Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação

Previsão de Séries Temporais via Seleção de Variáveis, Reconstrução Dinâmica, ARMA-GARCH e Redes Neurais Artificiais

Autor: Antonio Airton Carneiro de Freitas

Orientador: Prof. Dr. Márcio Luiz de Andrade Netto

Co-orientadores: Prof. Dr. José Roberto Securato e Profa. Dra. Alessandra de Ávila Montini

Tese de Doutorado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos para obtenção do título de Doutor em Engenharia Elétrica. Área de concentração: **Engenharia de Computação**.

Banca Examinadora

Márcio Luiz de Andrade Netto, Prof. Dr. ..... DCA/FEEC/UNICAMP
Ricardo Ribeiro Gudwin, Prof. Dr. ..... DCA/FEEC/UNICAMP
Fernando José Von Zuben, Prof. Dr. ..... DCA/FEEC/UNICAMP
Celma de Oliveira Ribeiro, Profa. Dra. ..... DEP/POLI/USP
Ivan Nunes da Silva, Prof. Dr. ..... DEE/EESC/USP

Campinas, SP

Janeiro/2007

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

F884p	Freitas, Antonio Airton Carneiro de Previsão de séries temporais via seleção de variáveis, reconstrução dinâmica, ARMA-GARCH e redes neurais artificiais. /Antonio Airton Carneiro de Freitas. - - Campinas, SP: [s.n.], 2007. Orientadores: Márcio Luiz de Andrade Netto, José Roberto Securato, Alessandra de Ávila Montini. Tese (doutorado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação. 1. Previsão de séries temporais. 2. Seleção de variáveis 3. Redes neurais artificiais. 4. Econometria. 5. Câmbio. I. Andrade Netto, Márcio Luiz de. II. Securato, José Roberto. III. Montini, Alessandra de Ávila. IV. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. V. Título

Título em Inglês: Time series prediction by means of variable selection, dynamic reconstruction, ARMA-GARCH and artificial neural networks

Palavras-chave em Inglês: Time series prediction, variable selection, ARMA, GARCH, artificial neural networks, exchange rates

Área de concentração: Engenharia de Computação

Titulação: Doutor em Engenharia Elétrica e Computação

Banca examinadora: Ricardo Ribeiro Gudwin, Fernando José Von Zuben, Celma de Oliveira Ribeiro e Ivan Nunes da Silva

Data da defesa: 27/02/2007

Programa de Pós-Graduação: Engenharia Elétrica e Computação

Resumo

A inferência sobre a previsibilidade de sistemas dinâmicos não lineares multivariados tem sido freqüentemente realizada a partir de testes que podem induzir à conclusões equivocadas. Isto porque em muitas pesquisas realizadas os testes utilizados são o de autocorrelação, o da razão de variância e do espectro, que só verificam a existência ou não da correlação serial de componentes lineares. Neste trabalho, também são utilizados testes para avaliar a correlação serial de componentes não lineares. Busca-se provar empiricamente se as classes de modelos ARMA-GARCH e neurais, bem como a combinação deles, tem qualidade de previsão superior ao modelo diferença Martingale em previsões na média condicional dos retornos da taxa de câmbio brasileira e da umidade em microclima. Um método de seleção de variáveis é proposto para melhorar os resultados obtidos com modelos de previsão multivariados não baseados em teoria. As não linearidades negligenciadas durante o ajuste dos modelos neurais são avaliadas por meio do teste de Blake and Kapetanios (2003). O teste de White (2000) é utilizado para comparar os modelos de previsão propostos em conjunto com o modelo *benchmark*. Foi constatado empiricamente que os dois processos analisados não são do tipo diferença Martingale.

Palavras-chave: 1. Previsão de séries temporais. 2. Seleção de variáveis. 3. Redes neurais artificiais. 4. Econometria. 5. Câmbio.

JEL Classification: C2, C5, F3.

Abstract

The inference on predictability of nonlinear multivariate systems has been done with some possible misleading conclusions when the test statistics are insignificant because autocorrelation, variance ratio and spectrum tests check only serial uncorrelatedness (linear components). This work empirically explores the non linear components and if the ARMA-GARCH, neural network models, as well as their combination, outperform a Martingale model in the conditional mean out-of-sample forecasts. It is proposed a variable selection method to improve the results obtained with multivariate models without a priori knowledge. The neglected nonlinearities and data snooping bias were avoided applying respectively the Blake and Kapetanios (2003) and the White (2000) reality check tests. The empirical results indicate that the Brazilian exchange rates and the microclimate humidity are not Martingale differences.

Keywords: 1. Time series prediction. 2. Variable selection. 3. ARMA. 4. GARCH. 5. Artificial neural networks. 6. Exchange rates. 7. Humidity.

Non-linearity begets completeness; misjugdment creates linearity.
Lao Tzu

Aos meus pais Luiz e Maria, meus filhos Felipe e Carol, e minha esposa Ana

Agradecimentos

Ao meu orientador Prof. Dr. Márcio Luiz de Andrade Netto (FEEC-UNICAMP) sou muito grato pela oportunidade, orientação segura e eficaz, sabedoria, tratamento amável e paciência.

À minha co-orientadora, Profª. Drª. Alessandra de Ávila Montini (FEA-USP) sou muito grato pela co-orientação objetiva e eficaz, amizade e paciência.

Aos Profs. Drs. Antonio Freitas (IBMEC-RJ e FGV-RJ) e José Roberto Securato (USP e FIA) pelas oportunidades concedidas e a amizade.

Aos Profs. Drs. Fernando Von Zuben (FEEC-UNICAMP), à Profª. Drª. Vera Fava (FEA-USP) e à Doutoranda (FEEC-UNICAMP) Wanessa Gazzoni por suas contribuições nesta tese.

Ao Prof. Dr. Ivan Nunes da Silva (EESC-USP) que me orientou no mestrado de forma eficaz, paciente e amiga; e à minha irmã Maria Vilanir Carneiro de Lima que possibilitou minha iniciação na área de educação.

Aos colegas da FIA e FEA-USP pelo companheirismo, críticas e sugestões para tentar tornar o estudo aplicável ao mercado.

Aos demais colegas da FEEC-UNICAMP, EMBRAPA e UTFPR pelo apoio necessário para que este trabalho fosse concretizado.

Aos integrantes do Laboratório de Bioinformática e Computação Bio-inspirada (LBiC/UNICAMP) e ao grupo de pesquisa em sistemas inteligentes (UTFPR/CP) pela ajuda e motivação.

À minha esposa Ana e minha filha Carol pelo apoio fundamental à conclusão desta jornada; e principalmente ao meu filho Paulo Felipe pela ajuda inestimável.

A Deus.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

ARCH	-	*autoregressive conditional heteroscedasticity*
ARFIMA	-	*autoregressive fractionary integrated moving average*
ARIA	-	*adaptive radius immune algorithm*
ARIMA	-	*autoregressive integrated moving average*
ARMA	-	*autoregressive moving average*
BDS	-	teste de BROCK, DECKER E SCHEINKMAN (1996)
DM	-	diferença Martingale
EGARCH	-	*exponential generalized autoregressive conditional heteroscedasticity*
EXPAR	-	*exponential autoregressive*
FAC	-	função de autocorrelação
FACP	-	função de autocorrelação parcial
FIR	-	*finite impulse response*
GARCH	-	*generalized autoregressive conditional heteroscedasticity*
IGARCH	-	*integrated generalized autoregressive conditional heteroscedasticity*
JPD	-	janela de previsão dinâmica
JPID	-	janela de previsão inteligente dinâmica
LMS	-	*least mean square*
MAFE	-	*mean absolute forecast error*
MCFD	-	*mean correct forecast direction*
MFTR	-	*mean forecast trading return*
MIMO	-	*multiples inputs and multiples outputs*
MISO	-	*multiples inputs and single output*
MSE	-	*mean square error*
MSFE	-	*mean square forecast error*
NMSE	-	*normalized mean square error*
PA	-	passeio aleatório
PCA	-	*principal components analysis*
PMC	-	*perceptron multi-camadas*
RBF	-	*radial basis function*
RNA	-	redes neurais artificiais
TGARCH	-	*threshold generalized autoregressive conditional heteroscedasticity*
TVAR	-	*threshold* VAR

$SARIMA$	-	*sazonal autoregressive integrated moving average*
$STVAR$	-	*smooth transition* VAR
VAR	-	*vector autoregressive moving average*
$VARMA$	-	*vectors autoregressives*
VE	-	volatilidade estocástica
VEC	-	*vector error correction*
WN	-	*white noise*
WSS	-	*wide sense stationary*

Trabalhos Publicados Pelo Autor

A. A. Carneiro de Freitas, A. A. Montini and M. L. A. Netto. "Inference on predictability of brazilian exchange rates via arma-garch and neural network models". in: third brazilian conference on statistical, 2007, Maresias. Third Brazilian Conference on Statistical 2007.

A. A. Carneiro de Freitas, J. R. Securato e R. H. Rocha. "Árvores binomiais aditivas para estimar custos e benefícios da administração do risco do negócio". In: SECURATO, José Roberto. (Org.). Árvore binomial e a formação de preços de direitos contigenciais. são paulo, 2006, v. 1, p. 201-218.

A. A. Carneiro de Freitas e I. N. da Silva. "Aplicação de redes neurais na estimação da temperatura interna de transformadores de distribuição imersos em óleo". Sba Controle e Automação, Campinas-SP, p. 266-274, 01 set. 2002.

A. A. Carneiro de Freitas, J. R. Securato, and M. L. A. Netto. "Brazilian Exchange Rates Modeling: Macro and Microstructure Variable Selection by Means of the Analysis of Relevance and Redundancy". Global Finance Conference, 2006.

A. A. Carneiro de Freitas, J. R. Securato, and M. L. A. Netto. "A nonlinear methodology to predict brazilian exchange rates by means of macro and microstructures variables via neural network". ENANPAD, 2005.

A. A. Carneiro de Freitas, J. R. Securato, and M. L. A. Netto. "Nonlinear ponder autoregressive neural network methodology to predict brazilian exchange rates". IV Encontro Brasileiro de Finanças. SBFIN, 2004.

A. A. Carneiro de Freitas, J. C. Luxo, and M. L. A. Netto. "Asset-backed securitization of Brazilian companies: fuzzy and implicative statistical analysis of debt level in profitability". Global Finance Conference, 2006.

J. S. Ferranti, A. P. Chaves e A. A. Carneiro de Freitas. "Previsão da Umidade na Região de Londrina por meio da Seleção de Variáveis e de Redes Neurais Artificiais Visando o Combate à Ferrugem Asiática". SBIAGRO, 2005.

A. A. Carneiro de Freitas e I. N. da Silva. "Aplicação de redes neurais na estimação da temperatura interna de transformadores de distribuição imersos em óleo". in: congresso brasileiro de automática, 2000, florianópolis. 2000.

A. A. Carneiro de Freitas e I. N. da Silva. "Monitoring and identification of processes related to liquid immersed distribution transformers by artificial neural networks". in: 11th ifac workshop control applications of optimization, 2000, Saint Petersburg. 2000.

A. A. Carneiro de Freitas, E. R. Brinhole, J. F. Z. Destro; N. P. de Alcantara JR. "Determination of Resonant Frequencies of Triangular and Retangular Microstrip Antennas, Using Artificial Neural Networks". In: PIERS 2005, 2005, Hangzhou. Progress In Electromagnetics Research Symposium. Progress In Electromagnetics Research Symposium, 2005.

F. Pereira Junior, M. H. Santos, J. A. Martini, A. A. Carneiro de Freitas. "Seleção de Variáveis e Características como Aplicação Paralela em Cluster MPI". ERI, 2006.

J. S. Ferranti, A. P. Chaves e A. A. Carneiro de Freitas. "Applying arima and fir filters in the context of agriculture and industries". UNINDU, 2005.

Capítulo 1

Introdução

1.1 Situação problema, justificativa e relevância

Várias questões desafiadoras surgem quando se deseja implementar a previsão de uma série temporal associada a um processo cuja forma funcional é desconhecida, como por exemplo as séries da taxa de câmbio do Brasil e da umidade em microclimas. O desafio consiste em identificar uma função somente a partir de pares de amostras entradas-saídas. Neste contexto, os métodos de predição podem ser divididos em três categorias: **multivariados** - utilizam somente as informações contidas nos dados disponíveis, não se sabe qual é a forma funcional, não são traçados cenários e não existe teoria consolidada sobre o assunto; **teóricos** - baseados em teoria, projeção de cenários, como os métodos econométricos; **julgamentais** que utilizam a cognição (intuição) humana. Esta tese tem o foco especificamente nos métodos multivariados.

A classe dos modelos não lineares ficou mais popular nos últimos anos, seja porque os dados exibem não linearidades inequívocas, seja pela disponibilidade de classes de modelos não lineares bem especificados. Por exemplo, os modelos não lineares GARCH e volatilidade estocástica (VE) já são utilizados com sucesso para estimar a volatilidade condicional (risco) em séries financeiras. Como no mercado financeiro retorno e risco estão fortemente relacionados, é importante investigar métodos que incluam não linearidades na média para estimar os retornos. As redes neurais artificiais (RNA) são candidatas naturais para realizar esta tarefa. Entretanto, os modelos ARMA-GARCH ainda são os mais utilizados para explicar o comportamento dos retornos médios.

A seleção de variáveis e características é importante na modelagem de sistemas multivariados tipo MISO (*multiples inputs and single output*) e MIMO (*multiples inputs and multiples outputs*) com um número grande de variáveis candidatas. Esta abordagem está de acordo com o princípio proposto por William de Ockham: "*Pluralitas non est ponenda sine neccesitate*". Este princípio diz que se dois modelos geram resultados de previsão semelhantes, é preferível escolher o mais simples. Logo, deve-se eliminar as variáveis de entrada e os parâmetros desnecessários do modelo já que a complexidade afeta a sua generalização. Esta seleção também pode ser útil para evitar o problema do argumento seletivo que distorce ou não contempla as variáveis de entrada que realmente influem na variável de saída.

Um problema na literatura de previsão de séries temporais ainda não totalmente resolvido é o ajuste da janela de predição dinâmica. A reconstrução dinâmica a partir dos dados disponí-

veis é uma opção a ser considerada para estimar esta janela e consiste no ajuste dos parâmetros dimensão de imersão (M) e tempo de atraso (*lag* L). Estes dois parâmetros definem uma janela de predição que se torna dinâmica ao se deslocar ao longo da série no tempo. A reconstrução dinâmica geralmente é sub-ótima, depende basicamente da série temporal analisada e dos métodos utilizados para ajustar estes parâmetros. Cada um destes métodos pode ser adequado para determinadas aplicações, podendo não funcionar para outras. Entretanto, são alternativas ao método da escolha arbitrária dos parâmetros da janela de predição.

A motivação por trás da utilização das redes neurais artificiais (RNA) é a possibilidade de se encontrar soluções eficazes para problemas de difícil tratamento, já que o cenário atual exige soluções cada vez mais competitivas. Entretanto, as RNA só podem ser devidamente exploradas por meio de procedimentos refinados de análise e síntese, ou seja, os recursos de processamento devem ser aplicados na medida certa e na situação apropriada. Tem-se que avaliar os ganhos de desempenho na presença de incrementos de complexidade. A complexidade da implementação de um modelo via RNA pode aumentar e tornar difícil encontrar a solução global ótima. Isto ocorre principalmente quando não se consegue encontrar o subconjunto de variáveis de entrada apropriado para se estimar adequadamente a janela de predição dinâmica de um sistema variante no tempo. Entretanto, nas competições patrocinadas pelo Santa Fé Institute, a classe de modelos neurais apresentou os melhores resultados para a previsão de séries temporais multivariadas, não-lineares e variantes no tempo [WEIGEND and GERSHENFELD, 1994].

O critério de escolha do subconjunto de testes é importante para a qualidade do ajuste do modelo e das previsões. Inicialmente são aplicados os testes de estacionariedade, de dependência temporal e da presença de não linearidades. Como os investidores buscam maximizar lucros, os testes que avaliam as possibilidades de retornos são mais relevantes para esta área que aqueles que avaliam somente a precisão. O teste utilizado para avaliar a habilidade preditiva do modelo em relação ao passeio aleatório (PA) será o NMSE (*normalized mean square error*), que é muito usado em todas as comunidades de previsão de séries temporais. A função de utilidade dos testes pode influenciar na avaliação do modelo, ou seja, um modelo pode ser superior para uma determinada função de utilidade, mas não para outra.

Entretanto, mesmo com a utilização destes critérios, a avaliação da qualidade das previsões é uma tarefa difícil que pode levar facilmente a conclusões equivocadas. Este trabalho utiliza o teste de WHITE (2000) para comparar um grupo de modelos de previsão na média condicional a um modelo *benchmark*. Este teste é não paramétrico e serve para verificar se pelo menos um modelo de um conjunto de modelos comparados gera previsões com superioridade significativa sobre um modelo *benchmark*. Em [HANSEN, 2005] foi proposto que este teste torna os resultados sensitivos à inclusão de modelos com resultados de previsão pobres entre os modelos comparados.

A literatura sobre a previsibilidade da taxa de câmbio é extensa. Parte dela busca provar se a série no tempo dos retornos da taxa de câmbio é ou não um processo Martingale. A importância desta pesquisa está nas suas implicações econômicas. Quando este processo não é Martingale, abrem-se as possibilidades de *overshooting* ou *undershooting* na taxa de câmbio, aversão ao risco e a intervenção oficial no mercado de moedas estrangeiras (ou indiretamente via operações de mercado aberto). Quando este processo não é Martingale, esta série tem algum nível de previsibilidade e a paridade do poder de compra não ocorre. No Brasil, que teve uma forte desvalorização do real, com posterior adoção do regime de câmbio flutuante, é relevante

analisar sua previsibilidade após estes eventos.

Pesquisas sobre a umidade na troposfera apontam para uma relação entre a umidade específica e a temperatura. Na troposfera, tanto na parte baixa como na alta, quando a temperatura aumenta a umidade também aumenta. Já uma relação inversa intensa ocorre quando a umidade relativa é comparada [SUN and OORT, 1995]. Por outro lado, a umidade é a quantidade de vapor d'água presente no ar e é razoável que durante uma chuva a umidade seja fortemente influenciada, resta saber a importância desta influência ao longo do tempo já que não chove na terra continuamente. Neste trabalho é implementado um modelo de previsão da umidade horária no microclima da região de Londrina-PR, utilizando um método de seleção de variáveis. Os dados coletados sobre a umidade nesta região sugerem que as variações são rápidas, muito bruscas e de intensidade elevada. A relevância social deste estudo está na possível economia de bilhões de dólares no setor da agricultura, principalmente na cultura da soja, já que o custo dos produtos químicos utilizados no controle da ferrugem asiática, a principal ameaça à esta cultura, pode ser reduzido substancialmente. Além disso, como a utilização dos produtos químicos diminui, o meio ambiente é menos agredido.

Finalmente, as pesquisas nas áreas de previsão de séries temporais e de redes neurais artificiais têm caráter multidisciplinar e qualquer contribuição que esta tese possa representar refletirá de forma abrangente em outros setores de pesquisa.

1.2 Questão principal, objetivos e hipótese geral da tese

1.2.1 Questão principal

A variação da taxa de câmbio brasileira é um processo estocástico Martingale? E a variação da umidade no microclima da região de Londrina-PR?

Para responder a estas questões foi definido um objetivo principal, desdobrado em objetivos secundários, de maneira a estabelecer os passos da investigação capaz de respondê-las.

1.2.2 Objetivos

Objetivo principal:
Verificar empiricamente se a variação da taxa de câmbio brasileira e da umidade no microclima da região de Londrina-PR são ou não processos tipo Martingale. Investigar quais modelos fornecem a melhor qualidade de previsão para cada função de perda, as limitações destes melhores modelos e se os mesmos têm aplicações práticas.

Objetivos secundários:

- Provar por meio de testes que existem não linearidades na média condicional e dependência temporal (linear e não linear) nestas séries.
- Propor uma metodologia para a seleção de variáveis.
- Ajuste da janela de previsão por meio da reconstrução dinâmica e de RNA.

- Investigar empiricamente se as magnitudes das não linearidades influem na especificação de modelos de previsão tipo ARMA-GARCH e neurais.

1.2.3 Hipótese geral

A variação da taxa de câmbio brasileira e da umidade no microclima da região de Londrina-PR não são processos tipo Martingale.

1.3 Metodologia

1.3.1 Metodologias de previsão de séries temporais

Do início do século passado até 1920, a predição de séries temporais era realizada a partir da extrapolação de dados no domínio do tempo. Coube a YULE no início do século passado, pesquisando as manchas solares, propor a técnica autoregressiva em [YULE, 1926] e ser o primeiro a abordar o problema das regressões sem sentido e espúrias. A técnica autoregressiva era puramente linear e consistia em utilizar a soma ponderada das observações anteriores para determinar o valor previsto. Durante aproximadamente cinqüenta anos, exceto pela aplicação do filtro adaptativo linear de WIDROW e HOFF (1960) por Hu, em estudos de previsão climática, o modelo baseado em um filtro autoregressivo acrescido do ruído foi praticamente o único a ser utilizado nesta área.

No final da década de 1960, os professores George E. P. BOX e G. M. JENKINS publicaram vários trabalhos sobre a teoria de controle e de análise de séries temporais. Em 1970 publicaram o livro *Time series analysis, forecasting and control* [BOX and JENKINS, 1970] apresentando uma metodologia para a análise de séries temporais, e em 1976 e 1994 foram lançadas versões revisadas desse livro [BOX and JENKINS, 1976] e [BOX et al., 1994] que normalmente são as mais mencionadas. A grande importância desse trabalho foi reunir as técnicas existentes numa metodologia para construir modelos ARMA que descreviam com uma certa precisão e de forma parcimoniosa o processo gerador da série temporal. Esta classe de modelos tem obtido considerável sucesso nas áreas econômica e financeira, mas nem sempre consegue lidar com os fatos estilizados característicos de dados financeiros (conglomerados de valores extremos, assimetrias e excesso de curtose).

Os modelos de equações simultâneas foram mais utilizados nas décadas de 1960 e 1970, quando modelos refinados da economia americana baseados em equações simultâneas dominaram a previsão econômica. Entretanto, em [SIMS, 1980] foi sugerido que a decisão sobre a escolha das variáveis era muito subjetiva. Ele achava que se há uma verdadeira simultaneidade entre um conjunto de variáveis, todas elas devem ser tratadas igualmente; não deve haver distinção *a priori* entre variáveis de entrada e saída. Foi com este espírito que SIMS (1980) apresentou a classe de modelos lineares VAR (*vectors autoregressives*). Os conceitos base associados ao VAR são: dependência temporal; impacto dinâmico de um distúrbio aleatório; seleção de modelos (AKAIKE, SCHARWZ).

Em [GRANGER and NEWBOLD, 1974] foi apresentada uma análise criteriosa sobre regressões espúrias. A regressão de uma variável sobre uma ou mais variáveis muitas vezes pode

fornecer resultados sem sentido ou espúrios. Uma maneira de se prevenir é testar se as séries temporais são cointegradas. Basicamente, cointegração significa que a combinação de duas ou mais séries individualmente não estacionárias pode resultar em uma série estacionária. O teste apresentado em [ENGLE and GRANGER, 1987] pode ser utilizado para verificar se duas ou mais séries são cointegradas, ou seja, sugere se há ou não uma relação entre elas a longo prazo (equilíbrio). Neste mesmo trabalho foi proposto um mecanismo de correção de erro para conciliar o comportamento a curto prazo de uma variável com seu comportamento a longo prazo, surgindo a classe de modelos lineares VEC (*vector error correction*). Posteriormente, em [JOHANSEN and JUSELIUS, 1990] foi sugerida uma complementação ao teste de Granger. Quando não existe cointegração, pode-se utilizar um modelo linear VAR. Caso contrário, aplica-se um modelo linear VEC. No artigo [SIMS et al., 1990] foi analisada a escolha do tamanho do *lag* nos testes de raízes unitárias, em modelos VAR e testes de cointegração. Os modelos lineares ARMA, VAR e VEC são muito utilizados nas áreas financeira e econômica.

No início dos anos 80, a comunidade estatística propôs alternativas aos modelos lineares de memória curta existentes (ARMA e VAR):

a) Modelos não lineares na média condicional: em [TONG and LIM, 1980] foram propostos os modelos autoregressivos com regimes determinados por limiares (*threshold* VAR, TVAR) ou por uma função de transição suave (*smooth transition* VAR, STVAR) que têm os regimes definidos por uma variável observada e por uma função de transição e combina dois ou mais modelos lineares de uma forma não-linear; no artigo [SUBBA and GABR, 1984] foram apresentados os modelos bilineares; em [OZAKI, 1980] foram sugeridos os modelos EXPAR (*exponential autoregressive*). Raras são as aplicações destes três modelos na literatura desta área. Não tão raras são as aplicações em finanças que utilizam mudanças de regime Markoviano abordado na referência [HAMILTON, 1994].

b) Modelos não-lineares na variância condicional: ENGLE (1982) e BOLLERSLEV (1986) apresentaram modelos não-lineares na variância (ARCH - *autoregressive conditional heteroskedasticity* e GARCH - *generalized ARCH*). O objetivo de Engle era descrever o comportamento persistente da volatilidade da série de retornos de um ativo. Nos modelos ARCH a variância condicional muda com o tempo enquanto a variância não condicional permanece constante. O objetivo de BOLLERSLEV (1986) era a generalização dos modelos ARCH como o próprio nome sugere. Posteriormente surgiram variações destes modelos, como o IGARCH, EGARCH e TGARCH. Em HARVEY et al. (1994) foi proposto o modelo de Volatilidade Estocástica (VE), com origem no modelo estrutural proposto pelo mesmo autor [HARVEY, 1992], que modela a variância por meio de um processo não observado que tenta captar informações que chegam ao mercado.

No final da década de 1980, pesquisadores ligados às comunidades de sistemas dinâmicos e dos físicos apresentaram metodologias para desenvolver modelos não-lineares de previsão de séries temporais via espaço de estados motivados pelo fenômeno constatado do caos. Estas comunidades abordaram questões muito interessantes ligadas à área de séries temporais, gerando contribuições, como o método BDS [BROCK et al., 1996] para avaliar a existência de dependência temporal (linear e não linear) em uma série temporal.

O foco da questão passou então a incluir a possibilidade da obtenção de informações por meio de técnicas de sistemas dinâmicos não-lineares para auxiliar as metodologias estatísticas consagradas de previsão. Fortalecendo a idéia de que deve prevalecer a performance de previsão

em que, caso fenômenos dinâmicos estejam envolvidos, aspectos temporais estarão presentes. Em síntese, existe alguma lei que rege o comportamento entrada-saída para um determinado domínio de interesse. Caso contrário, estarão em jogo apenas aspectos da distribuição espacial dos dados.

Nesta mesma década, com o crescimento da capacidade de processamento e de memória dos computadores, viabilizaram-se os estudos de séries temporais com grandes conjuntos de dados a partir de modelos mais complexos. Em [LAPEDES and FARBER, 1986] foi utilizado um perceptron multi-camadas (PMC) na predição de problemas populares na comunidade que estudava sistemas caóticos. [UTANS and MOODY, 1991] desenvolveram uma metodologia que incluía uma parcela de erro para penalizar o número efetivo de parâmetros de um modelo não-linear qualquer. Em [WEIGEND et al., 1990] foi apresentada uma técnica para penalizar parâmetros extras de um PMC, resultando em um modelo mais parcimonioso e mais eficiente que um modelo TAR correspondente. No artigo [REFENES et al., 1993] foi proposto um método para adicionar unidades de neurônios e foi mostrado que este método superou um modelo ARMA equivalente na predição de taxa de câmbio.

Em 1990, WEIGEND e GERSHENFELD tiveram a idéia de realizar uma competição patrocinada pelo Santa Fé Institute, cujo objetivo era o desenvolvimento de pesquisas na área de séries temporais. Esta competição envolvia pesquisadores que utilizavam séries temporais em estudos das áreas de biologia, economia, física pura e experimental, astrofísica, análise numérica, estatística aplicada e sistemas dinâmicos. A idéia foi um sucesso, e culminou, em 1992, num encontro patrocinado pela OTAN que reuniu os participantes do desafio e demais interessados. A maioria das contribuições relevantes resultantes desse encontro foram a partir de métodos conexionistas.

Não se deve perder a noção dos fatos e observar que, naturalmente, não se sabe com certeza qual será o valor futuro previsto de uma série temporal, principalmente daquelas geradas por processos multivariados e de forma funcional desconhecida. Finalmente, pode-se alegar que as simplificações feitas levam a modelos com premissas irreais. Isso sempre pode acontecer com qualquer modelo, que é, por definição, uma simplificação da realidade. No entanto, estes modelos nos permitem algumas previsões, o que para um primeiro modelo de análise e avaliação já é de grande relevância informacional. Desta forma, estes modelos de previsão podem também ser utilizados com eficiência em aplicações práticas.

As metodologias de predição na média apresentadas nesta tese são baseadas em RNAs tipo RBF. A primeira utiliza uma janela de predição inteligente dinâmica (JPID) para atribuir os centros por meio da análise de componentes principais (PCA) e a variância é ajustada via fator de dispersão adaptativo para a função spline fina. Os parâmetros da janela de previsão são inicialmente estimados por meio da reconstrução dinâmica e o ajuste final é realizado via RNA.

A segunda utiliza a seleção de variáveis e tem os centros atribuídos pelo algoritmo ARIA (*Adaptive Radius Immune Algorithm*), apresentado em [BEZERRA et al., 2005], com origem em conceitos de redes imunes [CASTRO and ZUBEN, 2001]. Por meio da seleção de variáveis, busca-se identificar conteúdo informacional nas variáveis de entradas candidatas mais relevantes e naquelas não tão relevantes mas não redundantes (predominantes), para ser transferido para os parâmetros livres da RNA.

Busca-se classes de modelos de previsão de séries temporais que apresentem convergência rápida da rede (tempo de processamento pequeno) e pouca demanda por memória e que as-

segure. Estas metodologias tornam-se mais competitivas à medida que o valor absoluto das componentes não lineares do sistema sejam mais expressivas em relação às componentes lineares.

1.3.2 Base de dados

Os dados utilizados neste trabalho foram fornecidos pela Economática, Bloomberg, Banco Central do Brasil (BCB) e pela Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA). As séries temporais analisadas são: retornos da taxa cambial brasileira em regime de câmbio flutuante, a partir de Janeiro de 2000 até Fevereiro de 2005, fornecida pelo BCB; vinte séries financeiras e econômicas, de Janeiro de 2000 até Fevereiro de 2005, utilizadas na seleção das variáveis, fornecidas pela Economática e Bloomberg; os dados sobre a variação da umidade na região de Londrina-PR, de Janeiro de 1999 a Dezembro de 2000, foram fornecidos pela EMBRAPA. Estas séries temporais serão apresentadas com mais detalhes no Capítulo 6.

1.4 Organização da tese e contribuições

Organização dos capítulos e suas principais contribuições:

- **Capítulo 2**: Aborda sucintamente o problema de predição de séries temporais estacionárias. Apresenta os métodos ARMA, GARCH e o filtro FIR, indicados para lidar com sistemas estacionários, visando posteriormente confrontar os resultados obtidos por estes métodos com os baseados em redes neurais artificiais.

- **Capítulo 3**: Propõe uma metodologia para a seleção de variáveis por meio de um filtro seguido por um *wrapper*. A principal contribuição é a sugestão de uma metodologia para implementar um filtro baseado na teoria da informação, em que a seleção de variáveis utiliza os conceitos de relevância e redundância. Este filtro é utilizado para selecionar um subconjunto de variáveis preliminares que serão avaliadas por um *wrapper*, baseado em redes neurais, que fará a seleção do subconjunto de variáveis considerado ótimo.

- **Capítulo 4**: Apresenta métodos para ajustar a janela de predição dinâmica por meio da reconstrução dinâmica, análise de dinâmica não linear e mineração de *clusters* com capacidade preditiva.

- **Capítulo 5**: Aborda o problema de predição de séries temporais não-lineares via redes neurais artificiais tipo RBF. Apresenta um tipo de rede RBF com função de base spline fina, cujos centros são estimados por meio da análise de componentes principais (PCA) e um fator de variância adaptativo implementado via otimização do desempenho de predição. O outro tipo de rede RBF utilizada tem os centros determinados via algoritmo ARIA. A principal contribuição do capítulo é a dedução matemática do método de ajuste otimizado do fator de variância adaptativo para uma função spline fina. Faz-se o ajuste dos centros da rede via conceitos de redes imunes (algoritmo ARIA), e o ajuste

dos parâmetros (pesos) da RNA por meio de testes estatísticos que garantem que as não-linearidades não foram negligenciadas. Também são apresentados os testes utilizados na avaliação de desempenho de predição para modelos não aninhados e aninhados (*nested*).

- **Capítulo 6**: Apresenta os resultados das análises e das previsões da variação da taxa de câmbio brasileira e da umidade no microclima da região de Londrina-PR.

- **Capítulo 7**: Apresenta as conclusões e perspectivas para trabalhos futuros. Também explicita as limitações e a aplicabilidade prática dos modelos propostos.

Capítulo 2

Previsão de Séries Temporais Estacionárias

2.1 Introdução

Historicamente, a teoria sobre os preditores lineares deriva principalmente dos seguintes trabalhos: [YULE, 1926], [WIENER, 1949], [KOLMOGOROV, 1957], [WIDROW and HOFF, 1960] e [BOX and JENKINS, 1970]. A literatura sobre o assunto já é vasta. Um livro que se tornou um clássico é a edição revisada de BOX e JENKINS (1976). Posteriormente, surgiu uma revisão desta edição realizada em conjunto com REINSEL [BOX et al., 1994]. Uma referência importante sobre o assunto, inclusive utilizada por alguns econometristas que trabalham no mercado, é [HAMILTON, 1994], que não tem sido atualizada. Já o livro de ENDERS (2004) foi atualizado e é bastante direcionado para aplicações. No Brasil, uma referência bastante expressiva e com edição recente é [MORETTIN and TOLOI, 2004].

Na prática, sabe-se que a maioria das séries temporais resulta de experiências que não poderão ser repetidas e geralmente ficam melhor representadas por processos estocásticos. O trabalho [NELSON and PLOSSER, 1982] sugeriu que a maioria das séries temporais econométricas são integradas de primeira ou segunda ordem, transformando-se em um marco na econometria porque até então pensava-se que estas séries eram deterministas. Em [PRIESTLY, 1989] e [TERÄSVIRTA and GRANGER, 1993] são apresentadas várias justificativas para a escolha de modelos estocásticos, sendo que TERÄSVIRTA e GRANGER ressaltam que os processos caóticos deterministas têm apresentado pouca importância em aplicações práticas em finanças, exceto se o tratamento for realizado por meio de técnicas associadas a processos estocásticos. Entretanto, os estudos sobre o caos determinista resultaram em ferramentas de análise úteis como a estatística BDS de BROCK, DECHERT e SHEINKMAN (1996).

Um processo estocástico ergódico e estacionário pode ser representado por um modelo projetado a partir de uma amostra apropriadamente coletada. Caso contrário, uma amostra somente não é suficiente para generalizar a respeito do processo. Assim, a análise preliminar das bases de dados possibilita a assunção de hipóteses necessárias para a especificação e a implementação de modelos mais eficazes.

Os modelos lineares e estacionários de uma série no tempo têm duas características particularmente desejáveis: podem ser mais facilmente compreendidos e existem métodos bem

direcionados para implementá-los. A contrapartida para estas conveniências é que podem ser inteiramente impróprios para sistemas com moderadas complicações. Entretanto, estes métodos ainda são bastante utilizados na predição de séries temporais.

As análises preliminares das séries temporais, apesar de importantes, não apresentam como produto final as relações bem definidas. Será a modelagem que identificará o tipo de relação funcional entre as variáveis e estimará os parâmetros que fazem parte desta relação. Busca-se com este tipo de análise a verificação empírica de algumas hipóteses de comportamento e sabe-se que uma relação estatística, por mais forte e sugestiva que seja, jamais pode estabelecer uma relação causal que deve vir de outra teoria (econômica, meteorológica e outras).

Uma regressão simples é somente um estudo das relações entre variáveis. O principal objetivo é, conhecida esta relação, poder estimar o valor de uma variável a partir da outra. Isto pode até ser útil na escolha das variáveis de entrada de um modelo matemático, embora não se trata de um método de previsão de séries temporais.

Os modelos *autoregressive moving average* (ARMA) [BOX and JENKINS, 1970], dão ênfase em analisar as propriedades estocásticas da série no tempo, deixando que os dados falem por si mesmos. Isto concorreu para que esta classe de modelos fosse entendida por alguns como *ateórica*, ou seja, não derivava de nenhuma teoria, como os modelos econométricos que são baseados na teoria econômica. Os modelos ARMA integrados (ARIMA) são utilizados quando torna-se possível remover as tendências, restando ao final somente um erro estocástico. Neste caso a equação é homogênea, mas quando não se consegue remover as tendências trata-se de uma equação heterogênea. Nas séries com sazonalidades, além de remover as tendências não sazonais, aplica-se um número adequado de diferenças sazonais.

O estudo da componente do erro estocástico (ou irregular), após serem extraídas as componentes tendência, ciclo e sazonalidade, muitas vezes torna-se o foco principal do problema de previsão de séries temporais. Caso esta componente possa ser bem caracterizada por uma função de densidade de probabilidade (fdp), e o processo seja linear, estacionário e com curta dependência temporal, o que nem sempre ocorre na prática, um modelo ARMA poderá possibilitar boas previsões. Quando se trata de um processo linear, estacionário, mas com longa dependência, uma rede neural ou um modelo ARFIMA poderá ser uma alternativa melhor. Caso exista uma componente sazonal, utiliza-se um modelo SARIMA.

Nos mercados financeiros, além de estimar os retornos do ativo, também é importante avaliar o risco (volatilidade). Quando a variância condicional de uma série temporal financeira não é constante, mas a média e a variância não condicional de longo prazo são constantes, pode-se estimar a volatilidade (incerteza) associada a este ativo. A família dos modelos GARCH foi criada para este fim. Em [HARVEY et al., 1994] foi apresentada uma classe de modelos alternativa à família ARCH denominada de volatilidade estocástica (VE), incluindo um termo estocástico que torna o valor de volatilidade calculado mais suavizado que os estimados por meio dos modelos ARCH.

O teste das raízes unitárias é utilizado para se verificar se a série temporal é estacionária, ou seja, se é explosiva ou estável. A estabilidade da série é uma condição necessária para utilização de modelos ARMA, mas não suficiente. É necessário também que haja dependência temporal entre as amostras e que a série seja linear. Este assunto será abordado em detalhes posteriormente.

Os modelos de previsão de séries temporais podem ser divididos em dois grandes grupos:

modelos paramétricos e não-paramétricos. A metodologia proposta por BOX e JENKINS (1970 e 1976) é classificada como paramétrica e tem sido largamente utilizada para predições em áreas como economia, finanças, meteorologia, hidrologia e outras. Os modelos baseados em redes neurais são exemplos de modelos não-paramétricos.

Em aplicações práticas, não linearidades, não estacionariedades e longa dependência temporal podem ser individualmente ou em conjunto intrínsecas ao processo estocástico estudado. Assim, nem sempre as premissas assumidas durante a modelagem do problema são totalmente verdadeiras e o modelo é geralmente sub-ótimo. Logo, é vantajoso abandonar um modelo em detrimento de outro mais acurado, ou seja, que represente melhor a realidade, principalmente quando esta realidade se altera ao longo do tempo.

O filtro linear adaptativo FIR (*Finite Impulse Response*), da classe dos modelos AR, tem grande capacidade de adaptação e serviu como uma das fontes de inspiração para a área de redes neurais. Uma arquitetura neural com um filtro FIR de ordem p substituindo as conexões sinápticas, distribuída no tempo, venceu a competição patrocinada pelo *Santa Fé Institute* [WEIGEND and GERSHENFELD, 1994]. Este tipo de rede neural pode lidar diretamente com a previsão de séries temporais não lineares. Entretanto, o filtro linear adaptativo FIR convencional é indicado somente para processos lineares e estacionários. Como este tipo de filtro tem grande capacidade de adaptação para acompanhar variações bruscas de sinais, tem sido bastante utilizado nas áreas de antenas e radares.

Os objetivos principais deste capítulo são: apresentar os modelos ARIMA, GARCH e os filtros adaptativos lineares (tipo FIR); analisar as vantagens e limitações destes modelos. Nesta tese, faz-se uma escolha pela modelagem em tempo discreto em razão da sua conveniência, simplicidade em relação à de tempo contínuo e principalmente porque, mesmo que os processos geradores sejam contínuos, as coletas das séries temporais são normalmente em períodos de tempo discreto, como por exemplo, os dados das séries econômicas. Considera-se que as observações são feitas em intervalos regulares de tempo e que este intervalo tem duração de uma unidade de tempo.

Este capítulo foi organizado como se segue: na Seção 2.2, as definições de processos estocásticos lineares e estacionários são apresentadas; na Seção 2.3, ilustra-se os modelos ARIMA e os testes relacionados; na Seção 2.4 é analisado sucintamente o modelo GARCH. Na Seção 2.5, aborda-se os filtros lineares adaptativos.

2.2 Definições

As definições básicas necessárias para a construção e fundamentação deste trabalho são apresentadas nesta seção.

2.2.1 Modelos lineares

Definição 2.2.1: Um processo estocástico $Y_{t,\theta} = \{Y_{t,\theta};\ t \epsilon T,\ \theta \epsilon \Omega\}$ é um conjunto de v.as. definidas sobre um espaço de probabilidade em (Ω, Υ, P) e tem como índices dos elementos da amostra os valores de $t \epsilon T$. O conjunto de elementos da amostra é normalmente composto de números inteiros ou reais. Já uma série temporal é definida a partir de um determinado $\theta \epsilon \Omega$, e

por questões de simplicidade será representada por $\{Y_t\}$. Assim, é uma parte de uma trajetória entre as muitas que poderiam ser associadas a um processo estocástico. Nas séries com a mesma estrutura, cada série temporal é uma realização possível do processo em questão. Logo, trajetória e série temporal têm o mesmo significado. Observa-se que uma variável discreta y é dita uma variável estocástica (randômica) se para qualquer número real r existe a probabilidade que y terá um valor menor ou igual a r, ou seja, $P(y \leq r) < 1$. Caso exista algum r para o qual $P(y = r) = 1$, y será determinístico em vez de randômico.

Definição 2.2.2: As distribuições finito-dimensionais de uma série temporal $\{Y_t\}$ são dadas por:

$$F(y_1, y_2, \ldots, y_N; t_1, t_2, \ldots, t_N) = P(Y_{t_1} \leq y_1, Y_{t_2} \leq y_2, \ldots, Y_{t_N} \leq y_N). \tag{2.1}$$

Definição 2.2.3: As funções média, variância, autocovariância e autocorrelação serão apresentadas a seguir

$$E(Y_t) = \int_{-\infty}^{\infty} Y dF(y;t) = \mu_{yt}, \tag{2.2}$$

$$Var(Y_t) = \int_{-\infty}^{\infty} (Y - \mu_{yt})^2 dF(y;t) = \gamma_y(t,t), \tag{2.3}$$

$$Cov(y_t, y_k) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (Y_t - \mu_{yt})(Y_k - \mu_{yk}) dF(y_t, y_k; t, k) = \gamma_y(t,k), \tag{2.4}$$

$$Corr_y(t,k) = \frac{\gamma_y(t,k)}{\sqrt{\gamma_y(t,t)}\sqrt{\gamma_y(k,k)}} = R_y(t,k). \tag{2.5}$$

Definição 2.2.4: Uma seqüência $\{\varepsilon_t\}$ é um ruído branco (*White Noise* - WN) se e somente se $E(\varepsilon_t) = 0$, $Var(\varepsilon_t) = \sigma^2$, não correlacionada com todas as outras realizações, ou seja, não tem memória.

Definição 2.2.5: Seja Y_t um processo estocástico, este é dito linear se pode ser representado na forma: $Y_t = \mu_{yt} + \sum_{j=-\infty}^{\infty} \Psi_j \varepsilon_{t-j}, \forall t$, em que $\varepsilon_t \sim WN(0, \sigma^2)$ e Ψ_j é uma seqüência de constantes tais que $\sum_{j=-\infty}^{\infty} |\Psi_j| < \infty$. Observa-se que a condição de não correlação é diferente daquela dada para uma seqüência de v.a.'s independentes. Quando se trata de um processo gaussiano esta distinção desaparece.

Definição 2.2.6: Um processo estocástico Y_t é um Martingale se e somente se $E(|Y_t|) \leq \infty$ e $E(Y_t|Y_{t-1}, Y_{t-2}, \ldots) = Y_{t-1}$ para todo t ou, de forma equivalente, $E(Y_t - Y_{t-1}|f(Y_{t-1}, Y_{t-2}, \ldots)) = 0$, e neste caso será chamado de diferença Martingale. A diferença Martingale impõe uma condição mais forte do que aquela dada para um processo não autocorrelacionado serialmente. Este tipo de processo não pode ser previsto com base em uma função linear de seus valores passados. Já uma diferença Martingale (DM) não pode ser prevista nem por uma seqüência linear e nem tampouco por uma não linear.

2.2.2 Modelos estacionários

Definição 2.2.7: Seja Y_t um processo estocástico, este é dito fracamente estacionário, ou estacionário de segunda ordem, se a média $\mu_{yt} = E(Y_t)$, a variância $Var(Y_t) = \sigma^2$ e a covariância

$\gamma_y(t,k) = \gamma(|t-k|)$, $\forall t \quad \epsilon \quad T$, ou seja, é independente do tempo e, consequentemente, a correlação também será. Caso $Y_t = \beta t + \varepsilon_t$ represente uma tendência no tempo mais um ruído branco gaussiano, este processo não será estacionário já que $E(Y_t) = \beta t$, ou seja, a esperança da média depende do tempo.

No contexto desta tese, quando se fala que um determinado processo estocástico é estacionário, assume-se que o processo estocástico é fracamente estacionário (WSS - *Wide Sense stationary*).

Definição 2.2.8: Seja Y_t um processo estocástico, este é dito estritamente estacionário se $(Y_1, Y_2, \ldots, Y_k) \cong (Y_{1+h}, Y_{2+h}, \ldots, Y_{k+h}), \forall (k,h) \geq 1$, em que $\cong$ indica que os dois vetores aleatórios são identicamente distribuídos.

Uma estacionariedade fraca não implica que esta seja estrita e que um processo é estritamente estacionário quando $E(Y_t^2) < \infty$. Um caso em que ocorre um processo estritamente estacionário de segunda ordem é aquele que trata de uma distribuição gaussiana, já que esta distribuição pode ser determinada apenas pelos dois primeiros momentos (média e variância). Observa-se que para um processo estocástico estacionário a média deste processo tende para a média populacional da amostra e quando não é estacionário a média passa a ser uma variável aleatória.

A questão da ergocidade de um processo estocástico é importante porque as definições de estacionariedade são desenvolvidas a partir da esperança $E(Y_t)$ e, como na prática o que se tem geralmente é uma única amostra, o que se pode calcular efetivamente é a média no tempo desta amostra, de acordo com a fórmula abaixo.

$$\overline{y} = (1/T) * \sum_{t=1}^{T} y_t, \quad t = 1, 2, \ldots, T. \tag{2.6}$$

em que T é o tamanho da amostra.

Caso esta média no tempo convirja eventualmente para $E(Y_t)$, e se tratar de um processo estacionário, pode haver ergocidade.

Definição 2.2.9: um processo estacionário na variância é tido como ergódico na média quando a fórmula anterior da média converge em probabilidade para $E(Y_t)$ quando $T \to \infty$ ou obedece à seguinte condição: $\sum_{k=0}^{\infty} |\gamma_y(t,k)| < \infty$, ou seja, todas as raízes estão dentro do círculo unitário, satisfazendo a condição de invertibilidade. Um processo ergódico na média significa que a autocovariância $\gamma_y(t,k)$ converge rápido para zero, para k suficientemente grande.

Definição 2.2.10: um processo estacionário na variância é tido como ergódico na variância se

$$(1/(T-j)) * \sum_{t=j+1}^{T} (y_t - \bar{y})(y_{t-j} - \bar{y}) \to \gamma_y(t, t-j) \tag{2.7}$$

para todo j. No caso de processos gaussianos estacionários, a condição dada pela definição 2.2.8 para as autocovariâncias é suficiente para garantir ergocidade para todos os momentos.

Em muitas aplicações, os conceitos de ergocidade e estacionariedade podem ser confundidos se não forem tratados com critério. Para ilustrar a diferença entre os mesmos, apresenta-se um exemplo em que o processo é estacionário, mas não ergódico. Supondo-se uma média $\mu^{(i)}$ estimada a partir de uma realização gerada por uma distribuição $N(0, \lambda^2)$, ou seja

$$Y_t^{(i)} = \mu^{(i)} + \varepsilon_t$$

com ε_t gerado por um ruído branco gaussiano independente de $\mu^{(i)}$. Note-se que

$$\mu_t = E(\mu^{(i)}) + E(\varepsilon_t) = 0$$

é a média não condicional. A variância e a covariância são dadas por

$$\gamma_y(t,t) = E(\mu^{(i)} + \varepsilon_t)^2 = \lambda^2 + \sigma^2$$

e

$$\gamma_y(t,t-j) = E(\mu^{(i)} + \varepsilon_t)(\mu^{(i)} + \varepsilon_{t-j}) = \lambda^2$$

para $j \neq 0$. Este processo é estacionário já que a média, variância e covariância não dependem do tempo, mas não satisfaz à condição de ergocidade na média apresentada anteriormente já que a média de $Y_t^{(i)}$ no tempo é dada por

$$(1/T) * \sum_{t=1}^{T} Y_t^{(i)} = (1/T) * \sum_{t=1}^{T} (\mu^{(i)} + \varepsilon_t) = \mu^{(i)}$$

ou seja, converge para a média de $Y_t^{(i)}$ e não para zero.

2.3 Modelos de previsão de séries temporais estacionárias

2.3.1 Modelos ARMA

Na teoria das equações lineares de diferenças com componentes estocásticas está a base teórica para o estudo das séries temporais econométricas. Isto é válido também para aplicações de outras áreas (física, engenharia, biologia etc). Uma equação linear de diferenças especial de n-ordem com coeficientes constantes é apresentada em seguida.

$$y_t = \alpha_0 + \sum_{i=1}^{n} \alpha_i y_{t-i} + x_t \tag{2.8}$$

em que n é a ordem da equação linear de diferenças e os coeficientes α_i são funções das variáveis ditadas pela teoria subjacente. O termo x_t pode ser visto como o processo forçador (*forcing process*). A forma deste processo pode ser bem geral; x_t pode ser qualquer função do tempo, valores atuais e defasados de outras variáveis, e/ou distúrbios estocásticos. A equação é linear porque as potências de cada variável dependente são unitárias.

Esta teoria das equações lineares de diferenças pode ser útil para representar o processo gerador x_t por meio de um modelo estocástico. É possível combinar um processo de médias móveis com uma equação linear de diferenças, resultando em modelo autoregressivo associado a médias móveis. Assim, a equação (2.8) transforma-se em

$$y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=0}^{q} \beta_i \varepsilon_{t-i}. \tag{2.9}$$

Resultando na metodologia de Box and Jenkins (1976) em que y_t é gerado inteiramente por um processo estocástico cuja representação matemática simplificada do modelo ARMA(p,q) pode ser dada por

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \ldots + \alpha_p y_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q} \tag{2.10}$$

em que o termo α_0 representa uma constante no modelo; $\alpha_1, \ldots, \alpha_p$ são parâmetros que ajustam os valores passados de y do instante imediatamente anterior até o mais distante representado por p; os valores de ε_t representam uma seqüência de choques aleatórios e independentes uns dos outros e é uma porção não controlável do modelo; os parâmetros $\beta_1, \ldots, \beta_q$ possibilitam escrever a série em função dos choques passados. Quando $p = 0$ tem-se um modelo de média móvel (MA) puro e quando $q = 0$ resulta em um modelo autoregressivo (AR) puro.

Um modelo ARMA, quando $\alpha_0 = 0$, também pode ser representado de forma compacta por

$$\alpha(B) Y_t = \beta(B) \varepsilon_t \tag{2.11}$$

em que o operador autoregressivo de ordem p é dado por

$$\alpha(B) = (1 - \alpha_1(B) - \alpha_2(B^2) - \ldots - \alpha_p(B^p))$$

e o operador de médias móveis de ordem q, com $\beta_0 = 1$, é dado por

$$\beta(B) = (1 + \beta_1(B) + \beta_2(B^2) + \ldots + \beta_q(B^q))$$

em que $B^m Y_t = Y_{t-m}$.

A parte estocástica da equação (2.8) é sempre estacionária se a seqüência $\{x_t\}$ for estacionária. As raízes da equação homogênea da parte autoregressiva desta equação determinam se a seqüência $\{Y_t\}$ é estacionária. Uma maneira de se verificar as condições de estacionariedade e de invertibilidade é encontrar a equação característica inversa, partindo-se da equação (2.9), após algumas manipulações algébricas, resulta em

$$y_t = \frac{\alpha_0}{1 - \sum_{i=1}^{p} \alpha_i} + \frac{\varepsilon_t}{1 - \sum_{i=1}^{p} \alpha_i B^i} + \frac{\beta_1 \varepsilon_{t-1}}{1 - \sum_{i=1}^{p} \alpha_i B^i} + \frac{\beta_2 \varepsilon_{t-2}}{1 - \sum_{i=1}^{p} \alpha_i B^i} + \ldots . \tag{2.12}$$

Verifica-se, a partir da equação anterior, que as condições de estacionariedade e invertibilidade de um modelo ARMA(p,q) requerem que todas as raízes da equação invertida $(1 - \sum_{i=1}^{p} \alpha_i B^i)$ estejam fora do círculo unitário.

É importante destacar que os modelos ARMA não apresentam especificidade gráfica, ou seja, somente a análise do gráfico da série temporal não é suficiente para identificar totalmente qual o modelo que a representa. Entretanto, a correlação temporal geralmente é o principal guia para sua identificação.

O processo prático de modelagem pelo método de BOX e JENKINS (1976) pode ser implementado em um ciclo de 2 (dois) estágios recursivos:

- **Identificação** - análise exploratória da série temporal (função de autocorrelação e função de autocorrelação parcial).

- **Estimação** - estimação dos parâmetros, análise dos resíduos e critérios de ajuste (AIC e BIC).

Este ciclo, com estágios recursivos, pode ser visualizado com maior facilidade e clareza por meio do fluxograma apresentado na Figura 2.1.

Nesta figura, são observadas as seguintes fases: teste da não estacionariedade na média; teste de dependência; teste de linearidade; escolha (identificação) do modelo; estimação dos parâmetros. Após o ajuste do modelo, é verificada a dependência (correlação) entre os resíduos. Caso o modelo passe por esta bateria de testes estará com grandes possibilidades de obter boas previsões.

Na prática, não se conhece exatamente a média, a variância e as autocorrelações da série no tempo, mas quando se trata de uma série estacionária pode-se utilizar os respectivos valores amostrados. Para testar se as autocorrelações são significativamente maiores que zero pode-se utilizar a estatística Q [ENDERS, 2004] dada por

$$Q = N(N+2)\sum_{k=1}^{s} R_y(t,k)^2/(N-k). \tag{2.13}$$

em que N é o número de amostras e s é a defasagem que também é utilizada para estimar os graus de liberdade.

Caso o valor amostrado de Q exceda o valor crítico de χ^2 com s graus de liberdade, então pelo menos um valor de $R_y(t,k)$ é estatisticamente diferente de zero para um nível de significância especificado. Este teste serve também para testar se os resíduos de um modelo ARMA(p,q) formam uma seqüência que pode ser considerada como um ruído branco, com o cuidado de considerar $s-p-q$ graus de liberdade. Caso haja mais uma constante no modelo deve-se considerar $s-p-q-1$.

Normalmente ajusta-se mais de um modelo e os erros de previsão destes modelos são comparados aos pares, chegando-se ao modelo com melhor desempenho em previsões fora da amostra. Geralmente os erros dos dois modelos (ε_{1t}, ε_{2t}) comparados são altamente correlacionados. Por exemplo, uma realização negativa de ε_{t+1} tenderá tornar a previsão dos dois modelos muito alta. Infelizmente, a violação da hipótese de não correlação entre os resíduos invalida a utilização da estatística F, para T previsões, dada por

$$F = \sum_{i=1}^{T} \varepsilon_{1i}^2 / \sum_{i=1}^{T} \varepsilon_{2i}^2 \tag{2.14}$$

ou seja, não se pode garantir que haja uma distribuição F. Observa-se que a razão F deriva da divisão entre os erros quadrados médios de previsão fora da amostra.

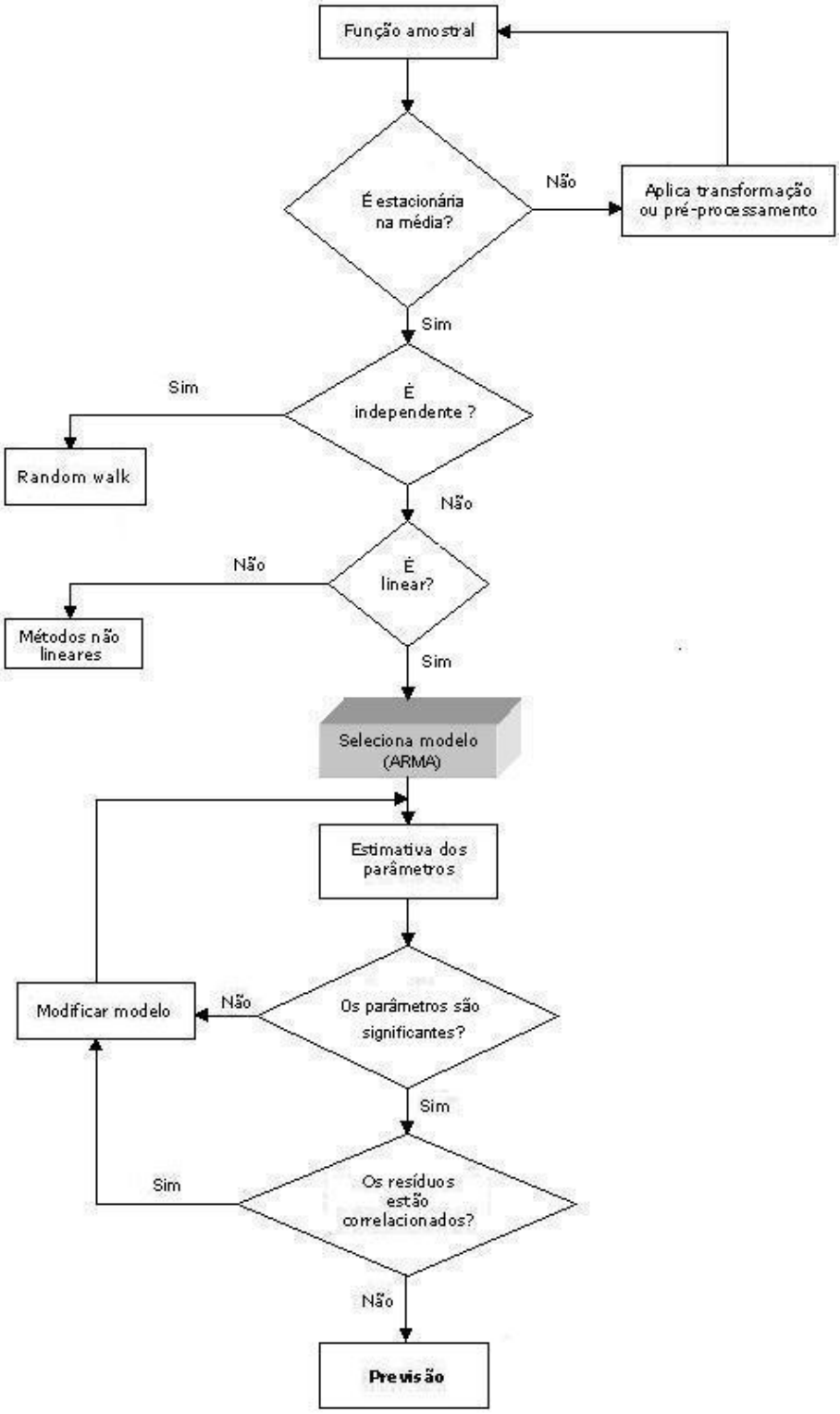

Fig. 2.1: Fluxograma para ajuste do modelo

Para superar este problema, GRANGER and NEWBOLD (1974) e DIEBOLD and MARIANO (1995) criaram outros testes, descritos em [ENDERS, 2004], para se obter a estatística t e verificar se o teste é estatisticamente significante e, conseqüentemente, se os modelos são estatisticamente diferentes ou não. Entretanto, este teste pode ser sensível à escolha da ordem q da variância e resultar em valores negativos de variâncias. Posteriormente, vários estudos surgiram propondo extensões ao teste apresentado na referência [DIEBOLD and MARIANO, 1995], como o trabalho de [NEWEY and WEST, 1987]. Este teste e as suas extensões serão apresentadas e discutidas em detalhes no Capítulo 5.

Um modelo ARMA pode lidar simultaneamente com tendência, sazonalidade e estrutura de curta dependência. No caso de um sistema não estacionário que pode se tornar homogêneo após uma transformação, utiliza-se o modelo ARIMA (autoregressivo (AR), integrado (I) e média móvel (MA)). Este modelo não necessariamente é estacionário na variância e na covariância, mesmo depois da série ser transformada, mas é necessário que o seja na média. Quando se trata de componente sazonal, utiliza-se um modelo SARIMA dado por uma composição de uma parte não sazonal ARIMA(p,d,q) e outra parte sazonal ARIMA(P,D,Q), que pode ser aditiva ou multiplicativa. Na longa dependência é mais indicado usar um modelo ARFIMA (ARIMA fracionário).

Nas séries temporais com componentes sazonais (não estacionariedades sazonais) pode-se também tentar torná-las estacionárias por meio de uma transformação nos dados, definindo-se uma nova variável:

$$z = \frac{(y_t - \bar{y})}{S} \tag{2.15}$$

em que $\bar{y}$ e S são respectivamente a média e o desvio padrão amostrados. A estacionariedade é induzida pela equação anterior, ou seja, a distribuição de probabilidade $p(y_t)$ se reduz a $p(z)$, uma vez que a distribuição de probabilidade para esta nova variável é a mesma para todo tempo t. Considerando que o processo é discreto, a forma da distribuição de probabilidade, p(z), pode ser descrita de acordo com os dados observados $\{z_1; z_2; \ldots; z_N\}$. Esta transformação é aplicada para que a classe de modelos lineares e estacionários possam ser empregados, uma vez que já existem testes estatísticos bem fundamentados nesta área.

Os modelos ARMA descrevem séries caracterizadas pela independência ou quase independência entre observações distantes no tempo. Os modelos ARIMA, por sua vez, descrevem séries em que as correlações tem um decaimento bastante lento, mas quando derivadas apresentam curta dependência. Ás vezes, a função de correlação não apresenta um decaimento tão lento quanto o do modelo ARIMA e nem tão rápido como o do modelo ARMA. Neste caso, ocorre o que se chama de efeito de longa dependência. Os modelos ARFIMA (*autoregressive fracionary integrated moving average*) foram criados para lidar com este tipo de problema e têm profunda associação com sistemas caóticos.

A característica que tornou o modelo ARMA(p,q) bastante popular consiste no fato de ele poder descrever uma série estacionária por meio de um modelo que envolve menos parâmetros que um AR ou um MA puro. A metodologia ARMA de predição tem sido bastante utilizada em estudos econométricos, principalmente, porque vários resultados expressivos foram alcançados nesta área, incluindo as propriedades dos estimadores de amostras finitas. Acredita-se que, entre as principais vantagens desta classe de modelos, estão as seguintes: é conceitualmente

sólido; o método de estimativa dos parâmetros permite calcular o erro associado e estabelecer intervalos de confiança; permite estabelecer relações causais considerando o tempo. Entre as desvantagens, destacam-se as seguintes: requer muita experiência na modelagem; normalmente só apresenta bom desempenho para processos lineares, estacionários e com curta dependência; logo, não consegue captar assimetrias, conglomerados de valores extremos, longa dependência e outras peculiaridades que podem ocorrer nas séries no tempo.

2.3.2 Modelos GARCH

O fato estilizado que consiste na presença de conglomerados de valores extremos em séries no tempo foi primeiramente comentado em [MANDELBROT, 1963]. Posteriormente, este efeito também foi observado em várias séries temporais financeiras, como nas variações diárias do índice composto da NYSE (*New York Stock Exchange*). Este fenômeno influi na variância condicional da série temporal.

Em 1982, em estudos sobre a antecipação racional da inflação do Reino Unido, ENGLE analisava o comportamento da variância desta série e constatou a presença de heteroscedasticidade não condicional. Sugeriu no seu artigo seminal [ENGLE, 1982] uma classe de modelos para expressar este fenômeno formulado a partir do modelo autoregressivo dado pela equação (2.11) com y_t dado por

$$y_t = \beta x_t + \varepsilon_t \tag{2.16}$$

em que ε_t é um ruído branco que tem correlação serial entre os quadrados de seus elementos. O valor de ε_t foi definido como

$$\varepsilon_t = v_t\sqrt{h_t} \tag{2.17}$$

em que v_t e h_t são independentes com $v_t \sim i.i.d.(0,1)$. O modelo ARCH(m) foi concebido por ENGLE como aquele em que h_t é dado por

$$h_t = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\varepsilon_{t-2}^2 + ... + \alpha_m\varepsilon_{t-m}^2$$

em que as variâncias devem ser positivas e finitas, o que é satisfeito a partir da seguinte condição: $\alpha_0 > 0; \alpha_1, \alpha_2, ..., \alpha_m \geq 0$; $\sum_{i=1}^{m} \alpha_i < 1$. Este modelo não inclui o erro, logo é determinista.

Engle também criou o teste ARCH-LM, baseado na autocorrelação, a partir da equação apresentada a seguir

$$\varepsilon_t^2 = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \alpha_2\varepsilon_{t-2}^2 + ... + \alpha_m\varepsilon_{t-m}^2 + \nu_t.$$

com a hipótese nula

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_m = 0$$

Caso a hipótese nula seja verdadeira, logo $E[\varepsilon_t^2] = \alpha_0$ e não existe heteroscedasticidade condicional. Caso contrário, estaria confirmada a ocorrência do efeito que foi denominado de *autoregressive conditional heteroskedasticity* (ARCH) e a variância condicional poderia ser expressa como função dos choques aleatórios dos m instantes imediatamente anteriores.

O modelo de ENGLE(1982) torna possível explicar a tendência de agrupamentos, nos quais ocorre a persistência dos valores da volatilidade nas séries de altas freqüências. Ou seja, valores altos de volatilidades são seguidos por outros valores altos de volatilidade e valores baixos de volatilidade são seguidos por outros valores baixos de volatilidades, formando agrupamentos, ora de altas, ora de baixas volatilidades. Ressalte-se que ENGLE não utilizou os valores passados da variância, somente os valores dos quadrados dos resíduos passados. Também é bom lembrar que, nos modelos ARCH(m), ε_t não é um processo autocorrelacionado, mas seus valores não são independentes porque os seus segundos momentos estão relacionados.

Uma extensão aos modelos ARCH foi a sua generalização proposta em [BOLLERSLEV, 1986], dando origem à classe de modelos (GARCH) *generalized autoregressive conditional heteroskedasticity*, parcimoniosos, que se tornaram os modelos mais utilizados da família ARCH. BOLLERSLEV contornou a limitação do modelo MA de ENGLE, incluindo no novo modelo os valores passados da variância, ou seja, uma variância condicional que pode ser representada por um modelo ARMA. O modelo autoregressivo com heterocedasticidade condicional generalizado - GARCH(q,p), dado na sua forma multiplicativa, foi definido como

$$\varepsilon_t = (\sqrt{\sigma_t^2})\nu_t$$

fazendo

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \tag{2.18}$$

também com a suposição que a variável aleatória ν_t possui média zero, variância 1 e valores não correlacionados. Observando a equação anterior é possível notar que se trata de um processo ARMA denominado por GARCH(q,p) e esta expressão pode englobar um ARCH(1) que é igual a um GARCH(1,0). Entretanto, é necessário que as variâncias sejam estacionárias, o que é satisfeito a partir da seguinte condição:

$$\sum_{t=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1. \tag{2.19}$$

Também, se faz necessário que h_t seja positivo, ou seja

$$\alpha_0 > 0, \alpha_i \geq 0 \;\; e \;\; \beta_j \geq 0.$$

Decorre deste modelo que a variância condicional depende do choque aleatório ocorrido no instante $t-i$. Caso ε_{t-i} for grande (pequeno), h_t também será grande (pequeno). Deste modo, o modelo pode captar o agrupamento das volatilidades, que por sua vez caracteriza que existe alguma dependência temporal no processo estocástico.

Em [ENGLE et al., 1987] foi sugerido que a variância condicional das séries financeiras afetava a média, dando origem aos modelos ARCH-M. Na prática significava que o aumento do risco associado a um título levaria ao aumento do rendimento do mesmo. Matematicamente, para o excesso de rendimento, tem-se

$$y_t = \mu_{yt} + \varepsilon_t \tag{2.20}$$

em que $E(y_t) = \mu_{yt}$ e representa o prêmio de risco.

A idéia básica de ENGLE foi assumir que

$$\mu_{yt} = \beta + \delta h_t \tag{2.21}$$

com $(\delta > 0)$ e h_t representado por

$$h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i h_{t-i}. \tag{2.22}$$

Esta abordagem coloca a volatilidade na média e foi denominada por ENGLE de modelos ARCH-M. Naturalmente h_t pode ter outras representações mais gerais.

A percepção da presença de assimetrias na série de retornos de diversos ativos financeiros, que geralmente tem origem nos impactos das boas e das más notícias, motivou a utilização de modelos não-lineares na variância direcionados para lidar com este tipo de problema. Observa-se que os valores negativos e positivos de mesma magnitude tem o mesmo valor absoluto, mas no mercado financeiro tem efeitos diferentes, ou seja, são assimétricos.

Primeiramente, em [NELSON, 1991] foi proposto um modelo não-linear alternativo para o fenômeno. Este modelo foi denominado de EGARCH (exponential GARCH) que matematicamente é da forma:

$$\log(h_t) = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}/\sqrt{h_{t-i}} + \gamma(|\varepsilon_{t-i}/\sqrt{h_{t-i}}| - E|\varepsilon_{t-i}|) + \sum_{j=1}^{q} \beta_j \log h_{t-j}. \tag{2.23}$$

Outros modelos também tentaram explicar o fenômeno da assimetria, como o TARCH (threshold ARCH), proposto separadamente em [ZAKOIAN, 1990] e [GLOSTEN et al., 1993]. A variância condicional é dada por

$$h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{q} \delta_i \varepsilon_{t-i}^2 d_{t-i} + \sum_{j=1}^{p} \beta_j h_{t-j} \tag{2.24}$$

em que $d_t = 1$ se $\varepsilon_t < 0$, caso contrário, então, $d_t = 0$. Assim, há uma distinção entre choques positivos e negativos, ocasionando efeitos diferentes sobre a variância condicional, ou seja, os choques positivos têm os impactos ponderados pelos $\alpha's$, enquanto os choques negativos são ponderados pelas somas de $\alpha's + \beta's$. Caso $\delta > 0$, os choques negativos terão maior impacto sobre a variância condicional.

Finalmente, como os investidores em ativos financeiros estão interessados na volatilidade futura dos retornos durante o período de aplicação de cada ativo e não na volatilidade histórica, e a variância dos erros pode ser interpretada como a incerteza associada aos valores médios dos retornos, o modelo ARCH e aqueles derivados dele passaram a ser muito utilizados pelo mercado financeiro para estimar a incerteza dos retornos dos ativos. Obviamente, esta classe de modelos pode ser utilizada para séries no tempo de outras áreas (biologia, engenharia, física etc) que apresentem este mesmo fato estilizado.

2.3.3 Os filtros lineares adaptativos

A teoria de filtragem linear já se encontra bem estabelecida e tem sido aplicada com sucesso em campos tão diversos como sistemas de comunicações, sistemas de controle, radar, sonar, antenas e outros [HAYKIN, 1989, HAYES, 1996]. O método tradicional para se projetar este tipo de filtro é ajustar os seus parâmetros (pesos) por meio de uma função objetivo que utiliza o erro quadrático ou o erro quadrático médio.

Um modelo puramente autoregressivo de ordem p, AR(p), ocorre quando se representa uma série temporal na forma $y_{t+1} = \sum_{i=0}^{p-1} w_i y_{t-i} + \varepsilon_t$, em que ε_t é um ruído branco. Assim, a série y_t é escrita a partir dos seus valores passados, supondo que a variável aleatória y_t é linearmente correlacionada com seus próprios valores defasados. Um parâmetro crítico e difícil de estimar é a ordem p do modelo.

A teoria aplicada no ajuste dos pesos dos filtros lineares adaptativos pode ser utilizada para incluir a variável tempo nas redes neurais tipo PMC e RBF, fazendo parte do contexto desta tese. As maiores susceptibilidades destes tipos de filtro estão relacionadas com as não linearidades e as não estacionariedades presentes em alguns processos geradores das séries no tempo. Logo, um caminho natural é a utilização de filtros não lineares (redes neurais artificiais) para lidar com o problema das não linearidades.

Estes filtros são bastante flexíveis e capazes de filtrar ruídos relativamente pequenos, e também podem ser utilizados associados a outros recursos, como por exemplo o filtro de KALMAN. Eles operam uma transformação linear sobre a entrada e, no caso de processos não estacionários, aplica-se o operador diferença até tornar a série temporal estacionária, resultando num modelo autoregressivo integrado (ARI). A Figura 2.2 faz um tipo de representação destes filtros, no instante n, possibilitando a compactação das equações que se seguem, em que $\mathbf{x}$ é o vetor de entradas com defasagem p no instante n, $y(n+1)$ é o valor observado no instante $n+1$, $\widehat{y}(n+1)$ é o valor estimado um passo à frente, $e(n)$ é o erro e J é a função objetivo. O filtro FIR foi projetado originalmente para não ter *bias*.

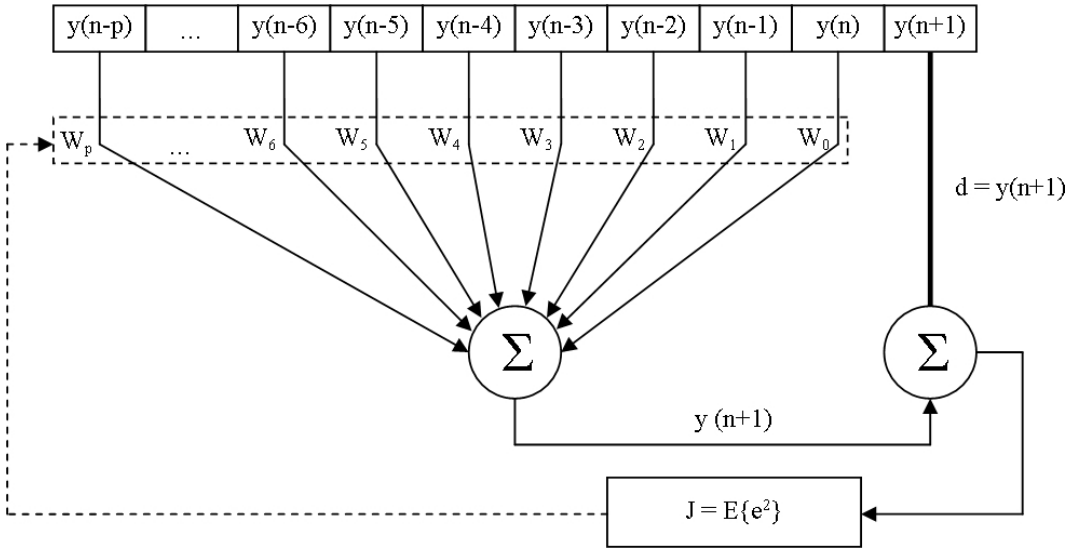

Fig. 2.2: Filtro Linear Adaptativo

Quando se trata de uma tendência estocástica e/ou de uma mudança estrutural, caracterizando uma não estacionariedade mais difícil de ser trabalhada por meio de uma transformação ou pré-processamento dos dados de entrada, a capacidade de adaptação destes filtros torna-se limitada. Entretanto, a capacidade de adaptabilidade deste tipo de filtro a cada nova informação torna-o uma ferramenta útil no processamento adaptativo de sinais e no controle adaptativo. A aplicação prática destes filtros se torna mais adequada em sistemas com ruídos, mas com forte relação causal, como em controles, radares, antenas, sonares etc.

Esta capacidade relativa de adaptabilidade pode gerar instabilidades. Após ser ajustado, o filtro deve ser capaz de ignorar perturbações espúrias sem deixar de responder às mudanças significativas no processo e, ainda, manter-se estável. É o dilema da estabilidade abordado em [GROSSBERG, 1982].

Caso o sinal y_t seja transformado em x_t, o grafo do fluxo de sinal do filtro linear adaptavivo poderá ser representado pela Figura 2.3, que será utilizada para analisar os efeitos da realimentação na estabilidade do filtro.

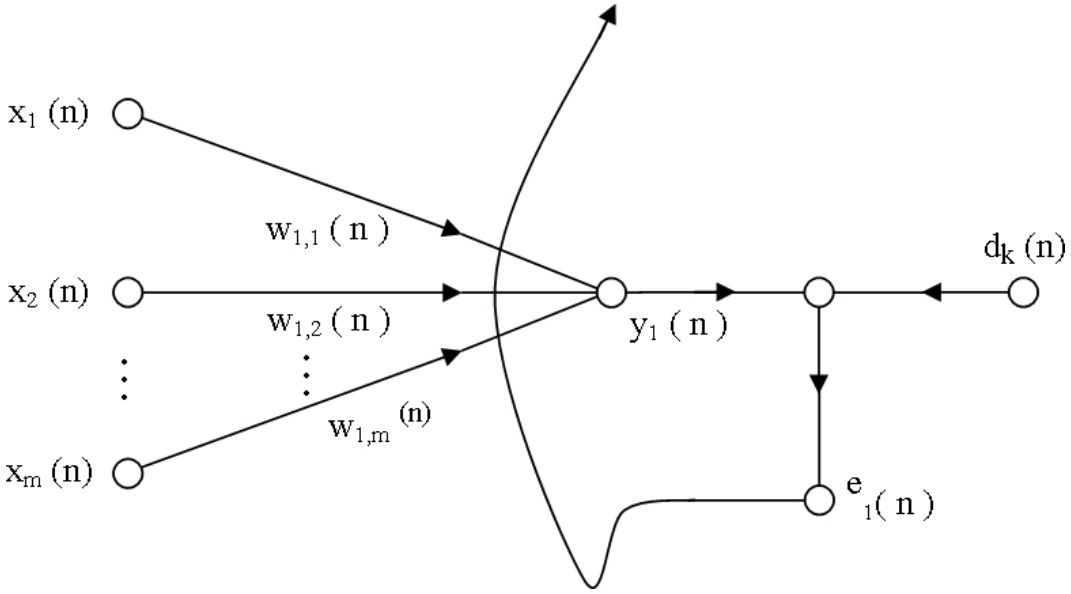

Fig. 2.3: Grafo do fluxo de sinal do filtro linear adaptavivo

A partir da Figura 2.3 chega-se à Figura 2.4 que representa um sistema realimentado de laço único por meio de um grafo de fluxo de sinal que será utilizado para analisar os detalhes dos efeitos da adaptação dos pesos por meio do erro de um determinado elemento. A adaptação dos pesos é realizada por meio da realimentação do erro que, em sistemas dinâmicos, consiste naquela parcela de influência exercida pela saída sobre a entrada de um determinado elemento. A realimentação, como será visto mais adiante, exerce uma influência importante na estabilidade do filtro.

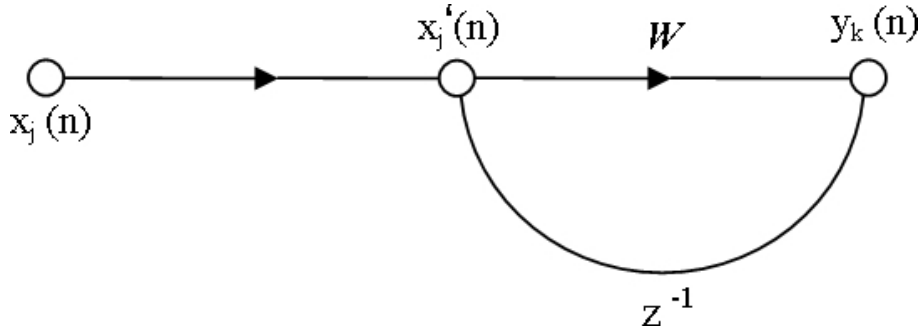

Fig. 2.4: Grafo do fluxo de sinal de um sistema realimentado com laço único

A partir da figura anterior chega-se facilmente às seguintes relações entrada-saída:

$$y_k(n) = w[x'_j(n)] \tag{2.25}$$

em que

$$x'_j(n) = x_j(n) + z^{-1}[y_k(n)] \tag{2.26}$$

em que z^{-1} é um operador de atraso unitário. Substituindo a equação (2.26) em (2.25) e desenvolvendo, chega-se à seguinte relação final de entrada e saída:

$$y_k(n) = \frac{w}{1 - wz^{-1}}[x_j(n)]. \tag{2.27}$$

Utilizando a expansão polinomial em $(1 - wz^{-1})^{-1}$, obtém-se:

$$\frac{w}{1 - wz^{-1}} = w \sum_{l=0}^{\infty} w^l z^{-1}. \tag{2.28}$$

Substituindo a equação (2.28) em (2.27), resulta:

$$y_k(n) = w \sum_{l=0}^{\infty} w^l z^{-1}[x_j(n)] = \sum_{l=0}^{\infty} w^{l+1}[x_j(n-l)] \tag{2.29}$$

em que se pode perceber que o comportamento do sistema é controlado pelo peso w e pelo valor da entrada. Distinguem-se os seguintes casos específicos:

1. $|w| < 1$, torna o sinal de saída exponencialmente convergente, isto é, o sistema é estável.
2. $|w| > 1$, torna o sinal de saída exponencialmente divergente, isto é, o sistema é instável. Caso $|w| = 1$, a divergência é linear.

Logo, a estabilidade tem destaque no estudo de sistemas realimentados e a atualização dos pesos influi na estabilidade do sistema. A memória para $|w| < 1$, embora seja infinita, é esvaecente já que a influência de uma amostra passada se reduz exponencialmente com o tempo n.

Quando se trata de um processo ergódico e estacionário, o filtro original de Wiener (WIENER, 1949) pode ser aplicado e as médias das amostras de longo prazo podem ser substituídas por expectativas (operador *esperança* $E\{.\}$). Para este caso, utilizando a abordagem vetorial (geométrica) de minimização do vetor erro, em que o vetor de erro mínimo deve ser ortogonal ao vetor da entrada, resulta a seguinte equação

$$E\{[d(n) - \sum_{k=0}^{M-1} w^{(k)} x(n-k)]x(n-i)\} = 0 \tag{2.30}$$

em que $k = 0, 1, 2, ..., p-1$ e $i = 0, 1, 2, ..., p-1$. Distribuindo os produtos e re-arranjando, tem-se:

$$\sum_{k=0}^{M-1} w^{(k)} E\{x(n-k)x(n-i)\} = E\{d(n)x(n-i)\}. \tag{2.31}$$

O lado esquerdo da igualdade anterior expressa a função de autocorrelação do processo estocástico da entrada (AR(p)), e o lado direito a função de correlação cruzada entre o processo estocástico que descreve a saída desejada $d(n) = x(n+1)$ e o processo estocástico apresentado à entrada. Para melhor ilustrar o significado das funções de autocorrelação e correlação cruzada, considere-se um exemplo para $M = 3$ em que o processo U é representado por $\mathbf{u}(n) = \begin{bmatrix} u(n) & u(n-1) & u(n-2) \end{bmatrix}^T$. A matriz de autocorrelação $\mathbf{R}$ pode ser dada por:

$$\mathbf{R} = E\left[\ \mathbf{u}(n)\mathbf{u}(n)^T\ \right]\} =$$

$$= E\{\begin{bmatrix} u(n) \\ u(n-1) \\ u(n-2) \end{bmatrix} \left[\ u(n) \quad u(n-1) \quad u(n-2)\ \right]\} =$$

$$= \begin{bmatrix} R_{uu}(0) & R_{uu}(-1) & R_{uu}(-2) \\ R_{uu}(1) & R_{uu}(0) & R_{uu}(-1) \\ R_{uu}(2) & R_{uu}(1) & R_{uu}(0) \end{bmatrix}. \tag{2.32}$$

Como $R_{uu}(x) = R_{uu}(-x)$, R pode ser expressa por:

$$\mathbf{R} = \begin{bmatrix} R_{uu}(0) & R_{uu}(1) & R_{uu}(2) \\ R_{uu}(1) & R_{uu}(0) & R_{uu}(1) \\ R_{uu}(2) & R_{uu}(1) & R_{uu}(0) \end{bmatrix}. \tag{2.33}$$

A correlação cruzada pode ser definida por um vetor p dado por:

$$\mathbf{p} = E\{\mathbf{d}(n)\mathbf{u}(n)\}. \tag{2.34}$$

$$= \left[\ E\{d(n)u(n)\} \quad E\{d(n)u(n-1)\} \quad ... \quad E\{d(n)u(n-M+1)\}\ \right] =$$

$$= \left[\ P(0) \quad P(-1) \quad ... \quad P(1-M)\ \right]^T.$$

Seja o vetor de pesos dado por:

$$\mathbf{w} = \left[\ w_0 \quad w_1 \quad ... \quad w_{M-1}\ \right]^T.$$

Assim, tem-se que:

$$\mathbf{R}\mathbf{w} = \mathbf{p}. \tag{2.35}$$

Pode-se então deduzir a equação que fornece o vetor peso, associado ao erro quadrado mínimo, apresentada a seguir:

$$\mathbf{w} = \mathbf{R}^{-1}\mathbf{p}. \tag{2.36}$$

Esta equação ficou conhecida como a equação de WIENER-HOPF (WIENER, 1949), em homenagem a NORBERT WIENER. Quando a matriz de correlação $\mathbf{R}$ é não-singular para um determinado M, então existe uma solução. Sendo o processo linear e estacionário, a precisão com que $\mathbf{R}$ e $\mathbf{p}$ representam as correlações envolvidas será tanto maior quanto maior for N_t com relação a M. Assim, quando as funções de autocorrelações não são conhecidas, o operador $E\{.\}$ pode ser substituído pela média dos vetores das M componentes envolvidas no cômputo de $\mathbf{R}$ e $\mathbf{p}$, média esta realizada sobre o intervalo de N_t amostras totais conhecidas da série temporal. Entretanto, o número de amostras por intervalos nem sempre é suficiente para expressar com fidelidade o comportamento do sistema.

O algoritmo de LEVINSON-DURBIN, inicialmente, foi muito utilizado como algoritmo de inversão de $\mathbf{R}$. Nos dias atuais, geralmente a inversão é implementada a partir da pseudo-inversão de MOORE-PENROSE, utilizando a técnica *Singular Value Decomposition* (SVD),

muito estável do ponto de vista numérico e freqüentemente adequada para contornar o fato de que a matriz $\mathbf{R}$, muitas vezes, é quase singular.

No contexto da não estacionariedade, pode-se assumir que $\mathbf{w}(n) = \mathbf{R}(n)^{-1}\mathbf{p}(n)$, ou seja, com $\mathbf{w}$ variando com n, não tendo valor fixo como na equação original de WIENER-HOPF, criada no contexto de processos fracamente estacionários. Esta abordagem pode se tornar impraticável para muitas aplicações reais. Para contornar este problema, WIDROW e HOFF (1960) criaram a regra delta, e o fizeram para formular o elemento linear adaptativo (Adaline - Adaptive Linear Element). Este tipo de filtro se tornou o carro chefe da filtragem linear adaptativa e fonte de inspiração para as redes neurais artificiais.

A maneira como o sinal do erro é utilizado para controlar o ajuste dos pesos é determinada pela função de custo utilizada para derivar o algoritmo de filtragem adaptativa de interesse e está intimamente ligada ao método de otimização utilizado. Como exemplos de métodos de otimização, pode-se citar algumas técnicas de otimização irrestritas clássicas relacionadas com o assunto: descida mais íngreme (gradiente), Newton e Gauss-Newton. No desenvolvimento que será apresentado a seguir, aplica-se o método do gradiente (descida mais íngreme). Este algoritmo faz o ajuste (adaptação) dos pesos do filtro usando a minimização do erro quadrático médio (MSE, do inglês *mean square error*).

A função objetivo baseada no MSE tem a forma quadrática e garante um ponto de mínimo. Logo, o vetor de pesos $\mathbf{w}^*$, associado ao ponto de mínimo, existe e a soma do erro quadrático (ξ) sobre todo o conjunto de entradas é mínima e possível de ser determinada. Matematicamente, tem-se que $\exists\xi \mid \xi(\mathbf{w}^*) \leq \xi(\mathbf{w}), \forall\mathbf{w} \epsilon R$.

O erro de predição $e(n)$ pode ser expresso por:

$$e(n) = d(n) - y(n). \tag{2.37}$$

A função objetivo, ou seja, a função soma dos erros quadráticos pode ser convenientemente definida por:

$$\xi(\mathbf{w}) = E[|e(n)|^2]. \tag{2.38}$$

A minimização do erro pode ser feita com a aplicação do vetor gradiente em $\xi(\mathbf{w})$, dado por:

$$\nabla\xi(\mathbf{w}) = \nabla E[|e(n)|^2] = E[\nabla|e(n)|^2] = E[e(n)\nabla e^*(n)]$$

e

$$\nabla e^*(n) = -\mathbf{x}^*(n).$$

Resultando em:

$$\nabla\xi(\mathbf{n}) = -E[e(n)\mathbf{x}^*(n)].$$

Como a adaptação (ajuste) dos pesos deve ser efetuada na direção oposta ao gradiente, tem-se:

$$\triangle\mathbf{w} = \eta E[e(n)\mathbf{x}^*(n)]. \tag{2.39}$$

Utilizando a equação de atualização dos pesos, tem-se:

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \eta E[e(n)\mathbf{x}^*(n)]. \tag{2.40}$$

Assim, pode-se atualizar $\mathbf{w}$ discretamente após a apresentação de cada época de treinamento. O parâmetro η é a taxa de aprendizagem. No caso do filtro linear adaptativo, via gradiente descendente, conforme pode ser observado na Figura 2.2, a amostra predita $\hat{y}(n+1)$ é dada por:

$$\hat{y}(n+1) = y(n) = \sum_{k=0}^{M-1} w^k(n)y(n-k) = \mathbf{w}^T(n)\mathbf{x}(n). \tag{2.41}$$

Foram realizadas operações de filtragem (comparação do sinal observado com o estimado) e de adaptação (ajuste dos pesos em função do sinal do erro). Para um processo fracamente estacionário, este algoritmo converge para a equação de WIENER-HOPF, quando a taxa de amostragem $\eta < 2/\lambda_{max}$, em que λ_{max} [HAYES, 1996] é o maior autovalor da matriz de autocorrelação $\mathbf{R}$.

O filtro de Wiener pode ser visto como um modelo linear AR puro, adequado a sistemas lineares fracamente estacionários ou àqueles sistemas que podem ser considerados estacionários em um determinado intervalo de tempo considerado suficientemente grande. Assumir esta hipótese tem desvantagens: sistemas com variações rápidas podem gerar intervalos relativamente pequenos para estimar apropriadamente os parâmetros do filtro; não absorvem facilmente mudanças estruturais de nível; impõe um modelo estacionário para dados que derivam de um processo não estacionário.

Este método é pouco utilizado em aplicações práticas embora tenha significado muito para o desenvolvimento teórico desta área. A razão para isto é que para computar o vetor gradiente é necessário que $E[e(n)\mathbf{x}^*(n)]$ seja conhecida. Assim as matrizes de autocorrelação de $\mathbf{x}(n)$ e correlação cruzada de $d(n)$ e $\mathbf{x}(n)$ devem ser conhecidas e nem sempre isto acontece, sendo necessário estimá-las a partir dos dados disponíveis.

O algoritmo LMS (*least mean square*) surgiu na esteira deste problema [WIDROW, 1976] calculando os valores da função de custo a partir de amostras disponíveis no momento, ou seja

$$\xi(\mathbf{w}) = \frac{1}{2}e^2(n) \tag{2.42}$$

em que $e(n)$ é o erro no instante n. A esperança $E[|e(n)|^2]$ do método do gradiente descendente é substituída pelo erro quadrado instantâneo $|e(n)|^2$. Diferenciando $\xi(\mathbf{w})$ em relação ao vetor peso $\mathbf{w}$, resulta:

$$\frac{\partial \xi(\mathbf{w})}{\partial \mathbf{w}} = e(n)\frac{\partial e(n)}{\partial \mathbf{w}}. \tag{2.43}$$

Como o algoritmo LMS opera associado a um neurônio linear, pode-se expressar o sinal de erro por meio da seguinte equação:

$$e(n) = d(n) - \mathbf{x}^T(n)\mathbf{w}(n). \tag{2.44}$$

Derivando a equação anterior, tem-se:

$$\frac{\partial e(n)}{\partial \mathbf{w}(n)} = -\mathbf{x}(n). \tag{2.45}$$

Substituindo (2.45) em (2.43), chega-se à equação:

$$\frac{\partial \xi(\mathbf{w})}{\partial \mathbf{w}(n)} = -\mathbf{x}(n)e(n). \tag{2.46}$$

Utilizando a equação anterior como o valor do vetor gradiente e substituindo na equação de atualização dos pesos, chega-se ao algoritmo LMS

$$\hat{\mathbf{w}}(n+1) = \hat{\mathbf{w}}(n) + \eta \mathbf{x}(n)e(n) \tag{2.47}$$

em que η é a taxa de amostragem, variando entre 0 e 1. A atualização dos pesos é por meio da realimentação do filtro e influi na estabilidade do sistema. O parâmetros taxa de amostragem η e as entradas influem na atualização dos pesos que por sua vez determinam a estabilidade do sistema.

A realimentação do filtro LMS em torno do vetor de peso estimado $\hat{\mathbf{w}}$, de acordo com a Figura 2.2, pode funcionar como um filtro passa-baixas, deixando passar as freqüências baixas do erro e atenuando as altas freqüências [WIDROW and STERNS, 1985]. Também, o inverso de η é uma medida de memória do algoritmo LMS, ou seja, para valores menores de η um número maior de dados passados será então recordado pelo algoritmo LMS, resultando em uma ação de filtragem mais efetiva.

No algoritmo com gradiente de descida mais íngreme, o vetor peso $\mathbf{w}$ segue uma trajetória bem definida de encontro à solução ótima de Wiener ($\mathbf{w}_0$) para uma determinada taxa η. Por outro lado, no algoritmo LMS, o vetor de peso $\hat{\mathbf{w}}$ traça uma trajetória aleatória. Por esta razão, também é chamado de algoritmo do gradiente estocástico. Assim, o algoritmo pode realizar uma caminhada aleatória para um ponto próximo da solução ótima de Wiener ($\mathbf{w}_0$).

Posteriormente, surgiram outros modelos que derivaram do LMS original. Também foi desenvolvido o filtro recursivo, que geralmente converge mais rápido que o filtro LMS. O filtro recursivo tem desempenho de previsão semelhante ao LMS, mas é computacionalmente mais complexo e numericamente mais instável [HAYES, 1996].

O conhecimento adquirido com os estudos sobre este tipo de filtro foi importante para o desenvolvimento do processo de aprendizagem de redes neurais artificiais associadas à previsão de séries temporais. Este tipo de rede depende da extração e da transferência eficaz das informações contidas nas amostras para os seus parâmetros livres. Um objetivo natural seria utilizar métodos que manipulassem diretamente a informação e requeressem somente a disponibilidade das informações nos dados, não necessitando de suposições *a priori* sobre as distribuições dos dados. O ponto crucial era encontrar a metodologia apropriada para identificar o potencial de informações contido na série temporal e transferi-las tão eficientemente quanto possível para os parâmetros livres da rede neural artificial.

No próximo capítulo utiliza-se uma aproximação não paramétrica para estimar a entropia, na qual a integração da entropia quadrática de Renyi, da inequação de Cauchy-Schwartz e da janela de Parzen fornecem um método para estimar a informação mútua. Este método é muito prático e ajuda a contornar o problema da estimativa da função de densidade de probabilidade.

Estes valores estimados de informação mútua, juntamente com os conceitos de relevância e de redundância, serão utilizados para implementar uma metodologia para a seleção de variáveis e características.

Capítulo 3

Seleção de Variáveis e Características

3.1 Introdução

Em modelos de predição para mercados financeiros e meteorologia existe um grande número de variáveis exógenas candidatas. Surge, então a questão de se determinar qual o subconjunto de variáveis oferece melhor qualidade de previsão. A seleção de variáveis também é importante para áreas de pesquisas como a modelagem de sistemas complexos, mineração de dados e reconhecimento de padrões. Os conjuntos de dados destas áreas geralmente apresentam alta dimensão e isto cria problemas para os algoritmos de aprendizagem. Logo, a redução da dimensão ou a seleção de um subconjunto de variáveis e características tornam o modelo mais parcimonioso. O algoritmo de aprendizado fica melhor, mais rápido e mais fácil de ser entendido. Também pode evitar o problema de excesso de treinamento (*overfitting*) e lidar melhor com o problema da alta dimensão. Um modelo pode ser considerado em *overfitting* quando o seu erro de treinamento é muito pequeno, mas tem resultados de previsão pobres para dados fora da amostra de treinamento.

No contexto da previsão de séries no tempo, os objetivos principais da seleção de variáveis geralmente são: melhorar a performance de predição dos modelos; diminuir os custos de processamento; prover um melhor entendimento sobre o processo analisado. O objetivo, portanto, é escolher estatisticamente um subconjunto mínimo de variáveis a partir do conjunto original das possíveis variáveis de entrada do processo [KOHAVI and JOHN, 1997]. Todavia, encontrar um subconjunto ótimo é um problema às vezes intratável [GUYON and ELISSEEFF, 2003]. Muitos problemas relacionados com a seleção de variáveis foram caracterizados como de difícil solução [BLUM and LANGLEY, 1997].

Visto de uma maneira não formal, o conceito de causalidade de GRANGER [GRANGER, 1969] para séries temporais baseia-se na capacidade de previsibilidade, ou seja, Y_t é causado por X_t se a previsão de Y_t é melhor quando se utiliza valores de X_t do que somente com valores do próprio Y_t. Este conceito será formalizado na Seção 3.2 e será utilizado para fundamentar o método de seleção de variáveis proposto neste capítulo. Entretanto, é muito difícil se fazer inferência sobre a existência de relações de causalidade. Observa-se que a existência de correlação entre duas variáveis não implica em uma relação de causa e efeito entre as mesmas já que ambas podem estar relacionadas com uma terceira.

Os métodos econométricos criados para as áreas de economia e finanças, baseados em re-

gressão (VAR, VEC-M e outros), utilizando a análise de co-integração e construção de cenários (conhecimento *a priori*), são mais utilizados pelos profissionais destas áreas que os métodos multivariados. Observa-se que os modelos econométricos já dispõem de ferramentas para estimar relevância (teste t, test F, função de verossimilhança e correlação canônica) e redundância (teste de multicolinearidade, coeficientes de correlação, regressão aos pares, análise de componentes principais e análise de fatores).

A metodologia de seleção de variáveis proposta neste capítulo faz a avaliação da relevância e da redundância. O objetivo é contribuir para melhorar os resultados obtidos com modelos de previsão multivariados tipo MISO (*multiple inputs and single output*) e MIMO (*multiple inputs and multiple output*) não baseados em teoria ou naqueles casos em que exista uma teoria, mas o conhecimento a priori sobre o assunto varia no tempo (taxa de câmbio) ou não está disponível em tempo viável (dados intra-diários do mercado financeiro). Também pode ser combinada com métodos baseados em teoria, como os métodos econométricos, aumentando a capacidade de captar mais informações.

Este capítulo foi organizado como segue: a Seção 3.2 apresenta o conceito de causalidade de GRANGER (1969) entre séries temporais e vetores de séries temporais; a Seção 3.3 faz uma introdução aos métodos de seleção de variáveis e características; a Seção 3.4 apresenta uma metodologia para a seleção de variáveis.

3.2 Definições formais de causalidade entre séries temporais

As definições a seguir são baseadas no conceito de causalidade de GRANGER (1969) e apresentados em [CUNHA, 1997]. Estes conceitos são importantes neste contexto porque serão utilizados para dar sustentação teórica à escolha do método de seleção das variáveis e características proposto nesta tese. Os testes de Granger não serão aplicados nesta tese porque são adequados somente para sistemas lineares, mas o conceito pode ser aplicado neste contexto.

Sejam $\{\mathbf{x}_t, t = 0, \pm1, \pm2, ...\}$ e $\{\mathbf{y}_t, t = 0, \pm1, \pm2, ...\}$ processos estocásticos estacionários. É dado que $\overline{\mathbf{x}_t} = \{\mathbf{x}_{t-j}, j = 1, 2, ...\}$ é o conjunto dos valores passados de $\mathbf{x}_t$ e $\overline{\overline{\mathbf{x}_t}} = \{\mathbf{x}_{t-j}, j = 0, 1, 2, ...\}$ é o conjunto dos valores passados e presentes de $\mathbf{x}_t$. As mesmas definições são aplicadas a $\overline{\mathbf{y}_t}$ e $\overline{\overline{\mathbf{y}_t}}$.

Seja $\mathbf{A}_t$ um conjunto de informações que inclui as séries $\mathbf{x}_t$ e $\mathbf{y}_t$, com informações disponíveis até o instante $t = 0$ e $\overline{\mathbf{A}_t}$ até o instante $t = 1$. Seja também $\mathbf{A}_t - \mathbf{x}_t$ o conjunto de informações que exclui as informações de $\mathbf{x}_t$.

Seja também $P(\mathbf{y}_t/\mathbf{A}_t)$ um previsor de mínimos quadrados ótimo não viciado de $\mathbf{y}_t$, usando o conjunto de informação de $\mathbf{A}_t$ [HAMILTON, 1994]. Dado também que o erro de previsão é $\varepsilon(\mathbf{y}_t/\mathbf{A}_t) = \mathbf{y}_t - P(\mathbf{y}_t/\mathbf{A}_t)$ e a variância do erro de previsão é $\sigma^2(\mathbf{y}_t/\mathbf{A}_t)$.

Definição 3.2.1:: Causalidade entre séries temporais
$\mathbf{x}_t$ causa $\mathbf{y}_t$ se

$$\sigma^2(\mathbf{y}_t/\overline{\mathbf{A}}_t) < \sigma^2(\mathbf{y}_t/(\overline{\mathbf{A}}_t - \overline{\mathbf{x}}_t)) \tag{3.1}$$

ou seja, o previsor da série temporal $\mathbf{y}_t$ pode apresentar melhores resultados de previsão usando toda a informação disponível, isto é, utilizando o passado de $\mathbf{y}_t$ e de $\mathbf{x}_t$. Diz-se que $\mathbf{x}_t$ antecede $\mathbf{y}_t$.

Definição 3.2.2: Causalidade instantânea

$\mathbf{x}_t$ **causa** $\mathbf{y}_t$ **instantaneamente** se

$$\sigma^2(\mathbf{y}_t/(\overline{\mathbf{A}}_t, \overline{\overline{\mathbf{x}}}_t)) < \sigma^2(\mathbf{y}_t/\overline{\mathbf{A}}_t). \tag{3.2}$$

O valor presente de $\mathbf{y}_t$ pode ser melhor previsto se o valor presente de $\mathbf{x}_t$ também for utilizado.

Definição 3.2.3: Efeito de retro-alimentação instantâneo

Quando $\mathbf{x}_t$ **causa** $\mathbf{y}_t$ e também $\mathbf{y}_t$ **causa** $\mathbf{x}_t$ **instantaneamente** tem-se o **efeito de retro-alimentação instantâneo**.

Observa-se que as inter-relações entre duas séries no tempo $\mathbf{x}_t$ e $\mathbf{y}_t$ tem três dimensões: $\mathbf{x}_t$ **causa ou não** $\mathbf{y}_t$; $\mathbf{y}_t$ **causa ou não** $\mathbf{x}_t$; ocorre **causalidade instantânea**. Assim, o espaço das relações de causalidade é de oito possibilidades, como ilustrado de forma binária e em notação matemática compacta na Tabela 3.1.

Possibilidade	Notação matemática	Notação binária
$\mathbf{x}_t$ e $\mathbf{y}_t$ são independentes	$\mathbf{x}_t \perp \mathbf{y}_t$	000
causalidade instantânea	$\mathbf{x}_t - \mathbf{y}_t$	001
$\mathbf{x}_t$ causa $\mathbf{y}_t$ somente	$\mathbf{x}_t \rightarrow \mathbf{y}_t$	100
$\mathbf{x}_t$ causa $\mathbf{y}_t$ instantaneamente	$\mathbf{x}_t \Rightarrow \mathbf{y}_t$	101
$\mathbf{y}_t$ causa $\mathbf{x}_t$ somente	$\mathbf{y}_t \rightarrow \mathbf{x}_t$	010
$\mathbf{y}_t$ causa $\mathbf{x}_t$ instantaneamente	$\mathbf{y}_t \Rightarrow \mathbf{x}_t$	011
retro-alimentação, não instantaneamente	$\mathbf{x}_t \leftrightarrow \mathbf{y}_t$	110
retro-alimentação e causalidade instantânea	$\mathbf{x}_t \Leftrightarrow \mathbf{y}_t$	111

Tab. 3.1: Relações de causalidade entre $\mathbf{x}_t$ e $\mathbf{y}_t$

Quando as séries no tempo $\mathbf{x}_t$ e $\mathbf{y}_t$ não são estacionárias aplica-se uma transformação que preserve as relações de causalidade (transformação linear; diferenças simples e sazonais; outras). Caso a série ao ser diferenciada se torne estacionária, diz-se que a mesma é homogênea, caso contrário, é chamada de heterogênea.

Definição 3.2.4: Causalidade entre vetores de séries temporais

Dado que $\{\mathbf{X}_t, t = 0, \pm1, \pm2, ...\}$ e $\{\mathbf{Y}_t, t = 0, \pm1, \pm2, ...\}$ são dois vetores com dimensão b_1 e b_2, respectivamente, de séries no tempo estacionárias de segunda ordem, em que $\mathbf{X}_t = \mathbf{x}_{1t}, \ldots, \mathbf{x}_{b_1 t}$ e $\mathbf{Y}_t = \mathbf{y}_{1t}, \ldots, \mathbf{y}_{b_2 t}$.

Definindo que $\{\mathbf{A}_t, t = 0, \pm1, \pm2, ...\}$ é um conjunto de informação contendo os vetores $\mathbf{X}_t$ e $\mathbf{Y}_t$, e as informações passadas de $\mathbf{A}_t$ como $\overline{\mathbf{A}}_t = \{\mathbf{A}_s : s < t\}$.

Observa-se que para algum conjunto de informação $\mathbf{B}_t$ contido em $\mathbf{A}_t$, o melhor previsor de mínimos quadrados de $\mathbf{y}_{it}$ baseado em $\mathbf{B}_t$ é dado por $P(\mathbf{y}_{it}/\mathbf{B}_t)$, $\varepsilon_{it}(\mathbf{y}_{it}/\mathbf{B}_t) = \mathbf{y}_{it} - P(\mathbf{y}_{it}/\mathbf{B}_t)$ é o erro de previsão correspondente e $\sigma^2_{it}(\mathbf{y}_{it}/\mathbf{B}_t)$ é a variância de ε_{it}.

O previsor $P(\mathbf{y}_{it}/\mathbf{B}_t)$ é uma projeção ortogonal de $\mathbf{y}_{it}$ no espaço gerado por $\mathbf{B}_t$. Quando se trata de um processo gaussiano, $P(\mathbf{y}_{it}/\mathbf{B}_t) = E(\mathbf{y}_{it}/\mathbf{B}_t)$ [BROCKWELL and DAVIS, 1991].

O melhor previsor linear de $\mathbf{Y}_t$ dado $\mathbf{B}_t$ é o vetor

$$P(\mathbf{Y}_t/\mathbf{B}_t) = (P(\mathbf{y}_{1t}/\mathbf{B}_t), \ldots, P(\mathbf{y}_{b_2 t}/\mathbf{B}_t))^T. \tag{3.3}$$

e o vetor erro de previsão é

$$\varepsilon_t(\mathbf{Y}_t/\mathbf{B}_t) = (\varepsilon_{1t}(\mathbf{y}_{1t}/\mathbf{B}_t), \ldots, \varepsilon_{b_2 t}(\mathbf{y}_{b_2 t}/\mathbf{B}_t))^T. \tag{3.4}$$

e a matriz de covariância de ε_t é dada por $\sum(\mathbf{Y}_t/\mathbf{B}_t)$. E seja o conjunto $\overline{\mathbf{B}}_t - \overline{\mathbf{X}}_t$ que representa todas as informações em $\overline{\mathbf{B}}_t$ menos as informações em $\overline{\mathbf{X}}_t$.

As definições que serão apresentadas a seguir para expressar a causalidade entre vetores de séries temporais são extensões simples das noções de causalidade entre duas séries univariadas definidas anteriormente.

Definição 3.2.5: O vetor $\mathbf{X}_t$ **não causa** o vetor $\mathbf{Y}_t$ se

$$\sigma^2(\mathbf{y}_{it}/\overline{\mathbf{A}}_t) = \sigma^2(\mathbf{y}_{it}/(\overline{\mathbf{A}}_t - \overline{\mathbf{X}}_t)), i = 1, \ldots, b_2. \tag{3.5}$$

Logo o vetor $\mathbf{X}_t$ causa o vetor $\mathbf{Y}_t$ se

$$\sigma^2(\mathbf{y}_{it}/\overline{\mathbf{A}}_t) < \sigma^2(\mathbf{y}_{it}/(\overline{\mathbf{A}}_t - \overline{\mathbf{X}}_t)), i = 1, \ldots, b_2 \tag{3.6}$$

para pelo menos um valor de i.

Definição 3.2.6: Outra definição de que o vetor $\mathbf{X}_t$ **não causa** o vetor $\mathbf{Y}_t$ se

$$\sum(\mathbf{Y}_t/\overline{\mathbf{A}}_t) = \sum(\mathbf{Y}_t/(\overline{\mathbf{A}}_t - \overline{\mathbf{X}}_t)). \tag{3.7}$$

Neste contexto, as equações (3.5) e (3.7) são equivalentes e o conceito de causalidade também pode ser expresso em termos de projeções.

$$P(\mathbf{y}_{it}/\overline{\mathbf{A}}_t) = P(\mathbf{y}_{it}/(\overline{\mathbf{A}}_t - \overline{\mathbf{X}}_t)), i = 1, \ldots, b_2 \tag{3.8}$$

ou

$$P(\mathbf{Y}_t/\overline{\mathbf{A}}_t) = P(\mathbf{Y}_t/(\overline{\mathbf{A}}_t - \overline{\mathbf{X}}_t)). \tag{3.9}$$

Pode-se demonstrar que a equação (3.9) implica a equação (3.7) e que esta implica a equação (3.5). Logo as equações (3.5), (3.7) e (3.9) são equivalentes. Os conceitos apresentados nesta seção estão relacionados com as definições associadas ao filtro e ao *wrapper* que serão abordados mais adiante.

3.3 Métodos de seleção de variáveis e características

O aprendizado de máquinas geralmente começa com uma representação apropriada que alcance uma reconstrução melhor dos dados. No contexto da seleção de variáveis e características, existem três abordagens tradicionais para se lidar com este problema de representação: a **transformação de variáveis** que consiste em converter os dados para uma representação de dimensão mais baixa do que a original; a **seleção de variáveis** descarta algumas variáveis originais e se chega a um subconjunto das variáveis originais sem transformar suas coordenadas; a **ponderação de variáveis** que é uma generalização da seleção de variáveis.

Geralmente, o termo **variável** é atribuído às variáveis de entrada com os dados ainda brutos. Enquanto o termo **característica** está relacionado com variáveis construídas (*clustering*, PCA/SVD, Fourier e outras) para serem variáveis de entrada. Entretanto, algumas características resultam do pré-processamento de variáveis brutas explicitamente computadas para serem variáveis de entrada [GUYON and ELISSEEFF, 2003]. A **extração** de características resulta tanto da construção de características como da seleção de variáveis. Como esta tese trata da predição de séries temporais, quando não houver impacto no algoritmo de seleção, o termo variável será utilizado.

Entre os métodos mais conhecidos de transformação de variáveis estão: PCA (*Principle Component Analysis*), ICA (*Independent Component Analysis*) e a análise de fatores (*Factor*

Analysis). Já entre os métodos mais conhecidos de seleção de variáveis estão: o filtro, o *wrapper* e o *embedded* [UNCU and TÜRKSEN, 2006].

No PCA, a variabilidade dos dados pode ser descrita usando um número menor de vetores de base. Este método funciona melhor quando as características são correlacionadas, entretanto, as novas dimensões (componentes principais) serão não correlacionadas. O PCA comprime os dados, tornando as novas características (fatores) não interpretáveis, e não é indicado para aplicações que necessitam que estas características sejam interpretáveis; inclusive, tem limitações em classificações. Observa-se que no PCA, o espaço de transformação das variáveis tem a forma esférica e cada nova variável é o resultado do produto interno da variável original por um autovetor.

A análise de fatores é a generalização do PCA e a principal diferença entre eles é que a análise de fatores permite que se adicione ruídos às variáveis para que o espaço de transformação não tenha a forma esférica. O objetivo principal, tanto do PCA como da análise de fatores, é transformar as coordenadas do sistema tal que a correlação entre as variáveis do sistema seja minimizada.

A utilização do PCA é equivalente a aplicar uma SVD (*Singular Value Decomposition*) nos dados. O PCA descobre as dimensões que são não correlacionadas. A ICA descobre as dimensões que são independentes, que consiste em uma propriedade muito mais forte, podendo ser utilizada para a separação de dois sinais utilizando o fato de que eles realmente são independentes. Nesta tese somente o método de transformação PCA será utilizado. As componentes principais serão estimadas por meio do método da SVD.

Os métodos de seleção de variáveis (filtro, *wrapper* e *embedded*) geralmente são classificados de acordo com a relação entre a estrutura de seleção e o algoritmo de indução. Esta decomposição é para ajudar a comparar diferentes abordagens que podem ser vistas como da mesma categoria. A Figura 3.1 ilustra sucintamente estes métodos de seleção de variáveis.

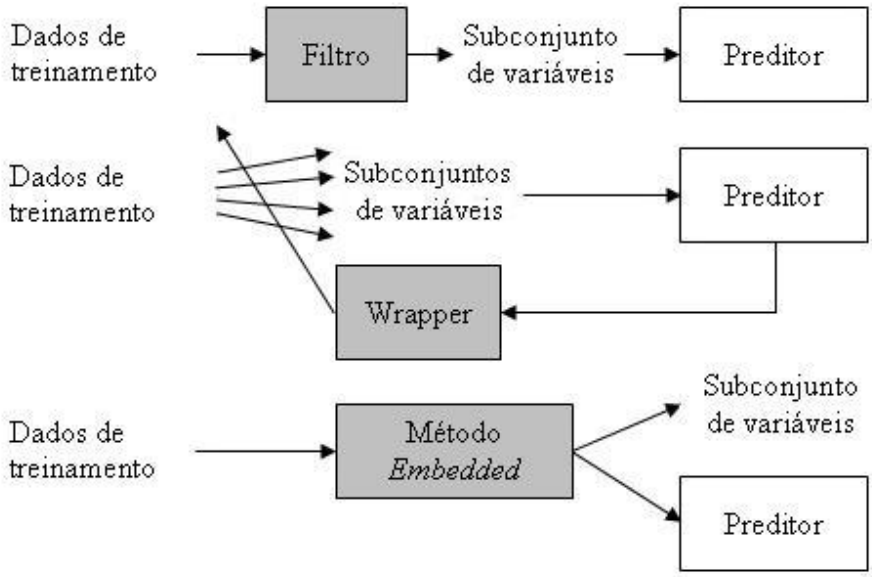

Fig. 3.1: Métodos de seleção de variáveis e características

A estrutura de seleção pode ser subdividida em três passos. O primeiro consiste na **busca** que determina o(s) ponto(s) de partida no espaço de busca, quais influenciam a direção de busca e quais operadores serão utilizados para gerar os estados sucessores. A decisão sobre a organização da busca também é importante; por exemplo, uma busca exaustiva geralmente é impraticável porque o espaço de busca para n variáveis é de $2^n - 1$ subconjuntos de variáveis. O segundo passo consiste na escolha do **critério** de avaliação dos subconjuntos de variáveis; e. g., utilizando o NMSE. O último passo consiste em decidir quando parar a **avaliação** do modelo durante o processo de seleção; e. g., parar de adicionar e retirar variáveis quando nenhuma das alternativas melhora a predição.

O método de seleção de variáveis tipo filtro faz o ordenamento das variáveis individualmente ou em subconjuntos, independentemente do preditor, e geralmente é robusto a *overfitting*, mas falha em encontrar o subconjunto de variáveis mais promissor. O *wrapper* utiliza o preditor para avaliar os subconjuntos e, idealmente encontra o subconjunto de variáveis mais promissor embora seja propenso ao *overfitting*. O *embedded* é similar ao *wrapper*, a diferença está na busca que é guiada pelo processo de aprendizado, tornando-o menos propenso ao *overfitting*, computacionalmente atraente, mas é complexo.

Cada um destes métodos tem os seus respectivos tipos de busca, critério e avaliação dos subconjuntos de variáveis. A Tabela 3.2 ilustra as características principais destes métodos.

Os principais tipos de busca são: exponencial, randômica e seqüencial. As buscas exponenciais podem ser do tipo exaustiva, busca em árvore e outras. Este tipo de busca garante uma solução ótima, mas é a mais complexa computacionalmente. A busca em árvore diminui o tempo de busca por meio da eliminação de alguns ramos da árvore de busca. Para melhorar o tempo de busca, pode-se utilizar heurísticas determinísticas ou randômicas. Uma heurística é dita determinista quando fornece o mesmo resultado em todas as execuções. Já a randômica pode fornecer diferentes resultados para cada semente do algoritmo gerador de números randô-

Tipos	Busca	Critério	Avaliação
Filtro	ordena variá./subconjuntos	mérito variá./subconjuntos	teste estatísticos
wrapper	espaço de subconjuntos	avalia subconjuntos	validação cruzada
embedded	guiada pelo aprendizado	avalia subconjuntos	validação cruzada

Tab. 3.2: Tipos de busca, critério e avaliação do filtro, *wrapper* e *embedded* dos subconjuntos de variáveis

micos. Entre os métodos seqüenciais mais conhecidos estão: busca direta, eliminação para trás, busca bi-direcional, seleção flutuante e a primeira melhor busca. Os algoritmos busca direta e eliminação para trás são largamente utilizados [UNCU and TÜRKSEN, 2006].

O algoritmo de busca direta pode ser sumarizado como: inicializar o conjunto de variáveis significantes com um conjunto vazio. Assim, entre todos os possíveis subconjuntos de variáveis com mais uma variável de entrada, seleciona a combinação de variáveis de entrada que fornece a melhor função de avaliação baseada na função erro. Este processo interativo deve continuar até que a função de avaliação da melhor combinação de variáveis de entrada da interação corrente é pior que a melhor da anterior.

O algoritmo de eliminação para trás pode ser sumarizado como: inicializar o conjunto de variáveis significantes com o conjunto das variáveis de entrada. Assim, entre todos os possíveis subconjuntos de variáveis com menos uma variável de entrada, selecione a combinação de variáveis de entrada que fornece a melhor função de avaliação baseada na função erro. Este processo interativo deve continuar até que todas as variáveis sejam analisadas separadamente ou até que a função de avaliação da melhor combinação de variáveis de entrada atinja um valor ótimo.

Apesar dos métodos seqüenciais serem muito utilizados, computacionalmente atrativos e fáceis de implementar, o seu algoritmo pode ficar preso em mínimos locais. Quando isto ocorre, uma possível solução para este problema é utilizar métodos estocásticos (*simulated annealing*, *genetic algorithms*, *probabilistic hill-climbing* e outros) [BLUM and LANGLEY, 1997]. A principal vantagem destes métodos estocásticos é fugir dos mínimos locais.

Um filtro baseia-se em uma função de mérito. É criado um índice que indica a capacidade de previsão de cada variável, detectando as variáveis com alto ganho, por exemplo, via teoria da informação. Esta abordagem baseada em filtros tradicionais tem alguns problemas, por exemplo: viés de modelos, em que diferentes características sugerem diferentes modelos de indução (filtros lineares, redes neurais etc); características dependentes, que se forem consideradas em conjunto, podem ser redundantes, mas pode ocorrer que uma variável necessite da outra para fornecer uma boa previsão. A Figura 3.2 ilustra com mais detalhes um filtro.

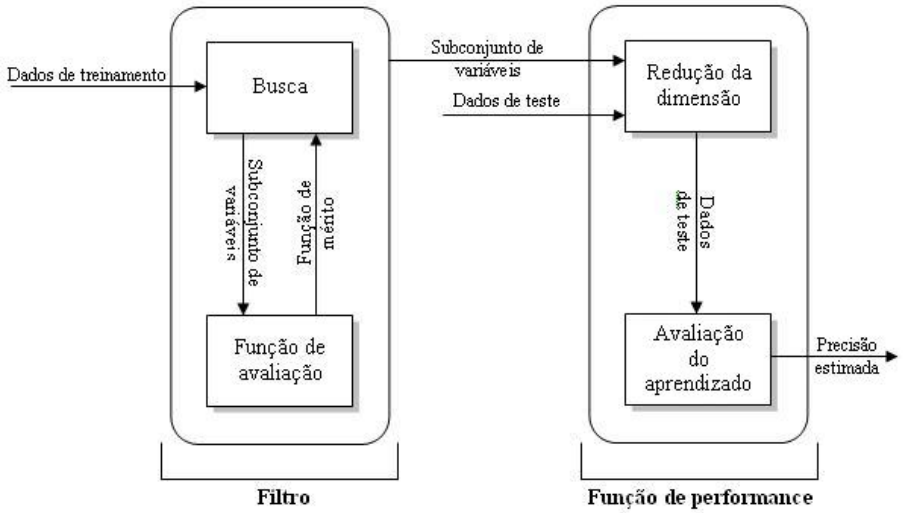

Fig. 3.2: Método de seleção de variáveis e características tipo filtro

A abordagem conhecida como *wrapper* (empacotamento) consiste em utilizar um algoritmo de indução para fazer a avaliação dos subconjuntos de variáveis. As vantagens principais deste método são: levar em conta o viés do algoritmo de indução; considerar as variáveis dentro do contexto. A princípio a busca é exponencial, mas pode-se implementar buscas estocásticas (algoritmos genéticos, *simulated annealing* e outras) ou seqüenciais (busca direta, eliminação para trás e outras). A performance da eliminação para trás é ligeiramente superior à busca direta porque ela considera as variáveis no contexto. A Figura 3.3 ilustra com mais detalhes o método *wrapper*.

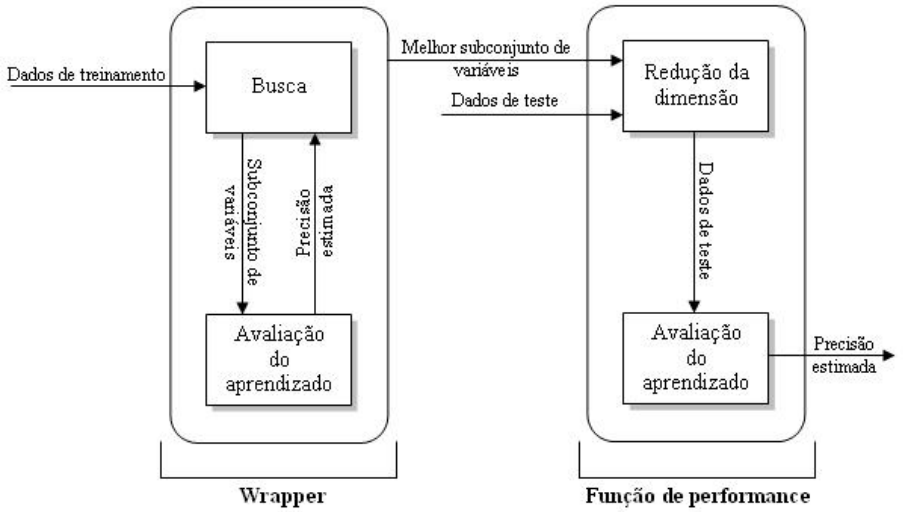

Fig. 3.3: Método de seleção de variáveis e características tipo *wrapper*

A abordagem conhecida como *embedded* consiste em um algoritmo em que o mecanismo de seleção de variáveis fica embutido/encaixado no algoritmo de indução. Geralmente este mecanismo fica ligado a um algoritmo mais complexo. Observa-se que a única diferença entre

este método e o *wrapper* é que neste a seleção de variáveis fica ao lado e no *embedded* fica junto ao algoritmo de indução. Esta abordagem funciona bem quando as variáveis relevantes tem pequenas interações. Entretanto, quando há interações entre as variáveis relevantes, este método tem pouco poder de discriminação entre uma variável relevante e uma pouco relevante isoladamente. Experimentos comprovaram que quando se acrescenta variáveis irrelevantes, este método perde precisão.

Durante o aprendizado supervisionado de máquina, quando se utiliza os métodos *wrapper*, é importante se obter uma boa generalização, utilizando uma criteriosa seleção do melhor modelo. O problema da generalização se torna mais crítico quando os dados são incompletos ou carregam ruídos. Este assunto será abordado mais adiante neste capítulo e em detalhes no Capítulo 5.

3.4 Método proposto para a seleção de variáveis e características

A metodologia de seleção de variáveis proposta utiliza um filtro no primeiro estágio e um *wrapper* no segundo estágio. Inicialmente, o filtro elimina os seguintes tipos de variáveis: as irrelevantes; as pouco relevantes, mas redundantes. Em seguida, o *wrapper* faz a escolha do melhor subconjunto de variáveis, utilizando algoritmos de indução baseados em redes neurais tipo RBF, de acordo com o conceito de causalidade de GRANGER (1969) apresentado anteriormente. Esta seleção de variáveis avalia a contribuição das variáveis de entrada em conjunto para a previsão da variável dependente. A Figura 3.4 ilustra este método.

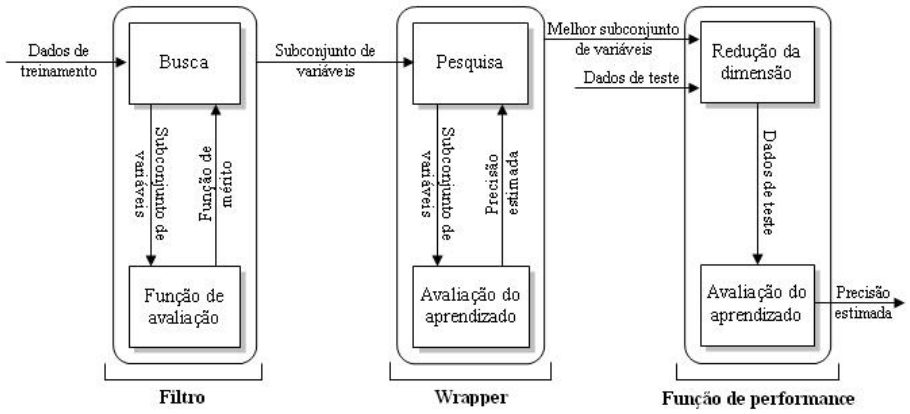

Fig. 3.4: Método de seleção de variáveis e características com filtro e *wrapper*

A seleção de variáveis e características possibilita a transferência eficaz das informações contidas nos dados para o modelo. Na teoria, mais variáveis e características deveriam prover mais poder de discriminação, mas na prática, a excessiva quantidade de variáveis e características não só torna a aprendizagem do processo mais lenta como também causa excessivo custo de processamento e, muitas vezes, confunde a aprendizagem do modelo.

No estágio do filtro faz-se a análise do conceito de relevância em conjunto com o conceito de redundância via teoria da informação, já que a análise de relevância (ordenamento) sozinha

é insuficiente para uma eficiente seleção de variáveis. Esta seleção objetiva contribuir para a melhora dos resultados obtidos com modelos de previsão multivariados não baseados em teoria ou, mesmo com aqueles em que exista uma teoria, mas o conhecimento a priori sobre o assunto varia no tempo ou não está disponível em tempo viável para previsões de curto prazo.

A referência [YU and LIU, 2004] apresentou um método de seleção de variáveis tipo filtro que utiliza somente os conceitos de relevância e redundância aplicados em conjunto. Os conceitos de relevância e redundância são geralmente ligados ao conceito de correlação que no plano linear pode ser estimado via coeficiente de correlação. Duas variáveis são consideradas totalmente redundantes se seus valores forem completamente correlacionados e não existirá relação linear entre elas se o coeficiente de correlação for zero. As definições formais de relevância e redundância serão apresentadas na subseção 3.4.2.

Esta tese apresenta um método para implementar o filtro de YU e LIU (2004) por meio da informação mútua (IM) já que em sistemas não lineares geralmente a relação entre variáveis é estimada via informação mútua (IM). Entretanto, a IM é difícil de estimar e, às vezes é necessário regularizar os dados, utilizando um método de suavização. Caso o objetivo seja fazer previsões pode ser que algumas variáveis possam ser incluídas indevidamente, logo é importante ter o método *wrapper* na saída do filtro para fazer a avaliação em conjunto das variáveis e eliminar as variáveis indesejáveis.

Observa-se que quando não se conhece a função de densidade de probabilidades (fdp) da distribuição dos dados, cria-se um histograma, mas a freqüência dos valores amostrados tem que ser adequadamente baixa para se obter valores precisos [CELLUCCI et al., 2003]. Também pode-se fazer hipóteses simplificadoras como considerar que a fdp é Normal. O filtro proposto nesta tese estima a informação mútua entre variáveis, diretamente de dados discretos, sem a necessidade de fazer hipóteses sobre a distribuição a priori dos dados, tendo vital importância prática. Isto pôde ser alcançado com a utilização da inequação de Cauchy-Schwartz, que é uma substituta do divergente de Kullback-Leibler, integrada a uma Janela de PARZEN [PARZEN, 1962]. Este procedimento será apresentado na próxima subseção.

3.4.1 Método para estimar a informação mútua

SHANNON [SHANNON, 1948] provocou um impacto significativo na área de tecnologia de informação (TI). Apesar da sua origem prática, tratava-se de uma teoria matemática profunda relacionada com a essência do processo de transferência de informação [PRÍNCIPE, 1998]. Já RENYI [RENYI, 1976] formalizou matematicamente nossa noção intuitiva de informação contida em dados. Se os dados forem totalmente conhecidos a priori, seu índice de informação é zero. Entretanto, quanto menos conhecidos são os dados, maior é seu índice de informação.

Em SHANNON (1948), tem-se uma aproximação axiomática para entropia de uma função de distribuição de probabilidades $P = \{p_1, p_2, ..., p_N\}$ como:

$$H_S(P) = \sum_{k=1}^{N} p_k \log \frac{1}{p_k} \tag{3.10}$$

em que H_S é a entropia de Shanon com $\sum_{k=1}^{N} p_k = 1$ e $p_k \geq 0$. Esta equação fornece a quantidade de informação média contida em uma única variável aleatória $Y = \{y_1, y_2, ..., y_N\}$

e com as distribuições das probabilidades dadas por $p_k = P(y = y_k)$, onde $k = 1, 2, ..., N$. A entropia também pode ser vista como a quantidade de informação faltante em Y, cuja distribuição a priori é conhecida.

A definição de entropia pode ser também derivada da teoria geral das médias, em que a média dos números reais $\{y_1, y_2, ..., y_N\}$, com pesos positivos (não necessariamente densidades de probabilidades) $\{p_1, p_2, ..., p_N\}$, tem a seguinte fórmula:

$$\bar{y} = \varphi^{-1}(\sum_{k=1}^{N} p_k \varphi(y_k)) \tag{3.11}$$

em que $\varphi(y)$ é dado pela função de Kolmogorov-Nagumo, que é uma função arbitrária contínua, estritamente monotônica e definida nos números reais. No geral, é uma medida da entropia (H) e obedece à relação:

$$H(P) = \varphi^{-1}(\sum_{k=1}^{N} p_k \varphi(I(p_k))) \tag{3.12}$$

em que $I(p_k) = -\log(p_k)$ é a medida de informação de Hartley [HARTLEY, 1928]. A fim de ter uma medida de informação, $\varphi(.)$ não pode ser escolhida arbitrariamente porque a informação tem que ser aditiva. Para satisfazer a condição de aditividade, pode ser escolhida entre as seguintes famílias de funções: $\varphi(y) = y$ ou $\varphi(y) = 2^{(1-\alpha)y}$. Caso for selecionada a primeira família, tem-se a entropia de Shannon, caso contrário, tem-se a entropia de Renyi de ordem α, que pode ser expressa por

$$H_{R_\alpha}(P) = \frac{1}{1-\alpha} \log(\sum_{k=1}^{N} p_k^\alpha) \tag{3.13}$$

com $\alpha > 0$ e $\alpha \neq 1$.

Resultando que:

$$H_{R_\alpha} \geq H_S \geq H_{R\beta}$$

caso $1 > \alpha > 0$ e $\beta > 1$.

Considerando a distribuição de probabilidades $P = \{p_1, p_2, ..., p_N\}$ como um ponto em um espaço dimensional de N e lembrando as condições impostas pelas leis das probabilidades $p_k \geq 0$, $\sum_{k=1}^{N} p_k = 1$, então, P encontra-se num hiperplano localizado no primeiro quadrante, com suas N dimensões podendo alcançar as coordenadas de valor 1. A distância de P à origem é a α raiz de V_α que é dada por:

$$V_\alpha = \sum_{k=1}^{N} p_k^\alpha = \|P\|^\alpha. \tag{3.14}$$

A raiz α de V_α é chamada de norma da distribuição de probabilidades. Logo, a entropia de Renyi pode ser escrita em função de V_α :

$$H_{R\alpha}(P) = \frac{1}{1-\alpha} \log V_\alpha. \tag{3.15}$$

Quando valores diferentes de α são selecionados na família de funções de Renyi, o resultado final é a seleção de diferentes α-normas. A entropia de Shannon pode ser considerada como um exemplo limite da norma da distribuição de probabilidades. Observa-se que o limite fornece uma indeterminação, mas o resultado existe e é dado pela entropia de Shannon. Com esta visão, a entropia de Renyi é uma função monotônica da α-norma da fdp e é essencialmente uma função monotônica da distância da distribuição de probabilidades à origem. Tem-se a liberdade para escolher a α-norma e quando $\alpha = 2$ resulta:

$$H_{R2}(P) = -\log \sum_{k=1}^{N} p_k^2 \tag{3.16}$$

em que o valor de $H_{R2}(P)$ é chamado de entropia quadrática de Renyi e corresponde à norma $L2$ da fdp P. A entropia de Renyi já foi utilizada com sucesso para estimar a dimensão da correlação de atratores de sistemas dinâmicos não-lineares. Em [SCHREIBER, 1998] foi sugerido que esta ferramenta é uma medida adequada para tal fim.

O cálculo da entropia quadrática de Renyi é a partir da soma das potências das probabilidades, ou seja, é a norma da função de densidade de probabilidae (fdp), logo pode ser estimada diretamente dos dados.

Assim, seja $a_i \in R^m$, $i = 1, 2, ...N$ um conjunto de amostras pertencendo a uma variável randômica $Y \in R^m$, ou seja, a um espaço m-dimensional. A entropia de Renyi associada a este conjunto de amostras pode ser estimada a partir de uma fdp ajustada a estes dados por meio de uma janela de Parzen dada por:

$$f_y(y) = \frac{1}{N} \sum_{i=1}^{N} G(y - a_i, \sigma^2 I) \tag{3.17}$$

em que G é uma função Gaussiana, σ^2 é a variância e I é a matriz quadrada de identidade de ordem m. A janela de Parzen é uma generalização da técnica dos k-vizinhos mais próximos de um ponto dado (teste). A partir do ponto de teste, distribui-se os pesos pelas curvas de nível no plano de forma que os que estão mais próximos do ponto de teste tem peso maior. Define-se o grau de pertinência como o limite das curvas de nível e tem-se então o núcleo da função. Neste capítulo, a função Gaussiana está no núcleo, o que implica que os pesos decrescem exponencialmente com o quadrado da distância, de forma que os pontos mais distantes são irrelevantes. O espalhamento (variância) dos pontos determina a diferença entre os pesos dos pontos mais próximos em relação aos mais distantes. A utilização desta metodologia torna a estimativa de entropia menos complexa do que se fosse estimada por meio da entropia de Shannon.

Por outro lado, a entropia quadrática de Renyi no tempo contínuo pode ser expressa pela seguinte equação

$$H_{R2}(y) = -log(\int f_y^2(y)dy)$$

e utilizando a janela de Parzen, sabendo que uma soma de gaussianas converge para uma gaussiana, resulta

$$\int f_y^2(y)dy = \int [\frac{1}{N}\sum_{i=1}^{N} G(y-a_i,\sigma^2 I)]^2 dy = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int G(y-a_i,\sigma^2)dy \int G(y-a_j,\sigma^2)dy$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(a_i-a_j,2\sigma^2) \tag{3.18}$$

e definindo $P(\{a_i\}) = \int f_y^2(y)dy$, resulta:

$$P(\{a_i\}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(a_i-a_j,2\sigma^2) \tag{3.19}$$

em que o termo $P(\{a_i\})$ pode ser interpretado como o potencial de informação contido nas amostras a_i e $G(a_i-a_j,2\sigma^2 I)$ como o potencial de informação das amostras a_i sobre as amostras a_j, e vice-versa. Assim, minimizar a entropia equivale a minimizar o potencial de informação contido nas amostras.

A informação mútua é uma idéia mais geral que a idéia de entropia e, às vezes, mais necessária por fornecer uma medida de independência entre duas variáveis randômicas representadas pelas suas respectivas fdps $f(x)$ e $g(x)$. O divergente de Kullback-Leibler dado pela equação abaixo é uma medida de informação mútua utilizada na área da teoria da informação.

$$K(f,g) = \int f(x)\log\left(f(x)/g(x)\right)dx. \tag{3.20}$$

A medida correspondente de divergente de Renyi pode ser dada por:

$$R_\alpha(f,g) = \log\left(\int (f(x)^\alpha/g(x)^\alpha)dx\right)/(\alpha-1). \tag{3.21}$$

Nenhum destes divergentes são fáceis de ser integrados com a janela de Parzen, apresentada anteriormente. Assim, a referência [XU and PRÍNCIPE, 1998] propôs uma métrica, baseada na inequação de Cauchy-Schwartz, para medir a independência entre duas fdps $f(x)$ e $g(x)$:

$$C(f,g) = \log\frac{(\int f(x)^2dx)(\int g(x)^2dx)}{(\int f(x)g(x)dx)^2}. \tag{3.22}$$

Pode ser verificado que $C(f,g) \geq 0$ e que a inequação torna-se verdadeira se e somente se $f(x) = g(x)$. Para duas variáveis randômicas Y_1 e Y_2 (com fdps marginais $f_{y_1}(y_1), f_{y_2}(y_2)$ e fdp conjunta $f_{y_1y_2}(y_1,y_2)$, resulta na seguinte medida de independência

$$C(Y_1,Y_2) = \log\frac{(\int\int f_{y_1y_2}(y_1,y_2)^2dy_1dy_2)(\int\int f_{y_1}(y_1)^2 f_{y_2}(y_2)^2dy_1dy_2)}{(\int\int f_{y_1y_2}(y_1,y_2)^2 f_{y_1}(y_1)^2 f_{y_2}(y_2)^2dy_1dy_2)}. \tag{3.23}$$

Como $C(Y_1,Y_2) \geq 0$, então, Y_1 e Y_2 serão estatisticamente independentes se e somente se $C(Y_1,Y_2) = 0$. Assim, seja um conjunto de dados $\{a_i, i = 1,2,...,N\}$, com a_{i_1} e a_{i_2} no mesmo espaço conjunto, a equação (3.23) pode ser usada para estimar a independência entre estes dois conjuntos de dados.

Para viabilizar um procedimento que utilize a equação anterior no tempo discreto, faz-se um desenvolvimento matemático para se chegar a uma equação que permita a sua implementação. Assim, é fato que

$$\int\int f_{y_1y_2}(y_1,y_2)^2 dy_1 dy_2 = \int f_{y_1}(y_1)^2 dy_1 \int f_{y_2}(y_2)^2 dy_2$$

por definição

$$f_{y_1y_2}(y_1,y_2)^2 = f_{y_1}(y_1)^2 f_{y_2}(y_2)^2$$

e a parcela $f_{y_1y_2}(y_1,y_2)^2$ do denominador da equação (3.23) pode ser substituída por $f_{y_1}(y_1)^2 f_{y_2}(y_2)^2$ e o denominador desta equação resulta em

$$\int\int f_{y_1y_2}(y_1,y_2)^2 f_{y_1}(y_1)^2 f_{y_2}(y_2)^2 dy_1 dy_2$$

$$= \int\int f_{y_1}(y_1)^4 f_{y_2}(y_2)^4 dy_1 dy_2 = [\int\int f_{y_1}(y_1)^2 f_{y_2}(y_2)^2 dy_1 dy_2]^2.$$

Se $\mathbf{a}_i$, $i = 1, 2, ..., N$ é um conjunto de amostras subdividido em dois vetores $\mathbf{a}_{i_1}$ e $\mathbf{a}_{i_2}$ na forma $\mathbf{a}_i = [\mathbf{a}_{i_1}\ \mathbf{a}_{i_2}]$ e os valores de Y são divididos em Y_1 e Y_2. Segundo este modelo, Y_1 e Y_2 não têm correlação explícita (caso contrário, a definição de $\mathbf{a}_i = [\mathbf{a}_{i_1}\ \mathbf{a}_{i_2}]$ não estaria correta). Assim, definindo

$$P(\{\mathbf{a}_i\}) = \int\int f_{y_1y_2}(y_1,y_2)^2 dy_1 dy_2$$

e

$$P(\{\mathbf{a}_{i1}\}) = \int\int f_{y_1}(y_1)^2 dy_1$$

e

$$P(\{\mathbf{a}_{i2}\}) = \int\int f_{y_2}(y_2)^2 dy_2.$$

Finalmente, a equação (3.23) resulta na equação abaixo que permite a implementação do cálculo de informação mútua diretamente dos dados.

$$C(\mathbf{a}_{i1}, \mathbf{a}_{i2}) = \log \frac{P(\{\mathbf{a}_i\}) P_1(\{\mathbf{a}_{i1}\}) P_2(\{\mathbf{a}_{i2}\})}{P_c(\{\mathbf{a}_i\})^2} \tag{3.24}$$

em que $P(\{\mathbf{a}_i\})$ é o potencial total de informação das amostras, $P_l(j, \{\mathbf{a}_i\}) = \frac{1}{N}\sum_{i=1}^{N} G(\mathbf{a}_{j_l} - \mathbf{a}_{i_l}, 2\sigma^2 I_l)$ é o potencial marginal de informação das amostras ($l = 1, 2$) e $P_c(\{\mathbf{a}_i\}) = \frac{1}{N}\sum_{j=1}^{N} P_1(j, \{\mathbf{a}_i\}) P_2(j, \{\mathbf{a}_i\})$ é o potencial de informação cruzada entre amostras. A independência entre duas variáveis requer: baixo potencial total de informação das amostras, baixo potencial de informação marginal e alto potencial de informação cruzada.

As variáveis podem ser binárias ou contínuas amostradas e quantizadas no espaço dos números reais. É sempre possível normalizar os vetores associados às variáveis em questão. Isto porque se trata de equações baseadas em funções gaussianas bem definidas por meio da janela de Parzen, mas deve-se utilizar o mesmo método de normalização em todas as variáveis. Entretanto, é uma metodologia que tem alta demanda por dados, ou seja, os resultados mais confiáveis são obtidos a partir de séries temporais longas.

3.4.2 Filtro

Na seleção do subconjunto de variáveis de entrada o objetivo é eliminar uma variável caso a mesma não forneça nenhuma informação adicional além daquelas fornecidas pelas outras restantes. Tradicionalmente, a pesquisa nesta área foi direcionada para procurar por variáveis relevantes [KOHAVI et al., 1994]. Embora alguns trabalhos publicados já indicassem a existência e o efeito da redundância nas características, havia poucos trabalhos sobre o tratamento explícito deste assunto até a referência [YU and LIU, 2004] apresentarem uma definição formal de redundância.

Seja um conjunto de variáveis $F = \{F_1, F_2, \ldots, F_N\}$ uma amostra de variáveis dentro de um determinado contexto com F_i sendo uma variável deste conjunto. Seja também um conjunto de classes $C = \{C_1, C_2, \ldots, C_k\}$ de modelos utilizados para a previsão de um determinado processo. Considerando G um subconjunto de F, geralmente o objetivo da seleção de variáveis é selecionar um subconjunto mínimo G tal que $P(C|G)$ é igual ou tão próxima possível de $P(C|F)$, em que $P(C|G)$ é a distribuição de probabilidade de C dado o subconjunto G e $P(C|F)$ é a distribuição de probabilidade de C dado F [KOLLER and SAHAMI, 1996].

As definições formais das categorias de relevância (forte, média e fraca) serão apresentadas em seguida. Assim, seja o subconjunto $S_i = F - \{F_i\}$.

Definição 3.4.1: relevância forte
A variável F_i é fortemente relevante se e somente se

$$P(C|F_i, S_i) \neq P(C|S_i).$$

Definição 3.4.2: relevância fraca
A variável F_i é fracamente relevante se e somente se
$P(C|F_i, S_i) = P(C|S_i), \exists S_i' \subset S_i$, tal que $P(C|F_i, S_i') \neq P(C|S_i')$.

Definição 3.4.3: irrelevante
A variável F_i é irrelevante se e somente se

$$\forall S_i' \subseteq S_i, P(C|F_i, S_i') = P(C|S_i').$$

Antes de apresentar a definição formal de redundância será apresentada a definição de cobertura de Markov de acordo com a referência KOLLER e SAHAMI (1996).

Definição 3.4.4: cobertura de Markov
Dada uma variável F_i e o subconjunto $M_i \subset F$ (F_i não pertence a M_i), M_i é dita uma cobertura de markov para F_i se e somente se

$$P(F - M_i - \{F_i\}, C|F_i, M_i) = P(F - M_i - \{F_i\}, C|M_i).$$

A cobertura de Markov possibilita que o subconjunto M_i inclua não somente a informação que F_i tem de C, mas também de outras variáveis.

A definição formal de redundância será apresentada em seguida de acordo com a referência YU e LIU (2004).

Definição 3.4.5: redundância

Seja G um conjunto corrente de variáveis, uma variável é redundante e pode ser removida de G se e somente se é fracamente relevante e tem uma cobertura de Markov M_i em G.

Na realidade, não se pode determinar a redundância diretamente de uma variável somente quando esta variável está correlacionada (talvez parcialmente) com um subconjunto de variáveis.

Para definir o algoritmo que faz a seleção automática de variáveis são necessárias as seguintes definições: C-informação mútua, F-informação mútua, cobertura aproximada de Markov e variável predominante.

Definição 3.4.6: C-informação mútua

A informação mútua entre todas as variáveis candidatas F_i e a classe C (variável dependente), denotada por $SU_{i,c}$.

Definição 3.4.7: F-informação mútua

A informação mútua entre qualquer par de variáveis de entrada F_i e F_j (i diferente de j), denotada por $SU_{i,j}$.

Definição 3.4.8: cobertura aproximada de Markov

Para duas variáveis relevantes de entrada F_i e F_j (i diferente de j), F_i forma uma cobertura aproximada de Markov para F_j se e somente se $SU_{i,c} \geq SU_{j,c}$ e $SU_{i,j} \geq SU_{i,c}$.

Definição 3.4.9: variável predominante

As características predominantes são aquelas que não tem nenhuma cobertura aproximada de Markov no conjunto atual e não podem ser removidas em nenhuma hipótese.

A seleção de variáveis por meio de um filtro no primeiro estágio, utilizando os conceitos de relevância e redundância, resulta na categorização das variáveis em quatro classes: relevantes; pouco relevantes e não redundantes; pouco relevantes e redundantes; irrelevantes. A Figura 3.5 ilustra estes quatro tipos de variáveis. O filtro busca eliminar os seguintes tipos de variáveis: pouco relevantes e redundantes; irrelevantes. O subconjunto das variáveis promissoras é formado pelas variáveis: relevantes; pouco relevantes e não redundantes.

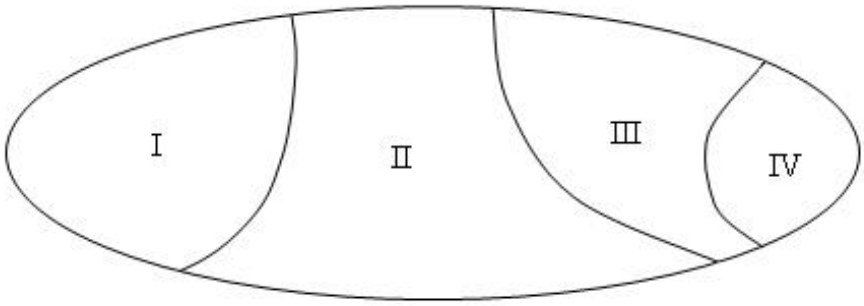

Fig. 3.5: Tipos de variáveis: I - irrelevantes; II - fracamente relevantes e redundantes; III - fracamente relevantes e não redundantes; IV - relevantes

O filtro escolhe um subconjunto de variáveis a partir do algoritmo apresentado na Figura 3.6. Este algoritmo envolve duas etapas conectadas:

- selecionando um subconjunto de variáveis relevantes;
- selecionando as variáveis predominantes entre as pouco relevantes, ou seja, as fracamente relevantes e não redundantes.

Entretanto, podem existir variáveis não controláveis (algumas não observáveis e algumas não conhecidas) e também as amostras finitas de dados podem não explicar o comportamento do processo.

Na implementação do algoritmo do filtro de seleção de variáveis, primeiramente são definidos dois tipos de informação mútua: a informação mútua entre todas as variáveis candidatas F_i e a classe C, chamada de C-informação mútua, denotada por $SU_{i,c}$; a informação mútua entre qualquer par de variáveis de entrada F_i e F_j (i diferente de j), é chamada de F-informação mútua, denotada por $SU_{i,j}$. Na análise de relevância, calcula-se a C-informação mútua para cada variável, e heuristicamente assume-se que uma característica F_i é relevante se tiver um valor alto de informação mútua com a classe C, isto é, se $SU_{i,c} > \delta$, em que δ é um patamar inicial de relevância determinado pelo usuário de acordo com a aplicação. Assim, determina-se um subconjunto de variáveis relevantes e, por exclusão, as irrelevantes.

Na análise de redundância, pode-se avaliar a informação mútua entre variáveis individuais sem considerar a informação mútua entre vários subconjuntos de variáveis. Entretanto, há pelo menos duas desvantagens em se determinar a redundância entre pares de variáveis por meio do cálculo da F-informação mútua: (1) quando duas variáveis não são completamente correlacionadas, pode tornar-se difícil determinar a redundância entre elas e qual delas vai ser removida; (2) requer o cálculo da F-informação mútua para um total de N(N-1) pares, o que é ineficiente para dados de dimensão muito elevada.

A Figura 3.6 apresenta o algoritmo do filtro para um conjunto de variáveis S com N variáveis e uma classe C, o algoritmo encontra um subconjunto de variáveis relevantes S_{filtro}. Na primeira etapa (linhas 1-6), calcula-se o valor de $SU_{i,c}$ para cada variável e são selecionadas as variáveis relevantes para formar a lista S_{lista}, que então é ordenada de forma descendente de acordo com valores de $SU_{i,c}$. Na segunda etapa (linhas 7-20), processa-se a S_{lista} para selecionar as características predominantes e formar o subconjunto de variáveis selecionadas pelo filtro S_{filtro}. Observa-se que neste algoritmo são utilizados os conceitos de cobertura aproximada de Markov e de variáveis predominantes.

```
entradas: S (F1, F2, ..., FN, C) // Conjunto de dados de treinamento.
          δ                      // Um nível pré-definido pelo usuário.
saída:    S filtro               // Subconjunto de variáveis selecionadas pelo filtro.

1         início
2          for i = 1 até N
3              calcule SUi,c para cada Fi
4              se (SUi,c > δ) então anexar Fi a S lista
5          end
6          ordenar S list em ordem descendente de SUi,c
7          Fj = buscar primeiro elemento de S lista
8          Fi = buscar próximo elemento de S lista
9             se (Fj <> vazio) então faça
10                se (Fi <> vazio) então faça
11                   se (SUi,j ≥ SUi,c) então remova Fi de S lista
12                   end
13                Fi = buscar próximo elemento de S lista
14             end até (Fi = = vazio)
15             Fj = buscar próximo elemento de S lista
16          end até (Fj = = vazio)
17         S filtro = S lista
18         end
```

Fig. 3.6: Algoritmo do filtro de seleção de variáveis via relevância e redundância

Este algoritmo explicitamente lida com relevância e a redundância por meio de duas etapas: primeiramente, a análise da relevância determina o subconjunto de características relevantes e remove as irrelevantes, e em segundo lugar, a análise da redundância determina e elimina características redundantes, mas pouco relevantes, desacoplando a análise de relevância da análise de redundância.

A Figura 3.7 ilustra de forma resumida e desacoplada as duas fases do algoritmo do filtro proposto neste capítulo para a seleção de variáveis: a primeira fase faz a análise de relevância e a segunda a análise de redundância.

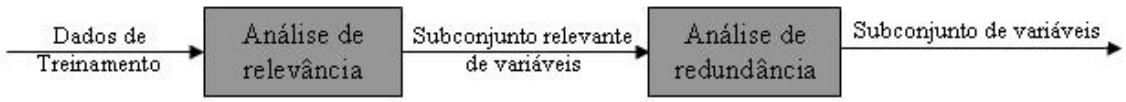

Fig. 3.7: Representação sucinta e desacoplada do filtro proposto para a seleção de variáveis baseado na análise de relevância e redundância

Na saída da Figura 3.7 que representa o filtro proposto para seleção de variáveis restará somente o subconjunto das variáveis selecionadas que irão ser avaliadas no *wrapper*. Como várias candidatas já foram eliminadas pelo filtro o trabalho do *wrapper* será facilitado.

3.4.3 *Wrapper*

A seleção do subconjunto final de variáveis é via *wrapper*, utilizando uma rede neural como algoritmo de indução. Esta seleção está de acordo com o conceito de causalidade de GRANGER (1969), formalizado na Seção 3.2 deste capítulo.

O método de eliminação para trás (*backward selection*) é utilizado para encontrar o subconjunto de variáveis que fornece as melhores previsões por meio de modelos neurais propostos nesta tese. Esta opção deve-se ao fato das aplicações analisadas neste trabalho terem relativamente poucas variáveis selecionadas pelo filtro apresentado anteriormente. Este método foi implementado em [ORR, 1996, ORR, 1999] a partir de redes RBF, mas este conceito pode ser aplicado a outros tipos de modelos.

Este tipo de busca direta ajusta o tamanho da rede RBF e o parâmetro de regularização (λ) em conjunto. Este método permite que múltiplos valores de λ sejam gerados e comparados para evitar que o valor ajustado pelo algoritmo caia em um mínimo local. O λ, nesta tese, será estimado via validação cruzada generalizada [GOLUB et al., 1979]. O método de eliminação para trás é computacionalmente tratável e pode ser utilizado em várias abordagens que controlam a complexidade de modelos neurais, inclusive naqueles em que não há regularização.

Existem outros métodos como o que combina árvores de regressão e redes RBF que tem como idéia básica particionar recursivamente o espaço de entrada em dois e aproximar uma função em cada partição. Este método permite investigar a relevância das variáveis de entrada, ou seja, as mais relevantes tendem a ser subdividas primeiro e com mais freqüência. Não há o problema de decidir quando parar de fazer crescer a rede ou quantos neurônios têm que ser podados; ou seja, o problema de balancear o bias e a variância, comum nos métodos de regressão não paramétrica, já é automaticamente resolvido.

Capítulo 4

Mineração de Dados em Séries Temporais

4.1 Introdução

As publicações originais [BACHELIER, 1900] e [POINCARÉ, 1952] deram origem a duas correntes de opiniões sobre o estudo das séries temporais. Na área de finanças a primeira corrente defende que os rendimentos dos ativos financeiros tem caráter estocástico já que depende da ocorrência de múltiplas variáveis tipicamente imprevisíveis. O trabalho de BACHELIER viriam a ter desdobramentos que deram origem aos principais trabalhos na área de finanças, como por exemplo, ao modelo de análise de portfólio apresentado em [MARKOWITZ, 1959].

POINCARÉ (1952) fez pesquisas na área das equações diferenciais, formalizando processos deterministas não lineares, ganhando relevância principalmente na análise de séries temporais caóticas. Nestes sistemas, a ocorrência de erros nas condições iniciais seriam ampliadas pela existência de uma realimentação no processo. Logo, a previsão a longo prazo seria impossível devido à existência de sensível dependência às condições iniciais, enquanto as previsões de curto prazo poderiam ser viáveis. Poincaré estava na origem daquilo que se chamaria mais tarde de teoria do caos.

Os sistemas dinâmicos deterministas, inclusive os caóticos, podem ter sua reconstrução dinâmica a partir de uma série temporal escalar [TAKENS, 1981]. Entretanto, nem sempre estamos certos se a série temporal é determinista. Para tentar lidar melhor com este possível problema, após estimar os parâmetros da reconstrução dinâmica (*lag* e dimensão de imersão), é realizado o ajuste fino destes parâmetros em função da qualidade das previsões (NMSE) por meio de um modelo baseado em uma rede neural tipo RBF regularizada [POGGIO and GIROSI, 1990b], visando somente previsões de curto prazo.

Os sistemas caóticos deterministas, apesar de aparentar um comportamento aleatório, são governados por leis deterministas. As características invariantes da dinâmica caótica analisadas nesta tese são a dimensão da correlação que expressa a complexidade do sistema e o expoente de LYAPUNOV que sinaliza a dependência às condições iniciais. Um processo caótico, por definição, é aquele que tem pelo menos um expoente de LYAPUNOV positivo. Isto sinaliza que a amostra é gerada por um sistema determinista não linear. O inverso deste expoente define o horizonte de previsibilidade de curto prazo da série temporal.

Um interesse maior na existência de dinâmica caótica em séries no tempo surgiu na década de 1980. Vários procedimentos importantes de testes foram estabelecidos nesta época, na

maior parte ajustando as ferramentas usadas na física e em outras ciências naturais. Um resumo dos resultados obtidos neste período, para a área de finanças, pode ser encontrado em [LEBARON, 1994].

Por outro lado, a mineração de dados em séries temporais apóia-se em vários campos da ciência: análise de séries temporais, análise dinâmica não-linear, estatística aplicada e outras [POVINELLI, 1999]. A mineração de dados é a busca de relações e padrões escondidos nos dados. No contexto desta tese, a mineração de dados consiste justamente na busca de relações e *clusters* de padrões temporais com capacidade preditiva.

O objetivo deste capítulo é fazer a reconstrução dinâmica e, a partir dela, ajustar a janela de previsão dinâmica via modelos neurais e implementar a identificação de relações (lineares e não lineares) e de *clusters* com capacidade preditiva inclusive para determinar os centros das redes RBF utilizadas nas previsões. Este capítulo foi organizado como segue: na Seção 4.2 faz-se uma revisão sobre a reconstrução dinâmica; na Seção 4.3 é apresentado o método utilizado para o ajuste da janela de previsão inteligente dinâmica (JPID); A Seção 4.4 analisa a possibilidade de ocorrência de dinâmica não-linear na série temporal; A Seção 4.5 aborda a identificação de *clusters* de padrões temporais, buscando encontrar principalmente aqueles com capacidade preditiva. Os métodos abordados neste capítulo serão utilizados também para gerar os resultados que serão apresentados no Capítulo 6.

4.2 Reconstrução dinâmica

A reconstrução dinâmica a partir de uma série temporal escalar gerada por um sistema determinista se baseia principalmente no teorema de TAKENS (1981) embora tenha sido em [PACKARD et al., 1980] que primeiro se visualizou a possibilidade da reconstrução dinâmica. Entretanto, em [CELLUCCI et al., 2003] foi observado que é temeroso fazer a reconstrução dinâmica somente por meio de critérios baseados na própria série temporal. Recomendam, por exemplo, um estudo analítico sobre o sistema dinâmico analisado, que permita a confirmação dos resultados obtidos por meio de dados experimentais, evitando a armadilha da lógica circular. Isto fornece mais uma motivação para que o ajuste da janela de previsão seja a partir da qualidade das previsões (NMSE).

A implementação da reconstrução dinâmica consiste no ajuste dos parâmetros tempo de atraso (*lag* L) e dimensão de imersão (M) a partir de uma série temporal escalar observada. Estes parâmetros servem para estimar vetores no $\Re^M$, expressos matematicamente na forma $Y_t = \{y(t), y(t+L), y(t+2L), ..., y(t+(M-1)L)\}$. Esta reconstrução dinâmica será utilizada para ajudar a determinar o que se chama janela de predição que se torna dinâmica ao se deslocar ao longo da série no tempo. Como é determinada por meio de RNA, será denominada de janela de predição inteligente. A reconstrução dinâmica apropriada é também necessária para a determinação da dimensão da correlação e do expoente de LYAPUNOV que serão utilizados posteriormente para investigar a presença de caos na série temporal.

A estratégia adequada de reconstrução dinâmica depende basicamente da série temporal amostrada e dos métodos utilizados para ajustar os parâmetros L (informação mútua, função de autocorrelação e outros) e M (falsos vizinhos mais próximos globais, RNA's e outros). Um destes métodos pode ser adequado para uma determinada aplicação, mas pode não funcionar

para outras. Não existe respostas simples para este assunto. Entretanto, é uma alternativa à determinação arbitrária dos parâmetros.

Na reconstrução dinâmica de um sistema pode ocorrer que este sistema seja caótico. Dessa forma, apesar das equações representarem um sistema determinista exibem um comportamento aparentemente estocástico e o poder de previsão não é mais absoluto. Assim, trata-se de um sistema determinista não previsível no longo prazo, independente de quão precisas são as condições iniciais. Entretanto, há um limite para o caos que, contrariamente ao uso comum da palavra, que significa o oposto da ordem, em um sistema físico representa um comportamento que, apesar de complexo, não é inteiramente desorganizado. Há uma persistência no comportamento do sistema, uma obediência a certos vínculos, que não existe em um sistema aleatório.

Para diversos experimentos é impossível registrar todo o conjunto de variáveis independentes simultaneamente a fim de se construir o atrator. Entretanto, de acordo com o teorema de TAKENS (1981), o atrator pode ser reconstruído a partir da medida de uma única série temporal. Ou seja, é possível definir um espaço de fase que capture a dinâmica do sistema em uma estrutura geométrica imersa nesse espaço. O conjunto geométrico imerso é chamado de atrator reconstruído e ele é topologicamente equivalente ao atrator que seria produzido pela evolução do sistema dinâmico, caso suas equações fossem conhecidas. A dimensão da correlação e o expoente de LYAPUNOV devem ser aproximadamente os mesmos tanto para o atrator original como para o atrator reconstruído.

A qualidade do atrator reconstruído é bastante sensível ao valor escolhido para o tempo de atraso (L). Por qualidade do atrator, entende-se quão similar o atrator reconstruído é do atrator original. Na prática, atratores gerados com L pequeno são fechados e mal definidos, valores elevados de L geram atratores dispersos, ao passo que valores adequados de L geram atratores com dinâmica bem definida.

O teorema de TAKENS (1981) foi concebido para a reconstrução de sistemas dinâmicos deterministas com dimensão finita a partir de uma série escalar no tempo observada $\{y_t\}_{t=1}^{N}$. Esta série pode representar, por exemplo, tanto um sinal de tensão medido durante um exame com eletro-encefalograma (EEG) como a cotação diária da taxa de câmbio brasileira. No caso discreto, a série pode ser representada por um conjunto de dados observados no tempo: $\{y(1), y(2), y(3), ..., y(N)\}$, $y_t \epsilon \Re$. Quando o valor do *lag* é igual a 1 (valores de *lag* diferentes de um serão analisados mais adiante), estes valores observados podem ser utilizados para criar um conjunto de pontos suficientes para implementar a reconstrução dinâmica via $\mathbf{Y}_t \epsilon R^M$, dada por

$$\mathbf{Y}_t = \{y(t), y(t+1), y(t+2), ..., y(t+M-1)\} \tag{4.1}$$

em que M é a dimensão de imersão (*embedding*). O comportamento no tempo de $\mathbf{Y}_t$ é a trajetória em um espaço de estados de dimensão M que pode ser expresso por $\mathbf{Y}_1 \rightarrow \mathbf{Y}_2 \rightarrow \mathbf{Y}_3 \rightarrow ...$. Isto significa que as propriedades dinâmicas do sistema são refletidas em $\mathbf{Y}_t$ ajustado a partir do sinal observado.

Considerando que o sinal observado é gerado por um sistema dinâmico com ω variáveis reais, às vezes, nem todas as variáveis são observáveis. Como uma função do tempo, o sistema dinâmico pode se mover em um espaço de estados compacto dado por P, o qual é um subconjunto de $\Re^{\omega}$. A hipótese de espaço de estados compacto (intervalo fechado) é considerada

para P, entretanto, em determinadas aplicações pode ser que não possa ser confirmada a partir somente dos dados experimentais disponíveis. A partir do espaço de estados P, o sistema dinâmico pode ser pensado como um mapeamento contínuo atuando no espaço Ψ: $P \rightarrow P$. Para qualquer ponto com valor inicial $\boldsymbol{Y}_t$, $\boldsymbol{Y}_t \epsilon$ P $\subseteq \Re^{\omega}$, o estado do sistema no tempo t é dado por $\Psi(\mathbf{Y}_t)$. O objetivo da análise é inferir propriedades do mapeamento Ψ a partir das amostras escalares $\{y_t\}_{t=1}^{N}$, utilizadas para ajustar o vetor $\mathbf{Y}_t$.

Seja $\mathbf{Y}_t \epsilon$ P, que denota a situação real do sistema no instante t, e $y_t \epsilon \Re$ que é o valor escalar da variável no instante t. Considerando que $\mathbf{Y}_t$ é relacionado a y_t por meio de um mapeamento suave dado por $c : P \rightarrow \Re$, tal que $c(\mathbf{Y}_t) = y_t$, para qualquer t. Considerando também que o conjunto de $\mathbf{Y}_t's$ correspondentes aos $y_t's$ formam um conjunto representativo de P. Assim, para qualquer M, com $M > 2\omega$, define-se o mapeamento $\Phi : P \subseteq \Re^{\omega} \rightarrow \Re^{M}$, logo

$$\Phi(\mathbf{Y}_t) = \{c(\mathbf{Y}_t), c(\Psi(\mathbf{Y}_t)), c(\Psi^2(\mathbf{Y}_t)), ..., c(\Psi^{M-1}(\mathbf{Y}_t))\} \tag{4.2}$$

como $\Psi(\mathbf{Y}_t) = \mathbf{Y}_{t+1}$ e $c(\mathbf{Y}_t) = y_t$, então

$$\Phi(\mathbf{Y}_t) = \{y(t), y(t+1), y(t+2), ..., y(t+M-1)\}. \tag{4.3}$$

Isto resulta que:

- para alguns valores de Ψ e c, Φ é uma reconstrução dinâmica do sistema se P, sob o difeomorfismo (uma função diferenciável com a sua inversa também diferenciável), tem sua imagem em Φ;
- o mapeamento contínuo $\mathbf{Y}_t \rightarrow \mathbf{Y}_{t+1}$ corresponde ao mapeamento original Ψ. Assim, a trajetória observada $\mathbf{Y}_t \rightarrow \mathbf{Y}_{t+1}$ é intimamente ligada ao mapeamento original Ψ. As propriedades do mapeamento $\mathbf{Y}_t \rightarrow \mathbf{Y}_{t+1}$, estabelecidas a partir dos dados observados serão também verdadeiras para Ψ.

Se estas condições são estabelecidas, pode-se fazer inferências sobre Ψ a partir de $\mathbf{Y}_t$, ou seja, a dinâmica do espaço de estados reconstruído pode conter as mesmas informações topológicas do espaço de estados original.

Assim, pode-se fazer a análise de um sistema dinâmico com dimensão w somente baseando-se em uma série temporal escalar. Entretanto, no mundo real, as condições do teorema anterior geralmente não são estabelecidas totalmente. A hipótese crucial é que o conjunto de $\mathbf{Y}_t's$ correspondentes aos valores observados $y_t's$ formam um subconjunto compacto no espaço P.

Ao incorporar o *lag* L, com L ϵ Z^+, tem-se mais informações para auxiliar a encontrar o espaço de estados da trajetória. A partir da equação (4.1), incorpora-se o *lag* L, resultando na equação do atrator

$$\mathbf{Y}_t = \{\mathbf{y}(t), \mathbf{y}(t+L), \mathbf{y}(t+2L), ..., \mathbf{y}(t+(M-1)L)\}. \tag{4.4}$$

As limitações impostas pelo tamanho finito de $\{y_t\}_{t=1}^{N}$ podem ser contornadas por meio da observação de mais de uma variável. A imersão pode ser aplicada a dados multicanais. Supondo que foram observados K canais e caso $\{y_t^i\}_{t=1}^{N}$ expresse a série temporal observada no canal i, então

$$\mathbf{y}_t^i = \{y_1^i, y_2^i, y_3^i, ..., y_N^i\}. \tag{4.5}$$

Considerando que o procedimento mais fácil é aquele utilizado para construir o espaço de imersão em $\Re^K$:

$$\mathbf{Y}_t = \{\mathbf{y}_t^1, \mathbf{y}_t^2, ..., \mathbf{y}_t^K\}. \tag{4.6}$$

O procedimento de imersão de dados escalares observados para uma dimensão arbitrária pode ser generalizado, resultando em

$$\mathbf{Y}_t = \{\mathbf{y}_t^1, \mathbf{y}_t^2, ..., \mathbf{y}_t^K, \mathbf{y}_{t+1}^1, \mathbf{y}_{t+1}^2, ..., \mathbf{y}_{t+1}^K, ...\}. \tag{4.7}$$

Este procedimento poderá falhar se o número de variáveis observadas K for menor que a dimensão efetiva do sistema dinâmico gerador dos dados. Considerando que as variáveis observadas sejam w, z e y, uma representação mais simples de $\mathbf{Y}_t$ pode ser formada no $\Re^3$, de acordo com a equação abaixo.

$$\mathbf{Y}_t = \{\mathbf{w}_t, \mathbf{z}_t, \mathbf{y}_t\}. \tag{4.8}$$

4.3 Janela de previsão inteligente dinâmica (JPID)

A função de autocorrelação linear tem sido utilizada de forma extensiva para determinar o *lag* L, entretanto não capta as relações não lineares. Em [ABARBANEL, 1993] foi observado que o primeiro mínimo da informação mútua é uma escolha mais apropriada para determinar o *lag* L, já que a informação mútua pode ser considerada como uma análoga não linear da função de autocorrelação. Os conceitos associados à informação mútua já foram abordados no capítulo anterior e destaca-se somente que o método utilizado nesta seção é o apresentado em [CELLUCCI et al., 2005].

A determinação da dimensão de imersão M é realizada por meio do conceito dos K vizinhos mais próximos. Nos modelos de previsão de séries temporais de alta freqüência, o algoritmo dos vizinhos mais próximos se apresenta como uma ótima opção e já tem um histórico de bons resultados [ALEXANDER, 2005]. Neste algoritmo, cada ponto é mapeado no espaço $\Re^M$, onde M é a dimensão de imersão. É criada uma biblioteca de dados no $\Re^M$. Para fazer uma previsão no instante t é mapeada a biblioteca na dimensão de imersão M. Assim, são encontrados os vizinhos mais próximos no $\Re^M$ de modo que as coordenadas desses pontos possam ser usadas como dados da variável explicativa.

Nos métodos convencionais de previsão de séries no tempo são utilizados os pontos imediatamente precedentes, tomados em intervalos sucessivos e espaçados igualmente no tempo. Já nos métodos de previsão que utilizam vizinhos mais próximos é utilizada uma seleção de pontos que estão sendo escolhidos porque são similares ao valor da série no instante da previsão, respeitando a dimensão de imersão. A menos que alguém resolva investigar detalhadamente o algoritmo, não se sabe exatamente onde ocorreram os pontos que estão sendo escolhidos na previsão de cada ponto. Eles podem ter ocorrido em qualquer tempo e não necessariamente caminham consecutivamente ao longo do tempo.

Assim, seja um conjunto de pontos Q no espaço M-dimensional, é definido como vizinho mais próximo de um ponto de referência p aquele ponto q pertencente a Q que tem a menor distância de p. A questão mais geral de encontrar mais de um vizinho é chamada do problema dos K vizinhos mais próximos. Em geral, o ponto de referência p é um ponto arbitrariamente localizado, mas é também possível que p seja um membro do conjunto Q. A Figura 4.1 ilustra esta definição.

Existem duas maneiras de se implementar o algoritmo dos K vizinhos mais próximos:

- considerar um número fixo de K vizinhos mais próximos, vide Figura 4.1;
- adotar uma esfera de raio fixo na dimensão de imersão M em torno de p, de acordo com a Figura 4.2.

No raio fixo, considera-se todos os valores dentro da esfera e quanto maior for o raio da esfera, maior é o número de vizinhos mais próximos. Cada valor de p pode ter um número diferente de vizinhos mais próximos. A previsão utilizará muitos ou poucos valores, dependendo do comportamento do processo. No método da esfera, o valor da distância (por exemplo a euclidiana) será menor ou maior dependendo do mercado estar mais estável ou mais agitado.

Os dois métodos necessitam do cálculo da distância, e geralmente é utilizada a euclidiana. Obviamente, se a quantidade de amostra for muito grande, gasta-se muito tempo de processamento. Algumas vezes, o método da distância de busca é chamado de procedimento de tamanho fixo, enquanto a busca de K vizinhos mais próximos é chamada de procedimento de massa fixa. A Figura 4.1 ilustra um gráfico com um exemplo de vizinhança mais próxima do tipo K vizinhos mais próximos.

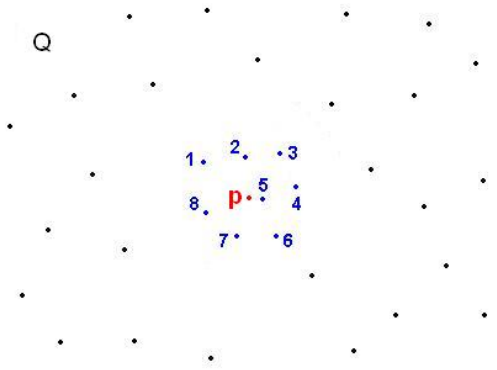

Fig. 4.1: Gráfico da vizinhança mais próxima

O algoritmo da circunferência, ou seja, aquele que permite um pequeno erro ε, de acordo com a Figura 4.2, estima os vizinhos em volta do ponto de referência p com distância possivelmente menor ou maior que a distância estipulada.

O erro relativo máximo permitido ε é dado como um parâmetro do algoritmo. Para $\varepsilon = 0$, a busca retorna a distância exata dos vizinhos mais próximos. Computar os vizinhos mais próximos por meio da distância exata de conjuntos de dados com dimensão fractal maior que 6 (seis) parece ser uma tarefa que consome bastante esforço computacional. Poucos algoritmos tem performance melhor que simplesmente computar todas as distâncias. Entretanto, com

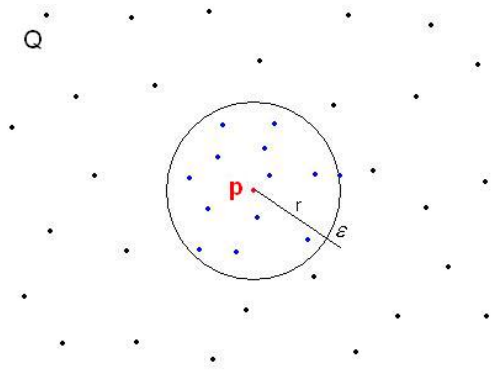

Fig. 4.2: Gráfico da distância da vizinhança mais próxima

o aperfeiçoamento dos algoritmos e a capacidade de processamento das máquinas, tem sido verificado que computar os vizinhos mais próximos aproximados possibilita alcançar tempos de processamento significativamente menores com erros de distância relativamente menores.

Seja $\mathbf{Y}_i$ um elemento com M e L ajustados, e $\mathbf{Y}_i^{NN} = \{\mathbf{y}_i^{NN}, \mathbf{y}_{i+L}^{NN}, ..., \mathbf{y}_{i+(M-1)L}^{NN}\}$ que expressa seu vizinho mais próximo. A distância euclideana entre estes dois pontos no $\Re^M$ é dada por:

$$|\mathbf{Y}_i - \mathbf{Y}_i^{NN}|_M^2 = \sum_{k=0}^{M-1} (\mathbf{y}_{i+kL} - \mathbf{y}_{i+kL}^{NN})^2. \tag{4.9}$$

A distância euclideana entre a projeção destes dois pontos no $\Re^{M+1}$ é dada por

$$|\mathbf{Y}_i - \mathbf{Y}_i^{NN}|_{M+1}^2 = |\mathbf{Y}_i - \mathbf{Y}_i^{NN}|_M^2 + (\mathbf{y}_{i+ML} - \mathbf{y}_{i+ML}^{NN})^2. \tag{4.10}$$

Em ABARBANEL et al. (1993) e CELLUCCI et al. (2003) foi definido um parâmetro R como uma medida da distância entre $\mathbf{Y}_i$ e $\mathbf{Y}_i^{NN}$, normalizada de acordo com a sua distância em $\Re^M$, inicialmente dada por

$$R = \{\frac{|\mathbf{Y}_i - \mathbf{Y}_i^{NN}|_{M+1}^2 - |\mathbf{Y}_i - \mathbf{Y}_i^{NN}|_M^2}{|\mathbf{Y}_i - \mathbf{Y}_i^{NN}|_M^2}\}^{1/2}. \tag{4.11}$$

Uma forma mais simples de expressar R é

$$R = \frac{|\mathbf{y}_{i+ML} - \mathbf{y}_{i+ML}^{NN}|}{|\mathbf{Y}_i - \mathbf{Y}_i^{NN}|_M^2}. \tag{4.12}$$

Logo, $\mathbf{Y}_i^{NN}$ será julgado um falso vizinho mais próximo no R^M se R exceder à constante R_{tot}. Nesta tese as recomendações de ABARBANEL et al. (1993) e CELLUCCI et al. (2003) de fazer $R_{tot} = 15$ será seguida.

O método dos falsos vizinhos mais próximos globais para determinar a dimensão de imersão é implementado neste capítulo por meio do seguinte procedimento:

- O *lag* L é ajustado via informação mútua.

- O R_{tot} é arbitrariamente fixado em 15.
- A quantidade de vizinhos falsos mais próximos é calculada em função de M, usando o seguinte procedimento: (a) Para cada ponto $\mathbf{Y}_i \in \Re^M$, os falsos vizinhos mais próximos $\mathbf{Y}_i^{NN}$ serão determinados. (b) O valor correspondente de R é calculado. (c) Se $R > R_{tot}$, $\mathbf{Y}_i^{NN}$ será considerado um falso vizinho mais próximo de $\mathbf{Y}_i$.
- O valor de M será aumentado até que os falsos vizinhos mais próximos não sejam mais observados ou até que a sua freqüência fique abaixo de uma percentagem aceitável.

O método dos vizinhos mais próximos pode ser utilizado também na análise de *clusters*. Este método poderá não funcionar para séries no tempo escalares com quantidade de amostras pequenas, mas pode ser utilizado em séries temporais multivariadas.

Resumindo, o procedimento para ajustar a janela de previsão inteligente dinâmica (JPID) consiste nos seguintes estágios:

- **Primeiro estágio**: o cálculo da informação mútua proposta em CELLUCCI et al. (2003), para o nível de confiança de 0,05, é utilizado para estimar inicialmente o *lag* L; a partir do valor de L, o algoritmo dos falsos vizinhos mais próximos globais é utilizado para estimar a dimensão de imersão M.
- **Segundo estágio**: rede RBF com validação cruzada generalizada [GOLUB et al., 1979] é utilizada para o ajuste final do *lag* L e da dimensão de imersão M em função da qualidade das previsões.

Neste contexto é importante observar que o processo de aprendizagem de uma rede neural depende da transferência eficaz das informações contidas nas amostras para os parâmetros da mesma. O objetivo principal é capturar a informação e concentrá-la na rede, manipulando-a diretamente, não necessitando de suposições a priori. A JPID está em sintonia com este tipo de aprendizagem e também pode ser utilizada para a identificação de *clusters* com capacidade preditiva.

Nesta janela a rede RBF busca o mapeamento ótimo entre a entrada e a saída em função do NMSE (*normalized mean square error*), ou seja, o L e o M que permite capturar o máximo de informação disponível nos dados. Este método busca o erro quadrado médio normalizado (NMSE) mínimo de predição. A validação cruzada generalizada [GOLUB et al., 1979] e a determinação do espalhamento dos *clusters* são utilizados para otimizar este ajuste e serão apresentados no Capítulo 5.

4.4 Análise dinâmica não-linear

Até 1963, os sistemas dinâmicos eram classificados em três categorias, segundo o padrão de variação no tempo: estáveis, convergindo para um valor fixo; periódicos, estabelecendo-se em oscilações periódicas; ou imprevisíveis, caracterizados por flutuações irregulares. Os sistemas imprevisíveis eram também denominados aleatórios ou ruidosos.

EDWARD LORENZ, quando estudava um modelo de previsão do tempo, fez uma grande descoberta que surpreendeu o mundo. Em [LORENZ, 1963] apresentou um modelo que seguia um curso que não se enquadrava como aleatório, periódico ou convergente, exibindo um comportamento bastante complexo, embora fosse definido apenas por poucas e simples equações diferenciais. A dinâmica gerada pelo modelo exibia uma característica não usual: dois pontos localizados a uma distância muito pequena seguiam ao longo do tempo rotas bastante divergentes. Esta observação levou LORENZ a conjecturar que a previsão do tempo em um intervalo de tempo longo não seria possível. Sistemas como o de LORENZ são denominados caóticos deterministas ou simplesmente caóticos. Embora apresentem um comportamento aperiódico e imprevisível, a sua dinâmica é governada por equações diferenciais deterministas.

Não existe uma definição geral de caos, mas em [THEILER, 1990] é tido como um comportamento irregular de equações simples. Estes sistemas possuem trajetórias no espaço de estados por onde convergem para atratores estranhos cuja dimensão fractal sinaliza o número efetivo de graus de liberdade e nível de complexidade. Assim, quando um atrator de um sistema dinâmico é fracionário (fractal), este sistema é caótico e o seu atrator é conhecido como estranho. A sensibilidade às condições iniciais, uma das características dos sistemas caóticos, pode ser analisada por meio do expoente de Lyapunov.

O termo fractal foi cunhado em [MANDELBROT, 1963] e representa para a matemática, uma forma geométrica complexa, detalhada e auto-semelhante (cada porção é uma réplica reduzida do todo) que não pode ser tratada como tendo uma, duas, ou qualquer outra dimensão inteira, mas alguma dimensão fracionária. A sua característica principal é poder ser descrito por uma dimensão não inteira. Podem ser arbitrariamente divididos em duas categorias: objetos sólidos ou atratores estranhos. As nuvens ou regiões costeiras são exemplos bem conhecidos de objetos sólidos. O atrator estranho, em contraste, é um conceito abstrato, mas é a dimensão fracionária de atratores de sistemas dinâmicos caóticos.

O número efetivo de graus de liberdade pode ser utilizado para se distinguir os sistemas estocásticos (muitos graus de liberdade efetiva) dos sistemas deterministas (poucos graus de liberdade efetiva). Ferramentas teóricas, como por exemplo o coeficiente de LYAPUNOV e a dimensão da correlação, fornecem informações globais sobre o processo, sinalizando para um sistema estocástico, determinista ou caótico-determinista. Geralmente são estimados a partir de dados experimentais obtidos de séries no tempo com o objetivo de reconhecer estados físicos diferentes do sistema.

A capacidade de auto-organização pode estar presente em sistemas dinâmicos dissipativos cujo transiente evolui para poucos graus de liberdade. Assim, um sistema pode ter vários graus de liberdade nominal, mesmo com poucos graus de liberdade efetiva. Por outro lado, a divergência de rotas bastante próximas observada por LORENZ chamou a atenção para a sensibilidade à variação das condições iniciais. É uma característica que diferencia os sistemas caóticos deterministas dos sistemas estocásticos. Para sistemas estocásticos, a mesma condição inicial pode conduzi-los a estados bastantes distintos em pequenos intervalos de tempo, o que não ocorre nos sistemas caóticos deterministas, nos quais as trajetórias próximas crescem praticamente de forma independente e a distância entre elas aumenta exponencialmente.

Embora sistemas conservativos possam exibir comportamento caótico somente sistemas dissipativos possuem atratores estranhos. Os graus de liberdade efetiva dos sistemas podem ser estimados por meio do conceito da dimensão da correlação dos atratores estranhos. Entretanto,

em séries temporais, para se determinar este parâmetro é preciso antes fazer a reconstrução dinâmica do sistema.

A teoria do caos é utilizada geralmente como uma ferramenta para analisar sistemas mal compreendidos do ponto de vista determinístico, tais como fenômenos sociais, turbulência em fluidos, econômicos, climáticos, epidemiológicos e outros. Um sistema estocástico é normalmente abordado por meio da teoria da probabilidade, já que tem muitos graus de liberdade e torna-se muito difícil de ser descrito por um sistema de equações diferenciais.

4.4.1 Dependência temporal não-linear

A verificação da existência da dependência temporal que possa incluir não linearidades é relevante no contexto da predição de séries temporais porque os gráficos de autocorrelação e autocorrelação parcial só detectam a dependência temporal linear. É bom lembrar que uma série temporal que não apresenta nenhum tipo de dependência (linear e não linear) temporal é uma Martingale e não é viável se fazer previsões sobre esta série.

O teste de BROCK, DECHERT E SCHEINKMAN (1996), estatística BDS, testa a hipótese nula de que as amostras da distribuição da variável aleatória são estocasticamente independentes, indicando ou não a presença de dependência linear ou não linear. Este teste também sinaliza que esta série tem algum nível de previsão e também a possibilidade da presença de não-linearidades na série, ou seja, é um teste indireto de não linearidade, mas para que seja aceito para este fim, é necessário antes filtrar as componentes lineares. Por exemplo, nos resíduos que resultaram de um ajuste de um modelo ARIMA, pode ser utilizado um teste BDS para detectar não linearidades. Um teste para detectar não-linearidades bastante utilizado na literatura é o de HSIEH (1989). Este teste pode ser utilizado em conjunto com o teste BDS para uma análise mais apropriada da existência ou não de não linearidades. No caso da dependência, o teste BDS não informa se é de curto, médio ou longo prazo.

Os valores da dimensão da correlação e do expoente de LYAPUNOV podem ser estimados a partir da reconstrução dinâmica do processo. O expoente de LYAPUNOV faz a avaliação da sensibilidade como este sistema reage às condições iniciais, sinalizando se o mesmo é estocástico ou caótico. A dimensão da correlação sinaliza o nível de complexidade, indicando também se o processo é estocástico, determinista ou caótico-determinista. Existem outras ferramentas de análise (seções de POINCARÉ, transformada rápida de Fourier, dimensão de LYAPUNOV e outras) que podem ser ou não utilizadas, dependendo do contexto da aplicação.

Na área de finanças, a maioria das publicações mais relevantes analisam os mercados financeiros como um processo aleatório, permitindo seu estudo estatístico. O modelo apresentado em [BLACK and SCHOLES, 1973], a teoria moderna da gestão de carteiras [SHARPE, 1963, SHARPE, 1964] e a diminuição do risco via cálculo das covariâncias [MARKOWITZ, 1959] são baseados nesta hipótese. A consideração de processos persistentes na área de finanças, mesmo apoiada por matemáticos de renome, por exemplo, MANDELBROT e PRIGOGYNE, sofreu vários revezes como também a que considera os processos financeiros aleatórios (ALEXANDER, 2005). É importante observar que a discussão sobre se os processos de formação de preços têm memória curta, o que os conduziriam a ser processos aleatórios; ou de passado distante, que resultaria em processos persistentes, ainda não foi bem resolvida.

Assim, a metodologia adotada para a análise de séries temporais de sistemas dinâmicos

não-lineares consistirá em:

- aplicar testes BDS, com distribuição assintótica padrão ($\hat{R} < 1.96$, com intervalo de confiança 0,05) para investigar a dependência temporal; e o teste de HSIEH, com distribuição assintótica padrão, para detectar não-linearidades aditivas (na média) e multiplicativas (na variância);
- estimar o atrator por meio da dimensão da correlação a partir da reconstrução dinâmica da série temporal escalar observada para verificar o nível de complexidade do processo, sinalizando se o mesmo é estocástico ou caótico;
- determinar o expoente de LYAPUNOV para analisar a sensibilidade do processo às condições iniciais. Este parâmetro sinaliza se o processo é estocástico ou caótico.

Uma questão prática é a caracterização do processo em aleatório ou caótico com a intenção de estabelecer uma metodologia de modelagem.

4.4.2 Dimensão da correlação

A dimensão dos estados ativos pode ser estimada por meio da dimensão de correlação [GRASSBERGER and PROCACCIA, 1983a, GRASSBERGER and PROCACCIA, 1983b]. É uma medida no espaço de fase com o objetivo de determinar a complexidade do sistema gerador, estimando o número de variáveis que afetam a evolução futura do sistema. Em [TARAMASCO and ISABELLE, 1997] é observado que um sistema dinâmico de natureza caótica só existe se tem um atrator de dimensão finita, mesmo que este seja fracionário. Caso contrário, tem-se um processo aleatório. A qualidade do atrator reconstruído, entendida como quão bem esta trajetória estimada expressa o comportamento dinâmico real do sistema, é bastante sensível ao tempo de atraso escolhido (lag). Na prática, se L é muito pequeno os atratores são fechados e mal definidos, caso seja muito grande geram valores dispersos. Quando é adequado gera atratores abertos e bem definidos.

Observa-se que o espaço de fases consiste em um sistema de coordenadas associado às variáveis independentes (explicativas) que descrevem a dinâmica do sistema. Por exemplo, no caso de um pêndulo simples, o seu espaço de fases é dado pelas coordenadas compostas por sua posição e sua velocidade. O atrator é a representação da dinâmica de um sistema no espaço de fases.

O conceito de dimensão da correlação (d_c) estimado, calculado por meio da integral da correlação, pode ser assumido como a dimensão efetiva do sistema dinâmico. Esta medida sinaliza o número de modos ativos, ou seja, o número efetivo de graus de liberdade do sistema. Este parâmetro fornece uma indicação se o sistema complexo é estocástico (muitos graus de liberdade) ou é determinista (poucos graus de liberdade). Para que um sistema seja caótico é também necessário que exista uma sensível dependência às condições iniciais, além da existência de uma baixa dimensão fractal.

Assim, faz-se a generalização do conceito da dimensão inteira dos objetos para a dimensão fracionária, ou seja, uma dimensão não inteira. As dimensões inteiras (por exemplo: uma, duas, três ou mais dimensões inteiras) são estabelecidas facilmente e são intuitivamente óbvias. Como por exemplo, a medida do volume V, que varia de acordo com a equação:

$$V \approx \tau^d \tag{4.13}$$

em que τ é o comprimento numa determinada escala, por exemplo, o comprimento de um lado de um cubo ou do raio de uma esfera, e d é a dimensão do objeto. Para um fractal generalizado, é natural supor que uma relação como a equação 4.13 é verdadeira e pode evoluir para a seguinte relação:

$$d = \frac{\log V}{\log \tau}. \tag{4.14}$$

E para uma série no tempo observada em R^1, representada por $\{y_t\}_{t=1}^N$, TAKENS (1981) demonstrou que para um M suficientemente grande, dentro de certas condições genéricas de medidas, um sistema com atrator de dimensão d_c em seu espaço de estados, o subespaço $\Re^{d_c}$ deste atrator estará contido no espaço de imersão $\Re^M$.

Para calcular d_c, em [GRASSBERGER and PROCACCIA, 1983a] foi sugerido uma expressão intuitiva para a função de correlação integral desta série, que pode ser representada por

$$C(\tau, N) = (1/N^2)[pares(i,j) \mid (\parallel v_i - v_j \parallel \leq \tau)]. \tag{4.15}$$

Matematicamente, pode ser expressa por

$$C(\tau, N) = (1/N^2)(\sum_{i=1}^{N} \sum_{j=i+1}^{N-n} H(\tau - \parallel v_i - v_j \parallel) \tag{4.16}$$

em que $H(X)$ é a função degrau e $\parallel . \parallel$ pode ser a norma euclidiana ou qualquer outra norma considerada mais conveniente.

Entretanto, muitas vezes, este método apresentava distorções tanto para sistemas estocásticos como para sistemas deterministas não-lineares, principalmente, para aqueles fortemente autocorrelacionados. As referências [THEILER, 1986, THEILER, 1987] apresentaram extensões ao método anterior, corrigindo praticamente as distorções e incrementando computacionalmente o algoritmo de cálculo. Para corrigir as distorções, a equação 4.16 foi modificada para

$$C(\tau, N, W) = (1/N^2)(\sum_{i=W}^{N} \sum_{j=i+1}^{N-n} H(\tau - \parallel v_i - v_j \parallel) \tag{4.17}$$

em que W é um parâmetro que incrementa a convergência do algoritmo tradicional de GRASSBERGER e PROCCACIA. O algoritmo apresentado em THEILER (1986) calcula as distâncias entre todos os pares de pontos, exceto aqueles que estão mais próximos no tempo do que o valor W assumido. Assim, elimina as distorções que são comuns ao algoritmo original de GRASSBERGER e PROCCACIA. Em Theiler (1986) é recomendado que W seja igual ao *lag* L. Note-se que caso $W = 1$ esta equação volta à equação original de GRASSBERGER e PROCCACIA. Quando o sistema é estocástico, a equação anterior pode ser reescrita, substituindo a somatória interna sobre a função de Heaviside pelo valor da esperança, resultando em

$$C(\tau, N, W) = (1/N^2)(\sum_{n=W}^{N} (N-n)P(\parallel v_{i+n} - v_i \parallel \leq \tau). \tag{4.18}$$

Note-se que não é estritamente necessário, mas facilita a análise entre os valores x_i e x_{i+L}, caso seja imposto que $L >> \tau$. Isto permitirá assumir que x_i e x_{i+L} sejam independentes e simplifica a expressão da probabilidade de que dois vetores são separados por uma distância menor que τ.

Caso os pontos sejam distribuídos uniformemente dentro de um objeto, esta soma é proporcional ao volume da intersecção de uma esfera de raio τ com este objeto, e $C(\tau, N, W)$ é proporcional à média deste volume. Comparando com a equação 4.17, tem-se

$$C(\tau, N, W) \approx \tau^{d_c} \tag{4.19}$$

em que d_c é a dimensão do objeto. Considerando a equação (4.17), é natural definir d_c como

$$d_c(\tau, N, W) = \lim_{\tau \to 0} \lim_{N \to \infty} \frac{\log C(\tau, N, W)}{\log \tau}. \tag{4.20}$$

Caso as derivadas existam, resulta na equação

$$d_c(\tau, N, W) = \lim_{\tau \to 0} \lim_{N \to \infty} \frac{d[\log C(\tau, N, W)]/d\tau}{d(\log \tau)/d\tau}. \tag{4.21}$$

A escala de $C(\tau, N)$ é escolhida de modo que melhor expresse uma estimativa do volume médio de um objeto ajustado dentro de uma esfera de raio τ, em relação a um determinado ponto, ou seja, preferivelmente uma estimativa da probabilidade que dois pontos escolhidos aleatoriamente estejam dentro de uma distância τ, um do outro. A diferença entre o volume e a probabilidade é somente uma constante de proporcionalidade. Caso os pontos sejam distribuídos uniformemente, esta constante desaparece nos limites da equação 4.21. A razão para se escolher a probabilidade em detrimento do volume é que o conceito de dimensão ainda continua fazendo sentido e normalmente generaliza melhor para situações nas quais os pontos da amostra não são distribuídos uniformemente dentro do objeto.

Pode-se dizer, dessa maneira, que a dimensão de correlação d_c é uma medida de densidade (ou dispersão) do atrator dentro de um espaço de fase. Assim, a integral da correlação $C(\tau, N, W)$ mede o número de pontos em uma esfera de raio τ no $\Re^M$, conforme a defasagem da série no tempo aumenta. A dimensão da correlação d_c mede a taxa de crescimento de $C(\tau, N, W)$ com relação a τ. Assim, a dimensão de correlação d_c é normalmente estimada encontrando-se uma aproximação finita para cada uma das integrais da correlação $C(\tau, N, W)$ para diferentes valores de τ e, logo a seguir, é traçado o gráfico $\log(C(\tau, N, W))$ versus $\log(\tau)$. Um valor típico da integral de correlação conterá uma região da escala sobre a qual a inclinação deste gráfico permanece relativamente constante. Isto produz uma função que é constante sobre a região da escala e o gradiente desta região do gráfico deve se aproximar da dimensão da correlação.

Números de amostras necessárias	Dimensão da correlação estimada (d_c)
1.000	7,83
2.000	7,94
5.000	8,30
10.000	8,56
30.000	9,11
100.000	9,73

Tab. 4.1: Amostras necessárias para estimar a dimensão da correlação (d_c)

Infelizmente, há diversos problemas para se determinar a dimensão da correlação. O mais óbvio destes é que a escolha da região de escala é inteiramente subjetiva. Para muitas séries de dados, uma ligeira mudança na região usada pode conduzir a resultados substancialmente diferentes. Assim, para uma quantidade de dados relativamente pequena ou de dimensão elevada, o gráfico poderá saltar de forma irregular para valores pequenos de τ. Para evitar problemas, olha-se preferivelmente para o comportamento deste gráfico para valores moderados de τ. Entretanto, este método continua bastante popular, principalmente por causa de sua simplicidade computacional, embora existam outros métodos mais sofisticados que incluem testes estatísticos, como o de KOLMOGOROV-SMIRNOV.

Outra questão importante é a quantidade de amostras necessárias (N) para estimar uma determinada dimensão de correlação d_c. Em [SMITH, 1988] foi sugerido que, para uma precisão de 5%, seriam necessárias a quantidade de amostras constantes na Tabela 4.1.

Observa-se que o número de amostras necessárias aumenta substancialmente com o aumento da dimensão a ser estimada. Constata-se o fenômeno da maldição da dimensionalidade. Logo, para dimensões elevadas, torna-se muito difícil estimar precisamente a dimensão da correlação. Uma técnica que pode ser utilizada para incrementar o cálculo deste parâmetro é o *bootstrap*.

4.4.3 Expoente de Lyapunov

O expoente de LYAPUNOV (λ) é um parâmetro de caracterização dinâmica do processo. Estima a taxa de divergência de órbitas vizinhas (e consecutivas), quantificando a sensibilidade do sistema a variações nas condições iniciais. Analogamente, pode-se dizer que este expoente fornece uma indicação de quão rápido perde-se informação, movendo-se ao longo da trajetória no espaço de estados. Em sistemas caóticos, associados a um atrator estranho, a dependência à variação das condições iniciais implica na existência de pelo menos um expoente positivo.

Em séries temporais, o ponto de partida para o cálculo dos expoentes de LYAPUNOV é o atrator reconstruído em uma dimensão de imersão adequada. Uma vez reconstruído o atrator, define-se uma trajetória fiducial a partir da seqüência de vetores reconstruídos. A seguir, deve-se analisar o que ocorre com pontos vizinhos desta trajetória. Com as informações sobre as taxas de divergência destes pontos, pode-se obter, então, estes expoentes.

O inverso do expoente de LYAPUNOV dá-nos a capacidade de previsão existente. Sensível dependência às condições iniciais e dimensão fractal finita (estimada pelo valor da dimensão de correlação) são duas condições essenciais para averiguar a existência de um sistema caótico.

Daí advém o fato da previsão a longo prazo ser praticamente impossível dadas as bifurcações e efeitos de realimentação existentes num sistema caótico.

Existem vários métodos para o cálculo de tais expoentes, os quais diferem na maneira de analisar a dinâmica ao longo da trajetória fiducial. Entre os métodos mais conhecidos estão os de: [WOLF et al., 1985] e [ECKMANN and RUELLE, 1992]. Atualmente, o método de [SHINTANI and LINTON, 2004] está sendo considerado entre os mais eficientes e é capaz de fornecer o intervalo de confiança.

Neste trabalho, o maior expoente positivo de LYAPUNOV será estimado a partir de um método que trata da estimativa destes expoentes não negativos de uma série experimental. Este expoente é obtido assumindo condições iniciais independentes e que o limite existe.

$$\lambda_i := lim_{n\to\infty}\frac{1}{n}ln[|m_i(t)|], i = 1, 2, ..., n. \tag{4.22}$$

O expoente de LYAPUNOV pode ser encontrado em um ponto de equilíbrio y_{eq}. Seja $\{\lambda_i\}_{i=1}^n$ e $\{\eta_i\}_{i=1}^n$ os autovalores e autovetores respectivamente de $Df(y_{eq})$. A matriz de transição neste caso é

$$\Psi_t(y_{eq}) = e^{Df(y_{eq})t}$$

e segue que o variacional $m_i(t) = e^{\lambda_i t}$ e

$$\lambda_i = lim_{t\to\infty}\frac{1}{t}ln|e^{\lambda_i t}|. \tag{4.23}$$

$$lim_{t\to\infty}\frac{1}{t}Re[\lambda_i]t = Re[\lambda_i].$$

Assim, neste caso especial, os expoentes de LYAPUNOV são iguais às partes reais dos autovalores no ponto de equilíbrio. Quando este expoente é positivo as trajetórias divergem e quando é negativo convergem. Sendo a função f desconhecida, ou seja, tudo o que se observa são as realizações da série no tempo, então devem ser projetados algoritmos para se estimar estes expoentes. Neste trabalho, este expoente é calculado por meio de um algoritmo similar ao de [WOLF et al., 1985], que computa a média exponencial do crescimento da distância entre órbitas vizinhas por meio do erro de previsão. O incremento do erro de predição versus o tempo de predição permite estimar o maior expoente.

Isto nos dá uma medida quantitativa da estabilidade da série, permitindo observar como ela converge ou diverge no tempo. Quanto mais negativo é o expoente de LYAPUNOV, mais rápido a série converge para os valores finais, quando o expoente é positivo, o sistema apresenta comportamento caótico.

Existem métodos matemáticos analíticos, como o apresentado em [BENETTIN et al., 1980], para a determinação do expoente de LYAPUNOV. Esta análise não será implementada neste trabalho. Em relação aos métodos empíricos, existem pontos importantes que devem ser observados em relação a este algoritmo:

- Conclusões erradas podem ser tiradas mesmo com conjuntos de dados grandes.

- Os dados não devem conter muito ruído por que os métodos não conseguem detectar o caos subjacente, mesmo que a quantidade de ruído seja pequena.
- As séries no tempo devem ser estacionárias para que os resultados sejam corretamente interpretados.

Assim, faz-se necessário uma grande quantidade de dados pré-branqueados para que esses algoritmos possam ser utilizados eficazmente.

Em [VANDROVYCH, 2005] a dinâmica de seis taxas de câmbio dos principais países desenvolvidos são analisadas. As estimativas da dimensão de correlação indicam a complexidade elevada em toda a série, sugerindo que as séries são processos estocásticos ou processos deterministas com dimensões elevadas. Embora tenha sido obtido um número de estimativas positivas do exponente de LIAPUNOV, são valores muito pequenos e o autor acredita que é mais apropriado interpretar estes dados como um indicador de origem estocástica da série.

4.5 Identificação de *clusters* de padrões temporais com capacidade preditiva

A partir da reconstrução dinâmica, faz-se a mineração de dados (informações) na série temporal em que se busca principalmente a identificação de *clusters* de padrões temporais com capacidade preditiva. Na mineração de dados para a previsão de séries temporais existem basicamente duas abordagens: a que não utiliza conhecimento a priori [POVINELLI et al., 1999]; a que utiliza conhecimento a priori sobre os padrões temporais e representa estes padrões temporais por meio de estruturas pré-definidas. Nesta tese utiliza-se a primeira abordagem.

A mineração de dados consiste no processo de extrair informações não conhecidas *a priori*, válidas, utilizáveis, de grandes bancos de dados e, então usá-las para a tomada de decisões cruciais nos negócios. Uma boa analogia pode ser feita entre a mineração de dados e a de ouro. Esta busca pepitas de ouro e a outra procura pepitas de informação. Um ponto importante é que estas pepitas (padrões) podem levar a uma grande jazida de ouro (evento). Enquanto o ouro está escondido embaixo da terra as informações estão escondidas nos dados. Para um mineiro experiente o tamanho das pepitas (padrões) faz uma grande diferença na abordagem utilizada para encontrar a jazida de ouro, ou seja, ele segue os sinais positivos, interpretando-os a partir de conhecimento *a priori*. Entretanto, se uma empresa está procurando ouro ou petróleo, o processo de mineração é diferente. Isto sinaliza que é importante definir claramente o que se procura descobrir, ou seja, qual é o evento. Assim, caso este não seja bem definido, não se sabe se foram encontrados indícios de ouro ou de petróleo.

Outro ponto importante é que o mineiro aprende aonde procurar o ouro. Ele aprende como os outros mineiros tiveram sucesso. Similarmente, tem-se que saber aonde se pode encontrar pepitas de informação. Em séries temporais, estas sinalizações que identificam estas informações são os padrões temporais. Padrão, pela estrutura que o identifica, temporal, pela natureza temporal do problema. O ponto comum entre a mineração de dados e a de ouro é que os padrões temporais que sinalizam a ocorrência de eventos não precisam ser perfeitos, só necessitam que contribuam para a descoberta das informações. Observa-se que eventos, em séries temporais, são definidos como ocorrências importantes, como por exemplo a ocorrência de abalos sísmicos.

Os conceitos utilizados na mineração de dados de séries temporais serão abordados em seguida.

4.5.1 Padrão temporal, *cluster* de padrões temporais e evento

Um evento em uma série temporal é uma ocorrência importante e, como já mencionado anteriormente, é necessário defini-lo claramente para que se possa identificar padrões temporais associados ao mesmo. Um padrão temporal é uma estrutura escondida na série temporal que é característica e é capaz de possibilitar a previsão de eventos. É definido como um vetor real $\boldsymbol{p}$ de comprimento M, ou seja, este vetor será representado como um ponto em um espaço M dimensional nos números reais, por exemplo, $\boldsymbol{p} \in R^M$. Os ruídos podem fazer com que os padrões temporais não sejam perfeitamente iguais às observações da série temporal que precedem o evento. Para superar esta limitação, um *cluster* de padrões temporais de série temporal univariada é definido como um conjunto de pontos dentro de uma hiperesfera de dimensão M, raio δ e centro $\boldsymbol{a}$, ou seja

$$\mathbf{P} = \{\mathbf{a} \in R^M : d(\boldsymbol{p}, \mathbf{a}) \leq \ \delta\} \tag{4.24}$$

em que d é uma distância ou métrica definida no espaço.

As séries temporais utilizadas para ilustrar os conceitos apresentados nesta seção serão apresentados no espaço de fase de ordem dois para facilitar a visualização das informações.

A Figura 4.3 apresenta o gráfico dos dados brutos e da primeira diferença do terremoto Nisqually, estação de Olympia, WA, USA, em 28 de fevereiro de 2001. A variável analisada é a aceleração do terremoto em cm/seg/seg (eixo y) e o tempo é em segundos (eixo x). Neste gráfico pode-se observar claramente o pico do terremoto e o padrão de comportamento típico deste evento.

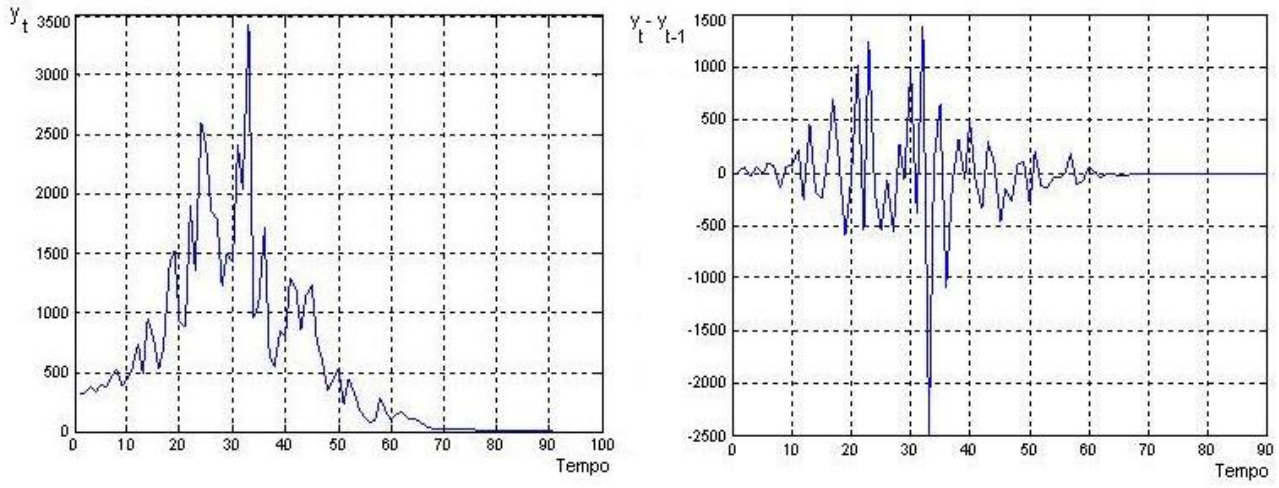

Fig. 4.3: Dados brutos e a primeira diferença do terremoto Nisqually, estação de Olympia, WA, USA, em 28 de fevereiro de 2001

A Figura 4.4 apresenta o gráfico dos dados brutos e dos retornos da taxa de câmbio brasileira em relação ao dólar americano para o período de 22 de março de 2002 a 03 de maio de 2004. Este período foi marcado por variações bruscas na taxa de câmbio devidas principalmente à eleição do Presidente Lula. Observa-se que ocorre um padrão de comportamento típico das séries financeiras, ou seja, percebe-se a formação de conglomerados de valores extremos, indicando a presença de heterocedasticidade condicional.

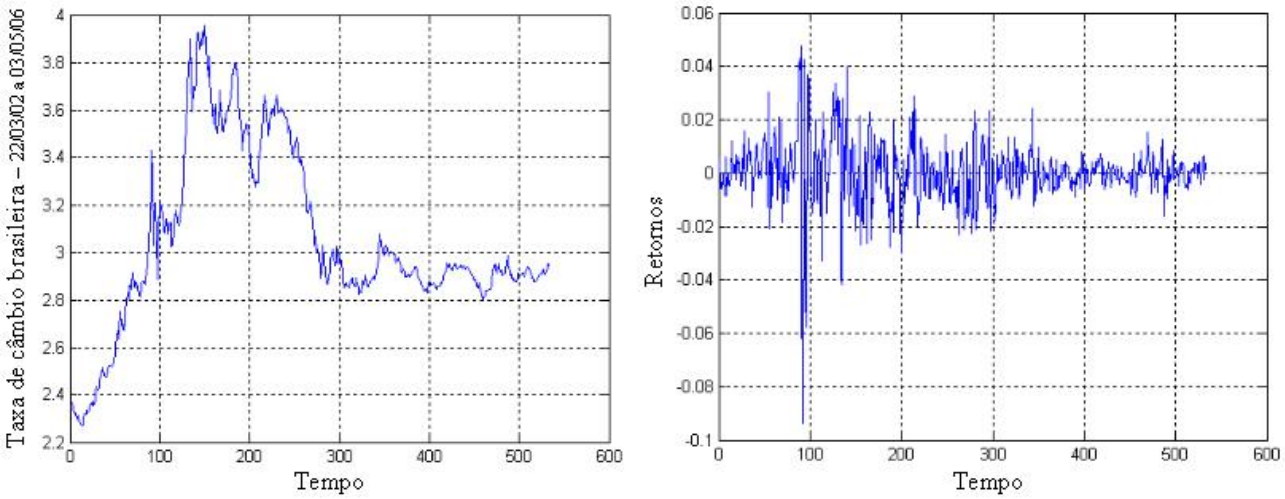

Fig. 4.4: Dados brutos e retornos da taxa de câmbio brasileira em relação ao dólar americano para o período de 22 de março de 2002 a 03 de maio de 2004

É importante usar uma notação em que o tempo seja particionado em passado, presente e futuro. Os padrões temporais ocorrem no passado e se completam no presente. Os eventos ocorrem no futuro. Um espaço de fase reconstruído, como já visto anteriormente, permite que seqüências possam ser comparadas a padrões temporais. Por exemplo, para um *lag* L, as observações $\{\mathbf{y}(t+(M-1)L), ..., \mathbf{y}(t-2L), \mathbf{y}(t-L), \mathbf{y}(t)\}$ formam uma seqüência que pode ser comparada a padrões temporais, em que $\mathbf{y}(t)$ representa a informação presente e $\{\mathbf{y}(t+(M-1)L), ..., \mathbf{y}(t-2L), \mathbf{y}(t-L)\}$ representam as informações passadas. Seja L um inteiro positivo, se t representa o índice atual, então $t-L$ é um índice no passado e $t+L$ é um índice no futuro. Assim, o tempo é particionado em passado, presente e futuro.

4.5.2 Função de caracterização de evento

Para ligar os padrões temporais (passado e presente) aos eventos (futuro) uma função de caracterização de evento é criada. Esta função é definida *a priori*, depende da aplicação e representa as possibilidades de um determinado evento ocorrer em um determinado tempo $t+i$ a partir do tempo t. Um exemplo simples de função de caracterização de evento na previsão de séries temporais é $g(t) = y_{t+1}$. Esta função captura o objetivo de prever séries temporais um passo a frente no futuro. No mercado financeiro, pode-se utilizar uma função de caracterização de evento para decidir sobre a compra ou venda de um ativo. Por exemplo, utilizando a taxa de retorno, apresentada em seguida:

$$g(t) = \frac{y_{t+1} - y_t}{y_t} \tag{4.25}$$

que atribui a percentagem de variação do preço do ativo para o próximo dia.

4.5.3 Espaço de fase estendido

O conceito de espaço de fase estendido deriva dos conceitos de função de caracterização de evento e de espaço de fase, no qual a série temporal é reconstruída. Este espaço tem dimensão $M+1$, ou seja, é composto pelo espaço de fase e tem a função de caracterização de evento g(.) como a dimensão extra. Cada ponto do espaço de fase estendido é um vetor $\mathbf{Y}_t, g(t) \in R^{M+1}$.

Antes de implementar o gráfico do espaço de fase estendido é necessário definir o que se chama função objetivo.

4.5.4 Função objetivo para padrões temporais univariados

Esta função caracteriza a eficácia que um *cluster* de padrões temporais tem para prever um evento, ordenando os *clusters* de padrões temporais P de acordo com suas habilidades em prever os eventos. É construída de tal maneira que o P^* ótimo atinge a melhor previsão.

A forma da função objetivo é dependente da aplicação e várias funções objetivo diferentes podem alcançar o mesmo resultado. Entretanto, antes de apresentar a função objetivo que será utilizada nesta tese, são necessárias algumas definições.

O conjunto indexador Λ é o conjunto de todos os índices t dos pontos no espaço de fase.

$$\Lambda = \{t : t = (M-1)L + 1, ..., N\} \tag{4.26}$$

em que $(M-1)L$ é o último vetor da reconstrução dinâmica e N é o número de observações da série no tempo.

O conjunto indexador Q é o conjunto de todos os índices t dos padrões temporais que estão dentro do *cluster* de padrões temporais, cuja representação matemática é

$$Q = \{t : \mathbf{Y}_t \in P, t \in \Lambda\}. \tag{4.27}$$

Similarmente, $\overline{Q}$, é o complemento de Q. A média dos valores de $g(t)$ dos pontos no espaço de fase que estão dentro do *cluster* P é

$$\mu_Q = \frac{1}{c(Q)} \sum_{t \in Q} g(t) \tag{4.28}$$

em que $c(Q)$ é a cardinalidade de Q.

A média dos valores de $g(t)$ dos pontos no espaço de fase que estão fora do *cluster* P é

$$\mu_{\overline{Q}} = \frac{1}{c(\overline{Q})} \sum_{t \in Q} g(t). \tag{4.29}$$

A média dos valores de $g(t)$ em todos os pontos no espaço de fase é

$$\mu_y = \frac{1}{c(\Lambda)} \sum_{t \in \Lambda} g(t). \tag{4.30}$$

As variâncias correspondentes são

$$\sigma_Q^2 = \frac{1}{c(Q)} \sum_{t \in Q} (g(t) - \mu_Q)^2. \tag{4.31}$$

$$\sigma_{\overline{Q}}^2 = \frac{1}{c(\overline{Q})} \sum_{t \in Q} (g(t) - \mu_{\overline{Q}})^2. \tag{4.32}$$

$$\sigma_y^2 = \frac{1}{c(\Lambda)} \sum_{t \in \Lambda} (g(t) - \mu_y)^2. \tag{4.33}$$

A partir destas definições, vários tipos de funções objetivo podem ser construídas. Esta função objetivo tem a capacidade de ordenar os *clusters* de padrões temporais de acordo com suas habilidades para auxiliar na predição de séries temporais e, no mínimo, gerar alguns eventos. Como o objetivo deste capítulo é somente escolher os *clusters* que serão utilizados para determinar os centros da rede RBF, logo eles não serão apresentados com as suas respectivas capacidades preditivas.

$$f(C) = \begin{cases} \mu_Q & \text{se} \quad \frac{c(Q)}{c(\Lambda)} \geq \beta \\ (\mu_Q - g_0)\frac{c(Q)}{\beta c(\Lambda)} + g_0 & \text{se} \quad \frac{c(Q)}{c(\Lambda)} < \beta \end{cases}$$

em que β é a percentagem mínima de cardinalidade do *cluster* de padrões temporais e g_0 é a mínima capacidade de ocorrência dos pontos no espaço, por exemplo,

$$g_0 = min\{g_t : t \in \Lambda\}. \tag{4.34}$$

Logo, a função objetivo representa um valor ou *fitness* de um *cluster* de padrões temporais.

A Figura 4.5 apresenta o gráfico do espaço de fase estendido dos dados brutos e da primeira diferença do terremoto Nisqually, estação de Olympia, WA, USA, em 28 de fevereiro de 2001. No gráfico da esquerda, dos dados brutos, observa-se o pico do terremoto, ou seja, a maior aceleração (círculo A). Já no gráfico da direita, das primeiras diferenças, observa-se as duas maiores acelerações (círculo A) e a maior desaceleração (círculo B), sinalizando o final do período mais crítico.

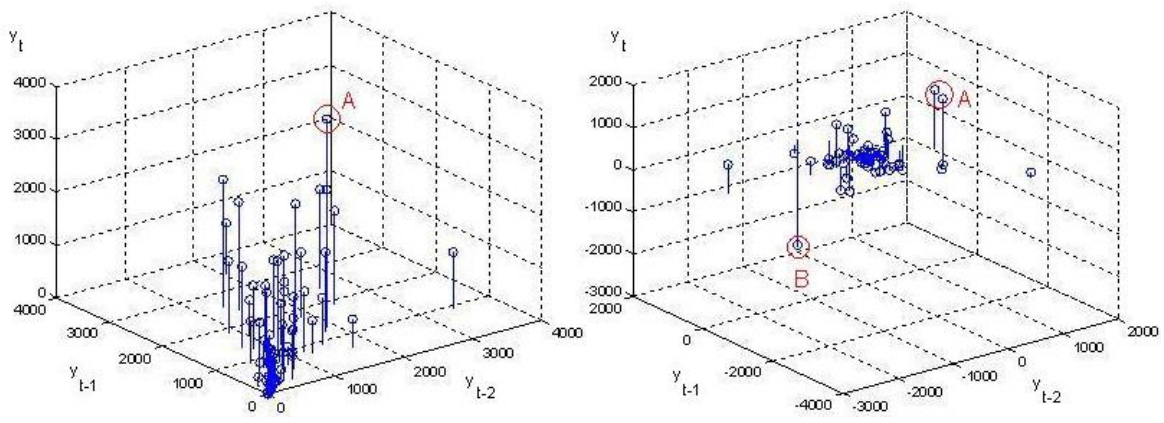

Fig. 4.5: Espaço de fase estendido dos dados brutos e da primeira diferença do terremoto Nisqually, estação de Olympia, WA, USA, em 28 de fevereiro de 2001

4.5.5 Função objetivo para padrões temporais multivariados

O procedimento para se fazer a reconstrução dinâmica de dados multicanais já foi apresentado anteriormente. Intuitivamente, sensores adicionais podem acrescentar informações, considerando que não estão captando estas informações da mesma variável de estado. A Figura 4.6 ilustra um método de MDST para séries temporais multivariadas, buscando identificar *clusters* de padrões temporais.

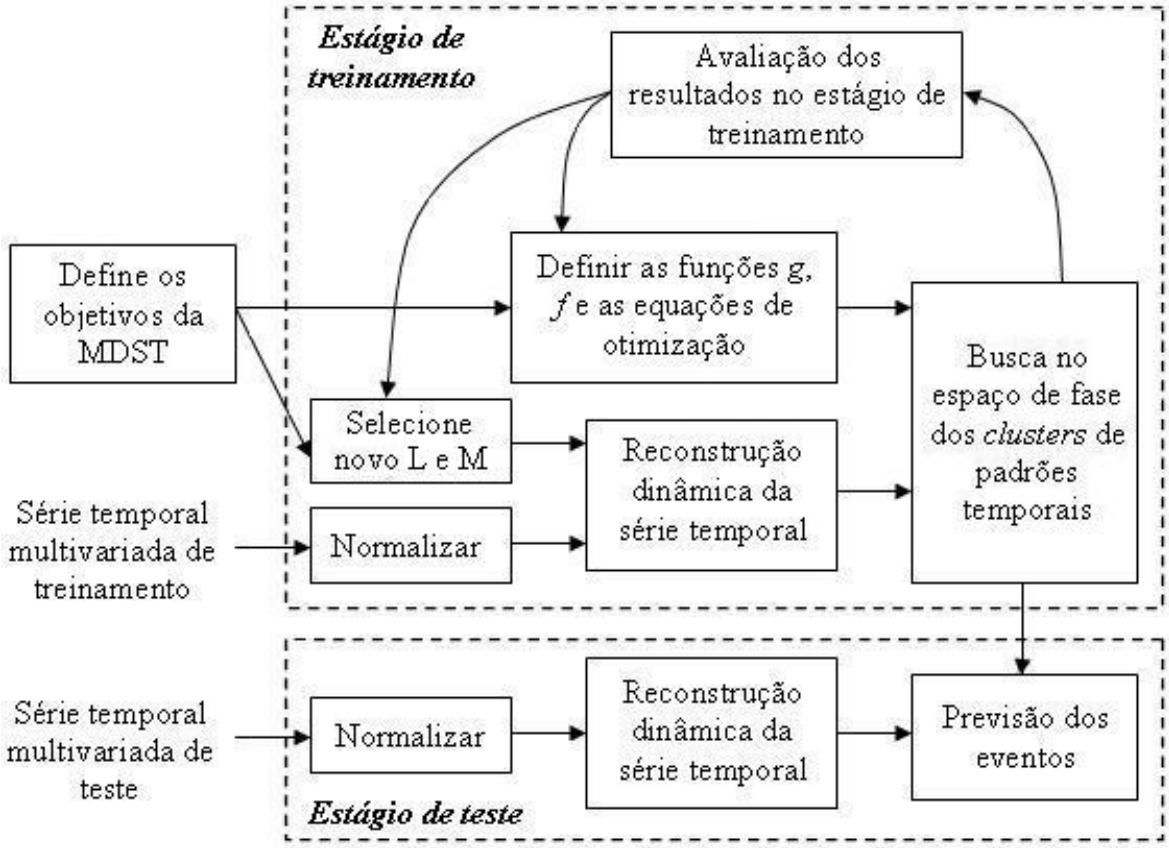

Fig. 4.6: Método de MDST para séries temporais multivariadas

Observa-se que é necessária a normalização, vide Figura 4.6, para forçar que cada variável esteja na mesma faixa (*range*). Esta normalização não muda a topologia do espaço de estados, mas mapeia cada série temporal no mesmo *range* e permite que se use similares tamanho de passo de busca para todas as dimensões do espaço de fase. A normalização auxilia nas rotinas de otimização e as constantes de normalização são retidas no estágio de treinamento para serem utilizadas nos eventos de predição no estágio de teste.

A função objetivo para incluir os pontos do espaço de fase dentro de cada *cluster* de padrões temporais é: $P_i \in C, i = 1, 2, ..., n$. A função utilizada é a mesma apresentada anteriormente, ou seja,

$$f(C) = \begin{cases} \mu_Q & \text{se} \quad \frac{c(Q)}{c(\Lambda)} \geq \beta \\ (\mu_Q - g_0)\frac{c(Q)}{\beta c(\Lambda)} + g_0 & \text{se} \quad \frac{c(Q)}{c(\Lambda)} < \beta \end{cases}$$

em que o indexador Q é generalizado e expresso por

$$Q = \{t : \mathbf{Y}_t \in P_i, t \in \Lambda\}. \tag{4.35}$$

em que $P_i \in C, i = 1, 2, ..., n$. Observa-se que $\overline{Q}$ é o conjunto de todos os índices t quando $\mathbf{Y}_t$ não está em nenhum dos $P_i \in C$.

Finalmente, a otimização pode ser dada a partir da formulação abaixo.

$$Max \ \ _{\mathbf{p}_{i,\delta}} \ \ f(C). \tag{4.36}$$

4.5.6 Escolha dos *clusters* para determinar os centros das redes RBF

A escolha dos *clusters* para determinar os centros das redes RBF é fundamental para garantir a qualidade das previsões por meio deste tipo de rede neural. Nesta seção, a identificação dos *clusters* será ilustrada via algoritmo *k-means* e algoritmo EM (*expectation maximization*). Estes métodos têm a característica de localidade geométrica.

No próximo capítulo serão apresentados os métodos de identificar *clusters* baseados no PCA e no algoritmo ARIA, com o foco na previsão de séries temporais via redes RBF. Os centros obtidos por meio da análise de componentes principais não tem a característica de localidade geométrica. O ajuste dos centros das funções de base por meio do PCA tem como primeiro passo calcular a matriz de covariância com o objetivo de captar a dinâmica temporal do processo associado à série.

O algoritmo ARIA (*Adaptive Radius Immune Algoritm*) [BEZERRA et al., 2004] tem a característica de localidade geométrica, mas permite uma visão geral mais completa e melhor dos dados porque preserva a densidade dos mesmos. A maioria dos métodos de clusterização não avalia as densidades locais dos dados. Por exemplo, os algoritmos SOM (Self-Organizing Maps), apresentados em [KOHONEN, 1988], e o aiNet apresentado em DE CASTRO e VON ZUBEN (2001) estão nesta categoria. Este problema será analisado com maior profundidade no próximo capítulo.

As Figuras 4.7 e 4.8 apresentam respectivamente os *clusters* dos dados brutos e das primeiras diferenças do terremoto Nisqually, estação de Olympia, WA, USA, em 28 de fevereiro de 2001, estimados pelo algoritmo *k-means* e EM .

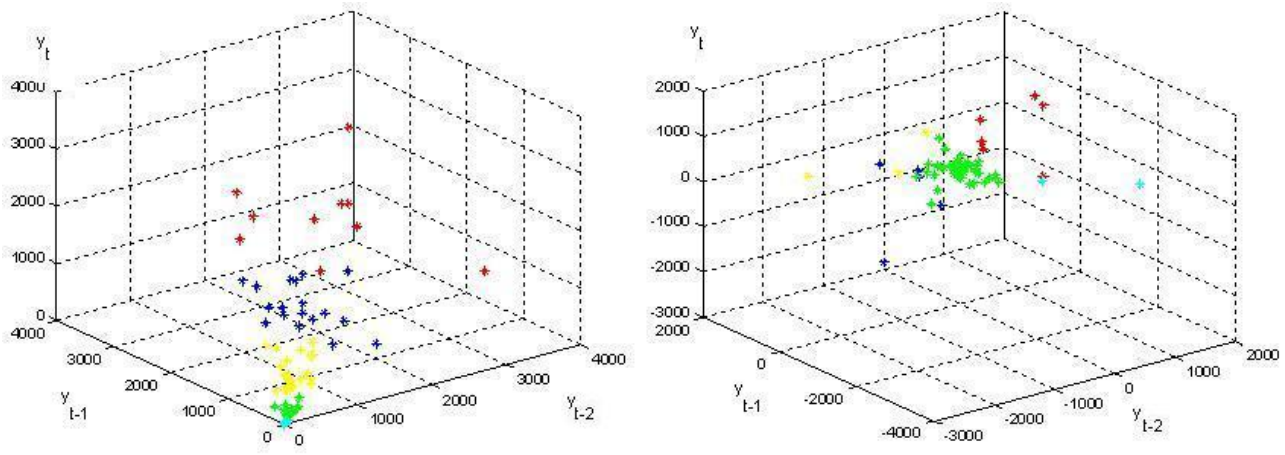

Fig. 4.7: *Clusters* estimados pelo algoritmo *k-means*, respectivamente dos dados brutos e das primeiras diferenças do terremoto Nisqually, estação de Olympia, WA, USA, em 28 de fevereiro de 2001

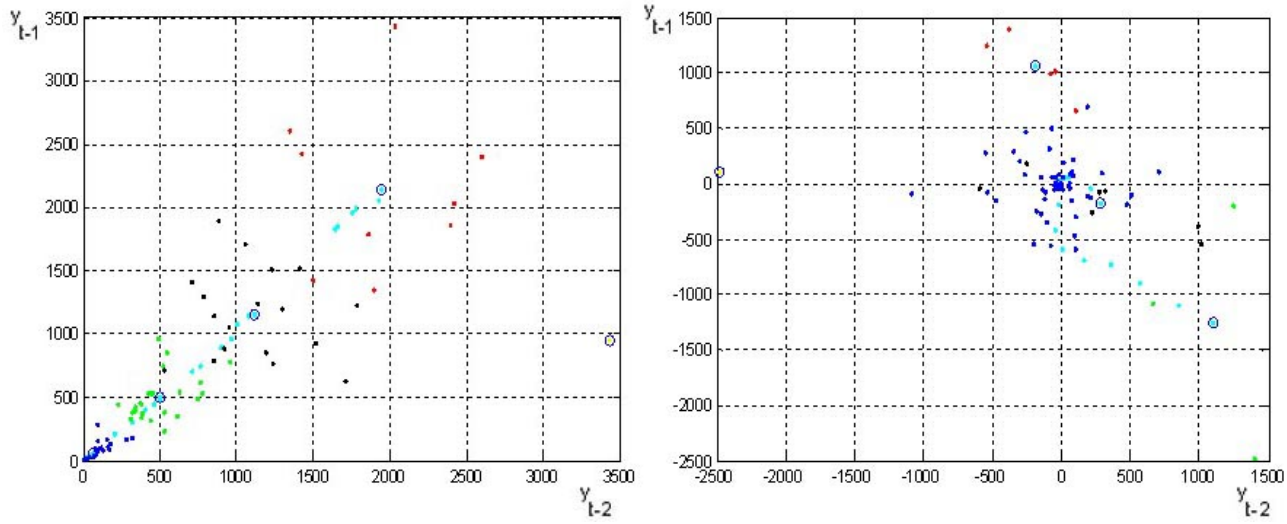

Fig. 4.8: *Clusters* estimados pelo algoritmo EM, respectivamente dos dados brutos e das primeiras diferenças do terremoto Nisqually, estação de Olympia, WA, USA, em 28 de fevereiro de 2001

As Figuras 4.9 e 4.10 apresentam os *clusters*, estimados pelo algoritmo *k-means* e EM, respectivamente dos dados brutos e dos retornos da taxa de câmbio brasileira para o período de 22 de março de 2002 a 03 de maio de 2004.

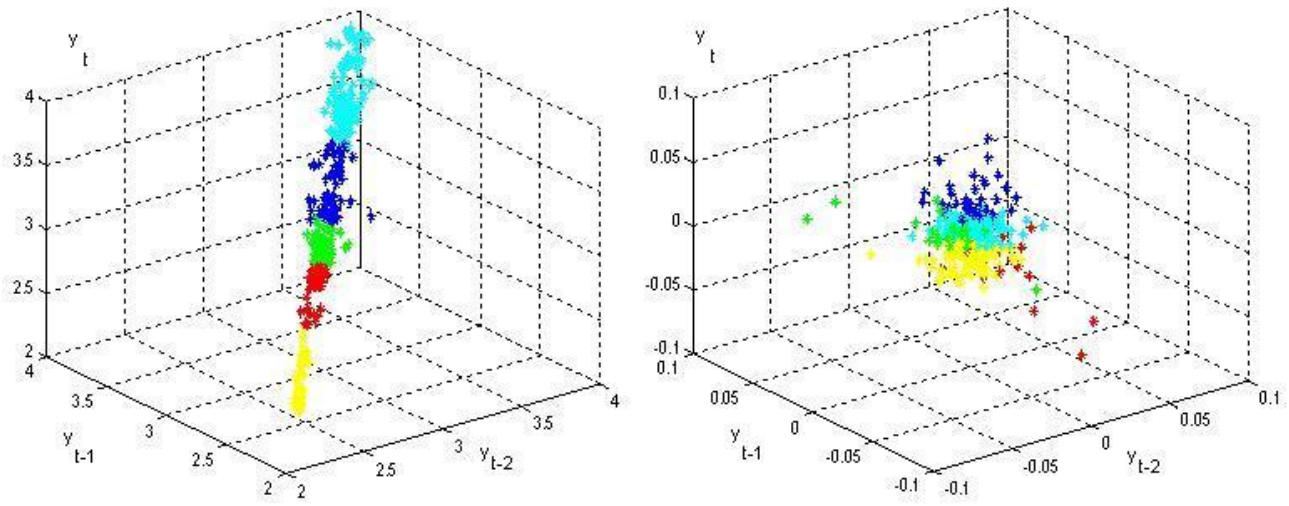

Fig. 4.9: *Clusters* estimados pelo algoritmo *k-means*, respectivamente dos dados brutos e dos retornos da taxa de câmbio brasileira em relação ao dólar americano para o período de 22 de março de 2002 a 03 de maio de 2004

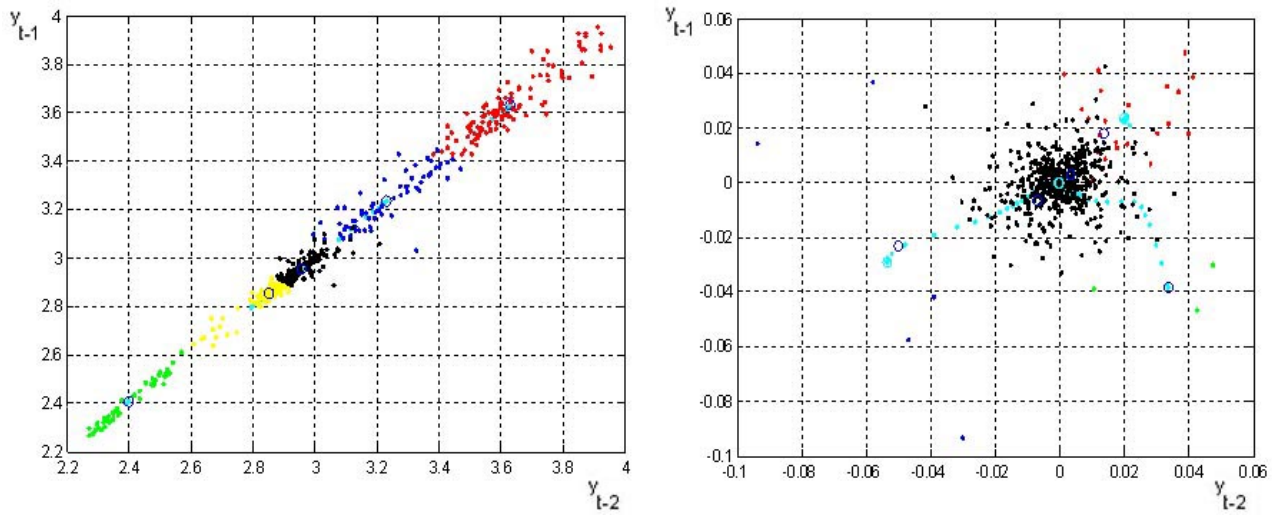

Fig. 4.10: *Clusters* estimados pelo algoritmo EM, respectivamente dos dados brutos e dos retornos da taxa de câmbio brasileira em relação ao dólar americano para o período de 22 de março de 2002 a 03 de maio de 2004

Os gráficos apresentados nesta seção ilustram a característica de localidade geométrica na identificação dos *clusters* via algoritmo *k-means* e algoritmo EM (*expectation maximization*). Estes métodos não identificam os modos dinâmicos do processo associado à série e não avaliam as densidades locais dos dados, logo não permitem uma visão geral mais completa e melhor dos dados. Assim, a escolha dos *clusters* para determinar os centros das redes RBF por meio destes métodos pode comprometer a qualidade das previsões via redes RBF.

No próximo capítulo será apresentado um método de clusterização que avalia as densidades

locais dos dados para determinar os centros de algumas das redes tipo RBF utilizadas nas previsões das séries temporais que serão analisadas neste trabalho. Finalmente, os resultados obtidos nas previsões de séries temporais por meio de redes RBF, com centros determinados via métodos de clusterização, podem sinalizar se existem ou não *clusters* de padrões temporais com capacidade preditiva.

Capítulo 5

Previsão de Séries no Tempo via Redes RBF

5.1 Introdução

A incorporação do tempo na estrutura de uma rede neural artificial (RNA) é condição necessária para capacitá-la a fazer previsões de séries temporais. A memória tem o papel importante de transformar uma rede estática em dinâmica, ou seja, uma rede tipo perceptron multi-camadas (PMC) estática pode incorporar memória (o tempo) em sua estrutura por meio de defasagens no tempo. Neste caso, existe uma separação clara de funções em que a PMC incorpora as não linearidades e a memória é responsável pelo tempo.

As classes de redes neurais artificiais (RNA) mais utilizadas para implementar modelos de previsão de séries temporais foram: as redes alimentadas adiante atrasadas no tempo (focadas e distribuídas) e as redes recorrentes com um ou mais laços de realimentação externos ou internos (Jordan e Elman). As PMCs focadas são adequadas somente para processos estacionários. As distribuídas e o perceptron de múltiplas camadas recorrente podem lidar com processos não estacionários. Entretanto, se o treinamento for feito pelo algoritmo *backpropagation* pode apresentar dificuldades típicas dos algoritmos de otimização baseados em gradiente. Estas dificuldades podem ser de velocidade de convergência ou susceptibilidade a mínimos locais, aumentando o esforço computacional e diminuindo a interpretabilidade e a transparência. As redes recorrentes também podem apresentar problemas de estabilidade ou de extinção de gradiente.

Nesta tese, utiliza-se a classe das redes RBF (*Radial Basis Function*) com regularização e centros determinados via PCA associada a uma janela de previsão inteligente dinâmica (JPID) que incorpora o tempo (dinâmica do processo) na previsão de séries temporais. Esta janela de previsão é inicialmente identificada por meio da reconstrução dinâmica e o seu ajuste final é via RNA. Implementa-se também uma rede RBF com centros determinados via algoritmo ARIA em que *clusters* de padrões temporais multivariados com capacidade preditiva são identificados a partir de todos os dados disponíveis, tendo um caráter mais probabilístico.

A classe das redes RBF tem origem no teorema apresentado em [MICCHELLI, 1986], propondo que o único pré-requisito para a inversão de uma matriz de regressão gerada por funções de base radial seria que as N amostras fossem distintas. A referência [POWELL, 1985] foi

quem primeiramente sugeriu as funções de base radial para resolver problemas de interpolação multivariada em que o número de centros era igual ao número de amostras. O artigo de [BROOMHEAD and LOWE, 1988] questionou a capacidade de generalização da estratégia anterior e criaram as redes neurais tipo RBF em que o número de centros era menor do que o número de amostras. Já o trabalho de [MOODY and DARKEN, 1991] contribuiu para a determinação dos centros das redes RBF e para a evolução da solução do problema de ajuste de curvas. Destaca-se também a contribuição do trabalho de [POGGIO and GIROSI, 1990a] na aproximação via regularização de mapeamentos mal-formulados.

O desempenho de uma rede RBF depende principalmente da escolha do número de funções de base e das respectivas posições dos centros. O procedimento mais simples para o treinamento da rede é assumir que as funções de ativação têm centros fixos e em seguida aplicar uma regressão linear para determinar os pesos da camada de saída. Esta estratégia geralmente não possibilita uma boa generalização. Neste trabalho, para contornar este problema, foram utilizados os seguintes algoritmos para determinar os centros: PCA e ARIA. O algoritmo ARIA (*Adaptive Radius Immune Algorithm*) avalia (preserva) a densidade dos *clusters* e tem caráter mais estocástico. O algoritmo baseado em PCA, com a variância controlada por meio de um fator adaptativo, para uma função de base tipo spline fina, preserva os modos dinâmicos (autovalores e autovetores) e tem caracter mais determinista.

Este capítulo tem o objetivo de apresentar uma metodologia para fazer previsões de curto prazo (um passo adiante) de sistemas não lineares e não estacionários por meio de redes RBF. O capítulo foi organizado como segue: na Seção 5.2 aborda-se o aprendizado como aproximação a partir de exemplos; na Seção 5.3 faz-se uma introdução às redes RBF exatas e generalizadas; na Seção 5.4 propõe-se uma metodologia para previsão de séries no tempo por meio de redes neurais RBF regularizadas por meio da função spline fina, com os centros ajustados via PCA e o espalhamento (variância) é controlado por meio de um fator adaptativo proposto; utiliza-se também uma rede RBF com centros ajustados por meio do algoritmo ARIA (redes imunes); na Seção 5.5 apresenta-se os testes estatísticos utilizados para auxiliar no treinamento e na avaliação da capacidade preditiva dos modelos de previsão.

5.2 Aprendizado como aproximação a partir de exemplos

Uma interpolação estrita considera que o conjunto $D = \{f(\mathbf{x}_i; y_i) \in \mathbf{X}, \mathbf{Y}\}_{i=1}^{N}$ é uma amostra de uma função multivariada f, que depende do espaço de entradas e de saídas, $\mathbf{X}$ e $\mathbf{Y}$. Considera também que existe alguma relação entre estas entradas e as saídas, e que os pares $(\mathbf{x}_i; y_i)$ são exemplos observados do processo. Portanto, existe um mapeamento $f : \mathbf{X} \rightarrow Y$ com a seguinte propriedade: $f(\mathbf{x}_i) = y_i$ com $i = 1, ..., N$. Entretanto, esta abordagem pode não ser a mais adequada para o treinamento das redes RBF já que a generalização pode ser pobre pelas seguintes razões: quando o número de amostras de treinamento é muito maior que o número de estados do processo físico isso torna o sistema indeterminado; a generalização também pode degradar porque os ruídos podem gerar variações enganosas.

A aprendizagem como um problema de reconstrução de uma hipersuperfície mal-formulado surgiu como uma solução para este problema de interpolação. Entenda-se como um problema bem formulado aquele que tem: existência, unicidade e continuidade. Caso alguma destas

condições não seja satisfeita, o problema será mal-formulado. Intuitivamente, pode-se dizer que um problema mal-formulado é aquele que, mesmo possuindo uma grande quantidade de dados, pode conter pouca informação acerca da solução do problema.

Uma função multi-dimensional de aproximação $F(\mathbf{w};\mathbf{x})$ pode aproximar uma função multi-dimensional de forma funcional $F(\mathbf{x})$ desconhecida a partir de exemplos, mesmo que estes sejam esparsos e contenham ruídos. O aprendizado como aproximação a partir de exemplos esparsos pode ser abordado como um problema de reconstrução de uma hipersuperfície. POGGIO e GIROSI (1990) observaram que uma representação exata não existe; entretanto, pode fornecer uma aproximação razoavelmente boa e geral.

Dois problemas principais surgem decorrentes desta abordagem:

- O problema da representação que consiste em saber qual função $F(\mathbf{x})$ pode ser aproximada efetivamente por qual função $F(\mathbf{w};\mathbf{x})$. Deste problema decorre a complexidade da aproximação (número de termos, escolha das escalas e altas dimensões).

- O problema da escolha do algoritmo para encontrar os valores ótimos dos parâmetros $\mathbf{w}$ para uma dada escolha de $F(\mathbf{w};\mathbf{x})$.

Estes problemas podem levar a uma função de aproximação $F(\mathbf{w};\mathbf{x})$ com representação pobre de $F(\mathbf{x})$, mesmo que os parâmetros $\mathbf{w}$ sejam otimizados. As soluções são geralmente sub-ótimas. Em particular, já que o problema de aproximar uma superfície a dados esparsos é mal-formulado, a regularização pode ser uma abordagem indicada. A teoria da regularização leva naturalmente à formulação do princípio do variacional, a partir do qual é possível derivar um esquema de aproximação bem conhecido, como o relacionado com as funções de base radial que geralmente são apropriadas para o aprendizado de máquina.

O aprendizado, ou generalização, significa que o modelo é capaz de estimar a função nos pontos do seu domínio $\mathbf{X}$ em que não se tem dados disponíveis. Isto significa estimar a função f entre os dados esparsos. Portanto, a partir deste ponto de vista, o problema do aprendizado é equivalente ao da reconstrução de uma hipersuperfície suavizada. A generalização não é possível sem suavização. Dessa forma, se o mapeamento for totalmente randômico não existirá generalização.

Um fenômeno físico que gera dados de treinamento (por exemplo, voz, abalo sísmico) é um problema direto bem-formulado. Entretanto, aprender a partir destes dados, um problema de reconstrução de uma hipersuperfície, geralmente é um problema inverso mal-formulado já que as condições de existência, unicidade e continuidade, em conjunto ou individualmente, podem ser violadas. Não há como superar este problema se não houver alguma informação *a priori* sobre o mapeamento entrada-saída, ou seja, não existe truque matemático que contorne a falta de informação.

No caso de processos que geram séries no tempo, na prática, primeiramente observa-se uma realização do processo e, a partir dela, faz-se a reconstrução dinâmica da forma

$$\mathbf{Y}_t = \{\mathbf{y}(t), \mathbf{y}(t+L), \mathbf{y}(t+2L), ..., \mathbf{y}(t+(M-1)L)\}$$

em que M é a dimensão de imersão (*embedding*) e L é o *lag*, como já visto no Capítulo 4. TAKENS (1981) propôs que esta representação preserva as propriedades topológicas do atrator.

Entretanto, em [BROOMHEAD and KING, 1985] foi observado que esta não é a única técnica que pode ser utilizada para representar o estado de um sistema dinâmico.

Portanto, na previsão de séries temporais, o mapeamento pode ser representado por

$$f_T : R^{LxM} \rightarrow R, \quad f_T(\mathbf{y}(t)) \rightarrow \mathbf{y}(t+T).$$

Vários autores [BROOMHEAD and LOWE, 1988, UTANS and MOODY, 1991] já aplicaram as redes RBF na predição de séries no tempo com relativo sucesso. Entretanto, existem dois problemas cruciais: a escolha das variáveis de entrada; estimar a dimensão do atrator. Estes dois assuntos já foram abordados nos Capítulos 3 e 4 desta tese, respectivamente.

5.2.1 Regularização como solução do problema de aproximação

No aprendizado via aproximação, o problema de aprender um mapeamento suave a partir de exemplos com ruídos é mal-formulado no sentido de que a informação contida nos dados geralmente não é suficiente para reconstruir um mapeamento único em regiões onde não há dados disponíveis. A característica crucial para este mapeamento é a suavização, sem a qual não há esperança de se conseguir a generalização. Para encontrar uma solução única para este mapeamento é necessário algum conhecimento *a priori* e, neste contexto, a suavização é uma informação importante. A teoria da regularização foi que unificou as pesquisas sobre este assunto. Segundo esta teoria, a solução de um problema mal-formulado pode ser obtida a partir do conceito de variacional que embute os dados e as informações *a priori*.

Portanto, considerando que o conjunto $D = \{f(\mathbf{x}_i; y_i) \in R^d \rightarrow R\}_{i=1}^N$ é uma amostra com ruído de uma função multivariada f, a solução do problema de ajustar esta função f aos dados D, via teoria da regularização [GOLUB et al., 1979, POGGIO and GIROSI, 1990b], utilizando um variacional, pode ser encontrada por meio da minimização do seguinte funcional

$$H(f) = \sum_{i=1}^{N}(y_i - f(\mathbf{x}_i))^2 + \lambda\psi(f) \tag{5.1}$$

em que $\sum_{i=1}^N(y_i - f(\mathbf{x}_i))^2$ é o termo do erro padrão, $\lambda\psi(f)$ é o termo de regularização, λ é um número positivo que é chamado usualmente de parâmetro de regularização e $\psi(f)$ é uma função de custo que restringe o espaço das possíveis soluções de acordo com alguma forma de conhecimento *a priori*. A forma mais comum de conhecimento *a priori* é a suavização. Em [POGGIO and GIROSI, 1990a] foi utilizada uma classe geral de funcional suavizado, invariante à translações e à rotações, expresso por

$$\psi(f) = \int_{R^d} \frac{|\overline{f}(\mathbf{s})|^2}{\overline{G}(\mathbf{s})} d\mathbf{s} \tag{5.2}$$

em que $\overline{f}$ e $\overline{G}$ são as transformadas generalizadas de Fourier de f e G, em que G é uma função de base radial definida positiva que tende a zero quando $s \rightarrow \infty$. Em GIROSI, JONES e POGGIO (1995) foi proposto que a função que minimiza o funcional da equação 5.1 é

$$f(\mathbf{x}) = \sum_{i=1}^{N} c_i G(\mathbf{x} - \mathbf{x}_i) + \sum_{j=1}^{K} d_j \gamma_j(\mathbf{x}) \tag{5.3}$$

em que N é o número de amostras de $\mathbf{x}$ e $\{\gamma_j\}_{j=1}^{K}$ é uma base no espaço polinomial de no máximo ordem m (a ordem de uma função definida positiva G) , e c_i e d_j são coeficientes que têm que ser determinados. Se a equação anterior é substituída na equação 5.1, a equação $H(f)$ passa a ser $H(\mathbf{c}, \mathbf{d})$, ou seja, uma função das variáveis c_i e d_j. Minimizando $H(\mathbf{c}, \mathbf{d})$ em função destas variáveis obtém-se uma equação linear

$$(G + \lambda I)\mathbf{c} + \Gamma^T \mathbf{d} = \mathbf{y} \tag{5.4}$$

$$\Gamma \mathbf{c} = 0$$

em que I é a matriz identidade, $(\mathbf{y})i = y_i$, $(\mathbf{c})i = c_i$, $(\mathbf{d})i = d_i$, $(\Gamma)ji = \gamma_j(\mathbf{x}_i)$ e $(G)ij = G(\mathbf{x}_i - \mathbf{x}_j)$. Exemplos clássicos de funções de base G são a gaussiana e a multiquadrática inversa. Observa-se que para $\lambda = 0$ as equações anteriores se tornam as equações das funções de base radial e as condições de interpolação $f(\mathbf{x}_i) = y_i$ são satisfeitas. Caso os dados tenham ruídos, o λ deve ser proporcional à quantidade de ruídos presente nos dados. O valor ótimo de λ pode ser determinado por meio da validação cruzada generalizada [HAYKIN, 1999].

5.2.2 Extensões ao procedimento de regularização

A técnica baseada em funções de base radial revelou-se como uma das mais apropriadas para aproximar funções em um espaço multidimensional. Entretanto, tem-se que superar dois problemas: capacidade de lidar com sistemas lineares mal-formulados; encontrar normas ponderadas adequadas.

Estes dois problemas são analisados em seguida:

a) **Sistemas lineares mal-formulados**: as técnicas iterativas via mínimos quadrados podem contornar as instabilidades numéricas. Portanto, a seminorma $\psi(f)$ da equação 5.1 é associada a uma função de Green G [HAYKIN, 1999], simétrica, coberta pelo Teorema de MICCHELLI. Logo, uma função de base radial com expansão dos mínimos quadrados é uma solução para este problema

$$f^*(\mathbf{x}) = \sum_{i=1}^{n} c_i G(||\mathbf{x} - \mathbf{t}_i||) \tag{5.5}$$

em que $\{\mathbf{t}_i\}_{i=1}^{n}$ é um conjunto fixo de vetores, chamados de centros, cujas localizações podem ou não coincidir com algum dado observado, n é o número de funções de base e N é o número de amostras, com $n < N$. Substituindo a equação 5.5 (expansão dos mínimos quadrados) na equação 5.1 (funcional) e minimizando a equação resultante $H(f^*)$ em função de $\mathbf{c}$ a seguinte equação linear é obtida

$$(G^T G + \lambda \mathbf{g})\mathbf{c} = G^T \mathbf{y} \tag{5.6}$$

em que se define os seguintes vetores e matrizes: $(\mathbf{y})_i = y_i$, $(\mathbf{c})_i = c_i$, $(G)_{ji} = G(||\mathbf{x}_i - \mathbf{t}_j)$ e a matriz $(\mathbf{g})_{jk} = G(||\mathbf{t}_j - \mathbf{t}_k||)$, para se chegar a estas equações foi usado o fato de

$$\psi(f) = \int_{R^d} \frac{|\overline{f}^*(\mathbf{s})|^2}{\overline{G}(\mathbf{s})} d\mathbf{s} = \int_{R^d} \frac{1}{\overline{G}(\mathbf{s})} \sum_{j,k=1}^{n} c_j c_k e^{is(\mathbf{t}_j - \mathbf{t}_k)(\overline{G}(\mathbf{s}))^2} d\mathbf{s}$$

$$= \sum_{j,k=1}^{n} c_j c_k \int_{R^d} e^{is(\mathbf{t}_j - \mathbf{t}_k)\overline{G}(\mathbf{s})} d\mathbf{s} = \sum_{j,k=1}^{n} c_j c_k G(||\mathbf{t}_j - \mathbf{t}_k||).$$

As equações 5.6 e 5.4 são similares; entretanto, a matriz dos mínimos quadrados a ser invertida é nxn em vez de NxN. Esta técnica de regularização com n funções de base reduz a complexidade computacional, mas depende da escolha dos centros das funções de base, que será abordada mais adiante. Para $\lambda = 0$, a solução da equação 5.6 só depende das condições de invertibilidade da matriz $G^T G$. Observando as equações, percebe-se que existem dois parâmetros de suavização: λ e os n números de funções de base. É comum fazer $\lambda = 0$, como fizeram BROOMHEAD e LOWE (1988), que julgavam que os n números de funções de base eram suficientes para se fazer a suavização da aproximação. Entretanto, os efeitos de λ e n na suavização são diferentes e quando se opta por $\lambda = 0$, torna-se a solução da equação subótima.

b) Normas ponderadas: a norma $||\mathbf{x} - \mathbf{t}_i||$ pode passar a ser uma norma ponderada $||\mathbf{x} - \mathbf{t}_i||^2_{\mathbf{Q}} = (\mathbf{x} - \mathbf{t}_i)^T \mathbf{Q}^T \mathbf{Q} (\mathbf{x} - \mathbf{t}_i)$ em que $\mathbf{Q}$ é uma matriz quadrada e o expoente T indica que esta matriz é transposta. No caso simples em que $\mathbf{Q}$ é uma matriz diagonal com elementos q_{ij}, cada elemento atribui um peso respectivamente a cada entrada. Quando $\mathbf{Q}$ é uma matriz identidade, tem-se a norma euclidiana. A matriz $\mathbf{Q}$ é geralmente utilizada quando diferentes tipos de entradas estão presentes, desde que as escalas relativas das componentes sejam arbitrariamente diferentes.

Quando a matriz Q é conhecida *a priori* o trabalho é simplificado. Como o principio da regularização consiste em encontrar uma função f que minimiza o funcional da equação 5.1, logo a solução aproximada do problema pode ser da seguinte forma

$$f^*(\mathbf{x}) = \sum_{i=1}^{n} c_i G(||\mathbf{x} - \mathbf{t}_i||^2_Q). \tag{5.7}$$

Supondo que os valores de Q não são conhecidos, o problema pode ser formulado para encontrar f e Q que minimize o funcional $H_Q(f)$, encontrando o Q ótimo que corresponda à transformação linear ótima. Como exemplo, apresenta-se o caso mais simples, no qual a matriz Q é diagonal e $G(x) = e^{-x^2}$. Logo,

$$G(||\mathbf{x}||^2_Q) = e^{-x_1^2 q_1^2} e^{-x_2^2 q_2^2} ... e^{-x_n^2 q_n^2}$$

em que os elementos da diagonal da matriz Q são proporcionais ao inverso das respectivas variâncias (σ^2) de cada componente da gaussiana multidimensional.

5.3 As redes RBF exatas e as generalizadas

Existem várias técnicas disponíveis para a aproximação de superfícies (regressão linear multivariada, splines cúbicas e outras), mas para algumas aplicações, como a previsão de séries temporais multivariadas, esta tarefa tem características típicas que fazem com que várias destas técnicas não sejam indicadas. Entre as principais características estão:

(1)**Alta dimensão**: os problemas que envolvem visão geralmente tem alta dimensão e em algumas séries temporais este problema também poderá ocorrer já que o número de amostras pode ser relativamente pequeno.

(2) **Número de amostras relativamente pequeno**: para uma dimensão 30, tem-se que ter 2^{30} vértices, ou seja, 10^9 vértices e geralmente tem-se no máximo 10^4 amostras. Isto é o que se chama de "maldição da dimensionalidade"e impõe limites a qualquer técnica de aproximação. No caso de séries temporais com atrator de ordem 10 são necessárias 1.024 amostras.

(3) **Dados ruidosos**: já que todos os dados observados, como as séries temporais, são ruidosos, uma aproximação é preferível a uma interpolação pura.

É natural que surja a pergunta: é possível aproximar uma função com 30 variáveis de entrada a partir de 10^4 amostras ruidosas ou menos? A resposta está relacionada mais com as características da função que será aproximada, principalmente o seu grau de suavização e o nível de precisão requerido.

A rede RBF tem grande capacidade de aproximar funções multivariadas não lineares em mapeamentos contínuos. Ela tem uma arquitetura simples e seu algoritmo de aprendizagem corresponde à solução de um problema de regressão linear, resultando em um processo de treinamento rápido. Ela tem três propriedades importantes para aplicações práticas: lida bem com altas dimensões; pode ser atribuído um variacional à mesma e, portanto, lida melhor com dados ruidosos; processamento com paralelismo.

Entre as funções de base radial se destaca a função gaussiana que normalmente é a mais utilizada em aplicações práticas. Entretanto, as funções tipo spline fina são conhecidas como exemplos de aproximadores polinomiais por partes relativamente suaves, estáveis, fáceis de manipular e calcular em um computador. Em [BORS, 2004] é destacado que a spline fina é mais indicada para a predição e a gaussiana é mais apropriada para a classificação de padrões. A Tabela 5.1 apresenta exemplos de funções de base radial.

Nome	Equação matemática	Condição
Gaussiana (localizada)	$h(r) = e^{(\frac{-r^2}{2\sigma^2})}$	$\sigma > 0$ e $r \in R$
Multiquadrática (não localizada)	$h(r) = (c^2 + r^2)^{\alpha}$	$\alpha > 0$, $c > 0$ e $r \in R$
Multiquadrática inversa (localizada)	$h(r) = \frac{1}{(c^2+r^2)^{\alpha}}$	$\alpha > 0$, $c > 0$ e $r \in R$
Spline fina (localizada)	$h(r) = \frac{r}{c^2}\log(\frac{r}{c})$	$c > 0$ e $r \in R$

Tab. 5.1: Exemplos de funções de base radial

O parâmetro σ controla o raio de influência de cada função. É particularmente evidente, no caso da função gaussiana, que esta função é localizada e monotônica decrescente ($\Theta(\alpha) \rightarrow 0$ à medida que $\alpha \rightarrow \infty$). O parâmetro σ determina o quão rapidamente o valor desta função cai a zero à medida em que se afasta do centro. No caso da função de base radial do tipo gaussiana, o parâmetro σ é o próprio desvio padrão. Assim, σ define a distância euclidiana média (raio médio) que mede o espalhamento dos dados representados pela função de base radial em torno de seu centro.

A rede RBF exata é aquela em que são utilizadas tantas funções de base radial quantos sejam os padrões representativos da função a ser aproximada. Em 1988, BROOMHEAD e LOWE sugeriram que nem todos os vetores de entrada (padrões do conjunto de dados) necessitavam ter uma função de base radial associada e que não havia necessidade de que a escolha dos centros fosse restrita ao conjunto original de vetores. Este tipo de rede foi denominada de generalizada e tida como um estimador de mínimos quadrados. As redes RBF exatas inicialmente não incluíam um *bias*. Entretanto, a partir do modelo sugerido por BROOMHEAD e LOWE (1988), foi acrescido um termo constante de polarização (*bias*).

A consciência de que o comportamento das redes RBF depende fortemente do número de neurônios da camada escondida e da atribuição dos seus centros foi criada. Os métodos tradicionais de se determinar os centros são: escolher randomicamente os vetores de entrada dos dados de treinamento; algoritmos de clusterização não supervisionados para obter os vetores dos centros a partir dos dados de entrada; métodos supervisionados para a escolha dos centros.

Posteriormente, surgiram as redes RBF com regressão generalizada, que são utilizadas para aproximação de funções multivariadas não lineares. Já foi provado que dada uma quantidade apropriada de neurônios na camada escondida este tipo de rede pode aproximar uma função contínua dentro de uma precisão requerida [DEMUTH and BEALE, 1998]. Entretanto, este tipo de rede exige um esforço computacional bem maior que as outras redes RBF.

A razão de freqüentemente se fazer com que a dimensão do espaço oculto de uma rede RBF seja alta (expansão) é que este parâmetro está relacionado à capacidade da rede de aproximar um mapeamento entrada-saída suave e quanto mais alta for a dimensão do espaço oculto, mais precisa será a aproximação [NIYOGI and GIROSI, 1996]. Esta questão tem origem no artigo original de COVER [HAYKIN, 1999], sinalizando que a classificação de padrões dispostos em um espaço de alta dimensionalidade é mais provável de ser linearmente separável que em um espaço de baixa dimensionalidade.

A arquitetura *feedforward* das redes RBF exatas, de acordo com a Figura 5.1, é composta de uma camada de nós fonte (que conectam a rede a seu ambiente externo), à qual é apresentado

o vetor de entrada $\mathbf{u}_i \in R^L$. Uma única camada intermediária com K neurônios não-lineares, cada um deles computando uma função distância entre o vetor de entrada e o respectivo centro da função de base radial, opera a transformação não linear e constitui a chamada camada escondida. Na camada de saída tem-se uma função de ativação linear com o objetivo de não limitar (saturar) os valores de saída e incorporar técnicas estatísticas de regressão linear. Uma função de ativação tipo sigmóide limitaria a saída entre 0 e +1 e uma função tipo tangente hiperbólica entre -1 e +1, podendo degradar a previsão.

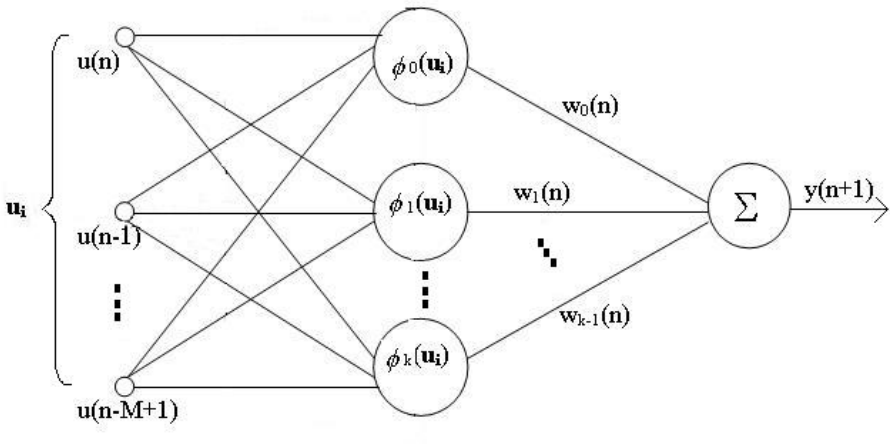

Fig. 5.1: Arquitetura da rede RBF

Pode-se acrescentar um *bias* à figura anterior. O mapeamento não-linear da camada escondida pode ser expresso, por exemplo, via funções de ativação gaussianas, da forma:

$$\varphi_j(\mathbf{u}_i(n)) = \exp(-\frac{1}{2\sigma_j^2(n)}\|\mathbf{u}_i(n) - \mathbf{t}_j(n)\|^2) \tag{5.8}$$

em que $\mathbf{u}_i(n) \in R^L$ e representa o vetor de entrada que pertence ao processo estocástico $\mathbf{U}$, no instante n. Também, $\mathbf{t}_j(n) \in R^L$ e representa o vetor centro da j-ésima função de base radial com $j = 0, 1, ..., K$, em que K é o número de funções de base radial (neurônios) que por sua vez é igual (redes RBF exatas) ou menor (redes RBF generalizadas) ao número de amostras. A variância $\sigma_j^2(n) \in R$ e é associada a cada uma das funções de base radial no instante n.

Como a camada de saída da rede neural RBF é formada por um único neurônio linear, este neurônio que compõe a camada de saída é definido como um combinador linear das funções de base radial. Logo, a saída y da rede RBF é a soma das saídas de cada gaussiana, ponderadas pelos respectivos pesos sinápticos w_k, de tal forma que a combinação linear é expressa por

$$\hat{y}(n+1) = \sum_{k=0}^{k-1} w_k(n) \exp(-\frac{1}{2\sigma_k^2(n)}\|\mathbf{u}_i(n) - \mathbf{t}_k(n)\|^2) \tag{5.9}$$

$$= \mathbf{w}^T(n).\varphi(n)$$

em que $\varphi_k(\mathbf{u}_i(n), \mathbf{t}_k(n), \sigma_k^2(n))$ representa a k-ésima função de base radial. A função gaussiana φ_k computa o quadrado da distância euclidiana (ou outra norma) $d_k^2 = \|\mathbf{u}_i - \mathbf{t}_k\|^2$ entre um

vetor de entrada $\mathbf{u}_i$ e o centro $\mathbf{t}_k$ da k-ésima função de base radial. O sinal de saída produzido pelo k-ésimo neurônio escondido é devido à função exp(.) e ao operador $(.)^2$, uma função não-linear da distância d_k. Os pesos w_k conectam o k-ésimo neurônio escondido ao nó de saída da rede.

A equação a seguir é utilizada na rede RBF como uma solução formal para o mapeamento com bias

$$f(\mathbf{u}_i) = w_0 + \sum_{j=1}^{K} w_j \varphi_j(\|\mathbf{u}_i - \mathbf{t}_j\|^2) \tag{5.10}$$

em que vetor $\mathbf{w}$ representa os pesos e w_0 expressa o bias.

Embora as variâncias das funções de base radial de uma rede RBF possam assumir diferentes valores, pode-se utilizar uma variância comum a todos os neurônios [MOODY e DARKEN, 1989]. É comum a crença de que isto já é suficiente para que a rede aproxime qualquer função contínua, desde que haja número suficiente de funções de base radial. O trabalho de [BISHOP, 1994] sugere que os valores das variâncias das funções de base radial afetam somente as propriedades numéricas dos algoritmos de aprendizado e a capacidade de predição da rede, mas não afetam a capacidade geral de aproximação das redes RBF. Como este trabalho trata da previsão de séries temporais, esta crença não será aceita e propõe-se um método para ajustar a variância a partir da qualidade de previsão.

O ajuste dos centros das funções de base de uma rede RBF pode seguir vários métodos. Nesta tese, por exemplo, na escolha dos centros das funções de base, são utilizados os seguintes métodos: análise das componentes principais (PCA); algoritmo ARIA. Nestes métodos, os procedimentos usados para determinar os centros são independentes dos ajustes das matrizes de ponderação da norma associada à camada oculta e dos ajustes dos pesos da camada de saída.

A variância é ajustada por meio de um parâmetro denominado de fator de variância que consiste em uma constante de proporcionalidade que define o valor de $2\sigma_k^2$, a partir do quadrado da máxima distância euclidiana entre os centros, resultando que $2\sigma_k^2(n) = \xi(n) \max\{\|\mathbf{t}_i(n) - \mathbf{t}_j(n)\|^2\}$. Também é utilizado $\sigma_k = \max\{\|\mathbf{t}_i(n) - \mathbf{t}_j(n)\|\}/\sqrt{2m}$, em que m é o número de centros. A escolha apropriada de σ_k garante que as funções de base individuais não tenham um espalhamento com pico acentuado e nem com pouca declividade (*flat*). É razoável admitir que o valor de ξ influencia os valores dos erros de aproximação. Assim, a equação (5.9) pode ser reescrita como

$$\hat{y}(n+1) = \sum_{j=0}^{k-1} w_j(n) \exp(-\frac{1}{\xi(n) \max\{\|\mathbf{t}_l(n) - \mathbf{t}_j(n)\|^2\}} \|\mathbf{u}_i(n) - \mathbf{t}_j(n)\|^2). \tag{5.11}$$

Observa-se então que nas redes RBF a não-linearidade no argumento é introduzida via expansão do vetor de entrada. Caso o vetor $\mathbf{u}_i$ tenha sofrido um pré-processamento, resultando em um vetor $\mathbf{x}(n) = [x(n) \quad x(n-1) \quad ... \quad x(n-L+1)]^T \in R^L$ e, a partir deste vetor, uma transformação não linear no argumento é gerada na saída das funções de ativação dos neurônios da camada escondida, tem-se

$$\varphi(n) = [\varphi_1(\mathbf{x}(n)) \quad \varphi_2(\mathbf{x}(n)) \quad ... \quad \varphi_k(\mathbf{x}(n))] \in R^K. \tag{5.12}$$

em que $\varphi_j(\mathbf{x}(n))$ é a j-ésima transformação não-linear de $R^L \rightarrow R$ sobre o sinal de entrada. Definindo a matriz de estado como $\Phi = [\varphi(1) \quad \varphi(2) \quad ... \quad \varphi(M)]^T$, onde cada $\varphi_j(\mathbf{x}(n))$ é dado pela equação $\varphi_j(\|\mathbf{x}_n - \mathbf{t}_j\|^2)$, uma função escalar radial simétrica, tendo $\mathbf{t}_j$ como centro.

Em aplicações práticas a probabilidade de se encontrar matrizes mal condicionadas é grande. Logo, geralmente os pesos da rede são ajustados a partir da pseudoinversa, de acordo com a equação abaixo. Este método torna os resultados de previsão menos sensíveis aos parâmetros de regularização.

$$\mathbf{w} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}. \tag{5.13}$$

Os pesos $\mathbf{w}$ armazenam as informações *a priori* sobre o modo como o próximo elemento na série é gerado a partir de seus estados prévios. Logo, os parâmetros da rede RBF que devem ser ajustados são os centros e a variância das funções de base e os pesos da camada de saída. Estes parâmetros podem também ser ajustados pelo algoritmo LMS. O número K de neurônios da camada escondida deve ser grande o suficiente para que a matriz de estados possa armazenar todos os estados significativos e a dimensão L dos vetores de estado do processo deve ser também suficientemente grande para captar as informações necessárias para ajustar os pesos do modelo.

A saída de uma rede RBF generalizada pode ser considerada então como um filtro de predição linear com matriz de interpolação definida pela matriz de estados Φ. Esta abordagem concorreu para a redução do custo computacional e possibilitou a aplicação das redes RBF na predição de séries temporais.

5.4 Redes RBF para a previsão de séries temporais

Na formalização dos modelos neurais de previsão propostos, baseados em redes RBF com funções de ativação tipo gaussiana e spline fina, são assumidas as seguintes condições :

- A série temporal $\{\mathbf{y}\}_{t=1}^N$, para cada valor predito, pode ser associada a um processo estocástico $\mathbf{Y}$ com número de amostras N, com taxa de amostragem ótima T_S, tanto para a metodologia baseada em redes RBF com centros ajustados via PCA como para a que utiliza o algoritmo ARIA.

- O treinamento da rede RBF a torna capaz de antecipar temporalmente o valor do processo estocástico $\mathbf{Y}$ que ocorre em $n+1$, ou seja, o valor um passo a frente representado por $\hat{y}(n+1)$.

- A janela de predição inteligente dinâmica $JPID$ associada às redes RBF com centros ajustados via PCA e as amostras que são utilizadas pelo algoritmo ARIA representam o processo estocástico $\mathbf{Y}$ e têm abrangência suficiente para que os modos de variação básicos de $\mathbf{Y}$ possam ser captados.

Um aspecto prático importante na implementação de um modelo neural é a normalização dos dados de entrada e de saída. Esta normalização é relevante porque se pelo menos uma entrada da rede tiver valores muito distintos das outras, provavelmente gerará valores de erros não proporcionais, comprometendo o ajuste dos parâmetros. Uma maneira eficaz de normalizar os dados de entrada e saída de redes tipo PMC é fazer $A(k) = (A(k) - A_{min})/(A_{max} - A_{min})$; em que A_{max} e A_{min} são, respectivamente, os valores máximos e mínimos encontrados dentro do espaço de amostras utilizado no treinamento da RNA. Já as redes tipo RBF geralmente têm suas entradas e saídas normalizadas de maneira a tornar a média das entradas igual a zero e o desvio padrão unitário.

5.4.1 Identificação dos centros das funções de base via PCA

Os centros obtidos por meio da análise de componentes principais não tem a característica de localidade geométrica. O ajuste dos centros das funções de base por meio do PCA tem como primeiro passo calcular a matriz de covariância γ_y, a partir da janela de predição inteligente dinâmica $JPID$, extraída de $\mathbf{Y}$, captando a dinâmica temporal do processo associado à série temporal analisada.

Para se calcular a matriz de covariância γ_y, primeiramente se calcula o vetor média do conjunto $\mathbf{Y}$, de acordo com

$$\bar{\mathbf{y}}(n) = \frac{1}{M} \sum_{i=0}^{M-1} \mathbf{y}(n-i). \tag{5.14}$$

Posteriormente, forma-se a matriz $\mathbf{X}$, composta por vetores $\mathbf{x}_i$, determinados a partir da diferença entre os vetores que compõem $\mathbf{Y}$ e o vetor média, calculado por meio da equação anterior. Assim, tem-se a seguinte equação:

$$\mathbf{x}_i(n) = \mathbf{y}(n-i) - \bar{\mathbf{y}}(n), \qquad i = 0, 1, ..., M-1. \tag{5.15}$$

Resultando na matriz de covariância γ_y dada por

$$\gamma_y(n) = \frac{1}{M} \sum_{i=0}^{M-1} \mathbf{x}(n-i)\mathbf{x}(n-i)^T. \tag{5.16}$$

Finalmente, resulta a matriz quadrada de ordem M, com $\gamma_y(n)$ dada por

$$\gamma_y(n) = \begin{bmatrix} \gamma_{0,0}(n) & \gamma_{0,1}(n)... & \gamma_{0,(M-1)}(n) \\ \gamma_{1,0}(n) & \gamma_{1,1}(n)... & \gamma_{1,(M-1)}(n) \\ \vdots & \vdots & \vdots \\ \gamma_{(M-1),0}(n) & \gamma_{(M-1),1}(n)... & \gamma_{(M-1),(M-1)}(n). \end{bmatrix} \tag{5.17}$$

Os autovalores e autovetores da matriz de covariância $\gamma(n)$ são obtidos por meio da transformada de Kahunen-Loéve, a partir da solução da seguinte equação

$$\gamma(n).\mathbf{e}_k(n) = \lambda_k(n).\mathbf{e}_k(n) \tag{5.18}$$

em que $k = 1, ..., K$ representa o número de autovalores e autovetores.

A partir da determinação dos autovalores são calculados os autovetores associados. Quando o esforço computacional demandado para calcular os autovalores é grande pode-se utilizar a metodologia de Householder associada à transformação matricial QL.

Finalmente, os centros das funções de base radial da rede RBF são determinados de acordo com a seguinte equação:

$$\begin{aligned}
\Omega_{1a}(n) &= \sqrt{\lambda_1(n)}.\mathbf{e}_1(n),\\
\Omega_{1b}(n) &= -\sqrt{\lambda_1(n)}.\mathbf{e}_1(n),\\
\Omega_{2a}(n) &= \sqrt{\lambda_2(n)}.\mathbf{e}_2(n),\\
\Omega_{2b}(n) &= -\sqrt{\lambda_2(n)}.\mathbf{e}_2(n),\\
&\vdots\\
\Omega_{Ka}(n) &= \sqrt{\lambda_K(n)}.\mathbf{e}_K(n),\\
\Omega_{Kb}(n) &= -\sqrt{\lambda_K(n)}.\mathbf{e}_K(n).
\end{aligned} \tag{5.19}$$

Quando a série temporal está carregada com ruído e este ruído pode ser aproximado por um ruído branco, pode-se filtrá-lo, desprezando os autovalores menos significativos. Neste caso, o número de neurônios pode ser menor que $2K$. Destaque-se que os ruídos com baixa correlação são representados nos subespaços de menores autovalores.

Finalmente, estes centros obtidos por meio da análise de componentes principais não têm a característica de localidade geométrica, mas têm a habilidade de captar a dinâmica temporal do processo que é fundamental para a previsão da série temporal estudada.

5.4.2 Identificação dos centros das funções de base via algoritmo ARIA

A maioria dos métodos de clusterização não avaliam as densidades locais dos dados. Por exemplo, os algoritmos SOM (Self-Organizing Maps), sugeridos em [KOHONEN, 1988], e o aiNet que foi proposto para atribuir os centros de redes RBF em [CASTRO and ZUBEN, 2001] estão nesta categoria. Estes métodos podem apresentar distorções devido, principalmente, aos seguintes fatos: quando os *clusters* são colocados relativamente próximos um do outro; quando a densidade varia de *cluster* para *cluster*; quando suas bordas são nebulosas e superpostas. Isto pode gerar uma representação não realista dos dados.

O algoritmo ARIA (*Adaptive Radius Immune Algorithm*) [BEZERRA et al., 2005], foi criado para preservar a densidade e é baseado nos mecanismos de expansão clonal e supressão, em conjunto com a informação de densidade dos dados. Os tamanhos dos raios dos *clusters* são inversamente proporcionais às respectivas densidades, ou seja, alta densidade tem raios pequenos e baixa densidade raios grandes. Este método gera *clusters* que permitem uma visão geral mais completa e melhor dos dados.

As três fases principais do ARIA são:

a) **Maturidade da afinidade**: os antígenos (padrões) são apresentados aos anticorpos, os quais sofrem uma forte mutação para se ajustar melhor aos antígenos (interações A_g -A_b).

b) **Expansão clonal**: os anticorpos que são mais estimulados são selecionados para serem clonados e a rede imune cresce.

c) **Supressão na rede**: a interação entre os anticorpos é quantificada e se um anticorpo reconhece outro, um dos dois é eliminado do conjunto de células (interações A_b -A_b).

Em seguida apresenta-se o pseudo código deste algoritmo:

1 Iniciar as variáveis.
2 Para a iteração de 1 a N padrões (antígenos), faça.
  2.1 Para cada antígeno A_g, faça.
    2.1.1 Selecionar o anticorpo A_b com melhor *matching*.
    2.1.2 Fazer a mutação de A_b com a taxa mi.
    end
  2.2 Eliminar os A_b que não foram estimulados.
  2.3 Clonar os A_b que reconhecem antígenos a uma distância maior que o raio R.
  2.4 Calcular a densidade local para cada A_b.
  2.5 Estimar o limiar de eliminação (raio) de cada A_b, fazendo $R_{ab} = r*(den_{max}/den)^{1/dim}$.
  2.6 Elimina anticorpos dando prioridade de vida para aqueles com menor R.
  2.7 Faz $E = mean(R)$.
  2.8 Se a generalização atual é maior que $gen/2$.
    2.8.1 Reduz mi, fazendo $mi = mi * decay$.
    end
  end

Os símbolos utilizados no pseudo código acima são: R_{ab} é o raio de cada anticorpo (limiar de eliminação); r^* é o multiplicador do raio, determina o tamanho do menor raio; mi^* é a taxa de mutação; $decay^*$ é uma constante ($0 < decay < 1$) usada para diminuir a taxa de mutação; E é o raio que define o vizinho para estimar a densidade; gen^* é o número de iterações; dim é a dimensão dos dados de entrada. Os parâmetros assinalados com asterisco são dados de entrada fornecidos pelo usuário do algoritmo.

O mecanismo de mutação para um anticorpo A_b, em função de um antígeno A_g, é apresentado na equação abaixo

$$A_b = A_b + mi * rand * (A_g - A_b) \tag{5.20}$$

em que a taxa de mutação mi é inicialmente 1 e $rand$ é um número randômico entre 0 e 1. Quando mi e rand são iguais a 1, então o anticorpo é exatamente igual ao A_g.

O procedimento de clonagem também utiliza a equação anterior. Um clone é uma simples cópia de seu anticorpo parente. Como um anticorpo pode ser estimulado por vários antígenos, utiliza-se o primeiro antígeno que o estimula para se obter o clone. A clonagem deve ocorrer depois do anticorpo sofrer mutação para gerar diversidade, caso contrário, a diversidade pode ser expressivamente afetada.

O raio adaptativo do anticorpo apresentado no ARIA é capaz de capturar a informação da densidade relativa, preservando as distâncias relativas após a compressão dos dados, tolerando ruídos e com performance superior ao *k-means*, EM, aiNet e o SOM. O ARIA tem somente um

parâmetro (r) importante para ser ajustado pelo usuário e a determinação exata do seu valor ainda é uma questão em aberto. Cada problema tem uma resolução diferente e este parâmetro serve para lidar com isto, cabendo ao usuário fazê-lo já que o ajuste automático do seu valor ainda é uma questão em aberto.

5.4.3 Ajuste do espalhamento da função spline fina

Este algoritmo é desenvolvido para a função de base tipo spline fina da camada escondida da rede RBF. Esta função é dada pela seguinte equação

$$\varphi_k(n) = \frac{\beta(n)}{\sigma^2(n)} \log \frac{\beta(n)}{\sigma(n)} \tag{5.21}$$

em que, para cada instante n, $\beta(n)$ é dado por $\| \mathbf{x}(n) - \mathbf{t}_k(n) \|^2$, estimado pela norma euclidiana (ou outra norma) entre o vetor de entrada e o centro do agrupamento, e $\sigma(n)$ controla o espalhamento do agrupamento.

É fato já bem conhecido que um baixo erro de aproximação não significa necessariamente um baixo erro de predição. Já um alto erro de aproximação quase sempre implica em um alto erro de predição. Portanto, é razoável que se busque ir direto ao ponto principal da questão, ou seja, uma relação entre a função que avalia a predição e um fator que possa ajustar a dispersão.

O espalhamento do agrupamento pode ser representado pela equação abaixo

$$\xi(n). \max_{a,b}\{\| \mathbf{t}_a(n) - \mathbf{t}_b(n) \|^2\}$$

em que a dispersão é ajustada de acordo com o maior valor do quadrado da distância euclidiana entre pares de centros dos agrupamentos, multiplicado por um fator de ajuste da dispersão. A partir do fator de variância ξ para uma função gaussiana, a redução do erro de aproximação por meio do ajuste de ξ implica na redução de erro de predição. Logo, como os erros de aproximação e predição se correlacionam via ξ, o uso do $NMSE(n)$, que avalia a qualidade da predição, serve como referência para ajuste experimental de ξ. Assim, atua-se no erro de predição e não na eventual redução do erro de aproximação.

Especificamente, dada a janela de predição $P(n)$ e os respectivos centros $\mathbf{t}_k(n)$, os pesos e o espalhamento são ajustados em função de $\xi(n)$. Seja $\hat{y}(n+1)$ o valor predito, calcula-se o $NMSE(n)$ e processa-se a otimização, ou seja, a minimização do $NMSE(n)$. Os centros permanecem constantes e os pesos e o espalhamento variam com o novo valor de ξ.

O procedimento recursivo para ajustar $\xi(n)$, no instante n, objetivando minimizar o $NMSE(n)$, utilizará a equação abaixo, que expressa o valor do neurônio linear de saída da rede RBF, com função spline fina na camada escondida.

$$\hat{y}(n+1) = \sum_{k=0}^{k-1} w_k(n)\varphi_k(n) = \sum_{k=0}^{k-1} w_k(n) * \frac{\|\mathbf{x}(n) - \mathbf{t}_k(n)\|}{[\xi(n). \max_{a,b}\{\|\mathbf{t}_a - \mathbf{t}_b\|^2\}]^2} \log[\frac{\|\mathbf{x}(n) - \mathbf{t}_k(n)\|}{\xi(n). \max_{a,b}\{\|\mathbf{t}_a - \mathbf{t}_b\|^2\}}]$$

com

$$\sigma(n) = \xi(n). \max_{a,b}\{\|\mathbf{t}_a(n) - \mathbf{t}_b(n)\|^2\}$$

$$C_1 = \|\mathbf{x}(n) - \mathbf{t}(n)\|$$

$$C_2 = \max_{a,b}\{\|\mathbf{t}_a(n) - \mathbf{t}_b(n)\|^2\}.$$

Resultando na seguinte equação

$$\hat{y}(n+1) = \sum_{k=0}^{k-1} w_k(n)\frac{c_1}{\xi(n)^2(c_2)^2}\log[\frac{c_1}{\xi(n).c_2}] = \sum_{k=0}^{k-1} w_k(n)\frac{c_1}{\xi(n)^2(c_2)^2}[\log c_1 - \log c_2 - \log \xi(n)]. \tag{5.22}$$

O ajuste recursivo de $\xi(n)$, será realizada a partir de

$$\xi(s+1) = \xi(s) - \eta\nabla(s) \tag{5.23}$$

em que s é o passo de recursão, η é a taxa de aprendizagem e $\nabla(s) = \frac{\partial\{NMSE(s)\}}{\partial\xi(s)}$. Para a otimização no mesmo passo s de recursão, o $NMSE(s) = \frac{\sum_{i=0}^{n}(y(n+1)-\hat{y}(s))^2}{\sum_{i=0}^{n}(y(n+1)-y(n))^2}$, e $\hat{y}(s)$ é o valor estimado (predito) e $y(n+1)$ é o valor observado da série temporal. Logo,

$$\hat{y}(s) = \sum_{k=0}^{k-1} w_k(s) * \frac{c_1}{\xi^2(s).c_2^2}[\log c_2 - \log c_2 - \log \xi(s)].$$

A função que se deseja minimizar é o $NMSE(s)$, para o passo n. Caso seja aplicado o operador gradiente ao mesmo, resulta:

$$\nabla(s) = \frac{\partial}{\partial\xi(s)}\{\frac{\sum_{i=0}^{n}(y(n+1) - y(s))^2}{\sum_{i=0}^{n}(y(n+1) - y(n))^2}\}.$$

Mas, como $\xi(s)$ é ajustado no instante n, o operador $\partial\{.\}$ anula todos os termos em $n+1$, isto é, o gradiente resulta em

$$\nabla(s) = \frac{\frac{\partial}{\partial\xi(s)}[y(n+1) - \sum_{k=0}^{k-1} w_k(s).\frac{c_1}{\xi^2(s).c_2^2} * [\log c_1 - \log c_2 - \log \xi(s)]]^2}{\sum_{i=0}^{n}(y(i+1) - y(i))^2}. \tag{5.24}$$

Como

$$e(s) = y(n+1) - \hat{y}(s)$$

então

$$\frac{\partial e(s)^2}{\partial\xi(s)} = 2.e(s).\frac{\partial e(s)}{\partial\xi(s)} \tag{5.25}$$

portanto

$$\frac{\partial e(s)}{\partial\xi(s)} = \frac{\partial}{\partial\xi(s)}\{y(n+1) - w_0(s).f_0 - w_1(s).f_1 - w_{k-1}(s).f_{k-1}\}$$

mas

$$\frac{d[f(x).g(x)]}{dx} = f'(x).g(x) + f(x).g'(x) = [\frac{\partial w_0(s)}{\partial \xi(s)}.f_0(s) + w_0(s).\frac{\partial f_0(s)}{\partial \xi(s)}]$$
$$+[\frac{\partial w_1(s)}{\partial \xi(s)}.f_1(s) + w_1(s).\frac{\partial f_1(s)}{\partial \xi(s)}] + \ldots + [\frac{\partial w_k(s)}{\partial \xi(s)}.f_k(s) + w_k(s).\frac{\partial f_k(s)}{\partial \xi(s)}]$$
$$= \sum_{k=0}^{k-1}[\frac{\partial w_k(s)}{\partial \xi(s)}.f_k(s) + w_k(s).\frac{\partial f_k(s)}{\partial \xi(s)}] \tag{5.26}$$

desenvolvendo

$$\frac{\partial f_k(s)}{\partial \xi(s)} = \frac{\partial}{\partial \xi(s)}[\frac{c_1}{\xi^2(s).c_2^2}(\log c_1 - \log c_2 - \log \xi(s))]$$
$$= -\frac{2\xi(s).c_1}{\xi(s)^4.c_2^2}.[\log c_1 - \log c_2 - \log \xi(s)] + \frac{c_1}{c_2^2.\xi^2(s)}.[\frac{-1}{\xi(s)}] =$$
$$= -\frac{2c_1}{c_2^2.\xi(s)^3}.\log[\frac{c_1}{c_2\xi(s)}] - \frac{c_1}{c_2^2\xi(s)^3}.$$

Como a taxa de aprendizagem geralmente é pequena, torna-se razoável a seguinte aproximação

$$\frac{\partial w_k(s)}{\partial \xi(s)} = \frac{w_k(s) - w_k(s-1)}{\xi(s) - \xi(s-1)}. \tag{5.27}$$

Aplicando (5.25), (5.26) e (5.27) em (5.24), resulta:

$$\nabla(s) = -\frac{2e(s)}{\sum_{i=0}^{n}(y(i+1) - y(i))^2}$$
$$\sum_{k=0}^{k-1}\{w_k(s)[\frac{2c_1}{c_2^2.\xi(s)^3}\log(\frac{c_1}{\xi(s).c_2}) + \frac{c_1}{c_2^2.\xi^3(s)}] - \frac{w_k(s) - w_k(s-1)}{\xi(s) - \xi(s-1)}\frac{c_1}{\xi^2(s).c_2}\log(\frac{c_1}{c_2\xi(s)})\}.$$

Com a fórmula de ajuste da dispersão sendo dada por:

$$\xi(s+1) = \xi(s) - \eta\nabla(s)$$

acrescentado um *momentum* (α) para tentar fugir dos mínimos locais, resulta:

$$\xi(s+1) = \xi(s) - \eta\nabla(s) + \alpha\triangle\xi(s)$$

onde

$$\triangle\xi(s) = \xi(s) - \xi(s-1).$$

Finalmente, tem-se a equação de ajuste do fator de dispersão:

$$\xi(s+1) = \xi(s) + \alpha\triangle\xi(s) + \frac{2\eta e(s)}{\sum_{i=0}^{n}(y(i+1) - y(i))^2}$$

$$* \sum_{k=0}^{k-1} [w_k(s)[\frac{2c_1}{c_2^2 \xi^3(s)}.\log(\frac{c_1}{c_2.\xi(s)}) + \frac{c_1}{c_2^2 \xi^3(s)}] - \frac{w_k(s) - w_k(s-1)}{\xi(s) - \xi(s-1)}.\frac{c_1}{c_2.\xi^2(s)}.\log(\frac{c_1}{c_2}.\xi(s))]. \quad (5.28)$$

O mesmo desenvolvimento matemático pode ser feito para a função de base radial gaussiana e se chega à equação (5.29) - vide referência [DECASTRO, 2000]. Esta equação também é utilizada nas simulações que serão apresentadas posteriormente.

$$\xi(s+1) = \xi(s) + \alpha \triangle \xi(s) + \frac{2\eta e(s)}{\sum_{i=0}^{n}(y_i(i+1) - y_i(i))^2} * \\ [\frac{\sum_{j=0}^{k-1} \varphi_j(s) w_j(s) \|y_i(s) - t_k(n)\|^2}{\xi^2(s) \max_{a,b}[\|t_b(n) - t_a(n)\|^2]}] + \sum_{j=0}^{k-1} \varphi_j(s) \frac{w_j(s) - w_j(s-1)}{\xi(s) - \xi(s-1)}. \quad (5.29)$$

em que $e(s)$, α e η são respectivamente o erro calculado pela diferença entre o valor estimado no passo s e o valor observado para a próxima amostra $y(n+1)$, a quantidade de *momentum* aplicada à trajetória utilizada para evitar os mínimos locais e a taxa de aprendizagem. Note-se que o algoritmo não garante um erro mínimo global e que as trajetórias em direção ao erro mínimo local são estocásticas.

5.4.4 A matriz de transição e a determinação dos pesos

A partir do ajuste dos centros e do fator adaptativo de variância pode-se calcular os valores das saídas das funções de ativação por meio da equação (5.30). O valor da saída da rede pode ser dado pelo produto interno entre as saídas dos neurônios da camada escondida e os parâmetros de ajuste (pesos), resultando em:

$$\hat{y} = \varphi^T w. \quad (5.30)$$

Entretanto, quando se faz a predição de um passo à frente, primeiramente se ajustam os pesos e, para fazê-lo, é necessário antes determinar a matriz de estados que pode ser representada por:

$$\Phi(n) = \begin{bmatrix} \varphi(n-M+1)^T \\ \vdots \\ \varphi(n-1)^T \\ \varphi(n)^T. \end{bmatrix} \quad (5.31)$$

Na predição de séries temporais, a matriz de estado Φ armazena as informações sobre os estados básicos do processo a ser predito. A cada estado armazenado em Φ é associada uma saída desejada, de tal forma a se definir um vetor de saídas desejadas $\mathbf{d}$. Cada elemento de $\mathbf{d}$ é definido pelo elemento que está uma posição à frente na série temporal com respeito ao vetor de entrada que gerou o correspondente vetor de estados em Φ.

Os pesos do modelo proposto, utilizando o conceito de regularização, são ajustados de acordo com a equação abaixo

$$\mathbf{w} = (\Phi^T \Phi + \lambda g)^{-1} \Phi^T \mathbf{d}. \quad (5.32)$$

em que a matriz $(\mathbf{g})_{jk} = G(||\mathbf{t}_j - \mathbf{t}_k||)$ e λ é parâmetro de regularização calculado pelo método da validação cruzada generalizada.

Os pesos $\mathbf{w}$ armazenam as informações *a priori* sobre o modo como o próximo elemento na série é gerado a partir de seus estados prévios. Deslizando a matriz Φ uma posição à frente na janela de predição e usando a informação de transição contida em $\mathbf{w}$, pode-se estimar o próximo elemento na série.

No caso das redes RBF com centros determinados via PCA, o aprendizado para a determinação dos centros das funções de base radial é não supervisionado (PCA) e a dispersão em relação aos centros é ajustada de maneira supervisionada por meio do fator de dispersão adaptativo, apresentado anteriormente.

Estas redes neurais têm a capacidade para armazenar conhecimento experimental e torná-lo disponível para o uso. Assemelham-se ao cérebro em dois aspectos: (1) o conhecimento é adquirido pela rede por meio de um processo de aprendizado, a partir das informações disponíveis; (2) forças de conexões entre neurônios, conhecidos como pesos sinápticos, são utilizados para armazenar o conhecimento adquirido. Logo, a essência de uma rede neural está na sua capacidade de aprender e no armazenamento deste aprendizado.

5.5 Testes para o ajuste e a avaliação dos modelos de previsão

No Capítulo 4 já foram abordados os testes de não linearidades e de dependência temporal linear e não linear. Nesta seção serão abordados os testes de detecção de não linearidades localizadas e de não linearidades negligenciadas durante o ajuste do modelo neural. Também serão apresentados os testes de avaliação da capacidade preditiva dos modelos. A avaliação preditiva dos modelos propostos será em relação a um modelo Martingale: $y_t = \mu + \varepsilon_t$, exceto para o NMSE. Nesta comparação é importante verificar a dependência entre os modelos. WHITE (2000) propôs um teste para evitar este tipo de problema (*data snooping via reality check*). Este teste será abordado mais adiante neste capítulo.

5.5.1 Testes para a detecção localizada de não-linearidades

O procedimento para testes de detecção localizada de não-linearidades geralmente é implementado a partir de dados discretos amostrados. A dificuldade na análise da não-linearidade aumenta quando os dados estão contaminados com ruídos. Neste caso, a detecção de não-linearidades deve ser realizada por meio de técnicas estatísticas. Caso seja detectada alguma característica de sistema não-linear determinista, mesmo que sejam sinais de não-linearidades fracas, pode-se utilizar alguns dos conceitos associados à teoria dos sistemas não-lineares. Entretanto, em [SMALL and TSE, 2003] foi sugerido que os melhores resultados obtidos por estes métodos são nas análises típicas de processamento de sinais e não na predição de séries no tempo. Em [SCHREIBER, 1998] foi observado que estes resultados, mesmo que sejam positivos, devem ser analisados com cuidado.

Nesta tese são apresentados e utilizados dois testes de detecção e localização de não-linearidades para identificar interações não-lineares nos dados. O primeiro método apresen-

tado é o proposto em [BILLINGS and VOON, 1983] e [BILLINGS and VOON, 1986], em que a equação

$$E\{(y(t) - E\{y(t)\})(y^2(t-\tau) - E\{y^2(t)\})\} = 0, \forall\tau \tag{5.33}$$

só é válida se e somente se o sistema original for linear. Assim, pode-se identificar não linearidades e, mais ainda, estabelecer os limites de um intervalo de confiança de 0,05: $\pm 1,96 * N$, em que N é o comprimento do registro de dados disponíveis. A função de correlação dada pela equação anterior pode ser estimada utilizando os dados disponíveis. Logo, se os valores da função saírem fora dos limites estabelecidos, o sistema que gerou estes dados é não-linear, pelo menos no intervalo analisado.

Um critério importante para a escolha deste tipo de teste é sua capacidade de discriminação. A capacidade de discriminação (potência) do teste é definida como a probabilidade de se rejeitar a hipótese nula quando realmente a mesma é falsa, dependendo de quão intensamente os dados atuais se desviam da hipótese nula. Um indicador cujo poder de discriminação é particularmente eficiente para este fim é o de detecção de não-linearidades via assimetrias sob reversão no tempo. Seja a equação

$$\phi^{rev} = \frac{\sum_{n=\tau+1}^{N}(y_n - y_{n-\tau})^3}{[\sum_{n=\tau+1}^{N}(y_n - y_{n-\tau})^2]^{\frac{3}{2}}}. \tag{5.34}$$

Esta estatística permite detectar possíveis assimetrias sob reversão no tempo. As estatísticas de processos estocásticos lineares são simétricas sob reversão no tempo. Em [SCHREIBER, 1998] foram comparados quantitativamente os testes mais populares de não linearidades e conclui que este método apresenta melhores resultados empíricos. O resultado positivo destes testes significa somente uma indicação de não-linearidade, mas não que o sistema é determinista. Dessa forma, este teste pode ser aplicado a processos deterministas e estocásticos. Estes dois testes podem ajudar na escolha do modelo (linear ou não linear) para um determinado contexto em que a variável analisada esteja inserida, ou seja, se linear ou não linear.

5.5.2 Detecção de não linearidades negligenciadas

Os modelos não lineares ficaram populares nos últimos anos, seja porque os dados exibem não linearidades inequívocas, seja pela disponibilidade de modelos não lineares que podem ser bem especificados. Entretanto, a hipótese de linearidade (logarítmica) ainda é muitas vezes mantida principalmente por que este tipo de modelo é mais fácil de estimar e de interpretar. Logo, erros significativos de especificação deste tipo de modelo podem ocorrer pelo fato de se ignorar as não linearidades.

No ajuste de modelos não lineares é prudente testar se alguma não linearidade gerada pelo processo foi negligenciada pelo modelo, assegurar que as relações funcionais foram propriamente ajustadas e que nenhuma informação com capacidade explanatória foi negligenciada. Neste contexto, existem duas categorias de testes: aqueles projetados para testar formas específicas de não linearidades, tendo uma hipótese alternativa paramétrica; os que não tem uma alternativa paramétrica, mas objetivam detectar não linearidades nos dados. Este último tipo de teste

pode identificar modelos mal especificados que não captaram as não linearidades na média condicional.

BLAKE e KAPETANIOS (2003) propuseram um teste estatístico baseado em redes RBF para detectar não linearidades negligenciadas pelo modelo, ou seja, um teste que verifica se o modelo capta as não linearidades ou se alguma delas foi negligenciada. Em [LEE et al., 1993] foi proposto um teste com este mesmo objetivo e é bastante popular. Foram apresentados estudos analíticos em [TERÄSVIRTA and GRANGER, 1993] sobre este teste e sugeriram que este teste tem as melhores propriedades de potência. Entretanto, os estudos empíricos de BLAKE e KAPETANIOS (2003) sugerem que o teste proposto por eles tem maior potência que o teste de LEE et al. (1993).

BLAKE e KAPETANIOS (2003) legaram duas contribuições nesta área: utilizaram uma rede RBF que, uma vez que os parâmetros da camada escondida são determinados, reduz a estimativa dos pesos a um problema de mínimos quadrados lineares; escolhe a ordem do modelo via critérios de informação, ou seja, não necessita que a arquitetura da rede RBF seja precisa. O ajuste de redes RBF via critério de informação é uma prática já bastante conhecida na literatura. Entretanto, a principal contribuição de BLAKE e KAPETANIOS (2003) reside na sua aplicação em testar se alguma não linearidade foi negligenciada. A utilização do *bootstrap* pode corrigir alguma distorção no teste, mas pode acarretar que este teste perca potência.

O teste para não linearidades na média de uma série temporal $\{y_t\}_1^N$ condicionada às entradas $\mathbf{x}_t$, assumindo que a média condicional pode ser expressa por uma função linear de $\mathbf{x}_t$, tem a seguinte hipótese nula

$$P[E(y_t|\mathbf{x}_t) = \theta^{'}\mathbf{x}_t] = 1 \tag{5.35}$$

em que θ é um vetor de constantes. A alternativa é

$$P[E(y_t|\mathbf{x}_t) = \theta^{'}\mathbf{x}_t] < 1 \tag{5.36}$$

para qualquer θ. A forma genérica de aproximação de uma RNA aplicada neste contexto pode ser dada por

$$E(y_t|\mathbf{x}_t) = \theta^{'}\mathbf{x}_t + \sum_{j=1}^{K} w_j\varphi_j(\|\mathbf{x}_i - \mathbf{t}_j\|^2). \tag{5.37}$$

O teste consiste em verificar se $w_1, w_2, \ldots, w_q = 0$ a partir da equação abaixo

$$y_t = \theta^{'}\mathbf{x}_t + \sum_{j=1}^{K} w_j\varphi_j(\|\mathbf{x}_i - \mathbf{t}_j\|^2) + \varepsilon_t \tag{5.38}$$

em que ε_t é um ruído branco. Esta equação possibilita um teste para linearidades negligenciadas. Assume-se também, sob a hipótese nula, que

$$y_t = \theta^{'}\mathbf{x}_t + \varepsilon_t, \quad t = 1, 2, \ldots, N. \tag{5.39}$$

Neste teste, os parâmetros da camada escondida da rede RBF não são utilizados. Entretanto, existe o problema de identificação relacionado com a determinação dos parâmetros da camada escondida. O teste de LEE et al. (1993) também tem problema de identificação, que

pode ser devido a uma possível multicolinearidade gerada pela função logística. Para contornar este problema em redes RBF, BLAKE e KAPETANIOS (2003) sugerem que a ordem do modelo (número de neurônios), os centros e as variâncias das funções de base podem ser ajustados por um critério de informação [AKAIKE, 1974].

Uma estatística WALD padrão é utilizada para testar a hipótese nula de que $w_1, w_2, \ldots, w_q = 0$. Este teste tem a seguinte forma

$$\frac{1}{\sigma^2}\mathbf{w}'[\mathbf{R}'(\Phi'\Phi)^{-1}\mathbf{R}]^{-1}\mathbf{w} \tag{5.40}$$

em que Φ é a matriz dos regressores da equação (5.37), R é a matriz de restrições para os coeficientes das funções escondidas ($\mathbf{w} = [w_1, w_2, \ldots, w_q]'$) e σ^2 é a variância dos resíduos. Este teste tem distribuição assintótica semelhante à distribuição chi-quadrada (χ^2_q). Finalmente, este tipo de teste objetiva evitar erros potenciais de aproximação, principalmente aquele devido ao desconhecimento do mapeamento exato da relação, que geralmente é inevitável.

5.5.3 Testes estatísticos de habilidade preditiva

O principal critério de avaliação adotado em [WEIGEND and GERSHENFELD, 1994] e considerado uma referência para a comunidade da área de previsão de séries temporais via redes neurais artificiais é o critério baseado no erro médio quadrático normalizado ($NMSE$ - *normalized mean square error*). Consiste na razão entre os erros médios quadráticos de dois modelos que estão sendo comparados. Freqüentemente, o modelo do passeio aleatório é utilizado como referência padrão para avaliar novos modelos de previsão de séries temporais. A equação apresentada em seguida representa o $NMSE$, tendo como referência o passeio aleatório.

$$NMSE = \frac{\sum_{i=1}^{n}(\hat{y}(i) - y(i))^2}{\sum_{i=1}^{n}(y(i) - y(i-1))^2} \tag{5.41}$$

em que o valor obtido para o numerador é resultante da soma dos quadrados das n diferenças entre os valores efetivamente observados ($y(i)$) e os respectivos valores obtidos pelo preditor ($\hat{y}(i)$). O denominador expressa a soma dos quadrados das diferenças entre os valores atuais ($y(i)$) e imediatamente anteriores da amostra ($y(i-1)$). Uma razão inferior a 1 corresponde a uma predição melhor do que aquela obtida pela simples repetição do valor efetivamente observado para a mostra anterior àquela a ser predita - limiar que qualifica um preditor que pretenda ser útil. Tal critério para o MSE é tido como normalizador e sinaliza se o previsor gera melhores previsões do que o passeio aleatório (*random walk*).

Outros dois testes estatísticos utilizados serão o MSFE (*mean square forecast error*) e MAFE (*mean absolute forecast error*) dados pelas equações

$$MSFE = (1/n)\sum_{i=1}^{n}(\hat{y}(i) - y(i))^2. \tag{5.42}$$

$$MAFE = (1/n)\sum_{i=1}^{n}(|\hat{y}(i) - y(i)|). \tag{5.43}$$

Entretanto, quando se utiliza estes testes, nem sempre é possível saber se a superioridade de um dos modelos comparados deve-se efetivamente à superioridade em termos de performance

ou se esta superioridade se deve somente a alguma variabilidade ligada à amostra coletada. Para evitar este tipo de problema utiliza-se o teste de [WHITE, 2000].

O teste de DIEBOLD e MARIANO (1995), muito utilizado na comunidade econômica, incorporando os conceitos propostos em NEWEY e WEST (1987), resulta na equação que incorpora a função autocovariância dada por

$$DM = \frac{\bar{d}}{\sigma_d} \sim N(0,1) \tag{5.44}$$

em que $\bar{d} = \frac{1}{m}\sum_{j=1}^{m} d_j$ e m é o número de amostras, ou seja, é a média da função perda dada pela diferença entre os erros quadráticos dos modelos A e B avaliados, expressa por $d_j = e_A^2 - e_B^2$. O valor do desvio padrão σ_d dos $d_j s$ é dado por

$$\sigma_d = \sqrt{DP + 2\sum_{j=1}^{q} w_j(q)\hat{\gamma}(j)} \tag{5.45}$$

em que DP é o desvio padrão comum dos $d_j s$, $\hat{\gamma}(j)$ é a função de autocovariância de ordem j dos $d_j s$ e $w_j(q) = 1 - \frac{j}{q+1}$, com $q < m$, m é o número de amostras e q é o número de autocorrelações impostas no modelo. Assim, $w_j(q)$ é a função núcleo proposta por NEWEY e WEST (1987) para assegurar, por meio da atribuição de pesos às autocorrelações dos $d_j s$, que a matriz de autocovariância será positiva definida.

A grande vantagem deste teste reside no fato de o mesmo apresentar distribuição assintótica normal com média zero e variância igual a um. Caso, a título de exemplo, o valor dessa estatística seja maior que 1.65, rejeita-se a hipótese nula de que a diferença entre esses dois modelos A e B se deve à aleatoriedade, com um grau de confiança de 95%. Agrega-se ao critério o nível de confiança os efeitos das possíveis autocorrelações entre os valores residuais dos modelos A e B. Entretanto, este teste não é aplicável na comparação entre modelos aninhados.

Em determinados contextos, é mais importante avaliar se os modelos são bons para prever se haverá uma ascensão ou queda no valor da variável predita, ainda que não se saiba o valor exato desta queda ou subida. Na economia e no mercado financeiro essa informação pode modificar as eventuais decisões de investimento já que os investidores desejam mais maximizar lucros do que minimizar erros de predição. O teste MFTR (*mean forecast trading returns*), que será apresentado em seguida, é mais importante que os testes MSFE e MAFE para a área da economia já que fornece uma porcentagem média dos lucros. Já o teste MCFD (*mean correct forecast direction*) é uma medida econômica relacionada com movimentos (*timing*) do mercado, fornecendo a porcentagem média de acertos da direção correta, possibilitando aos administradores de ativos prever a direção das mudanças, gerando mais lucros que a média do mercado. Estes testes têm as seguintes equações

$$MFTR = (1/n)\sum_{i=1}^{n} sign(\hat{y}(i))y(i)). \tag{5.46}$$

$$MAFE = (1/n)\sum_{i=1}^{n} \mathbf{I}(sign(\hat{y}(i))sign(y(i))) > 0) \tag{5.47}$$

em que I é a função impulso unitário.

A lógica desenvolvida por PESARAN e TIMMERMANN (1992) é bastante intuitiva: se o produto entre o valor observado ($y(i)$) e o valor predito ($y(\hat{in})$) for positivo, então o modelo acertou o sentido das previsões. A idéia básica é construir uma função indicadora para a taxa de sucesso (SR = *success ratio*), uma indicação de quantas vezes foi acertada a escolha, na média, a direção para a qual o mercado estava caminhando. O SR é dado pela equação

$$SR = \frac{1}{m}\sum_{i=1}^{m} I[y(i) * \hat{y}(i) > 0]. \tag{5.48}$$

Esta taxa é utilizada na construção da estatística teste de PESARAN e TIMMERMANN (1992), denominada de DA (*direction of accuracy*), juntamente com as funções p e $\hat{p}$, dadas pelas equações

$$p = \frac{1}{m}\sum_{j=1}^{m} I[y(i) > 0]. \tag{5.49}$$

$$\hat{p} = \frac{1}{m}\sum_{j=1}^{m} I[\hat{y}(i) > 0]. \tag{5.50}$$

A função p é o percentual de vezes em que o valor observado da variável em estudo é maior que zero. Chama-se de $\hat{p}$ o percentual de vezes em que os valores das previsões são positivos. Sabe-se que a probabilidade de sucesso no caso em que os eventos forem independentes é dada por

$$SRI = p\hat{p} + (1-p)(1-\hat{p}). \tag{5.51}$$

A taxa de sucesso estimada é estatisticamente significativa em relação à taxa de sucesso para eventos independentes? PESARAN e TIMMERMANN (1992) construíram uma estatística para lidar com esta questão utilizando as seguintes equações

$$Var(SRI) = \frac{1}{m}[(2\hat{p}-1)^2 p(1-p) + (2p+1)^2\hat{p}(1-\hat{p}) + \frac{4}{m} + p\hat{p}(1-p)(1-\hat{p})] \tag{5.52}$$

e

$$Var(SR) = \frac{1}{m}SRI(1-SRI). \tag{5.53}$$

A estatística DA (*direction of accuracy*) é dada por:

$$DA = \frac{SR - SRI}{\sqrt{Var(SR) - Var(SRI)}} \sim N(0,1). \tag{5.54}$$

Ressalta-se que, tal como nas estatísticas desenvolvidas por DIEBOLD e MARIANO (1995), a estatística deste teste é unicaudal. Assim, uma vez que essa estatística tem distribuição assintótica, com $N(0,1)$, caso seu valor seja maior do que 1.65, pode-se rejeitar com um nível de confiança de 95% a hipótese nula de que os acertos na direção obtidos pelo modelo avaliado se devem à aleatoriedade.

5.5.4 Teste de White via *boostrap*

Quando um conjunto de dados é utilizado mais de uma vez na comparação de modelos a um modelo *benchmark* por meio de uma estatística pode ocorrer que os resultados satisfatórios são devidos somente à chance de ocorrer e não ao mérito do modelo. Este problema é difícil de evitar na análise de séries temporais já que muitas vezes tem-se uma única trajetória para ser analisada. WHITE (2000) criou um teste para a comparação de múltiplos modelos, a partir de uma mesma realização, em previsões fora da amostra denominado por White de *reality check* para evitar bias na mineração de dados. Este método testa a hipótese nula que o melhor modelo encontrado não tem superioridade preditiva sobre um modelo *benchmark*. Isto permite que se tenha um grau de confiança no resultado obtido, evitando resultados equivocados gerados pelas chances em vez do mérito genuíno do modelo. Este teste será utilizado em conjunto com os testes MSFE, MAFE, MFTR e MCFD, em que as funções de avaliações destes métodos são as funções integradas ao teste de WHITE (2000) que por sua vez fornece o nível de significância da melhor estatística.

Supondo que n é o número de previsões realizadas de $t = R, \ldots, T$ e que l modelos serão avaliados, o próximo passo é determinar o número re-amostragens N e o parâmetro de suavização q associados à implementação do *bootstrap*. O valor de N influencia na precisão do valor estimado de p-valor e geralmente é um número entre 500 a 1000 devido ao custo computacional. A dependência temporal da série $\{y_t\}_{t=1}^{T}$ é administrada por meio da variável q e quanto maior a dependência menor é q. Por exemplo, uma diferença Martingale terá $q = 1$.

Em seguida, aplica-se o *bootstrap* para gerar ramdomicamente os N conjuntos de amostras de comprimento n, $\{\theta_i(t) = R, \ldots, T\}, i = 1, \ldots, N$. Estes índices são gerados um de cada vez e a amostra correspondente a este índice é escolhida entre aqueles conjuntos de amostras ainda não selecionados. Assim, os dados que são necessários são R, T, q e N, e a armazenagem e manipulação de dados é proporcional a l, número de modelos avaliados, e não a l^2, que é requerido pelo método de Monte Carlo.

O teste é realizado de forma recursiva e começa com a estimativa da performance do modelo *benchmark*. Por exemplo, utilizando o negativo dos quadrados dos erros como a função de avaliação:

$$\hat{h}_{0,t+1} = -(y_{t+1} - \hat{y}_{0,t+1})^2, t = R, \ldots, T.$$

Em seguida, obtém-se a performance do primeiro modelo a ser comparado

$$\hat{h}_{1,t+1} = -(y_{t+1} - \hat{y}_{1,t+1})^2, t = R, \ldots, T.$$

A partir dos valores anteriores calcula-se

$$\hat{f}_{1,t+1} = \hat{h}_{1,t+1} - \hat{h}_{0,t+1}$$

e

$$\bar{f}_1 = n^{-1} \sum_{t=R}^{T} \hat{f}_{1,t+1}.$$

Utilizando os valores gerados pelo *bootstrap* calcula-se

$$\bar{f}^*_{1,i} = n^{-1} \sum_{t=R}^{T} \hat{f}_{1,\theta_i(t)+1}, i = 1, \ldots, N.$$

Fazendo

$$\bar{V}_1 = n^{1/2} \bar{f}_1$$

e

$$\bar{V}^*_{1,i} = n^{1/2}(\bar{f}^*_{1,i} - \bar{f}_1), i = 1, \ldots, N.$$

A inferência para avaliar o primeiro modelo em relação ao modelo *benchmark* é comparando os valores de $\bar{V}_1$ e $\bar{V}^*_{1,i}$. A avaliação da performance do segundo modelo a ser comparado com o *benchmark* é dada por

$$\hat{h}_{2,t+1} = -(y_{t+1} - \hat{y}_{2,t+1})^2, t = R, \ldots, T$$

faz-se

$$\hat{f}_{2,t+1} = \hat{h}_{2,t+1} - \hat{h}_{0,t+1}$$

e

$$\bar{f}_2 = n^{-1} \sum_{t=R}^{T} \hat{f}_{2,t+1}.$$

Utilizando os valores gerados pelo *bootstrap* (POLITIS e ROMAN0, 1994) tem-se

$$\bar{f}^*_{2,i} = n^{-1} \sum_{t=R}^{T} \hat{f}_{2,\theta_i(t)+1}, i = 1, \ldots, N$$

fazendo

$$\bar{V}_2 = max(n^{1/2} \bar{f}_2, \bar{V}_1)$$

e

$$\bar{V}^*_{2,i} = max(n^{1/2}(\bar{f}^*_{2,i} - \bar{f}_2, \bar{V}^*_{1,i}), i = 1, \ldots, N.$$

Para testar se o melhor dos dois modelos é superior ao modelo *benchmark*, compara-se $\bar{V}_2$ e $\bar{V}^*_{2,i}$. Procedendo recursivamente desta maneira de $k = 3, \ldots, l$, testando se o melhor dos k modelos analisados é superior ao modelo *benchmark*. Isto pode ser expresso por

$$\bar{V}_k = max(n^{1/2} \bar{f}_k, \bar{V}_{k-1})$$

e

$$\bar{V}^*_{k,i} = max(n^{1/2}(\bar{f}^*_{k,i} - \bar{f}_k, \bar{V}^*_{k-1,i}), i = 1, \ldots, N.$$

Os valores estatísticos de $\bar{V}^*_{l,i}$ são ordenados como $\bar{V}^*_{l,1}, \bar{V}^*_{l,2}, \ldots, \bar{V}^*_{l,N}$. Para encontrar o p-valor, primeiramente encontra-se o valor de M para $\bar{V}^*_{l,M} < \bar{V}_l < \bar{V}^*_{l,M+1}$. Finalmente, o p-valor do teste será

$$P_{RC2} = 1 - M/N.$$

Este teste para amostras finitas serve para comparar l modelos ajustados a um modelo *benchmark*. Cada número dos l valores de P_{RC2} fornece o p-valor para a hipótese nula que o melhor modelo dos l primeiros modelos não tem superioridade preditiva sobre o modelo benchmark. O último p-valor de P_{RC2} indica se o melhor de todos os modelos ajustados tem superioridade preditiva sobre o modelo *benchmark* para um determinado nível de significância adotado, ou seja, se o valor de P_{RC2} é menor que o nível de significância adotado (5 por cento, 10 por cento e outros). O valor do teste diminui a medida que o número de amostras (T) cresce. A potência do teste melhora com o número de previsões (n) da série temporal e o número de re-amostragens geradas pelo *bootstrap* (N) que tem como contrapartida o esforço computacional.

Considerando que a estatística cresce com a qualidade das previsões, o valor do teste decresce. Já se a estatística decresce com a qualidade das previsões, o valor do teste cresce, ou seja, o problema é simétrico.

Na avaliação entre os modelos ajustados e o *benchmark*, que no caso desta tese é um modelo Martingale, é importante analisar este p-valor em conjunto com o do *bootstrap* básico (*naive*), dado por

$$P_{RC1} = (\sum_{n=1}^{N} I(S_n^* => S)/N.$$

A diferença entre cada P_{RC1} e o último valor de P_{RC2} fornece uma indicação do efeito *data mining bias*, possibilitando quantificar as conseqüências de uma especificação sem informações *a priori*, evitando também confundir o espúrio com o relevante. Entretanto, este teste é sensível à inclusão de modelos com qualidade de previsão pobre, produzindo valores inconsistentes de p-valores. Hansen (2005) abordou esta deficiência e criou um teste alternativo para superar este problema.

5.5.5 Testes para modelos aninhados

No problema geral de seleção de modelos, a escolha de qual teste estatístico deve-se utilizar depende também da existência de modelos competidores aninhados. Pode-se dizer que o modelo A está aninhado dentro do modelo B se o modelo A for um caso especial de B. Quando um modelo tem um modelo aninhado, a diferença entre sua distribuição de teste (chi-quadrada) é assintoticamente independente da estatística do modelo aninhado.

Observa-se que para modelos aninhados, caso as estatísticas originais de teste sigam distribuições chi-quadrada, a diferença também é uma distribuição chi-quadrada. Se as estatísticas originais de teste seguirem distribuições chi-quadrada não centralizada, então a diferença é também uma distribuição chi-quadrada não centralizada. Os graus de liberdade para a diferença são iguais aos graus de liberdade para as duas estatísticas originais de teste. Os parâmetros da distribuição chi-quadrada não centralizada da diferença é igual à diferença dos parâmetros das distribuições chi-quadrada não centralizadas das duas estatísticas originais de teste.

Isto sugere um método de comparação que geralmente é o mais utilizado para comparar o ajuste de dois modelos aninhados. Testa-se a hipótese nula de que não existe nenhuma diferença significativa no ajuste avaliando se a diferença da chi-quadrada é significativa, para os graus de liberdade dados e para um determinado nível de significância. Caso a diferença

seja significativa, a hipótese nula será rejeitada. Este tipo de aproximação está limitada às comparações de modelos aninhados, e a sua interpretação torna-se difícil no caso de uma distribuição chi-quadrada não centralizada.

Observa-se que o teste de causalidade apresentado em [GRANGER, 1969] foi criado para avaliações dentro da amostra e que o teste de DIEBOLD e MARIANO (1995) não tem a capacidade de avaliar a habilidade preditiva de modelos aninhados embora tenha habilidade de avaliar modelos não aninhados de previsão fora da amostra. O teste apresentado em [CHAO et al., 2000] foi criado para suprir esta deficiência e avalia modelos lineares e não lineares aninhados em previsões fora da amostra, inclusive modelos baseados em redes neurais. A referência anterior ilustra empiricamente como a causalidade de GRANGER pode ser afetada em testes fora da amostra, ou seja, este teste pode apresentar resultados diferentes para testes dentro e fora da amostra.

CHAO et al. (2000) criaram uma versão mais geral do teste de causalidade de GRANGER, construindo um teste baseado em $\frac{1}{\sqrt{P}} \sum_{t=R}^{T} \hat{e}_{t+1} h(\nu, \mathbf{x}_t)$, em que ν são parâmetros não identificados sob a hipótese nula, como, por exemplo, os parâmetros das funções de base da camada escondida das redes RBF. Este teste utiliza a mesma equação apresentada em LEE et al. (1993) e BLAKE e KAPETANIOS (2003) para testar não linearidades negligenciadas dentro da amostra, representada por

$$h(\gamma, \mathbf{x}_t) = \nu_1 \mathbf{x}_t + G(\nu_2, \mathbf{x}_t) \tag{5.55}$$

em que ν engloba os parâmetros da parcela linear (ν_1) e os parâmetros da função de base G (ν_2). Neste contexto, CHAO, CORRADI e SWANSON (2000) sugeriram a seguinte estatística para previsões um passo adiante

$$\frac{1}{\sqrt{N}} \sum_{t=R}^{T} \hat{e}_{t+1} h_j(\nu, \mathbf{x}_t)$$

em que $j = 1, 2, \ldots, K$ e, no caso de uma rede RBF, K é o número de funções de base da camada escondida. As hipóteses nula e alternativa são

$$H_0 : E(\hat{e}_{t+1} h(\nu, \mathbf{x}_t)) = 0$$

e

$$H_a : E(\hat{e}_{t+1} h(\nu, \mathbf{x}_t)) \neq 0.$$

Esta estatística tem distribuição normal padrão e também pode ser utilizada para modelos lineares, bastando para isso substituir a matriz das funções de base ($\mathbf{H}(\nu, \mathbf{x}_t)$)) pela matriz de entradas ($\mathbf{X}_t$) do modelo linear, já com os respectivos *lags*. Finalmente, este teste possibilita a avaliação da previsibilidade de modelos lineares e não lineares aninhados para previsões um passo adiante, ou seja, fora da amostra, sem restrições à razão entre o número de previsões e o número de amostras disponíveis, que pode ser representada por $n/N > 0$ ou $n/N = 0$, em que n representa o número de previsões e N o número de amostras disponíveis.

Capítulo 6

Resultados

6.1 Introdução

No contexto de séries no tempo é relevante distinguir um processo diferença Martingale (DM) de um passeio aleatório (PA). O primeiro implica no segundo, mas a recíproca não é verdadeira. Uma série temporal pode ser serialmente não correlacionada, mas ter uma média condicional diferente de zero na sua história passada. Quando uma série temporal é uma diferença Martingale isto implica na sua não previsibilidade.

Vários estudos têm incorporado a hipótese da diferença Martingale na modelagem da taxa de câmbio. Entretanto, freqüentemente, autocorrelação, taxa de variância e potência de espectro são utilizados para testar se uma série temporal da taxa de câmbio segue um processo diferença Martingale. Mas estes testes checam mais a existência da não correlação serial (processo PA) do que uma diferença Martingale e conclusões equivocadas podem ter ocorrido. Isto talvez possa explicar parte dos resultados conflitantes da literatura sobre o assunto.

Em [HSIEH, 1993] foi destacada a detecção de não linearidades nos retornos taxa de câmbio diária devido às variações da volatilidade no tempo, sem efeito ARCH-M, não implicando em capacidade de previsão na média. Já em [HONG and LEE, 2002] foram analisadas estas variações via momentos espectrais generalizados, capazes de distinguir um processo DM de um PA, em cinco das mais importantes taxas de câmbio, e concluíram que freqüentemente estas taxas são não correlacionadas serialmente, mas apresentam não linearidades fortes na média condicional, sinalizando que estas taxas não são seqüências DM e, portanto, são previsíveis.

Neste trabalho, investiga-se inicialmente as características das séries temporais dos retornos diários da taxa de câmbio brasileira e da umidade horária na região de Londrina-PR. Posteriormente, os modelos de previsão na média condicional ARMA-GARCH e os baseados em redes neurais e suas combinações são comparados com um modelo diferença Martingale ($y_t = \mu_t + \varepsilon_t$, em que μ_t é a média no tempo e ε_t é um ruído branco) para previsões fora da amostra. A comparação destes modelos deverá fornecer informações se os modelos ARMA-GARCH, neural e combinado podem fornecer ou não melhores previsões que o modelo diferença Martingale, possibilitando investigar empiricamente se a série é ou não uma diferença Martingale. É importante saber também qual destes modelos fornece a melhor qualidade de previsão.

Na prática, é muito difícil encontrar um modelo que tenha alta performance em todos os períodos. Para incrementar a qualidade das previsões individuais foram criados os modelos

combinados. Neste trabalho é utilizado o método de combinação utilizado em HONG e LEE (2002). Esta metodologia de combinação é representada pelas equações 6.1 e 6.2.

$$\hat{y}(n+1) = \sum_k p^k(n+1)\hat{y}_k(n+1). \tag{6.1}$$

em que $\hat{y}_k$ é o valor estimado de y pelo modelo k e o peso $p^k(n+1)$ é calculado por meio de

$$p^k(n+1) = \frac{exp[-\lambda(n) * \sum_{j=1}^{n}(y(j) - \hat{y}^k(j))^2]}{\sum_{j=1}^{n} exp[-\lambda(n) * \sum_{j=1}^{n}(y(j) - \hat{y}^k(j))^2]} \tag{6.2}$$

em que $\lambda(n) = 1/(2S^2)$ e S^2 é a variância amostral. Intuitivamente, numa janela n, o modelo que teve performance melhor terá pesos maiores, já o modelo que teve performance fraca terá pesos menores.

Estas classes de modelos são ajustadas para implementar a previsão na média condicional das séries temporais dos retornos da taxa de câmbio e da umidade para previsões fora da amostra, um passo adiante. Especificamente são implementados os seguintes modelos: diferença Martingale na média (*benchmark*); ARMA-GARCH; rede neural tipo RBF com função de base gaussiana e centros ajustados via PCA (RBF PCA GAUSS); rede neural tipo RBF com função de base spline e centros ajustados via PCA (RBF PCA SPLINE); rede neural tipo RBF com função de base gaussiana e centros ajustados via algoritmo ARIA (RBF ARIA); e a combinação dos melhores modelos (**combinado**). O modelo RBF ARIA é multivariado com variáveis selecionadas por meio do método de seleção de variáveis proposto no Capítulo 3.

Durante as avaliações das previsões, podem ser obtidos resultados devido somente às chances numéricas (sorte) e não ao mérito do modelo. Este problema é chamado de *bias* na mineração de dados e merece atenção. Para evitar este tipo de problema, o teste de WHITE (2000) é utilizado em conjunto com os métodos de avaliação das previsões para a comparação de múltiplos modelos em previsões fora da amostra, incorporando a dependência entre os modelos comparados. O teste de WHITE (2000) via bootstrap [POLITIS and ROMANO, 1994], gerando o bootstrap p-valor, é implementado neste trabalho. Como algumas não linearidades na média geralmente são negligenciadas durante o ajuste do modelo neural, para evitar este tipo de problema, o teste de BLAKE e KAPETANIOS (2003) será utilizado.

Os critérios utilizados na avaliação das previsões são: NMSE (*normalized mean square error*), MSFE (*mean square forecast error*), MAFE (*mean absolute forecast error*), MFTR (*mean forecast trading returns*) and MCFD (*mean correct forecast direction*). O NMSE compara a habilidade preditiva dos modelos em relação ao passeio aleatório. O MSFE e MAFE são critérios estatísticos de precisão das previsões. O MSFE e o MAFE são potencialmente geradores de enganos em avaliações de modelos porque não são invariantes às transformações dos dados. Os testes MFTR e MCFD são importantes para os investidores que almejam maximizar lucros.

O capítulo foi organizado como segue: na Seção 6.2 são ilustradas as séries temporais, as fontes dos dados, e os seus respectivos gráficos e kernels; na Seção 6.3 são fornecidos os resultados dos testes de estacionariedade; na Seção 6.4 são apresentados os resultados dos testes de não linearidades; na Seção 6.5 são fornecidos os resultados dos testes de dependência temporal; na Seção 6.6 faz-se a seleção das variáveis de entradas dos modelos neurais multivariados tipo RBF ARIA GAUSS; na Seção 6.7 estima-se as dimensões das correlações, expoentes de

LYAPUNOV, *lags* e as dimensões de imersão (*embbeding*); na Seção 6.8 são apresentados os modelos ajustados; na Seção 6.9 chega-se aos resultados das previsões e dos testes, e faz-se as análises.

6.2 Séries temporais utilizadas e as respectivas fontes de dados

A taxa de câmbio nominal resulta do jogo entre oferta e procura por dólares. Se houver mecanismos capazes de conduzir oferta e procura a um equilíbrio, a taxa de câmbio se estabiliza temporariamente. Parte da oferta e procura por dólares resulta do comércio internacional, mas o dólar também é um ativo, logo a oferta e procura por ele estão sujeitas às instabilidades dos mercados de ativos financeiros (bolsas, dívidas públicas e privadas etc). Como este mercado depende de um grupo de variáveis externas e internas, que variam de acordo com o contexto macroeconômico, das microestruturas de mercados e do apetite por risco dos mercados de capitais de países emergentes, a previsão da taxa de câmbio tem sido um grande desafio para a comunidade de séries temporais.

A quantidade de variáveis explicativas relacionadas com as variações cambiais encontradas na literatura chega em torno de uma centena e meia [KAMINSKY and REINHART, 1996, KAMINSKY et al., 1998]. Entretanto, um número reduzido de variáveis, consideradas as mais expressivas no contexto atual, foi selecionado por sua relevância e disponibilidade. As variações diárias do real em relação ao dólar americano, a partir de Janeiro de 2000 até Fevereiro de 2005, foram retiradas do banco de dados do Banco Central do Brasil (BCB). As séries das variáveis de entrada candidatas, a partir de Janeiro de 2000 até Fevereiro de 2005, foram retiradas dos bancos de dados da Economática e da Bloomberg.

Foram escolhidos indicadores de microestruturas de mercados, macroeconômicos, e de mercados de capitais emergentes para testar a metodologia de seleção de variáveis proposta. A Tabela 6.1 apresenta as séries temporais candidatas a variáveis de entrada.

Séries temporais candidatas a variáveis de entrada do modelo para a previsão dos retornos da taxa de câmbio brasileira
Microestruturas de mercados
High e *low* do primeiro vencimento do futuro (forward)
Diferença entre o *high* e o *low* do primeiro vencimento do futuro (forward)
Macroeconômicas
Taxas de câmbio nominais do real/dólar USA, yene/dólar USA, libra/dólar USA, franco suíço/dólar USA e euro/dólar USA
Índice ibovespa
Juros nominais (selic e os dozes vértices da estrutura a termo da taxa de juros)
Diferença entre as taxas de juros internas (cupom cambial) e externas (*fed funds*)
Cotações do aço
Cotações da soja
Mercados de capitais de países emergentes (WEM - world emergents markets)
EMBIPLUS - índice dos mercados emergentes que indica o apetite para ativos nestes mercados
EMBIBR - índice do mercado brasileiro (risco Brasil) que fornece o apetite para ativos neste mercado

Tab. 6.1: Séries temporais candidatas a variáveis de entrada do modelo para a previsão dos retornos da taxa de câmbio brasileira

A variação horária da umidade na região de Londrina-PR, durante o período de Janeiro de 1999 a Dezembro de 2000, foi fornecida pela EMpresa BRAsileira de Pesquisa Agropecuária (EMBRAPA). As séries temporais candidatas a variáveis de entrada do modelo para a previsão da umidade obtidas, durante o período de Janeiro de 1999 a Dezembro de 2000, foram em pequeno número, mas optou-se por utilizá-las. As séries temporais horárias disponíveis na EMBRAPA Soja de Londrina-PR, sobre o microclima desta região, são apresentadas na Tabela 6.2. Esta região tem as melhores terras do país para o cultivo de soja e os produtores utilizam tecnologia avançada nesta cultura.

Séries temporais candidatas a variáveis de entrada do modelo para a previsão da umidade no microclima de Londrina-PR
Precipitações pluviométricas (chuvas)
Temperatura mínima
Temperatura média
Temperatura máxima

Séries temporais candidatas a variáveis de entrada do modelo para a previsão da umidade no microclima de Londrina-PR

Tab. 6.2:

Os gráficos dos dados brutos, logaritmo, primeira diferença e retornos diários da taxa cambial brasileira são ilustrados na Figura 6.1. Observando a figura dos retornos diários, nota-se a presença de conglomerados de valores extremos comuns em séries financeiras e que esta série está aparentemente estabilizada na média. A linha vermelha transversal apresentada nesta figura e na Figura 6.2 separa os dados de treinamento dos dados de testes dos modelos de previsão.

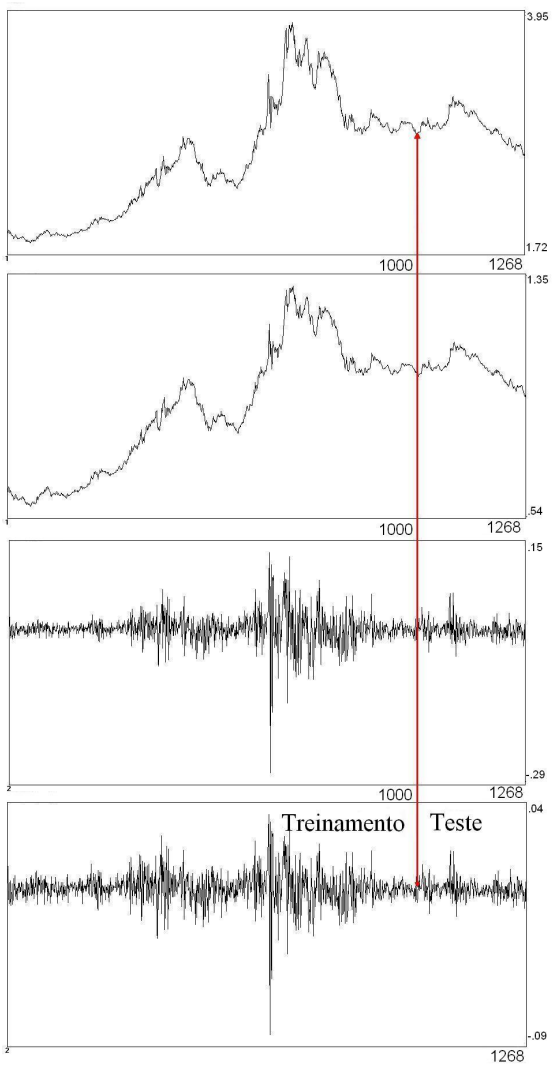

Fig. 6.1: Dados brutos, logaritmo, primeira diferença e retornos diários da taxa cambial brasileira

Os dados brutos, logaritmo, primeira diferença e retornos horários da umidade no microclima de Londrina-PR são apresentados na Figura 6.2 e aparentemente a série dos dados brutos está estabilizada na média.

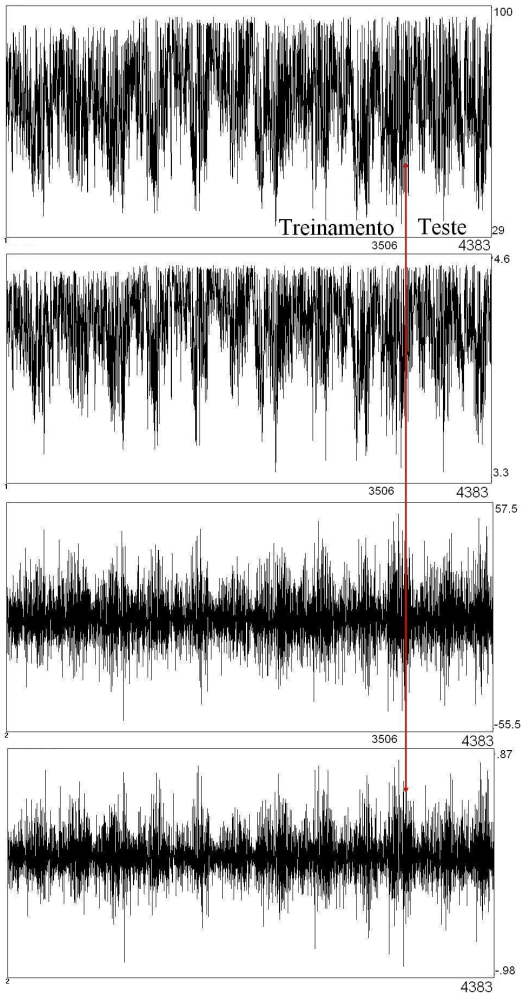

Fig. 6.2: Dados brutos, logaritmo, primeira diferença e retornos horários da umidade no microclima de Londrina-PR

A função de densidade de probabilidade (fdp) é um conceito importante na análise de dados, como, por exemplo, em séries temporais univariadas. O seu papel é encapsular as variações randômicas dos dados em um padrão e isto não é explicado por outras técnicas estruturais de modelagem. O mais antigo e utilizado estimador não paramétrico de densidade é o histograma, mas ele tem desvantagens como: estima todas as densidades via função impulso dentro de um intervalo; cria a necessidade de escolher o número de intervalos (bins); e a perda de informação devido ao fato de ter de colocar cada amostra de x_t no ponto central do intervalo (bin) ao qual ela está associada.

Nesta tese utiliza-se o kernel da normal. A Figura 6.3 apresenta as fdps resultantes do *kernel* da normal padrão (linha tracejada) e dos dados brutos, logaritmo, primeira diferença e retornos diários da taxa cambial brasileira. Existem assimetrias nas séries dos dados brutos e do logaritmo. As caudas pesadas estão na primeira diferença e nos retornos.

Esta série financeira apresentou peculiaridades como conglomerados de valores extremos, assimetrias e excesso de curtose (caldas pesadas).

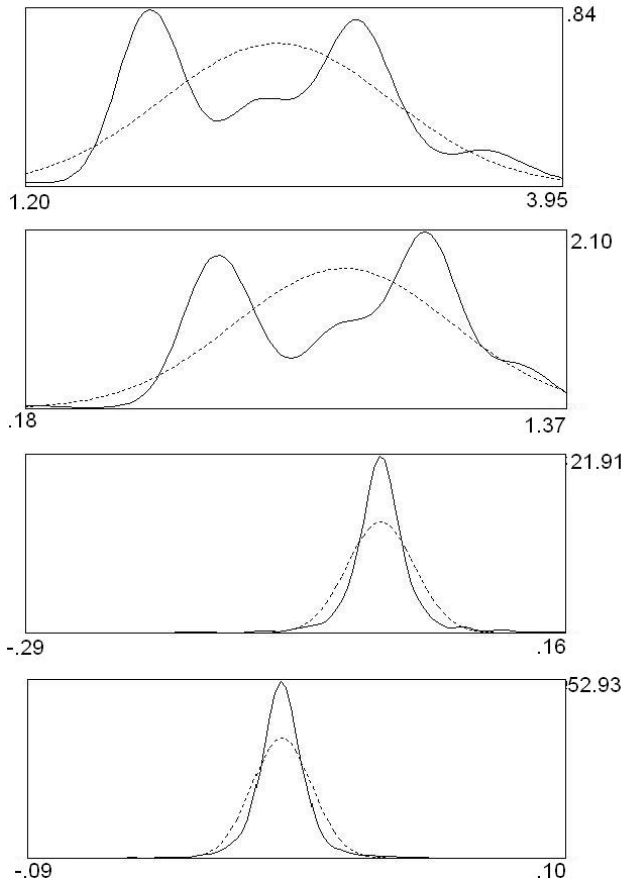

Fig. 6.3: Funções de densidade de probabilidade (fdp - kernel da normal - linha tracejada) dos dados brutos, logaritmo, primeira diferença e retornos diários da taxa cambial brasileira

A Figura 6.4 apresenta as fdps resultantes do *kernel* da normal padrão (linha tracejada) e dos dados brutos, logaritmo, primeira diferença e retornos diários da umidade na região de Londrina-PR. Os gráficos dos dados brutos e da primeira diferença da série apresentam semelhanças com uma distribuição normal, mas não necessariamente se trata de um processo gaussiano.

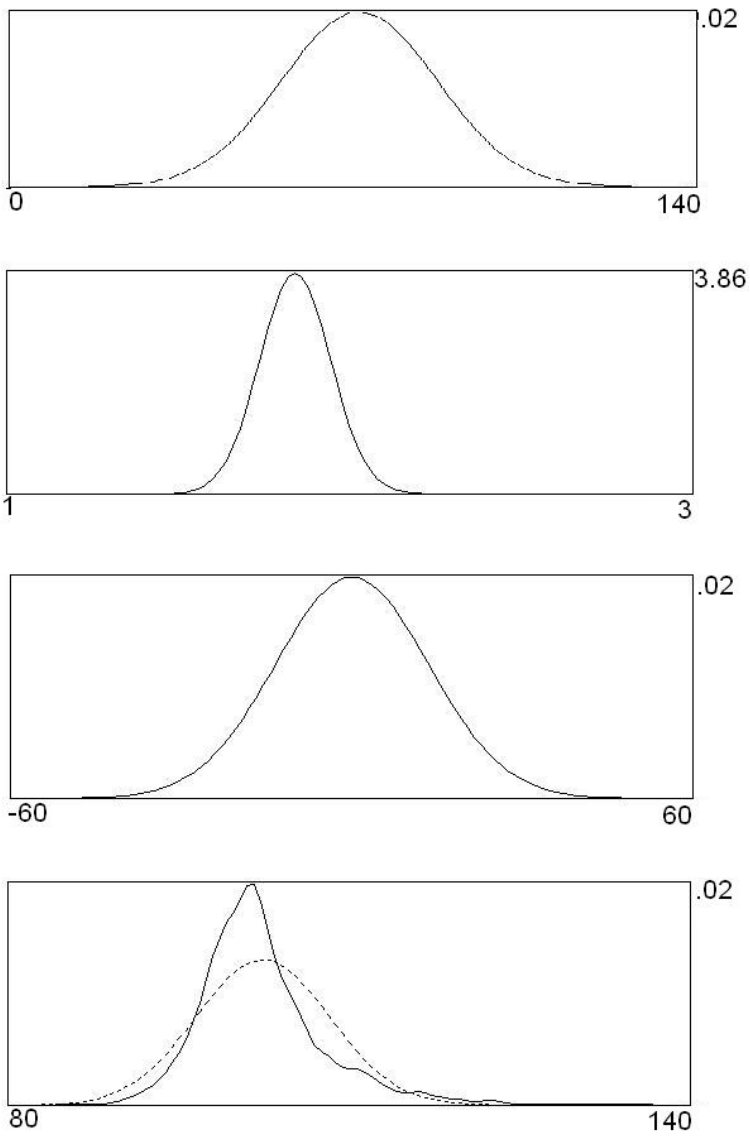

Fig. 6.4: Funções de densidade de probabilidade (fdp - kernel da normal - linha tracejada) dos dados brutos, logaritmo, primeira diferença e retornos diários da umidade na região de Londrina-PR

6.3 Detecção de não estacionariedades

Os processos não estacionários são ilustrados por meio das equações 6.3 e 6.4. A equação 6.3 expressa uma tendência determinista no tempo, é simplesmente uma função linear no tempo somada ao último distúrbio aleatório e representa um processo determinista não estacionário

na média e na variância. Este tipo de série muitas vezes pode ser ajustada por um modelo de regressão padrão. A equação 6.4 representa uma tendência estocástica, é uma função que acumula os últimos choques, ou seja, os últimos distúrbios aleatórios ($\sum_n \varepsilon_t$). Caso um processo com tendência estocástica, ao ter sua série diferenciada, torne-se um processo estocástico estacionário, diz-se que este é não estacionário de origem e homogêneo via diferenciação. Esta homogeneidade não depende do nível original da série.

$$y_t = \mu + \delta t + \varepsilon_t. \tag{6.3}$$

$$y_t = \delta + y_{t-1} + \varepsilon_t. \tag{6.4}$$

Matematicamente, a razão principal para se transformar os dados originais de uma série temporal geralmente é a necessidade de se tornar a série estacionária. A condição de estacionariedade é importante porque é necessária para a especificação de modelos como o ARMA e VAR.

O teste de [DICKEY and PANTULA, 1987] foi aplicado nas duas séries analisadas e os resultados apontaram que a realização da taxa de câmbio brasileira tem uma raiz unitária e que a umidade não tem raízes unitárias. A partir desta informação, aplicou-se os testes de [DICKEY and FULLER, 1979] e [PHILLIPS and PERRON, 1987, PHILLIPS and PERRON, 1988] nas duas séries, cuja hipótese nula é que a série não é estacionária contra a hipótese alternativa que a série é estacionária. Utilizou-se também o teste de KPSS [KWIATKOWSKI et al., 1992], cuja hipótese nula é que a série é estacionária contra a hipótese alternativa que a série é não estacionária. O nível de significância de todos os testes de estacionariedade é 0,05.

Estes testes de estacionariedade para a taxa de câmbio brasileira, logaritmo, 1ª diferença e taxa de retornos são apresentados na Tabela 6.3.

Testes	Brutos	P_{valor}	Log	P_{valor}	1ª diferença	P_{valor}	Retornos	P_{valor}
$ADF-1^{(1)}$	$\tilde{\text{N}}$ rejeita	0,78	$\tilde{\text{N}}$ rejeita	0,84	Rejeita	0,00	Rejeita	0,00
$ADF-2^{(2)}$	$\tilde{\text{N}}$ rejeita	0,53	$\tilde{\text{N}}$ rejeita	0,52	Rejeita	0,00	Rejeita	0,00
$ADF-3^{(3)}$	$\tilde{\text{N}}$ rejeita	0,71	$\tilde{\text{N}}$ rejeita	0,98	Rejeita	0,00	Rejeita	0,00
$PP-1^{(1)}$	$\tilde{\text{N}}$ rejeita	0,74	$\tilde{\text{N}}$ rejeita	0,74	Rejeita	0,00	Rejeita	0,00
$PP-2^{(2)}$	$\tilde{\text{N}}$ rejeita	0,59	$\tilde{\text{N}}$ rejeita	0,68	Rejeita	0,00	Rejeita	0,00
$PP-3^{(3)}$	$\tilde{\text{N}}$ rejeita	0,92	$\tilde{\text{N}}$ rejeita	0,98	Rejeita	0,00	Rejeita	0,00
$KPSS^{(4)}$	Rejeita	0,04	Rejeita	0,03	$\tilde{\text{N}}$ rejeita	0,94	$\tilde{\text{N}}$ rejeita	0,98

Tab. 6.3: Testes de estacionariedade para a taxa de câmbio brasileira (brutos), logaritmo (log), 1ª diferença e taxa de retornos

(1) Este teste é raramente aplicável porque a hipótese nula é que se trata de um processo com raiz unitária e a alternativa é que se trata de um processo estacionário com média zero, podendo resultar em um teste de baixa potência.

(2) A hipótese nula é que o processo tem raiz unitária e a hipótese alternativa é que o processo é estacionário. Este teste pode apresentar baixa potência.

(3) A hipótese nula é que se trata de um processo com raiz unitária e *drift* e a alternativa é que se trata de um processo estacionário, podendo também resultar em um teste de baixa potência.

(4) A hipótese nula é que se trata de um processo estacionário e com média zero e a alternativa é que se trata de um processo raiz unitária e *drift*, podendo também resultar em um teste de baixa potência.

Analisando os testes de estacionariedade da taxa cambial brasileira, observa-se que os testes ao nível de significância de 0,05 são unânimes em apontar a existência de não estacionariedade nos dados brutos e na transformação logarítmica e estacionariedade para a primeira diferença e para a taxa de retorno. Logo, esta série é não estacionária de origem e homogênea via diferenciação de primeira ordem e transformação para a taxa de retorno. A série dos retornos ilustrada na Figura 6.1 sugere heteroscedasticidade na taxa de retornos e que um modelo GARCH poderá eventualmente ser útil.

Os testes de estacionariedade para a umidade na região de Londrina-PR, logaritmo, 1ª diferença e taxa de retornos são apresentados na Tabela 6.4 e as mesmas observações feitas para os testes aplicados nos retornos da taxa de câmbio também são válidas para a umidade.

Testes	Brutos	P_{valor}	Log	P_{valor}	1ª diferença	P_{valor}	Retornos	P_{valor}
$ADF - 1^{(2)}$	Rejeita	0,00	Rejeita	0,00	Rejeita	0,00	Rejeita	0,00
$ADF - 1^{(3)}$	Rejeita	0,00	Rejeita	0,00	Rejeita	0,00	Rejeita	0,00
$PP - 2^{(2)}$	Rejeita	0,00	Rejeita	0,00	Rejeita	0,00	Rejeita	0,00
$PP - 3^{(3)}$	Rejeita	0,00	Rejeita	0,00	Rejeita	0,00	Rejeita	0,00

Tab. 6.4: Testes de estacionariedade para a umidade na região de Londrina-PR(brutos), logaritmo (log), 1ª diferença e taxa de retornos

Os testes não indicam a existência de não estacionariedade nas séries da umidade ao nível de significância de 0,05. A série é estacionária, mas isto não é suficiente para garantir uma boa previsão pelos métodos ARMA já que existem as questões da dependência temporal e das não linearidades.

6.4 Detecção de não linearidades

O teste para detecção de não linearidades utilizado foi o teste de HSIEH (1989) pelo interesse de saber se a não linearidade é na média ou na variância. Neste teste, para o nível de significância de 0,05, caso a estatística do teste esteja entre -1,96 e 1,96, a não linearidade é na variância, caso contrário, é na média. Ilustra-se também nesta seção a utilização de dois tipos de testes de não linearidades localizadas para o nível de significância de 0,05, caso a estatística do teste esteja fora dos limites -1,96 e 1,96. Estes testes podem também ajudar no ajuste de modelos neurais.

Observando a Tabela 6.5, nota-se que a série no tempo da taxa cambial brasileira apresenta não linearidades na variância para as séries dos dados brutos e logaritmos. Existem não linearidades na média para a primeira diferença e retornos. É razoável achar que a ocorrência destas não linearidades sinalizam que não se trata de um processo diferença Martingale e que talvez estas não linearidades possam comprometer os resultados de métodos lineares como o ARMA e o VAR.

Coeficientes	Dados brutos	Logaritmo	1ª diferença	Taxa de retornos
1 1	1,2605	-0,7893	-0,2239	0,7398
1 2	1,2980	-0,7623	3,0458	4,5972
1 3	1,3383	-0,7324	1,9232	2,6163
1 4	1,3828	-0,6967	-0,3379	0,2258
1 5	1,4301	-0,6543	2,1370	3,6713
1 6	1,4789	-0,6054	0,4510	4,7715
2 2	1,2789	-0,7813	1,0176	1,5007
2 3	1,3190	-0,7521	-1,5364	-1,2720
2 4	1,3615	-0,7191	-0,5586	0,2105
2 5	1,4071	-0,6807	0,6436	1,3986
2 6	1,4553	-0,6350	-0,1048	1,5731
3 3	1,3017	-0,7716	-0,5242	0,8407
3 4	1,3442	-0,7393	-1,0446	0,8398
3 5	1,3877	-0,7039	0,1333	0,8848
3 6	1,4343	-0,6625	0,6317	2,2179
4 4	1,3280	-0,7591	-0,7089	0,2004
4 5	1,3713	-0,7245	-2,0268	-0,6924
4 6	1,4159	-0,6863	1,1886	2,1962
5 5	1,3543	-0,7461	1,0895	1,2866
5 6	1,3988	-0,7089	2,2990	2,2301
6 6	1,3814	-0,7323	0,6434	1,2256

Tab. 6.5: Teste de HSIEH para a taxa cambial brasileira, logaritmo, 1ª diferença e taxa de retornos

Existem não linearidades, ao nível de significância de 0,05, na média condicional dos retornos da taxa de câmbio brasileira.

A partir da Tabela 6.6, verifica-se que a série no tempo da umidade na região de Londrina-PR apresenta não linearidades na média condicional, ao nível de significância de 0,05, para as séries dos dados brutos, logaritmos e primeira diferença. Há não linearidades na variância para a taxa dos retornos. A série dos dados brutos, que será utilizada nas previsões, apresenta não linearidades na média.

Coeficientes	Dados brutos	Log dos dados brutos	Primeira diferença	Retornos
1 1	-2,0866	-2,9515	-2,2785	-1,0691
1 2	-2,4651	-3,1214	0,6832	0,2094
1 3	-2,0290	-2,7778	0,6111	0,4596
1 4	-1,5454	-2,3132	0,2953	0,6075
1 5	-1,9416	-2,4942	0,9173	0,4533
1 6	-1,9994	-2,4797	-0,5364	-0,2121
2 2	-1,8191	-2,6821	-1,2485	-0,3913
2 3	-1,6662	-2,5574	1,1851	1,5566
2 4	-0,6036	-1,8709	-0,1576	-0,3666
2 5	-0,7400	-1,7021	-0,0195	0,0891
2 6	-0,2550	-1,4819	-0,2253	0,2842
3 3	-2,1777	-2,6620	-0,5449	-0,3116
3 4	-1,4563	-2,2949	0,4753	0,8673
3 5	-1,1882	-2,0621	-0,1565	0,6038
3 6	-0,3957	-1,4484	0,0159	-0,9393
4 4	-2,0433	-2,5199	0,3909	0,4355
4 5	-1,9039	-2,5485	-0,1004	-0,3441
4 6	-0,8707	-1,9081	-0,0486	-0,0606
5 5	-2,2877	-2,5355	-0,1543	0,0834
5 6	-1,5783	-2,2301	0,3877	0,3089
6 6	-1,8046	-2,1849	-0,2091	-0,1108

Tab. 6.6: Teste de HSIEH para a umidade na região de Londrina-PR, logaritmo, 1ª diferença e taxa de retornos

Como uma ilustração, os resultados dos testes de não linearidades localizadas para os retornos da taxa de câmbio brasileira são apresentados na Figura 6.5. Observa-se a existência de fortes não linearidades localizadas antes, durante e logo após a eleição de 2002, na qual Lula foi eleito Presidente pela primeira vez. Existem também períodos em que o comportamento da série temporal é linear, logo estes testes podem nos guiar para a escolha de um método de previsão linear ou não linear dependendo do contexto. No caso dos modelos baseados em redes neurais artificiais, estes testes podem também ajudar a escolher o tipo de função de ativação mais indicada para capturar as não linearidades detectadas pelos testes.

As análises até agora estão sendo feitas no domínio do tempo, mas poderiam ser realizadas também no domínio da freqüência. A razão da escolha de uma ou outra abordagem está na facilidade de apresentação e interpretação dos resultados e também de acordo com o objetivo da análise. Nesta tese, trabalha-se somente no domínio do tempo.

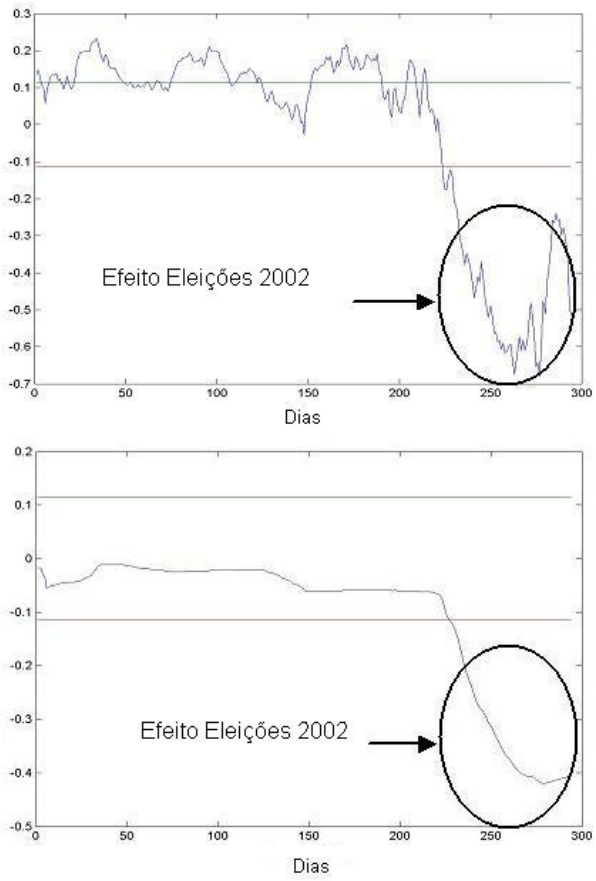

Fig. 6.5: Testes de não linearidades localizadas para os retornos da taxa de câmbio brasileira

6.5 Análise da dependência temporal linear e não linear

Nesta seção faz-se a análise da dependência temporal linear e não linear com o objetivo de descobrir mais informações para sinalizar se o processo é um passeio aleatório, uma diferença Martingale ou nenhum destes processos. Isto é feito com o intuito de verificar a previsibilidade do processo e facilitar a especificação do modelo.

6.5.1 Análise da dependência temporal via autocorrelação

A função de autocorrelação (FAC) e a função de autocorrelação parcial (FACP) dos retornos da taxa cambial brasileira, Figura 6.6, apresentam pequeno nível de correlação, decrescem rapidamente e graficamente sugerem que é possível ajustar um modelo ARMA. Entretanto, ainda falta investigar as informações contidas nas funções de autocorrelação e de autocorrelação parcial dos quadrados dos retornos.

A Figura 6.7 apresenta o gráfico da FAC e da FACP dos dados brutos da umidade na região de Londrina-PR.

A partir da função de autocorrelação e a função de autocorrelação parcial da umidade, para os dados brutos, pode-se intuir fortemente, a partir dos gráficos, que é possível ajustar

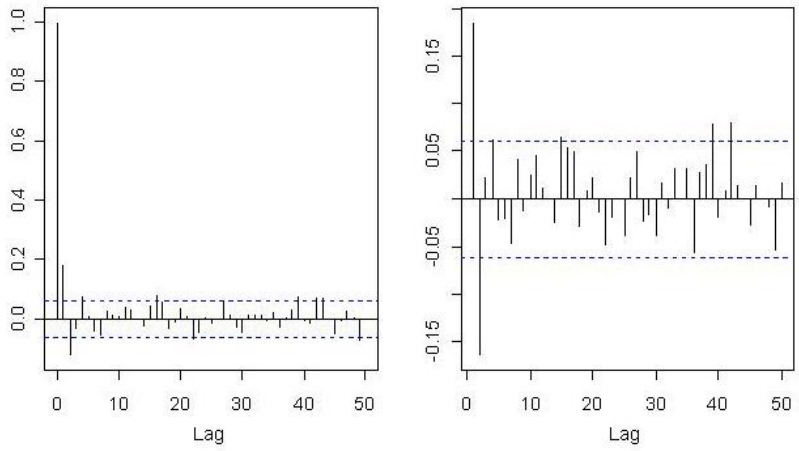

Fig. 6.6: FAC e FACP dos retornos da taxa de câmbio do Brasil

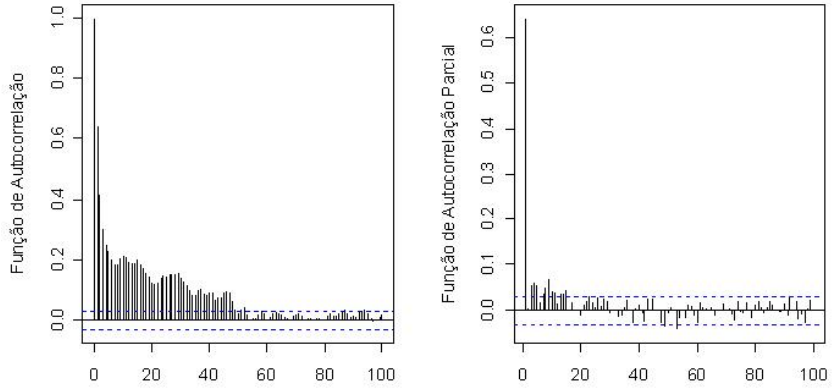

Fig. 6.7: FAC e FACP dos dados brutos da umidade

um modelo ARMA. Isto não impede que as informações dos quadrados dos dados brutos possam melhorar a precisão do modelo. Existe também a possibilidade de que não linearidades significativas comprometam a qualidade das previsões.

A existência de autocorrelação nos quadrados dos resíduos das previsões dos retornos da taxa de câmbio do Brasil possibilita a especificação de um modelo GARCH para melhorar a performance de previsão e, quando for possível, deve-se combinar um modelo GARCH com um modelo ARMA. A Figura 6.8 apresenta o gráficos da FAC e da FACP dos quadrados dos retornos da taxa de câmbio do Brasil.

Esta figura sugere graficamente que será possível ajustar um modelo ARMA-GARCH para o caso dos retornos da taxa de câmbio Brasil.

A função de autocorrelação e a função de autocorrelação parcial dos quadrados dos resíduos das previsões dos dados brutos da umidade é apresentada abaixo.

A Figura 6.9 sugere graficamente que possivelmente um modelo ARMA-GARCH poderá ser ajustado para se obter a previsão da umidade. As informações contidas nesta seção serão utilizadas para especificar os modelos ARMA-GARCH que serão apresentados mais adiante.

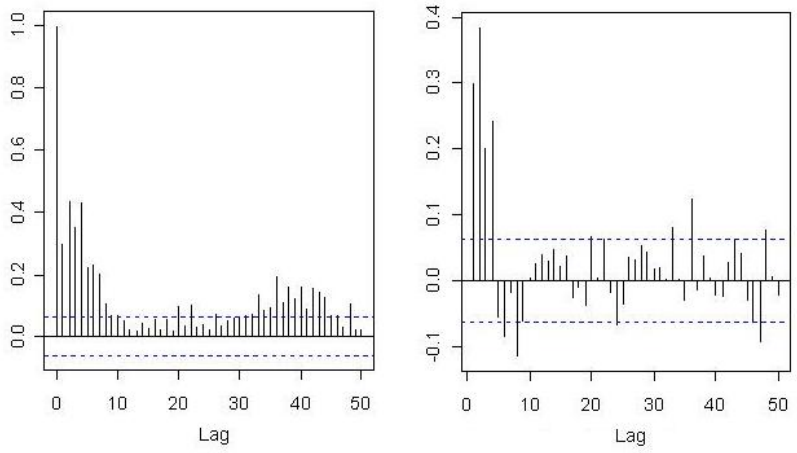

Fig. 6.8: FAC e da FACP dos quadrados dos resíduos das previsões dos retornos da taxa de câmbio do Brasil

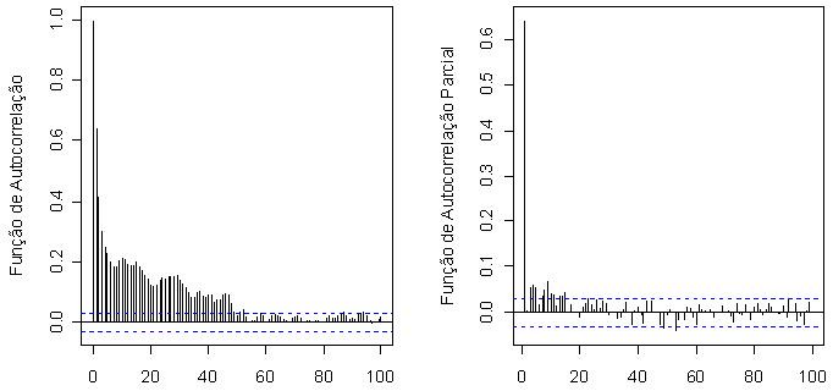

Fig. 6.9: FAC e FACP dos quadrados dos resíduos das previsões dos dados brutos da umidade

Entretanto, existe também a possibilidade de que não linearidades significativas comprometam a qualidade das previsões.

6.5.2 Análise da dependência temporal via teste BDS

O teste de dependência temporal de BROCK, DECKER e SHEINKMAN (1996), conhecido como teste BDS, consiste basicamente em testar a hipótese nula de que os valores amostrados da série são estatisticamente independentes (linear e não linear), utilizando a dimensão da correlação para avaliar a existência de dependência temporal em processos não lineares. A análise da dependência temporal linear e não linear é importante neste contexto porque, caso haja algum destes tipos de dependência temporal serial, tem-se uma sinalização de que não se trata de um processo diferença Martingale. Os resultados da aplicação deste teste para a taxa cambial brasileira, logaritmo, 1ª diferença e taxa de retornos são apresentados na Tabela 6.7.

m-dimensão	Dados brutos	Logaritmo	1ª diferença	Taxa de retornos
2	128,13	142,22	15,38	15,51
3	125,08	140,59	15,38	15,50
4	122,48	139,55	15,37	15,49
5	122,09	139,58	15,36	15,48
6	122,11	139,61	15,37	15,49
7	122,13	139,64	15,30	15,41
8	122,15	139,66	15,35	15,48
9	122,17	139,69	15,35	15,48
10	23,37	23,43	2,27	1,74

Tab. 6.7: Testes BDS para a taxa cambial brasileira, logaritmo, 1ª diferença e taxa de retornos

A estatística BDS para a taxa cambial brasileira foi calculada variando a dimensão da correlação de dois até dez. Existe uma forte dependência serial nesta série no tempo ao nível de significância de 0,05, cuja faixa de aceite da hipótese nula é de - 1,96 a 1,96. Até este ponto as hipóteses de passeio aleatório ou de diferença Martingale são rejeitadas. Para todas as dimensões os testes foram significantes. Entretanto, para o teste validar a dependência seria necessário somente que pelo menos uma delas estivesse fora da faixa de aceite para 0,05. Este teste não informa se a série tem dependência curta, média ou longa.

Os resultados da aplicação deste teste para a umidade na região de Londrina-PR, logaritmo, 1ª diferença e taxa de retornos são apresentados na Tabela 6.8.

m-dimensão	Dados brutos	Logaritmo	1ª diferença	Taxa de retornos
2	75,24	67,13	8,13	2,17
3	72,29	64,63	8,13	2,13
4	68,60	61,70	6,52	2,12
5	66,28	60,03	6,56	2,16
6	64,94	59,44	6,52	2,13
7	64,22	59,36	6,56	2,17
8	63,92	59,39	6,42	2,12
9	63,81	59,34	6,58	2,17
10	3,64	4,54	0,37	0,13

Tab. 6.8: Testes BDS para a umidade na região de Londrina-PR, logaritmo, 1ª diferença e taxa de retornos

Existe uma forte dependência serial na série no tempo da umidade na região de Londrina-PR ao nível de significância de 0,05. Até este ponto as hipóteses de passeio aleatório ou de diferença Martingale são rejeitadas.

Os resultados dos testes aplicados nesta seção indicam que, a partir dos gráficos da autocorrelação e autocorrelação parcial, os retornos da taxa cambial brasileira e os dados brutos da umidade indicaram clara dependência temporal linear. As duas séries apresentaram não linearidades ([HSIEH, 1988, HSIEH, 1989, HSIEH and KLEIDON, 1996]) e dependência temporal (linear e/ou não linear) via teste BDS. Logo, estas duas séries possivelmente apresentarão algum nível de previsão. É razoável que se utilize métodos de previsão não lineares como complementares ou substitutos dos métodos lineares.

6.6 Seleção das variáveis de entrada

Em [MEESE and ROGOFF, 1983] foram apontadas as dificuldades para o ajuste de modelos macroeconômicos para previsão da taxa de câmbio. Em [DORNBUSCH and FRANKEL, 1988] e [FRANKEL and FROOT, 1990] foi sugerido que a série temporal da taxa de câmbio apresenta uma variabilidade superior àquela detectada nas variáveis macroeconômicas fundamentais e, como os ganhos no mercado financeiro estão associados à volatilidade do ativo, é natural que existam forças especulativas atuando no mercado de câmbio, inclusive aquelas que não são refletidas em modelos baseados unicamente em fundamentos macroeconômicos.

Os trabalhos baseados em variáveis de microestruturas de mercado de câmbio, como as referências [EVANS and LYONS, 2002] e [GUIMARÃES and TABAK, 2004], encontraram conteúdo informacional neste tipo de variáveis, principalmente o conteúdo informacional relacionado aos movimentos especulativos atuantes no mercado. Já em [TABAK, 2002], utilizando a taxa de variância, encontrou evidências empíricas de que a variação diária e semanal da taxa cambial brasileira, já em regime de câmbio flutuante, tem comportamento de passeio aleatório. Entretanto, as referências [CARNEIRO and NETTO, 2005, CARNEIRO et al., 2004], utilizando métodos não lineares, apresentaram indícios de que existe algum grau de determi-

nismo e de previsibilidade.

Como existem mais de cento e cinqüenta variáveis explicativas catalogadas relacionadas às variações cambiais [KAMINSKY and REINHART, 1996], é razoável que se faça uma seleção de variáveis. Portanto, um número reduzido de variáveis, consideradas as mais expressivas no contexto atual, serão selecionadas pelo método de seleção de variáveis apresentado no Capítulo 3, a partir das seguintes áreas: macroeconômica, microestruturas de mercados e que refletem o apetite dos mercados internacionais por risco. Isto torna o modelo mais eficaz e parcimonioso.

Em relação à umidade na região de Londrina-PR, sabe-se que a temperatura e as chuvas guardam relação com esta variável. Entretanto, pouco se tem de informações sobre as outras candidatas possíveis a variáveis de entrada do modelo. Poucas séries relacionadas com esta variável foram obtidas. O interesse nesta série se deve à sua importância para o combate à doença conhecida como ferrugem asiática que causa prejuízos da ordem de bilhões de dólares aos agricultores dos países produtores de soja.

Faz-se a pré-seleção das variáveis via filtro e depois o método *wrapper* via redes neurais é utilizado para escolher o subconjunto de variáveis dentre aquelas pré-selecionadas. Este método de seleção de variáveis pode ser utilizado também para escolher as variáveis explicativas de modelos estruturais de previsão de séries temporais associadas a processos sobre o qual não existe conhecimento *a priori*.

6.6.1 Taxa de câmbio brasileira

O filtro e o *wrapper* não interagem, ou seja, o filtro seleciona os subconjuntos de características independentemente do algoritmo de aprendizagem. Uma rede neural é utilizada para realizar a seleção final das variáveis que, em conjunto, possibilitam a melhor qualidade de previsão. Isto se deve ao fato de que a abordagem baseada somente em filtros tradicionais tem alguns problemas, como: viés, em que diferentes características sugerem diferentes modelos de indução (filtros lineares, redes neurais e outros); características dependentes, que se forem consideradas aos pares, podem ser redundantes, mas pode ocorrer que uma variável necessite da outra para fornecer uma boa previsão.

Observa-se que a ptax é uma taxa de câmbio calculada ao final de cada dia pelo Banco Central do Brasil, é a taxa média de todos os negócios com dólares realizados naquela data no mercado interbancário de câmbio e não pode ser confundida com preço de fechamento - que podem ter um viés.

A Tabela 6.9 foi preenchida com os valores estimados de informação mútua entre a taxa de câmbio brasileira (ptax é a classe) e as candidatas a variáveis de entrada. Todas as variáveis são normalizadas antes de serem analisadas pelo filtro. A informação mútua entre a variável dependente (classe) e a candidata a variável de entrada (C-informação mútua, S(i,c), em que c representa a classe e i as variáveis candidatas) expressa a relevância da candidata a variável de entrada. Observa-se que a variável que apresentou maior valor de informação mútua (relevância) foi a taxa de câmbio euro/dólar US.

Os valores de informação mútua entre a variável mais relevante (euro/dólar US) e as outras variáveis de entrada restantes são apresentadas na Tabela 6.10.

Variável	Informação mútua
ptax e euro/dólar US	0,6063
ptax e libra/dólar US	0,6039
ptax e embiplus	0,5968
ptax e franco/dólar US	0,5926
ptax e yene/dólar US	0,5718
ptax e diferença ptax fed funds	0,5633
ptax e ibovespa	0,5248
ptax e vértice mês 5 da ettj	0,5070
ptax e vértice mês 6 da ettj	0,5054
ptax e vértice mês 4 da ettj	0,5040
ptax e vértice mês 7 da ettj	0,5038
ptax e vértice mês 8 da ettj	0,5003
ptax e vértice mês 9 da ettj	0,4965
ptax e vértice mês 3 da ettj	0,4932
ptax e vértice mês 10 da ettj	0,4928
ptax e vértice mês 11 da ettj	0,4879
ptax e vértice mês 12 da ettj	0,4826
ptax e vértice mês 2 da ettj	0,4812
ptax e UC1CurncyPxLow	0,4791
ptax e UC1CurncyPxHigh	0,4763
ptax e vértice mês 2 da ettj	0,4723
ptax e selic	0,4642
ptax e cotação soja	0,4513
ptax e T-bill 180 dias	0,4351
ptax e embiBR	0,3481
ptax e DiferencaHighLow	0,3415

Tab. 6.9: Cálculo da informação mútua (relevância, C-informação mútua, S(i,c)) entre a variável dependente (ptax) e as candidatas a variáveis de entrada

Variável	Informação mútua
euro/dólar US e libra/dólar US	0,4354
euro/dólar US e embiplus	*0,3893*
euro/dólar US e franco/dólar US	0,5036
euro/dólar US e yene/dólar US	0,5095
euro/dólar US e ibovespa	0,6201
euro/dólar US e vértice mês 5 da ettj	*0,5503*
euro/dólar US e vértice mês 6 da ettj	*0,5512*
euro/dólar US e vértice mês 4 da ettj	*0,5447*
euro/dólar US e vértice mês 7 da ettj	*0,5509*
euro/dólar US e vértice mês 8 da ettj	*0,5485*
euro/dólar US e vértice mês 9 da ettj	*0,5458*
euro/dólar US e vértice mês 3 da ettj	*0,5328*
euro/dólar US e vértice mês 10 da ettj	*0,5427*
euro/dólar US e vértice mês 11 da ettj	*0,5381*
euro/dólar US e vértice mês 12 da ettj	*0,5329*
euro/dólar US e vértice mês 2 da ettj	*0,5182*
euro/dólar US e UC1CurncyPxLow	*0,6036*
euro/dólar US e UC1CurncyPxHigh	*0,6069*
euro/dólar US e vértice mês 2 da ettj	*0,5079*
euro/dólar US e selic	*0,4691*
euro/dólar US e cotação soja	0,3205
euro/dólar US e T-bill 180 dias	*0,5231*
euro/dólar US e embiBR	*0,3893*
euro/dólar US e DiferencaHighLow	*0,4032*
euro/dólar US e Diferença ptax fed funds	0,5522

Tab. 6.10: A informação mútua entre a variável de entrada mais relevante neste passo (euro/dólar US) e as outras variáveis de entrada (F-informação mútua, S(i,j))

Aquelas variáveis da Tabela 6.10 com maior F-informação mútua (S(i,j) que a C-informação mútua (S(i,c)) correspondente são consideradas redundantes e são eliminadas. Estas variáveis são destacadas em itálico nesta tabela. Caso contrário, são consideradas não redundantes e não são eliminadas.

A segunda variável mais relevante restante (S(i,c)) da Tabela 6.9 é a taxa de câmbio libra/dólar US. A Tabela 6.11 apresenta a S(i,j)) entre esta variável e as variáveis restantes da Tabela 6.10. As variáveis redundantes são aquelas com S(i,j) maior que as respectivas S(i,c), estão escritas em itálico e serão eliminadas. A terceira variável mais relevante restante (S(i,c)) é o embiplus. A Tabela 6.12 apresenta a S(i,j)) entre esta variável e as variáveis restantes da Tabela 6.11. As variáveis redundantes são aquelas com S(i,j) maior que as respectivas S(i,c).

Variável	Informação mútua
libra/dólar US e ibovespa	*0,6159*
libra/dólar US e embiplus	0,5246
libra/dólar US e franco/dólar US	0,4777
libra/dólar US e yene/dólar US	0,4904
libra/dólar US e cotação soja	0,2627
libra/dólar US e diferença ptax fed funds	0,5323

Tab. 6.11: A informação mútua entre a segunda variável de entrada mais relevante (libra/dólar US) e as outras variáveis de entrada (F-informação mútua, S(i,j)) restantes

Variável	Informação mútua
embiplus e franco/dólar US	0,5722
embiplus e yene/dólar US	0,5706
embiplus e cotação soja	*0,4797*
embiplus e diferença ptax fed funds	0,5620

Tab. 6.12: A informação mútua entre a terceira variável de entrada mais relevante (embiplus) e as outras variáveis de entrada (F-informação mútua, S(i,j)) restantes

A quarta variável mais relevante restante (S(i,c)) é a taxa de câmbio franco/dólar US. A Tabela 6.13 apresenta a S(i,j)) entre esta variável e as variáveis restantes da Tabela 6.12. As variáveis redundantes são aquelas com S(i,j) maior que as respectivas S(i,c) e serão eliminadas.

Até o presente momento, os resultados alcançados na seleção de variáveis são coerentes com o contexto do mercado de câmbio observado no período analisado. A diferença entre cupom cambial e a *fed funds* sinaliza a grande diferença entre as taxas de juros internas e externas. Esta pesquisa constatou que o *high-low* não tem expressão para ser escolhido, corroborando vários estudos internacionais que sugerem que a cotação do dólar *forward* não influencia na cotação do dólar *spot*, embora no mercado doméstico tradicionalmente o *high-low* tenha sido considerado um excelente sinalizador de expectativas de curto prazo. As taxas de câmbio euro/dólar US e libra/dólar US apresentaram expressiva relevância, sinalizando a influência dos mercados internacionais de câmbio.

Variável	Informação mútua
franco/dólar US e yene/dólar US	0,5417
franco/dólar US e diferença ptax fed funds	0,5630

Tab. 6.13: A informação mútua entre a quarta variável de entrada mais relevante (franco/dólar US) e as outras variáveis de entrada (F-informação mútua, S(i,j)) restantes

Variável	Informação mútua
yene/dólar US e diferença ptax fed funds	0,5582

Tab. 6.14: A informação mútua entre a quinta variável de entrada mais relevante (yene/dólar US) e as outras variáveis de entrada (F-informação mútua, S(i,j)) restantes

A quinta variável mais relevante restante (S(i,c)) é a taxa de câmbio yene/dólar US. A Tabela 6.14 apresenta a S(i,j)) entre esta variável e as variáveis restantes da Tabela 6.13. As variáveis redundantes são aquelas com S(i,j) maior que as respectivas S(i,c).

A Tabela 6.15 apresenta o subconjunto de variáveis escolhido por meio do filtro. Este subconjunto de variáveis é utilizado para os testes com o *wrapper* para a previsão da variação da taxa de câmbio.

O algoritmo de eliminação para trás é utilizado para implementar o *wrapper*. Assim, entre todos os possíveis subconjuntos de variáveis com menos uma variável de entrada, seleciona-se a combinação de variáveis de entrada que fornece a melhor função de avaliação baseada no NMSE. Este processo interativo deve continuar até que todas as variáveis sejam analisadas separadamente ou até que a função de avaliação da melhor combinação de variáveis de entrada da interação corrente seja pior que a melhor da anterior. Na seleção de variáveis da taxa de câmbio brasileira via *wrapper*, como são poucas variáveis que devem ser analisadas, todos os subconjuntos de variáveis possíveis foram analisados, evitando que o algoritmo pudesse ficar preso em mínimos locais.

A Tabela 6.16 apresenta o melhor subconjunto de variáveis escolhido pelo *wrapper* via modelos RBF ARIA GAUSS para as previsões da série temporal dos retornos da taxa cambial brasileira. Logo, as variáveis escolhidas via *wrapper* são: a própria ptax; o euro; a diferença entre a taxa do cupom cambial e a taxa da *fed funds*. Assim, fica confirmada a influência da diferença entre as taxas de juros do cupom cambial e da *fed funds*, ou seja, a grande diferença entre as taxas de juros internas e externas. A taxa de câmbio euro/dólar US confirmou sua expressiva relevância, sinalizando a influência dos mercados internacionais de câmbio.

Variável
ptax, euro/dólar US, libra/dólar US, embiplus, franco/dólar US
yene/dólar US, diferença cupom cambial e fed funds

Tab. 6.15: Subconjunto de variáveis utilizado para os testes com o *wrapper* para a previsão da variação da taxa de câmbio

Método	Candidatas a variáveis entrada	r	Neur.	NMSE
RBF ARIA GAUSS	ptax, euro e dif fed funds	0,002	22	0,5850

Tab. 6.16: Melhor subconjunto de variáveis escolhido pelo *wrapper* via modelos RBF ARIA GAUSS para as previsões da série temporal dos retornos da taxa cambial brasileira

Nestas simulações foram utilizadas 1.000 (hum mil) amostras para ajustar os parâmetros da rede e 200 (duzentas) amostras para os testes.

6.6.2 Umidade na região de Londrina-PR

Já existem estudos sobre a umidade na troposfera que apontam para uma relação entre a umidade específica e a temperatura. Na troposfera, tanto na parte baixa como na alta, quando a temperatura aumenta a umidade também aumenta. Já uma relação inversa intensa ocorre quando a umidade relativa é comparada [SUN and OORT, 1995]. Por outro lado, a umidade é a quantidade de vapor de água no ar e é razoável que durante uma chuva a umidade seja fortemente influenciada. Resta saber a importância desta influência ao longo do tempo já que não chove na terra continuamente.

O *wrapper* implementado faz seleção de variáveis de entrada para o modelo de previsão da umidade horária no microclima da região de Londrina-PR. Os dados coletados sobre a umidade nesta região sugerem que as variações são rápidas, muito bruscas e de intensidade elevada.

Foram obtidas poucas variáveis candidatas: temperaturas mínima, média e máxima e a chuva. Utilizando o mesmo procedimento anterior, a Tabela 6.17 apresenta o subconjunto de variáveis escolhido por meio do filtro, composto por variáveis que são muito relevantes e por aquelas que são pouco relevantes e não redundantes. Este é o subconjunto de variáveis considerado o mais indicado para os testes com o *wrapper* em modelos neurais multivariáveis para a previsão da umidade.

Na seleção de variáveis da umidade, como são poucas variáveis que devem ser analisadas, todos os subconjuntos de variáveis possíveis foram analisados, evitando que o algoritmo pudesse ficar preso em mínimos locais. A Tabela 6.18 apresenta o melhor subconjunto de variáveis escolhido pelo *wrapper* via modelos RBF ARIA GAUSS para as previsões da série temporal da umidade: a própria umidade; a temperatura mínima; a temperatura média.

Foi confirmada a relação entre a umidade e a temperatura, entretanto a temperatura máxima não apresentou sinergia com as outras variáveis analisadas e pode ter ocorrido que a análise aos pares realizada pelo filtro não captou a redundância desta variável com o conjunto das outras variáveis. Já a chuva não foi selecionada pelos critérios utilizados.

Variável
umidade
temperatura mínima
temperatura média
temperatura máxima

Tab. 6.17: Subconjunto de variáveis considerado o mais indicado para os testes com o *wrapper* em modelos neurais multivariáveis para a previsão da umidade

Método	Candidatas a variáveis entrada	r	Neur.	NMSE
RBF ARIA GAUSS	umid, tmin e tmed	0,000097	97	0,8558

Tab. 6.18: Melhor subconjunto de variáveis escolhido pelo *wrapper* via modelos RBF ARIA GAUSS para as previsões da série temporal da umidade

6.7 Reconstrução dinâmica e análise dinâmica não linear

Um princípio antigo da ciência era que todos os sistemas deterministas eram previsíveis. Posteriormente, descobriu-se que apesar dos sistemas lineares terem solução fechada, poucos sistemas não lineares tinham solução fechada. Posteriormente, foram descobertos os sistemas caóticos que são previsíveis somente no curto prazo. No Capítulo 4 foram apresentadas algumas técnicas empíricas para se fazer a reconstrução dinâmica de séries temporais e para o reconhecimento do comportamento caótico. Nesta seção, apresentam-se os resultados das aplicações de algumas destas técnicas.

Inicialmente, investiga-se a dimensão da correlação e o coeficiente de LYAPUNOV. A Tabela 6.19 apresenta os valores estimados destes parâmetros para o modelo de previsão dos retornos da taxa de câmbio brasileira.

A dimensão da correlação é um tipo de dimensão probabilística que expressa o nível de complexidade do sistema. Um atrator simples tem dimensão inteira; já um atrator estranho (fractal) tem dimensão da correlação fracionária.

Variável	Dim. da correlação	Maior exp. de LYAPUNOV
ptax	2,52	0,0237
euro/dólar US	4,13	0,0135
Dif. cupom cambial e fed funds	4,33	0,0145

Tab. 6.19: Valores estimados para a dimensão da correlação e do expoente de LYAPUNOV para o modelo de previsão dos retornos da taxa de câmbio brasileira

O expoente de LYAPUNOV é a generalização dos autovalores associados ao ponto de equilíbrio, ou seja, é igual à parte real do autovalor próximo ao ponto de equilíbrio e indica a taxa de contração ($\lambda_i < 0$) e de expansão ($\lambda_i > 0$). Este expoente indica a taxa de divergência entre as trajetórias e fornece a sensibilidade às condições iniciais, que é uma das características dos sistemas caóticos. Conseqüentemente sinaliza o tipo de estabilidade do sistema. Um expoente de LYAPUNOV positivo indica que o sistema tem sensibilidade às condições iniciais e quanto maior for este expoente maior é a sensibilidade e menor é a capacidade de previsão.

O inverso do expoente de LYAPUNOV sinaliza a capacidade de previsão existente. Sensível dependência às condições iniciais e dimensão fractal finita (estimada pelo valor da dimensão de correlação) são condições essenciais para se averiguar a existência de um sistema caótico. Daí advém, incluindo também os efeitos de realimentação, o fato da previsão a longo prazo ser praticamente impossível.

Os resultados apresentados acima para o exponente de LIAPUNOV são semelhantes aos ilustrados em [VANDROVYCH, 2005] em que a dinâmica de seis taxas de câmbio dos principais países desenvolvidos são analisadas. Embora tenha obtido um número de estimativas positivas do exponente de LIAPUNOV, são valores muito pequenos e o autor acredita que é mais apropriado interpretar estes dados como um indicador de origem estocástica da série. Entretanto, os dados não devem conter muito ruído porque os métodos não conseguem detectar o caos subjacente, mesmo que a quantidade de ruído seja pequena.

A partir da Tabela 6.19, observa-se que a dimensão da correlação das séries são fracionárias, com valores médios. Estes valores são um pouco diferentes dos apresentados em VANDROVYCH (2005) em que as estimativas da dimensão de correlação indicaram a complexidade elevada em toda a série, sugerindo que as séries são processos estocásticos ou processos deterministas com dimensões elevadas. Quando é pequeno indica algum grau de determinismo na série. Os coeficientes positivos de Lyapunov são muito pequenos. Isto indica sensibilidade às condições iniciais, mesmo que pequena. Estes resultados em conjunto indicam que é mais apropriado interpretar estes dados como um indicador de origem estocástica da série.

A Tabela 6.20 apresenta os valores estimados para a dimensão da correlação e o expoente de Lyapunov da variação da umidade.

Variável	Dim. da correlação	Maior exp. de Lyapunov
umidade	4,0483	0,0046
temperatura mínima	3,0750	0,0220
temperatura média	5,5159	0,0200

Tab. 6.20: Valores estimados para a dimensão da correlação e do expoente de LYAPUNOV para o modelo de previsão da umidade

A Tabela 6.20 apresenta os valores médios para a dimensão da correlação de todas as séries e os maiores coeficientes de LYAPUNOV de cada série, que embora muito pequenos, são positivos. Os resultados do cálculo do exponente máximo de LYAPUNOV sugerem qualitativa e quantitativamente que a série tem pequeno grau de determinismo. Isto também indica uma pequena sensibilidade às condições iniciais. Estes resultados sinalizam que a umidade apresenta algum grau de determinismo.

Em [TAKENS, 1981] foi proposto que se uma série temporal é obtida de um sistema determinista, uma reconstrução dinâmica existe, consistindo no *lag* e na dimensão de imersão (*embedding*), que ao se deslocar no tempo, forma o que se chama janela de previsão dinâmica (JPD). Nem sempre estamos seguros se a série pertence a um sistema determinista, embora já existam métodos estatísticos para isto, mas que necessitam de um grande número de amostras para testar esta hipótese. Assim, este método também é usualmente utilizado em sistemas com um certo grau de estocasticidade.

A Tabela 6.21 apresenta a reconstrução dinâmica do subconjunto de variáveis escolhido para o início das simulações com modelos de previsão da variação da taxa de câmbio. O *lag* é estimado pelo método apresentado em [WOLF et al., 1985] via o software tstool do Instituto Max Planck da Alemanha, pelo de CELLUCCI et al. (2005) para o cálculo da informação mútua e por meio do método RBF PCA proposto. A dimensão de *embedding* é estimada por meio do conceito de falsos vizinhos mais próximos, também via o software tstool, e por meio do método RBF PCA proposto.

O valor do *lag* estimado via software tstool e método de CELLUCCI et al. (2005) são representados por L(tstool) e L(Cellu.), respectivamente. O valor de L estimado pelo método RBF PCA é representado por L(RNA) e chega-se a este valor ajustando os parâmetros da rede em função da otimização da função objetivo (NMSE) associada à qualidade de previsão. O software tstool, utilizando o método dos falsos vizinhos mais próximos, foi utilizado para determinar o valor da dimensão de imersão M(tstool) e a rede RBF PCA para estimar M(RNA). Observando a Tabela 6.21, nota-se que os valores dos *lags* L(Cellu.) e L(RNA) são semelhantes, mas os valores da dimensão de imersão M(tstool) e M(RNA) são diferentes.

Variável	L(tstool)	L(Cellu.)	L(RNA)	M(tstool)	M(RNA)
ptax	3	2	2	2	28
euro/dólar US	5	3	-	2	-
Dif. cup. camb. e fed funds	5	2	-	2	-

Tab. 6.21: Reconstrução dinâmica do subconjunto de variáveis considerado para o início das simulações com modelos de previsão dos retornos da taxa de câmbio do Brasil

O valor de L deve ser o menor possível, mas suficientemente grande para minimizar a autocorrelação entre os componentes da série no tempo. Na análise de séries temporais não-lineares, a informação mútua, que consegue medir a dependência entre duas variáveis não-lineares, fornece uma estimativa melhor para o *lag* que a função de autocorrelação, que considera apenas a dependência serial linear entre variáveis.

A Tabela 6.22 apresenta a reconstrução dinâmica do subconjunto de variáveis para o início das simulações com modelos de previsão da variação da umidade. O mesmo procedimento usado para a taxa de câmbio brasileira é utilizado para a umidade.

Os valores dos *lags* e das dimensões de imersão das Tabelas 6.21 e 6.22 são utilizados como valores iniciais nas simulações dos modelos para determinar o subconjunto ótimo de variáveis de entrada (*wrapper*) e a janela de previsão dinâmica dos modelos neurais que serão apresentados na próxima seção. A partir das simulações com os modelos neurais, serão determinadas as janelas de previsão com melhor desempenho em função da qualidade preditiva dos modelos.

Ressalta-se a diferença entre um fenômeno essencialmente determinista e um essencialmente aleatório, em que não é possível encontrar nenhum modelo que permita descrever seu comportamento sob condições arbitrárias. A *priori*, um processo não explicável adequadamente pela teoria disponível poderá ter origem tanto essencialmente determinista como essencialmente aleatória. Qualquer processo, determinista ou não, permite que se associe a ele uma estatística, isto é, estimativas de valores médios, variância, correlações entre as variáveis envolvidas. O fato de se associar ao processo uma estatística não o transforma em um processo essencialmente aleatório. Já para um processo essencialmente aleatório, a única modelagem possível é a probabilística [KOVÁCS, 2002].

Variável	L (tstool)	L (Cellucci)	L (RNA)	M (tstool)	M (RNA)
umidade	4	2	2	4	50
temperatura mínima	7	3	-	4	-
temperatura média	6	2	-	4	-

Tab. 6.22: Reconstrução dinâmica do subconjunto de variáveis considerado para o início das simulações com modelos de previsão da umidade

Os valores dos *lags* e das dimensões de imersão das Tabelas 6.21 e 6.22 indicam que o método apresentado em [CELLUCCI et al., 2005] para o cálculo da informação mútua gerou *lags* semelhantes aos ajustados via RNA, já as dimensões de imersão calculadas via tstool são significativamente diferentes dos ajustados via RNA.

6.8 Ajuste dos modelos de previsão

As seguintes classes terão modelos ajustados: diferença Martingale, ARMA-GARCH, neural e combinação dos melhores modelos. O objetivo principal é verificar se as séries no tempo analisadas são diferenças Martingale. Mais especificamente são implementados os seguintes modelos: uma diferença Martingale (*benchmark*) na média; ARMA-GARCH; rede neural tipo RBF com função de base gaussiana e centros ajustados via PCA (RBF PCA GAUSS); rede neural tipo RBF com função de base spline e centros ajustados via PCA (RBF PCA SPLINE); rede neural tipo RBF com função de base gaussiana e centros ajustados via algoritmo ARIA (RBF ARIA); e a combinação dos melhores modelos (**combinado**), descartando aqueles modelos com qualidade de previsão muito inferior.

O modelo RBF ARIA é multivariado com variáveis selecionadas por meio do método de seleção de variáveis proposto nesta tese, ou seja, o filtro já foi implementado anteriormente e os resultados consistiram em um subconjunto de variáveis pré-selecionadas por meio da relevância e da redundância. Finalmente, o *wrapper* foi implementado e escolheu o melhor subconjunto de variáveis dentre aquelas pré-selecionadas pelo filtro, avaliadas em conjunto, o que resulta em um modelo mais parcimonioso e com melhor qualidade de previsão.

6.8.1 ARMA-GARCH

Nas análises anteriores da série temporal dos retornos da taxa cambial brasileira, a Figura 6.6 apresentou o gráfico da FAC e da FACP da série temporal dos retornos da taxa cambial brasileira. A análise desta figura indica a existência de uma dependência linear e sinaliza que o ajuste de um modelo autoregressivo pode ser viável. A Figura 6.8 apresentou o gráfico da FAC e da FACP da série temporal dos quadrados dos resíduos e também indica a existência de uma dependência linear. Percebe-se também a existência de heteroscedasticidade, sugerindo desta forma o ajuste de um modelo ARMA-GARCH.

Na Tabela 6.23 é apresentado o modelo ARMA-GARCH ajustado para os retornos da taxa cambial brasileira e na Tabela 6.24 são ilustradas as estatísticas deste modelo.

Parâmetros do modelo	Coeficiente	Erro padrão	Estatística Z	P-valor
ϕ_1	0,1267	0,0365	3,4750	0,0005
ϕ_2	-0,0745	0,0327	-2,2753	0,0229
C	1,3e-6	4,3e-7	3,0266	0,0025
α_1	0,2018	0,0267	7,5326	0,0000
β_1	0,7978	0,0234	33,9692	0,0000

Tab. 6.23: Modelo ARMA-GARCH ajustado para fazer as previsões da série temporal dos retornos da taxa cambial brasileira

Estatística	Valor	Estatística	Valor
Erro padrão	0,0103	AIC	-6,8474
MSE	0,1065	SIC	-6,8125

Tab. 6.24: Estatísticas de ajuste do modelo ARMA-GARCH especificado para fazer as previsões da série temporal dos retornos da taxa cambial brasileira

Os valores do erro padrão e do MSE estão relativamente baixos e os critérios de informação AIC e SIC estão numa faixa de valores muito boa. Ou seja, as estatísticas do modelo anterior indicam que este método é muito promissor para esta aplicação e será avaliado em conjunto com os outros modelos em termos de qualidade de previsão.

A Figura 6.7 apresentou o gráfico da FAC e da FACP da série da umidade. A análise desta figura indica a existência de uma dependência linear entre as observações e que um modelo autoregressivo possivelmente será adequado. Analisando o gráfico da Figura 6.9, que apresentou o gráfico da FAC e da FACP da série dos quadrados dos resíduos, percebe-se a existência de heteroscedasticidade, sugerindo também que se pode ajustar um modelo ARMA-GARCH. O modelo ARMA-GARCH ajustado para a umidade é o apresentado na Tabela 6.25. As estatísticas do modelo da Tabela 6.25 estão ilustradas na Tabela 6.26.

	Coeficiente	Erro padrão	Estatística Z	P-valor
C	73,5646	0,9556	76,9802	0,0000
ϕ_1	0,6065	0,0137	0,0137	0,0000
ϕ_5	0,4333	0,0127	0,0127	0,0007
ϕ_9	0,0525	0,0137	0,0137	0,0001
ϕ_{11}	0,0318	0,0137	0,0137	0,0204
ϕ_{15}	0,0469	0,0127	0,0127	0,0002
ϕ_{28}	0,0389	0,0121	0,0121	0,0013
ϕ_{46}	0,0220	0,0122	0,0122	0,0710
ϕ_{49}	-0,0327	0,0123	0,0123	0,0080
C	2,1157	0,6978	3,0319	0,0024
α_1	0,0360	0,0065	5,4985	0,0000
β_1	0,9494	0,0098	96,2658	0,0000

Tab. 6.25: Modelo ARMA-GARCH ajustado para fazer as previsões da série temporal da umidade

Estatística	Valor	Estatística	Valor
Erro padrão	11,9854	AIC	7,7694
MSE	565837,2	SIC	7,7885

Tab. 6.26: Estatísticas de ajuste do modelo ARMA-GARCH ajustado para fazer as previsões da série temporal da umidade

Os valores do erro padrão e do MSE estão excessivamente altos e os critérios de informação AIC e SIC estão fora da faixa de valores razoáveis. Ou seja, as estatísticas do modelo anterior indicam que este método não é promissor para esta aplicação.

6.8.2 Rede RBF PCA

Foram ajustados os seguintes modelos com rede RBF PCA: rede RBF com funções de base gaussiana com centros ajustados via PCA (RBF PCA GAUSS); rede RBF com funções de base spline fina com centros ajustados via PCA (RBF PCA SPLINE). Os *lags* L e a dimensão de *embedding* M destes métodos são ajustados durante o treinamento das redes neurais. A tabela abaixo apresenta estes valores e os valores dos *lags* são semelhantes aos valores fornecidos pelo método de CELLUCCI et al. (2005). Já os valores de M são maiores que os valores estimados via reconstrução dinâmica apresentados na Tabela 6.21. Acredita-se que o método RBF PCA consegue capturar informações nos vetores formados por amostras passadas mais distantes no tempo, vide FACP da Figura 6.6, e armazená-las na rede. A Tabela 6.27 apresenta o melhor modelo e ilustra o ajuste de outros modelos neurais tipo rede RBF PCA para as previsões da série dos retornos da taxa cambial. Chega-se aos valores de L e M ajustando os parâmetros da rede em função da otimização da função objetivo (NMSE) associada à qualidade de previsão.

Método	L	M	NMSE	$BK^{(1)}$	$PT^{(2)}$	$DM^{(3)}$
RBF PCA GAUSS	2	28	0,7715	0,6250	0,9826	0,0012
RBF PCA SPLINE	2	28	0,7439	0,0113	1,1279	0,0010
RBF PCA GAUSS fator adaptativo	2	28	0,7302	0,1578	1,2697	0,0016
RBF PCA SPLINE fator adaptativo	2	28	0,3252	0,1562	1,3156	0,0063
RBF PCA GAUSS fator adapt. e reg.	2	28	0,5860	0,1473	1,1232	0,0023
RBF PCA SPLINE fator adap. e reg.	**2**	**28**	**0,3250**	**0,1239**	**0,7016**	**0,0063**

Tab. 6.27: Ajuste dos modelos neurais tipo rede RBF PCA para as previsões da série temporal dos retornos da taxa cambial brasileira

(1) Teste apresentado em [BLAKE and KAPETANIOS, 2003];
(2) Teste apresentado em [PESARAN and TIMMERNANN, 1992];
(3) Teste apresentado em [DIEBOLD and MARIANO, 1995].

O modelo RBF PCA SPLINE com fator adaptativo e regularização foi o que apresentou os melhores resultados, com as estatísticas de testes dentro dos valores críticos, mas o modelo RBF PCA SPLINE com fator adaptativo sem regularização apresentou resultado bem próximo, sinalizando aparentemente que a regularização nesta aplicação influiu pouco. Mesmo assim, a utilização da função spline, do fator adaptativo e da regularização foram positivos em todos os sentidos. A normalização com média e desvio padrão não melhorou os resultados. Talvez porque se trate de valores de baixa magnitude e de mesma escala.

No caso da umidade, a Tabela 6.28 apresenta os valores de L e M, ajustados durante o treinamento das redes neurais, sendo que os valores dos *lags* são próximos aos valores fornecidos pelo método de CELLUCCI et al. (2005). O valor da dimensão de imersão M ajustada via *wrapper*, utilizando o modelo neural RBF PCA, de acordo com a Tabela 6.22, também é muito maior que o valor estimado via reconstrução dinâmica. Acredita-se que o modelo neural consegue capturar o efeito da longa dependência, indicado na Figura 6.7, presente nos dados da umidade.

Método	L	M	NMSE	$BK^{(1)}$	$PT^{(2)}$	$DM^{(3)}$
RBF PCA GAUSS	2	50	0,5666	0,2343	1,9411	9,1683
RBF PCA SPLINE	2	50	0,4577	0,1223	1,8976	6,9346
RBF PCA GAUSS fator adapt.	2	50	0,5553	0,1593	1,2684	7,1859
RBF PCA SPLINE fator adapt.	2	50	0,4390	0,0374	1,4224	9,3066
RBF PCA GAUSS fator adapt. e reg.	2	50	0,4640	0,0015	3,8258	8,6588
RBF PCA SPLINE fator adap. e reg.	**2**	**50**	**0,3450**	**0,0457**	**0,8145**	**9,1147**

Tab. 6.28: Ajuste dos modelos neurais tipo rede RBF PCA para as previsões da série temporal da umidade

(1) Teste apresentado em [BLAKE and KAPETANIOS, 2003];
(2) Teste apresentado em [PESARAN and TIMMERNANN, 1992];
(3) Teste apresentado em [DIEBOLD and MARIANO, 1995].

O modelo RBF PCA SPLINE com fator adaptativo e regularização foi o que apresentou os melhores resultados, com as estatísticas de testes dentro dos valores críticos, sinalizando aparentemente que a regularização nesta aplicação possibilitou melhores previsões. A utilização da função spline, do fator adaptativo e da regularização foram positivos em todos os sentidos. A normalização com média e desvio padrão não melhorou os resultados. Talvez pelos mesmos motivos expostos acima.

6.8.3 Rede RBF ARIA

A abordagem conhecida como *wrapper* (empacotamento) consiste em utilizar um algoritmo de indução para fazer a avaliação dos subconjuntos de variáveis. As vantagens principais deste método são: levar em conta o viés do algoritmo de indução; considerar as variáveis dentro do contexto. A princípio a busca é exponencial, mas pode-se implementar buscas estocásticas (algoritmos genéticos, *simulated annealing* e outras) ou seqüenciais (busca direta, eliminação para trás e outras). A Figura 3.3 ilustra com mais detalhes o método *wrapper*.

Durante o aprendizado supervisionado de máquina, quando se utiliza os métodos *wrapper*, é importante se obter uma boa generalização, utilizando uma criteriosa seleção do melhor modelo. O problema da generalização se torna mais crítico quando os dados são incompletos ou carregam ruídos.

Foram ajustados modelos neurais multivariados em que os centros são ajustados via algoritmo ARIA (rede RBF ARIA). Estes modelos têm como variáveis de entrada aquelas selecionadas na Seção 6.6. Os resultados sinalizam que a normalização via média e desvio padrão melhoram a qualidade das previsões. São ajustadas somente redes RBF com funções de base gaussianas.

A Tabela 6.29 apresenta o melhor modelo e ilustra o ajuste de outros modelos neurais tipo rede RBF ARIA GAUSS para as previsões dos retornos da taxa cambial brasileira

Candidatas a variáveis entrada	r	Neurônios	NMSE
ptax, euro, libra, franco, yene e dif fed funds	0,0100	285	0,7483
ptax, euro, libra, franco e yene	0,0100	189	0,6974
ptax, euro, libra, franco e dif fed funds	0,0100	134	0,6346
ptax, euro, libra, yene e dif fed funds	0,0100	129	0,7256
ptax, euro, franco, yene e dif fed funds	0,0100	119	0,6584
ptax, libra, franco, yene e dif fed funds	0,0100	199	0,7379
ptax, euro, libra e yene	0,0100	159	0,7359
ptax, euro, libra e dif fed funds	0,0100	149	0,6114
ptax, euro, yene e dif fed funds	0,0100	187	0,6578
ptax, libra e dif fed funds	0,0010	82	0,6349
ptax, euro, libra e yene	0,0010	211	0,6679
ptax, euro e dif fed funds	**0,0020**	**22**	**0,5850**
ptax e dif fed funds	0,0008	182	0,7342
ptax e euro	0,0008	111	0,7289

Tab. 6.29: Ajuste dos modelos RBF ARIA GAUSS para as previsões da série temporal dos retornos da taxa cambial brasileira

O modelo RBF ARIA GAUSS, tendo como variáveis de entrada a ptax, o euro e a diferença entre o cupom cambial e a *fed funds*, foi o que apresentou os melhores resultados, indicando que a escolha das variáveis de entrada no *wrapper* da série no tempo dos retornos da taxa cambial brasileira era coerente. O parâmetro r de ajuste do algoritmo ARIA foi estimado em 0,002 e a rede neural RBF ficou com 22 neurônios.

Os resultados obtidos nas previsões desta série temporal com este tipo de modelo sinalizam que existem *clusters* com capacidade preditiva, mas a série têm características estocásticas.

A Tabela 6.30 apresenta o melhor modelo e ilustra o ajuste de outros modelos neurais tipo rede RBF ARIA GAUSS para as previsões da série temporal da umidade. A escolha do melhor modelo para a umidade segue o mesmo método (filtro seguido de *wrapper* baseado em redes neurais) utilizado para escolher o melhor modelo de previsão para a taxa de câmbio.

Candidatas a variáveis entrada	r	Neurônios	NMSE
umid, tmin, tmed, tmax e chuva	0,0003	193	1,9740
umid, tmin, tmed e tmax	0,0002	101	1,2740
umid, tmin e tmed	**0,000097**	**97**	**0,8558**
umid e tmed	0,0002	89	1,6256

Tab. 6.30: Ajuste dos modelos RBF ARIA GAUSS para as previsões da série temporal da umidade

O modelo RBF ARIA GAUSS, tendo como variáveis de entrada a umidade, a temperatura mínima e temperatura média, foi o que apresentou os melhores resultados, indicando que a escolha das variáveis de entrada no *wrapper* da série no tempo da umidade era coerente. O parâmetro r de ajuste do algoritmo ARIA foi estimado em 0,000097 e a rede neural RBF ficou com 97 neurônios.

6.9 Avaliação conjunta dos modelos de previsão

Os investidores tentam mais maximizar lucros do que minimizar erros, logo os critérios de precisão como o MFSE (*mean square forecast error*) e o MAFE (*mean absolute forecast error*) podem não ser os mais apropriados para avaliar a predição de uma série temporal financeira.

É mais importante utilizar medidas econômicas de avaliação da previsão de uma série temporal financeira como o MFTR (*mean forecast trading returns*) do que um critério estatístico que minimiza erros como o MSFE, que pode gerar distorções de acordo com a transformação utilizada. O MCFD (*mean correct forecast direction*) é bastante associado a uma medida econômica já que ele é relacionado com o tempo (*timing*) de mercado e será utilizado. Nos resultados obtidos por meio das simulações realizadas, que serão apresentados neste trabalho, os impostos, os diferenciais de taxas de juros e os custos de transação são ignorados e assume-se que não se tem restrições de orçamento.

Um método estatístico de habilidade preditiva muito aplicado nas áreas de séries temporais não lineares é o NMSE (*normalized mean square error*) baseado no passeio aleatório. Na avaliação fora da amostra dos modelos, todas as tabelas, exceto as relacionadas com o NMSE, reportam o teste de White (2000), em que o P_{RC1} é o *bootstrap* p-valor para comparar individualmente cada modelo com o modelo diferença Martingale e o P_{RC2} é o p-valor do teste que compara os l modelos com o modelo diferença Martingale. A hipótese nula do P_{RC2} é que o melhor dos l modelos não tem superioridade preditiva sobre o modelo diferença Martingale. O último número de P_{RC2} checa se o melhor modelo dos modelos comparados tem previsibilidade superior sobre a diferença Martingale. A diferença entre cada P_{RC1} e o último P_{RC2} fornece uma idéia do *bias* na mineração de dados. Este teste possibilita a quantificação do efeito da busca cega de especificação do modelo, separando o espúrio do relevante.

A Tabela 6.31 apresenta os resultados do NMSE das previsões dos retornos da taxa cambial brasileira.

Método	NMSE	$BK^{(1)}$	$PT^{(2)}$	$DM^{(3)}$
Martingale	0,5841	-	2,9734	1,5600
ARMA-GARCH	0,5790	-	2,4240	2,3445
RBF PCA GAUSS	0,5860	0,1473	0,4095	0,0023
RBF PCA SPLINE	**0,3250**	**0,1563**	**1,3157**	**0,0064**
RBF ARIA	0,3869	0,0089	2,0729	2,8973
Combinado	0,5691	-	1,7720	1,7589

Tab. 6.31: Resultados do NMSE para as previsões da série temporal dos retornos da taxa cambial brasileira

(1) Teste apresentado em [BLAKE and KAPETANIOS, 2003];
(2) Teste apresentado em [PESARAN and TIMMERNANN, 1992];
(3) Teste apresentado em [DIEBOLD and MARIANO, 1995].

O modelo RBF PCA SPLINE foi o que apresentou as melhores previsões dos retornos da taxa cambial brasileira, batendo o passeio aleatório, inclusive com valor bem menor que a diferença Martingale. Os resultados do teste de BLAKE e KAPETANIOS (2003) indicam que as redes neurais não negligenciaram as não linearidades. Os valores dos testes de DIEBOLD e MARIANO (1995) e PESARAM e TIMMERMANN (1992) são menores que os valores críticos.

A determinação da defasagem (*lag*) e da dimensão de imersão (*embedding*) foi inicialmente realizada por meio da reconstrução dinâmica, e posteriormente o ajuste final foi feito via RNA. Os resultados obtidos na reconstrução dinâmica para o valor de L foi mais próximo do valor ajustado pela rede neural. Isto se deve ao fato desta defasagem ser estimada via método apresentado em [CELLUCCI et al., 2005] que leva em conta a distribuição dos *bins* no histograma. A capacidade atual de processamento dos recursos computacionais facilita o ajuste automático dos valores de L e M, restando para a reconstrução dinâmica o papel da análise da dinâmica não linear do processo.

As Tabelas 6.32 e 6.33 apresentam respectivamente os resultados do MSFE e do MAFE das previsões da série temporal dos retornos da taxa cambial brasileira.

Método	MSFE	P_{RC1}	P_{RC2}
Martingale	3,3814e-005	-	-
ARMA-GARCH	6,5494e-009	0,0160	0
RBF PCA GAUSS	0,3737e-004	0,7460	0
RBF PCA SPLINE	0,3374e-004	0,4900	0
RBF ARIA	0,2240e-004	0,0030	0
Combinado	0,3539e-004	0,5920	0

Tab. 6.32: Resultados do MSFE das previsões da série temporal dos retornos da taxa cambial brasileira

Método	MAFE	P_{RC1}	P_{RC2}
Martingale	0,0044	-	-
ARMA-GARCH	3,3635e-005	0,0110	0,0010
RBF PCA GAUSS	0,0044	0,7154	0,0010
RBF PCA SPLINE	0,0046	0,4210	0,0010
RBF ARIA	0,0035	0,0260	0,0010
Combinado	0,0044	0,4360	0,0010

Tab. 6.33: Resultados do MAFE das previsões da série temporal dos retornos da taxa cambial brasileira

O modelo ARMA-GARCH foi o que apresentou a melhor estatística MSFE e os valores de P_{RC1} e do último valor de P_{RC2} são significantes, indicando que os resultados de previsão são consistentes. O modelo ARMA-GARCH realmente foi superior neste critério estatístico e não ocorreu *bias* na mineração dos dados. O modelo ARMA-GARCH também foi o que apresentou a melhor estatística para o MAFE, os valores de P_{RC1} e do último valor de P_{RC2} são significantes, indicando que os resultados de previsão são consistentes.

Advoga-se que o sucesso dos modelos ARMA-GARCH nesta aplicação pode ter sido devido à magnitude das não linearidades, ou seja, os valores das amostras dos retornos são pequenos. Assim, mesmo tendo sido constatado a presença de não linearidades, os modelos ARMA-GARCH conseguiram captá-las.

A escolha da função de perda (NMSE e MSFE) já afetou os resultados das avaliações dos modelos. Os resultados sinalizam que foram superados os modelos passeio aleatório (NMSE) e a diferença Martingale na média (MSFE e MAFE).

As Tabelas 6.34 e 6.35 apresentam respectivamente os resultados do MFTR e MCFD das previsões da série temporal dos retornos da taxa cambial brasileira.

Método	MFTR	P_{RC1}	P_{RC2}
Martingale	2,8593e-005	-	-
ARMA-GARCH	4,1900e-004	0,1710	0,2380
RBF PCA GAUSS	6,4503e-004	0,0370	0,2210
RBF PCA SPLINE	-9,7304e-005	0,9220	0,3140
RBF ARIA	**9,4119e-004**	**0,0660**	**0,0940**
Combinado	4,1074e-004	0,1560	0,4620

Tab. 6.34: Resultados do MFTR das previsões da série temporal dos retornos da taxa cambial brasileira

Método	MCFD	P_{RC1}	P_{RC2}
Martingale	0,4750	-	-
ARMA-GARCH	0,5000	0,2300	0,2820
RBF PCA GAUSS	0,5250	0,1470	0,3980
RBF PCA SPLINE	0,4850	0,9350	0,5280
RBF ARIA	**0,5350**	**0,0890**	**0,5450**
Combinado	0,5100	0,1480	0,0960

Tab. 6.35: Resultados do MCFD das previsões da série temporal dos retornos da taxa cambial brasileira

O modelo RBF ARIA foi o que apresentou o melhor retorno (0,00094) de negociação. Não parece um resultado de negociação atraente. Entretanto, houve uma queda significativa da taxa de câmbio no período analisado e isto sinaliza a possibilidade de que se pôde neutralizar os prejuízos com o câmbio. Os valores do P_{RC1} e do último P_{RC2} indicam que não ocorreu *bias* na mineração dos dados.

O modelo RBF ARIA foi o que apresentou a melhor capacidade de acertar a direção (53.50 por cento) da próxima variação. O valor do P_{RC1} é significante e o valor do último P_{RC2} indica que não ocorreu um *bias* na mineração dos dados.

Os modelos RBF ARIA com a seleção de variáveis têm superioridade preditiva econômica em relação aos outros modelos ajustados. O modelo RBF ARIA com a seleção de variáveis também é superior ao passeio aleatório. Assim, foi provado empiricamente que a taxa de câmbio brasileira não é um passeio aleatório e nem uma diferença Martingale. Logo, tem algum nível de previsibilidade.

A Tabela 6.36 apresenta os resultados do NMSE das previsões da série temporal da umidade.

Método	NMSE	$BK^{(1)}$	$PT^{(2)}$	$DM^{(3)}$
Martingale	1,2798	-	0,9810	0,8756
ARMA-GARCH	511,58	-	0,0908	0,0743
RBF PCA GAUSS	0,4640	0,0015	3,8258	8,6588
RBF PCA SPLINE	**0,4385**	**0,0457**	**0,8146**	**1,5833**
RBF ARIA	0,8558	0,4980	1,5497	1,1890
Combinado	0,6091	-	1,6720	1,7853

Tab. 6.36: Resultados do NMSE das previsões da série temporal da umidade

(1) Teste apresentado em [BLAKE and KAPETANIOS, 2003];
(2) Teste apresentado em [PESARAN and TIMMERNANN, 1992];
(3) Teste apresentado em [DIEBOLD and MARIANO, 1995].

Os modelos baseados em redes neurais apresentaram melhores previsões que a diferença Martingale. Entretanto, o modelo RBF PCA SPLINE apresentou as melhores previsões. Os testes de BLAKE e KAPETANIOS (2003) indicam que as redes neurais não negligenciaram as não linearidades. Os valores dos testes de DIEBOLD e MARIANO (1995) e PESARAM e TIMMERMANN (1992) são menores que os valores críticos.

A determinação da defasagem (*lag*) e da dimensão de imersão (*embedding*) também foi inicialmente realizada por meio da reconstrução dinâmica, e o ajuste final foi feito via RNA. O resultado obtido na reconstrução dinâmica para o valor de L também foi mais próximo do valor ajustado pela rede neural. Isto também se deve ao fato desta defasagem ser estimada via método de CELLUCCI et al.(2005) que leva em conta a distribuição dos *bins* no histograma.

Foi constatado que a determinação da dimensão M de imersão (*embedding*) e da defasagem L (*lag*) por meio da reconstrução dinâmica, no contexto desta tese, foi de pouca valia. Os valores ajustados via redes neurais apresentaram melhor qualidade de previsão, rapidez no processamento das informações e custo computacional baixo.

As Tabelas 6.37 e 6.38 apresentam os resultados do MSFE e do MAFE das previsões da série temporal da umidade.

Método	MSFE	P_{RC1}	P_{RC2}
Martingale	228,4135	-	-
ARMA-GARCH	9,1561	0,9860	0,5080
RBF PCA GAUSS	0,0192	0,1660	0,7580
RBF PCA SPLINE	0,0257	0,7220	0,8760
RBF ARIA	0,0142	0,0026	0,9360
Combinado	0,0220	0,3840	0,9560

Tab. 6.37: Resultados do MSFE das previsões da série temporal da umidade

Método	MAFE	P_{RC1}	P_{RC2}
Martingale	12,6378	-	-
ARMA-GARCH	273,7130	0,9968	0,5340
RBF PCA GAUSS	9,9025	0,0101	0,7380
RBF PCA SPLINE	11,5095	0,0940	0,8700
RBF ARIA	9,7267	0,0030	0,9200
Combinado	10,5740	0,0040	0,9560

Tab. 6.38: Resultados do MAFE das previsões da série temporal da umidade

O modelo RBF ARIA foi o que apresentou a melhor estatística MSFE, o valor do P_{RC1} é significante. Entretanto, o valor do último P_{RC2} indica que pode ter ocorrido um *bias* na mineração dos dados.

O modelo RBF ARIA também foi o que apresentou a melhor estatística MAFE, o valor do P_{RC1} é significante. Entretanto, o valor do último P_{RC2} também indica que pode ter ocorrido um *bias* na mineração dos dados.

Entretanto, o teste apresentado em [WHITE, 2000] é muito conservador e sensível a modelos com previsões pobres em relação aos outros modelos comparados, que é o caso das previsões de baixa qualidade dos modelos ARMA-GARCH apontadas pelas estatísticas MSFE e MAFE. Pode ocorrer que o valor do P_{RC2} significante não estar refletindo os fatos, logo este teste é mais indicado para garantir que não ocorreu *bias* na mineração de dados do que para confirmá-lo. Especificamente, este teste foi construído como sendo um teste de hipótese nula simples, comparando somente dois modelos, enquanto sua aplicação geralmente é de fato conjunta, avaliando mais de dois modelos ao mesmo tempo. HANSEN (2005) detalha este e outros problemas, observando que este teste pode gerar p-valores inconsistentes e criou um teste alternativo.

A Tabela 6.39 apresenta os resultados do MCFD das previsões da série temporal da umidade. Já a utilização do MFTR para esta aplicação, que não é da área econômica, não faz muito sentido.

Método	MCFD	P_{RC1}	P_{RC2}
Martingale	0,6181	-	-
ARMA-GARCH	0,4673	0,9870	0,4900
RBF PCA GAUSS	0,4523	0,9702	0,7500
RBF PCA SPLINE	0,4070	0,9911	0,8840
RBF ARIA	0,6985	0,0000	0,2140
Combinado	0,4472	0,9867	0,0920

Tab. 6.39: Resultados do MCFD das previsões da série temporal da variação da umidade

O modelo RBF ARIA foi o que apresentou a melhor estatística, o valor do P_{RC1} é significante e o valor do último P_{RC2} indica que não ocorreu um *bias* na mineração dos dados.

No caso da umidade, os modelos neurais RBF PCA SPLINE para NMSE e RBF ARIA, com a seleção de variáveis, para MSFE, MAFE e MCFD, tem superioridade preditiva estatística e econômica em relação aos outros modelos ajustados. Entretanto, os testes de WHITE (2000) não garantem os resultados para o MSFE e MAFE. O modelo RBF PCA SPLINE é superior ao passeio aleatório e aos outros modelos. Assim, esta pesquisa empírica sinaliza que a série temporal da variação da umidade não é um passeio aleatório e nem uma diferença Martingale, logo tem algum nível de previsibilidade. Os resultados obtidos nas previsões das duas séries temporais com este tipo de modelo sinalizam que existem *clusters* de padrões temporais com capacidade preditiva, mas ambas têm características estocásticas.

Os modelos ARMA-GARCH não apresentaram boa qualidade de previsão para a umidade. Para os retornos da taxa de câmbio os resultados foram muito bons, e inclusive os testes de WHITE (2000) foram significantes para o MSFE e o MAFE. Advoga-se que pode ter sido a magnitude das não linearidades, ou seja, no caso dos retornos da taxa de câmbio, os valores das amostras são pequenos e, no caso da umidade, são grandes.

A escolha da função de perda afetou os resultados das avaliações dos modelos. No caso da função de perda NMSE, foi fácil superar o modelo passeio aleatório. As duas séries temporais são de difícil previsão, mas os resultados comprovam que estas séries não são processos de diferença Martingale, tendo, portanto, algum nível de previsibilidade.

Os modelos neurais tipo RBF ARIA com entradas determinadas por meio do método de seleção de variáveis proposto nesta tese, no geral, foram os que apresentaram melhores resultados de previsão para critérios econômicos. Os modelos ARMA-GARCH apresentaram boa qualidade de previsão para os retornos da taxa de câmbio brasileira nos critérios estatísticos de precisão. A determinação da dimensão de *embedding* e da defasagem (*lag*) por meio da reconstrução dinâmica, no contexto desta tese, foi de pouca valia. Os resultados obtidos mais próximos dos ajustados pelas redes neurais foram as defasagens estimadas via método de CELLUCCI et al.(2005).

Capítulo 7

Conclusões e Trabalhos Futuros

7.1 Conclusões

7.1.1 Conclusões principais

- A hipótese geral de que a variação da taxa de câmbio brasileira e da umidade no microclima da região de Londrina-PR não são processos tipo Martingale foi confirmada empiricamente, mas estas duas séries temporais, todavia, são de difícil previsão, principalmente porque as informações disponíveis sobre as séries temporais, que possivelmente seriam as variáveis de entrada dos modelos, são incompletas. Por exemplo, neste trabalho, não conseguimos montar uma série temporal para expressar as informações geradas pelos *hedge funds* e nem as informações sobre as reuniões diárias entre o Banco Central do Brasil e os *dealers*, que por força da lei, não são disponibilizadas ao público.

- O método de seleção de variáveis proposto neste trabalho para modelos que representam sistemas complexos de forma funcional desconhecida e sem conhecimento *a priori* sobre as distribuições das variáveis aleatórias contribuiu para a qualidade das previsões. Este método ajuda a lidar com a busca de informações para possibilitar uma melhor modelagem deste tipo de processo.

- O método de previsão proposto neste trabalho, baseado em redes neurais tipo RBF spline fina com fator adaptativo da variância, demonstrou capacidade de aprendizagem principalmente para prever a direção das variações ocorridas nestes processos. O baseado em redes neurais tipo RBF com os centros ajustados via algoritmo ARIA também demonstrou capacidade de aprendizagem.

- **Limitações deste trabalho**: o fato de lidar com informações incompletas e com amostras que podem estar carregadas de ruídos dificulta e limita a qualidade das previsões. O teste de WHITE (2000), como já mencionado, pode apresentar algumas distorções. Já o artigo que poderia possibilitar a implementação do teste de HANSEN (2005) e evitar as distorções do teste WHITE (2000) não está disponível ao público no momento. Em relação às limitações do método de reconstrução dinâmica, observa-se que este método tem

apresentado melhores resultados com processos carregados de inércia e não apresentou resultados expressivos para processos ligados aos sistemas financeiros e econômicos.

- **Aplicabilidade do trabalho**: apesar das limitações apresentadas acima, esta metodologia pode ajudar a identificar mais informações e fazer a transferência das mesmas para o modelo que representa o processo, incrementando a qualidade das previsões. Por exemplo, em determinadas aplicações, um incremento de ganho percentual pequeno (0,5 por cento) devido ao incremento da qualidade das previsões pode implicar em um montante expressivo de ganhos no mercado financeiro (milhões de dólares). O mesmo se aplica na economia de produtos químicos utilizados na cultura da soja.

7.1.2 Conclusões secundárias

- Os gráficos da autocorrelação e autocorrelação parcial dos retornos da taxa cambial brasileira e dos dados brutos da umidade indicaram dependência temporal linear. As duas séries apresentaram não linearidades (teste de HSIEH) e dependência temporal via teste BDS. Dessa forma, estas informações (FAC, FACP e Hsieh) sinalizam que as duas séries possivelmente não são processos Martingale. É razoável que se utilize métodos de previsão não lineares como complementares ou substitutos dos métodos lineares. Entretanto, a detecção de não linearidades na amostra não significa que as previsões fora dela tenham que ser realizadas por modelos não lineares. Estas não linearidades podem estar somente nas amostras de ajuste. As não linearidades também podem não ser fortes o suficiente para exigir um modelo não linear que é difícil de se lidar. As não linearidades também podem ser exógenas, derivando de *outliers*, mudanças estruturais, intervenções do governo, que podem ser captadas pelos testes, mas que não contribuem para previsões fora da amostra.

- A determinação da defasagem (*lag*) e da dimensão de imersão (*embedding*) por meio da reconstrução dinâmica, no contexto desta tese, foi de pouca valia para determinar a janela de previsão, já que o esforço computacional para ajustá-los de forma automática é relativamente baixo. Os resultados obtidos mais próximos daqueles ajustados pelas redes neurais foram as defasagens estimadas via método de CELLUCCI et al. (2005), que leva em conta a distribuição dos *bins* no histograma. A dimensão de imersão dos retornos da taxa de câmbio e da umidade estimados via RNA são elevados. A dimensão da correlação das séries é fracionária e com valores médios. Os maiores coeficientes de LYAPUNOV de cada série, embora muito pequenos, são positivos. Isto indica sensibilidade às condições iniciais, mesmo que não muito elevada. Estes resultados, em conjunto, indicam que as séries, por definição, são caóticas, já que têm pelo menos um expoente de LYAPUNOV positivo. Entretanto é mais apropriado interpretar estes dados como um indicador de origem estocástica da série.

- Os modelos ARMA-GARCH não apresentaram boa qualidade de previsão para a umidade. Para os retornos da taxa de câmbio, os resultados foram muito bons, tendo, inclusive os testes de White (2000) sido não significantes para o MSFE e para o MAFE. Advoga-se que pode ter sido a magnitude das não linearidades, ou seja, no caso dos retornos da taxa

de câmbio, os valores das amostras são pequenos e, no caso da umidade, são grandes. Será observado ao longo destas conclusões que a escolha da função de perda afeta os resultados das avaliações dos modelos. Ou seja, a escolha do melhor modelo depende da função de perda que é utilizada.

- O modelo RBF ARIA GAUSS, tendo como variáveis de entrada a ptax, o euro e a diferença entre as taxas internas e a *fed funds*, foi o que apresentou os melhores resultados para os testes econômicos, indicando que estas são as variáveis de entrada escolhidas no *wrapper* da série dos retornos da taxa cambial brasileira. O modelo RBF ARIA GAUSS, tendo como variáveis de entrada a umidade, a temperatura mínima e temperatura média, foi o que apresentou os melhores resultados, indicando que estas são as variáveis de entrada escolhidas no *wrapper* da série temporal da umidade. Portanto, o método de seleção de variáveis proposto (filtro e *wrapper*) mostrou-se operacional e útil. As duas séries temporais apresentaram *clusters* com capacidade preditiva, mas ambas têm características estocásticas.

- Os modelos RBF PCA SPLINE com fator adaptativo e regularização apresentaram os melhores resultados de previsão para o critério do NMSE, sendo superior ao passeio aleatório, sugerindo que o fator adaptativo para a função spline proposto nesta tese e a regularização concorreram para a qualidade das previsões.

- Os resultados do teste de BLAKE e KAPETANIOS (2003) indicam que as redes neurais não negligenciaram as não linearidades. Os valores dos testes de DIEBOLD e MARIANO (1995) e PESARAM e TIMMERMANN (1992) foram úteis para verificar se as estatísticas do critério NMSE são significantes.

- O teste de WHITE (2000) mostrou-se muito conservador e sensível a modelos com previsões pobres em relação aos outros modelos comparados, mas no geral foi bastante útil. Isto ocorreu com os modelos ajustados para a previsão da umidade. Especificamente, este teste foi construído como sendo uma simples hipótese nula enquanto de fato é composta. HANSEN (2005) detalha este e outros problemas, observando que o teste de WHITE (2000) pode gerar p-valores inconsistentes e criou um teste alternativo para lidar com estas limitações.

7.2 Trabalhos futuros

Os esforços serão direcionados para realizar as seguintes pesquisas:

- Estimar a informação mútua conjunta entre N diferentes variáveis aleatórias.

- Implementar testes de sensibilidade dos índices.

- Realizar previsões com múltiplos passos que geralmente apresentam expressiva utilidade prática.

- Implementar métodos que combinam árvores de regressão e redes RBF. O problema de balancear o *bias* e a variância, comum nos métodos de regressão não paramétrica, já pode ser automaticamente resolvido.

- Aplicar esta metodologia na previsão de outras séries temporais relevantes para a sociedade.

- Investigar as microestruturas e os dados de alta freqüência do mercado financeiro brasileiro, no qual os componentes estocásticos são dominantes.

- Implementar o método de HANSEN (2005) e comparar os resultados obtidos com este teste com aqueles fornecidos pelo teste de WHITE (2000).

Referências Bibliográficas

[ABARBANEL, 1993] ABARBANEL, H. D. I. (1993). The analysis of observed chaotic data in physical systems. *Reviews of Modern Physics*, 4:1331–1392.

[AKAIKE, 1974] AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, 19:716–723.

[ALEXANDER, 2005] ALEXANDER, C. (2005). *Modelos de Mercado*. Editora Saraiva.

[BACHELIER, 1900] BACHELIER, L. (1900). Théorie de la speculation. *Annales de Lécole Normale Supérieure*, pages 21–86.

[BENETTIN et al., 1980] BENETTIN, G., GALGANI, L., GIORGILLI, A., and STRELCYN, J. M. (1980). Lyapunov characteristic exponents for smooth dynamical systems and for hamiltonian systems a method for computing all of them. *Meccanica - Part II: Numerical application*, 15:21–30.

[BEZERRA et al., 2005] BEZERRA, G., BARRA, T., CASTRO, L. N., and VONZUBEN, F. J. (2005). Adaptive radius immune algorithm for data clustering. *C. Jacob et al. Eds. ICARIS 2005*, page 290.

[BILLINGS and VOON, 1983] BILLINGS, S. A. and VOON, W. S. F. (1983). Structure detection and model validity tests in the identification of nonlinear systems. *IEEE Proceedings*, 130(4):193–1999.

[BILLINGS and VOON, 1986] BILLINGS, S. A. and VOON, W. S. F. (1986). Correlation based model validity tests for nonlinear models. *International Journal of Control*, 44(1):235–244.

[BISHOP, 1994] BISHOP, C. M. (1994). Mixture density networks. Technical report, Neural Computing Reserch Group - Dept. of Computer Science and applied Mathematics - Aston University.

[BLACK and SCHOLES, 1973] BLACK, F. and SCHOLES, M. (1973). The pricing of options and corporates liabilities. *Journal of Political Economy*, 81:637–659.

[BLAKE and KAPETANIOS, 2003] BLAKE, A. P. and KAPETANIOS, G. (2003). A radial basis function neural network test for neglected nonlinearity. *Econometrics Journal*, 6:357–373.

[BLUM and LANGLEY, 1997] BLUM, A. and LANGLEY, P. (1997). Selection of relevant features and examples of machine learning. *Artificial Intelligence*, 97(1-2):245–271.

[BOLLERSLEV, 1986] BOLLERSLEV, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327.

[BORS, 2004] BORS, A. G. (2004). Introduction of the radial basis function (rbf) networks. Working paper, Department of Computer Science - University of York.

[BOX et al., 1994] BOX, G. E. P., JENKINS, J. E., and REINSEL, G. C. (1994). *Times Series Analysis - Forecasting and Control*. Holden Day.

[BOX and JENKINS, 1970] BOX, G. E. P. and JENKINS, J. E. (1970). *Times Series Analysis - Forecasting and Control*. Holden Day.

[BOX and JENKINS, 1976] BOX, G. E. P. and JENKINS, J. E. (1976). *Times Series Analysis - Forecasting and Control*. Holden Day.

[BROCKWELL and DAVIS, 1991] BROCKWELL, P. J. and DAVIS, R. A. (1991). *Times series - theory and methods*. Springer-Verlag.

[BROCK et al., 1996] BROCK, W. A., DECHERT, W. D., and Scheinkman, J. A. (1996). A test for independence based on the correlation dimension. *Econometric Reviews*, 15(197):235.

[BROOMHEAD and KING, 1985] BROOMHEAD, D. S. and KING, G. P. (1985). Extracting qualitative dynamics from experimental data. *Physica 20D*, pages 217–236.

[BROOMHEAD and LOWE, 1988] BROOMHEAD, D. S. and LOWE, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355.

[CARNEIRO and NETTO, 2005] CARNEIRO, A. A. C. and NETTO, M. L. A. (2005). A nonlinear methodology to predict brazilian exchange rates by means of macro and microstructures variables via neural network. *ENANPAD*.

[CARNEIRO et al., 2004] CARNEIRO, A. A. C., SECURATO, J. R., and NETTO, M. L. A. (2004). Non-linear ponder autoregressive neural network methodology to predict brazilian exchange rates. *IV Encontro Brasileiro de Finanças, Brazil, SBFIN*.

[CASTRO and ZUBEN, 2001] CASTRO, L. N. and ZUBEN, F. J. V. (2001). An immunological approach to initialize centers of radial basis function neural networks. *Proceedings of V Brazilian Conference on Neural Networks - V Congresso Brasileiro de Redes Neurais*, pages 79–84.

[CELLUCCI et al., 2003] CELLUCCI, C. J., ALBANO, A. M., and RAPP, P. E. (2003). Comparative study of embedding methods. *Phys. Rev. E*, 67:1063.

[CELLUCCI et al., 2005] CELLUCCI, C. J., ALBANO, A. M., and RAPP, P. E. (2005). Statistical validation of mutual information calculations: Comparison of alternative numerical algorithms. *Phys. Rev. E*, 71:1539–3755.

[CHAO et al., 2000] CHAO, J., CORRADI, V., and SWANSON, N. (2000). An out of sample test for granger causality. *Journal of non-linear dynamics*.

[CUNHA, 1997] CUNHA, M. S. (1997). Causalidade em séries temporais. Tese de mestrado, IME, USP.

[DECASTRO, 2000] DECASTRO, M. C. F. (2000). Previsão de séries temporais via redes neurais tipo rbf. Tese de doutorado, Faculdade de Engenharia Elátrica e Computação, UNICAMP.

[DEMUTH and BEALE, 1998] DEMUTH, H. and BEALE, M. (1998). *Neural Network Toolbox*. The Mathworks Inc.

[DICKEY and FULLER, 1979] DICKEY, D. A. and FULLER, W. (1979). Distribution of the estimators for times series regressions with a unit root. *Journal of the American Statistical Association*, 74:427–431.

[DICKEY and PANTULA, 1987] DICKEY, D. A. and PANTULA, S. S. (1987). Determining the order of differencing in autoregressive processes. *Journal of Business and Statistics*, 5:455–461.

[DIEBOLD and MARIANO, 1995] DIEBOLD, F. X. and MARIANO, R. S. (1995). Comparing predictive accuracy. *Business Economic Estat.*, 13:253–263.

[DORNBUSCH and FRANKEL, 1988] DORNBUSCH, R. and FRANKEL, J. (1988). *The flexible exchange rate system: experience and alternatives. In: Borner S. Ed. International trade and finance in a polycentric world*. New York: St. Martin Press.

[ECKMANN and RUELLE, 1992] ECKMANN, J. and RUELLE, D. (1992). Fundamental limitations for estimating dimensions and lyapunov exponents in dynamical systems. *Physica D*, 56:185–187.

[ENDERS, 2004] ENDERS, W. (2004). *Applied Econometric Time Series*. John Wiley and Sons, USA.

[ENGLE and GRANGER, 1987] ENGLE, R. F. and GRANGER, C. W. J. (1987). Co-integration and error correction: Representation, estimation and testing. *Econometrica*, 55:251–276.

[ENGLE et al., 1987] ENGLE, R., LILIEN, D., and ROBBINS, R. (1987). Estimating time-varying risk premia in the term structure: the arch-m model. *Econometrica*, 55:251–276.

[ENGLE, 1982] ENGLE, R. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of united kingdon inflation. *Econometrica*, 50(4):987–1007.

[EVANS and LYONS, 2002] EVANS, M. and LYONS, R. (2002). Order flow and exchange rate dynamics. *Journal of International Economics*, 110(1):170–180.

[FRANKEL and FROOT, 1990] FRANKEL, J. and FROOT, K. (1990). *Chartists fundamentalists and the demand for dollars.* Oxford University Press.

[GLOSTEN et al., 1993] GLOSTEN, L. R., JAGANNATHAN, R., and RUNKLE, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48:1779–1801.

[GOLUB et al., 1979] GOLUB, G. H., HEATH, M., and WAHBA, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

[GRANGER and NEWBOLD, 1974] GRANGER, C. W. J. and NEWBOLD, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, 2:111–120.

[GRANGER, 1969] GRANGER, C. W. J. (1969). Investigating causal relations by econometric models and cross spectral methods. *Econometrica*, 37:424.

[GRASSBERGER and PROCACCIA, 1983a] GRASSBERGER, P. and PROCACCIA, I. (1983a). Characterization of strange attractors. *Phys. Rev. Lett.*, 50:346–349.

[GRASSBERGER and PROCACCIA, 1983b] GRASSBERGER, P. and PROCACCIA, I. (1983b). Measuring the strangeness of strange attractors. *Physica D*, 9:189–208.

[GROSSBERG, 1982] GROSSBERG, S. (1982). *Studies of mind and brain.* Reidel Press.

[GUIMARÃES and TABAK, 2004] GUIMARÃES, A. M. F. and TABAK, B. M. (2004). Testando o conteúdo informacional em variáveis de microestrutura de mercado para a taxa de câmbio. In Sbfin, editor, *IV Encontro Brasileiro de Finanças*, Rio de Janeiro-Brazil.

[GUYON and ELISSEEFF, 2003] GUYON, I. and ELISSEEFF, A. (2003). An introduction to variable and feature selection. *Journal of machine learning*, 3:1157–1182.

[HAMILTON, 1994] HAMILTON, J. D. (1994). *Time series analysis.* Princeton University Press.

[HANSEN, 2005] HANSEN, P. R. (2005). A test for superior predictive ability. Working paper, Stanford University - apud HONG, H. and LEE, T. (2002).

[HARTLEY, 1928] HARTLEY, R. V. (1928). Transmission of information. *Bell Sys. Tech. J.*, 7.

[HARVEY et al., 1994] HARVEY, A. C., RUIZ, E., and SHEPHARD, N. (1994). Multivariate stochastic variance models. *Review of economic studies*, (47):247–264.

[HARVEY, 1992] HARVEY, A. C. (1992). *Forecasting structural time series model and Kalman filter.* Cambridge university press.

[HAYES, 1996] HAYES, M. H. (1996). *Statistical digital signal processing and modeling.* John Wiley and Sons.

[HAYKIN, 1989] HAYKIN, S. (1989). *Modern filters*. Macmillan publishing company.

[HAYKIN, 1999] HAYKIN, S. (1999). *Neural Networks*. Prentice Hall.

[HONG and LEE, 2002] HONG, H. and LEE, T. (2002). Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. *Journal of Nonlinear Dynamics*.

[HSIEH and KLEIDON, 1996] HSIEH, D. A. and KLEIDON, A. W. (1996). Bid-ask spreads in foreign exchange markets: implications for models of asymmetric information. In Frankel and eds., J., editors, *The microstructure of foreign exchange markets*, pages 261–294. University of Chicago Press.

[HSIEH, 1988] HSIEH, D. A. (1988). The statistical properties of daily foreign exchange rates: 1974 to 1983. *Journal of International Economics*, 24:129–145.

[HSIEH, 1989] HSIEH, D. A. (1989). Testing for nonlinear dependence in daily foreign exchange rates. *Journal of Business*, 62(3):339–368.

[HSIEH, 1993] HSIEH, D. A. (1993). Implications of nonlinear dynamics for financial risk management. *Journal of Business*, 28:41–64.

[JOHANSEN and JUSELIUS, 1990] JOHANSEN, S. and JUSELIUS, K. (1990). Maximum likehood estimation and inference on co-integration with applications to the demand for money. *Oxford Bulletin of Economics and Statistics*, 52:169–210.

[KAMINSKY et al., 1998] KAMINSKY, G. L., LIZONDO, C., and LEIDERMAN, L. (1998). Leading indicators of currency crises. 45 1, International Monetary Fund.

[KAMINSKY and REINHART, 1996] KAMINSKY, G. L. and REINHART, C. M. (1996). The twin crises: The causes of banking and balance-of-payments problems. Technical Report 544, Board of Governors of the Federal Reserve System. International Finance Discussion Papers.

[KOHAVI et al., 1994] KOHAVI, R., JOHN, G., and PFLEGER, K. (1994). Irrelevant feature and the subset selection problem. *Conference on machine learning*, pages 121–129.

[KOHAVI and JOHN, 1997] KOHAVI, R. and JOHN, G. (1997). Wrappers for feature selection. *Artificial Intelligence*, 97(1-2):273–324.

[KOHONEN, 1988] KOHONEN, T. (1988). *The self-organization and associative memory*. Springer-Verlag.

[KOLLER and SAHAMI, 1996] KOLLER, D. and SAHAMI, M. (1996). Toward optimal feature selection. *Conference on machine learning*, pages 284–292.

[KOLMOGOROV, 1957] KOLMOGOROV, A. N. (1957). On the representation of continuous functions. Technical Report 114, Dokl. Acad. Naut. USSR.

[KOVÁCS, 2002] KOVÁCS, Z. L. (2002). *Redes neurais artificiais*. Editora da Física.

[KWIATKOWSKI et al., 1992] KWIATKOWSKI, D., PHILLIPS, P., SCHMIDT, P., and SHIN, Y. (1992). Testing the null of stationarity against the alternative of a unit root. *Journal of Econometrics*, (542):159–178.

[LAPEDES and FARBER, 1986] LAPEDES, A. and FARBER, R. (1986). Programming a massively parallel, computation universal system: static behavior. A, In neural networks for computing. New York - American institute of physics.

[LEBARON, 1994] LEBARON, B. (1994). Chaos and nonlinear forecastability in economics and finance. *Philosophical Transactions of the Royal Society of London*, 348:397–404.

[LEE et al., 1993] LEE, T. H., GRANGER, C., and WHITE, H. (1993). Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests. *Journal of Econometrics*, 56:269–290.

[LORENZ, 1963] LORENZ, N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20:130–141.

[MANDELBROT, 1963] MANDELBROT, B. (1963). The variation of certain speculative prices. *Journal of Business*, 36:394–416.

[MARKOWITZ, 1959] MARKOWITZ, H. M. (1959). *Portfolio Selection: Efficient Diversification of Investments*. John Wiley and Sons eds.

[MEESE and ROGOFF, 1983] MEESE, R. and ROGOFF, K. (1983). Empirical exchange rate models of the seventies. *Journal of International Economics*, 14:3–24.

[MICCHELLI, 1986] MICCHELLI, C. A. (1986). Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2(11):22.

[MOODY and DARKEN, 1991] MOODY, J. and DARKEN, C. (1991). Note on learning rate schedules for stochastic optimization. In Moody, L. and eds., T., editors, *Managing QoS in Multimedia Networks and Services*. Morgan Kaufmann.

[MORETTIN and TOLOI, 2004] MORETTIN, P. A. and TOLOI, C. M. (2004). *Séries temporais*. Atual editora.

[NELSON and PLOSSER, 1982] NELSON, C. I. and PLOSSER, C. I. (1982). Trends and random walks in macroeconomic time series: Some evidence and implications. *Journal of Monetary Economics*, 10:139–162.

[NELSON, 1991] NELSON, D. (1991). Conditional heteroskedasticity in assets returns: new approach. *Econometrica*, 59(2):347–370.

[NEWEY and WEST, 1987] NEWEY, W. and WEST, K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55:703–708.

[NIYOGI and GIROSI, 1996] NIYOGI, P. and GIROSI, F. (1996). On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions. *Neural Computation*, 8:819–842.

[ORR, 1996] ORR, M. J. L. (1996). Introduction to radial basis functions networks. Technical report, University of Edinburgh.

[ORR, 1999] ORR, M. J. L. (1999). Recent advances in radial basis functions networks. Technical report, University of Edinburgh.

[OZAKI, 1980] OZAKI, T. (1980). Non-linear time series models for non-linear random vibration. *Journal of applied probability*, 17:84–93.

[PACKARD et al., 1980] PACKARD, N., CRUTCHFIELD, J., FARMER, J., and SHAW, R. (1980). Geometry from a time series. *Phys. Rev. Lett.*, 45:712–716.

[PARZEN, 1962] PARZEN, E. (1962). Estimation of probability density function and mode. *Annals of mathematical statistics*, 33:1065–1076.

[PESARAN and TIMMERNANN, 1992] PESARAN, M. H. and TIMMERNANN, A. (1992). A simple non-parametric test of predictive performance. *Journal of Business and Economic Statistics*, 10:461–465.

[PHILLIPS and PERRON, 1987] PHILLIPS, P. and PERRON, P. (1987). Testing forwith a unit root im time series regressions. *Biometrika*, 75:335–346.

[PHILLIPS and PERRON, 1988] PHILLIPS, P. and PERRON, P. (1988). Testing for a unit root in time series regression. *Biometrica*, (75):335–346.

[POGGIO and GIROSI, 1990a] POGGIO, T. and GIROSI, F. (1990a). Networks and the best approximation property. Technical Report 63, Biological Cybernetics.

[POGGIO and GIROSI, 1990b] POGGIO, T. and GIROSI, F. (1990b). Networks for approximation and learning. In IEEE, editor, *Proceedings of the IEEE*, pages 1481–1497.

[POINCARÉ, 1952] POINCARÉ, H. (1952). *Science and Method*. Dover Press - Originally published in 1908.

[POLITIS and ROMANO, 1994] POLITIS, D. N. and ROMANO, J. P. (1994). The stationary bootstrap. *J. Amer. Statist. Assoc.*, 89:1303.

[POVINELLI, 1999] POVINELLI, J. (1999). Times series data mining: identifying temporal patterns for characterization and prediction of times series events (phd). *Marquette University, Milwakee, Wisconsin*.

[POWELL, 1985] POWELL, M. (1985). Radial basis functions for multivariable interpolation: review. *IMA Conference on algorithms for approximation of functions and data*, 2:143–167.

[PRIESTLY, 1989] PRIESTLY, M. B. (1989). *Nonlinear and Non Stationary Time Series Analysis*. London Academic.

[PRÍNCIPE, 1998] PRÍNCIPE, J. C. (1998). *Information-theoretic learning*. John Wiley and Sons, New York-USA.

[REFENES et al., 1993] REFENES, A. N., AZEMA-BARAC, M., CHEN, L., and KAROUSSOS, S. A. (1993). Currency exchange rate prediction and neural network design strategies. *Neural Computing and Applications*, 1(1):46–58.

[RENYI, 1976] RENYI, A. (1976). Selected papers of alfred renyi. Technical Report 2, Akademia Kiado, Budapest.

[SCHREIBER, 1998] SCHREIBER, T. (1998). Interdisciplinary application of nonlinear time series methods. Technical Report 42097, Wuppertal, Germany.

[SHANNON, 1948] SHANNON, C. E. (1948). A mathematical theory of communication. J 27, Bell Sys. Tech.

[SHARPE, 1963] SHARPE, W. E. (1963). A simplired model of portfolio analysis. *Management Science*, 9:277–293.

[SHARPE, 1964] SHARPE, W. E. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19:425–442.

[SHINTANI and LINTON, 2004] SHINTANI, M. and LINTON, O. (2004). Is there chaos in the world economy? a nonparametric test using consistent standard errors. *International Economic Review*, 44:331–358.

[SIMS et al., 1990] SIMS, C. A., J., S., and W., W. M. (1990). Inference in linear time series models with some units roots. *Econometrica*, 58:113–144.

[SIMS, 1980] SIMS, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48:1–48.

[SMALL and TSE, 2003] SMALL, M. and TSE, C. K. (2003). Determinism in financial times series. studies in nonlinear dynamics and econometrics. *Journal of International Economics*, 7(3).

[SMITH, 1988] SMITH, L. A. (1988). Intrinsic limits on dimension calculations. *Physical Let. A*, 133(6):283–288.

[SUBBA and GABR, 1984] SUBBA, R. and GABR, M. (1984). *An introduction to bispectral analysis and bilinear time series models*. Springer-Verlag.

[SUN and OORT, 1995] SUN, D. and OORT, A. H. (1995). Humidity-temperature relationships in the tropical troposphere. *Journal of Climate*.

[TABAK, 2002] TABAK, B. M. (2002). Testing the random walk hypothesis for emerging markets exchange rates. Technical report, Banco Central do Brasil.

[TAKENS, 1981] TAKENS, F. (1981). Detecting strange attractors in turbulence. *Lect. Notes Math.*, 898:366–381.

[TARAMASCO and ISABELLE, 1997] TARAMASCO, O. and ISABELLE, J. P. (1997). Les rentabilités à la bourse de paris sont-elles chaotiques? *Revue Economique*, (2):215–238.

[TERÄSVIRTA and GRANGER, 1993] TERÄSVIRTA, C. F. L. and GRANGER, C. W. J. (1993). Power of the neutral network linearity test comparing predictive accuracy. *Journal Time Series Anal*, 14(2):309–323.

[THEILER, 1986] THEILER, J. (1986). Spurius dimension from correlation algorithms applied to limited time-series data. *Physical Rev. A*, 34(3):2427–2432.

[THEILER, 1987] THEILER, J. (1987). Efficient algorithm for estimating the correlation dimension from a set of discrete points. *Physical Rev. A*, 36(9):4456–4462.

[THEILER, 1990] THEILER, J. (1990). Estimating fractal dimension. *J. Opt. Soc. Am.*, A(7):1055.

[TONG and LIM, 1980] TONG, H. and LIM, K. (1980). *Threshold autoregression limit cycles and cyclical data - with discussion.* New York - Oxford Univ. Press.

[UNCU and TÜRKSEN, 2006] UNCU, O. and TÜRKSEN, I. B. (2006). A novel feature selection approach: combining feature wrappers and filters. *Information Sciences (submitted).*

[UTANS and MOODY, 1991] UTANS, J. and MOODY, J. (1991). Selecting neural network architectures via the prediction risk - application to corporate bond rating prediction. In Press, I. C. S., editor, *in Proceedings of the First International Conference on Artificial Intelligence Applications on Wall Street*, pages 1–20, Los Alamitos CA.

[VANDROVYCH, 2005] VANDROVYCH, V. (2005). Study of nonlinearities in the dynamics of exchange rates: Is there any evidence of chaos? *Preliminary draft: June 11 2005.*

[WEIGEND et al., 1990] WEIGEND, A. S., RUMELHART, D. E., and HUBERMAN, B. A. (1990). Predicting the future - a connectionist approach. *International Journal of Neural System*, 1:193–209.

[WEIGEND and GERSHENFELD, 1994] WEIGEND, A. and GERSHENFELD, N. (1994). *Time series prediction: forecasting the future and understanding the past.* Addison Wesley.

[WHITE, 2000] WHITE, H. (2000). A reality check for data snooping. *Econometrica*, 68(5):1097–1126.

[WIDROW and HOFF, 1960] WIDROW, B. and HOFF, M. E. (1960). Adaptive switching circuits. *IRE WESCON Convention Record*, (96-104).

[WIDROW and STERNS, 1985] WIDROW, B. and STERNS, S. D. (1985). *Adaptive signal processing).* Prentice Hall.

[WIDROW, 1976] WIDROW, B. (1976). Stationary and nonstationary learning characteristics of the lms adaptive filter. *IEEE Proceedings*, 64(1151-1162).

[WIENER, 1949] WIENER, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. New York: Wiley.

[WOLF et al., 1985] WOLF, A., SWIFT, J., SWINNEY, H., and VASTANO, J. (1985). Determining lyapunov exponents from a time series. *Physica D*, 16:285–317.

[XU and PRÍNCIPE, 1998] XU, D. and PRÍNCIPE, J. C. (1998). A novel measure for independent component analysis (ica). *IEEE international conference on acoustics. Speech and signal processing*, 2:1145–1148.

[YULE, 1926] YULE, G. U. (1926). Why do we sometimes get nonsense correlations between time series. *Journal of Royal Statistical Society*, 89:2–9.

[YU and LIU, 2004] YU, L. and LIU, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, (5):1205–1224.

[ZAKOIAN, 1990] ZAKOIAN, J. M. (1990). Threshold heteroskedasticity models. *CREST-INSEE*.