$$

Uma série temporal é estacionária na variância se $\sigma^2(t) = \sigma^2$ é uma constante. Quando uma série temporal é estacionária na média e na variância, dizemos que a série é *estacionária*

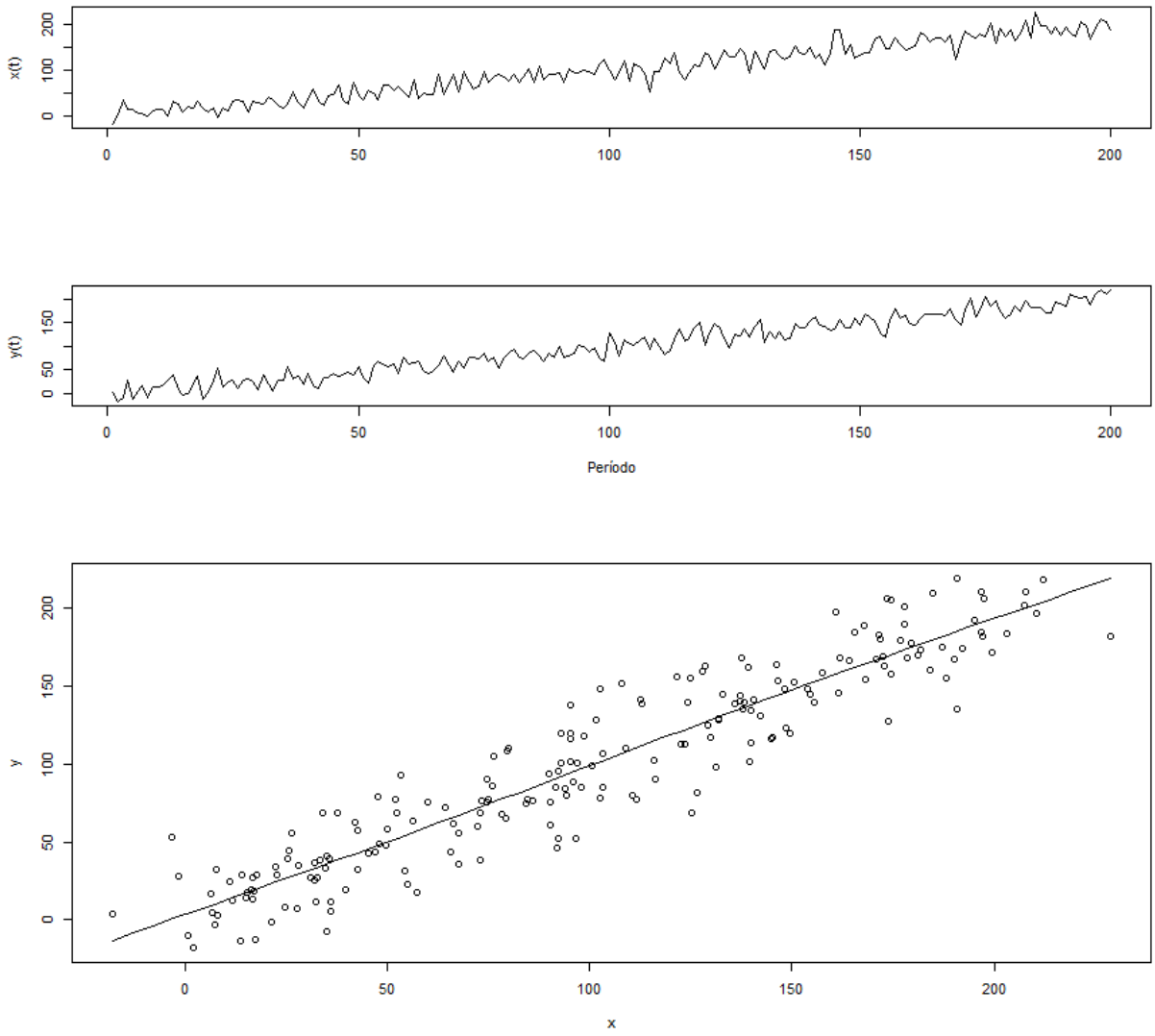

Figura 4 – Covariância vs Correlação

Fonte: Elaboração do Autor

de segunda ordem (METCALFE; COWPERTWAIT, 2009, p. 33) caso a correlação entre observações sequenciais seja função apenas da quantidade de defasagens entre as mesmas. Para séries estacionárias de segunda ordem definimos a covariância serial, ou simplesmente autocovariância , da defasagem k como

$$C_k = E[(x_t - \mu)(x_{t+k} - \mu)].$$

Definimos a correlação serial, ou simplesmente autocorrelação, da defasagem k para séries estacionárias de segunda ordem como a autocovariância normalizada pelo produto dos desvios padrão e a denotamos por $\rho_k = C_k/\sigma^2$. Finalmente, definimos a função de autocovariância

amostral como

$$c_k = \frac{1}{n}\sum_{t=1}^{n-k}(x_t - \bar{x})((x_{t+k} - \bar{x}),$$

e a função de autocorrelação amostral como

$$r_k = \frac{c_k}{c_0}.$$

Podemos visualizar a autocorrelação a partir de um diagrama de correlação, ou simplesmente correlograma, onde plotamos os resultados da função de autocorrelação para valores sequenciais da defasagem $k = 0, 1, \cdots, n$. Com base em Box, Jenkins e Reinsel (1978), a Figura 5 mostra o correlograma da função de autocorrelação e da função de autocorrelação parcial para a série mensal do IPCA

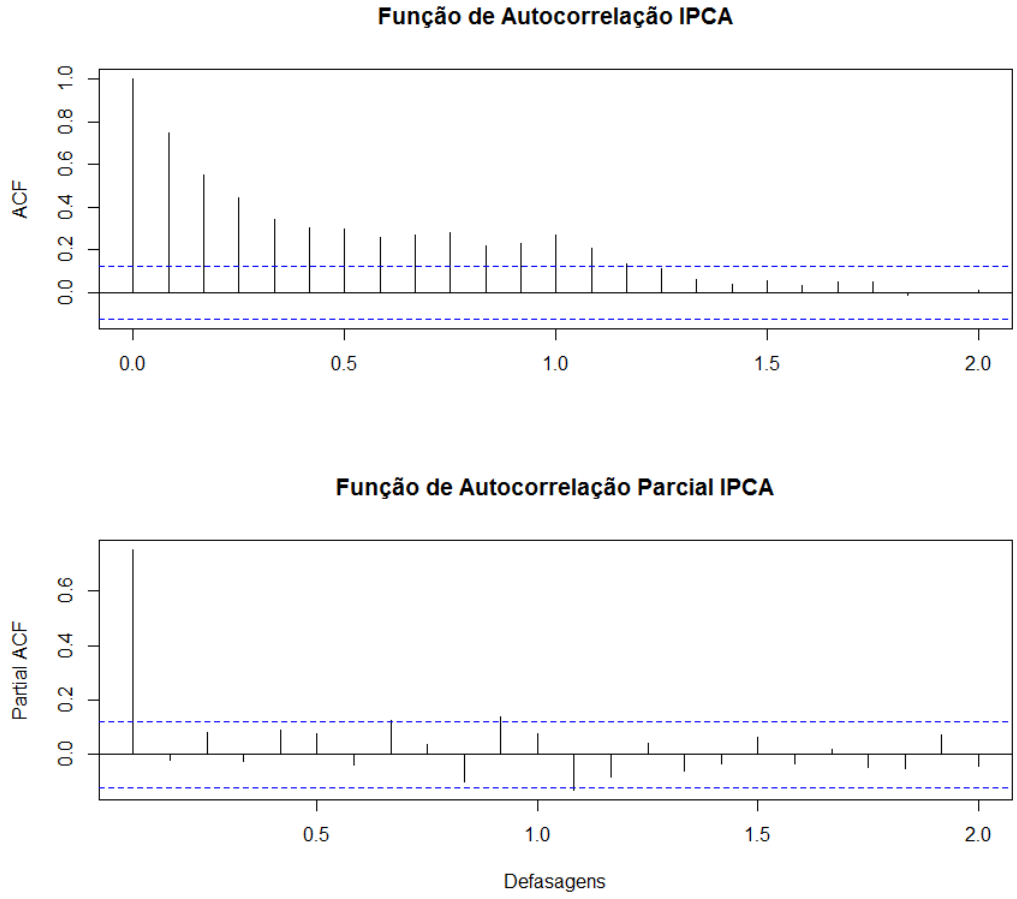

Figura 5 – Diagrama de Correlação IPCA

Fonte: Banco Central do Brasil, Elaboração do Autor

O diagrama de correlação pode ser utilizado para a detecção de autocorrelação após a remoção dos componentes de tendência e sazonais, para avaliar autocorrelação entre os termos de erro e ainda para identificar os próprios componentes sazonais.

Como exemplo, a Figura 6 mostra a série $w(t)$, com $t = 1, 2, \cdots, 150$ e período $p = 15$ adicionada um termo de pertubação normal, onde percebemos que a função de autocorrelação consegue capturar o efeito periódico associado à série

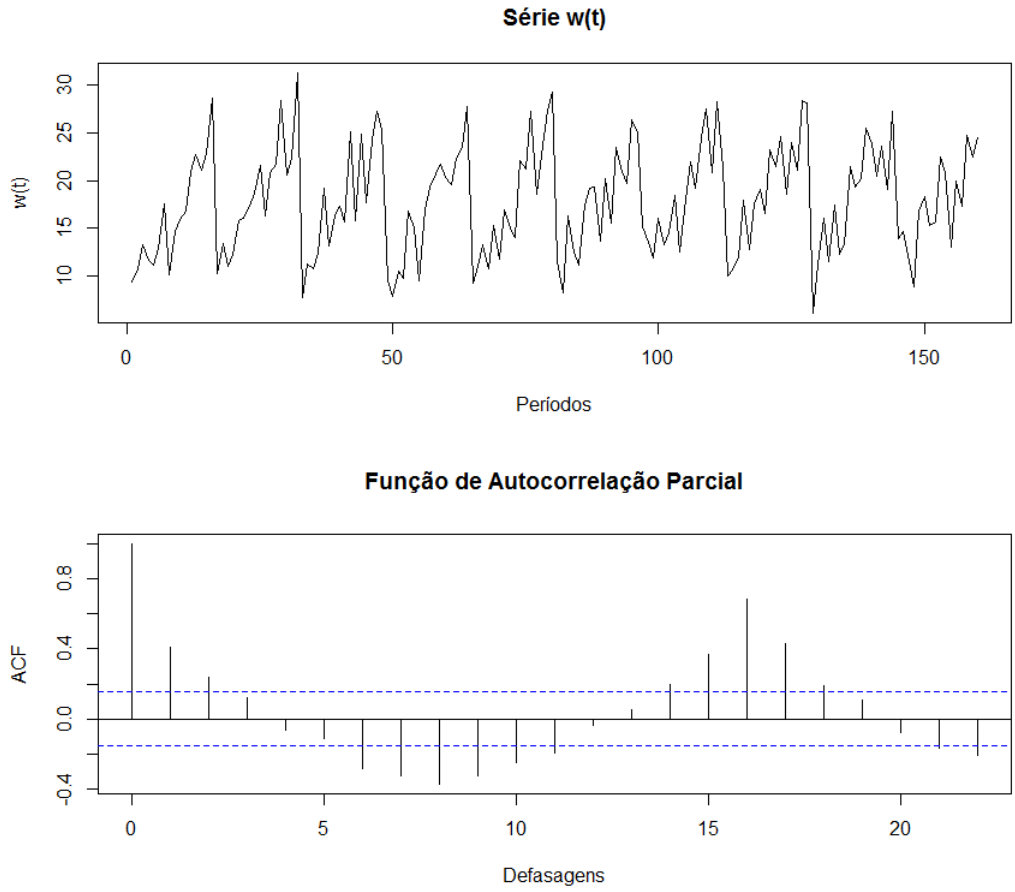

Figura 6 – Diagrama de Correlação da Série w(t)

Fonte: Elaboração do Autor

No contexto na análise de séries temporais, tratamos da modelagem de séries temporais. Podemos definir um modelo como uma representação matemática que tenta explicar o processo gerador de uma série, ou até mesmo a estrutura de relacionamento entre séries distintas. Quando modelamos uma série, podemos ter como objetivo a realização de previsões e é importante que tenhamos critérios para avaliar a qualidade das mesmas. Sempre que fazemos uma previsão, geramos ao mesmo tempo um erro de previsão. Não é imprudente afirmar que os erros de previsão trazem tanta (ou até mais) informação quanto os coeficientes estimados ou até a própria previsão.

A série de erros residuais e_t é definida como a diferença entre os valores observados y_t e os valores previstos $\hat{y}$ a partir de um modelo de séries temporais, em um determinado período t. , de forma que

$$e_t = y_t - \hat{y}_t.$$

Caso nosso modelo seja capaz de explicar a correlação serial entre as observações, então os erros de previsão serão serialmente não-correlacionados. Isto implica que cada elemento da série de resíduos serialmente não-correlacionados seja uma realização independente de uma certa distribuição de probabilidade. Tal resultado é importante pois, na ausência de erros correlação entre os erros, podemos assumir que a média amostral seja uma boa aproximação para a média populacional. Há ainda o resultado importante de que a medida quantidade de amostras aumenta, caso os erros sejam não correlacionados, nosso estimador se aproximará cada vez mais do valor "real". Tal resultado por ser verificado a partir da de um experimento cujos resultados podem ser considerados realizações independentes de uma certa distribuição de probabilidade.

Com base em Rakhshan e Pishro-Nik (2014), consideremos U uma variável aleatória que segue distribuição Uniforme(0,1). Definimos X como uma variável aleatória que segue distribuição de Bernoulli como

$$X = \begin{cases} 1 & \text{if } U < p \\ 0 & \text{if } U \geq p \end{cases},$$

de forma que $P(H) = P(X = 1) = P(U < p) = p$ e X segue uma distribuição Bernoulli com parâmetro p. Considerando $p = 0.5$ temos um experimento equivalente ao lançamento de uma moeda não viesada, e chegamos a uma demonstração do funcionamento da lei dos grandes números.

A Figura 7 mostra 3 simulações de um experimento composto pelo lançamento de uma moeda não-viesada por 500 vezes. Percebemos que como os lançamentos são independentes, a probabilidade tende a convergir para 0.5 a medida que a quantidade de lançamentos aumenta.

Retornando à série de erros residuais $e(t)$, caso os elementos da série sejam serialmente não-correlacionados, os denominados independentes e identicamente distribuídos. Portanto, se estamos interessados em criar modelos de séries temporais que consigam explicar correlações seriais, é natural iniciarmos por um processo que produza variáveis aleatórias independentes a partir de alguma distribuição de probabilidade.

Com base em Metcalfe e Cowpertwait (2009), consideremos uma série $w_t : t = 1, \cdots, n$, onde os elementos w_i são independentes e identicamente distribuídos, com média zero, variância σ^2 e sem a presença de correlação serial. Denominamos tal série como um processo de ruído

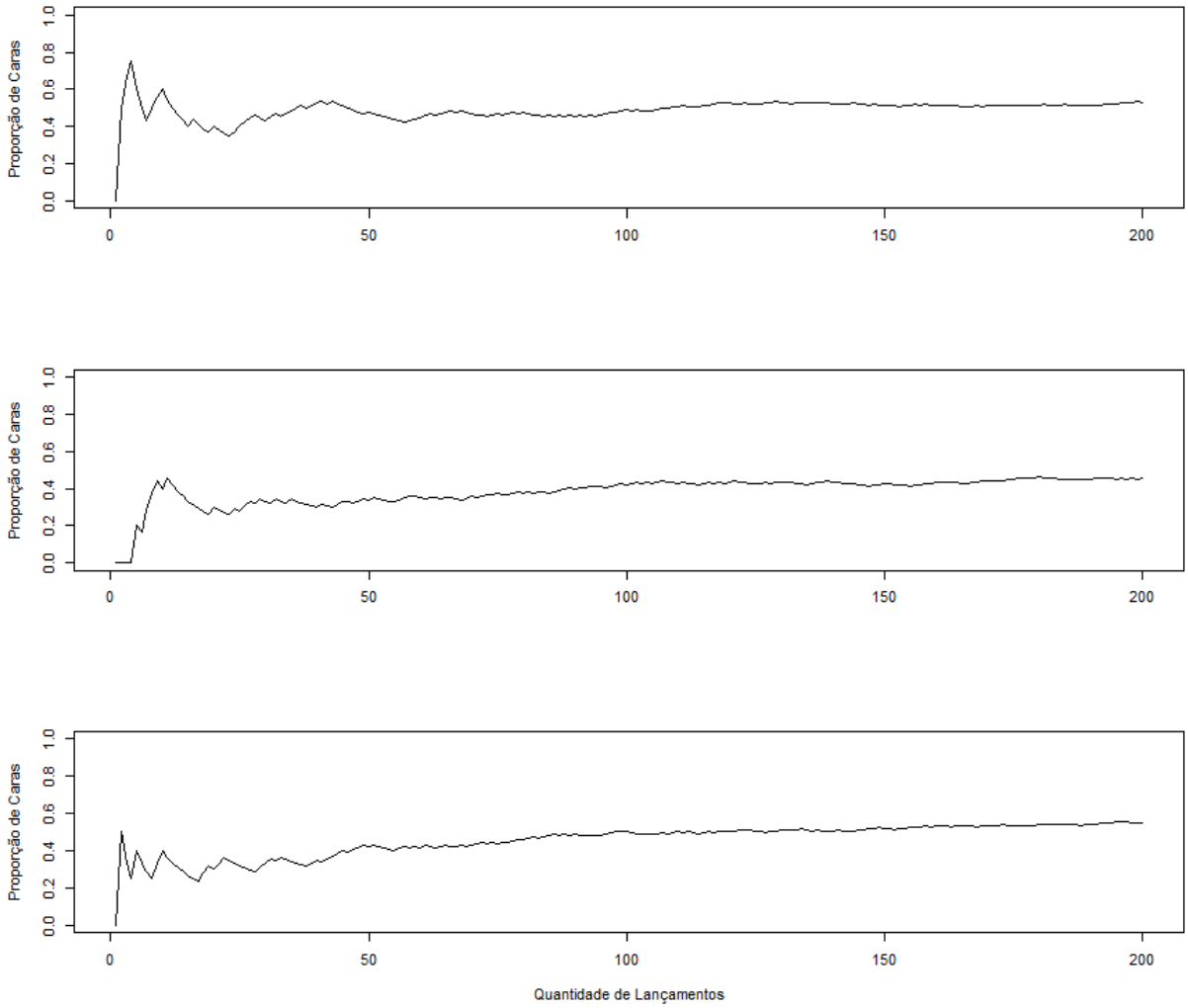

Figura 7 – Simulação de Lançamentos de uma Moeda

Fonte: Elaboração do Autor

branco discreto, e caso os elementos w_i sejam extraídos de uma distribuição normal padrão, a série é denominada um processo de ruído branco Gaussiano.

A Figura 8 mostra a simulação de um processo de ruído branco Gaussiano e a diferença entre a distribuição de uma amostra e de sua respectiva população. Mais especificamente, simulamos a 100 variáveis as quais seguem uma distribuição normal padrão, o que, segundo Metcalfe e Cowpertwait (2009) é equivalente à simulação de um processo de ruído branco Gaussiano de extensão 100.

Consideremos agora uma série x_t. Dizemos que x_t é um passeio aleatório caso

$$x_t = x_{t-1} + w_t,$$

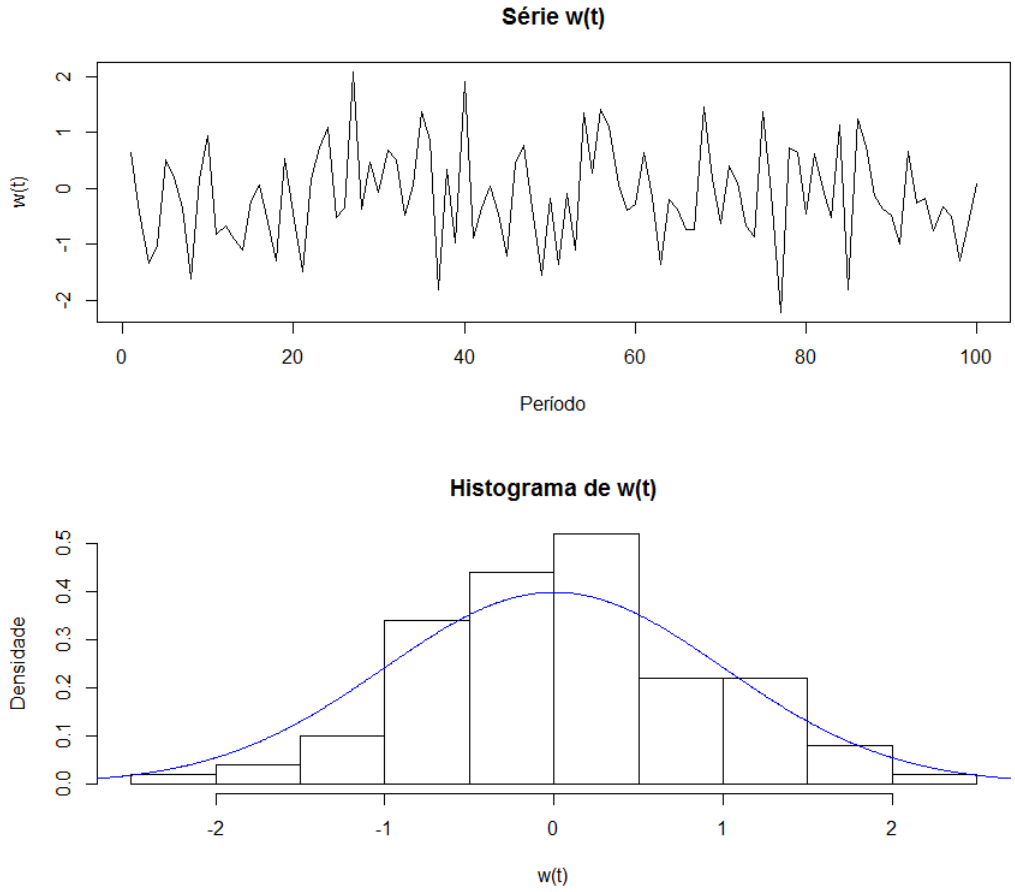

Figura 8 – Simulação de um Processo de Ruído Branco Gaussiano

Fonte: Elaboração do Autor

onde w_t é uma série de ruído branco.

Substituindo $x_{t-1} = x_{t-2} + w_{t-1}$ na equação acima, e assim sucessivamente para as outras defasagens de x_t obtemos

$$x_t = w_1 + w_2 + \cdots + w_t.$$

Desta forma, como os termos w_i são elementos de um processo de ruído branco, a realização de previsões sobre o valor de x_{t+1} não parece muito promissor. Em um contexto de mercado, podemos considerar que o processo de geração de preços de um determinado ativo siga um processo como o passeio aleatório. Neste caso, temos então um *martingale*. Em essência, de acordo com Campbell et al. (1997), um martingale é um processo estocástico P_t que satisfaz a propriedade

$$E[P_{t+1}|P_t, P_{t-1}, \cdots] = P_t,$$

ou, de forma equivalente, que

$$E[P_{t+1} - P_t | P_t, P_{t-1}, \cdots] = 0.$$

Se esta condição for satisfeita, a implicação prática é que a melhor previsão para o preço de um ativo no período $t+1$ é o preço do ativo no período t. Segundo Campbell et al. (1997), a hipótese de martingale foi por muito tempo considerada uma condição necessária para que um mercado fosse eficiente, de forma que toda informação contida no passado é refletida instantaneamente e completamente no nível atual de preços dos ativos. Se um mercado é eficiente, então não é possível realizar lucro a partir de operações realizadas com base em informações contidas no histórico de preços. Quanto mais eficiente for o mercado, mais aleatória será a sequência de mudanças de preço geradas pelo mercado, e o mercado mais eficiente é aquele onde toda e qualquer mudança no nível de preços dos ativos é completamente imprevisível e aleatória (CAMPBELL et al., 1997, p. 30). Quando avaliado em tempo contínuo, o processo estocástico que da origem ao passeio aleatório forma um movimento browniano. A Figura 9 representa a simulação de um movimento browniano em uma e duas dimensões.

Voltando ao tempo discreto, um processo de passeio aleatório possui média $\mu_x = 0$ e covariância dependente do tempo, tornando o passeio aleatório um processo não estacionário. Há diversos processos baseados em passeios aleatório, podendo variar de acordo com o processo gerador subjacente e com as características dos termos de erro. (CAMPBELL et al., 1997) introduzem três tipos de passeio aleatório, denominados PA1, PA2 e PA3.

O PA1 é a versão mais simples de um passeio aleatório, onde os incrementos são independentes e identicamente distribuídos, de forma que a dinâmica do processo P_t é dada por

$$P_t = \mu + P_{t-1} + \varepsilon_t,$$

onde μ representa a mudança de preços esperada, ou *drift*. Com base em Pfaff (2008), a Figura 10 representa a simulação de um processo de passeio aleatório com drift $\mu = 0.1$ e seu respectivo diagrama de autocorrelação parcial.

No segundo tipo de passeio aleatório, PA2, consideramos incrementos independentes mas não identicamente distribuídos. Desta forma, o processo PA2 permite heterocedasticidade incondicional nos incrementos ε_t, sendo um caso mais geral de do que PA1.

O terceiro tipo de passeio aleatório, PA3, descreve a versão mais testada da hipótese de passeio aleatório no contexto da literatura empírica de finanças (CAMPBELL et al., 1997), relaxando a hipótese de independência entre os incrementos, permitindo assim processos com incrementos dependentes, porém não correlacionados. Um exemplo de PA3 é um processo cuja

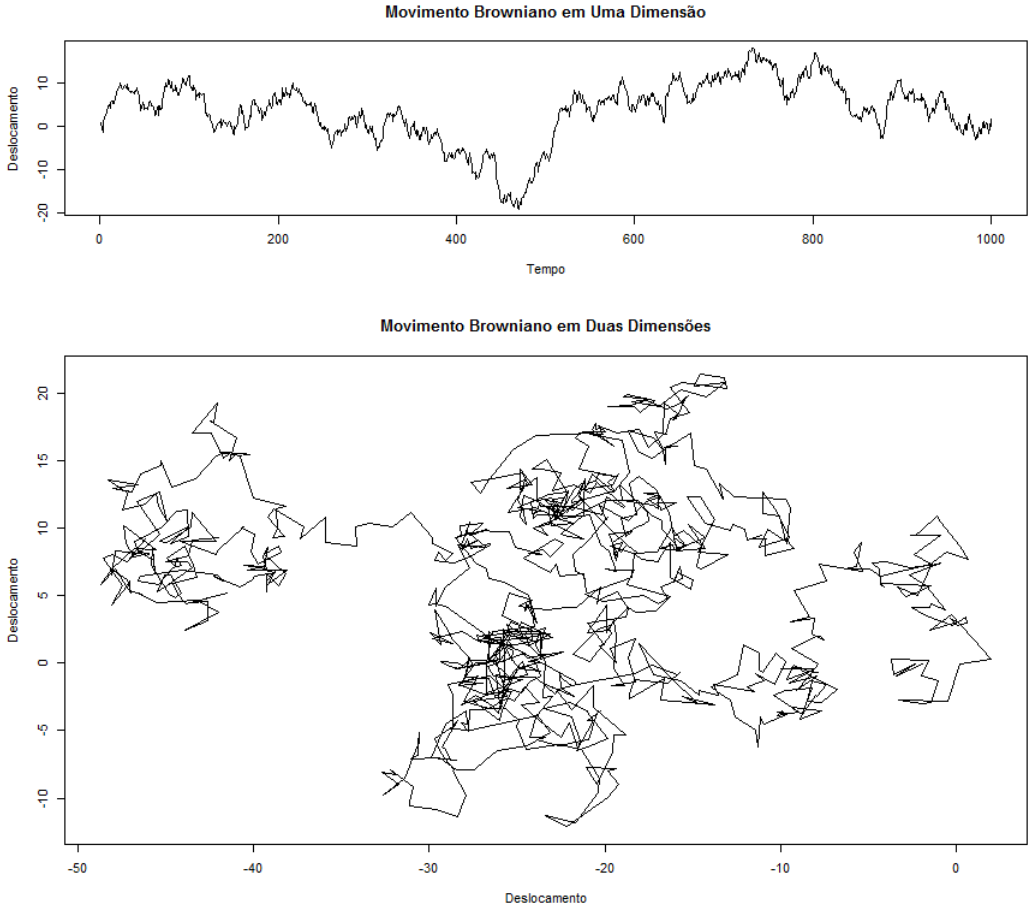

Figura 9 – Simulação de um Movimento Browniano

Fonte: Elaboração do Autor

covariância dos incrementos seja igual à zero para todos os valores de k, mas cuja covariância do quadrado dos incrementos seja diferente de zero, para algum k.

Tal característica está relacionada a uma propriedade conhecida de algumas séries temporais financeiras conhecida como agrupamento de volatilidade, o que de forma simplificada significa que a volatilidade da série não é constante no tempo. Tecnicamente, este comportamento é denominado como *heterocedasticidade condicional* e caso uma série apresente volatilidade variante no tempo, a mesma não será estacionária na variância, por definição.

Com base em Metcalfe e Cowpertwait (2009), uma série x_t é *estritamente estacionária* caso a distribuição conjunta dos elementos $x_{t_1}, \cdots, x_{t_n}$ seja a mesma do que aquela dos elementos $x_{t_1+m}, \cdots, x_{t_n+m}$ para todos os valores de (t_i, m).

A definição de estacionariedade estrita implica que a distribuição da série não se altera por conta de mudanças no tempo. Desta forma, séries estritamente estacionárias apresentam média e variância constantes, com a covariância entre x_t e um dado x_s dependendo somente da diferença absoluta entre t e s.

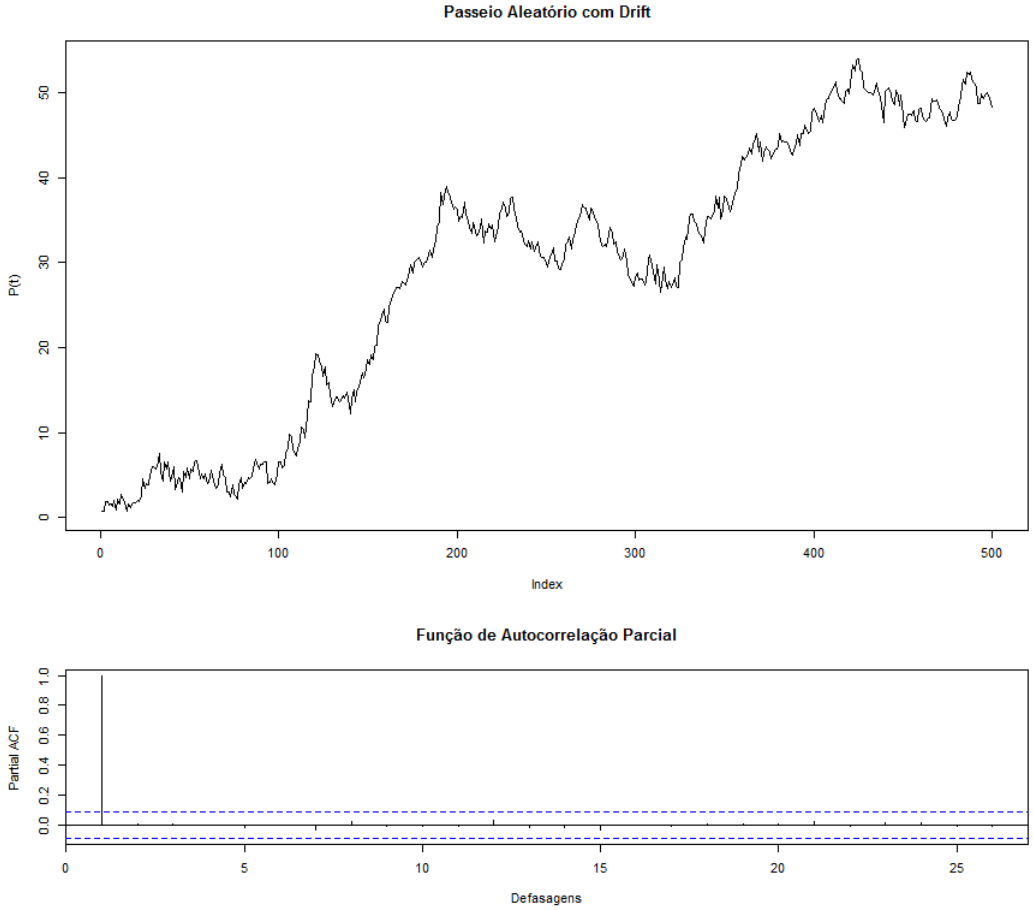

Figura 10 – Simulação de um Passeio Aleatório com Drift

Fonte: Elaboração do Autor

Partindo da definição de um processo de passeio aleatório, onde cada termo x_t depende somente o termo anterior x_{t-1} e de um termo de ruído branco w_t, temos que um modelo autoregressivo é simplemente uma extensão de um processo de passeio aleatório, permitindo a inclusão de termos anteriores no modelo. Em geral, os modelos autoregressivos apresentam uma estrutura linear, de forma que o modelo depende linearmente dos termos passados, com coeficiente representando a magnitude desta relação.

Definimos uma série x_t como um processo autoregressivo de ordem p, ou simplesmente AR(p) se

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_p x_{t-p} + w_t,$$

onde w_t é um ruído branco. Percebemos que o passeio aleatório é um caso especial de processo autoregressivo, com $\alpha = 1$.

A Figura 11 representa a simulação de um processo autoregressivo de ordem 1, com

coeficientes $\alpha_1 = 0.8$ e $\alpha_1 = -0.8$.

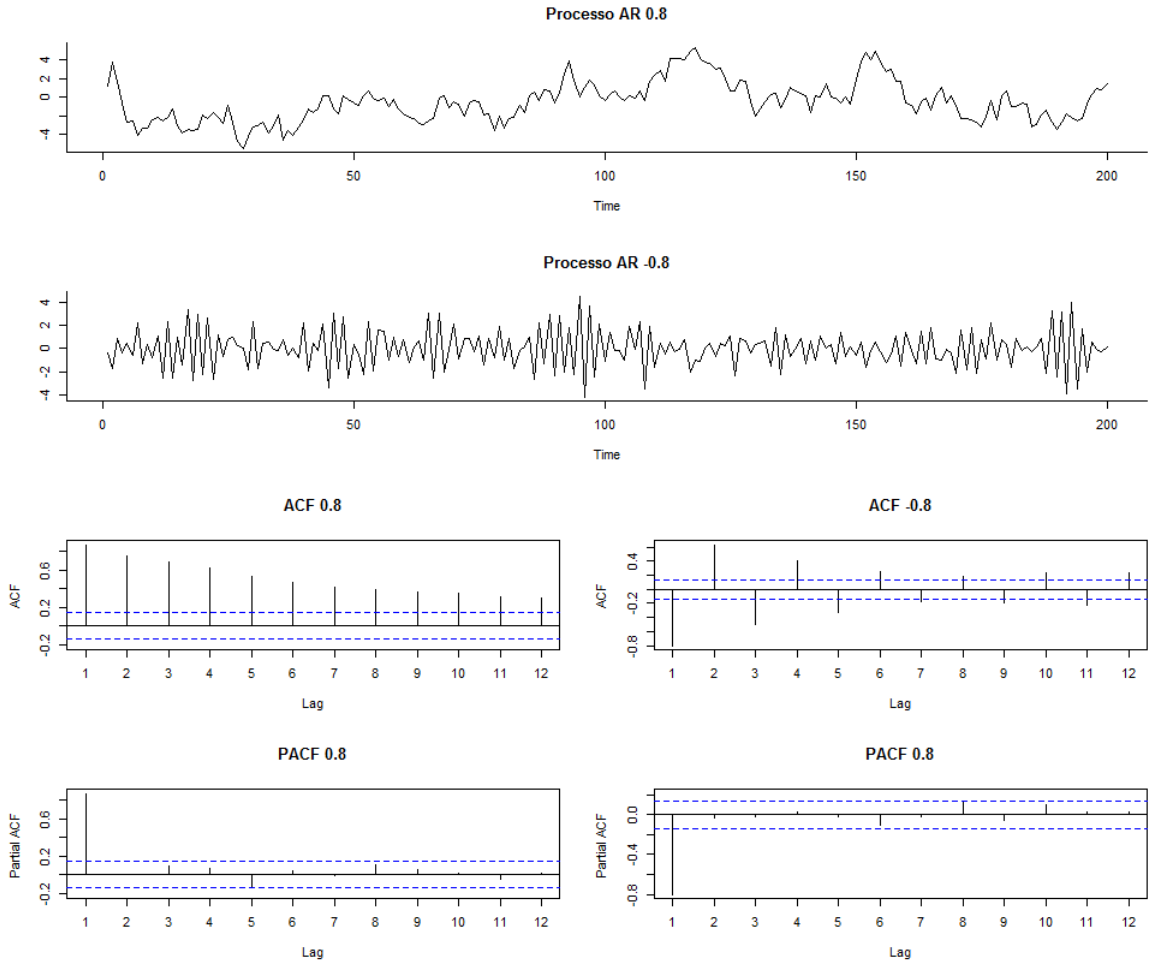

Figura 11 – Simulação de um Processo AR(1)

Fonte: Elaboração do Autor

Adicionalmente, definimos um processo de média móvel de ordem q, ou simplesmente MA(q) como uma combinação linear do termo de ruído branco atual com os q termos de ruído branco passados mais recentes, de forma que

$$x_t = w_t + \beta_1 w_{t-1} + \cdots + \beta_q w_{t-q},$$

onde w_t é um termo de ruído branco com média zero e variância σ_w^2. Como um processo MA(q) é composto pela soma termos de ruído branco estacionários, o mesmo é estacionário com média e autocovariância invariantes no tempo.

De forma complementar, definimos que uma série x_t segue um processo autoregressivo de média móvel de ordem (p,q), ou simplesmente ARMA(p,q), se

$$x_t = \alpha_1 x_{t-1} + \cdots + \alpha_p x_{t-p} + w_t + w_t + \beta_1 w_{t-1} + \cdots + \beta_q w_{t-q}$$

A Figura 12 representa a simulação de processos ARMA com diferentes coeficientes autoregressivos e de média móvel.

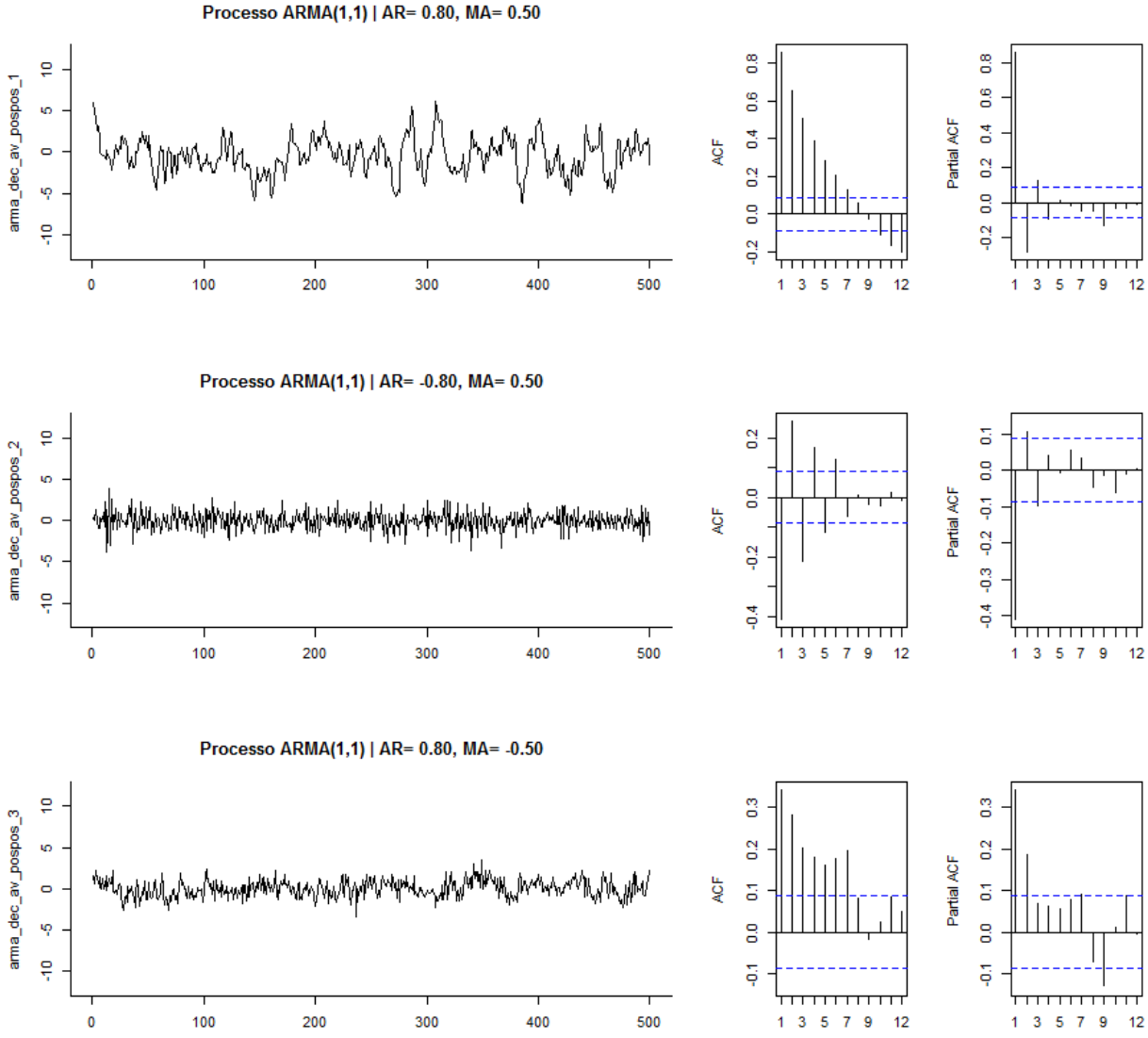

Figura 12 – Simulação de um Processo ARMA

Fonte: Elaboração do Autor

Os processos ARMA apresentados até o momento são estacionários. Caso não o sejam, é possível transformar uma série não estacionária em estacionária a partir da diferenciação (METCALFE; COWPERTWAIT, 2009, p. 137). A diferenciação de uma série tem o efeito de remover tendências, sejam elas estocásticas, como em um passeio aleatório, ou determinísticas.

Finalmente, dizemos que uma série x_t segue um processo autoregressivo integrado de média móvel de ordem (p,d,q), ou simplesmente ARIMA(p, d, q) caso a d-ésima diferença da série x_t seja um processo ARMA(p, q).

Consideremos um conjunto de variáveis aleatórias, como elementos em um modelo de séries temporais. Dizemos que tal conjunto é heterocedástico caso certos grupos, ou subconjuntos de variáveis, apresentem variância diferente das outras. Como exemplo, em séries não estacionárias que possuem efeitos de sazonalidade ou tendência, é comum encontrar que o nível de variância aumenta com o passar do tempo.

Como exemplo, consideremos novamente a série do Índice Bovespa, agora representada em termos de retornos diários na Figura 13

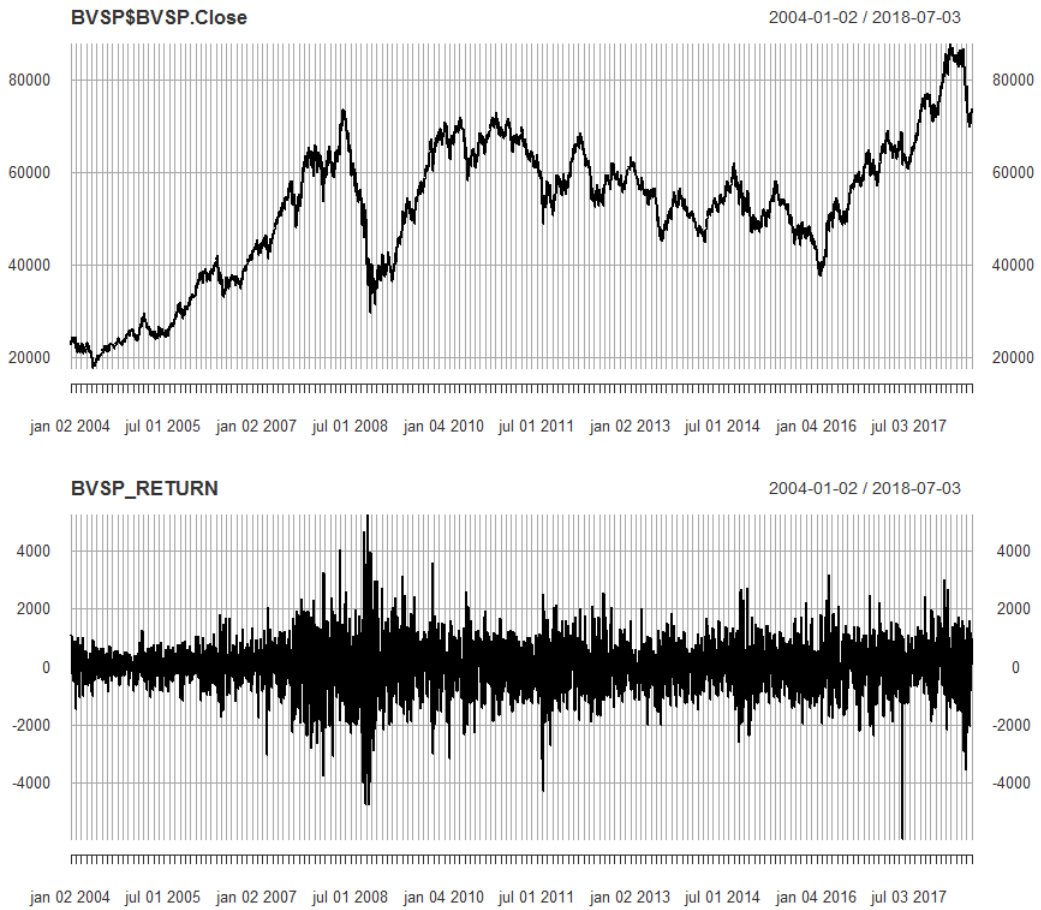

Figura 13 – Retornos Índice Bovespa

Fonte: Yahoo Finance; Elaboração do Autor

A partir de uma inspeção visual, percebemos que a variância parece ser maior entre os anos de 2007 à 2010 do que nos anos de 2004 ou 2013 e este é um indício de que a série apresenta heterocedasticidade. Detalhando, caso a variância não seja constante ao longo do tempo, mas varie de forma regular, dizemos que a série é heterocedástica. Caso a série exiba períodos de

aumento na variância, de forma que a variância seja correlacionada ao longo do tempo, dizemos que a série apresenta heterocedasticidade condicional (METCALFE; COWPERTWAIT, 2009).

A figura mostra os diagramas de correlação para os retornos do Índice Bovespa e para a variância dos retornos. Percebemos que os retornos não apresentam forte autocorrelação, apresentando resultados parecidos aos obtidos em um passeio aleatório simples, já a autocorrelação entre a variância dos retornos apresenta valores mais elevados, o que é um forte indício da presença de heterocedasticidade condicional.

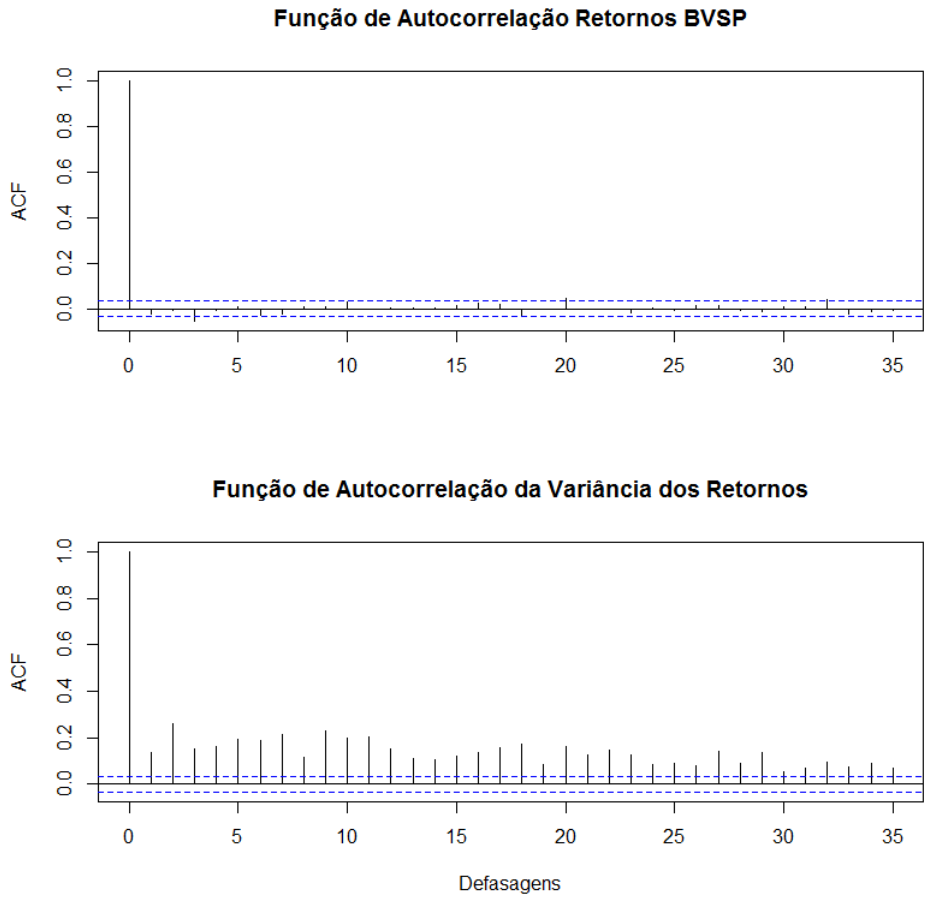

Figura 14 – Função de Autocorrelação da Variância dos Retornos do Índice Bovespa

Fonte: Yahoo Finance; Elaboração do Autor

Para lidarmos com a característica de heterocedasticidade condicional, é necessário que tenhamos um modelo que permita, em sua especificação, mudanças condicionais na variância. Uma alternativa é a utilização de um modelo autoregressivo para o processo de variância. Dizemos que uma série ε_t é um processo autoregressivo de heterocedasticidade condicional de ordem 1, ou simplesmente ARCH(1), se

$$\varepsilon_t = w_t \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2},$$

onde w_t é um processo de ruído branco. Tomando a variância de ε_t, obtemos que

$$Var(\varepsilon_t) = \alpha_0 + \alpha_1 Var(\varepsilon_{t-1})$$

Podemos estender um ARCH(1) para um modelo ARCH(p) a partir da inclusão de defasagens de ordem mais altas. Denotamos um processo ARCH(p) por

$$\varepsilon_t = w_t \sqrt{\alpha_0 + \sum_{i=1}^{p} \alpha_p \varepsilon_{t-i}^2},$$

onde novamente w_t é um processo de ruído branco.

Finalmente, generalizamos um processo ARCH(p) para um processo autoregressivo de heterocedasticidade condicional generalizado, ou simplesmente GARCH(q, p). Uma série $\varepsilon_t = w_t \sqrt{h_t}$ é um processo GARCH(p, q) se

$$\varepsilon_t = w_t \sqrt{h_t},$$

onde

$$h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}.$$

Com base em Metcalfe e Cowpertwait (2009), simulamos um processo GARCH(1,1), denotado por

$$e_t = w_t \sqrt{h_t},$$

onde

$$h_t = \alpha_0 + \alpha_1 e_{t-1} + \beta_1 h_{t-1}.$$

Para a simulação, definimos $(\alpha_0, \alpha_1, \beta_1 = (0.1, 0.3, 0.2)$. O termo w_t é um ruído branco gaussiano e a simulação é realizada para 5000 períodos.

A figura 15 mostra as séries geradas, e o diagrama de correlação para e(t). Percebemos que a série de fato se encaixa nas características de um GARCH, de forma que a ACF da variância da série apresenta valores não nulos; o que é nosso indicativo para a presença de heterocedasticidade condicional.

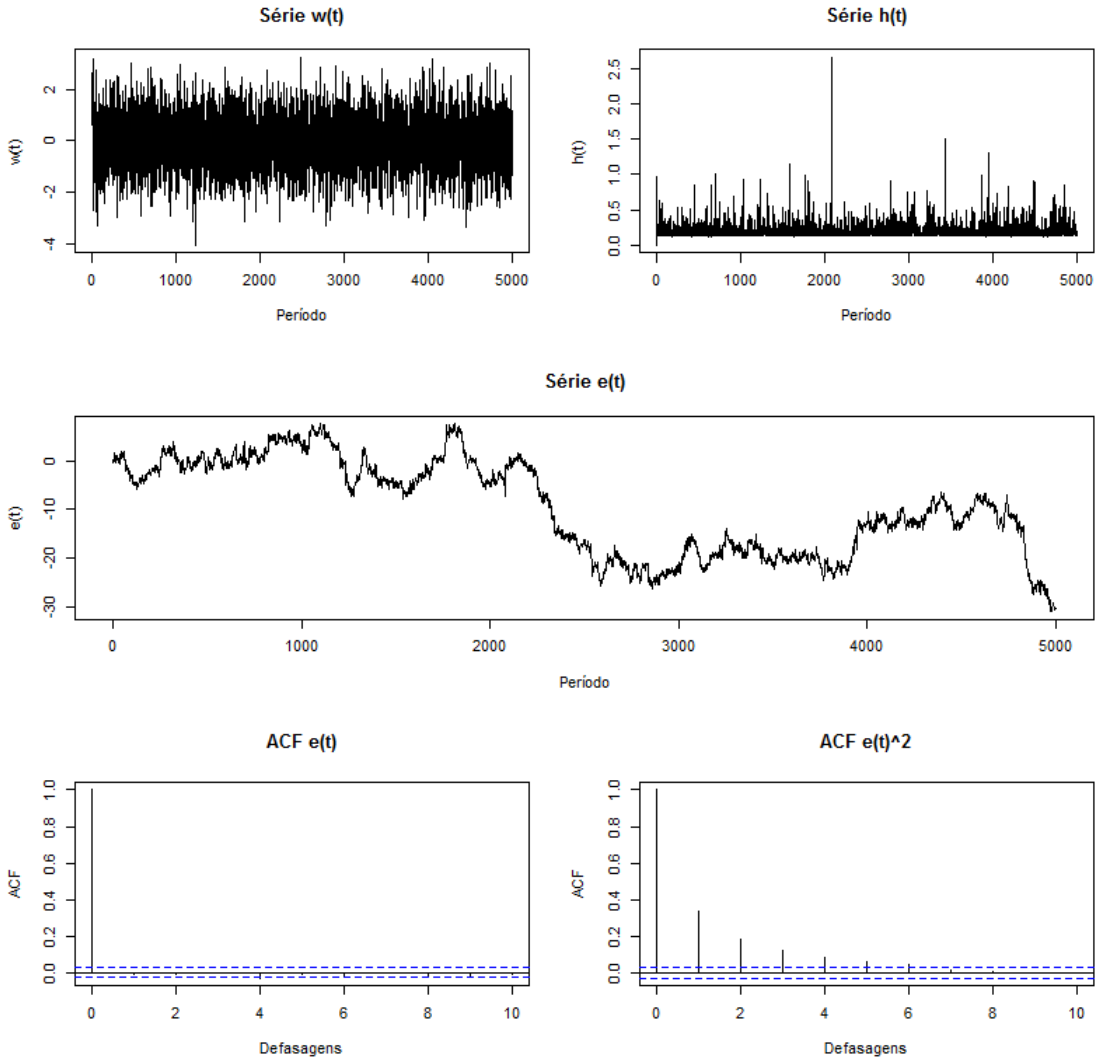

Figura 15 – Simulação de um Processo GARCH(1,1)

Fonte: Elaboração do Autor

Dadas as características de um modelo GARCH, a expectativa é que a partir do encaixe do modelo a uma série que apresente heterocedasticidade condicional, obtenhamos resíduos estacionários, que se comportem como ruídos brancos.

A Figura 16 mostra a série dos resíduos obtidos a partir do encaixe de um modelo GARCH à série de retornos do Índice Bovespa. Percebemos que agora os resíduos de fato se comportam como ruídos brancos, e podemos concluir que o modelo GARCH de fato consegue capturar as principais características da série.

Por vezes, a uma série depende não apenas de seus valores defasados, mas também dos valores defasados de outras série. Nestas situações, precisamos de um modelo que possibilite a avaliação de processos autoregressivos em forma de sistema, onde os valores atuais e defasados de uma variável possam afetar os valores das outras variáveis.

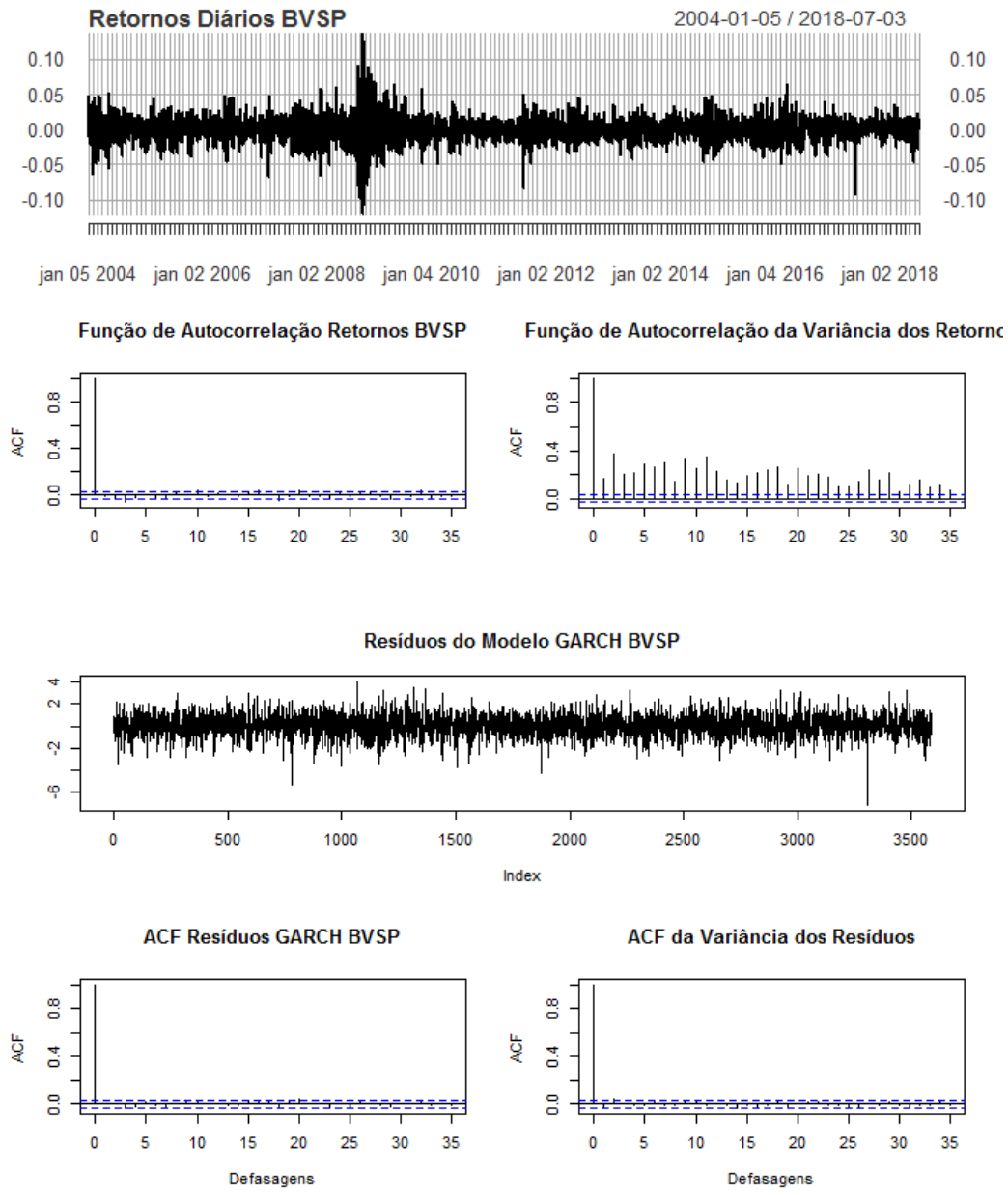

Figura 16 – Diagrama de Correlação dos Resíduos do Modelo GARCH BVSP

Fonte: Elaboração do Autor

Tais modelos existem, sendo denominados vetores autoregressivos, ou simplesmente VAR. Com base em Enders (2008), representamos um VAR de n variáveis como

$$\begin{bmatrix} x_{1t} \\ x_{2t} \\ \vdots \\ x_{nt} \end{bmatrix} = \begin{bmatrix} A_{10} \\ A_{20} \\ \vdots \\ A_{n0} \end{bmatrix} + \begin{bmatrix} A_{11}(L) & A_{12}(L) & \cdots & A_{1n}(L) \\ A_{21}(L) & A_{22}(L) & \cdots & A_{2n}(L) \\ \vdots & \vdots & \vdots & \vdots \\ A_{n1}(L) & A_{n2}(L) & \cdots & A_{nn}(L) \end{bmatrix} \begin{bmatrix} x_{1t-1} \\ x_{2t-1} \\ \vdots \\ x_{nt-1} \end{bmatrix} + \begin{bmatrix} e_{1t} \\ e_{2t} \\ \vdots \\ e_{nt} \end{bmatrix},$$

onde os termos A_{i0} representam os interceptos das equações, os temos $A_{ij}(L)$ representam

os polinômios no operador de defasagem L e todas as equações possuem a mesma extensão, gerando polinômios de mesmo grau. Isto gera uma condição sensível, a qual envolve a definição da extensão apropriada de defasagens.

Com base em Metcalfe e Cowpertwait (2009), simulamos um processo VAR bivariado de ordem 1, onde as séries y_t e z_t são geradas aleatoriamente a partir de ruídos brancos bivariados gaussianos, com matriz de covariância dada por

$$\sum = \begin{bmatrix} 1 & 0,65 \\ 0,65 & 1 \end{bmatrix},$$

adicionalmente, para a simulação, definimos a matriz de coeficientes na forma do VAR bivariado de (ENDERS, 2008)

$$\begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} a_{10} \\ a_{20} \end{bmatrix} + \begin{bmatrix} 0,4 & 0,3 \\ 0,2 & 0,1 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix}. \tag{2}$$

A Figura 17 mostra a o diagrama de correlação-cruzada entre os ruídos brancos bivariados e as séries geradas pela simulação em 10000 períodos

Em diversas situações, as variáveis cujas relações são investigadas por um econometrista são não-estacionárias. Uma saída é buscar encontrar uma combinação linear destas variáveis não estacionárias que seja estacionária. Nestes casos, o conceito chave é o de *cointegração*.

De acordo com Pfaff (2008), a ideia por trás da do conceito de cointegração é encontrar uma combinação linear entre duas variáveis integradas de mesma ordem I(d) que resulte em uma variável com ordem mais baixa de integração. Com base em Pfaff (2008), definimos que os componentes de um vetor x_t são cointegrados de ordem *d, b*, denotado por $x_t \sim CI(d,b)$, se todos os componentes de x_t são integrados de ordem d, I(d), e se existe um vetor $\alpha \neq 0$ de forma que

$$z_t = \alpha' x_t \sim I(d,b), \;\; b < 0,$$

onde denominamos α como o *vetor de cointegração*.

Dados que encontramos duas séries cointegradas, podemos utilizar a especificação de um modelo de correção de erros. Com base em Pfaff (2008), a especificação geral de um modelo de correção de erros é dada por

$$\begin{aligned} \Delta y_t = &\quad \psi_0 + \gamma_1 \hat{z_{t-1}} + \textstyle\sum_{i=1}^{K} \psi_{1,i} \Delta x_{t-i} + \sum_{i=1}^{L} \psi_{2,i} \Delta y_{t-i} + \varepsilon_{1,t} \\ \Delta x_t = &\quad \xi_0 + \gamma_2 \hat{z_{t-1}} + \textstyle\sum_{i=1}^{K} \xi_{1,i} \Delta y_{t-i} + \sum_{i=1}^{L} \xi_{2,i} \Delta x_{t-i} + \varepsilon_{1,t} \end{aligned} \tag{3}$$

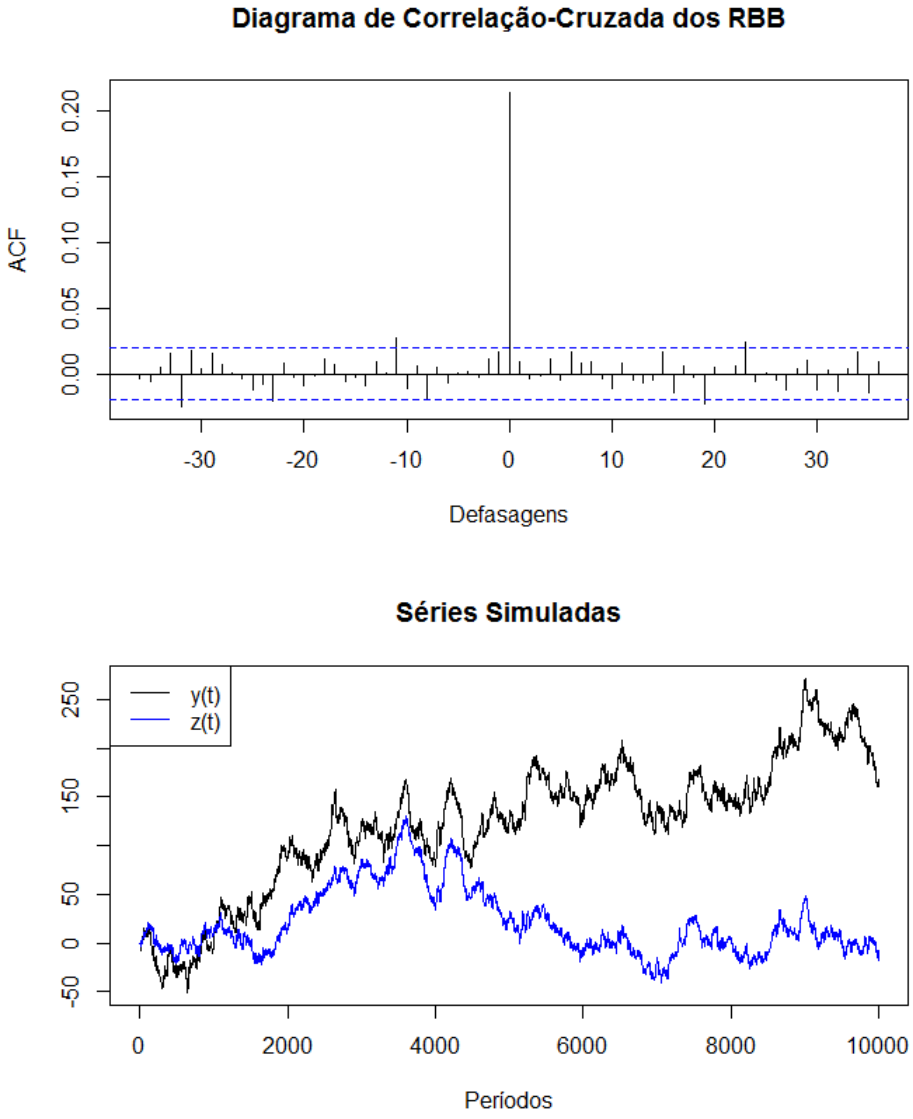

Figura 17 – Diagrama de Correlação-Cruzada dos Ruídos Brancos Bivariados

Fonte: Elaboração do Autor

onde z_t representa o erro da regressão e os termos $\varepsilon_{1,t}$ e $\varepsilon_{2,t}$ são processos de ruído branco. Simplificadamente, a equação acima implica que as mudanças em y_t são explicadas por sua própria história, representada por $\sum_{i=1}^{L} \psi_{2,i}\Delta y_{t-i}$, mudanças defasadas de x_t, representadas por $\sum_{i=1}^{K} \psi_{1,i}\Delta x_{t-i}$, e pelo erro com relação ao equilíbrio de longo prazo do período anterior, representado por $\gamma_1 \hat{z_{t-1}}$.

1 Probabilidade: Teoria e Aplicações

> "The day will come when, by study pursued through several ages, the things now concealed will appear with evidence; and posterity will be astronished that thruts so clear had scaped us"
>
> Lucius Annaeus Seneca

1.1 Fundações Conceituais

"A curva a qual descreve uma simples molécula de ar ou vapor é regulada de maneira tão certa quanto órbitas planetárias; a única diferença entre as duas é a parte que vem da nossa ignorância"(LAPLACE, 1820).

De acordo com Mises (1939), a definição clássica de probabilidade foi dada por Laplace e foi repetida, até 1930, em praticamente todos os livros texto sobre a teoria da probabilidade com sua forma praticamente inalterada; Segundo Mises (1939, p. 66) tal definição afima que a probabilidade é a razão do número de casos favoráveis em relação ao total de casos igualmente possíveis.

A partir dos anos 30, matemáticos proeminentes começaram a se atentar de forma sistemática à inadequação da definição de Laplace. Poincaré (1912, p. 24) afirma que dificilmente é possível dar uma definição satisfatória de probabilidade. Segundo Mises (1939, p. 67) um desenvolvimento lógico da teoria com base na definição clássica de Laplace nunca foi realizado; a maior parte dos autores começam com a suposição de 'casos igualmente prováveis', somente para abandonar este ponto de vista em um momento favorável e se voltarem para a definição de probabilidade baseada em frequência.

Quando sabemos que de três ou mais eventos possíveis, apenas um ocorrerá, temos um estado de indecisão e é impossível anunciar o que ocorrerá com certeza. Em suma, quando tratamos de eventos incertos, sua forma de expressão se dá em termos de probabilidade. Ainda segundo Laplace (1820), a teoria das chances consiste em reduzir todos os eventos do mesmo tipo a um certo número de casos igualmente possíveis, isto é, de forma que o observador esteja igualmente indeciso quanto à existência de cada um dos casos bem como quanto à determinação do número de casos favoráveis em relação ao evento cuja probabilidade é procurada. A razão entre este número de casos favoráveis em relação à todos os casos possíveis é a medida desta probabilidade, a qual é simplesmente uma fração cujo numerador é a quantidade de casos favoráveis e o denominador é a quantidade de casos possível - tal noção de probabilidade supõe

que a probabilidade se mantém constante com o aumento da quantidade de casos favoráveis e desfavoráveis na mesma proporção que a probabilidade observada inicialmente.

Mises (1939) define a teoria de probabilidade como uma ciência similar às outras. O argumento central se opõe àqueles que afirmam que a teoria de probabilidade é uma ciência fundamentalmente diferente das outras ciências, sendo governada por um tipo especial, ou diferente, de lógica. A principal distinção se da na forma como se chega às principais conclusões e resultados: "enquanto nas outras ciências os principais resultados são encontrados a partir do que se sabe, a ciência de probabilidade deriva os seus resultados mais importantes a partir do que não se sabe"(MISES, 1939, p. 30) . "Nossa absoluta falta de conhecimento à respeito das condições sobre as quais um dados cai, nos leva a concluir que cada lado do dado possui uma probabilidade de 1/6 de cair em determinada posição como, por exemplo, virado para cima"(MISES, 1939 apud CZUBER; BURKHARDT, 1921).

Como qualquer outra ciência natural, a teoria da probabilidade é iniciada a partir de observações, as ordena, classifica e deriva delas certos conceitos básicos e leis e, finalmente, por meio de sua lógica universal e aplicável, retira conclusões que podem ser testadas em comparação a resultados experimentais. Em suma, Von Mises define a teoria de probabilidade como uma ciência normal, distinta das demais por um objeto de estudo especial, e não por um método especial de raciocínio.

Portanto, como qualquer outra ciência, um estudo da teoria da probabilidade deve ser iniciado a partir de seus componentes mais fundamentais; iniciamos pela distribuição de elementos em um coletivo. Tais elementos se distinguem a partir de certos atributos, os quais podem ser números, no caso de um jogo de dados, cores, como em uma roleta ou outra propriedade observável. O menor número de diferentes atributos em um coletivo é 2; onde chamamos tal configuração de uma *alternativa simples*. Em tal coletivo simplificado, há apenas duas probabilidades, onde obviamente, a soma das mesmas deve ser igual à 1. O jogo de "cara ou coroa"com uma moeda é um exemplo de tal configuração, onde as faces da moeda representam os dois atributos distintos; em condições normais, cada um dos atributos possui a mesma probabilidade de ocorrência, 1/2. Ao se jogar um dado, pode-se obter seis resultados distintos, um para cada face do dado. Se a aparição de cada uma das faces do dado for igualmente verossímil, todas as probabilidades singulares possuem valor 1/6 e chamamos um dados deste tipo de um dado não-viesado. No entanto, tal dado pode ser viesado, e apesar de somarem um, as probabilidades singulares podem não ser iguais à 1/6.

É importante que tenhamos uma expressão curta para denotar o total de probabilidades associadas à diferentes atributos de um coletivo Mises (1964) propõe o termo *distribuição* para tal fim. Quando pensamos na distribuição das probabilidades em um jogo, as razões para a escolha do termo são facilmente entendidas.

Definido conceito de uma distribuição, podemos passar para o objeto de estudo da

teoria de probabilidade.Tal objeto de estudo é composto basicamente por longas sequência de experimentos ou observações, repetidas diversas vezes sob condições inalteradas. Em outras palavras, observamos os diversos resultados decorrentes de medidas de uma mesma quantidade, onde o mesmo procedimento de medida é repetido diversas vezes.

Feller (1968) define o resultado de experimentos ou observações que respeitam as condições citadas como *eventos*. Tais eventos podem ser ainda *simples* ou *compostos*, tal definição leva em consideração a possibilidade ou não da decomposição do evento em questão. Por exemplo, mostrar que o lançamento de dois dados resultou em um "soma igual à seis"é equivalente a dizer que o mesmo resultou em "(1, 5) ou (2,4) ou (3,3) ou (5,2) ou (5,1)", onde tal enumeração decompõe o evento "soma igual à seis"em cinco eventos simples. Como segundo exemplo, quando consideramos a idade de uma pessoa, cada valor particular x representa um *evento simples*, ao passo que a afirmação de que uma pessoa está na faixa dos 50 anos descreve o evento onde x se encontra entre 50 e 60 anos. A partir destes exemplos, vemos que todo *evento composto* é um agregado de certos *eventos simples*.

Feller (1968, p. 9) destaca que, para tratarmos de experimentos ou observações de forma teórica e livre de ambiguidade, devemos primeiro concordar com a abordagem onde eventos simples representam todos os resultados possíveis; tais eventos, por sua vez definem o *experimento idealizado*. Feller (1968) denota os *eventos simples* como *pontos amostrais*, ou somente *pontos*. Por definição, cada resultado indecomponível do experimento idealizado é representado por um, e apenas um, ponto amostral. O agregado de todos os pontos amostrais será denominado o *espaço amostral*. Fica claro a partir destas definições que só devemos nos referir à probabilidades quando tratando de um espaço amostral definido.

Meyer (1970) destaca que a fim de descrever um espaço amostral associado a um experimento, devemos ter uma ideia bastante clara do que estamos mensurando ou observando: "Devemos tratar de "um"espaço amostral associado a um experimento, e não de "o"espaço amostral"(MEYER, 1970, p. 11). Quanto ao estudo da quantidade de resultados em um espaço amostral, há três possibilidades: o espaço amostral pode ser finito, infinito enumerável ou infinito não-enumerável.

Com o conceito de espaço amostral definido, podemos nos voltar ao estudo dos eventos para uma definição mais robusta. De acordo com Feller (1968), o espaço amostral nos dá o modelo de um experimento ideal, no sentido de que todo resultado imaginável do experimento é completamente descrito por um, e somente por um, ponto amostral. Sendo assim, só faz sentido tratarmos de um evento A quando for possível verificar, para cada resultado do experimento, se o evento A ocorreu ou não. A coleção de todos os pontos amostrais que representam resultados nos quais A ocorreu descrevem completamente tal evento. Inversamente, qualquer agregado A contendo um ou mais pontos amostrais os quais representam resultados onde A ocorreu podem ser denominados como um evento. Meyer (1970), por sua vez, define um evento A como um

conjunto de resultados possíveis; utilizando a terminologia dos conjuntos, um evento é um *subconjunto* do espaço amostral.

Definidos o espaço amostral, eventos e pontos, só nos resta um elemento para podermos definir a noção de probabilidade formalmente: a frequência. Consideremos um espaço amostral finito S contendo k eventos distintos $a_1, a_2, ..., a_k$, e uma sequência x_j onde cada x_j é um dos eventos a_i. Considerando a_i como um evento fixo; dentre os n primeiros elementos da sequência x_j, haverá uma quantidade n_i de resultados iguais ao representado pelo evento a_i. Tal quantidade, n_i depende tanto de n quanto de a_i, além da própria sequência x_j. Definimos a *frequência*, ou *frequência relativa* de a_i dentre os primeiros n elementos de x_j como n_i/n.

Segundo Mises (1964), é necessário que o conceito de frequência seja independente da quantidade de elementos da sequência n . Para tal, definimos que uma sequência x_j pode ser indefinidamente estendida e que a frequência n_i/n se aproxima de um limite a medida que n tende ao infinito:

$$lim_{n\to\infty}\frac{n_i}{n} = p_i, \qquad i = 1,2,...,k. \tag{1.1}$$

Tal limite, para o qual, necessariamente, $0 \leq p_i \leq 1$ é então denominado *frequência de limitação* do evento a_i. Mises (1964) define então que a teoria matemática da probabilidade diz respeito a sequências infinitas x_j de eventos, nas quais cada evento distinto a_i possui uma *frequência de limitação* p_i, a qual também é denominada *probabilidade* do evento a_i dentro da sequência x_j.

A probabilidade de ocorrência do espaço amostral deve ser igual a 1, o que nos leva a construção do conceito de *probabilidade total*. Para o caso dos k eventos $a_1, a_2, ..., a_k$, chegamos à conclusão de que

$$\sum_{i=1}^{k} p_i = lim_{n\to\infty}\sum_{i=1}^{k}\frac{n_i}{n} = lim_{n\to\infty}\frac{n}{n} = 1 \tag{1.2}$$

Assumindo que mesmo para o caso de uma quantidade infinita enumerável de eventos distintos a série $\sum_{i=1}^{\infty} p_i$ converge, obtemos

$$\sum_{i=1}^{r} p_i = \sum_{i=1}^{r} lim_{n\to\infty}\frac{n_i}{n} = lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{r} n_i \leq lim_{n\to\infty}\frac{n}{n} = 1, \tag{1.3}$$

de forma que as somas parciais de $\sum_{i=1}^{\infty} p_i$ são vistas como não decrescentes e limitadas superiormente. Meyer (1970) ressalta que, dadas as propriedades definidas acima, se um determinado experimento for executado repetidas vezes, a frequência da ocorrência de um certo evento a_i tenderá a variar cada vez menos a medida que o número de repetições aumentar. Tal

característica é conhecida como *regularidade estatística* e é fundamental para a construção de alguns dos teoremas detalhados nas próximas seções.

1.2 Fundações Matemáticas

Para o estudo das distribuições de probabilidade, é de importante traduzir os elementos s_i de um espaço amostral em um números reais. Para tanto, é conveniente que tenhamos uma função que realize tal mapeamento. Uma função X, que associa a cada elemento $s \in S$ um número real, $X(s)$, é denominada *variável aleatória*. De acordo com Meyer (1970) a definição acima pode gerar bastante confusão pois, apesar de nos referirmos a uma "variável"aleatória, estamos tratando de uma função.

Considerando um espaço amostral discreto, uma forma conveniente de descrever uma distribuição de probabilidade é a partir de uma *função de distribuição* ou *função de distribuição acumulada*. Consideremos $a_1, a_2, ...$ eventos com probabilidades de ocorrência definidas por $p_1, p_2, ...$

A *funcão de distribuição acumulada* $F(x)$ é denotada por

$$F(x) = \sum_{a_i \leq x} p_i. \tag{1.4}$$

A interpretação é que a função $F(x)$ nos da a probabilidade de um resultado que menor ou igual à x; tal função é contínua à direita $F(x+0) = F(x)$, possui saltos de magnitude p_i nos pontos a_i e durante o intervalo entre os saltos, $F(x)$ é constante. Tal função descreve a distribuição em um espaço amostral S enumerável, onde a família de subconjuntos de S para os quais foram atribuídas probabilidades consiste em todos os subconjuntos de S. Para o caso de um espaço amostral não enumerável, por exemplo um intervalo finito, não podemos atribuir probabilidades a todos os subconjuntos do intervalo; pois isso geraria inconsistências. Precisamos então de uma abordagem diferente.

Consideremos S um conjunto de eventos, e T um conjunto de subconjuntos de S; onde T contém S como um elemento. Se A e B pertencem à T, o mesmo deve ser válido para $A^{'}$ e $B^{'}$, onde $A^{'} = S - A$ denota o complemento de A. $A^{'}$ contém todos os elementos de S que não pertencem à A e, como T contém S, também deve conter $S^{'}$; o qual é o conjunto vazio, denotado por ν. Uma função P é definida em uma família T de subconjuntos de S se é atribuído um número não negativo $P(A)$ para todo conjunto $A \in T$. Tal função P é então definida como a *distribuição de probabilidade* na σ-álgebra T, a qual forma com S um espaço mensurável; consultar Fernandez (1976) para um tratamento mais profundo sobre espaços mensuráveis e medidas de Jordan. Em se tratando de espaços amostrais não enumeráveis, Mises (1964) destaca

que, correspondendo a cada declaração sobre a probabilidade de ocorrência de um evento, existe uma forma de verificação através de um experimento de frequência; pelo menos em princípio.

Passamos então ao detalhamento das propriedades da função $F(x)$. Consideremos S um intervalo aberto $(-X,+X)$, e T, composto por todos os subconjuntos de S e por fim, consideremos P como a probabilidade definida sobre T. Se o conjunto A é o intervalo $(a,b]$, $P(A) = P(a < \xi \leq b)$ é definida uma *função intervalo*. Com base em Mises (1964, p. 92), dadas uma função intervalo $P(A)$ e alguma constante k, definimos uma função ponto $F(x;k)$ como

$$Pr(k < \xi \leq x) \quad para \quad x > k \tag{1.5}$$

$$F(x;k) = 0 \quad para \quad x = k \tag{1.6}$$

$$-Pr(x < \xi \leq k) \quad para \quad x < k. \tag{1.7}$$

Para qualquer valor de k, encontramos para um intervalo finito (a, b], tal que

$$F(b;k) - F(a;k) = P(a < \xi \leq b) = P(A) \geq 0 \tag{1.8}$$

que mostra que $F(x;k)$ é uma função não decrescente de x. Consideremos agora, que x representa b e y representa a. Temos então

$$F(x) - F(y) = P(y < \varepsilon \leq x) = P(A_{y,x}) = Pr(y < \varepsilon \leq x), \tag{1.9}$$

onde $A_{y,x}$ é o intervalo $y < \varepsilon \leq x$. À medida que $x \to -\infty$, $y \to -\infty$. Definimos então que

$$\lim_{x \to -\infty} F(x) = 0, \tag{1.10}$$

Agora, mantendo x fixo a medida que $y \to -\infty$, obtemos

$$\lim_{y \to -\infty} [F(x) - F(y)] = lim_{y \to -\infty} P(y < \varepsilon \leq x) = Pr(-\infty < \varepsilon \leq x) \tag{1.11}$$

onde pela equação 1.11

$$F(x) = Pr(-\infty < \varepsilon \leq x) = Pr(\varepsilon \leq x); \tag{1.12}$$

Desta forma, $F(x)$ representa a possibilidade de obtermos um resultado que é menor ou igual à x (MISES, 1964). $F(x)$ é denominada a *função de distribuição* da *distribuição de probabilidade* $P(A)$

Tal *funcão de distribuição* é definida como uma função monotônica não-decrescente, com $F(-\infty) = 0,\ F(+\infty) = 1$.

Assumimos que a função $F(x)$ é contínua em $[a,\ b]$ e que em todos os pontos de $[a,\ b]$ existe uma derivada contínua $F'(x)$. Com uma generalização deste caso, podemos assumir que $F(x)$ é absolutamente contínua, o que significa que $F(x)$ é uma integral indefinida no sentido de Riemann; consultar Edwards (1974) para um tratamento mais profundo das integrais de Riemann.

Consideremos E como qualquer conjunto em um intervalo i, que pode ser, por exemplo, $[a,\ b] : E \in i$, a *função característica* $\theta(x)$ de E é a função igual à 1 em E e igual à 0 no complemento de E, E'.

Com E sendo uma *medida de jordan* linear em i e considerando

$$\int_a^b \theta(x), f(x), dx \tag{1.13}$$

a integral de Riemann existe e obtemos

$$\int_a^b \theta(x), f(x), dx = \int_{(E)} f(x), dx = P(E), \tag{1.14}$$

onde $P(E)$ é a medida de Jordan com respeito à $p(x)$ de E. Se, como no caso em questão, $F(x)$ é absolutamente contínua e, em particular, se em qualquer $F'(x) = f(x)$ existe e é contínua, a distribuição é denominada *contínua* e $f(x)$ é denominada sua *densidade*, dada por

$$\int_{-\infty}^{+\infty} f(x), dx = 1, \tag{1.15}$$

onde a densidade é não negativa e obtemos

$$F(x) = \int_{-\infty}^{x} f(t), dt \tag{1.16}$$

Vemos então que com $F(x)$ contínua, e com $\theta(x)$ a *função característica* de E, a integral

$$\int_a^b \theta(x), dF(x) = \int_{(E)} dF(x) = P(E) \tag{1.17}$$

existe se, e somente se, E é uma medida de Jordan e se a função $F(x)$ é absolutamente contínua (MISES, 1964).

É importante notar que a função característica e a função de densidade possuem uma forte relação (MEYER, 1970, p. 69). A função característica de qualquer variável aleatória

que assume valores reais define completamente sua distribuição de probabilidade. A função característica é a transformada de Fourier Inversa da função de densidade (FIGUEIREDO, 2000, p. 206).

Por conta de suas características, a transformada de Fourier nos da acesso a uma rota alternativa para obtenção de resultados analíticos, comparativamente ao manuseio direto de funções de densidade ou funções acumuladas. No campo da Análise Matemática, nos referimos usualmente à (integral) transformada de Fourier e à séries de Fourier. De acordo com Figueiredo (2000), uma transformada de Fourier é a integral de Fourier de uma função, f, definida na reta $\mathbb{R}$. Considerando a função f como um sinal analógico, seu domínio de definição é o tempo contínuo, e como a informação espectral é dada em termos de frequência, o domínio de definição da transformada de Fourier, $\hat{f}$, o qual é $\mathbb{R}$, é denominado domínio de frequência.

1.3 Propriedades Básicas de Distribuições

Com base em Mises (1964), consideremos um conjunto composto por uma distribuição discreta, com variáveis aleatórias $x_1, x_2, ..., x_k$. Supondo que o *i-ésimo* evento x_i tenha ocorrido n_i vezes dentre os n primeiros elementos do conjunto, a expressão

$$\frac{1}{n}\sum_{i=1}^{k} n_i x_i = x_1\frac{n_1}{n} + x_2\frac{n_2}{n} + ... + x_k\frac{n_k}{n} \tag{1.18}$$

representa o valor médio para os n primeiros elementos. De acordo com nossa definição de probabilidade p_i, a medida que $n \to \infty$, a equação 1.18 tende ao limite

$$a = \sum_{i=1}^{k} x_i p_i = \sum_{i=1}^{k} x_i p(x_i) \tag{1.19}$$

A variável a definida pela equação 1.20 é denominada a *média* da distribuição em consideração. Para o caso de uma distribuição contínua unidimensional com densidade $p(x)$, a definição da *média* toma a forma

$$a = \int_{-\infty}^{+\infty} xp(x)d(x), \tag{1.20}$$

onde a média a só existe se a integral acima convergir.

De acordo com (MISES, 1964), se pensarmos na distribuição de probabilidade como a distribuição de uma massa, um conceito importante seria o grau de concentração da massa em torno de seu centro. Tal medida de concentração é definida como *variância*, que apresenta

relação próxima ao momento de inércia de uma massa. Definimos a variância para os casos discreto e contínuo por

$$s^2 = \sum_{i=1}^{k} (x_i - a)^2 p(x_i) \tag{1.21}$$

e

$$s^2 = \int_{-\infty}^{+\infty} (x-a)^2 p(x) d(x). \tag{1.22}$$

A raiz quadrada positiva da variância é denominada o *desvio padrão* da variável aleatória. Por vezes, nos referimos à variância ou desvio padrão da distribuição, sem mencionarmos diretamente a variável aleatória.

Continuando a caracterização das distribuições de probabilidade, consideremos $f(x_i)$ uma função definida para todos os valores x_i de uma distribuição discreta $p(x_i), i = 1, 2, ...$ Para o caso contínuo, é necessário que $f(x)$ seja uma função definida e contínua no domínio $p(x) > 0$. Definimos a *esperança* de f com relação à distribuição como

$$E[f] = \sum_i f(x_i) p(x_i) \tag{1.23}$$

ou

$$E[f] = \int f(x) p(x) dx, \tag{1.24}$$

onde assumimos que o somatório ou a integral convergem. O termo $f(x)$ pode ser considerado uma variável aleatória, e como E depende da função f, é denominado um *funcional*. Dadas as definições acima, a *média* e *variância* de uma distribuição podem ser escritas como

$$a = E[x], \qquad s^2 = E[(x-a)^2] = E[x^2] - (E[x])^2 \tag{1.25}$$

Em suma, a *média* é dada pela *esperança* de x e a variância é dada pela *esperança* de $(x-a)^2$.

Tais definições são extremamente importantes para que possamos interpretar os resultados de um experimento e extrair conclusões de uma a partir da observação de uma determinada série . É necessário, no entanto, que possamos interpretar os resultados decorrentes de experimentos que envolvem duas ou mais variável aleatórias.

Lembremos que para o caso de uma variável aleatória X assumindo valores $x_1, x_2, ...$, o agregado de todos os pontos amostrais nos quais X forma o evento $X = x_j$ possui probabilidade denotada por $P(X = x_j)$.

A distribuição da variável aleatória X é representada pela função

$$P(X = x_j) = f(x_j), \tag{1.26}$$

onde $\sum f(x_j) = 1$.

Consideremos agora, duas variáveis aleatórias X e Y definidas no mesmo espaço amostral e os valores assumidos por elas dados por $(x_1, x_2, ...)$ e $(y_1, y_2, ...)$, respectivamente. O agregado dos pontos onde as duas condições $X = x_j$ e $Y = y_k$ são satisfeitas forma um evento cuja probabilidade é denotada por $P(X = x_j, Y = y_k)$.

Denominamos a função

$$P(X = x_j, Y = y_k) = p(x_j, y_k) \qquad (j,k = 1,2,...) \tag{1.27}$$

como a *distribuição de probabilidade conjunta* de X e Y, onde

$$p(x_j, y_k) \geq 0, \qquad \sum_{j,k} p(x_j, y_k) = 1. \tag{1.28}$$

Para cada j constante, temos

$$p(x_j, y_1) + p(x_j, y_2) + p(x_j, y_3) + ... = P\{X = x_j\} = f(x_j) \tag{1.29}$$

e para cada k constante, temos

$$p(x_1, y_k) + p(x_2, y_k) + p(x_3, y_k) + ... = P\{Y = y_k\} = g(y_k). \tag{1.30}$$

Percebemos que a partir da soma das probabilidades obtemos as distribuições de probabilidade de X e Y, denominadas *distribuições marginais*.

Quando temos um sistema composto por duas ou mais variáveis aleatórias, a relação entre as mesmas passa a ser extremamente importante. Introduzimos então o conceito de *distri-*

buição de probabilidade condicional. Utilizando a notação da equação 1.28, a *probabilidade condicional* do evento $Y = y_k$ dado que $X = x_j$ é denotada por

$$P\{Y = y_k | X = x_j\} = \frac{p(x_j, y_k)}{f(x_j)}, \tag{1.31}$$

onde um número é associado a cada valor assumido por X e a equação acima define uma função de X. Denotamos então, a *distribuição condicional de probabilidade* por $P\{Y = y_k \mid X\}$.

Tal definição implica que as variáveis aleatórias X e Y são *estocasticamente dependentes*. O grau mais elevado de dependência é observado quando Y é uma função de X, ou seja, quando o valor de X determina completamente Y.

Quando $p(x_j, y_k) = f(x_j)g(y_k)$ para todas as combinações de x_j, y_k, os eventos $X = x_j$ e $X = y_k$ são denominados *eventos independentes*; neste caso tratamos de X e Y como *variáveis aleatórias independentes*. Tal resultado pode ser extrapolado para sistemas de várias variáveis aleatórias.

Feller (1968) define que uma variável aleatória X é uma função definida em um determinado espaço amostral, a qual atribui um número real para cada ponto amostral. Se duas variáveis X e Y são definidas no mesmo espaço amostral, sua distribuição de probabilidade conjunta é dada por pela equação 1.27 e atribui probabilidades a todas as combinações $(x_j, y_k$ de valores assumidos por X e Y. Tal definição se aplica para qualquer conjunto finito de variáveis aleatórias $X, Y, ..., W$ definidas em um mesmo espaço amostral. Tais variáveis são denominadas mutuamente independentes se, para cada combinação de valores $(x, y, ..., w)$ assumida por elas, obtivermos

$$P\{X = x, Y = y, \ldots, W = w\} = P\{X = x\}\, P\{Y = y\} \ldots P\{W = w\}. \tag{1.32}$$

Finalmente, se $X, Y, ..., W$ são variáveis aleatórias definidas em um mesmo espaço amostral, então qualquer função $F(X, Y, ...W)$ é também uma variável aleatória. Feller (1968) afirma que é conveniente pensar intuitivamente em probabilidades como *limites de frequências observáveis em experimentos repetidos*.

Dadas as definições relativas à sistemas compostos por mais de uma variável, podemos abordar a definição da média e da variância de distribuições que compostas por mais de uma variável aleatória. Se $X_1, X_2, \ldots, X_n$ são variáveis aleatórias com esperanças definidas, a esperança de sua soma existe e é definida como a soma de cada uma das esperanças individuais

$$E(X_1 + \ldots + X_n) = E(X_1) + \ldots + E(X_n) \tag{1.33}$$

pois

$$E(X)+E(Y)=\sum_{j,k}x_j p(x_j,y_k\ +\ \sum_{j,k}y_k p(x_j,y_k), \tag{1.34}$$

onde o somatório se estende sobre todos os valores possíveis de x_j, y_k e as duas séries convergem absolutamente. Podemos então escrever $\sum_{j,k}(x_j+y_k)p(x_j,y_k)$, que é a própria definição da esperança de **X + Y**.

No entanto, não há um teorema geral válido para produtos de variáveis aleatórias; por exemplo, $E(X^2)$ é geralmente diferente de $(E(X))^2$. Se **X** e **Y** são variáveis aleatórias com esperança finita, então seu produto é uma variável aleatória com esperança finita e temos

$$E(XY)=E(X)E(Y) \tag{1.35}$$

onde, para calcular $E(XY)$ devemos multiplicar cada valor possível de $x_j y_k$ por sua respectiva probabilidade de ocorrência e obtemos

$$E(XY)=\sum_{j,k}x_j y_k f(x_j)g(y_k)=\{\sum_j x_j f(x_j)\}\ \{\sum_k y_k g(y_k)\} \tag{1.36}$$

onde assumimos convergência absoluta para as séries e os resultados acima são válidos para qualquer quantidade de variáveis aleatórias mutuamente independentes.

Se X e Y são variáveis aleatórias com distribuição conjunta de probabilidade dada por 1.27, a esperança condicional $E(Y\mid X)$ de Y para um dado X é dada pela função a qual no ponto x_j assume o valor

$$\sum_k y_k P\{Y=y_k\mid X=x_j\}=\frac{\sum_k y_k p(x_j,y_k)}{f(x_j)} \tag{1.37}$$

onde assumimos novamente que as série convergem absolutamente e $f(x_j)>0\forall j$.

Ambas a média e a variância são exemplos de momentos de uma distribuição, e é adequado procedermos rumo a uma caracterização dos mesmos.

Feller (1968) define que sempre que o $r-esimo$ momento de uma variável aleatória X existir, o $(r\ -\ 1)-esimo$ momento também existirá. Considerando a variável aleatória X , com distribuição $\{f(x_j)\}$, e definindo $r>0$ um número inteiro, se a esperança da variável aleatória X^r , dada por

$$E(X^r)=\sum x_j^r f(x_j \tag{1.38}$$

existir, então a mesma é denominada o $r-esimo$ momento de X com relação à origem. De acordo com estas definições, a média é o primeiro momento de uma variável aleatória e a variância é o segundo. Meyer (1970) introduz uma definição que complementar, explicitando o racional por trás dos resultados descritos acima.

Sendo X uma variável aleatória discreta, a qual possui distribuição de probabilidade dada por $p(x_i) = P(X = x_i\,,\ i = 1,2,\ldots\,,$ a função

$$M_x(t) = \sum_{j=1}^{\infty} e^{tx_j} p(x_j) \tag{1.39}$$

é a *função geratriz de momentos* de X . Caso X seja uma variável aleatória contínua com função densidade de probabilidade f , a *função geratriz de momentos* se torna

$$M_x(t) = \int_{-\infty}^{+\infty} e^{tx} f(x) d(x). \tag{1.40}$$

Em ambos os casos, $M_x(t)$ é o valor esperado de e^{tx} :

$$M_x(t) = E(e^{tx}) \tag{1.41}$$

Para entendermos a razão da denominação da função *geratriz de momentos*, devemos recordar como se da o desenvolvimento de uma função e^x em uma série de MacLaurin. Com base em Meyer (1970), a expansão é dada por

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!} + \ldots \tag{1.42}$$

onde assumimos a série converge para todos os valores de x e obtemos

$$e^{tx} = 1 + x + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!} + \ldots + \frac{(tx)^n}{n!} + \ldots \tag{1.43}$$

Admitindo que algumas condições gerais são satisfeitas Com base em Meyer (1970, p. 229), temos que o valor esperado da soma é igual à soma dos valores esperados

$$M_x(t) = E(e^{tx}) = E(1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \ldots + \frac{(tX)^n}{n!} + \ldots) \tag{1.44}$$

e, dado que t é uma constante, obtemos

$$M_x(t) = 1 + tE(X) + \frac{t^2E(X^2)}{2!} + ... + \frac{t^nE(X^n)}{n!} + ... \tag{1.45}$$

Dado que M_x é uma função da variável real t, podemos tomar a derivada de $M_x(t)$ em relação à t, a qual denotamos $M'(t)$. Assumindo as mesmas condições gerais da equação , temos

$$M'(t) = E(X) + tE(X^2) + \frac{t^2E(X^3)}{2!} + \ ... \ + \frac{t^{n-1}E(X^n)}{(n-1)!} + \ ... \tag{1.46}$$

Quando $t = 0$, verificamos que

$$M'(0) = E(X) = \mu. \tag{1.47}$$

Ou seja, a derivada primeira da *função geratriz de momentos* nos dá a média da variável aleatória. Segue que a partir do cálculo da derivada segunda de $M_x(t)$, obtemos:

$$M''(t) = E(X^2) + tE(X^3) + ... + \frac{t^{n-2}E(X^n)}{(n-2)!} + ... \tag{1.48}$$

Quando avaliamos a expressão acima no ponto $t = 0$, obtemos

$$M''(0) = E(X^2). \tag{1.49}$$

Estendendo os cálculos, e admitindo que $M^n(0)$ exista, obtemos

$$M^{(n)}(0) = E(X^n), \tag{1.50}$$

o qual nos diz que a n-ésima derivada de $M_x(t)$, avaliada no ponto 0 fornece o *momento de ordem n* da variável aleatória $\mathbf{X}$, denotado por $E(X^n)$.

Com a definição dos momentos de primeira e segunda ordem, podemos agora introduzir dois conceitos extremamente importantes: a *covariância* e a *correlação*.

Com base em Feller (1968), consideremos duas variáveis aleatórias X e Y, pertencentes ao mesmo espaço amostral. Sabemos que $X + Y$ e XY são também variáveis aleatórias e

podemos obter suas distribuições a partir das distribuições de X e Y. Se a distribuição de X e Y é dada por $\{p(x_j, y_k\}$, a esperança de XY é dada por

$$E(XY) = \sum x_j y_k p(x_j, y_k), \tag{1.51}$$

assumindo convergência absoluta das séries.

A esperança de XY existe se $E(X^2)$ e $E(Y^2)$ existirem. Adicionalmente, temos que os valores esperados de X e Y existem e são iguais às médias, definidas como

$$\mu_x = E(X), \qquad \mu_y = E(Y), \qquad e \qquad E(X-\mu_x) = E(Y-\mu_y) = 0. \tag{1.52}$$

O produto das médias é dado por

$$E((X-\mu_x)(Y-\mu_y)) = E(XY) - \mu_x E(Y) - \mu_y E(X) + \mu_x\mu_y = E(XY) - \mu_x\mu_y \tag{1.53}$$

e a *covariância* de X e Y

$$Cov\,(X, Y) = E((X-\mu_x)(Y-\mu_y)) = E(XY) - \mu_x\mu_y. \tag{1.54}$$

existe quando X e Y possuem variâncias finitas.

A partir desta definição, chegamos a alguns resultados importantes. Em primeiro lugar, sabemos que se X e Y são variáveis aleatórias independentes, temos que a $E(XY) = E(X)E(Y)$ e obtemos $Cov\,(X, Y) = 0$. Adicionalmente, se $X_1, X_2, ..., X_n$ são variáveis aleatórias com variância $\sigma_1^2, ..., \sigma_n^2$ e soma $S_n = X_1 + ... + X_n$, obtemos

$$Var(S_n) = \sum_{k=1}^{n} \sigma_k^2 + 2\sum_{j,k} Cov(X_j, X_k) \tag{1.55}$$

onde $Var(S_n) = \sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2$ se as variáveis aleatórias X_j forem mutuamente independentes.

Definida a covariância, passamos para a definição do conceito de *correlação*. Feller (1968) descreve o uso de coeficientes de correlação como uma representação "sofisticada"da covariância.

Consideremos novamente duas variáveis aleatórias, X e Y, as quais possuem média $\mu_x e \mu y$ e variância σ_x^2 e σ_y^2.

Definimos as variáveis padronizadas como

$$X^* = (X-\mu)/\sigma \qquad e \qquad Y^* = (Y-\mu)/\sigma \tag{1.56}$$

e definimos suas respectivas covariâncias como o *coeficiente de correlação* de X , Y , o qual denotamos por

$$\rho(X,Y) = Cov(X^*,\, Y^*) = \frac{Cov(X,\, Y)}{\sigma_x \sigma_y}. \tag{1.57}$$

Note que quando X e Y são independentes $\rho(X,Y) = 0$.

Segundo Feller (1968), é importante notar que, dadas apenas a *média* e a *variância* de uma distribuição, não podemos determina-la por completo. No entanto, dado o conhecimento da média μ e da variância σ^2, podemos derivar uma desigualdade para a diferença $P(\mu+K) - P(\mu-K)$, onde K é um número positivo qualquer maior que σ^2 (MISES, 1964, p. 116). Portanto a probabilidade da variável aleatória X cair dentro do intervalo aberto $(\mu-K<X<\mu+K)$ possui um limite inferior definido.

Com base em Mises (1964), definindo A como o intervalo aberto $(\mu-K<X<\mu+K)$ e A' como a região complementar composta pelos pontos $\mu \leq \mu-K$ e $X \geq \mu+K$, obtemos

$$\sigma^2 = \int_{A'} (x-\mu)^2 p(x)d(x) + \int_A (x-\mu)^2 p(x)d(x) \geq \int_{A'} (x-\mu)^2 p(x)d(x). \tag{1.58}$$

Como $(x-\mu)^2$ e $p(x)$ são não-negativos e o valor mínimo de $(x-\mu)^2$ na região complementar A' é X^2, obtemos

$$\sigma^2 \geq X^2 \int_{A'} p(x)d(x) = X^2 P(A') = X^2(1-P(A)), \tag{1.59}$$

onde $P(A')$ e $P(A)$ denotam a probabilidade de que x caia dentro das regiões A' e A, respectivamente.

Resolvendo 1.59 para $P(A)$, chegamos na *Desigualdade de Tchebycheff*, denotada por

$$P(A) \geq 1 - \frac{\sigma^2}{X^2}. \tag{1.60}$$

Ao invés de escrever $P(A)$, podemos escrever $Pr(x \in A)$ ou $Pr(|x-a|<X)$ e obter

$$Pr(|x-a|<X) \geq 1 - \frac{\sigma^2}{X^2} \tag{1.61}$$

e ainda

$$P(A') = Pr(|x-a| \geq X) \leq \frac{\sigma^2}{X^2} \tag{1.62}$$

De acordo com Mises (1964), a grande importância da desigualdade de Tchebycheff está em sua validade "universal". Já Meyer (1970), afirma que a desigualdade de Tchebycheff nos fornece meios de compreender precisamente como a variância mede a variabilidade em relação ao valor esperado de uma variável aleatória. Uma generalização importante da desigualdade de Tchebicheff é dada pela *desigualdade de Kolmogorov*, a qual é peça central para a derivação da *lei dos grandes números*.

Consideremos $X_1, X_2, ..., X_n$ variáveis aleatórias mutuamente independentes com esperança $\mu_k = E(X_k)$ e variância σ_k^2. Adicionalmente, definimos

$$S_k = X_1 + ... + X_k \tag{1.63}$$

$$m_k = E(S_k) = \mu_1 + ... + \mu_k \tag{1.64}$$

$$s_k^2 = Var(S_k) = \sigma_1^2 + ... + \sigma_k^2. \tag{1.65}$$

Segue-se então que, para cada $t > 0$, a probabilidade $p(x_n)$ da realização simultânea das n inequações

$$|S_k - m_k| < ts_n, \qquad k = 1, 2, ..., n, \tag{1.66}$$

é dada por

$$p(x_n) \geq 1 - t^{-2}. \tag{1.67}$$

Por fim, percebemos que para $n = 1$ o teorema de Kolmogorov se reduz ao teorema de Tchebycheff.

A lei dos grandes números desempenha um papel central em em probabilidade e estatística. Meyer (1970) destaca o fato de que a medida que o número de repetições de um experimento cresce, a frequência relativa f_a de um evento A converge para a probabilidade teórica $P(A)$. Nesta seção, caracterizamos de forma mais precisa o que esta *convergência* significa e como ela ocorre.

Consideremos $\{X_k\}$ uma sequência de variáveis aleatórias mutuamente independentes as quais possuem uma mesma distribuição de probabilidade. Se o valor esperado $\mu = E(X_k)$ existe, então para cada $\varepsilon > 0$, à medida que $n \to \infty$:

$$P\left\{\left|\frac{X_1 + ... + X_n}{n} - \mu\right| > \varepsilon\right\} \to 0. \tag{1.68}$$

Adiconalmente, Feller (1968) afirma que o resultado acima implica que a probabilidade da média amostral S_n/n ser diferente do valor esperado por um valor arbitrariamente escolhido ε, inferior a um, tende à 1. Ademais, Meyer (1970) destaca que tal aproximação constitui apenas um caso particular de um resultado geral, o *Teorema Central do Limite*. De acordo com o autor, O Teorema Central do Limite define que sob certas condições, a soma de um grande número de variáveis aleatórias é aproximadamente normal e é um dos teoremas mais importantes da teoria da probabilidade.

Consideremos $\{X\}$ uma sequencia de variáveis aleatórias independentes, as quais possuem uma mesma distribuição de probabilidade. A média é dada por $\mu = E(X_k)$, a variância é dada por $\sigma^2 = Var(X_k)$ e a soma é dada por $S_n = X_1 + ... + X_n$. Segue que para qualquer valor fixado β:

$$P\left\{\frac{S_n - n\mu}{\sigma\sqrt{n}} < \beta\right\} \to \Re(\beta) \tag{1.69}$$

onde $\Re(x)$ é a distribuição normal. Uma relação importante entre a lei dos grandes números e o teorema central do limite, é que a primeira é válida mesmo quando as variáveis aleatórias X_k não possuem variância finita, sendo assim, um resultado mais geral.

1.4 Elementos de Análise Combinatória, Distribuições e Aplicações

De acordo com Feller (1968), nos estudos envolvendo jogos de probabilidade, procedimentos de amostragem e problemas de ocupação, usualmente lidamos com espaços amostrais finitos, onde a mesma probabilidade é atribuída para todos os pontos amostrais. Para computar a probabilidade de ocorrência de um evento A, dividimos a quantidade de pontos amostrais em A, pontos favoráveis, pela quantidade total de pontos amostrais, os quais representam todos os resultados possíveis.

Com m elementos $a_1, ..., a_m$ e n elementos $b_1, ..., b_n$, é possível formar mn pares (a_j, b_k) contendo um elemento de cada grupo.

Dados n_1 elementos $a_1, ..., a_{n_1}$ e n_2 elementos $b_1, ..., b_{n_2}$ até n_r elementos $x_1, ..., x_{n_r}$, é possível formar $n_1 \cdot n_2 \cdots n_r$ *r-uplas* $(a_{j_1}, b_{j_2}, \ldots, x_{j_r})$ contendo um elemento de cada grupo.

Consideremos a população de n elementos $a_1, a_2, ..., a_n$. Qualquer arranjo ordenado $a_{j_1}, a_{j_2}, ..., a_{j_r}$ de r elementos é denominado uma *amostra ordenada* de tamanho r retirada da população. Há dois tipos de amostragem: com e sem reposição. No primeiro caso, o mesmo elemento pode ser escolhido mais de uma vez e, no segundo, um elemento escolhido da população é removido da mesma, eliminando a possibilidade de escolhas repetidas de elementos. Na amostragem com reposição, a quantidade da amostra pode exceder a quantidade de elementos da população. Já na amostragem sem repetição, o tamanho da amostra está limitado ao tamanho da população em questão.

Na amostragem com repetição, cada um dos r elementos pode ser escolhido de n formas distintas. Portanto, temos n^r amostras possíveis, considerando que a probabilidade de ocorrência de cada um dos elementos é igual. Já para o caso de amostragem sem repetição, temos n possibilidades de escolha para o primeiro elemento, mas apenas $(n-1)$ possibilidades para o segundo e $(n-p)$ para o p-ésimo elemento. Logo, temos uma quantidade total de escolhas dada por $n(n-1)(n-2)(n-3)...(n-r+1)$ e definimos

$$(n)_r = n(n-1)(n-2)(n-3)...(n-r+1). \tag{1.70}$$

Segue que, para uma população de tamanho n e uma amostra de tamanho r , há n^r formas distintas de amostragem para o caso onde é permitida a reposição e $(n)_r$ formas distintas de amostragem para o caso onde não é permitida a reposição. Notamos que quando $n = r$, a amostragem sem reposição representa um reordenamento dos elementos. Como n elementos podem ser ordenados de

$$n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdots (n-r+1) \cdots 2 \cdot 1 \tag{1.71}$$

formas distintas, definimos o produto acima como *fatorial de n*, denotado por

$$n! = n(n-1) \cdots 2 \cdot 1. \tag{1.72}$$

Assumido probabilidades iguais para a ocorrência de cada um dos elementos, temos uma *amostragem aleatória*. O termo *escolha aleatória* está associado ao caso onde todos os resultados são igualmente verossímeis. Diversos experimentos do mundo real, como jogar um dado ou uma moeda repetidas vezes, podem ser caracterizados como aplicações de amostragem aleatória com reposição; onde as probabilidades numéricas são próximas da frequências observadas em experimentos onde há uma elevada quantidade de repetições.

Adicionalmente, também encontramos exemplos de amostragem sem repetição, como no caso da retirada sucessiva de cartas de um monte. Definimos então que, para uma amostragem

sem reposição, a probabilidade de que um dado elemento da população esteja presente na amostra aleatória é dada por

$$p(r_n) = 1 - \frac{(n-1)_r}{(n)_r} = 1 - \frac{n-r}{n} = \frac{r}{n} \tag{1.73}$$

e para o caso de amostragem com repetição, obtemos

$$p(r_n) = 1 - (1 - 1/n)^r. \tag{1.74}$$

O termo *população de tamanho n* denota um agregado de n elementos, não importando sua ordem. Definimos que duas populações são diferentes quando um ou mais elementos de uma população não estão contidos na outra. Considerando uma subpopulação de tamanho r, originada de uma população de n elementos, ao enumerar, obtemos uma amostra ordenada de tamanho r e podemos repetir este procedimento para obter outras amostras. De acordo com Feller (1968), como r elementos podem ser enumerados de $r!$ formas distintas, chegamos ao resultado de que há exatamente $r!$ amostras para cada grupo de subpopulações de tamanho r. Obtemos então que o número de subpopulações de tamanho r é dado por $(n)_r/r!$.

Definimos as expressões desta natureza como *coeficientes binomiais* , denotados por

$$\binom{n}{r} = \frac{n_r}{r!} = \frac{n(n-1)\cdots(n-r+1)}{1\cdot 2\cdots(r-1)\cdot r} \tag{1.75}$$

De acordo com a definição proposta por Feller (1968), fica claro que para uma população de tamanho n e um subconjunto de tamanho r, tal subconjunto pode assumir $\binom{n}{r}$ formas distintas. Adicionalmente, percebemos que um dado subconjunto pode ser completamente determinado tanto pelos elementos contidos, quanto pelos elementos não contidos no mesmo. Os elementos não contidos no subconjunto formam uma população de tamanho $n-r$. Há uma quantidade igual de subpopulações de tamanho r e $n-r$, e para o caso onde $1 \leq r \leq n$ obtemos que

$$\binom{n}{r} = \binom{n}{n-r} \tag{1.76}$$

e podemos denotar o coeficiente binomial por sua forma mais conhecida

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}. \tag{1.77}$$

Quando tratamos de amostras ordenadas, a posição de cada um dos elementos é um fator de diferenciação e obtemos resultados distintos, sendo necessária a definição de um outro tipo de coeficiente.

Consideremos $r_1, ...r_k$ números inteiros, tais que

$$r_1 + r_2 + \cdots + r_k = n, \qquad r_i \geq 0. \tag{1.78}$$

Com base em Feller (1968), a quantidade de formas distintas de divisão de uma população de n elementos em k pares ordenados, onde o primeiro par contém r_1 elementos, o segundo contém r_2 elementos e assim por diante é dada por

$$\frac{n!}{r_1!r_2!\cdots r_k!}, \tag{1.79}$$

onde a equação define *coeficientes multinomiais*. Observe que a ordem é importante, no sentido de que a quantidade de elementos em cada um dos pares ordenados transforma a distribuição. No entanto, a ordem de elementos dentro de cada par ordenado não é relevante.

Um exemplo de experimento enquadrado nesta configuração é o de colocar aleatoriamente r bolas em n compartimentos distintos. Tal experimento pode ser totalmente descrito pela *ocupação* dos compartimentos, representada pelos números $r_1, r_2, \cdots, r_n$ onde cada número r_n representa a quantidade de bolas contidas em cada um dos n compartimentos.

Cada n-úpla de números inteiros que satisfazem $r_1, r_2, r_3, \cdots, r_n = r$ descreve uma possível configuração para o experimento de ocupação. A quantidade de configurações distintas para um experimento de ocupação como o descrito acima é de

$$A_{r,n} = \binom{n+r-1}{r} = \binom{n+r-1}{n-1}. \tag{1.80}$$

Como exemplo, para o caso onde temos n inteiros que satisfazem $r_1 + r_2 + \cdots + r_n = r$, a quantidade de formas distintas de colocar r bolas em n compartimentos é dada por $\binom{n}{r}$. Assumindo que todos as n^r configurações possíveis são igualmente prováveis, a probabilidade de obter os números de ocupação $r_1 + r_2 + \cdots + r_n$ é igual à

$$\frac{r!}{r_1!r_2!\cdots r_n!} n^{-r}. \tag{1.81}$$

Tal configuração é conhecida em física como *Distribuição de Maxwell-Boltzmann*, nomeada em homenagem à James Clerk Maxwell e Ludwig Boltzmann e foi definida inicialmente para descrever a velocidade de partículas de gases, onde as partículas se moviam livremente dentro de um recipiente, sem que houvesse contato entre os gases, a não ser por colisões breves,

e geravam a troca de energia e momento entre os gases e o ambiente térmico (FELLER, 1968, p. 39). A distribuição avalia a probabilidade de ocorrência de uma determinada velocidade para uma partícula, selecionada aleatoriamente da população. Na teoria de probabilidade, a distribuição de Maxwell Boltzmann é uma distribuição *Chi* com três graus de liberdade e um parâmetro de escala. consultar Meyer (1970) para um tratamento mais profundo sobre distribuições de probabilidade, como a *Chi*.

Feller (1968) apresenta três abordagens distintas para problemas de ocupação, as quais são extensivamente utilizadas para descrever fenômenos da natureza e nos dão um bom fundamento para começarmos a tratar mais profundamente dos tipos diferentes de distribuição de probabilidade.

Consideremos um sistema mecânico composto por r partículas indistinguíveis e uma subdivisão do espaço amostral em n pequenas regiões ou células, de forma que cada partícula é atribuída a uma célula. Desta forma, todo o sistema é descrito em termos da distribuição de r partículas em n células. Segundo Feller (1968), a princípio todas as n^r configurações possíveis deveriam ter probabilidades iguais e, se este for o caso, temos um sistema que respeita a estatítica de Maxwell-Boltzmann. No entanto, apesar de várias tentativas de mostrar que partículas se comportam de acordo com a estatítica de Maxwell-Boltzmann, não foi possível mostrar que isso de fato acontece. A estatística não é aplicável para nenhuma partícula conhecida, o que significa que em nenhum caso todas as combinações possíveis são equiprováveis (FELLER, 1968, p. 41).

Uma configuração alternativa considera apenas configurações distinguíveis, onde ainda possuímos o mesmo sistema de r partículas em n células e a cada configuração possível é associada uma probabilidade $\frac{1}{A_{r,n}}$, onde:

$$A_{r,n} = \binom{n+r-1}{r} = \binom{n+r-1}{n-1} = \frac{(n-r-1)!}{(n-1)!\,(r)!} \tag{1.82}$$

Tal configuração é valida para explicar o movimento de photons, nucleos e átomos que contém uma quantidade par de partículas elementares e é denominada de *estatística de Bose-Einstein*. Definindo adicionalmente que não seja possível que duas ou mais partículas sejam atribuídas a uma mesma célula e que todas as configurações que satisfazem esta condição são equiprováveis, obtemos a *estatística de Fermi-Dirac*. O sistema é totalmente descrito a partir da definição de quais das n células contém uma partícula. Como há r partículas, as células correspondentes podem ser escolhidas de $\binom{n}{r}$ formas distintas, cada uma com probabilidade igual à $\binom{n}{r}^{-1}$. Estas distribuições formam uma base sob a qual podemos desenvolver experimentos mais complexos e chegar a resultados relevantes.

Consideremos agora o caso onde em uma população de n elementos, n_1 são azuis e $n_2 = n - n_1$ são vermelhos e, dentre estes, escolhemos um grupo de r elementos, aleatoriamente. O objetivo é então definir a probabilidade q_k de que o grupo de r elementos contenha exatamente

k elementos azuis, onde k é um número inteiro maior que zero e menor ou igual à n_1. Para encontrar a probabilidade q_k, devemos calcular a quantidade de formas de escolher um grupo contendo k elementos azuis e $r-k$ elementos vermelhos. Os elementos azuis podem ser escolhidos de $\binom{n_1}{k}$ formas distintas e os elementos vermelhos de $\binom{n-n_1}{r-k}$ formas distintas. Por definição, a escolha de uma determinada quantidade de elementos vermelhos implica na escolha da quantidade de elementos azuis, e obtemos

$$q_k = \frac{\binom{n_1}{k}\binom{n-n_1}{r-k}}{\binom{n}{n_1}}. \tag{1.83}$$

Tal distribuição é conhecida como *distribuição Hipergeométrica*, e pode ser generalizada para o caso onde a população original de tamanho n contém diversas classes de elementos. Consideremos o caso onde a população possui quatro classes de elementos, de tamanhos n_1, n_2, n_3 e $n-n_1-n_2-n_3$, respectivamente. Selecionando aleatroriamente uma amostra de tamanho r, a probabilidade de que a amostra contenha exatamente k_1 elementos da primeira, k_2 elementos da segunda, k_3 elementos da terceira e $r-k_1-k_2-k_3$ elementos da quarta classe é dada por

$$\frac{\binom{n_1}{k_1}\binom{n_2}{k_2}\binom{n_3}{k_3}\binom{n-n_1-n_2-n_3}{r-k_1-k_2-k_3}}{\binom{n}{r}}. \tag{1.84}$$

Tratamos agora do que pode ser considerado o átomo, ou componente fundamental de boa parte do que trataremos daqui até o final deste trabalho. Com base em Feller (1968), abordaremos a definição, os principais elementos e exemplos de aplicação deste componente.

Ensaios independentes são denominados *ensaios de Bernoulli* se há apenas dois resultados possíveis e a probabilidade de ambos os resultados permanece inalterada durante todas as etapas. Usualmente denotamos as duas probabilidades p e q por sucesso e fracasso e $p+q=1$.

O espaço amostral é formado pelos dois pontos amostrais, sucesso S e fracasso F e, consequentemente, o espaço amostral de n ensaios de Bernoulli, contém 2^n pontos amostrais, cada um representando um resultado possível dos ensaios do experimentos. Como os ensaios são independentes, as probabilidades são multiplicadas.

Uma variável aleatória X, que assume valores 1, no caso de sucesso e 0 no caso de fracasso, onde S e F são definidos a partir de ensaios de Bernoulli possui *distribuição de Bernoulli*, cuja função de densidade de probabilidade é dada por

$$P(X=x) = p^x(1-p)^{1-x}, \text{ para x} = 0,1 \; e \; p(X=x) = 0 \text{ para outros valores de x.} \tag{1.85}$$

A esperança de X é dada por $E(X)=p$ e a variância é dada por $Var(X)=p(1-p)$.

A partir da repetição de ensaios idênticos e independentes de Bernoulli, obtemos um processo de Bernoulli, o qual da origem à *distribuição binomial*. O modelo probabilístico binomial consiste de n ensaios independentes, onde cada ensaio possui apenas dois resultados possíveis. A variável aleatória, denotada por X, é então definida como a quantidade k de sucessos obtidos nos n ensaios. Ademais, quando $X_1, ...X_n$ são variáveis aleatórias independentes com distribuição de Bernoulli de parâmetro n, então $X = X_1, ...X_n$ é uma variável aleatória que possui distribuição binomial de parâmetros p e n, com funcão de densidade de probabilidade dada por

$$p(X = x) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad onde \qquad \binom{n}{k} = \frac{n!}{(n-k)!k!}. \tag{1.86}$$

O valor esperado de X é dado por $E(X) = np$ e a variância é dada por $Var(X) = np(1-p)$

Em algumas aplicações, lidamos com ensaios de Bernoulli onde n é grande e p é pequeno, com produto

$$\lambda = np. \tag{1.87}$$

Nestes casos, é conveniente utilizarmos uma aproximação à distribuição binomial $b(k;n,p)$, onde avaliamos a distribuição para $k = 0$ e obtemos

$$b(k;n,p) = (1-p)^n = \left(1 - \frac{\lambda}{n}\right)^n. \tag{1.88}$$

Passando logaritmos e utilizando uma expansão de Taylor, obtemos

$$log b(0;n,p) = n\, log\left(1 - \frac{\lambda}{n}\right) = -\lambda - \frac{\lambda^2}{2n} - \cdots \tag{1.89}$$

de forma que para um n suficientemente grande

$$b(0;n,p) \approx e^{-\lambda}. \tag{1.90}$$

Da definição da distribuição binomial, vemos que

$$\frac{b(k;n,p)}{b(k-1;n,p)} = \frac{(n-k+1)p}{kq} = 1 + \frac{(n+1)p-k}{kq} \tag{1.91}$$

de forma que o termo $b(k;n,p)$ é maior que o termo anterior $b(k-1;n,p)$ caso $k<(n+1)p$, e menor, caso $k>(n+1)p$. Se $(n+1)p=m$ for um número inteiro, de forma a obtermos $b(m;n,p)=b(m-1;n,p)$, podemos afirmar que existe um número inteiro m tal que:

$$(n+1)p-1<m\leq(n+1)p \tag{1.92}$$

e percebemos que, à medida que k vai de 0 n, os temos $b(k;n,p)$ inicialmente aumentam monotonicamente, para posteriormente decrescerem monotonicamente, alcançando seu valor mais elevado no ponto $k=m$. O termo $b(m;n,p)$ é denominado *termo central* e Feller (1968) destaca que tal termo representa a "quantidade mais provável de sucessos", lembrando que para grandes valores de n todos os termos $b(k;n,p)$ são pequenos.

De posse da definição do *termo central*, temos que para um valor fixo de k e para n suficientemente grande, obtemos

$$\frac{b(k;n,p)}{b(k-1;n,p)}=\frac{\lambda-(k-1)p}{kq}\approx\frac{\lambda}{k} \tag{1.93}$$

e concluímos que

$$b(k;n,p)\approx\frac{\lambda^k}{k!}e^{-\lambda}. \tag{1.94}$$

Tal expressão representa a *aproximação de Poisson para a distribuição binomial*.

A equação pode ser então reescrita como

$$p(k;\lambda)=e^{-\lambda}\frac{\lambda^k}{k!}, \tag{1.95}$$

representando uma *distribuição de Poisson*, a qual é uma aproximação da distribuição binominal $b(k;n,\lambda\ n)$. Se uma variável aleatória X possui distribuição de Poisson com parâmetro $\lambda>0$, então a esperança de X, $E(X)$ é igual a variância $Var(X)$ e obtemos

$$E(X)=Var(X)=\lambda \tag{1.96}$$

Por fim, outro tipo de aproximação à distribuição Binomial é a *distribuição normal*, a qual possui função de densidade de probabilidade

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2} \tag{1.97}$$

e a distribuição acumulada

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2} dy. \tag{1.98}$$

2 Aleatoriedade e Processos Estocásticos

2.1 Passeio Aleatório

Consideremos uma partícula a qual se encontra inicialmente na origem do eixo X_0 . No momento $n = 1$ tal partícula da um salto de magnitude Z_1, onde Z_1 é uma variável aleatória que possui uma distribuição de probabilidade conhecida. No momento $n = 2$, a partícula dá um salto de magnitude Z_2 , onde Z_2 é independente de Z_1 e possui a mesma distribuição. Então, a partícula se move sobre o eixo, de forma que após o primeiro salto ela se encontra no ponto $X_0 + Z_1$, após o segundo salto, no ponto $X_0 + Z_1 + Z_2$ e naturalmente, após n saltos a posição da partícula é

$$X_n = X_0 + Z_1 + Z_2 + ... + Z_n, \tag{2.1}$$

onde as variáveis Z_i são variáveis aleatórias mutuamente independentes e identicamente distribuídas. Cox e Miller (1968) definem que neste caso, o movimento de tal partícula é representado por um passeio aleatório unidimensional, onde

$$X_n = X_{n-1} + Z_n, \qquad n = 1, 2, \cdots . \tag{2.2}$$

Quando os saltos Z_i assumem somente os valores 0, 1 ou -1, com distribuição

$$prob(Z_i = 1) = p \quad , \quad prob(Z_i = 0) = 1 - p - q \quad , \quad prob(Z_i = -1) = q \tag{2.3}$$

obtemos um processo denominado *passeio aleatório simples*. De acordo com a definição utilizada por, Cox e Miller (1968), o passeio aleatório é um processo aleatório em tempo discreto. O espaço estado será contínuo se os saltos $\{Z_i\}$ forem variáveis aleatórias contínuas e discreto se os saltos forem restritos a valores inteiros, como no passeio aleatório simples.

Nos referimos à posição da partícula para definir seu *estado*. Quando a partícula pode se movimentar indefinidamente, obtemos um passeio aleatório *irrestrito*. Quando a particula $X_0 = 0$ está restrita a se movimentar no intervalo a acima e b abaixo da origem, de forma que o movimento é interrompido quando a partícula chega a algum destes pontos, denominamos a e b *barreiras de absorção*, e obtemos um passeio aleatório restrito.

Segundo Cox e Miller (1968, p. 25), o passeio aleatório é um tipo particular de *processo de Markov*, e um passeio aleatório irrestrito é uma soma de variáveis aleatórias independentes,

o que torna relevante o papel da teoria da probabilidade e mais especificamente os aspectos relacionados à soma de variáveis aleatórias independentes. Passamos agora a um detalhamento adicional do passeio aleatório simples, onde os saltos $Z_1, Z_2, \cdots,$ são independentes e as probabilidades são dadas por 2.3.

Considerando o caso *irrestrito*, consideremos que o passeio aleatório se inicia na origem e que a partícula é livre para se movimentar indefinidamente em qualquer uma das direções. Temos então que após n saltos (ou períodos), a posição da partícula é dada por

$$X_n = \sum_{t=1}^{n} Z_r \tag{2.4}$$

Por conta da definição da magnitude dos saltos e da possibilidade de movimentação livre em qualquer direção, temos que a qualquer momento n o conjunto de todas as posições possíveis, o qual podemos considerar como o espaço amostral é dado pelos pontos amostrais

$$k = 0 \pm 1, \pm 2, \cdots, \pm n \tag{2.5}$$

De forma que se formos avaliar o caminho percorrido para que a partícula chegasse a uma determinada posição k, devemos considerar que ela tenha passado por r_1 saltos positivos, r_2 saltos negativos e r_s saltos de magnitude zero. É importante notar que r_1, r_2 e r_s são quaisquer números inteiros não negativos, os quais devem satisfazer as equações

$$r_1 - r_2 = k \tag{2.6}$$

e

$$r_s = n - r_1 - r_2. \tag{2.7}$$

Se só fosse possível obter saltos de magnitude 1 ou 0, poderíamos enquadrar tal processo a partir de um processo de Bernoulli, e consequentemente, o passeio aleatório possuiria uma distribuição binomial, onde a posição k, seria a quantidade de sucessos, em n repetições com probabilidade p.

Para o nosso caso, com três magnitudes possíveis para cada salto, a probabilidade de $X_n = k$ é dada por

$$p(X_n = k) = \sum \frac{n!}{r_1! r_2! r_s!} p^{r_1} (1 - p - q)^{r_3} q^{r_2} \tag{2.8}$$

onde temos a soma de probabilidades multinomiais. A função ou geradora de momentos do salto Z_r é

$$G(z) = E(z^{Z_r}) = pz + (1-p-q) + qz^{-1} \tag{2.9}$$

e a de X_n é dada por

$$E(z^{X_n}) = \{G(z)\}^n. \tag{2.10}$$

Como $X_0 = 0$, definimos $G_0(z) = 1$ e introduzimos a função geradora

$$G(z,s) = \sum_{n=0}^{\infty} s_n \{G(z)\}^n = \frac{1}{1-sG(s)}(|sG(z)| < 1) = \frac{z}{-spz^2 + z\{1-s(1-p-q)\} - sq}, \tag{2.11}$$

onde $G(z,s)$ contém toda a informação sobre o passeio aleatório.

Denotando a média por μ e a variância por σ^2, temos que:

$$\mu = p - q \tag{2.12}$$

e

$$\sigma^2 = p + q - (p-q)^2. \tag{2.13}$$

Consequentemente, definimos os valores esperados para a média e variância por

$$E(X_n) = n\mu \quad e \quad Var(X_n) = n\sigma^2. \tag{2.14}$$

Apesar das definições acima serem necessárias para o entendimento de um processo estocástico, na maior parte dos problemas envolvendo economia e finanças não tratamos de pontos, e sim de intervalos. Tal situação pode ser encontrada por um analista que busca prever a probabilidade do preço de uma determinada ação estar em dado um intervalo em um período futuro específico, e também por um economista, que tenta encontrar um intervalo de referência para o valor da taxa de câmbio entre o real e o dólar americano para um determinado período. Tendo problemas desta natureza como referência, torna-se necessário avaliar a probabilidade de que no momento n a partícula se encontre em um dos estados

$$j, j+1, \cdots, k, \quad onde \quad X_n(j < k). \tag{2.15}$$

Com base em Cox e Miller (1968), para podermos identificar esta probabilidade, utilizamos a aproximação fornecida pelo teorema central do limite, onde X_n é uma variável aleatória que possui aproximadamente uma distribuição normal

$$X \sim N(n\mu, n\sigma^2) \tag{2.16}$$

com média $n\mu$, variância $n\sigma^2$ e para grandes valores de n

$$p(j \leq X_n \leq k) \simeq (2\pi\sigma^2 n)^{-\frac{1}{2}} \int_j^k exp\Big\{ - \frac{(x-n\mu)^2}{2n\sigma^2} \Big\} dx. \tag{2.17}$$

Para uma avaliação em tempo contínuo, utilizamos $j-c$ e $k+c$ como limites de integração, onde $c = \frac{1}{2}$ ou $c = 1$ a depender de $p+q<1$ ou $p+q=1$.

Segue então que

$$p(j \leq X_n \leq k) \simeq \Phi\Big(\frac{k+c-n\mu}{\sigma\sqrt{n}}\Big) - \Phi\Big(\frac{j-c-n\mu}{\sigma\sqrt{n}}\Big), \tag{2.18}$$

onde

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{1}{2}x^2} dx \tag{2.19}$$

é a distribuição normal padrão.

Ainda é possível que $p > q$ ou $p < q$. Para o caso onde a probabilidade p de um salto *positivo* é maior do que a probabilidade q de um salto negativo, para encontrarmos o ponto onde esperamos encontrar a partícula após n saltos precisamos de algumas definições adicionais. Sabemos que como $p > q$, a média μ é positiva e, de acordo com o teorema central do limite, a variável aleatória X_n se encontrará em um intervalo de até 3 desvios-padrão do valor esperado com probabilidade

$$p(n\mu - 3\sigma\sqrt{n} < X_n < n\mu + 3\sigma\sqrt{n}) \simeq 1 \tag{2.20}$$

e

$$X_n = n\mu + O(\sqrt{n}) = n\mu 1 + O(n^{-\frac{1}{2}}). \tag{2.21}$$

Como $\mu > 0$, a partir da lei dos grandes números obtemos que

$$p(X_n > j) \to 1 \quad a\ medida\ que \quad n \to \infty, \quad \forall j. \tag{2.22}$$

Para o caso onde $p = q$, obtemos que $E(X) = \mu = 0$ e pelo teorema central do limite encontramos que a partícula estará a uma distância de ordem $\sqrt{n}$ após n saltos, com alta probabilidade. Também é importante considerar movimentos de partículas em mais de um plano, o que da origem ao passeio aleatório multidimensional.

Consideremos uma partícula que se encontra inicialmente na origem e passa por saltos $Z_1, Z_2, \cdots, Z_i$, onde os saltos Z_i são vetores bidimensionais independentes. Após n saltos a posição da partícula será dada pelo vetor

$$X_n = Z_1 + \cdots + Z_n. \tag{2.23}$$

Generalizando para m dimensões e considerando cada salto individual como vetores m-dimensionais independentes e identicamente distribuídos com segundo momento finito, a posição esperada do vetor m-dimensional X_n após n saltos será dada por $X_n = Z_1 + \cdots + Z_n$ Adicionalmente, pelo teorema do limite central X_n possui distribuição normal (assintoticamente) com média $n\mu$ e matriz de dispersão $n\Sigma$.

2.2 Cadeias de Markov

Até então, lidamos com experimentos compostos por ensaios independentes, que podem ser descritos de forma geral por um conjunto de resultados possíveis $E_1, E_2, \cdots$, onde cada um possui uma probabilidade p_k associada. A probabilidade de uma dada sequência é definida por

$$P(E_{j_0}) = p_{j_0} \cdot p_{j_1} \cdot p_{j_3} \cdots p_{j_n}.$$

Já na teoria das *cadeias de Markov*, consideramos processos onde o resultado de um ensaio n depende apenas do resultado do ensaio imediatamente anterior $n-1$, não importando o tempo m . Os Processos que respeitam esta propriedade são denominados *homogêneos* . Portanto, nestes casos o resultado E_k não é mais associado à uma probabilidade fixa.

Com base em Cox e Miller (1968) , obtemos que para cada par (E_j, E_k) é atribuída uma *probabilidade condicional* dada por p_{jk}. Ou seja, dado que E_j ocorreu em um ensaio, a probabilidade de ocorrência de E_k no ensaio subsequente é p_{jk} . Além disso, definimos a pro-

babilidade de ocorrência do resultado E_k no primeiro ensaio por a_k e obtemos a probabilidade de sequências amostrais

$$P\{(E_{j_0}, E_{j_1}, \cdots, E_{j_n})\} = a_j \cdot p_{j_0 j_1} \cdot p_{j_1 j_2} \cdot p_{j_2 j_3} \cdots p_{j_{n-1} j_n}. \tag{2.24}$$

Nos referimos aos resultados $E(k)$ como os possíveis *estados do sistema* e à probabilidade p_{jk} como a *probabilidade de transição* de E_j para E_k, onde os ensaios são performados a uma taxa constante e os saltos podem ser utilizados como parâmetro temporal.

Arranjamos as probabilidades de transição p_{jk} em forma de uma matriz de probabilidades de transição:

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1n} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & p_{n3} & \cdots & p_{nn} \end{bmatrix} \tag{2.25}$$

onde P é uma matriz quadrada com elementos não-negativos e cujas linhas somam 1. Denotamos tal matriz como uma *matriz estocástica*, a qual, associada com a distribuição inicial $\{a_k\}$, define completamente uma cadeia de Markov com estados $(E_1, E_2, \cdots)$. Adicionalmente, uma cadeia de Markov é *finita* quando seu espaço de estado consiste de uma quantidade finita de pontos. De posse destes novos conceitos, podemos avaliar o caso de um *passeio aleatório com probabilidades variáveis*.

Até então, havíamos considerado somente passeios aleatórios os quais possuíam a propriedade de homogeneidade espacial, a qual implica que para um salto de dada magnitude e direção, sua probabilidade independe da posição a partir da qual o salto é dado. Consideremos agora um passeio aleatório simples, onde cada salto pode assumir os valores $(-1,\ 0,\ 1)$, com probabilidades

$$\phi_j,\ \psi_j,\ \theta_j, \quad \textit{onde} \quad (\phi_j + \psi_j + \theta_j = 1,\ j = 1, 2, \cdots), \tag{2.26}$$

onde j denota a posição a partir da qual cada salto é dado.

Considerando os elementos de transição de probabilidade

$$\begin{aligned} p_{j,j+1} = \theta_j, \quad p_{j,j} = \psi_j, \quad p_{j,j-1} = \phi_j \quad (j = 1, 2, ...), \\ p_{00} = \psi_0, \quad p_{01} = \theta_0 (\psi_0 + \theta_0 = 1), \\ p_{jk} = 0, \end{aligned}$$

obtemos a matriz de transição

$$P = \begin{bmatrix} \psi_0 & \theta_0 & 0 & 0 & \dots \\ \phi_1 & \psi_1 & \theta_1 & 0 & \dots \\ 0 & \phi_2 & \psi_2 & \theta_2 & \dots \\ 0 & 0 & \phi_3 & \psi_3 & \dots \\ \vdots & \vdots & \vdots & \vdots & \end{bmatrix} \tag{2.27}$$

onde cada um dos componentes a_{ij} representa a probabilidade de transição da linha i para a coluna j.

Voltando ao caso geral, consideremos uma cadeia de Markov homogênea com matriz de transição P, com probabilidade de ocupação para o estado inicial dada pelo vetor linha

$$p^(0) = (p_0^{(0)}, p_1^{(0)}, p_3^{(0)}, \cdots) \tag{2.28}$$

e probabilidade de ocupação para o estado no período n dada por

$$p^(n) = (p_0^{(n)}, p_1^{(n)}, p_3^{(n)}, \cdots), \tag{2.29}$$

onde os componentes de $p^{(n)}$ $(n = 0, 1, 2, \cdots)$ são definidos por

$$p_j^{(n)} = p(X_n = j) \qquad (n = 0, 1, 2, \cdots ;\ j = 0, 1, 2, \cdots) \tag{2.30}$$

e obtemos

$$p_k^{(n)} = \sum_{j=0}^{\infty} p_j^{n-1} p_{jk} \qquad (n = 1, 2, 3, \cdots). \tag{2.31}$$

Assumindo que a série acima é convergente, que $\{p_j^(n-1);\ j = 0, 1, \cdots\}$ é a distribuição de probabilidade para cada um dos estados possíveis e que p_{jk} é não negativa e limitada por 1 para todo jek, introduzimos a notação matricial

$$\mathbf{p^{(n)}} = \mathbf{p^{(n-1)P}}, \tag{2.32}$$

onde, por interação, obtemos que o vetor inicial de probabilidades $\mathbf{p^{(0)}}$ e o conjunto de probabilidades de transição p_{jk} são suficientes para determinar as distribuições marginais de $\mathbf{p^{(n)}}$.

Isto significa que para um determinado inicial estado inicial j, o vetor $\mathbf{p^{(0)}}$ possui apenas componentes iguais a zero em todas as posições, a não ser na j-ésima:

$$\mathbf{p^{(0)}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ \phi_j \\ \vdots \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix} \tag{2.33}$$

o que significa que para cada k, a probabilidade de transição $p_{jk}^{(n)}$ é dada pelo elemento na posição (j,k) elevado à n-ésima potência de $\mathbf{P}$

$$(p_{jk}^{(n)}) = \mathbf{P^n}. \tag{2.34}$$

Utilizando do fato de que

$$\mathbf{P^{m+n}} = \mathbf{P^m P^n}, \tag{2.35}$$

obtemos

$$p_{jk}^{(m+n)} = \sum_l p_{jl}^{(m)} p_{lk}^{(n)} \tag{2.36}$$

a qual é conhecida como a *relação de Chapman-Kolmogorov* para cadeias de Markov homogêneas (COX; MILLER, 1968, p. 91).

Para o caso não homogêneo, temos que

$$\{p_{jk}(r,s)\} = \mathbf{P(r)P(r+1)}\cdots\mathbf{P(s-1)}, \quad (\mathbf{s} > \mathbf{r}) \tag{2.37}$$

e a equação de *Chapman-Kolmogorov* assume a forma:

$$p_{jk}(r,t) = \sum_l p_{jl}(r,s)p_{lk}(n) \qquad (r < s < t). \tag{2.38}$$

Outro aspecto importante diz respeito aos diferentes estados que podem ser assumidos por uma cadeia de Markov. Cox e Miller (1968) definem seis tipos de estado para uma cadeia de Markov, de acordo com o comportamento limitante de cada um:

Tabela 1 – Classificação de estados em uma cadeia de Markov

Tipo de estado	Definição do Estado
Periódico	Retorna aos estados possíveis somente nos períodos t, 2t, 3t, ..., onde t > 1
Aperiódico	Não periódico
Recorrente	Certeza do retorno a um determinado estado
Transitório	Incerteza do restorno a um determinado estado
Recorrente - Positivo	Recorrente, com média do tempo médio retorno finita
Recorrente - Nulo	Recorrente, com média do tempo médio retorno infinita
Ergótico	Aperiódico, com recorrência positiva

2.3 Exemplos e Aplicações

Campbell et al. (1997), fornecem um bom exemplo da aplicação de passeios aleatórios e cadeias de Markov em finanças. Os definem três tipos de passeio aleatório, denominados Passeio Aleatório 1 (PA1), Passeio Aleatório 2 (PA2) e Passeio Aleatório 3 (PA3).

A notação utilizada considera os diversos tipos de dependência que podem existir entre os retornos de um ativo r_t e r_{t+k}, em dois períodos t e $t+k$. Definimos f_{r_t} e $g_{(r_t+k)}$ como variáveis aleatórias, onde $f(\cdot)$ e $g(\cdot)$ são duas funções arbitrárias e

$$Cov[f(r_t), g(r_{t+k})] = 0, \qquad \forall t, e \forall k \neq 0.. \tag{2.39}$$

Trazendo alguns destes problemas para o mercado financeiro, podemos passar a uma investigação de como alguns processos estocásticos são abordados, como se manifestam e as consequências geradas. Considerando que um dos maiores problemas em finanças passa pela previsão do retorno de um determinado ativo, consideramos a noção de um *jogo justo*, dada por Girolamo Cardano (CAMPBELL et al., 1997 apud CARDANO, 1961), onde define-se as bases para o que definimos como um *martingale*.

Com base em Campbell et al. (1997), um *martingale* é um processo estocástico $\{P_t\}$ o qual satisfaz

$$E[P_{t+1} \mid P_t, P_{t-1}, \cdots] = P_t, \quad e \quad E[P_{t-1} - P_t \mid P_t, P_{t-1}, \cdots] = 0 \tag{2.40}$$

onde, caso P_t represente a riqueza acumulada na data t, um jogo justo seria aquele onde a esperança da riqueza para o próximo período P_{t+1} fosse igual à riqueza do período P_t.

Quando P_t é o preço de um ativo na data t, a *hipótese de martingale* implica que o valor esperado do preço do ativo de *amanhã* é igual ao preço de *hoje*. Segundo Campbell et al. (1997), sob uma perspectiva de *previsão*, a hipótese de martingale implica que a melhor previsão, sob o critério de menor erro quadrado médio, para o preço futuro do ativo é o preço atual.

Nos voltando ao primeiro tipo de passeio aleatório, PA1, consideramos a versão mais simples da hipótese de passeio aleatório, onde os incrementos são independentes e identicamente distribuídos

$$P_t = \mu + P_{t-1} + \varepsilon_t, \quad \varepsilon_t \approx IID(0, \sigma^2) \tag{2.41}$$

onde μ é a mudança de preços esperada, ou *drift*.

A independência implica não somente que os incrementos são não correlacionados, mas também que quaisquer funções não lineares dos incrementos também serão não correlacionadas.

Considerando a média e variância condicionais em uma data t, condicionadas a um valor inicial P_0 na data 0, obtemos que

$$E[P_t \mid P_0] = P_0 + \mu t, \quad e \quad Var[P_t \mid P_0] = a^2 t \tag{2.42}$$

e fica claro que o PA1 é um processo não estacionário, onde tanto a média e variância condicionais são ambas lineares no tempo.

Assumindo que $\varepsilon_t \sim N(0, \sigma^2)$, obtemos um *movimento Browniano aritmético*. Para evitar alguns dos problemas decorrentes da suposição de normalidade para os incrementos, utilizamos a definição acima em forma logarítmica, onde $p_t \equiv logP_t$ e:

$$p_t = \mu + p_{t-1} + e_t. \tag{2.43}$$

Tal definição implica retornos contínuos compostos e nos da o modelo lognormal de Bachelier (CAMPBELL et al., 1997 apud BACHELIER, 1900) .

Campbell et al. (1997, p. 33) destacam que apesar da *elegância e simplicidade* do PA1, a hipótese de incrementos independentes e identicamente distribuídos não é plausível e não se encaixa nos dados de preços ao longo de períodos extensos. Os autores destacam que ao longo dos útimos 200 anos de história da Bolsa de Ações de Nova Iorque, houve diversas mudanças no ambiente tecnológico, institucional, regulatório, econômico e social, no qual os preços são determinados. Portanto, assumir que a distribuição dos retornos se manteve constante durante todo o período é imprudente.

Para solucionar esta fragilidade, no PA2 os autores relaxam algumas das hipóteses do PA1 e permitem que os incrementos sejam independentes, mas não necessariamente identicamente distribuídos. Percebemos que PA2 contém PA1 como um caso especial, mas também contém outros processos mais *gerais* e que sob esta nova hipótese permitimos heterocedasticidade incondicional nos incrementos ε_t , a qual é particularmente útil quando há variação da volatilidade no tempo.

Há ainda uma versão mais geral da hipótese de passeio aleatório, onde relaxamos a hipótese de independência de PA2 e incluímos processos com incrementos dependentes porém não correlacionados, chegando ao processo PA3. Os autores definem esta como a *forma fraca* da hipótese de passeio aleatório, onde PA3 contém PA1 e PA2 como casos especiais.

Um exemplo de caso onde um processo satisfaz PA3 mas não PA1 e PA2 é dado quando $Cov[\varepsilon_t, \varepsilon_{t-k}] = 0, \quad \forall k \neq 0$, mas $Cov[\varepsilon_t^2, \varepsilon_{t-k}^2] \neq 0, \quad paraalgumk \neq 0$.

Neste caso, obtemos incrementos não correlacionados, mas não independentes. Campbell et al. (1997) afirmam que há diversos testes cujo objetivo é verificar ou validar a existência de processos de passeio aleatório; trataremos de alguns deles a seguir.

Consideremos a versão logarítimica do processo 1, a qual denominamos *movimento Browniano geométrico* e assumimos que o logarítimo dos preços p_t é um processo composto por variáveis aleatórias independentes e identicamente distribuídas, e segue um passeio aleatório sem *drift*

$$p_t = p_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim IID(0, \sigma^2) \tag{2.44}$$

e I_t é uma variável aleatória denotada por

$$I_t = \begin{cases} 1 & \text{if } \; r_t \equiv p_t - p_{t-1} \; > \; 0; \\ 0 & \text{if } \; r_t \equiv p_t - p_{t-1} \; \leq \; 0. \end{cases} \tag{2.45}$$

a qual indica se o retorno r_t é positivo ou negativo.

Um dos primeiros testes para processos PS1 foi proposto por Cowles e Jones (CAMPBELL et al., 1997 apud 3RD; JONES, 1937) , consistindo na comparação da frequência de *sequências* e *reversões* em series históricas de retornos de ações. As sequências são definidas como pares de retornos consecutivos possuindo o mesmo sinal, e reversões são definidas como pares de retornos consecutivos os quais possuem sinais opostos.

Desta forma, dada uma amostra de $n+l$ retornos $r_1, \cdots, r_{n+l}$, o número de sequências

N_s e reversões N_r pode ser expresso como função das variáveis aleatórias

$$N_s \equiv \sum_{t=1}^{n} Y_t \qquad e \qquad N_r \equiv n - N_s. \tag{2.46}$$

Se o logaritmo dos preços segue um passeio aleatório sem drift e se a distribuição dos incrementos ε_t é simétrica, então a probabilidade do retorno r_t ser positivo ou negativo deve ser igualmente provável. Tal resultado pode ser comparado com o lançamento de uma moeda, e de forma mais geral, com um ensaio de Bernoulli. Definimos a *razão de Cowles-Jones* como

$$\widehat{CJ} \equiv \frac{N_s}{N_r} = \frac{\frac{N_s}{n}}{\frac{N_r}{n}} = \frac{\hat{\pi}_s}{1-\hat{\pi}_s} \approx CJ = \frac{\frac{1}{2}}{\frac{1}{2}} = 1, \tag{2.47}$$

onde π_s denota a probabilidade de uma sequência e $(1-\pi_s)$ denota a probabilidade de uma reversão.

De acordo com Campbell et al. (1997), a razão deve ser interpretada como um estimador consistente e deve ser aproximadamente igual à 1, para o caso de um processo PS1. Um resultado maior que 1 sugere que há alguma estrutura na série testada, mais especificamente, que há *drift* positivo ou negativo. A presença de drift torna sequências mais prováveis que reversões, o que gera um coeficiente $CJ > 1$. Para verificar tal resultado, consideremos que o logarítimo dos preços segue um passeio aleatório com drift, onde

$$p_t = \mu + p_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2) \tag{2.48}$$

e obtemos um processo equivalente ao lançamento de uma moeda viesada na direção do emph-drift:

$$I_t = \begin{cases} 1 & \text{com probabilidade } \pi \\ 0 & \text{com probabilidade } (1-\pi)\,. \end{cases} \tag{2.49}$$

Se o drift é positivo, então a probabilidade de um retorno positivo é $\pi > \frac{1}{2}$ e caso o drift μ seja negativo, obtemos $\pi < \frac{1}{2}$. De forma mais geral, obtemos que o coeficiente CJ para este caso é dado por

$$CJ = \frac{\pi^2 + (1-\pi)^2}{2\pi(1-\pi)} \geq 1. \tag{2.50}$$

Um observador atento pode constatar que o número de sequências N_s é uma variável aleatória Binomial, composta pela soma de n variáveis aleatórias de Bernoulli Y_t onde

$$Y_t = \begin{cases} 1 & \text{com probabilidade } \pi_s = \pi^2 + (1-\pi)^2 \\ 0 & \text{com probabilidade } (1-\pi). . \end{cases} \tag{2.51}$$

Desta forma, podemos aproximar a distribuição de N_s para valores elevados de n por uma distribuição normal com média $E[N_s] = nn$ e variância

$$\begin{aligned} Var[N_s] &= n\pi_s(1-\pi_s) + 2nCov[Y_t, Y_{t+1}] \\ &= n\pi_s(1-\pi_s) + 2(\pi^3 + (1-\pi)^3 - \pi_s^2) \end{aligned}$$

Aplicando uma aproximação de Taylor de primeira ordem à

$$\widehat{CJ} = \frac{N_s}{(n-N_s)}, \tag{2.52}$$

utilizando a aproximação assintótica para a distribuição de N_s , obtemos

$$\widehat{CJ} \sim N\Big(\frac{\pi_s}{1-\pi_s}, \frac{\pi_s(1-\pi_s) + 2(\pi^3 + (1-\pi)^3 - \pi_s^2)}{n(1-\pi_s)^4}\Big). \tag{2.53}$$

Para uma investigação mais profunda dos processos PA e do teste Coles-Jones, é importante avaliarmos o efeito causado por um distanciamento de um passeio aleatório.

Consideremos I_t uma cadeia de Markov com dois estados:

$$\begin{Vmatrix} (1-\alpha) & \alpha \\ \beta & (1-\beta) \end{Vmatrix} \tag{2.54}$$

onde α denota a probabilidade de que r_{t+1} seja negativo, condicional a um r_t positivo, e β denota a probabilidade de que r_{t+1} seja positivo, dado um r_t negativo. Conquanto $\alpha \neq 1-\beta$, I_t e os retornos r_t serão serialmente correlacionados, violando as condições de caracterização de um PA1. Neste caso, o valor de CJ é dado por

$$CJ = \frac{(1-\alpha)\beta + (1-\beta)\alpha}{2\alpha\beta}, \tag{2.55}$$

podendo assumir quaisquer valores reais não negativos. Percebemos então que a medida que α e β se aproximam de 1, a probabilidade de reversões aumenta, e a medida que α e β se aproximam de 0 probabilidade de sequências aumenta.

Um teste complementar a teste de Coles-Jones é o teste das *corridas*.Com base em Campbell et al. (1997), definimos uma corrida como a quantidade de sequências de retornos positivos ou negativos. Posteriormente, os resultados encontrados em uma série são comparados com os resultados obtidos a partir de um passeio aleatório.

Como exemplo, consideremos uma sequência de 10 retornos como 1011110100 , onde a mesma contém 3 corridas de 1s, de tamanho 1, 4 e 1, e três corridas de 0s, de tamanho 1, 1 e 2. Em contrapartida, sequência 1111000000 contém apenas 2 corridas. Consideremos que cada um das n observações sejam independentes e identicamente distribuiídas e assumam q possíveis valores com probabilidade $\pi_i, i = 1,2,\cdots,q$, onde $\sum_i \pi_i = 1$. Denotamos $N_c(i)$ como o numero total de corridas de extensão i , onde $i = 1,\cdots,q$, e N_c como a soma da quantidade de corridas de cada tipo $\sum_i N_c(i)$.

Utilizando de propriedades da distribuição multinomial, obtemos a distribuição discreta de $N_c(i)$ com base em Mood (1940), e obtemos os momentos

$$E[N_c(i)] = n\pi_i(1-\pi_i)+\pi^2 \tag{2.56}$$

$$Var[N_c(i)] = n\pi_i(1-4\pi_i+6\pi_i^2-3\pi_i^3)+\pi_i^2(3-8\pi_i^2) \tag{2.57}$$

$$Cov[N_c(i),N_c(j)] = -n\pi_i\pi_j(1-2\pi_i-2\pi_j+3\pi_i\pi_j)-\pi_i\pi_j(2\pi_i-2\pi_j-5\pi_i\pi_j). \tag{2.58}$$

Adicionamente, observamos que a distribuição do número de corridas converge assintoticamente para uma distribuição normal

$$x_i \equiv \frac{N_c(i)-n\pi_i(1-\pi_i-\pi_i^2)}{\sqrt{n}} \tag{2.59}$$

$$\sim \mathcal{N}(0,\pi_i(1-n)-3\pi_i^2(1-\pi_i)^2) \tag{2.60}$$

$$Cov[x_i,x_j] = -\pi_i\pi_j(1-2\pi_i-2\pi_j+3\pi_i\pi_j) \tag{2.61}$$

$$x = \frac{N_c-n(1-\sum_i \pi_i^2)}{\sqrt{n}} \tag{2.62}$$

$$\sim \mathcal{N}(0,\sum_{i=1}^{k}\pi_i^2(1+2\pi_i)-3(\sum_{i=1}^{k}\pi_i^2)^2). \tag{2.63}$$

As probabilidades podem ser estimadas diretamente a partir dos dados como as razões $\hat{\pi} \equiv \frac{n_i}{n}$, onde n_i representa a quantidade de corridas em uma amostra de tamanho n , de forma que $n = \sum_i n_i$ e $=^a$ significa que a igualdade é assintoticamente válida.

3 Abordagens Econométricas Modernas

3.1 Modelos ARMA, ARIMA

Consideremos a avaliação da diferença de uma função $y = f(t)$ em t_0 e $t_0 + h$. Tal diferença pode ser representada por:

$$\Delta y = f(t_0 + h) - f(t_0) \tag{3.1}$$

Retomando os conceitos apresentados no primeiro capítulo, é adequado considerar os elementos observados de uma série temporal $[y_0, y_1, y_2, \cdots, y_t]$ como realizações de um determinado processo estocástico.

Para entender um processo estocástico, é necessário introduzir um conceito importante, o *ruído branco*. Definimos uma sequência ε_t como um ruído branco se cada valor na sequência possui média zero, variância constante e se cada valor da sequência é serialmente não correlacionado. Formalmente, com base em Enders (2008), temos que

$$E(\varepsilon_t) = E(epsilon_{t-1}) = \cdots = 0 \tag{3.2}$$

$$E(\varepsilon_t^2) = E(\varepsilon_{t-1}^2) = \cdots = \sigma_2 \tag{3.3}$$

$$E(\varepsilon_t \varepsilon_{t-s}) = E(\varepsilon_{t-j} \varepsilon_{t-j-s}) = 0, \quad \forall j, s. \tag{3.4}$$

e podemos então construir séries temporais interessantes a partir de ruídos brancos.

Definindo coeficientes β_i, os qual são multiplicados por uma sequência de ruídos brancos ε_{t-i}, obtemos

$$x_t = \sum_{i=0}^{q} \beta_i \varepsilon_{t-i} \tag{3.5}$$

Denotamos tal processo como uma *média móvel de ordem q* e a denotamos por *MA(q)*. Percebemos também, que apesar de ε_t ser um ruído branco, a sequência x_t não será, caso dois ou mais dos coeficientes β_i sejam diferentes de zero.

Combinando o processo acima com uma equação de diferenças linear de ordem p

$$y_t = a_0 + \sum_{i=1}^{p} a_i y_{t-i} + x_t \tag{3.6}$$

onde x_t é o processo *MA(q)*, obtemos

$$y_t = a_0 + \sum_{i=1}^{p} a_i y_{t-i} + \sum_{i=0}^{q} \beta_i \varepsilon t - i. \tag{3.7}$$

Considerando que tal processo é estacionário, y_t é denominado um processo *auto-regressivo de média móvel*, ou simplesmente ARMA , onde o componente autoregressivo é dado pela equação de diferenças e o componente de média móvel é dado por x_t . Quando a parcela homogênea da equação de diferenças contém p defasagens e x_t contém q defasagens, denominamos o modelo por ARMA(p, q).

A solução de um modelo ARMA(p, q) expressando y_t em função da sequência ε_t é denominada a *representação de média móvel* de y_t . Para um processo AR(1), a representação de média móvel é dada por

$$y_t = \frac{a_0}{(1-a_1)} + \sum_{i=0}^{\infty} \tag{3.8}$$

e para um processo ARMA(p, q):

$$\left(1 - \sum_{i=1}^{p} a_i L^i\right) y_y = a_0 + \sum_{i=0}^{q} \beta_i \varepsilon_{t-i} \tag{3.9}$$

de forma que a solução particular para y_t é dada por:

$$y_t = \frac{a_0 + \sum_{i=0}^{q} \beta_i \varepsilon_{t-i}}{1 - \sum_{i=1}^{p} a_i L^i} \tag{3.10}$$

Segundo Enders (2008), para que a expressão acima seja válida, é necessário que as séries acima sejam convergentes, de forma que a equação de diferenças estocástica definida por 3.10 seja estável. A condição de estabilidade é que as raízes características do polinômio $(1 - \sum a_i L^i)$ devem estar fora do circulo unitário. Esta condição de estabilidade é necessária para que a série y_t seja estacionária. Formalmente, um processo estocástico que contém média μ e variância σ_y^2 finitas é *covariância estacionário* se para quaisquer t e $t-s$ se

$$E(y_t) = E(y_{t-s}) = \mu, \tag{3.11}$$

$$E[(y_t - \mu)^2] = E[(y_{t-s} - \mu)^2] = \sigma_y^2 \tag{3.12}$$

e

$$E[(y_t - \mu)(y_{t-s} - \mu)] = E[(y_{t-j} - \mu)(y_{t-j-s} - \mu)] = \gamma_s, \tag{3.13}$$

onde μ, σ_y^2, e γ_s são constantes.

Em suma, uma serie temporal é covariância estacionária quando sua média e todas as autocovariâncias não são afetadas por mudanças na origem do tempo observado. Para séries covariância estacionárias, definimos a autocorrelação entre y_t *e* y_{t-s} como

$$\rho \equiv \frac{\gamma_s}{\gamma_0} \tag{3.14}$$

e é importante notar que a autocorrelação entre y_t e y_{t-1} pode ser diferente da autocorrelação entre y_t e y_{t-2}, ao passo que autocorrelações devem ser idênticas quando a defasagem é igual, como nos casos entre y_t e y_{t-1} e y_{t-s} e y_{t-s-1}.

Para um modelo AR(1), obtemos

$$\gamma_0 = \frac{\sigma^2}{1 - (a_1^2)} \tag{3.15}$$

$$\gamma_s = \frac{\sigma^2 (a_i)^s}{1 - (a_1^2)}. \tag{3.16}$$

Formamos as autocorrelações a partir da divisão de cada um dos γ_s por γ_0, e obtemos

$$\rho_0 = 1, \rho_1 = a_1, \rho_2 = (a_1^2), \cdots, \rho_s = (a_1^s). \tag{3.17}$$

Para que um processo AR(1) seja estacionário, é necessário que $|a_1| < 1$ e a partir da plotagem dos ρ_s contra s, obtemos a *função de autocorrelação*, a qual deve convergir geometricamente para zero se a série for estacionária.

Caso o coeficiente a_1 seja positivo, a convergência será direta, e caso seja negativo, as autocorrelações seguirão uma trajetória oscilatória em torno de zero.

Expandindo para um processo ARMA(1, 1), denotado por

$$y_t = a_1 y_{t-1} + \varepsilon_t + \beta_1 \varepsilon_{t-1} \tag{3.18}$$

multiplicamos a equação de diferenças por y_{t-s} , $s = 0, 1, 2, \cdots$, e tomamos a esperança para obter

$$\begin{aligned}
Ey_t y_t = &\quad a_1 Ey_{t-1}y_t + E\varepsilon_t y_t + \beta_1 E\varepsilon_{t-1}y_t = \gamma_0 = a_1\gamma_1 + \sigma^2 + \beta_1(a_1+\beta_1)\sigma^2 \\
Ey_t y_{t-1} = &\quad a_1 Ey_{t-1}y_{t-1} + E\varepsilon_t y_{t-1} + \beta_1 E\varepsilon_{t-1}y_{t-1} = \gamma_1 = a_1\gamma_0 + \sigma^2 + \beta_1\sigma^2 \\
Ey_t y_{t-2} = &\quad a_1 Ey_{t-1}y_{t-2} + E\varepsilon_t y_{t-2} + \beta_1 E\varepsilon_{t-1}y_{t-2} = \gamma_2 = a_1\gamma_1 \\
&\quad\vdots \qquad\qquad\qquad\qquad \vdots \\
Ey_t y_{t-s} = &\quad a_1 Ey_{t-1}y_{t-s} + E\varepsilon_t y_{t-s} + \beta_1 E\varepsilon_{t-1}y_{t-s} = \gamma_s = a_s\gamma_{s-1}
\end{aligned} \tag{3.19}$$

Com base em Enders (2008), resolvendo simultaneamente para γ_0 *e* γ_1 , obtemos

$$\gamma_0 = \frac{1+\beta_1^2+2a_1\beta_1}{(1-a_1^2)}\sigma^2 \tag{3.20}$$

$$\gamma_1 = \frac{(1+a_1\beta_1)+(a_1+\beta_1)}{(1-a_1^2)}\sigma^2 \tag{3.21}$$

$$\rho_1 = \frac{(1+a_1\beta_1)(a_1+\beta_1)}{1+\beta_1^2+2a_1\beta_1} \tag{3.22}$$

$$\rho_s = a_1\rho_{s-1}, \qquad \forall s \geq 2. \tag{3.23}$$

Desta forma, chegamos à conclusão de que para um processo ARMA(1, 1), a função de autocorrelação é tal que a magnitude de ρ_1 depende tanto de a_1 quanto de β_1 .

Generalizando para um processo ARMA(p, q), os valores de ρ_i satisfazem

$$\rho_i = a_1\rho_{i-1} + a_2\rho_{i-2} + \cdots + a_p\rho_{i-p}. \tag{3.24}$$

É importante notar que para um processo AR(1), os valores em y_t e y_{t-2} estão correlacionados, mesmo considerando o fato de que y_{t-2} não aparece diretamente no modelo. Tal correlação se da por conta do efeito que y_{t-1} tem sobre y_t , e pelo efeito que y_{t-2} tem sobre y_{t-1} . Logo, observamos um efeito indireto sobre as defasagens, e tal efeito é capturado pela função de autocorrelação (ACF). No entanto, a informação do efeito direto de cada defasagem sobre y_t é relevante e, para que possamos obtê-la, introduzimos a função de *autocorrelação parcial*.

O que diferencia a função de autocorrelação parcial (PACF) da ACF, é que a primeira elimina os efeitos intermediários das defasagens entre y_t e y_{t-s} . Desta forma, em um processo

autogregressivo de ordem p, a autocorrelação entre quaisquer defasagens com intervalos superiores à y_{t-s} serão iguais a zero.

Para formarmos a função de autocorrelação parcial, subtraímos a média $y(\mu)$ de cada uma das observações y_t, obtendo

$$y_t^* \equiv y_t - \mu \tag{3.25}$$

e formamos a equação de autoregressão de primeira ordem

$$y_t^* = \phi_{11} y_{t-1}^* + e_t. \tag{3.25}$$

Expandindo, obtemos a equação de autocorrelação de segunda ordem

$$y_t^* = \phi_{21} y_{t-1}^* + \phi_{22} y_{t-2}^* + e_t, \tag{3.25}$$

onde ϕ_{22} é o coeficiente de autocorrelação parcial entre y_t e y_{t-2} e o erro definido por e_t pode não ser um ruído branco. Enders (2008) afima que o coeficiente ϕ_{22} representa a correlação entre y_t e y_{t-2}, controlando para o efeito de y_{t-1}.

Repetindo o processo acima para s defasagens adicionais, obtemos a função de autocorrelação parcial. Finalmente, podemos relacionar a ACF e PACF a partir de

$$\phi_{11} = \rho_1 \tag{3.26}$$
$$\phi_{22} = \frac{(\rho_2 - \rho_1^2)}{(1-p_1^2)} \tag{3.27}$$
$$\vdots \quad \vdots \tag{3.28}$$
$$\phi_{ss} = \frac{\rho_s - \sum_{j=1}^{s-1} \phi_{s-1,j}\rho_{s-j}}{1 - \sum_{j=1}^{s-1} \phi_{s-1,j} p_j}, \tag{3.29}$$
$$\tag{3.30}$$

onde

$$\phi_{sj} = \phi_{s-1,j} - \phi_{ss}\phi_{s-1,s-j}, \;\; j = 1, 2, \cdots, s-1. \tag{3.29}$$

Definidos os processos AR, MA e ARMA, e as ACF e PACF, passamos para a breve definição de um dos aspectos mais importantes envolvendo o trabalho de um econometrista, a previsão.

Atualizando um processo AR(1) por um período, obtemos

$$y_{t+1} = a_0 + a_1 y_t + \varepsilon_{t+1} \tag{3.29}$$

e se conhecemos os coeficientes a_0 e a_1, podemos realizar uma previsão do valor assumido por y_{t+j} no período t a partir de

$$E_t y_{t+j} = E(y_{t+j}|y_t, y_{t+1}, y_{t+2}, \cdots, \varepsilon_t, \varepsilon_{t-1}, \cdots) \tag{3.30}$$
$$E_t y_{t+j} = a_0(1 + a_1 + a_1^2 + \cdots + a_1^{j-1}) + a_1^j y_t \tag{3.31}$$

a qual denotamos *função de previsão*.

Apesar de úteis, de acordo com Enders (2008), as previsões derivadas de um processo ARMA não serão perfeitas, apesar dos erros não serem correlacionados. Realizando a previsão com base no período t, denotamos o erro da previsão para o período $t+j$, $f_t(j)$ por

$$f_t(j) \equiv y_{t+j} - E_t y_{t+j}. \tag{3.31}$$

Para a previsão de um período a frente, obtemos

$$f_t(1) = y_{t+1} - E_t y_{t+1} = \varepsilon_{t+1}, \tag{3.31}$$

onde o erro ε_t representa a parcela imprevisível de y_{t+1}, dada a informação disponível no período t.

Generalizando, para um processo AR(1), o erro da previsão de j períodos a frente é dado por

$$f_t(j) = \varepsilon_{t+j} + a_1\varepsilon_{t+j-1} + a_1^2\varepsilon_{t+j-2} + a_1^3\varepsilon_{t+j-3} + \cdots + a_1^{j-1}\varepsilon_{t+1}. \tag{3.31}$$

É importante notar que, apesar das estimativas serem não viesadas, o erro de previsão aumenta com o tempo, diminuindo a acurácia das previsões.

Dado que o valor esperado dos erros de previsão são dados por

$$E_t\varepsilon_{t+j} = E_t\varepsilon_{t+j-1} = \cdots = E_t\varepsilon_{t+1} = 0, \tag{3.31}$$

a variância do erro de previsão é dada por

$$Var[f_t(j)] = \sigma^2[1 + a_1^2 + a_1^4 + a_1^6 + \cdots + + a_1^{(2j-2)}], \tag{3.31}$$

e notamos que a variância do erro de previsão é uma função crescente da quantidade de períodos a frente j.

3.2 Modelos ARCH, GARCH

Por vezes, e de fato na maioria das vezes, series as quais capturam algum tipo de expressão do comportamento humano tendem a ser não lineares. Quando este for o caso, os conceitos e modelos apresentados na primeira seção serão pouco eficientes. Campbell et al. (1997) afimam que a modelagem de fenômenos não lineares é uma fronteira natural para a econometria, e que sua importância reside no fato de que a coleção de modelos não lineares é demasiadamente maior do que a coleção de modelos lineares. Os autores destacam ainda que a análise de modelos não lineares é uma tarefa complexa, e por vezes, a única forma de solução é a aproximação por análise computacional, o que pode ser estranho àqueles acostumados ao pensamento estritamente analítico.

Enders (2008) complementa, apresentando a alguns fatos estilizados sobre séries temporais.

1. A maior parte das séries contém uma tendência visível: O PNB real, seus subcomponentes e a oferta de instrumentos financeiros de curto prazo exibem uma tendência positiva. Séries como as de taxa de juros (i.e. Fed Funds Rate (FFR), Selic) e taxa de inflação apresentam tendências as quais variam entre positivas e negativas. Por conta destes fatores, não é possível afimar que tais séries contenham uma média invariante ao tempo. Portanto, tais séries são não estacionárias.

2. Algumas séries apresentam comportamento oscilatório: A taxa de câmbio entre a libra e o dólar não apresenta nenhuma tendência particular, positiva ou negativa. Este tipo de comportamento de passeio aleatório é típico de séries não estacionárias.

3. Quaisquer choques a uma série apresentam alto grau de persistência: A taxa FFR apresentou alto crescimento em 1973, permanecendo neste nível elevado por aproximadamente 2 anos. Da mesma forma, a produção industrial do Reino Unido caiu consideravelmente no fim da década de 70, sem retornar ao seu nível anterior até meados da década de 80.

4. A volatilidade de muitas séries não é constante no tempo: Os preços dos produtores americanos flutuaram descontroladamente durante a década de 70, quando em comparação com a flutuação suave das décadas de 60 e 80. O investimento real cresceu suavemente durante a maior parte da década de 60, porém se tornou altamente variável a partir da década de 70. Tais séries são denominadas *condicionalmente heteroscedásticas* caso a variância de longo prazo é constante, mas há períodos nos quais a variância é relativamente elevada.

5. Algumas séries apresentam comovimentos com outras séries: Grandes choques na

produção industrial americana aparentam acontecer ao mesmo tempo do que aqueles ocorridos no Reino Unido e no Canada. Taxas de juros de curto e longo prazo rastreiam uma à outra de forma próxima. A presença de tais comovimentos não deve causar surpresa, pois é esperado que as forças econômicas afetando os Estados Unidos também afetem as indústrias de outros países.

Com base em Hamilton (1994), denotamos um processo AR(1) por

$$y = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + u_t, \tag{3.31}$$

onde u_t é um ruído branco com valor esperado igual à zero e o valor esperado para a covariância entre o erro de dois períodos $E(u_t u_\tau)$ é igual à zero para $t \neq \tau$ e igual à σ^2 para $t = \tau$.

Um processo como o definido por 3.46 é covariância estacionário se as raízes do polinômio característico se encontram fora do círculo unitário. Denotando os coeficientes por ϕ_i, temos que a previsão linear ótima do valor de y_t para um processo AR(p) é dado por:

$$\hat{E}(y_t | y_{t-1}, y_{t-2}, \cdots) = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} \tag{3.31}$$

e dado que o processo é covariância estacionário, a média incondicional de y_t é constante, denotada por

$$E(y_t) = \frac{c}{1 - \phi_1 - \phi_2 - \phi_3 \cdots - \phi_p}. \tag{3.31}$$

Apesar da equação acima implicar que a variância incondicional do erro u_t é constante, igual à σ_2, a variância condicional pode mudar com o tempo.

Descrevendo o quadrado do erro u_t em função de m defasagens de si mesmo, obtemos um processo autoregressivo de ordem m, denotado por

$$u_t^2 = \zeta + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \alpha_3 u_{t-3}^2 + \cdots + \alpha_m u_{t-m}^2 + w_t, \tag{3.31}$$

onde w_t é um processo de ruído branco. Adicionalmente, denotamos o valor esperado e a autocovariância do processo AR(m) por

$$E(w_t w_\tau) = \begin{cases} \lambda^2, & \text{para } t \neq \tau \\ 0, & \text{caso contrário} \end{cases}.$$

Desta forma, o erro na previsão de y_t implica a dependência linear do erro quadrado u_t^2 sobre os m erros quadrados de previsão dos períodos anteriores e obtemos:

$$\hat{E}(u_t^2|u_{t-1}^2, u_{t-2}^2, \cdots,) = \zeta + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \alpha_3 u_{t-3}^2 + \cdots + \alpha_m u_{t-m}^2. \tag{3.31}$$

Processos de ruído branco da forma definida por (172) são descritos como *autoregressivos de heterocedasticidade condicional*, ou ARCH(m), para a sigla em inglês.

Quando todos os coeficientes α_j são não negativos, para que u_t^2 seja covariância estacionário é necessário que:

$$1 - \alpha_1 z - \alpha_2 z^2 - \cdots - \alpha_m z_m = 0 \text{e} \alpha_1 + \alpha_2 + \cdots + \alpha_m < 1. \tag{3.31}$$

Quando tais condições são satisfeitas, obtemos a variância incondicional de u_t , dada por

$$\sigma_2 = E(u_t^2) = \frac{\zeta}{1 - \alpha_1 - \alpha_2 - \cdots - \alpha_m}. \tag{3.31}$$

Há diversos métodos de estimação dos parâmetros de um modelo com perturbações ARCH. Passamos brevemente pela abordagem de máxima verossimilhança com incrementos Gaussianos, baseada em Hamilton (1994).

Consideremos a equação de regressão dada por:

$$y_t = x_t' \beta + u_t \tag{3.31}$$

onde x_t denota um vetor de variáveis explicativas. Definimos por Υ o vetor de observações obtidas até a data t :

$$\Upsilon_t = (y_t, t_{t-1}, \cdots, y_1, y_0, \cdots, y_{-m+1}, x_t', x_{t-1}', \cdots, x_1', x_0', \cdots, x_{-m+1}')' \tag{3.31}$$

contendo tanto os valores de y_i quanto os valores do vetor das variáveis explicativas, que pode conter tanto defasagens de y quanto outras variáveis relacionadas.

Caso o termo υ_t seja independente tanto de x quanto de Υ_{t-1} e seja uma sequência com média zero e variância unitária, então a distribuição condicional de y_t

$$f(y_t|x_t, \Upsilon_{t-1}) = \frac{1}{\sqrt{2\pi h_t}} exp\left(\frac{-(y_t - \mathbf{x_t'}\beta)^2}{2h_t}\right), \tag{3.31}$$

onde

$$\upsilon_t \sim N(0,1) \tag{3.31}$$

é gaussiana com média $x_t'\beta$ e variância h_t, onde

$$h_t = \zeta + \alpha_1(y_{t-1} - \mathbf{x'_{t-1}}\beta)^2, \cdots, \alpha_m(y_{t-m} - \mathbf{x'_{t-m}}\beta)^2 \equiv [\mathbf{z_t}(\beta)]'^{\delta} \tag{3.31}$$

para

$$\begin{aligned} \delta &\equiv (\zeta, \alpha_1, \alpha_2, \cdots, \alpha_m)' && (3.32)\\ z_t(\beta)' &\equiv [1, (y_{t-1} - x'_{t-1}\beta)^2, \cdots, (y_{t-m} - x'_{t-m}\beta)^2] && (3.33) \end{aligned}$$

Coletando os parâmetros desconhecidos em um vetor θ de dimensão $(a\ x\ 1)$, onde

$$\theta \equiv (\beta', \delta')' \tag{3.33}$$

obtemos a função de log-verossimilhança da amostra, condicional às m primeiras observações

$$\begin{aligned} \mathscr{L}(\theta) &= \sum_{t=1}^{T} log f(y_t | x_t, \Upsilon_{t-1}; \theta) \\ &= -\left(\frac{T}{2}\right) log(2\pi) - (\frac{1}{2}) \sum_{t=1}^{T} log(h_t) - \left(\frac{1}{2}\right) \sum_{t=1}^{T} \frac{(y_{t-1} - x'_t\beta)^2}{h_t} \end{aligned} \tag{3.33}$$

É importante destacar que a maior parte das séries financeiras não apresenta distribuição Gaussiana. Nestes casos, é possível utilizar uma abordagem semelhante à qual apresentamos para o caso de máxima verossimilhança, porém considerando outras distribuições.

Assumindo que os resíduos da equação de regressão são não correlacionados com as variáveis explicativas e que o erro quadrado das previsões é não correlacionado com os quadrados dos resíduos defasados, de forma que

$$\begin{aligned} E[(y_t - x'_t\beta)] &= 0 && (3.34)\\ E[(u_t^2 - h_t)z_t] &= 0, && (3.35) \end{aligned}$$

Podemos utilizar a abordagem do *Método Generalizado dos Momentos* para a estimação dos parâmetros de uma regressão ARCH. Com base em Hamilton (1994), podemos caracterizar

a regressão de um processo ARCH com a suposição de que os resíduos da equação de regressão são não correlacionados com as variáveis explicativas, de forma que

$$E[(y_t - x_t'\beta)x_t] = 0 \tag{3.35}$$

e complementarmente, que o quadrado do erro de previsão seja não correlacionado com o quadrado dos resíduos da equação de regressão

$$E[(u_t^2 - h_t)z_t] = 0, \tag{3.35}$$

o que torna os parâmetros de um modelo GARCH passíveis de estimação através do método generalizado de momentos.

A implementação passa pela escolha de

$$\theta = (\beta', \delta')' \tag{3.35}$$

com o objetivo de minimizar

$$[g(\theta; \upsilon_t)]' \hat{S_T^{-1}} [g(\theta; \upsilon_t)], \tag{3.35}$$

onde $\hat{S_T}^{-1}$ representa a matriz de erros padrão para as estimativas dos parâmetros e

$$g(\theta; \upsilon_t) = \begin{bmatrix} T^{-1} \sum_{t=1}^{T} (y_t - x_t'\beta)x_t \\ T^{-1} \sum_{t=1}^{T} \{(y_t - x_t'\beta)^2 - [z_t(\beta)]'\delta\} z_t(\beta) \end{bmatrix}. \tag{3.35}$$

Até então, caracterizamos um processo ARCH(m) por

$$u_t = \sqrt{h_t} \cdot \upsilon_t \tag{3.35}$$

onde as inovações υ_t são independentes e identicamente distribuídas com média zero e variância unitária, e h_t é representado por

$$h_t^2 = \zeta + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \alpha_3 u_{t-3}^2 + \cdots + \alpha_m u_{t-m}^2. \tag{3.35}$$

Generalizando para um processo onde a variância condicional depende de uma quantidade infinita de defasagens de u_{t-j}^2, obtemos

$$h_t = \zeta + \pi(L)u_t^2, \tag{3.35}$$

onde

$$\pi(L) = \sum_{j=1}^{\infty} \pi_j L^j. \tag{3.35}$$

Parametrizando $\pi(L)$

$$\pi(L) = \frac{\alpha(L)}{1-\delta(L)} = \frac{\alpha_1 L^1 + \alpha_2 L^2 + \cdots + \alpha_m L^m}{1 - \delta_1 L^1 - \delta_2 L^2 - \cdots - \delta_r L^r} \tag{3.35}$$

Multiplicando 3.72 por $1-\delta(L) = 0$ e assumindo que as raízes de $1-\delta(L) = 0$ se encontram fora do circulo unitário, obtemos

$$[1-\delta(L)]h_t = [1-\delta_1]\zeta + \alpha(L)u_t^2, \tag{3.35}$$

e, finalmente

$$h_t = \kappa + \delta_1 h_{t-1} + \delta_2 h_{t-2} + \cdots + \delta_r h_{t-r} + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \cdots + \alpha_m u_{t-m}^2, \tag{3.35}$$

para

$$\kappa = [1 - \delta 1 - \delta_2 - \cdots - \delta_r]\zeta. \tag{3.35}$$

Definimos um processo do tipo definido pela equação 3.76 pelo modelo *generalizado autoregressivo de heterocedasticidade condicional*, ou simplesmente GARCH para a sigla em inglês, e escrevemos

$$u_t \sim GARCH(r,m) \tag{3.35}$$

A intuição por traz de um modelo GARCH não é trivial. Para facilitar o entendimento do modelo, Hamilton (1994) propõe a adição de u_t^2 à ambos os lados da equação 3.76, de forma a obtermos

$$\begin{aligned} h_t = \quad & \kappa - \delta_1(u_{t-1}^2 - h_{t-1}) + \delta_2(u_{t-2}^2 - h_{t-2}) + \cdots + \delta_r(u_{t-r}^2) - h_{t-r} \\ & + \delta_1 u_{t-1}^2 + \delta_2 u_{t-2}^2 + \cdots + \delta_r u_{t-r}^2 + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \cdots + \alpha_m u_{t-m}^2 + u_t^2. \end{aligned} \tag{3.35}$$

Notamos que h_t é a previsão de u_t^2 com base em suas próprias defasagens, e podemos definir $w_t \equiv u_t^2 - h_t$ como o erro associado a tal previsão, onde w_t é um processo de ruído branco.

Reordenando os termos da equação e introduzindo $w_(t)$, obtemos:

$$\begin{aligned} u_t^2 = \quad & \kappa + (\delta_1 + \alpha_1)u_{t-1}^2 + (\delta_2 + \alpha_2)u_{t-2}^2 + \cdots + (\delta_p + \alpha_p)u_{t-p}^2 \\ & + w_t - \delta_1 w_{t-1} - \delta_2 w_{t-2} - \cdots - \delta_r w_{t-r} \end{aligned} \tag{3.35}$$

o qual é reconhecido como um processo ARMA(p, r) para u_t^2 . O coeficiente para o j-ésimo coeficiente autoregressivo é dado por $(\delta_j + \alpha_j)$ e o coeficiente para o j-ésimo coeficiente de média móvel é dado por $-\delta_j$.

Percebemos então que um processo GARCH, é nada mais que aquele onde u_t segue um processo ARMA(p,q).

3.3 Vetores Autoregressivos (VAR)

"Como os grandes modelos atuais contém muitas restrições, a pesquisa empírica com o objetivo de testar teorias macroeconômicas conflitantes usualmente procede em um *framework* de uma ou poucas equações. Por esta razão, parece promissor investigar a possibilidade de construir grandes modelos em um estilo que não tenda à acumulação de restrições... Deveria ser possível estimar modelos macroeconômicos de larga escala em forma reduzida irrestrita, tratando todas as variáveis como endógenas"(SIMS, 1980).

Em econometria, usualmente temos que avaliar o impacto de uma variável sobre outra e este efeito muitas vezes pode variar no tempo. Nesta seção, Definimos alguns conceitos importantes antes de generalizarmos o modelo de intervenção para modelos de função de transferência e modelos multivariados com endogeneidade dos parâmetros .

Com base em (ENDERS, 2008), definimos uma função de impulso resposta como:

$$(1 - \alpha_1 L)y_t = \alpha_0 + c_0 z_t + \varepsilon_t \tag{3.35}$$

onde

$$y_t = \frac{\alpha_0}{1 - \alpha_1} + c_0 \sum_{i=0}^{\infty} a_1^i z_{t-1} + \sum_{i=0}^{\infty} a_1^i \varepsilon_{t-1}, \tag{3.35}$$

z_t é uma variável de intervenção, $\alpha_0 = \alpha_0/(1 - \alpha_1)$ é a média de longo prazo da série, c_0 representa o efeito do impacto da variável z_t em y_t e L é o operador de defasagem.

Permitindo que a sequência $\{z_t\}$ assuma valores diferentes dos assumidos por uma dummy, obtemos uma extensão do modelo de intervenção, dada por

$$y_t = \alpha_0 + A(L)y_{t-1} + C(L)z_t + B(L)\varepsilon_t \tag{3.35}$$

onde $A(L), B(L),$ *e* $C(L)$ são polinômios no operador de defasagem L. Neste modelo, a variável de intervenção $\{z_t\}$ pode ser qualquer processo estocástico exógeno e denominamos o polinômio $C(L)$ como uma *função de transferência*, no sentido de que a mesma determina como uma mudança na variável exógena z_t afeta a trajetória temporal de da variável endógena y_t.

Em alguns casos podemos não estar confiantes de que a variável de intervenção seja exógena. Nestas situações, uma solução, extensão natural das funções de impulso resposta e de transferência, é tratar cada variável do modelo de forma simétrica.

Como exemplo inicial, em um modelo de apenas duas variáveis a trajetória temporal de y_t pode ser afetada pelo valor presente e valores defasados de z_t e a trajetória temporal de z_t pode ser afetada pelas realizações atual e defasadas de y_t

$$y_t = \quad b_{10} - b_{11}z_t + \gamma_{11}y_{t-1} + \gamma_{12}z_{t-1} + \varepsilon_{yt} \tag{3.36}$$

$$z_t = \quad b_{20} - b_{21}y_t + \gamma_{21}y_{t-1} + \gamma_{22}z_{t-1} + \varepsilon_{zt}, \tag{3.37}$$

onde assumimos que âmbos y_t e z_t são estacionários, ε_{yt} e ε_{zt} são processos de ruído branco não correlacionados e com variância σ_y e σ_z.

Denominamos a equação acima como *vetor autoregressivo* (VAR) de primeira ordem. A ordem diz respeito à maior defasagem observada e a estrutura do sistema incorpora retroalimentações, dado que y_t e z_t se afetam mutuamente. Por conta dos efeitos contemporâneos mútuos, as equações acima não se encontram em forma reduzida e é útil encontramos um formato mais amigável; recorremos à algebra linear para escrever o sistema em foma compacta

$$\begin{bmatrix} 1 & b_{12} \\ b_{21} & 1 \end{bmatrix} \begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} b_{10} \\ b_{20} \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{yt} \\ \varepsilon_{zt} \end{bmatrix}.$$

Podemos representar o sistema acima de forma simplificada por:

$$Bx_t = \Gamma_0 + \Gamma_1 x_{t-1} + \varepsilon_t \tag{3.37}$$

onde

$$B = \begin{bmatrix} 1 & b_{12} \\ b_{21} & 1 \end{bmatrix}, \Gamma_1 = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix}, \Gamma_0 = \begin{bmatrix} b_{10} \\ b_{20} \end{bmatrix}, x_t = \begin{bmatrix} y_t \\ z_t \end{bmatrix} \; e \; \varepsilon_t = \begin{bmatrix} \varepsilon_{yt} \\ \varepsilon_{zt} \end{bmatrix}. \tag{3.37}$$

Pré-multiplicando a equação 3.87 por B^{-1} chegamos à forma padrão do vetor autoregressivo (VAR) de primeira ordem, denotado por

$$x_t = A_0 + A_1 x_{t-1} + e_t \tag{3.37}$$

onde:

$$A_0 = B^{-1}\Gamma_0, \;\; A_1 = B_{-1}\Gamma_1, \;\; e_t = B^{-1}\varepsilon_t \tag{3.37}$$

e generalizando para o caso de um VAR de ordem p, obtemos

$$x_t = A_0 + \sum_{i=1}^{p} A_i x_{t-i} + e_t. \tag{3.37}$$

Definindo a_{i0} como o i-ésimo elemento do vetor A_0, a_{ij} como o elemento na i-ésima linha e j-ésima coluna da matriz A_1 e e_{it} como o iésimo elemento do vetor e_t, reescrevemos as equação equações 3.85 e 3.86 como:

$$y_t = \quad a_{10} - a_{11} y_{t-1} + a_{12} z_{t-1} + e_{1t} \tag{3.38}$$

$$z_t = \quad a_{20} - a_{21} y_{t-1} + a_{22} z_{t-1} + e_{2t}, \tag{3.39}$$

onde os termos de erro e_{1t} *e* e_{2t} são compostos pelos choques ε_{yt} e ε_{zt} e obtemos

$$e_{1t} = \quad (\varepsilon_{yt} - b_{12}\varepsilon_{zt})/(1 - b_{12}b_{21}) \tag{3.40}$$

$$e_{2t} = \quad (\varepsilon_{zt} - b_{21}\varepsilon_{yt})/(1 - b_{12}b_{21}) \tag{3.41}$$

Como ε_{yt} e ε_{zt} são ruídos brancos, ambos e_{1t} e e_{2t} possuem valor esperado igual à zero, variância constante e são individualmente serialmente não-correlacionados.

Tomando a esperança de e_{1t}, obtemos

$$Ee_{1t} = E(\varepsilon_{yt} - b_{12}\varepsilon_{zt})/(1 - b_{12}b_{21}) = 0 \tag{3.41}$$

e tomando a esperança de e_{1t}^2 , obtemos

$$Ee_{1t}^2 = E[(\varepsilon_{yt} - b_{12}\varepsilon_{zt})/(1-b_{12}b_{21})]^2 = (\sigma_y^2 + b_{12}^2\sigma_z^2)/(1-b_{12}b_{21})^2. \tag{3.41}$$

A variância é então invariante no tempo, o que torna o processo estacionário, e as autocovariâncias entre e_{1t} e e_{1t-i} são dadas por

$$Ee_{1t}e_{1t-i} = \frac{E(\varepsilon_{yt} - b_{12}\varepsilon_{zt})(\varepsilon_{yt-i} - b_{21}\varepsilon_{zt-i})}{(1-b_{12}b_{21})^2} = 0. \tag{3.41}$$

É importante notar que os termos e_{1t} e e_{2t} são correlacionados, com covariância dada por:

$$Ee_{1t}e_{2t} = \frac{E(\varepsilon_{yt} - b_{12}\varepsilon_{zt})(\varepsilon_{zt} - b_{21}\varepsilon_{yt})}{(1-b_{12}b_{21})^2} = \frac{-(b_{21}\sigma_y^2 + b_{12}\sigma_z^2)}{(1-b_{12}b_{21})^2}. \tag{3.41}$$

Tal covariância é usualmente diferente de zero, por conta dos efeitos contemporâneos de uma variável na outra, e definimos a matriz de variância/covariância de e_{1t} e e_{2t} como:

$$\sum = \begin{bmatrix} Var(e_{1t}) & Cov(e_{1t}, e_{2t}) \\ Cov(e_{1t}, e_{2t}) & Var(e_{2t}) \end{bmatrix}. \tag{3.41}$$

Passamos agora para uma avaliação das condições necessárias para a estabilidade de modelos VAR. Iterando a forma padrão de um vetor autoregressivo

$$x_t = A_0 + \cdots + A_1 x_{t-1} + e_t, \tag{3.41}$$

obtemos

$$x_t = A_0 + A_1(A_0 + A_1 x_{t-2} + e_{t-1}) + e_t \tag{3.41}$$

e após n iterações, obtemos

$$x_t = (I + A_1 + \cdots + A_1^n)A_0 + \sum_{i=0}^{n} A_1^i e_{t-i} + A_1^{n+1} x_{t-n-1}, \tag{3.41}$$

onde I é uma matriz identidade de dimensão 2.

Para que tal expressão seja convergente, é necessário que a expressão A_1^n desapareça a medida que $n \to \infty$. Em suma, para que o modelo seja estável, precisamos que as raízes do polinômio

$$(1-a_{11}L)(1-a_{22}L)-(a_{12}a_{21}L^2) \tag{3.41}$$

se encontrem fora do circulo unitário. Assumindo que tal condição seja satisfeita, chegamos à solução particular para x_t :

$$x_t = \mu + \sum_{i=0}^{\infty} A_1^i e_{t-i}, \tag{3.41}$$

onde

$$\mu = |\bar{y}\ \bar{z}|', \tag{3.41}$$

$$\bar{y} = [a_{10}(1-a_{22}) + a_{12}a_{20}]/\Delta \tag{3.41}$$

$$\bar{z} = [a_{20}(1-a_{11}) + a_{21}a_{10}]/\Delta \tag{3.41}$$

e

$$\Delta = (1-a_{11})(1-a_{22}) - a_{12}a_{21}. \tag{3.41}$$

Tomando a esperança da equação (233), obtemos que a média incondicional de x_t é dada por μ e segue que, para obtermos a matriz de variância/covariância

$$E(x_t-\mu)^2 = E[\sum_{i=0}^{\infty} A_1^i e_{t-i}]^2 \tag{3.41}$$

e com base na definição da matriz de variância covariância, onde

$$\sum = \begin{bmatrix} Var(e_{1t}) & Cov(e_{1t},e_{2t}) \\ Cov(e_{1t},e_{2t}) & Var(e_{2t}) \end{bmatrix} = \sum = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \tag{3.41}$$

chegamos à

$$Ee_t^2 = E \begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix} \begin{bmatrix} e_{1t} & e_{2t} \end{bmatrix} = \sum. \tag{3.41}$$

Como a autocovariância $Ee_te_{t-i} = 0$, para $i \neq 0$, obtemos

$$E(x_t - \mu)^2 = (I + A_1^2 + A_1^4 + A_1^6 + \cdots)\sum = (I - A_1^2)^{-1}\sum \tag{3.41}$$

onde as sequências y_t e z_t serão conjuntamente covariância estacionárias, com médias finitas e invariantes no tempo, se assumirmos que a condição de estabilidade é válida, de forma que $A_1^n \to 0$ a medida que $n \to \infty$

Observamos que da mesma forma que um modelo AR possui uma representação de média móvel, um modelo VAR também o possui. Denominamos tal representação como um vetor de média móvel (VMA). A representação VMA faz parte da metodologia definida por Sims (1980) e nos permite traçar a trajetória temporal de varios choques contidos em um sistema VAR. Com base em Enders (2008), e mantendo um modelo de primeira ordem de duas variáveis, obtemos

$$\begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} a_{10} \\ a_{20} \end{bmatrix} + \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix}. \tag{3.41}$$

e representando os termos e_{1t} e e_{1t} em termos das sequências ε_{yt} e ε_{zt}, obtemos

$$\begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix} = [1/(1 - b_{12}b_{21})] \begin{bmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_{yt} \\ \varepsilon_{zt} \end{bmatrix}. \tag{3.41}$$

Combinando a expressão acima com a representação matricial, obtemos

$$\begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} \overline{y_t} \\ \overline{z_t} \end{bmatrix} + [1/(1 - b_{12}b_{21})] \sum_{i=0}^{\infty} \begin{bmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}. \tag{3.41}$$

Simplificando a notação, ainda em termos das sequências ε_{yt} e ε_{zt}, obtemos

$$\begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} \overline{y_t} \\ \overline{z_t} \end{bmatrix} + \sum_{i=0}^{\infty} \begin{bmatrix} \phi_{11}(i) & \phi_{12}(i) \\ \phi_{21}(i) & \phi_{22}(i) \end{bmatrix} \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}, \tag{3.41}$$

onde

$$\phi_i = [A_1^i/(1 - b_{12}b_{21})] \begin{bmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{bmatrix}. \tag{3.41}$$

A representação de média móvel é útil pois a mesma pode ser utilizada para gerar os efeitos dos choques ε_{yt} e ε_{zt} em toda a trajetória temporal das sequências y_t e z_t, de forma que cada um dos elementos $\phi_{12}(i)$ é denominado um *multiplicador de impacto*.

O efeito cumulativo dos choques pode ser avaliado a partir do somatório das funções de impulso resposta. O efeito cumulativo dos choques ε_{zt} no valor da sequência y_t é dado por

$$\sum_{i=0}^{n} \phi_{12}(i) \tag{3.41}$$

de tal sorte que a medida que $n \to \infty$, obtemos o *multiplicador de impacto de longo prazo*.

Como as sequências são estacionárias $\sum_{i=0}^{\infty} \phi_{jk}^2(i)$ é finita, $\forall j,k$.

De acordo com Enders (2008) um problema associado à metodologia descrita acima é que por construção, o sistema VAR a ser estimado é mal especificado. Portanto, para identificarmos o sistema primitivo, devemos impor uma restrição ao sistema, de forma a obtermos as funções de impulso resposta.

Uma das possíveis restrições de identificação é a decomposição de *Choleski*, a qual representa um conjunto mínimo de suposições com o objetivo de identificar o sistema primitivo (ENDERS, 2008, p. 307).

Segundo Enders (2008), a ideia de impôr uma estrutura ao sistema VAR parece conflitante ao argumento de contrario as restrições de identificação de um modelo. Neste contexto, a decomposição de Choleski passa a ser uma forma de minimizar a intervenção do pesquisador, e consequentemente, a minimização de vieses.

Dadas a necessidade de imposição de alguma restrição, é importante que o pesquisador possa avaliar a relação entre as variáveis do modelo, de forma a identificar qual das restrições seria "mais adequada". Com este objetivo em mente, entender as propriedades dos erros de previsão se torna extremamente importante para entender as interdependências entre as variáveis do sistema.

Focando, por hora, na sequência y_t, o erro de previsão n períodos a frente é dado por

$$y_{t+n} - E_t y_{t+n} = \phi_{11}(0)\varepsilon_{yt+n} + \phi_{11}(1)\varepsilon_{yt+n-1} + \cdots + \phi_{11}(n-1)\varepsilon_{yt+1} \tag{3.42}$$

$$+\phi_{12}(0)\varepsilon_{zt+n} + \phi_{12}(1)\varepsilon_{zt+n-1} + cdots + \phi_{12}(n-1)\varepsilon_{zt+1} \tag{3.43}$$

e a variância do erro de previsão n períodos a frente para y_t é dada por

$$\sigma_y(n)^2 = \sigma_y^2[\phi_{11}(0)^2 + \phi_{11}(1)^2 + \cdots + \phi_{11}(n-1)^2] \tag{3.44}$$

$$+\sigma_z^2[\phi_{12}(0)^2 + \phi_{12}(1)^2 + \cdots + \phi_{12}(n-1)^2] \tag{3.45}$$

.

Portanto, a variância do erro de previsão y_t períodos a frente pode ser decomposta de acordo com a contribuição de cada um dos choques e, como os valores dos multiplicadores de impacto são não negativos, o erro aumenta à medida que o horizonte de previsão n aumenta.

A proporção de cada um dos choques y_t e z_t na variância $\sigma_y(n)^2$ é dada por

$$\frac{\sigma_y^2[\phi_{11}(0)^2+\phi_{11}(1)^2+\cdots+\phi_{11}(n-1)^2]}{\sigma_y(n)^2}, \tag{3.45}$$

e

$$\frac{\sigma_z^2[\phi_{12}(0)^2+\phi_{12}(1)^2+\cdots+\phi_{12}(n-1)^2]}{\sigma_y(n)^2}, \tag{3.45}$$

respectivamente.

Em suma, a *decomposição da variância do erro de previsão* nos dá a proporção dos movimentos em uma sequência decorrentes seus próprios choques em relação aos movimentos decorrentes dos choques em outra variável. Como exemplo, caso os choques ε_{zt} não possuam poder explicativo sobre a variância do erro de previsão da sequência, dizemos que a sequência y_t é exógena.

Segundo Enders (2008), é recomendado que a decomposição da variância seja examinada ao longo de diversos períodos pois, à medida que o horizonte n aumenta, as decomposições devem convergir.

Quando analisadas conjuntamente, as funções de impulso resposta e a decomposição da variância possuem grande importância na determinação de relações entre variáveis econômicas.

Seguindo os princípios propostos por Sims (1980), podemos construir um modelo VAR a partir da inclusão de um grande número de variáveis, de forma a obtermos um VAR de n equações, onde cada equação contém p defasagens de cada uma das n variáveis do sistema. Há, no entanto um *trade-off* associado à inclusão de novas variáveis.

Considerando um caso onde temos dados mensais de cada variável, com 12 defasagens, a inclusão de uma variável adicional implica a perda de 12 graus de liberdade.

Representamos um VAR de n variáveis como

$$\begin{bmatrix} x_{1t} \\ x_{2t} \\ \vdots \\ x_{nt} \end{bmatrix} = \begin{bmatrix} A_{10} \\ A_{20} \\ \vdots \\ A_{n0} \end{bmatrix} + \begin{bmatrix} A_{11}(L) & A_{12}(L) & \cdots & A_{1n}(L) \\ A_{21}(L) & A_{22}(L) & \cdots & A_{2n}(L) \\ \vdots & \vdots & \vdots & \vdots \\ A_{n1}(L) & A_{n2}(L) & \cdots & A_{nn}(L) \end{bmatrix} \begin{bmatrix} x_{1t-1} \\ x_{2t-1} \\ \vdots \\ x_{nt-1} \end{bmatrix} + \begin{bmatrix} e_{1t} \\ e_{2t} \\ \vdots \\ e_{nt} \end{bmatrix},$$

onde os termos A_{i0} representam os interceptos das equações, os temos $A_{ij}(L)$ representam os polinômios no operador de defasagem L e todas as equações possuem a mesma extensão, gerando polinômios de mesmo grau. Isto gera uma condição sensível, a qual envolve a definição da extensão apropriada de defasagens.

Para o caso de um sistema com n equações e p defasagens, temos $n \cdot p$ coeficientes. Se o valor de p é demasiadamente reduzido, o modelo se torna mal especificado; se p é demasiadamente grande, graus de liberdade são desperdiçados. Felizmente, podemos endereçar esta questão a partir de alguns testes. Apesar de não fazermos um tratamento completo dos mesmo, uma descrição breve já é suficiente para evidenciar sua utilidade.

Uma estratégia proposta por Enders (2008) é a de iniciar a construção a partir de valores mais elevados de defasagem, considerando as condições referentes à quantidade de graus de liberdade. Posteriormente, estimamos o VAR e formamos a matriz de variância/covariância dos resíduos.

Supondo como exemplo, que iniciamos com 12 defasagens para cada variável, podemos testar se uma quantidade menor, por exemplo de 8 defasagens é adequada. Poderíamos definir se esta redução na quantidade de defasagens é adequada a partir da utilização de um *teste de razão de probabilidade*. Para tanto, reestimamos o modelo VAR para o mesmo período utilizando apenas oito defasagens e obtemos a matriz de variância/covariância dos resíduos, denotada por $\sum_8$. A estatística de razão de probabilidade é dada por

$$(T)(log|\sum_{8}| - log|\sum_{12}|). \tag{3.45}$$

Adaptando para o tamanho das amostras que usualmente são encontradas em análise econômica, (SIMS, 1980) recomenda a utilização da estatística

$$(T-c)(log|\sum_{8}| - log|\sum_{12}|) \tag{3.45}$$

onde T representa a quantidade de observações utilizáveis, c representa a quantidade de parâmetros estimados em cada equação e $log|\sum_n|$ é o logarítimo natural do determinando de $\sum_n$.

Em nosso exemplo de 12 defasagens, $c = 12n+1$ dado que cada uma das equações do modelo irrestrito possuem 12 defasagens para cada variável, além de um intercepto.

Apesar de ser válido em algumas situações, para o caso de pequenas amostras o teste de razão de probabilidade não é o mais adequado; isto se deve principalmente pelo fato do teste estar baseado na teoria assintótica e por conta da limitação da aplicação do teste somente para os casos onde um modelo é uma versão restrita de outro. Nestes casos, sugere-se a aplicação dos

testes de avaliação de critérios de informação, como o critério de informação de Aikaike (AIC) e o critério de informação Bayesiano [1], ou simplesmente BIC. Há ainda os testes de causalidade, como o teste de caudalidade de *Granger*.

De acordo com Enders (2008) o principal objetivo do teste de causalidade de Granger, no contexto da modelagem de um VAR, é identificar se uma variável deve entrar na equação de outra variável. Considerando o modelo VAR de duas variáveis e p defasagens, consideramos que y_t *não Granger-causa* z_t se, e somente se, todos os coeficientes $A_{21}(L)$ forem iguais à zero. Desta forma, assumimos que caso y_t não melhore a performance da previsão de z_t, então y_t não Granger-causa z_t. Podemos ainda ter a situação onde y_t não Granger-causa z_t, consequentemente não aumentando o performance preditiva de z_t, mas onde $\phi_{21} \neq 0$; de tal modo que choques à y_{t+1} afetem o valor de z_{t+1}, mesmo que a sequência y_t não Granger-cause a sequência $\{z_t\}$.

Segundo Enders (2008), a forma direta de se determinar a causalidade de Granger se da a partir da aplicação de um teste F, de forma a testar a restrição

$$a_{21}(1) = a_{21}(2) = a_{21}(3) = \cdots = 0. \tag{3.45}$$

Para o caso de n variáveis, com $A_{ij}(L)$ representando os coeficientes dos valores defasados da variável j na variável i, dizemos que a variável j não Granger-causa a variável i se todos os coeficientes do polinômio $A_{ij}(L)$ puderem ser igualados a zero.

Sims et al. (1986) trata da utilização e validade de Vetores Autoregressivos (VARs) para análise de política econômica. O autor destaca a utilização de três termos, recorrentemente utilizados sob significados distintos por economistas e não economistas: Forma Reduzida, Estrutura e Identificação.

De acordo com Sims et al. (1986, p. 3), uma utilização recorrente é a definição de *forma reduzida* como um modelo o qual descreve como alguns dados históricos, denominados X, são gerados por algum mecanismo aleatório. Quando estimamos um modelo em forma reduzida, construímos alguma estatística que resuma toda a coleção de dados X. A forma reduzida pode ser interpretada como um racional para tipos particulares de resumos para coleções de dados. A estrutura, ou *modelo estrutural*, é considerado com um modelo o qual pode ser utilizado para a tomada de decisão. Usualmente, estes modelos geram previsões sobre os resultados Z de diversas ações A as quais podem ser tomadas.

As noções de estrutura e forma reduzida estão implícitas na maior parte dos usos de dados para a tomada de decisões, no entanto, quando um econometrista utiliza estas noções o mesmo geralmente explicita a forma reduzida a partir de uma distribuição de probabilidade $p(X;\beta)$, para os dados como uma função dos parâmetros em forma reduzida β e explicita a

[1] Consultar Pfaff (2008) para um tratamento mais profundo sobre AIC e BIC

estrutura como uma distribuição condicional $q(Z|A;\alpha)$ dos resultados, dadas determinadas ações, as quais por sua vez dependem de parâmetros estruturais α.

A *identificação* é considerada como o estabelecimento de uma conexão entre a *forma reduzida* e a *estrutura*, de tal sorte que as estimativas dos parâmetros em forma reduzida β possam ser utilizados para determinar os parâmetros estruturais α.

Em suma, a identificação é a interpretação de variações históricas em uma coleção de dados, de forma que tais variações possibilitem a previsão das consequências de ações ainda não tomadas. No entanto, Sims et al. (1986) destaca que por vezes devemos contemplar ações muito diferentes daquelas observadas historicamente e, nestes casos, a identificação se torna mais difícil e controversa.

Os modelos estruturais em economia usualmente são formulados de forma que todo parâmetro no vetor α possua uma interpretação econômica. Em contrapartida, os elementos do vetor β são usualmente mais difíceis de serem interpretados, pois os mesmos refletem a influência da combinação do comportamento de diversos setores econômicos.

De acordo com Enders (2008), a abordagem de Sims (1980) possui a propriedade desejável de que todas as variáveis são tratadas de forma simétrica, de forma que o econometrista não se precise se basear em alguma restrição. Assim, um modelo VAR pode ser útil para examinar as relações entre um conjunto de variáveis econômicas, apesar de ser criticado como desprovido de qualquer conteúdo econômico. Nesta configuração, de acordo com Enders (2008), o único papel do economista é o de sugerir quais variáveis devem ou não entrar no modelo; a partir deste ponto, o procedimento é praticamente mecânico.

Com base em Enders (2008), a não ser que a estrutura do modelo possa ser identificada a partir da forma reduzida do VAR, as inovações decorrentes de uma decomposição de Choleski não possuem uma interpretação econômica direta.

É importante lembrar que os termos de erro de um VAR com duas variáveis e_{1t} e e_{2t} são na verdade compostos pelos choques ε_{yt} e ε_{zt}, definidos como

$$\begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix} = [1/(1-b_{12}b_{21})] \begin{bmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_{yt} \\ \varepsilon_{zt} \end{bmatrix},$$

onde os choques não possuem uma interpretação estrutural. Deste fato, decorre que há uma diferença importante entre a utilização de modelos VAR para previsão em relação à utilização dos mesmos para análise econômica. De acordo com Enders (2008), se estamos apenas interessados em previsão, os componentes dos erros de previsão não são importantes, ao passo que, se estamos interessados em obter um função de impulso resposta, ou uma decomposição da variância para rastrear os efeitos de inovações em y_t ou z_t, é necessário utilizar os choques estruturais ε_{yt} e ε_{zt}, ao invés dos erros de previsão. Ainda segundo o autor, o objetivo de um modelo

VAR estrutural é a utilização da teoria econômica, ao invés da decomposição de Choleski, para recuperar as inovações estruturais dos resíduos e_{1t} e e_{2t} .

Ao escolhermos variáveis a serem incluídas no modelo, é provável sejam escolhidas variáveis as quais apresentem comovimentos relevantes. De acordo com Enders (2008), quando os resíduos do VAR forem correlacionados, não é adequado que utilizemos a tentativa e erro para identificar um ordenamento adequado; principalmente quando a quantidade de variáveis e defasagens é significativa.

Se utilizarmos uma decomposição de Choleski, onde selecionamos um ordenamento de tal sorte que $b_{21} = 0$, obtemos que as inovações podem ser recuperadas como $\varepsilon_{zt} = e_{2t}$ e $\varepsilon_{yt} = e_{1t} - b_{12}e_{2t}$. Percebemos então que, forçar $b_{21} = 0$ é equivalente a assumir que uma inovação em y_t não possui um efeito contemporâneo em z_t . Portanto, a não ser que tenhamos uma fundação teórica para esta suposição, tais choques não estarão identificados adequadamente; o que gera funções de impulso resposta e decomposição de variância impropriamente identificadas e, consequentemente, resultados duvidosos.

Segundo Enders (2008), uma solução para tal questão seria modelar as inovações a partir da utilização de análise econômica, examinando a relação entre os erros de previsão e as inovações estruturais em um VAR de n variáveis. Com base em Enders (2008), como tal relação independe da extensão das defasagens, consideramos inicialmente um modelo de primeira ordem com n variáveis

$$\begin{bmatrix} 1 & b_{12} & \cdots & b_{1n} \\ b_{21} & 1 & \cdots & b_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ b_{n1} & b_{n2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_{1t} \\ x_{2t} \\ \vdots \\ x_{nt} \end{bmatrix} = \begin{bmatrix} b_{10} \\ b_{20} \\ \vdots \\ b_{n0} \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nn} \end{bmatrix} \begin{bmatrix} x_{1t-1} \\ x_{2t-1} \\ \vdots \\ x_{nt-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \vdots \\ \varepsilon_{nt} \end{bmatrix}, \quad (3.45)$$

o qual representamos de forma compacta por

$$Bx_t = \Gamma_0 + \Gamma_1 x_{t-1} + \varepsilon_t. \quad (3.45)$$

Selecionamos então, os valores observados de e_t e restringimos o sistema de forma a recuperar os valores de ε_t . Restringimos o sistema de forma a recuperar os ε_{it} e preservar a estrutura de erros com relação a independência dos choques ε_{it} .

Para chegar à quantidade necessária de restrições, obtemos a matriz de variância/covariância

$$\sum = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdot & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdot & \sigma_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \sigma_{n1} & \sigma_{n2} & \cdot & \sigma_n^2 \end{bmatrix} \tag{3.45}$$

onde cada elemento de $\sum$ é dado por

$$\sigma_{ij} = (1/T)\sum_{i=1}^{T} e_{it} e_{jt}. \tag{3.45}$$

Como $\sum$ contém $(n^2+n)/n$ elementos distintos, e precisamos obter n^2 variáveis desconhecidas, é necessária a imposição de

$$n^2 - \left[\frac{(n^2+n)}{2}\right] = \frac{(n^2-n)}{2} \tag{3.45}$$

restrições para identificarmos o modelo estrutural de um dado VAR estimado.

3.4 Cointegração e Modelos de Correção de Erros

Em diversas situações, as variáveis cujas relações são investigadas por um econometrista são não-estacionárias. Uma saída é buscar encontrar uma combinação linear destas variáveis não estacionárias que seja estacionária. Nestes casos, o conceito chave é o de *cointegração*. Nesta seção, introduzimos o conceito de cointegração, bem como os modelos de correção de erros, baseados fortemente neste conceito. A maior parte da exposição é baseada principalmente em Pfaff (2008), com alguns exemplos de aplicações importantes retirados de Enders (2008).

"Teorias de equilíbrio envolvendo variáveis não estacionárias requerem a existência de uma combinações de tais variáveis que seja estacionária"(ENDERS, 2008). O autor defende que tal afirmação representa a ideia que tem dominado boa parte da literatura macroeconômica recente. Para exemplificar, iniciamos com um modelo simples para a demanda por moeda, dado por:

$$m_t = \beta_0 + \beta_1 p_t + \beta_2 y_t + \beta_3 r_t + e_t \tag{3.45}$$

onde m_t representa a demanda por moeda de longo prazo, p_t representa o nível de preços, y_t representa a renda real, r_t , representa a taxa de juros, e_t é um termo de erro estacionário e os termos β_i são os parâmetros a serem estimados.

Uma investigação rápida sugere que a partir da teoria econômica clássica, devemos encontrar coeficientes $\beta_1 = 1, \beta_2 > 0$ *e* $\beta_3 < 0$. No entanto, as propriedades da parcela não

explicada da demanda por moeda, a saber e_t , representam uma parcela importante da teoria, de modo que quaisquer desvios na demanda por moeda devem ser temporários por natureza. Fica claro que se e_t possui uma tendência estocástica, os erros no modelo serão cumulativos e não haverá uma convergência, impedindo a existência de um equilíbrio.

Por conta dos fatores explicitados acima, fica claro que a sequência do termo de erro deve ser estacionária. No entanto, as variáveis as quais representam a renda real, a oferta de modela e o nível de preços podem todas ser caracterizadas como não estacionárias. Ora, mas para que haja equilíbrio, é necessário que o termo de erro, e consequentemente a combinação linear das variáveis acima seja estacionária. Denotando a equação 3.134 em termos das variáveis

$$e_t = m_t - \beta_0 - \beta_1 p_t - \beta_2 y_t - \beta_3 r_t, \tag{3.45}$$

fica claro que para que a sequência e_t seja estacionária, a combinação dos termos no lado direito deve ser estacionária. Disto recorre que, as trajetórias temporais das variáveis não estacionárias m_t , p_t , y_t , e r_t , devem obrigatoriamente estar relacionadas.

Enders (2008) destaca que a função de demanda por moeda é apenas um exemplo de combinação estacionária de variáveis não estacionárias e que dentro de um arcabouço teórico de equilíbrio, quaisquer desvios do equilíbrio devem ser temporários. O autor apresenta três exemplos envolvendo combinações estacionárias de variáveis não estacionárias, encontrados em economia.

1. Teoria da Função de Consumo: uma versão simplificada da hipótese de renda permanente afirma que o consumo total, denotado por c_t é igual à soma do consumo permanente, c_t^p , com o consumo transitório c_t^t . Como o consumo permanente é proporcional à renda permanente y_t^p , é possível considerar β como uma constante de proporcionalidade e escrever

$$c_t = \beta y_t^p + c_t^t. \tag{3.45}$$

Como o consumo transitório é uma variável estacionária por definição e assumindo que o consumo e a renda permanente são variáveis integradas de ordem 1, I(1), a hipótese de renda permanente exige que a combinação linear das variáveis consumo total c_t , e renda permanente y_t^p , dada por $c_t - \beta y_t^p$ seja estacionária.

2. Hipótese não Viesada de Mercado Futuro: uma das formas da hipótese de mercados eficientes (EMH) afirma que o preço *forward* de um ativo deve ser igual ao valor esperado do preço *spot* deste ativo no futuro. Denotando o log do preço-um-período-a-frente (forward) da taxa de câmbio EUR/USD no período t por f_t e denotando o log do preço spot da taxa de câmbio no período t , a teoria afirma que

$$E_t s_{t+1} = f_t. \tag{3.45}$$

Se esta relação falhar, especuladores podem realizar lucros nas suas operações no mercado de câmbio. Se as expectativas dos agentes forem racionais, o erro de previsão para a taxa *spot* em $t+1$ terá média condicional igual à zero, de forma que

$$s_{t+1} - E_t s_{t+1} = \varepsilon_{t+1}, \tag{3.45}$$

onde $E_t \varepsilon_{t+1} = 0$.

Combinando as duas equações, obtemos:

$$s_{t+1} = f_t + \varepsilon_{t+1}. \tag{3.45}$$

Como as variáveis $\{s_t\}$ e $\{f_t\}$ são I(1), a hipótese de mercado futuro não viesado exige que exista uma combinação linear das variáveis não estacionárias de taxa de câmbio *spot* e *forward* que seja estacionária.

3. Arbitragem no Mercado de Commodities e Paridade do Poder de Compra: teorias de competição geográfica sugerem que no curto prazo, os preços de produtos similares em mercados distintos podem variar. No entanto, arbitradores irão prevenir tais preços de se distanciar demasiadamente, mesmo que sejam não estacionários. De forma equivalente, os preços de computadores de marcas diferentes têm exibido movimentos correlatos. A teoria econômica sugere que o comovimento do preço dos produtos não pode apresentar muita divergência. A paridade do poder de compra impõe restrições sobre os movimentos de níveis de preço não estacionários e taxas de câmbio. Denotando o logarítimo do preço da taxa de câmbio por e_t e os níveis de preço doméstico e externo por p_t e p_t^* , respectivamente, a paridade do poder de compra de longo prazo exige que a combinação linear $e_t + p_t^* - p_t$ seja estacionária.

De acordo com Enders (2008), todos os três exemplos acima ilustram o conceito de **cointegração**.O conceito de cointegração foi introduzido na literatura em Granger (1981), e seu caso de aplicação geral foi publicado em Engle e Granger (1987).

De acordo com Pfaff (2008), a ideia por trás da do conceito de cointegração é encontrar uma combinação linear entre duas variáveis integradas de mesma ordem I(d) que resulte em uma variável com ordem mais baixa de integração. Com base em Pfaff (2008), definimos que os componentes de um vetor x_t são cointegrados de ordem *d, b*, denotado por $x_t \sim CI(d,b)$, se todos os componentes de x_t são integrados de ordem d, I(d), e se existe um vetor $\alpha \neq 0$ de forma que

$$z_t = \alpha' x_t \sim I(d,b), \;\; b < 0, \tag{3.45}$$

onde denominamos α como o *vetor de cointegração*. O maior avanço proporcionado por tal desenvolvimento é a possibilidade de se detectar relações estáveis de longo prazo entre variáveis não estacionárias. Em termos econômicos, desvios em relação a um equilíbrio são permitidos, porém tais erros são caracterizados por reversão à média em direção ao equilíbrio estável de longo prazo.

Com relação à estimação do vetor de cointegração α, Engle e Granger (1987) propoem um procedimento de estimativa em duas etapas. Na primeira etapa, rodamos uma regressão nas variáveis I(1)

$$y_t = \alpha_1 x_{t,1} + \alpha_2 x_{t,2} + \cdots + \alpha_k x_{t,k} + z_t, \tag{3.45}$$

para $t = 1, \cdots, T$ e onde z_t representa o termo de erro. O vetor de cointegração estimado $\hat{\alpha}$ é dado por

$$\hat{\alpha} = (1, -\hat{\alpha}^*), \tag{3.45}$$

onde $\hat{\alpha} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_k)'$.

Nesta regressão estática, o vetor de cointegração pode ser estimado consistentemente, no entanto, com um viés amostral finito de magnitude $O_p(T^{-1})$, e os resíduos da regressão, $\hat{z}_t$, devem ser integrados de ordem zero. Tais resíduos representam o erro em relação à trajetória de equilíbrio de longo prazo das variáveis integradas de ordem um I(1). Podemos testar a estacionariedade dos termos de erro a partir do teste de *Dickey-Fuller* (DF) ou do teste de *Dickey-Fuller aumentado* (ADF); Consultar Pfaff (2008) para um tratamento mais profundo sobre os testes DF e ADF. Uma vez que tenhamos rejeitado a hipótese de raiz unitária na série $\hat{z}_t$, podemos passar para a segunda etapa, onde especificamos um *modelo de correção de erros*, ou simplesmente ECM.

Inicialmente, consideramos o caso bivariado, com duas variáveis cointegradas y_t e x_t, âmbas integradas de ordem um I(1). Com base em Pfaff (2008), a especificação geral de um modelo de correção de erros (ECM) é dada por

$$\Delta y_t = \psi_0 + \gamma_1 \hat{z_t}_{-1} + \sum_{i=1}^{K} \psi_{1,i} \Delta x_{t-i} + \sum_{i=1}^{L} \psi_{2,i} \Delta y_{t-i} + \varepsilon_{1,t} \tag{3.46}$$

$$\Delta x_t = \xi_0 + \gamma_2 \hat{z_t}_{-1} + \sum_{i=1}^{K} \xi_{1,i} \Delta y_{t-i} + \sum_{i=1}^{L} \xi_{2,i} \Delta x_{t-i} + \varepsilon_{1,t} \tag{3.47}$$

onde z_t representa o erro oriundo da regressão 3.141 e os termos $\varepsilon_{1,t}$ e $\varepsilon_{2,t}$ são processos de ruído branco. Simplificadamente, a equação 3.143 implica que as mudanças em y_t são explicadas por sua própria história, representada por $\sum_{i=1}^{L} \psi_{2,i}\Delta y_{t-i}$, mudanças defasadas de x_t , representadas por $\sum_{i=1}^{K} \psi_{1,i}\Delta x_{t-i}$, e pelo erro com relação ao equilíbrio de longo prazo do período anterior, representado por $\gamma_1 \hat{z_{t-1}}$.

O coeficiente γ_1 determina a velocidade de ajuste e deve sempre possuir sinal negativo pois, caso contrário, o sistema iria divergir de sua trajetória de equilíbrio. Vale ressaltar que é possível generalizar o a equação para o caso com mais de uma defasagem para o termo de erro com relação ao equilíbrio de longo prazo. Ademais, podemos concluir que para um caso de duas variáveis cointegradas I(1), deve existir causalidade de Granger em pelo menos uma direção.

Para tratarmos do caso com n variáveis, é necessário realizar algumas extensões dos conceitos apresentados até então.

Com base em Pfaff (2008), consideremos novamente o processo

$$y_t = TD_t + z_t, \tag{3.47}$$

onde a série $\{y_t\}$ representa a realização de uma tendência determinística e um componente estocástico.

Assumimos que y_t é um vetor de dimensão $(Kx1)$, para $t = 1, \cdots, T$ onde cada um de seus componentes pode ser representado por

$$y_{i,t} = TD_{i,t} + z_{i,t}, \tag{3.47}$$

para $i = 1, \cdots, K$ e $t = 1, \cdots, T$, onde $TD_{i,t}$ representa o componente determinístico da i-ésima variável e $z_{i,t}$ representa o componente estocástico como um processo ARMA, $\phi_i(L)z_{i,t} = \theta_i(L)\varepsilon_{i,t}$.

De forma geral, definimos um vetor de dimensão $(nx1)$ como *cointegrado* se existe pelo menos um vetor $\beta_i \neq 0$, de tal sorte que $\beta_i' y_t$ seja tendência estacionário; denotamos tal vetor β_1 como um *vetor de cointegração*. Se existirem r vetores $\beta_i (i = 1, \cdots, r)$, linearmente indepententes, definimos que a sequência y_t é cointegrada, com ordem de cointegração r. Finalmente, definimos a matriz de vetores de cointegração de dimensão $(n\,x\,r)$ como

$$\beta = (\beta_1, \cdots, \beta_r). \tag{3.47}$$

Os r elementos do vetor $\beta_i' y_t$ são tendência estacionários e β é denominada *matriz de cointegração*.

Consideremos agora um modelo vetor autoregressivo de ordem p:

$$y_t = \Pi_1 y_{t-1} + \cdots + \Pi_K y_{t-p} + \mu + \Phi D_t + \varepsilon_t, \tag{3.47}$$

onde y_t denota o vetor das K séries no período t, as matrizes $\Pi_i (i = 1, \cdots, p)$ são as matrizes de coeficientes, de dimensão $(K\ x\ K)$, das variáveis endógenas defasadas, μ é um vetor de constantes de dimensão $(K\ x\ 1)$, D_t é um vetor de variáveis não estocásticas, como variáveis dummy sazonais ou dummies de intervenção, e o termo ε_t, de dimensão $(K\ x\ 1)$, é o termo de erro $\varepsilon_t \sim N(0, \Sigma)$, o qual assumimos seguir uma distribuição normal, aproximadamente .

Pfaff (2008) propõe duas especificações do *Modelo de Correção de Erros Vetorial* (VECM), baseadas na equação 3.148.

Na primeira, introduzimos os valores de y_t com defasagem $t - p$ e obtemos

$$\Delta y_t = \Gamma_1 \Delta y_{t-1} + \cdots + \Gamma_{p-1} \Delta y_{t-p+1} + \Pi y_{t-p} + \mu + \Phi D_t + \varepsilon_t, \tag{3.47}$$

onde $\Gamma_i = -(I - \Pi_1 - \cdots - \Pi_i)$, para $i = 1, \cdots, p-1$, $\Pi = -(I - \Pi_1 - \cdots - \Pi_p)$ e I é a matriz identidade de dimensão $(K\ x\ K)$. As matrizes $\Gamma_i (i = 1, \cdots, p-1)$ contêm os impactos cumulativos de longo prazo, desta forma, esta especificação é denominada a *forma de longo prazo* do VECM.

A segunda especificação do toma a forma

$$\Delta y_t = \Gamma_1 \Delta y_{t-1} + \cdots + \Gamma_{p-1} \Delta y_{t-p+1} + \Pi y_{t-1} + \mu + \Phi D_t + \varepsilon_t, \tag{3.47}$$

onde $\Gamma_i = -(I - \Pi_{i+1} - \cdots - \Pi_p)$, para $i = 1, \cdots, p-1$ e $\Pi = -(I - \Pi_1 - \cdots - \Pi_p)$. Notamos que a matriz Π é a mesma da primeira especificação. No entanto, a matriz Γ_i se diferencia da matriz que introduzimos na primiera especificação, pois agora, a mesma mede efeitos transitórios. Definimos esta forma como *forma transitória* do VECM. É importante notar que nesta especificação os componentes em y_t entram no modelo defasados por um período.

O termo de correção de erros é dado por Πy_{t-p} na primeira especificação e por Πy_{t-1} na segunda, devendo ser estacionário em ambos os casos. É necessário então que tenhamos uma forma de avaliar a estacionariedade da matriz Π. Consideramos três casos.

No primeiro, a matriz Π apresenta posto igual à K (rank K), e denotamos $rk(\Pi) = K$. Neste caso, todas as K combinações lineares independentes devem ser estacionárias; o que só pode ser alcançado se os desvios de y_t com relação aos componentes determinísticos forem estacionários.

No segundo caso, o posto da matriz Π é igual à zero, de forma que não existe nenhuma combinação linear que transforme o termo Πy_t em estacionário, a não ser a solução trivial.

No terceiro caso, temos que

$$0 < rk(\Pi) = r < K, \tag{3.47}$$

onde, como a matriz não possui *rank completo* (full rank), existem duas matrizes α e β de dimensão $(K\,x\,r)$, tais que $\Pi = \alpha\beta'$. Desta forma, o termo $\alpha\beta' y_{t-p}$ é estacionário e o produto entre a vetor e matriz $\beta' y_{t-p}$ também é estacionário.

As r colunas linearmente independentes de β são os vetores de cointegração e o posto da matriz Π é igual ao posto do sistema y_t, de tal sorte que cada coluna representa a relação de longo prazo entre as séries individuais de y_t e os elementos da matriz α determinam a velocidade de ajuste do equilíbrio de longo prazo; α é denominada *matriz de ajuste*.

Os conceitos apresentados nas últimas seções, quando avaliados sem um contexto específico podem parecer um tanto quanto abstratos. O trabalho de Enders (2008) soluciona o problema da abstração a partir da introdução de alguns exemplos da generalização dos resultados e do estabelecimento da relação entre os modelos VAR e os modelos de correção de erros. De forma simplificada, em um modelo de correção de erros a dinâmica de curto prazo das variáveis do sistema é influenciada pelo desvio de longo prazo do equilíbrio.

Com o objetivo de evidenciar a relação entre os modelos VAR e os modelos de correção de erros, iniciamos com a representação de um modelo VAR bivariado simples, denotado por

$$y_t = a_{11} y_{t-1} + a_{12} z_{t-1} + \varepsilon_{yt} \tag{3.48}$$

$$z_t = a_{21} y_{t-1} + a_{22} z_{t-1} + \varepsilon_{zt} \tag{3.49}$$

onde os termos ε_{yt} e ε_{zt} são processos de ruído branco que podem estar correlacionados um com o outro. Em notação matricial, obtemos

$$\begin{bmatrix} \Delta y_t \\ \Delta z_t \end{bmatrix} = \begin{bmatrix} a_{11} - 1 & a_{12} \\ a_{21} & a_{22} - 1 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{yt} \\ \varepsilon_{zt} \end{bmatrix} \tag{3.49}$$

e as soluções para y_t e z_t em termos de operadores de defasagens L são dadas por

$$y_t = \frac{(1-a_{22}L)\varepsilon_{yt}+a_{12}L\varepsilon_{zt}}{(1-a_{11}L)(1-a_{22}L)-a_{12}a_{21}L^2} \tag{3.50}$$

$$z_t = \frac{a_{21}L\varepsilon_{yt}+(1-a_{11}L)\varepsilon_{zt}}{(1-a_{11}L)(1-a_{22}L)-a_{12}a_{21}L^2} \tag{3.51}$$

e percebemos que ambas as equações possuem as mesmas raízes para a equação característica inversa, denotada por

$$(1-a_{11}L)(1-a_{22}L)-a_{12}a_{21}L^2. \tag{3.51}$$

Para obtermos as raízes da equação, igualamo-as à zero e resolvemos para L. Definindo $\lambda = \frac{1}{L}$, podemos trabalhar diretamente com as raízes características e reescrevemos a equação característica como

$$\lambda^2 - (a_{11}+a_{22})\lambda + (a_{11}a_{22}-a_{12}a_{21}) = 0, \tag{3.51}$$

onde as raízes da equação acima definem as trajetórias temporais de y_t e z_t. Para garantir que as variáveis são cointegradas de ordem 1, $C(1,1)$, é necessário que uma das raízes características seja igual à unidade e que a outra seja menor que a unidade em termos absolutos. Obtemos então que para que tal fato seja observado, os coeficientes devem satisfazer

$$a_{11} = \frac{[(1-a_{22})-a_{12}a_{21}]}{(1-a_{22})} \tag{3.51}$$

e considerando a segunda raíz característica, como os coeficientes a_{12} e/ou a_{21} devem ser diferentes de zero, vale que

$$a_{22} > 1 \quad e \quad a_{12}a_{21} + (a_{22})^2 < 1, \tag{3.51}$$

dada a condição de que $|\lambda_2| < 1$.

Logo, podemos escrever 3.154 como

$$\Delta y_t = -[a_{11}a_{21}/(1-a_{22})]y_{t-1} + a_{12}z_{t-1} + \varepsilon_{yt} \tag{3.52}$$

$$\Delta z_t = a_{21}y_{t-1} - (1-a_{22})z_{t-1} + \varepsilon_{zt}, \tag{3.53}$$

de forma que a equação acima representa o modelo de correção de erros, obtido a partir da representação em forma de VAR.

De acordo com Enders (2008), caso a_{12} *e* a_{21} sejam diferentes de zero, podemos normalizar o vetor de cointegração com relação a cada variável. Normalizando com relação à y_t obtemos

$$\Delta y_t = \alpha_y(y_{t-1} - \beta z_{t-1}) + \varepsilon_{yt} \tag{3.54}$$
$$\Delta z_t = \alpha_z(y_{t-1} - \beta z_{t-1}) + \varepsilon_{zt}, \tag{3.55}$$

onde

$$\alpha_y = -a_{12}a_{21}/(1-a_{22}) \tag{3.56}$$
$$\beta = (1-a_{22})/a_{21} \tag{3.57}$$
$$\alpha_z = a_{21}. \tag{3.58}$$

Tais restrições são necessárias para garantir que as variáveis são $C(1,1)$ e mesmas asseguram a existência do modelo de correção de erros. Fica claro também que a existência de cointegração faz necessária a imposição de restrições ao modelo VAR.

Considerando

$$x_t = (y_t, z_t)' \tag{3.59}$$
$$\varepsilon_t = (\varepsilon_{yt}, \varepsilon_{zt})', \tag{3.60}$$

representamos a equação na forma

$$\Delta x_t = \pi x_{t-1} + \varepsilon_t. \tag{3.61}$$

Segundo (ENDERS, 2008), é inapropriado estimar um VAR de variáveis cointegradas utilizando apenas primeiras diferenças. Caso 3.171 seja estimada sem a expressão πx_{t-1} , estaremos eliminando a parcela de correção de erros do modelo e é importante notar que as linhas de π não serão linearmente independentes caso as variáveis sejam cointegradas.

Finalmente, em um sistema cointegrado, ambas as variáveis respondem a desvios em relação ao equilíbrio de longo prazo. No entanto, é possível que a velocidade de ajuste dos parâmetros de uma delas seja igual à zero. Nesta situação, tal variável não responde ao desequilíbrio e a outra variável é responsável por todo o ajuste. Como as duas variáveis são relacionadas, o

efeito acaba sendo transmitido de uma para a outra. Neste caso, Enders destaca que é necessário reinterpretar o conceito de *causalidade de Granger*: "Em um sistema cointegrado, $\{z_t\}$ não *Granger-causa* $\{y_t\}$ se os valores defasados Δz_{t-i} não entrem na equação Δy_t e se y_t não responde a desvios em relação ao equilíbrio de longo prazo"(ENDERS, 2008).

4 Abordagem Bayesiana e Modelos Lineares Dinâmicos

4.1 A Abordagem Bayesiana

Uma das principais características dos modelos lineares dinâmicos é o afastamento de uma visão de mundo determinística, em direção à visão estocástica dos sistemas estudados. A incerteza é um conceito fundamental, estando presente seja por variáveis omitidas, erros de mensuração ou imperfeições, todas avaliadas enquanto probabilidades. Segundo Campagnoli, Petrone e Petris (2009), sistemas lineares dinâmicos são baseados na ideia de descrever o resultado de um sistema dinâmico, como uma serie temporal, como função de um processo não observável, afetado por erros aleatórios.

Tais modelos vem sendo utilizados extensivamente dentro de uma abordagem Bayesiana, principalmente por conta da possibilidade de se chegar aos resultados através de métodos computacionais recursivos, como métodos de Monte Carlo. Campagnoli, Petrone e Petris (2009) afirma que o ponto central da estatística Bayesiana é o fato de que toda a incerteza com relação a um fenômeno que estamos investigando deve ser descrita em termos de probabilidade. Neste contexto, a probabilidade possui uma interpretação subjetiva, sendo um meio de expressar a incompletude de informações que um pesquisados possui sobre um evento em estudo.

Em Bayes, Price e Canton (1763), fica clara a importância da incerteza e da probabilidade como forma de expressão da mesma. No trecho:

> "From the preceding proposition it is plain, that in the case of such an event as I there call M, from the number of trials it happens and fails in a certain number of trials, without knowing any thing more concerning it, one may give a guess whereabouts it's probability is, and, by the usual methods computing the magnitudes of the areas there mentioned see the chance that the guess is right. And that the same rule is the proper one to be used in the case of an event concerning the probability of which we absolutely know nothing antecedently to any trials made concerning it, seems to appear from the following consideration: viz. that concerning such an event I have no reason to think that, in a certain number of trials, it should rather happen any one possible number of times than another. For, on this account, I may justly reason concerning it as if its probability had been at first unfixed, and then determined in such a manner as to give me no reason to think that, in a certain number of trials, it should rather happen any one possible number of times rather than another"(BAYES; PRICE; CANTON, 1763, p. 11).

percebemos que a abordagem difere daquela tradicional, frequentista. Em geral, uma abordagem de inferência frequentista não utiliza distribuições de probabilidade para os parâmetros desconhecidos, de forma que a inferência dos parâmetros é baseada na determinação de estimadores os quais possuem propriedades desejáveis, intervalos de confiança e testes de hipótese. É importante notar que os parâmetros não são interpretados como variáveis aleatórias, e portanto, não é possível atribuir probabilidades aos resultados dos mesmos. Na abordagem Bayesiana, adota-se a probabilidade subjetiva, onde os parâmetros são considerados como aleatórios por conta da incerteza do pesquisador.

Em geral, a abordagem Bayesiana com relação ao processo de aprendizado se resume à aplicação de leis de probabilidade, de forma que a probabilidade condicional de um evento, dadas as informações disponíveis, é computada. Instrumentalmente, aplica-se o teorema de Bayes para expressar as probabilidades condicionais.

Dados dois eventos A e B, a probabilidade conjunta de ocorrência de A e B é dada por

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A) \tag{4.0}$$

e o teorema de Bayes pode ser representado como

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{4.0}$$

Usualmente encontramos esta configuração quando A representa um evento de interesse para o pesquisador e B é um resultado experimental, o qual acredita-se conter informações relevantes sobre A. O problema de "aprender"sobre A a partir da evidência experimental B é resolvido a partir do cálculo da probabilidade condicional $P(A|B)$.

No campo de inferência estatística, o fato experimental é usualmente o resultado de um procedimento amostral, sendo descrito pelo vetor aleatório Y. É comum a utilização um modelo paramétrico, onde é atribuída uma distribuição de probabilidade à Y e onde a quantidade de interesse é representada pelo vetor dos parâmetros, denotado por θ. Já a inferência Bayesiana de θ consiste em calcular sua distribuição condicional, dados os resultados amostrais. Os conceitos expostos até o momento compõem boa parte da base da abordagem Bayesiana. Passamos agora para alguns conceitos cuja aplicação fica mais evidente. A notação e ordem de apresentação dos conceitos deste capítulo é baseada em Campagnoli, Petrone e Petris (2009).

Para o caso do vetor aleatório Y, introduzido acima, consideremos que, com base no seu conhecimento do problema, o pesquisador possa atribuir uma distribuição condicional $\pi(y|\theta)$ para Y dado θ, a verossimilhança e uma distribuição a priori $\pi(\theta)$, expressando sua

incerteza com relação ao parâmetro θ. Através da observação de $Y = y$, podemos utilizar a *fómula de Bayes* para calcular a densidade condiconal de θ dado y

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)}, \tag{4.0}$$

onde $\pi(y)$ representa a distribuição marginal de Y, denotada por

$$\pi(y) = \int \pi(y|\theta)\pi(\theta)d\theta. \tag{4.0}$$

Em diversas aplicações em economia e finanças, estamos interessados na *previsão* de algumas variáveis, de forma que o evento de interesse passa a ser a observação do valor futuro de uma variável Y^*, por exemplo. A solução da previsão de Y^* dadas as informações observadas y na abordagem Bayesiana, é dada pelo cálculo da distribuição condicional de Y^* dado que $Y = y$, a qual denominamos *distribuição preditiva* e denotamos por

$$\pi(y^*|y) = \int \pi(y^*, \theta|y)d\theta = \int \pi(y^*|y, \theta)d\theta. \tag{4.0}$$

Prosseguimos com a exposição de outro conceito central na abordagem Bayesiana, as estruturas de dependência. O objetivo não é fazer um tratamento formal de todas as estruturas de dependência, mas sim introduzir algumas estruturas simples e importantes.

Iniciamos com a *independência condicional*, em seguida introduzimos a *permutabilidade* e por fim tratamos da estrutura mais comum nos problemas reais, a de *dados heterogêneos*.

Segundo Campagnoli, Petrone e Petris (2009), a independência condicional é a estrutura de dependência mais simples. Em algumas aplicações, é razoável assumir que os termos $Y_1, \cdots, Y_n$ são condicionalmente independentes e identicamente distribuídos dado θ, de forma que

$$\pi(y_{1:n}|\theta) = \Pi_{i=1}^{n}\pi(y_i|\theta). \tag{4.0}$$

Considerando que as medidas dos termos Y_i são afetadas por um erro aleatório, obtemos

$$Y_i = \theta + \varepsilon_i, \tag{4.0}$$

onde os ε_i são erros aleatórios Gaussianos com média igual a zero e variância constante igual à σ. Obtemos que, condicionalmente à $theta$, os termos Y_i são independentes e identicamente distribuídos, com

$$Y_i|\theta \sim N(\theta, \sigma^2). \tag{4.0}$$

É importante notar que os termos Y_i são apenas condicionalmente independentes, de forma que a observação de y_i nos dá informações importantes sobre os valores desconhecidos dos parâmetros θ e, através de $theta$, nos da informações sobre a observação do próximo período Y_{n+1} .

Tal configuração implica que Y_{n+1} depende em termos probabilísticos das observações passadas $Y_1, \cdots, Y_n$ e computamos a densidade preditiva como:

$$\begin{aligned} \pi(y_{n+1}|y_{1:n}) &= \int \pi(y_{n+1}, \theta|y_{1:n}) d\theta \\ &= \int \pi(y_{n+1}|\theta, y_{1:n}) \pi(\theta|y_{1:n}) d\theta \\ &= \int \pi(y_{n+1}|\theta) \pi(\theta|y_{1:n}) d\theta \end{aligned} \tag{4.-1}$$

onde o termo $\pi(\theta|y_{1:n})$ é a densidade à posteriori de θ condicional aos dados $(y_1, \cdots, y_n)$.

A densidade à posteriori é calculada a partir da fórmula de Bayes

$$\pi(\theta|y_{1:n}) = \frac{\pi(y_{1:n}|\theta)\pi(\theta)}{\pi(y_{1:n})} \propto \Pi_{t=1}^{n} \pi(y_{1:n}|\theta)\pi(\theta), \tag{4.-1}$$

de forma que a densidade marginal $\pi(y_{1:n})$ não depende de θ, assumindo o papel de uma constante de normalização e implicando que a densidade à posteriori seja proporcional ao produto entre a verossimilhança e a densidade a priori.

É importante notar que assumindo independência condicional entre os termos, podemos calcular a densidade a posteriori recursivamente. Isso significa que não é necessário reprocessar o cálculo da densidade a cada nova medida. Descrevemos a informação sobre $theta$ disponível no período $(n-1)$ como a densidade condicional

$$\pi(\theta|y_{1:n-1}) \propto \Pi_{t=1}^{n-1} \pi(y_t|\theta)\pi(\theta), \tag{4.-1}$$

de forma que no período n , tal densidade assume o papel de distribuição a priori.

À medida a nova observação y_n se torna disponível, devemos calcular a verossimilhança, denotada por

$$\pi(y_n|\theta, y_{1:n-1}) = \pi(y_n|\theta), \tag{4.-1}$$

assumindo independência condicional e atualizar a distribuição a priori $\pi(\theta|y_{1:n-1})$ pela regra de Bayes, obtendo

$$\pi(\theta|y_{1:n-1},y_n) \propto \pi(\theta|y_{1:n-1})\pi(y_n|\theta) \propto \Pi_{t=1}^{n-1}\pi(y_t|\theta)\pi(\theta)\pi(y_n|\theta). \tag{4.-1}$$

A segunda estrutura de dependência mencionada em Campagnoli, Petrone e Petris (2009) é a *permutabilidade*. Cosideremos uma sequência infinita $Y_t : t = 1,2,...$ de vetores aleatórios e suponhamos que a ordem dos termos na sequência não seja relevante, de forma que para qualquer $n \geq 1$, o vetor $(Y_1,\cdots,Y_n)$ e qualquer uma das permutações de seus componentes possua a mesma distribuição. Neste caso, dizemos que a sequência $Y_t : t = 1,2,...$ é permutável. A suposição de permutabilidade é equivalente a assumir independência condicional e distribuições idênticas; tal resultado é proposto pelo *teorema de representação de Finetti* (CAMPAGNOLI; PETRONE; PETRIS, 2009, p. 9).

De forma simplificada, o teorema de representação de Finetti afirma que para uma sequência infinita de vetores aleatórios permutáveis, representada por $Y_t : t = 1,2,...$, a sequência de funções de distribuições empíricas:

$$F_n(y) = F_n(y;Y_1,\cdots,Y_n) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,y]}(Y_i) \tag{4.-1}$$

converge fracamente para uma função de distribuição aleatória F , à medida que $n \to \infty$, com probabilidade 1. Ademais, para qualquer $n \geq 1$, a função de distribuição de $Y_1,\cdots,Y_n$ pode ser representada como

$$P(Y_1 \leq y_1,\cdots,Y_n \leq y_n) = \int \Pi_{i=1}^{n}\pi(y_i)d\pi(F). \tag{4.-1}$$

Se assumimos que a sequência de variáveis observáveis Y_t é permutável, podemos pensar nas mesmas como independentes e identicamente distribuídas, condicionalmente à função de distribuição F . A distribuição a priori π é uma lei de probabilidade no espaço F de todas a funções de distribuição no espaço amostral Y e deve expressar nossa crença quanto ao limite das funções de distribuição empíricas. De acordo com Campagnoli, Petrone e Petris (2009), em diversos problemas podemos restringir a distribuição a priori a uma classe paramétrica

$$P_\Theta = \{\pi(\cdot|\theta), \theta \in \Theta\} \subset F \tag{4.-1}$$

onde $\Theta \subseteq \mathbb{R}^p$, e denominamos a distribuição a priori como paramétrica. Em resumo, para o caso de uma priori paramétrica, o teorema de representação de Finetti implica que os $Y_1, Y_2, \cdots$

são condicionalmente independentes e identicamente distribuídos dado θ , possuindo a mesma função de distribuição $\pi(\cdot|\theta)$ e θ possui distribuição a priori $\pi(\theta)$.

A terceira estrutura de dependência é a de *dados heterogêneos*. Quando permitimos heterogeneidade entre os dados assumimos que

$$Y_1,\cdots,Y_n|\theta_1,\cdots,\theta_n \sim \Pi_{t=1}^{n} f_t(y_t|\theta_t) \tag{4.-1}$$

de forma que os termos $Y_1,\cdots,Y_n$ são condicionalmente independentes dado um vetor $theta = (\theta_1,\cdots,\theta_n)$, com Y_t dependendo somente do valor correspondente θ_t.

Quando estamos interessados no cálculo das distribuições condicionais de probabilidade de parâmetros multivariados, é interessante que possamos apresentar uma síntese das distribuições preditivas e posteriori.

Consideremos o caso de inferência em um parâmetro multivariado $\theta = (\theta_1,\cdots,\theta_p)$. Caso após o cálculo da distribuição a posteriori de θ identificarmos que alguns elementos de θ são termos de perturbação, é possível atenuar o efeito causado pelos mesmos a partir do cálculo da distribuição a posteriori marginal dos parâmetros de interesse. Com base em Campagnoli, Petrone e Petris (2009), para $p = 2$ podemos marginalizar a posteriori $\pi(\theta_1,\theta_2|y)$ e calcular a densidade a posteriori marginal de θ_1, denotada por

$$\pi(\theta_1|y) = \int \pi(\theta_1,\theta_2|y)d\theta_2. \tag{4.-1}$$

A escolha de uma síntese para a distribuição a posteriori pode ser enquadrada como um problema de tomada de decisão. Em termos gerais, em um problema de decisão estatística, devemos escolher uma ação de um conjunto de ações possíveis A, o qual denominamos espaço de ação, baseado na amostra y . As consequências da ação escolhida são expressas através de uma função de perda $L(\theta,a)$. Considerando um conjunto de dados y, uma regra de decisão Bayesiana seleciona a ação disponível em A a qual minimiza a perda esperada condicional, denotada por:

$$E(L(\theta,a)|y) = \int L(\theta,a)L(\theta,y)d\theta. \tag{4.-1}$$

Sob esta perspectiva, a estimativa pontual Bayesiana pode ser encarada como um problema de decisão no qual o espaço de ação é igual ao espaço composto pelos parâmetros. Diferentes funções de perda resultam em diferentes estimativas Bayesianas de θ , sendo que duas das mais utilizadas são a função de perda *quadrática*, a função de perda *linear* e a função de perda *zero ou um* (CAMPAGNOLI; PETRONE; PETRIS, 2009, p. 11).

Definimos uma função de perda quadrática como

$$L(\theta, a) = (\theta - a)^2, \tag{4.-1}$$

onde θ é um escalar. Consequentemente, para o caso de uma distribuição a posteriori, o valor esperado da perda é dado por $E((\theta - a)^2|y)$, sendo minimizado no ponto $a = E(\theta|y)$. Este resultado implica que a estimativa Bayesiana de θ com função de perda quadrada é dada pelo valor esperado à posteriori de θ.

A função de perda linear é denotada por:

$$L(\theta, a) = \begin{cases} c_1 & |a - \theta| \quad \text{se } a \leq \theta \\ c_2 & |a - \theta| \quad \text{se } a > \theta \end{cases}$$

onde θ é um escalar e c_1 e c_2 são constantes positivas. Nesta configuração, a estimativa Bayesiana é o quartil $\frac{c_1}{(c_1+c_2)}$ da distribuição a posteriori. Para o caso especial onde $c_1 = c_2$, a estimativa Bayesiana é a mediana da posteriori.

Por fim, definimos a função de perda *zero ou um* como:

$$L(\theta, a) = \begin{cases} c & \text{se } a \neq \theta \\ 0 & \text{se } a = \theta \end{cases}$$

de forma que a estimativa Bayesiana é a moda da distribuição a posteriori.

Um aspecto importante na abordagem da econometria Bayesiana é a escolha das distribuições a priori, pois é através delas que introduzimos a nossa percepção na análise, para além dos dados existentes. A tese é que o conhecimento prévio do pesquisador sobre algum fenômeno que está sendo estudado é relevante, e na abordagem Bayesiana é possível introduzir explicitamente toda a informação que possuímos no processo de inferência.

Iniciando por algumas definições básicas, a escolha da distribuição a priori, ou simplesmente priori, é a escolha de um par $\pi(y|theta)$ e θ . A escolha de $\pi(y|theta)$ é denominada a *especifcação do modelo* e conjuntamente à especificação de $\pi(y|\theta)$ representa parte das escolhas subjetivas que devemos fazer para estudar um fenômeno, em uma abordagem Bayesiana.

Por questões de tratabilidade matemática e computacional, uma boa prática é a utilização de *prioris* conjugadas; onde denotamos por conjugada ao modelo $\pi(y|\theta)$, as densidades de θ cuja priori e posteriori fazem parte da mesma família de distribuições.

Como exemplo, caso utilizemos uma distribuição priori Gaussiana $N(m_0, C_0)$ e o resultado da posteriori seja também Gaussiano, com parâmetros $N(m_n, C_n)$ dizemos que a família Gaussiana é conjugada ao modelo $\pi(y|\theta) = N(y; \theta, \sigma^2)$.

Os conceitos introduzidos até o momento podem parecer um tanto quanto abstratos. Com o objetivo de exemplificar a aplicação desta abordagem, detalhamos a utilização da inferência Bayesiana no contexto de uma regressão linear. De acordo com Campagnoli, Petrone e Petris (2009), os modelos lineares dinâmicos (DLMs) podem ser interpretados como uma generalização do modelo de regressão linear padrão, permitindo que os coeficientes da regressão assumam valores distintos com o passar do tempo.

Definimos o modelo de regressão linear padrão como

$$Y_t = x_t'\beta + \varepsilon_t, \qquad t = 1, \cdots, n, \qquad \varepsilon_t \sim N(0, \sigma^2), \tag{4.-1}$$

onde Y_t é uma variável aleatória e os termos x_t e β são vetores de dimensão p. Em um modelo de regressão linear padrão, tratamos as variáveis x como determinísticas ou exógenas. Já em uma regressão estocástica, tratamos as variáveis x como variáveis aleatórias. No caso da regressão estocástica, possuímos um vetor aleatório (Y_t, X_t), de dimensão $(p+1)$. Precisamos então especificar uma distribuição conjunta e derivar o modelo de regressão linear a partir da mesma.

Por conveniência, assumindo que tal distribuição conjunta seja Gaussiana, obtemos que:

$$\begin{bmatrix} Y_t \\ X_t \end{bmatrix} \Big| \, \mu, \qquad \Sigma \sim N(\mu, \Sigma) \tag{4.-1}$$

onde

$$\mu = \begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix} \qquad \text{e} \qquad \Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma xx \end{bmatrix}. \tag{4.-1}$$

Decompondo a distribuição conjunta em um modelo marginal para X_t e um modelo condicional para Y_t, dado $X_t = x_t$, obtemos

$$X_t|\mu, \qquad \Sigma \sim N(\mu_x, \Sigma_{xx}) \tag{4.0}$$

$$Y_t|x_t, \quad \Sigma \sim N(\beta_1 + x_t'\beta_2, \sigma^2) \tag{4.1}$$

onde

$$\beta_2 = \Sigma_{xx}^{-1}\Sigma_{xy} \tag{4.2}$$

$$\beta_1 = \mu_y - \mu_x'\beta_2 \tag{4.3}$$

$$\sigma^2 = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}. \tag{4.4}$$

Se a distribuição a priori em (μ,Σ) é tal que os parâmetros do modelo marginal e aqueles do modelo condicional são independentes e, caso nosso foco de interesse seja a variável Y, podemos restringir as análises ao modelo condicional; neste caso o modelo de regressão descreve a distribuição condicional de Y_t dados (β,Σ) e x_t.

Finalmente, obtemos que

$$Y|X,\beta,V \sim N_n(X\beta,V), \tag{4.4}$$

onde $Y = (Y_1,\cdots,Y_n)$ e X é uma matriz de dimensão $n\,x\,p$. Assumindo que os termos Y_t sejam condicionalmente independentes, com mesma variância σ^2, obtemos uma matriz de covariância diagonal $V = \sigma^2 I_n$.

A inferência Bayesiana pode ser aplicada em situações distintas. Para a situação onde a matriz de covariância V é conhecida e queremos realizar inferências sobre os coeficientes de regressão β dado o conjunto de dados y, podemos obter a distribuição a priori para β a partir da utilização da verossimilhança como função de β.

A verossimilhança para o modelo definido pela equação 3.41 é dada por

$$\pi(y|\beta,V,X) = (2\pi)^{-n/2}|V|^{-1/2}exp\{-(1/2)(y-X\beta)'V^{-1}(y-XB)\} \tag{4.5}$$

$$\propto |V|^{-1/2}exp\{-(1/2)(y'V^{-1}y - 2\beta'X'V^{-1}y + \beta'X'V^{-1}X\beta)\} \tag{4.6}$$

e considerando o caso onde $\beta \sim N_p(m,C)$, obtemos

$$\pi(\beta) \propto exp\{-(1/2)(\beta-m)'C^{-1}(\beta-m)\} \tag{4.7}$$

$$\propto exp\{-(1/2)(-\beta'C^{-1} - 2\beta'C^{-1}m)\} \tag{4.8}$$

de forma que a verossimilhança, enquanto função de β é proporcional à densidade Gaussiana multivariada, com média $(X'V^{-1}X)^{-1}X'V^{-1}y$ e variância $(X'V^{-1}X)^{-1}$. Com uma distribuição a priori conjugada Gaussiana, a distribuição a posteriori também será Gaussiana, porém com parâmetros atualizados.

Com o objetivo de derivar a expressão para os parâmetros da distribuição a posteriori, podemos computar a densidade posterior com base na fórmula de Bayes

$$\pi(\beta|Y,X,V) \propto exp\{-(1/2)(\beta'(X'V^{-1}X+C_0^{-1})\beta - 2\beta'(X'V^{-1}y+C_0^{-1}m_0))\}, \tag{4.8}$$

a qual reconhecemos ser o *kernel* de uma densidade Gaussiana de p variáveis com parâmetros

$$\begin{aligned} m_n &= C_n(X'V^{-1}y+C_0^{-1}\mu_0) && (4.9)\\ C_n &= (C_0^{-1}+X'V^{-1}X)^{-1}. && (4.10) \end{aligned}$$

Percebemos então que a estimativa Bayesiana pontual de β, quando utilizamos uma função de perda quadrática é igual ao valor esperado da posteriori $E(\beta|X,y) = m_n$ e que m_n pode ser representa como uma combinação matricial ponderada entre a suposição a priori m_0, com peso proporcional a matriz a priori de precisão, denotada por C_0^{-1}, e a estimativa $\hat{\beta}$, cujo peso é proporcional à matriz de precisão $X'V^{-1}X$, de β.

Em diversos casos, é impossível calcular analiticamente a média ou variância de distribuições a posteriori. Uma alternativa é recorrer à métodos de simulação computacional, que são parte fundamental da abordagem Bayesiana. Considerando o caso onde temos os parâmetros $\psi_1,\cdots,\psi_n$ em uma distribuição a posteirori π, podemos aproximar a média de qualquer função $g(\psi)$, com expectativa finita da posteriori, pela média amostral denotada por

$$E_\pi(g(\psi)) \approx N^{-1}\sum_{j=1}^{N} g(\psi_j). \tag{4.10}$$

O problema é que, por vezes, não é possível obter amostras independentes de distribuições a posteriori. No entanto, para alguns tipos de cadeias de Markov, tal aproximação é válida. Quando aplicamos métodos de Monte Carlo para a simulação de variáveis aleatórias oriundas de uma cadeia de Markov, chegamos aos denominados métodos de Markov chain Monte Carlo; os quais representam parte fundamental da análise numérica necessária no contexto de análise Bayesiana de dados.

Quando temos uma cadeia de Markov $(\psi_t)_{t\geq 1}$ aperiódica e recorrente, com distribuição invariante π, de acordo com Campagnoli, Petrone e Petris (2009), é possível mostrar que para cada valor inicial ψ_1 a distribuição de ψ_t tende a π a medida que $t\to\infty$. Isto implica que para um valor de M suficientemente grande, os parâmetros $\psi_{M+1},\cdots,\psi_{M+N}$ são distribuídos de acordo com π e, conjuntamente, possuem propriedades similares àquelas que possuem as amostras independentes de π.

Ademais, quando os parâmetros ψ_j são simulados a partir de uma cadeia de Markov, a abordagem padrão para a estimativa da variância amostral a partir da média não é válida. A alternativa utilizada é, quando N é suficientemente grande, considerar

$$Var(\overline{g}_N) \approx N^{-1} Var(g(\psi_1))\tau(g) \tag{4.10}$$

onde

$$\tau(g) = \sum_{t=-\infty}^{\infty} \rho_t \tag{4.11}$$

$$\rho_t = corr(g(\psi_s), g(\psi_{s+t})) \tag{4.12}$$

e a estimativa de $Var(g(\psi_1))$ é dada pela variância amostral de $g(\psi_1), \cdots, g(\psi_N)$.

4.2 Modelos Lineares Dinâmicos e a Representação em Espaço de Estado

Modelos de Espaço de Estado consideram uma série temporal como resultado de um sistema dinâmico perturbado por perturbações aleatórias. Nesta abordagem, interpretamos a série temporal como uma combinação de componentes de tendência, sazonais ou autoregressivos e a representamos através uma estrutura probabilística.

Para problemas envolvendo estimativa e previões, a abordagem de Espaço de Estado envolve a computação recursiva da distribuição condicional das quantidades de interesse, dada a informação disponível. Uma vantagem dos modelos de Espaço de Estado é que é possível utilizá-los para modelar tanto séries univariadas quanto multivariadas, permitindo a presença de não estacionariedade, mudanças estruturais e padrões irregulares.

Quando estamos tratando de séries não estacionárias, como por exemplo a série de preços de uma ação, caso utilizemos a abordagem padrão de um modelo ARMA, é necessário diferenciar a série para que a mesma se torne estacionária. Apesar de simples, tal procedimento envolve a perda de informações importantes sobre a série, principalmente quando utilizada em um contexto de não isolamento; como por exemplo a construção de um portfólio. Uma das vantagens dos modelos de Espaço de Estado é que os mesmos envolvem modelos ARMA como casos especias, e podemos construir modelos que nos permitam analisar de forma mais direta dados que apresentem instabilidade nos valores da média e variância, sem a necessidade de transformações preliminares.

Com base em Campagnoli, Petrone e Petris (2009) consideremos uma série temporal $(Y_t)_{t\geq 1}$. A especificação das distribuições conjuntas dos termos $(Y_1, \cdots, Y_t)$ não é uma tarefa

simples, especialmente quando tratamos de séries temporais cujas estruturas de dependência não são claras. A dependencia Markoviana é uma das estruturas de dependência mais simples e, retomando algumas definições da seção de processos estocásticos, dizemos que uma série $(Y_t)_{t\geq 1}$. é uma cadeia de Markov caso

$$\pi(y_t|y_{1:t-1}) = \pi(y_t|y_{t-1}), \tag{4.12}$$

para qualquer valor de t. A equação 4.43 implica que a observação y_{t-1} carrega toda a informação contida em toda a série, até o período $t-1$ e que os termos Y_t e $Y_{1:t-2}$ são condicionalmente independentes dado y_{t-1}.

Definimos então as distribuições conjuntas para uma cadeia de Markov como:

$$\pi(y_{1:t}) = \pi(y_1)\Pi_{j=2}^{t}\pi(y_j|y_{j-1}). \tag{4.12}$$

Os modelos de Espaço de Estado utilizam a estrutura de dependência das cadeias de Markov para construir modelos mais complexos. Para construir um modelo de Espaço de Estado, assumimos que há uma cadeia de Markov não observável θ_t, a qual denominamos processo de estado, e que os termos Y_t representam uma mensuração imprecisa de θ_t.

Em econometria, θ_t é usualmente uma construção latente, e pensamos nos termos θ_t como séries temporais auxiliares as quais facilitam a tarefa de especificar a distribuição de probabilidade da série temporal observável Y_t. Um modelo de Espaço de Estado consiste de séries $\theta_t : t = 0, 1, \cdots$ e séries $Y_t : t = 1, 2 \cdots$, onde θ_t é uma cadeia de Markov e, condicionalmente à θ_t, os termos Y_t são independentes, onde cada termo Y_t depende apenas de θ_t.

De posse destas definições, podemos afimar que um modelo de Espaço de Estado é completamente especificado pela sua distribuição inicial $\pi(\theta_0)$ e por suas densidades condicionais $\pi(\theta_t|\theta_{t-1})$ e $\pi(y_t|\theta_t)$. Finalmente, para qualquer período $t > 0$, temos que

$$\pi(\theta_{0:t}, y_{1:t}) = \pi(\theta_0)\Pi_{j=1}^{t}\pi(\theta_j|\theta_{j-1})\pi(y_j|\theta_j). \tag{4.12}$$

Uma das classes mais importantes de modelos de Espaço de Estado é a classe de Modelos de Espaço de Estado Lineares Gaussianos, também denominados *modelos lineares dinâmicos*. Os modelos lineares dinâmicos, ou simplesmente DLM, são especificados por uma distribuição a priori Normal para o vetor de estados de dimensão p no tempo $t = 0$

$$\theta_0 \sim N_p(m_0, C_0), \tag{4.12}$$

conjuntamente a um par de equações para cada $t \geq 1$:

$$Y_t = F_t\theta_t + \upsilon_t, \qquad \upsilon_t \sim N_m(0, V_t) \tag{4.13}$$

$$\theta_t = G_t\theta_{t-1} + w_t, \qquad w_t \sim N_p(0, W_t) \tag{4.14}$$

onde G_t é uma matriz de ordem (pxp) e F_t é uma matriz de ordem (mxp) e os termos υ_t *e* w_t são sequências de vetores aleatórios Gaussianos com média zero e matrizes de covariância dadas por V_t *e* W_t, respectivamente.

Denominamos 4.13 como a *equação de observação* e 4.14 como *equação de estado* e assumimos que θ_0 é independente de υ_t *e* w_t .

De forma geral, definimos um modelo de Espaço de Estado por uma distribuição a priori para θ_0, conjuntamente as equações de observação e evolução, denotadas por

$$Y_t = h_t(\theta_t, \upsilon_t) \tag{4.15}$$

$$\theta_t = g_t(\theta_{t-1}, w_t). \tag{4.16}$$

Os modelos de Espaço de Estado Lineares especificam as funções h_t *e* g_t como funcões lineares e os modelos Lineares Gaussianos adicionam a estes as premissas de distribuições Gaussianas.

Uma aplicação útil de modelos de espaço de estado é a generalização de modelos de regressão que permitem que os parâmetros variem no tempo. Ademais, dizemos que um sistema é observável se for possível inferir os valores de todos os componentes do vetor de estados a partir de observações ruidosas. Desta forma, se o sistema for observável, poderemos distinguir entre previsão, filtragem e suavização.

Segundo Metcalfe e Cowpertwait (2009, p. 231), a previsão é dada pela previsão de valores de estado futuros, a filtragem está relacionada a obtermos a melhor estimativa dos valores atuais a partir das observações, incluido a observação atual e a suavização diz respeito à fazer a melhor estimativa dos valores passados de estado, dado um histórico de observações. Abaixo, definimos formalmente os conceitos de previsão, filtragem e suavização.

Com base em Metcalfe e Cowpertwait (2009), consideremos D_t o conjunto de dados observados até o período t. Podemos expressar D_t como a combinação entre dados observados até o período $t-1$ e a observação realizada no período t, de forma que $D_t = (D_{t-1}, y_t)$. Sob esta perspectiva, obtemos, a partir do teorema de Bayes

$$p(\theta_t|D_{t-1}, y_t) = \frac{p(y_t|\theta_t)p(\theta_t|D_{t-1})}{p(y_t)} \tag{4.16}$$

onde a densidade a posteriori do estado no período t, θ_t, dado o conjunto de dados observados até o período t, $D_t = (D_{t-1}, y_t)$ é proporcional ao produto entre densidade de probabilidade da observação no período t, dado o estado no período t e a densidade a priori do estado no período t, dado o conjunto de observações até o período $t-1$.

Para o caso univariado, a média da distribuição a posteriori é uma média ponderada da média da distribuição a priori e das observações, com pesos proporcionais a precisão de cada uma das partes; onde definimos a precisão a partir da medida da variância. Ademais, a precisão da distribuição a posteriori é igual a soma das precisões da distribuição a priori e a precisão das observações. A extensão deste resultado para uma distribuição multivariada normal nos leva ao resultado

$$\theta_t | D_t \; N(m_t, C_t), \tag{4.16}$$

5 A Abordagem de Aprendizado Estatístico

5.1 Seleção de Modelos

Nesta seção, tratamos de alguns métodos de seleção de modelos. Tais métodos representam melhorias aos modelos lineares usuais e, quando utilizados corretamente, podem apresentar ganhos consideráveis em relação à capacidade preditiva e interpretabilidade dos modelos. Tratamos especificamente de dois destes métodos, a *seleção de subconjunto* e os *métodos de encolhimento*.

De acordo com James et al. (2013), a abordagem de seleção de subconjunto envolve a identificação de um subconjunto dos p preditores, o qual acreditamos conter os preditores relacionados com a reposta que estamos tentando modelar. Após a seleção do subconjunto, aplicamos o método de mínimos quadrados para estimar o modelo constituído apenas pelo grupo reduzido de variáveis selecionadas. A seleção de subconjunto pode ser aplicada de formas distintas.

A abordagem de *seleção do melhor subconjunto* é aplicada a partir da estimativa de uma regressão por mínimos quadrados para cada uma das possíveis combinações dos p preditores considerados. Isto significa que é feita a estimativa de todos os p modelos que contém exatamente um preditor, todos os $\binom{p}{2} = \frac{p(1-p)}{2}$ modelos que contém 2 preditores e assim sucessivamente. Posteriormente avaliamos os resultados de todos eles, identificando o melhor a partir de um determinado critério de informação ou erro de previsão em validação cruzada.

A escolha do "melhor"modelo dentre os $M_0, \cdots, M_p$ modelos disponíveis não é uma tarefa trivial. Devemos notar que a medida que inserimos novas variáveis no modelo o R^2 aumenta monotonicamente. Desta forma, utilizando somente este critério escolheríamos sempre o modelo que envolve todas as variáveis.

Outro problema associado à abordagem de seleção do melhor conjunto é a grande demanda de processamento computacional. À medida que a quantidade de variáveis p aumenta, a quantidade de modelos possíveis 2^p aumenta exponencialmente, tornando inviável a avaliação de todos os resultados possíveis. Há, no entanto, outras abordagens que se propõem a resolver a mesma questão de forma mais eficiente, sob a perspectiva de processamento computacional.

Quando p é grande e não podemos aplicar a seleção do melhor subconjunto diretamente, seja por questões computacionais ou pelo risco de sobreajuste, uma alternativa é a *seleção passo a passo*, que explora uma quantidade restrita de modelos dentre todos os possíveis. A seleção passo a passo pode ser aplicada "para frente", onde iniciamos com o modelo M_0 e adicionamos

variáveis sequencialmente, ou "para trás", onde iniciamos com o modelo M_p e retiramos variáveis do modelo, sequencialmente. Apesar da diferença no sentido de aplicação, o racional é semelhante em ambos os casos.

Mais especificamente, na seleção passo a passo para frente, iniciamos com o modelo nulo M_0, o qual não contém nenhum preditor. Para cada passo $k = 0, \cdots, p-1$, consideramos todos os $p-k$ modelos que aumentam M_k com um preditor adicional e escolhemos o "melhor"dentre tais modelos, denotando-o por M_{k+1}; nesta abordagem, o melhor modelo é aquele que apresenta a menor SQR ou maior R^2.

Após a identificação do melhor modelo em cada um dos passos, selecionamos o melhor dentre os $M_0, \cdots, M_p$ utilizando novamente validação cruzada ou algum critério de informação como AIC, BIC ou C_p ou R^2 ajustado, os quais são detalhados ao fim desta seção.

Quanto aos aspectos computacionais, diferentemente da seleção do melhor modelo, a qual envolve a estimação de 2^p modelos, na seleção passo a passo a frente estimamos

$$\sum_{k=0}^{p-1} p-k = \frac{1+p(p+1)}{2} \tag{5.0}$$

modelos.

Uma outra alternativa é a *seleção passo a passo para trás*, onde começamos com o modelo completo M_p, o qual contém todos os p preditores e, para cada um dos passos $k = p, p-1, \cdots, 1$, consideramos todos os k modelos que contém um preditor a menos do que M_k, escolhendo o melhor dentre os k possíveis modelos a partir dos mesmos critérios utilizados na seleção passo a passo a frente. É importante notar que a seleção passo a passo para trás só pode ser utilizada quando $n > p$, ao passo que a seleção passo a passo a frente pode ser utilizada mesmo quando $n < p$.

Ambas as abordagens de seleção passo a passo podem chegar a modelos que não sejam os ideias, dentre todos os possíveis, principalmente por conta da natureza sequencial da metodologia, que acaba sempre escolhendo a melhor opção em cada um dos passo. Neste sentido, tais abordagens podem ser consideradas gananciosas, tendo um paralelo com os métodos de árvores detalhados na próxima seção. Ao tratar dos métodos de seleção, por vezes nos referimos ao que seria o "melhor"modelo dentre os disponíveis.

A abordagem padrão seria escolher o modelo com menor valor da soma do quadrado dos resíduos [1], ou simplesmente SQR, ou maior valor para a estatística R^2, no entanto, utilizando somente estas métricas sempre escolheríamos o modelo com todas as variáveis.

[1] para um detalhamento maior sobre a soma do quadrado dos resíduos e o R quadrado, consultar Wooldridge (2015)

Outro problema é que ambos R^2 e SQR são métricas relacionadas ao erro de treino. No entanto, devemos escolher modelos que apresentem bons resultados em uma base de teste, cujos valores sejam desconhecidos ao modelo, o que denominamos como fora da amostra. Por conta destes fatores o R^2 e a SQR não são as métricas mais adequadas para a seleção do melhor modelo.

Para selecionar o melhor modelo, considerando os resultados em relação a uma base de teste, precisamos estimar este erro de teste. James et al. (2013) destacam que há duas abordagens comuns. Uma delas é estimar o erro de teste indiretamente, realizando um ajuste ao erro de treino que leve em consideração o viés decorrente de sobreajuste. A outra é estimar o erro de teste diretamente, utilizando uma base de dados para validação, ou ainda uma abordagem de validação cruzada.

Em geral, o erro médio quadrado [2], ou simplesmente EMQ, de treino é uma subestimativa do EMQ de teste. James et al. (2013) sugerem a utlização de quatro métricas para o ajuste do EMQ de treino.

A estatística C_p adiciona uma penalidade de magnitude $2d\hat{\sigma}^2$ à SQR de treino para realizar o ajuste, e é denotado por

$$C_p = \frac{1}{n}(SQR + 2d\hat{\sigma}^2), \tag{5.0}$$

onde $\hat{\sigma}^2$ é uma estimativa da variância do erro de previsão ε, e é usualmente calculada utilizando o modelo completo, o qual contém todas os preditores.

O critério de informação de Aikaike, ou simplesmente AIC, é definido para uma grande classe de modelos estimados por máxima verossimilhança [3]. Para modelos com erros Gaussianos, onde a estimativa por máxima verossimilhança é equivalente à estimativa de mínimos quadrados o AIC é denotado por

$$AIC = \frac{1}{n\hat{\sigma}^2}(SQR + 2d\hat{\sigma}^2). \tag{5.0}$$

O critério de informação Bayesiano em geral penaliza mais fortemente modelos com várias variáveis, resultando em modelos "menores"do que os obtidos a partir da estatística C_p, por exemplo. Denotamos o BIC, para o caso de um modelo estimado por mínimos quadrados com d preditores, por

$$BIC = \frac{1}{n\hat{\sigma}^2}(SQR + log(n)d\hat{\sigma}^2). \tag{5.0}$$

[2] para um detalhamento maior sobre o erro médio quadrado, consultar Wooldridge (2015)
[3] para um detalhamento maior sobre a estimativa por máxima verossimilhança, consultar Wooldridge (2015)

A última métrica apresentada é o R^2 ajustado. O R^2 padrão é defindo como

$$R^2 = 1 - SQR/SQT, \tag{5.0}$$

onde

$$SQT = \sum (y_i - \bar{y})^2. \tag{5.0}$$

Como a SQR diminui a medida que novos preditores são adicionados, o R^2 aumenta. Para um modelo estimado por mínimos quadrados com d preditores, o R^2 ajustado é denotado por

$$R^{2*} = 1 - \frac{SQR/(n-d-1)}{SQT/(n-1)}. \tag{5.0}$$

De acordo com James et al. (2013), a intuição por trás do R^2 ajustado é que uma vez que todas as variáveis corretas tenham sido inclusas no modelo, a adição de variáveis de ruído levarão apenas a uma pequena diminuição no nível da SQR. Como a adição de uma nova variável leva a um aumento na quantidade de preditores d , o resultado será um aumento em $SQR/(n-d-1)$, consequentemente diminuindo o valor do R^2 ajustado do modelo.

5.2 Métodos de Encolhimento

Uma alternativa aos métodos de seleção de modelos abordados na seção anterior é a utilização de métodos de encolhimento, onde ao invés de tratar de seleção de modelos, tratamos da seleção de variáveis. Isto é possível graças a utilização de funções objetivo distintas das utilizadas no método de mínimos quadrados.

Uma característica comum aos métodos de encolhimento é a inclusão de termos de penalização. Tais termos de penalização podem dizer respeito tanto à magnitude dos coeficientes estimados quanto à quantidade de coeficientes que assumem valores diferentes de zero. Tais métodos são valiosos pois possibilitam que boa parte do viés relacionado à escolha de variáveis seja eliminado, possibilitando ao pesquisador a possibilidade de responder questões de pesquisa sem a obrigatoriedade de selecionar as variáveis que serão utilizadas na modelagem a priori.

Nesta seção, tratamos dos dois principais métodos de encolhimento: a regressão Ridge e o LASSO. Ambas representam abordagens semelhantes enquanto conceito, porém consideravelmente distintas enquanto forma de implementação. Um entendimento maduro de ambas as abordagens possibilita tratar da maior parte dos métodos de encolhimento utilizados para

regressão, pois boa parte destes métodos são derivados da regressão LASSO e Ridge, ou ainda, podem ser uma combinação dos dois métodos.

Com base em (JAMES et al., 2013), o procedimento de mínimos quadrados para a estimativa dos coeficientes $\beta_0, \beta_1, \cdots, \beta_p$ escolhe os valores que minimizam a soma dos erros quadrados, denotada por

$$SQR = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 . \tag{5.0}$$

Na regressão Ridge, estimamos os coeficientes com base na minimização de uma função levemente diferente. Particularmente, os coeficientes de uma regressão Ridge são aqueles que minimizam a função

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{5.0}$$

onde λ é um parâmetro de regulagem do modelo, definido pelo pesquisador.

A primeira parte da equação é simplesmente a SQR, e a segunda parte é o que denominamos *penalidade de encolhimento*. O parâmetro λ é utilizado para controlar o impacto da penalidade à magnitude dos coeficientes. É de se esperar que valores baixos de λ representem uma penalidade mais baixa à magnitude dos coeficientes e que valores elevados de λ representem uma forte penalidade, forçando coeficientes a assumir valores mais baixos, ou até mesmo zero.

Desta forma, quando $\lambda = 0$ o termo de penalidade não possui nenhum efeito e obtemos estimativas de mínimos quadrados dos coeficientes e quando $\lambda \to \infty$ os coeficientes Ridge tendem à zero e obtemos uma equação contendo apenas o intercepto β_0. É importante notar que a magnitude dos coeficientes passa a ser relevante, no contexto de uma regressão Ridge. Por conta disso, é recomendado que as variáveis sejam padronizadas, eliminando possíveis efeitos causados por pela utilização de unidades de medida distintas (JAMES et al., 2013).

Padronizamos as variáveis a partir da função

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}} \tag{5.0}$$

e desta forma, representamos todas as variáveis em uma mesma escala. Uma grande vantagem da utilização da regressão Ridge frente à estimativa de mínimos quadrados fica evidente em situações onde a quantidade de variáveis independentes de um modelo é próxima à quantidade de observações, especialmente quando $p > n$.

Quando a relação entre a variável dependente e as variáveis independentes é próxima a uma relação linear, a estimativa de mínimos quadrados dos coeficientes em geral apresentará um baixo viés e uma variância elevada. Quando temos valores de p próximos à n, as estimativas de mínimos quadrados dos coeficientes tendem a apresentar graus elevados de variação, de forma que pequenas mudanças nos dados de treino levam a mudanças consideráveis nas estimativas. Para o caso particular onde $p > n$, não é nem possível realizar a estimativa dos coeficientes a partir do métodos de mínimos quadrados,ao passo que na regressão Ridge, é possível assumir um pequeno aumento de viés a partir da eliminação de quedas consideráveis na quantidade de variância. Em suma, a regressão Ridge funciona melhor em situações onde as estimativas de mínimos quadrados apresentam variância elevada.

Uma desvantagem associada a regressão Ridge reside no fato de que o coeficiente de penalização atua sobre todas as variáveis consideradas no modelo, causando o efeito de encolhimento sobre o valor dos coeficientes. Colocando em perspectiva os métodos de seleção de modelos apresentados na seção anterior, a regressao Ridge estaria sempre tratando do modelo que inclui todas as variáveis M_p, o qual sabemos nem sempre ser o mais eficiente sob as métricas apresentadas.

Portanto, seria interessante se fosse possível, além de penalizar a magnitude dos coeficientes, penalizar a quantidade de coeficientes que assumirão valores diferentes de zero. Tal solução existe, sendo denominada LASSO, do inglês Least Absolute Shrinkage and Selection Operator, e pode ser interpretada como uma forma alternativa de seleção de modelos, particularmente a partir da seleção direta de variáveis cujos coeficientes serão diferentes de zero.

Diferentemente da regressão Ridge, a abordagem LASSO minimiza a função

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|, \tag{5.0}$$

a qual penaliza o valor absoluto dos coeficientes, tornando alguns deles exatamente iguais à zero.

A principal diferença entre a regressão Ridge e o LASSO está no termo de penalização. Enquanto na regressão Ridge utilizamos uma penalização $l_2 = ||\beta||_2 = \sqrt{\sum_{j=1}^{p}\beta_j^2}$, no LASSO utilizamos uma penalização $l_1 = ||\beta||_1 = \sum|\beta_j.|$

É possível formular o LASSO e a regressão Ridge de forma que a solução envolva a minimização de uma função, sujeita a uma restrição. Nesta configuração, fica claro que podemos encontrar uma solução para cada um dos valores de λ. Formalmente, estamos interessados em minimizar a função

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2, \tag{5.0}$$

sujeita à restrição

$$\sum_{j=1}^{p} |\beta_j| \leq s, \tag{5.0}$$

para o caso do LASSO, e sujeita à

$$\sum_{j=1}^{p} \beta_j^2 \leq s, \tag{5.0}$$

para a regressão Ridge.

Para o caso onde a quantidade de variáveis independentes é igual à 2, $p = 2$, os coeficientes LASSO serão aqueles que apresentem o menor valor da SQR e se encontrem na área definida pelo losango formado por $|\beta_1| + |\beta_2| \leq s$. Para a regressão Ridge, os coeficientes serão aqueles que apresentem o menor valor da SQR e se encontrem na área definida pelo círculo $\beta_{1^2} + \beta_2^2 \leq s$. No contexto do problema de otimização descrito acima, o coeficiente λ assume o papel de multiplicador de Lagrange da equação (JAMES et al., 2013).

Quanto á propriedade de seleção de variáveis, podemos dizer que o que diferencia o LASSO da regressão Ridge é o formato da região de restrição. No caso da regressão Ridge, esta região apresentará formato circular, e como um círculo não apresenta arestas, a interseção entre a elipse que representa pontos associados à um mesmo valor de SQR não ocorrerá encima de um eixo, impedindo que os coeficientes assumam um valor exatamente igual à 0. Já no LASSO, como a região de restrição apresenta arestas, a elipse da SQR geralmente fará a interseção com a região de restrição encima de um dos eixos, tornando os coeficientes de algumas das variáveis exatamente iguais à 0.

De acordo com James et al. (2013), a principal vantagem do LASSO em relação à regressão Ridge está no fato de que a primeira produz modelos mais simples e interpretáveis, os quais envolvem somente um subconjunto de todos os possíveis preditores.

Avaliando sob uma perspectiva de regressão Bayesiana, assumimos que o vetor de coeficientes β possui uma distribuição à priori $p(\beta)$. A verossimilhança das observações é denotada por $f(Y|X,\beta)$, onde $X = (X_1, \cdots, X_p)$ e, multiplicando a priori pela verossimilhança obtemos a distribuição a posteriori

$$p(\beta|X,Y) \propto f(Y|X,\beta)p(\beta|X) = f(Y|X,\beta)p(\beta). \tag{5.0}$$

Com base em James et al. (2013), consideremos o modelo linear padrão

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon, \tag{5.0}$$

onde assumimos que os erros são independentes e sujeitos a uma distribuição normal. Assumindo que a distribuição a priori do vetor de coeficientes β é dada por

$$p(\beta) = \Pi_{j=1}^{p} g(\beta_j), \tag{5.0}$$

onde g é uma função de densidade de probabilidade, percebemos que a regressão Ridge e o LASSO seguem dois casos particulares de g.

Quando g é uma distribuição de probabilidade com média zero e desvio padrão dado por uma função de λ, então o valor mais provável para β é dado pela solução da regressão Ridge. Já quando g é uma distribuição exponencial-dupla com média zero com parâmetro de escala dado por uma função de λ, , a moda da distribuição a posteriori para β é a solução LASSO.

Desta forma, sob um ponto de vista Bayesiano, a regressão Ridge e o LASSO são resultados diretos da suposição de um modelo linear com erros normais, conjuntamente a suposição de distribuições a priori específicas para β.

5.3 Árvores de Decisão, Bagging e Boosting

Os métodos de árvore de decisão podem ser aplicados tanto para regressão quanto para classificação. De acordo com James et al. (2013), os métodos baseados em árvores envolvem a *estratificação* ou segmentação do espaço preditor, representação do espaço amostral associado ao problema, em uma determinada quantidade de regiões simples. Como o conjunto das regras de estratificação do espaço preditor pode ser representado através de árvores, tais métodos são conhecidos como métodos de *árvore de decisão*.

Com relação ao processo de construção de uma árvore de decisão, James et al. (2013) propõem uma abordagem composta por duas etapas. Na primeira, dividimos o espaço preditor, o qual é o conjunto de todos os valores possíveis para $X_1, X_2, \cdots, X_p$ em J regiões distintas e não sobrepostas, denotadas por $R_1, R_2, \cdots, R_J$. Na segunda etapa, realizamos a mesma previsão para cada observação contida na região R_j. Tal previsão é simplesmente a média dos valores de resposta para cada uma das observações contidas em R_j.

Com relação ao método de divisão do espaço preditor, o método proposto por James et al. (2013) utiliza retângulos de alta dimensão, os quais denominamos *caixas*. O objetivo é encontrar caixas $R_1, \cdots, R_J$ que minimizem a soma dos quadrados dos resíduos, dada por:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2, \tag{5.0}$$

onde $\hat{y}_{R_j}$ representa a resposta média para as observações contidas na j-ésima *caixa*. Como não é possível computar toda e qualquer partição do espaço preditor em J caixas, é utilizada uma abordagem *gananciosa* conhecida como *divisão binária recursiva*. Nesta abordagem, partimos do início da árvore, onde todas as observações pertencem a uma mesma região, e sucessivamente dividimos o espaço preditor. A cada divisão, criamos dois novos *ramos*, os quais representam o pertencimento à duas novas caixas.

A característica gananciosa se deve ao fato de que a cada etapa do processo de construção da árvore, a melhor divisão é realizada considerando cada etapa em particular. A divisão leva em consideração o que é melhor para a etapa específica (minimizando ou maximizando uma função), diferentemente de uma abordagem que escolha a melhor divisão com base na expectativa de resultados melhores em epatas subsequentes.

Para performar a divisão binária inicialmente selecionamos o preditor X_j e o ponto de corte s de forma que, dividindo o espaço preditor em duas regiões $\{X|X_j < s\}$ e $\{X|X_j \geq s\}$, obtenhamos a maior redução possível na soma dos quadrados dos resíduos. Portanto, consideramos todos os preditores $X_1, \cdots, X(p)$ e todos os valores possíveis para o ponto de corte s para cada um dos preditores e escolhemos a dupla preditor/ponto de corte cuja árvore resultante possua o menor valor de SQR.

Temos então que, para quaisquer j e s , definimos o par de *meio-planos* como:

$$R_1(j,s) = \{X|X_j < s\} \text{ e } R_2(j,s) = \{X|X_j \geq s\} \tag{5.0}$$

e buscamos os valores de j e s que minimizem a equação

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2, \tag{5.0}$$

onde os termos $\hat{y}_{R_i}$ representam a resposta média para as observações de treino denotadas por $R_1(j,s)$. Cada nova divisão segmenta uma das regiões identificadas anteriormente e não todo o espaço preditor. Tal processo é repetido até que um critério de interrupção seja alcançado. Segundo James et al. (2013), usualmente este critério é definido pela quantidade mínima necessária de observações em cada região. Uma vez definidas as regiões $R_1, \cdots, R_j$, realizamos a previsão da resposta de uma dada observação de teste utilizando a média das observações de treino contidas na região a qual a observação de teste pertence.

Um ponto negativo associado ao processo descrito acima é que, apesar de produzir bons resultados na base de dados de treino, comumente ocorre sobreajuste, o que nos leva a observar resultados pobres quando aplicamos o modelo a dados de teste, fora da amostra. Dada esta situação, uma alternativa é construir árvores mais curtas, de forma que uma nova divisão só seja realizada caso a redução no nível de SQR decorrente da divisão exceda um determinado limite. O problema é que, por conta da característica *gananciosa* da abordagem de divisão binária, uma divisão sem muito impacto na redução do erro poderia ser seguida por uma outra divisão com muito impacto. Caso a exigência de se manter um nível mínimo de redução no nível de SQR seja mantida, há uma grande chance de que boas árvores sejam perdidas durante o processo.

Desta forma, uma estratégia alternativa é construir uma grande árvore, a qual denotamos por T_0 e posteriormente podá-la, suprimindo alguns ramos, com o objetivo de obter uma sub-árvore. Idealmente, devemos selecionar a sub-árvore que nos leve à menor taxa de erro durante os período de testes e, podemos estimar o erro de teste para cada sub-árvore a partir de validação cruzada. No entanto, em algumas aplicações, estimar o erro de validação cruzada para toda e qualquer subárvore é impossível, por conta do custo computacional associado a tal tarefa, dada a possibilidade de existência de uma quantidade extremamente elevada de possíveis sub-árvores.

Uma alternativa para selecionar um pequeno subconjunto de árvores a partir de todo o conjunto é a utilização de *"Poda por Custo de Complexidade"*, ou simplesmente PCC, onde, ao invés de considerarmos todas as sub-árvores possíveis, consideramos uma sequência de árvores indexada por um parâmetro de regulagem não negativo denotado por α. De forma simplificada, o procedimento implica na utilização de divisão binária para *crescer* uma grande árvore com base nos dados de treino, interrompendo o crescimento somente quando cada nó terminal possuir menos observações do que um valor limite preestabelecido. Em sequência, aplicamos PCC à grande árvore de forma a obtermos a melhor sequência de sub-árvores e utilizamos validação cruzada *dobra-K* para a escolha do parâmetro α. Desta forma, dividimos as observações de treino em K dobras e, para cada $k = 1, \cdots, K$, repetimos o procedimento de divisão binária e avaliação da média quadrada dos erros de previsão, como função de α. Finalmente, tiramos a média dos erros para cada um dos valores de α e escolhemos o valor de $alpha$ que minimize o erro médio, retornando a sub-árvore associada a este erro médio.

Cada valor de α possui então uma sub-árvore correspondente, de forma que:

$$\sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T| \tag{5.0}$$

seja tão pequeno quanto possível. De acordo com a notação utilizada por James et al. (2013), $|T|$ representa a quantidade de nós terminais da árvore T, R_m é o subconjunto do espaço preditor correspondente ao m-ésimo nó terminal e $\hat{y}_{R_m}$ é a resposta prevista associada à R_m. O parâmetro $alpha$ controla o *trade-off* entre a complexidade das sub-árvores e seu nível de

encaixe aos dados de treino. Quando $\alpha = 0$ a sub-árvore T será igual à árvore T_0 e a medida que o valor de α aumenta, passamos a observar uma penalização para a quantidade de nós em cada sub-árvore e obtemos subárvores menores como solução para a minimização da equação 5.21.

A abordagem de árvores pode ser utilizada tanto para problemas de regressão quanto de classificação. A principal diferença entre uma*árvore de classificação* e uma *árvore de regressão* é que na primeira, estamos interessados em prever uma resposta qualitativa, ao invés de quantitativa.

Em uma árvore de classificação, a resposta prevista para uma determinada observação, contida em uma região do espaço preditor, é dada pela resposta mais *comum* das observações pertencentes àquela região. Em um contexto de classificação, utilizamos também a divisão binária recursiva para crescer uma árvore de classificação e utilizamos a *taxa de erro de classificação* como critério para a realização das divisões binárias. A taxa de erro de classificação é definida como a fração das observações contidas em uma determinada região que não pertencem à classe mais comum. Denotamos a taxa de classificação por:

$$E = 1 - max(\hat{p}_{mk}), \tag{5.0}$$

onde $\hat{p}_{mk}$ representa a proporção de observações contidas na m-ésima região que pertencem à k-ésima classe. Os autores destacam que a taxa de erro de classificação não é suficientemente sensível para fins de crescimento das árvores e em aplicações práticas usualmente são utilizadas duas medidas alternativas, o *índice de Gini* e a *entropia*.

O índice de Gini é definido como

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{5.0}$$

e mede a variância total através das K classes. Caso os valores dos $\hat{p}_{mk}$ sejam próximos à zero ou um, obtemos um valor baixo para o índice de Gini. Por conta dessa característica, consideramos o índice de Gini como uma medida de pureza dos nós, de modo que valores baixos para o índice representam nós com predominância de observações de uma única classe.

A entropia é definida como

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} log \hat{p}_{mk}, \tag{5.0}$$

e como $\hat{p}_{mk}$ assume valores entre zero e um, a *entropia* assumirá valores próximos a zero caso os temos $\hat{p}_{mk}$ sejam próximos de zero ou um.

Portanto, tanto a entropia quanto o índice de Gini assumirão um valor baixo caso os nós sejam *puros*. De acordo com James et al. (2013), como ambos são mais sensíveis do que a taxa de erro de classificação, as duas medidas são utilizadas como critério para avaliar a qualidade de uma divisão particular, no contexto da construção de uma árvore de classificação. No entanto, de acordo com os autores, quando o objetivo é *podar* uma árvore tendo como objetivo a minimização do erro de previsão, a taxa de erro de classificação é uma medida mais adequada.

Há ainda métodos que complementam a abordagem de árvores de classificação e regressão. Tratamos agora dos métodos de bootstrap, bagging e boosting, com base em James et al. (2013).

O boostrap é uma ferramenta estatística que pode ser utilizada para quantificar a incerteza associada a um estimador associado a algum método de aprendizado estatístico. A utilidade do *boostrap* está associada ao fato de que o mesmo pode ser aplicado a uma vasta gama de métodos de aprendizado estatístico, incluindo alguns cuja medida de variabilidade é difícil de ser obtida mesmo por softwares estatísticos.

Na abordagem de bootstrap, podemos utilizar um computador para simular o processo de obtenção de novas amostras de dados, de forma que possamos estimar a variabilidade de um determinado parâmetro sem a necessidade de geração de amostras adicionais. Desta forma, ao invés de repetidamente obtermos bases de dados independentes de uma população, a abordagem implica na obtenção de bases de dados a partir da retirada de amostras da base de dados original de forma repetida.

Cosideremos o caso onde estamos interessados em medir um parâmetro denotado por α a partir de uma base de dados Z com n observações. Selecionamos aleatoriamente n observações de nossa base da dados para gerar uma base de dados *boostrap*, denotada por Z^{*1}. A amostragem é performada com repetição, o que significa que a mesma observação pode ocorrer mais de uma vez em uma base de dados *bootstrap*. Podemos então utilizar Z^{*1} para gerar uma nova estimativa bootstrap para α, a qual denominamos $\hat{\alpha}^{*1}$. Repetimos tal processo B vezes, de forma a produzirmos B diferentes bases de dados denotadas por $Z^{*1}, Z^{*2}, \cdots, Z^{*B}$, e estimativas correspondentes de α denotadas por $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \cdots, \hat{\alpha}^{*B}$. Finalmente, computamos o erro padrão destas estimativas *boostrap* utilizando a fórmula

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}'\right)^2} \tag{5.0}$$

e tal resultado pode ser utilizado como estimativa do erro padrão de $\hat{\alpha}$. (CONTINUAR DAQUI)

As árvores de decisão, em geral, sofrem do problema de grande variância. Isto significa que quando fazemos a divisão dos dados de treino em duas partes, aleatoriamente, e aplicamos um modelo de árvore de decisão a cada uma das partes, podemos obter resultados consideravelmente

diferentes. Em contraste, procedimentos com a característica de baixa variância apresentariam resultados semelhantes quando aplicados repetitivamente à bases de dados distintas. No entanto, podemos utilizar alguns procedimentos para minimar o efeito de grande variância, como a *agregação Boostrap*, também denominada *bagging*, a qual é utiizada frequentemente no contexto de árvores de decisão.

Com base em James et al. (2013), dado um conjunto de n observações independentes $Z_1, \cdots, Z_n$, cada uma com variância σ^2. A variância da média das observações Z é dada por σ^2/n. Tal resultado implica que utilizar a média de um conjunto de observações tem um efeito de redução da variância. Decorre disso, que uma forma natural de se reduzir a variância associada às previsões de um modelo de aprendizado estatístico é retirar tantas bases de treino quanto possíveis da população e construir um modelo de previsão com base em cada uma das bases de treino, retirando a média das previsões que resultaram de cada um dos modelos.

Desta forma, podemos calcular as previsões utilizando B diferentes bases de treino e utilizar a média das previsões de forma a obter um único modelo de baixa variância, o qual denotamos por

$$\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x), \tag{5.0}$$

onde os termos $\hat{f}^b(x)$ representam as previsões obtidas a partir de cada uma das bases de treino. Na prática, não temos acesso à diversas bases de treino. E para superar esta dificuldade, podemos utilizar a abordagem *bootstrap*, retirando, repetidamente, amostras de uma única base de dados de treino. Desta forma, geramos B bases de dados *bootstrapped*. Aplicamos nosso método na b-ésima base da dados de treino *bootstrapped*, obtendo as previsões $\hat{f}^b(x)$, e retirando a média de todas as previsões obtemos

$$\hat{f_{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x). \tag{5.0}$$

Um dos problemas associados à utilização de *bagging* em modelos de árvore de decisão é a diminuição da capacidade de interpretação dos resultados. Como utilizamos diversas árvores e posteriormente retiramos a média das previsões, deixa de ser possível visualizar os resultados utilizando apenas uma única árvore. Em suma, *bagging* aumenta a capacidade preditiva em detrimento da capacidade de interpretação dos resultados. Vale destacar que, ainda que não seja possível representar o modelo a partir de uma única árvore, podemos avaliar a importância das variáveis a partir do so efeito médio de cada uma na diminuição da soma dos erros quadrados.

Utilizando os procedimentos de bagging no contexto de modelos de árvore de decisão, damos origem ao que denominamos de *bagged trees*. Um problema associado às *bagged trees* é

a correlação entre as árvores construídas a partir de cada uma das amostras emphbootstrapped, pois sempre consideramos as mesmas variáveis para realizar as divisões. Uma alternativa para superar o problema de correlação entre as árvores é a utilizar amostragem aleatória das variáveis candidatas para a divisão. Nesta abordagem, quando vamos construir uma nova árvore, escolhemos apenas m dentre as p variáveis disponíveis para realizar cada divisão. Tal procedimento dá origem às emphrandom forests, as quais representam uma melhoria quando comparadas às *bagged trees*, por conta da menor correlação entre as previsões realizadas por cada uma das árvores. É importante notar que, quando as previsões são fortemente correlacionadas, a utilização da média das mesmas acaba não gerando o efeito de redução da variância esperado pela abordagem de emphbagging.

Em suma, a principal diferença entre as abordagens de *bagging* e emphrandom forests é a escolha to tamanho m do subconjunto de variáveis preditoras consideradas para cada divisão do espaço preditor. Em geral, para *random forests* $m < p$. Quando $m = p$ obtemos simplesmente a o resultado da abordagem de *bagging*. A utilização de um valor baixo de m será útil principalmente quando temos uma quantidade elevada de preditores correlacionados.

Ao passo que a abordagem de *bagging* envolve a criação de múltiplas cópias da base de dados de treino original, utilizando o *boostrap*, encaixando uma árvore de decisão para cada uma e posteriormente combinando todas as árvores para criar um único modelo preditivo, a abordagem de *boosting* utiliza alguns dos mesmos princípio, mas é aplicada de forma diferente. Na abordagem de *boosting*, construímos cada arvore utilizando informações decorrentes das árvores anteriores. Neste sentido, a principal característica da abordagem de *boosting* é a sua natureza sequencial. Diferentemente da abordagem de *boostrap*, em *boosting*, cada árvore é construída em uma versão modificada da base de dados original.

Utilizando a abordagem de *boosting*, temos o que denominado *aprendizado lento*. Tal característica é obtida pois, para cada árvore construída, construímos uma nova árvore com base nos resíduos da anterior, repetindo tal procedimento por diversas vezes. Desta forma, ao invés de construirmos cada árvore com base no resultado que desejamos prever, Y, construímos cada árvore com base nos resíduos de previsão da árvore anterior, posteriormente adicionando esta nova árvore ao modelo e atualizando as previsões. Por fim, algorítimos de *boosting* possuem três hiperparâmetros: a quantidade de árvores; o parâmetro de encolhimento λ que controla a taxa de aprendizado; e a quantidade de divisões em cada árvore, que controla a complexidade do modelo agregado e é também interpretada como a profundidade das interações entre as variáveis utilizadas para as divisões.

Referências

3RD, A. C.; JONES, H. E. Some a posteriori probabilities in stock market action. *Econometrica, Journal of the Econometric Society*, JSTOR, p. 280–294, 1937. Citado na página 69.

BACHELIER, L. *Théorie de la spéculation*. [S.l.]: Gauthier-Villars, 1900. Citado na página 68.

BAYES, T.; PRICE, R.; CANTON, J. An essay towards solving a problem in the doctrine of chances. C. Davis, Printer to the Royal Society of London London, U. K, 1763. Citado na página 107.

BOX, G.; JENKINS, G.; REINSEL, G. Time series analysis: Forecasting and control. 1994. *Princton-Hall International*, 1978. Citado na página 15.

CAMPAGNOLI, P.; PETRONE, S.; PETRIS, G. *Dynamic Linear Models with R*. [S.l.: s.n.], 2009. Citado nas páginas 107, 108, 109, 111, 112, 113, 116 e 117.

CAMPBELL, J. Y. et al. *The econometrics of financial markets*. [S.l.]: princeton University press Princeton, NJ, 1997. v. 2. Citado nas páginas 9, 19, 20, 21, 22, 67, 68, 69, 70, 72 e 79.

CARDANO, G. *The book on games of chance:(Liber de ludo aleae)*. [S.l.]: Holt, Rinehart and Winston, 1961. Citado na página 67.

COX, D.; MILLER, H. The theory of stochastic processes (methuen, london, 1965). *Google Scholar*, p. 224, 1968. Citado nas páginas 59, 62, 63, 66 e 67.

CZUBER, E.; BURKHARDT, F. *Die statistischen Forschungsmethoden*. [S.l.]: LW Seidel, 1921. v. 2. Citado na página 34.

EDWARDS, R. E. *What is the riemann integral?* [S.l.]: Dept. of Pure Mathematics, Dept. of Mathematics, Australian National University, 1974. v. 1. Citado na página 38.

ENDERS, W. *Applied econometric time series*. [S.l.]: John Wiley & Sons, 2008. Citado nas páginas 10, 11, 28, 30, 73, 74, 76, 77, 78, 79, 85, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 103, 104 e 105.

ENGLE, R. F.; GRANGER, C. W. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, JSTOR, p. 251–276, 1987. Citado na página 99.

FELLER, W. An introduction to probability theory and its applications: Volume 1. 1968. Citado nas páginas 35, 43, 44, 46, 47, 49, 50, 52, 53, 54, 55 e 56.

FERNANDEZ, P. J. *Medida e integração*. [S.l.]: IMPA, Instituto de Matemática Pura e Aplicada, CNPq, 1976. v. 2. Citado na página 37.

FIGUEIREDO, D. G. de. *Análise de Fourier e equações diferenciais parciais*. [S.l.]: Instituto de Matemática Pura e Aplicada, 2000. Citado na página 39.

GRANGER, C. W. Some properties of time series data and their use in econometric model specification. *Journal of econometrics*, North-Holland, v. 16, n. 1, p. 121–130, 1981. Citado na página 99.

HAMILTON, J. D. *Time series analysis*. [S.l.]: Princeton university press Princeton, NJ, 1994. v. 2. Citado nas páginas 80, 81, 82 e 84.

JAMES, G. et al. *An introduction to statistical learning*. [S.l.]: Springer, 2013. v. 112. Citado nas páginas 121, 123, 124, 125, 127, 128, 129, 130, 131, 132 e 133.

JOHNSON, N. L.; KOTZ, S.; BALAKRISHNAN, N. *Continuous univariate distributions, vol. 2 of wiley series in probability and mathematical statistics: applied probability and statistics*. [S.l.]: Wiley, New York,, 1995. Citado na página 13.

KENDALL, M.; STUART, A.; ORD, J. K. Classification: discrimination and clustering. *The advanced theory of statistics*, Griffin, v. 3, p. 370–421, 1983. Citado na página 11.

LAPLACE, P. S. *Théorie analytique des probabilités*. [S.l.]: Courcier, 1820. Citado na página 33.

MALKIEL, B. G.; FAMA, E. F. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, Wiley Online Library, v. 25, n. 2, p. 383–417, 1970. Citado na página 9.

METCALFE, A. V.; COWPERTWAIT, P. S. *Introductory time series with R*. [S.l.: s.n.], 2009. Citado nas páginas 11, 13, 17, 18, 22, 24, 25, 27, 30 e 119.

MEYER, P. L. Probabilidade: aplicações à estatística. In: *Probabilidade: aplicações à estatística*. [S.l.]: Livro Técnico, 1970. Citado nas páginas 35, 36, 37, 39, 44, 45, 48, 49 e 53.

MISES, R. v. Probability, statistics and truth. Macmillan, 1939. Citado nas páginas 33 e 34.

MISES, R. V. *Mathematical theory of probability and statistics*. [S.l.]: Academic Press, 1964. Citado nas páginas 34, 36, 37, 38, 39, 40 e 48.

MOOD, A. M. The distribution theory of runs. *The Annals of Mathematical Statistics*, JSTOR, v. 11, n. 4, p. 367–392, 1940. Citado na página 72.

PFAFF, B. *Analysis of integrated and cointegrated time series with R*. [S.l.]: Springer Science & Business Media, 2008. Citado nas páginas 21, 30, 93, 97, 99, 100 e 101.

POINCARÉ, H. *Calcul des probabilités*. [S.l.]: Gauthier-Villars, 1912. Citado na página 33.

RAKHSHAN, A.; PISHRO-NIK, H. Introduction to simulation using matlab. In: *Introduction to Probability, Statistics, and Random Processes: Statistics and Random Processes*. [S.l.]: Kappa Research, LLC, 2014. p. 703–723. Citado na página 17.

SEWELL, M. History of the efficient market hypothesis. *RN*, v. 11, n. 04, p. 04, 2011. Citado na página 9.

SIMON, D. *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. [S.l.]: John Wiley & Sons, 2006. Nenhuma citação no texto.

SIMS, C. A. Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, JSTOR, p. 1–48, 1980. Citado nas páginas 85, 90, 92, 93 e 95.

SIMS, C. A. et al. Are forecasting models usable for policy analysis? *Quarterly Review*, Federal Reserve Bank of Minneapolis, n. Win, p. 2–16, 1986. Citado nas páginas 94 e 95.

TSAY, R. S. *Analysis of financial time series*. [S.l.]: John Wiley & Sons, 2005. v. 543. Citado na página 9.

WICHMANN, B. A.; HILL, I. D. Algorithm as 183: An efficient and portable pseudo-random number generator. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, JSTOR, v. 31, n. 2, p. 188–190, 1982. Citado na página 13.

WOOLDRIDGE, J. M. *Introductory econometrics: A modern approach*. [S.l.]: Nelson Education, 2015. Citado nas páginas 122 e 123.