



**Universidade do Estado do Rio de Janeiro**

Centro de Tecnologia e Ciências

Instituto de Física Armando Dias Tavares

Paulo Ricardo Lourenço Alves

**A aproximação global no problema inverso das séries temporais**

Rio de Janeiro

2017

Paulo Ricardo Lourenço Alves

**A aproximação global no problema inverso das séries temporais**



Tese apresentada, como requisito parcial para obtenção do título de Doutor, ao Programa de Pós-Graduação em Física, da Universidade do Estado do Rio de Janeiro.

Orientador: Prof. Dr. Luis Antonio Campinho Pereira da Mota

Coorientador: Prof. Dr. Luiz Guilherme Silva Duarte

Rio de Janeiro

2017

CATALOGAÇÃO NA FONTE  
UERJ/ REDE SIRIUS /CTC/D

A474 Alves, Paulo Ricardo Lourenço.  
A aproximação global no problema inverso das séries  
temporais / Paulo Ricardo Lourenço Alves. - 2017  
124 f.: il.

Orientador: Luis Antonio Campinho Pereira da Mota.  
Coorientador: Luiz Guilherme Silva Duarte.  
Tese (doutorado) - Universidade do Estado do Rio de  
Janeiro, Instituto de Física Armando Dias Tavares.

1. Análise de séries temporais - Teses. 2. Física  
matemática - Teses. 3. Comportamento caótico nos sistemas -  
Teses. I. Mota, Luis Antonio Campinho Pereira da. II. Duarte,  
Luiz Guilherme Silva. III. Universidade do Estado do Rio de  
Janeiro. Instituto de Física Armando Dias Tavares. IV. Título.

CDU 519.6

Autorizo, apenas para fins acadêmicos e científicos, a reprodução total ou parcial desta tese desde que citada a fonte.

---

Assinatura

---

Data

Paulo Ricardo Lourenço Alves

## A aproximação global no problema inverso das séries temporais

Tese apresentada, como requisito parcial para obtenção do título de Doutor, ao Programa de Pós-Graduação em Física, da Universidade do Estado do Rio de Janeiro.

Aprovada em 24 de abril de 2017.

Banca Examinadora:

---

Prof. Dr. Luis Antonio Campinho Pereira da Mota (Orientador)  
Instituto de Física Armando Dias Tavares – UERJ

---

Prof. Dr. Luiz Guilherme Silva Duarte (Coorientador)  
Instituto de Física Armando Dias Tavares – UERJ

---

Prof. Dr. Henrique Pereira de Oliveira  
Instituto de Física Armando Dias Tavares – UERJ

---

Prof. Dr. Rafael Fernandes Aranha  
Instituto de Física Armando Dias Tavares – UERJ

---

Prof. Dr. Andrés Reinaldo Rodriguez Papa  
Observatório Nacional

---

Prof. Dr. Caio Henrique Lewenkopf  
Universidade Federal Fluminense

Rio de Janeiro

2017

## AGRADECIMENTOS

Aos professores Luis da Mota e Luiz Duarte, pela construtiva orientação recebida.

Aos docentes da Pós-Graduação, pela consistente formação proporcionada.

Aos meus colegas do programa de Pós-Graduação, pelo companheirismo.

Ao físico Sérgio Eduardo Silva Duarte, pelos diálogos científicos que despertaram meu interesse pelo estudo das séries temporais.

Ao secretário Rogerio Teixeira, pelo zelo profissional.

A minha esposa Marcia, pelo apoio recebido.

A solução de um problema inverso consiste em determinar causas com base na  
observação dos seus efeitos.

*O. M. Alifanov*

## RESUMO

ALVES, P.R.L. *A aproximação global no problema inverso das séries temporais*. 2017. 122 f. Tese (Doutorado em Física) – Instituto de Física Armando Dias Tavares, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2017.

O método da aproximação global é o eixo de uma teoria para predição e caracterização dinâmica de séries temporais. Teoremas da topologia diferencial formam a base matemática deste método. A minimização de uma função custo pelo método dos mínimos quadrados determina o preditor. Neste processo, o teste de Shapiro-Wilk avalia se a distribuição normal é uma boa aproximação para os resíduos do mapeamento global. As rotinas computacionais implementam a teoria desenvolvida e realizam o cálculo de uma correção que aumenta a acurácia da previsão pela técnica global. A teoria e os procedimentos computacionais admitem diferentes tempos de predição. O método estabelece o nível de confiança na predição, permite previsões acuradas e requer baixo tempo de execução. Critérios para o controle de qualidade dos mapeamentos globais podem ser implementados computacionalmente. Neste trabalho, um tipo de representação gráfica emprega resultados da aproximação global na caracterização da dinâmica de um sistema. Tal diagrama é a base para a construção de um novo quantificador que detecta a presença de caos a partir da série temporal. Os testes da teoria e das rotinas computacionais utilizam dados simulados ou que tem sua origem em fenômenos complexos do mundo real.

Palavras-chave: Séries Temporais. Caos. Previsibilidade. Computação Algébrica.

## ABSTRACT

ALVES, P.R.L. *The global approach to the inverse problem of time series*. 2017. 122 f. Tese (Doutorado em Física) – Instituto de Física Armando Dias Tavares, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2017.

The method of global approach is the axis of a theory for predicting and dynamic characteristics of time series. Theorems in the branch of differential topology are the mathematical basis for this method. The minimization of a cost function by the least squares method determines the predictor. In this process, the Shapiro-Wilk test evaluates if the normal distribution is a good approach to residuals in global mapping. The computational routines presented in this work implement the theory and perform the calculation of a quantity that increases the accuracy of the global forecast. The theory and computational procedures are compatible with different prediction times. The method establishes a confidence level in prediction, permits accurate forecasts and requires low runtimes. Criteria to control the quality of the global mappings are present in the computational routines. A kind of diagram developed in this paper uses results of the global approach to characterise the dynamics of a system from the time series. A computational program determines a new quantifier of chaos. Both the test of theory and the evaluation of the computational routines employ data that have their origin in simulated data or complex phenomena of real-world.

Keywords: Time Series. Chaos. Predictability. Algebraic Computation.

## LISTA DE FIGURAS

Figura 1 -	Manchas Solares . . . . .	11
Figura 2 -	Índice da Bolsa de Valores Dow Jones . . . . .	12
Figura 3 -	Séries temporais reais para o teste dos programas . . . . .	14
Figura 4 -	Fluxo circular de Couette . . . . .	21
Figura 5 -	Atrator estranho de Lorenz e Conjunto de Cantor . . . . .	22
Figura 6 -	Divergência de trajetórias com proximidade infinitesimal . . . . .	24
Figura 7 -	Teste de hipóteses e cálculo do p-valor . . . . .	33
Figura 8 -	Diagrama Acurácia-Desvio Logarítmico . . . . .	43
Figura 9 -	Quantificador de caos no plano complexo . . . . .	45
Figura 10 -	Variável dinâmica $X$ do Sistema de Lorenz . . . . .	65
Figura 11 -	Número de graus de liberdade e eficiência das rotinas <code>GfiTS</code> e <code>LinGfiTS</code> . . . . .	67
Figura 12 -	Mapeamentos distintos para o Sistema de Lorenz . . . . .	71
Figura 13 -	Distribuição dos resíduos no mapa <code>LinMap2</code> . . . . .	72
Figura 14 -	Esquema de geração do sinal caótico . . . . .	73
Figura 15 -	Mapa global do sinal elétrico caótico . . . . .	74
Figura 16 -	Desvio na predição das dezenas sorteadas na Mega-Sena . . . . .	76
Figura 17 -	Falsos vizinhos próximos e mútua informação média . . . . .	76
Figura 18 -	Distribuição dos resíduos com os preditores $\mathcal{P}_{pol}( x_r )$ e $\mathcal{P}_{log}( x_r )$ . . . . .	78
Figura 19 -	Desvios numa série temporal caótica . . . . .	80
Figura 20 -	Preditor polinomial aplicado na série temporal <code>LorenzX</code> . . . . .	80
Figura 21 -	Caracterização da série caótica experimental . . . . .	85
Figura 22 -	Testes com as séries periódica e aleatória . . . . .	86
Figura 23 -	Mapeamento global para os estudos de caso ( $\tau = 1$ ) . . . . .	88
Figura 24 -	Diagramas Acurácia-Desvio para os estudos de caso . . . . .	88
Figura 25 -	Expansão térmica do cobre . . . . .	106
Figura 26 -	Gráficos adicionais para a expansão térmica do cobre. . . . .	107
Figura 27 -	Comparação entre dois diferentes modelos de ajuste quanto ao seu desempenho . . . . .	118
Figura 28 -	Comparação entre o comando <code>PolynomialFit</code> e programa <code>LinFit</code> . . . . .	120
Figura 29 -	Desempenho dos comandos <code>PolynomialFit</code> e <code>NonlinearFit</code> . . . . .	121

## LISTA DE TABELAS

Tabela 1 - Exemplos de funções $\phi_i( x_r )$ . . . . .	28
Tabela 2 - Decisão estatística e p-valor . . . . .	35
Tabela 3 - Expoentes de Lyapunov espúrios . . . . .	39
Tabela 4 - Séries temporais de naturezas distintas . . . . .	40
Tabela 5 - Critérios de qualificação para os preditores . . . . .	41
Tabela 6 - Parâmetros para a qualificação dos preditores . . . . .	42
Tabela 7 - Caracterização dinâmica com o Diagrama Acurácia-Desvio Logarítmico . . . . .	45
Tabela 8 - Rotinas do pacote <code>TimeS</code> . . . . .	51
Tabela 9 - Os passos do algoritmo que aperfeiçoa a previsão global . . . . .	55
Tabela 10 - Os passos do algoritmo que aperfeiçoa a previsão global estendido . . . . .	57
Tabela 11 - Medida do tempo de execução das rotinas <code>GfiTS</code> e <code>LinGfiTS</code> . . . . .	67
Tabela 12 - Comparação entre preditores . . . . .	78
Tabela 13 - Qualificação dos preditores para as séries temporais caóticas . . . . .	87
Tabela 14 - Resultados para o quantificador de caos $Z_{dyn}$ . . . . .	87
Tabela 15 - Opções de análise disponíveis no programa <code>LinFit</code> . . . . .	109
Tabela 16 - Critérios para qualificação do <i>data fitting</i> . . . . .	109
Tabela 17 - Parâmetros para a detecção de <i>outliers</i> . . . . .	111
Tabela 18 - Comparação de grandezas estatísticas e detecção de <i>outliers</i> . . . . .	115
Tabela 19 - Data fitting com diferentes modelos de ajuste . . . . .	117
Tabela 20 - Comparação entre o desempenho dos comandos do pacote <code>Statistics</code> e do programa <code>LinFit</code> . . . . .	120

## SUMÁRIO

	INTRODUÇÃO . . . . .	10
1	RECONSTRUÇÃO DE TAKENS . . . . .	16
1.1	Mergulhos . . . . .	18
1.2	O Mergulho de Whitney . . . . .	19
1.3	Teoremas de Takens . . . . .	20
1.4	Atratores estranhos . . . . .	22
1.5	Séries temporais e o caos . . . . .	24
1.6	O problema inverso das séries temporais . . . . .	26
2	A APROXIMAÇÃO GLOBAL . . . . .	27
2.1	Previsão pela aproximação global . . . . .	27
2.2	Determinação do preditor . . . . .	29
2.3	Distribuição dos resíduos . . . . .	31
2.4	Teste de hipóteses e o p-valor . . . . .	32
2.5	Teste de Shapiro-Wilk . . . . .	35
2.6	Nível de confiança na predição . . . . .	36
3	CARACTERIZAÇÃO DINÂMICA . . . . .	38
3.1	Expoentes de Lyapunov a partir de séries temporais . . . . .	38
3.2	A caracterização dinâmica pela Aproximação Global . . . . .	39
3.3	Qualificação dos preditores . . . . .	41
3.4	Diagramas Acurácia-Desvio e Acurácia-Desvio Logarítmico . . . . .	42
3.5	Um novo quantificador para o caos . . . . .	44
4	ROTINAS COMPUTACIONAIS . . . . .	47
4.1	Pacote LinMapTS . . . . .	47
4.2	Sobre o pacote TimeS . . . . .	49
4.3	O algoritmo para o aperfeiçoamento da previsão global revisitado . . . . .	51
4.4	A extensão do algoritmo para diferentes tempos de predição . . . . .	55
4.5	A nova versão do pacote TimeS . . . . .	57
4.6	O programa DynCharTS . . . . .	60
4.7	Sobre os resultados do programa DynCharTS . . . . .	61
4.8	Problemas convencionais de <i>data fitting</i> . . . . .	62
5	RESULTADOS E DISCUSSÕES . . . . .	64
5.1	Aplicação do pacote LinMapTS . . . . .	64
5.2	Comparação dos tempos de execução das rotinas GfiTS e LinGfiTS . . . . .	68
5.3	Aperfeiçoamento na Aproximação Global Polinomial . . . . .	69
5.4	Série experimental caótica . . . . .	72
5.5	O desvio numa série randômica . . . . .	75

5.6	Viabilidade dos preditores não polinomiais . . . . .	77
5.7	Variável dinâmica $X$ do Sistema de Lorenz . . . . .	79
5.8	Série experimental caótica revisitada . . . . .	82
5.9	Sinal periódico e série randômica . . . . .	84
5.10	O Número de Manchas Solares e o Índice Dow Jones . . . . .	86
	<b>CONCLUSÃO</b> . . . . .	90
	<b>REFERÊNCIAS</b> . . . . .	92
	<b>APÊNDICE A</b> – Classe restrita de funções nos problemas de <i>data fitting</i> . . . . .	98
	<b>APÊNDICE B</b> – Abordagem analítica do ajuste de dados . . . . .	101
	<b>APÊNDICE C</b> – Análise gráfica . . . . .	105
	<b>APÊNDICE D</b> – O programa LinFit . . . . .	108
	<b>APÊNDICE E</b> – Aplicações do programa LinFit . . . . .	113
	<b>APÊNDICE F</b> – Computação Algébrica no <i>data fitting</i> . . . . .	117

## INTRODUÇÃO

Obter conhecimento quantitativo da natureza a partir da observação remete à origem do método científico. A combinação entre experiência e métodos matemáticos tem papel crucial desde a explicação da queda dos corpos até a interpretação dos resultados no contemporâneo acelerador de partículas LHC. Uma vez estabelecida a lei matemática para uma determinada categoria de fenômenos, o problema de determinar a evolução temporal de um observável depende das condições iniciais e da possibilidade de calcular a quantidade de interesse. Ocorre, contudo, que o conjunto de parâmetros necessários à determinação completa do observável não está disponível para um largo espectro de fenômenos. Também são de interesse científico as situações em que nem sequer existem propostas de uma lei determinística que possa reger o funcionamento do sistema. Em face da necessidade de investigar estas duas últimas categorias de fenômenos, emprega-se a metodologia de um *problema inverso*.

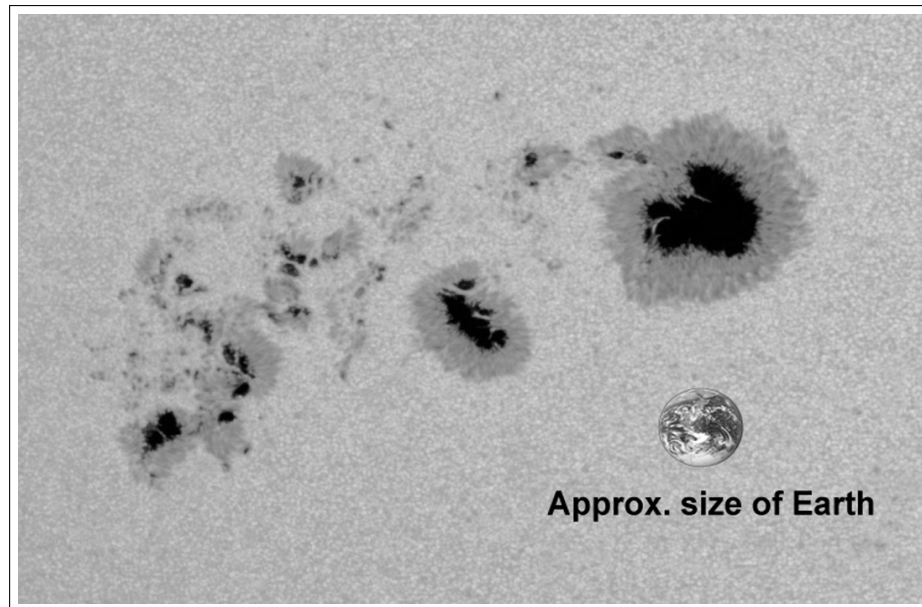
Historicamente, os *problemas diretos* têm sido extensivamente estudados, enquanto que os inversos são mais recentes e ainda não tão bem compreendidos. Determinar o potencial gravitacional da Terra numa posição da crosta, sendo conhecida a distribuição de massa, é um problema direto. Por outro lado, encontrar a distribuição de massa num determinado ponto do interior do planeta — a partir do potencial gravitacional — constitui o problema inverso. Neste exemplo, a abordagem inversa tem maior interesse porque pode indicar a presença de minério ou petróleo a partir das anomalias do campo gravitacional na superfície terrestre (KELLER, 1976).

Técnicas de manipulação de medidas e dados para que conclusões sejam estabelecidas a respeito de sistemas físicos são empregadas com o objetivo de resolver problemas inversos. Newton resolveu um problema desta natureza ao descobrir que a lei de força que rege o movimento dos planetas está em acordo com as Leis de Kepler. Atualmente, a teoria dos problemas inversos para equações diferenciais está sendo amplamente desenvolvida no domínio da Física Matemática. Transferência de calor, transferência de massa, teoria da elasticidade, teoria dos potenciais, física nuclear e hidrodinâmica são áreas de concentração desta linha de pesquisa (PRILEPKO, 2000).

Fenômenos complexos são candidatos em potencial para receberem abordagens na concepção dos problemas inversos. Um estudo de caso que motiva o desenvolvimento de uma metodologia inversa diz respeito a incidência das *manchas solares*.

O fenômeno das Manchas Solares foi observado por Galilei (1613), que o considerou como um evento superficial capaz de demonstrar o movimento de rotação do Sol. Manchas na superfície solar, como as obtidas pela NASA (2014) e mostradas na Figura 1, são muito provavelmente causadas por intensos campos magnéticos no interior do Sol (HATHAWAY, 2015).

Figura 1 - Manchas Solares



Legenda: A mancha solar gigante AR1944 — visualizada na escala da Terra — foi observada no início de janeiro de 2014.

Fonte: NASA, 2014.

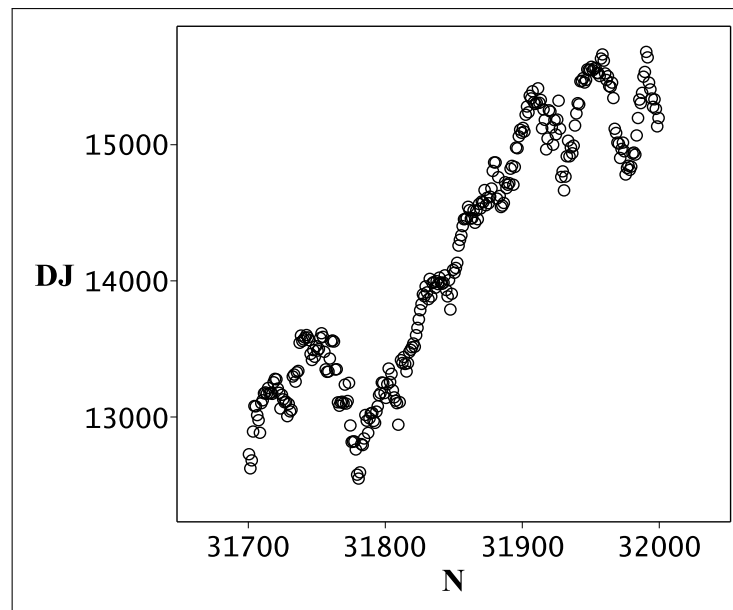
No trabalho seminal da Análise de Séries Temporais sobre a investigação da periodicidade em séries perturbadas, Yule (1927) faz referência especial ao Número de Manchas Solares de Wolf — um indicador da atividade solar com registros de longa data.

O interesse na evolução temporal da atividade solar ganha uma motivação adicional nos dias atuais em que a Ciência se debruça sobre o problema do aquecimento global. A duração do ciclo solar é um ingrediente importante na temperatura do planeta e também nas alterações climáticas. Existem dados de observações que sugerem uma correlação entre as variações de longo prazo na temperatura da Terra e as mudanças na duração do ciclo solar (FRIIS-CHRISTENSEN; LASSEN, 1991).

Um problema de natureza diferente da astrofísica que também desperta o interesse de físicos é a previsibilidade dos índices do mercado de ações (ver Figura 2). A aplicação de modelos da Física na interpretação do comportamento subjacente dos mercados levou a importantes resultados nas últimas décadas. Além das séries temporais, as leis de potência e de escala vem sendo empregadas no campo das Transições de Fase, Mecânica Estatística e Sistemas Não Lineares, numa área de pesquisa interdisciplinar chamada de *Econofísica* (MANTEGNA, 2000).

Naturalmente, o tratamento de sistemas com a complexidade dos dois exemplos apresentados deve contemplar as não linearidades que fazem parte tanto do comportamento do mercado de ações quanto do indicador da atividade solar. O método da **Reconstrução do Espaço de Fase** é capaz de descrever comportamentos pouco regulares,

Figura 2 - Índice da Bolsa de Valores Dow Jones



Legenda: O gráfico mostra os índices  $DJ$  em função de sua ordem  $N$  na série temporal.

Fonte: O autor, 2017.

inclusive quando o caos está presente. Este método tem sua motivação na Teoria dos Sistemas Dinâmicos, na qual usualmente os sistemas são representados por um conjunto de equações diferenciais de primeira ordem no espaço de fase. A existência e a unicidade das soluções destas equações possuem o suporte da Teoria das Equações Diferenciais e implicam no *determinismo* do espaço de fase (KANTZ; SCHREIBER, 2003).

## A Computação Algébrica na Análise de Séries Temporais

O uso de computadores para a manipulação de expressões matemáticas permite a execução de longas e complicadas tarefas algébricas num tempo de execução exíguo. O desenvolvimento de sistemas para a computação matemática de caráter simbólico tornou-se uma ativa área de pesquisa e implementação na década de 1961-1971. Atualmente, a *Computação Algébrica* constitui-se numa reconhecida área de pesquisa e ensino na Ciência da Computação e na Matemática. Os esforços nesta linha de investigação contemplam o desenvolvimento de *sistemas* — linguagens de programação e *software* associado para manipulação simbólica —, *algoritmos* — eficientes algoritmos matemáticos para a manipulação de polinômios e funções mais gerais —, e *aplicações* — que abrangem vastas áreas do conhecimento e motivam o desenvolvimento de sistemas e algoritmos (GEDDES, 1992).

Carli, Duarte e da Mota (2014b) desenvolveram um conjunto de procedimentos no **Sistema Maple** para a Análise de Séries Temporais . Este pacote computacional realiza a reconstrução dos vetores de estado a partir dos observáveis armazenados em um arquivo ou numa lista. A previsão de um observável desconhecido é realizada pela aplicação de um mapa polinomial e de um termo que pode melhorar a aproximação do valor verdadeiro.

Cabe, neste ponto, dar um exemplo que evidencie o poderio dos recursos da Computação Algébrica na predição. Para isto, um polinômio de segundo grau (1) é aplicado a um vetor de estado tridimensional  $|x_r\rangle$  no espaço reconstruído para o conjunto de índices Dow Jones (ver Figura 2), armazenado no arquivo ‘downjones’.txt:

$$\mathcal{P}(|x_r\rangle) = c_1x_{1r} + c_2x_{2r} + c_3x_{3r} + c_4x_{1r}^2 + c_5x_{1r}x_{2r} + c_6x_{1r}x_{3r} + c_7x_{2r}^2 + c_8x_{2r}x_{3r} + c_9x_{3r}^2. \quad (1)$$

O índice a ser previsto tem a ordem de entrada na série  $N = 32001$  e as coordenadas utilizadas no polinômio correspondem aos índices da bolsa que ocupam as respectivas posições na série temporal: 32000 ( $x_{1r}$ ), 31994 ( $x_{2r}$ ) e 31988 ( $x_{3r}$ ). Para a reconstrução do espaço de fase, o comando **VecTS** gera os vetores a partir dos dados armazenados num arquivo (CARLI; DUARTE; da MOTA, 2014a). No modo Worksheet do Maple, o *prompt* reproduzido abaixo tem como saída os vetores tridimensionais (argumento **Dim=3**), com defasagem de seis entradas na série entre suas coordenadas (argumento **TimeLag=6**). Os nove parâmetros  $c_i$  do polinômio (1) são determinados pelo *método dos mínimos quadrados* a partir do comando **LinGfiTS** (ALVES; DUARTE; da MOTA, 2016a).

```
[> V := VecTS('downjones'.txt, Dim=3, TimeLag=6):
[> Map1 := LinGfiTS(V, 32000, Degree=2):
```

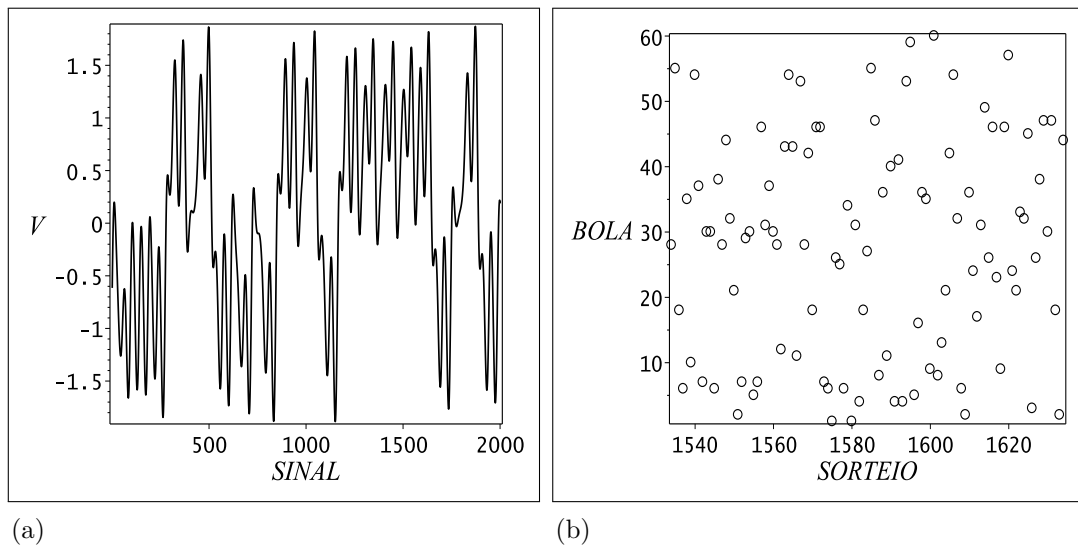
Acima, os argumentos correspondem respectivamente: à lista de vetores reconstruídos **V**, ao dado da série anterior ao que se deseja prever (**32000**) e ao grau da função preditiva (**Degree=2**). O polinômio gerado é

$$\begin{aligned} \text{Map1} : &= 0.146478 X_1 + 1.43872 X_2 - 0.397422 X_3 \\ &+ 0.000029714 X_1^2 + 0.0000348435 X_1X_2 - 0.000044795 X_1X_3 \\ &+ 0.0000412013 X_2^2 - 0.000215511 X_2X_3 + 0.000142166 X_3^2. \end{aligned}$$

Finalizando este exemplo, o resultado da previsão tem um erro de +0.23%. O valor verdadeiro do índice é dado pelo primeiro componente do vetor **V[32001]**.

```
[> ForecastTS(Vects=V, Map=Map1, Position=32000); V[32001][1];
15167.9039
15133.1362
```

Figura 3 - Séries temporais reais para o teste dos programas



Legenda: (a) Sinal elétrico caótico. O gráfico relaciona a voltagem  $V$  com a sua ordem na série temporal. (b) Dezenas sorteadas na loteria Mega-Sena.

Fonte: O autor, 2017.

## O plano desta tese

O primeiro capítulo tem o objetivo de apresentar a sustentação matemática do esquema da reconstrução. O ponto de partida é a abordagem da série temporal como resultado da evolução de um sistema dinâmico e da sua representação num espaço de fase. Em seguida, são definidos os objetos topológicos necessários para o acesso ao conteúdo dos Teoremas de Takens. Estes teoremas constituem o fundamento de toda a teoria e das rotinas computacionais que foram desenvolvidas e são exploradas ao longo do texto.

Na sequência deste capítulo inicial, existem três propostas. A primeira é tratar da escolha dos parâmetros que definem os vetores de estado no espaço reconstruído. A segunda consiste em partir da motivação experimental dos teoremas para desenvolver o conteúdo da Teoria do Caos que tem conexão com a reconstrução de Takens. E a última tem a finalidade de definir a natureza do problema a ser atacado nos capítulos seguintes.

Discutir e definir a Aproximação Global, para qualquer tempo de predição, é o objetivo inicial do segundo capítulo. O seguinte é apresentar o ganho na operacionalização — com vistas a análise em séries temporais — como consequência da limitação da forma imposta à função preditiva. O desenvolvimento realizado, a partir do método da máxima verossimilhança, se propõe a fornecer uma formulação matemática eficiente para a obtenção dos coeficientes responsáveis pela predição. A estratégia consiste em assumir e testar a forma de distribuição normal para os resíduos num processo de ajuste de dados e estabelecer o nível de confiança na previsão de um observável.

Apresentar uma teoria para a caracterização dinâmica — diretamente a partir dos dados armazenados na série temporal — está na proposta do terceiro capítulo. Quantidades estatísticas e um conjunto de critérios para assegurar a qualidade do mapeamento global são os ingredientes desta iniciativa de diagnosticar o tipo de evolução temporal experimentada pelo conjunto de observáveis. Um aspecto inovador neste trabalho contempla um tipo de representação gráfica que contém a assinatura da dinâmica na série temporal. Neste contexto, as ideias apresentadas constituem uma alternativa a abordagens tradicionalmente propostas e que resultam na construção de um novo quantificador para o caos.

A descrição detalhada dos programas — desenvolvidos por Alves, Duarte e da Mota (2016c) — que implementam computacionalmente a aproximação global está no cerne do capítulo seguinte. Esta parte do texto tem também a finalidade de revisitar o algoritmo que aperfeiçoa o resultado da predição obtido pela técnica da aproximação global e a sua extensão — na nova versão do pacote TimeS (ALVES; DUARTE; da MOTA, 2016d) — para diferentes *tempos de predição*.

No quinto capítulo, os pacotes computacionais são aplicados numa série caótica que tem sua origem em uma simulação numérica e em séries temporais obtidas do mundo real. O experimento que gera um sinal elétrico com evolução temporal caótica — representado no gráfico da Figura 3a — destina-se a ressaltar que as ferramentas e a teoria são compatíveis com o contexto da experiência.

Ao lado da representação gráfica do sinal caótico, a Figura 3b mostra as dezenas sorteadas em uma loteria. Esta série com dados aleatórios cumpre o papel de fornecer o ambiente para o teste da sensibilidade das rotinas. Em toda predição existe a intenção de obter um resultado o mais acurado possível. Todavia, o mecanismo de predição deve ser verossímil. Neste sentido, resultados de baixa acurácia para sistemas desprovidos de qualquer determinismo indica que o método empregado é capaz de descrever fielmente a realidade.

No mapeamento global de extensos intervalos de uma série temporal, polinômios de alto grau são investigados quando a sua capacidade de aumentar a acurácia da predição e o tempo necessário à sua implementação computacional. O desempenho de outras formas funcionais das coordenadas do espaço de fase reconstruído — como alternativas aos modelos polinomiais — é avaliado.

No final do capítulo, a teoria e as rotinas são testadas em séries de diferentes naturezas, mas que são previamente conhecidas. Já a evolução dos índices Dow Jones e o Número de Manchas Solares — mencionados no início desta Introdução — são retomados como estudos de caso que ilustram as possibilidades de aplicação do método desenvolvido em fenômenos complexos do mundo real.

## 1 RECONSTRUÇÃO DE TAKENS

O conhecimento e a compreensão da realidade física prescindem do conjunto formado pelas medidas relacionadas aos fenômenos de interesse científico. No caso das *teorias de primeiros princípios*, as proposições levam a resultados compatíveis com as observações empíricas, como é o caso da Mecânica Estatística. Por outro lado, nas *teorias fenomenológicas*, o conhecimento é estabelecido a partir de medições das grandezas físicas em questão. No desenvolvimento da Física, ocorre que teorias de primeiros princípios podem construir uma justificação profunda dos modelos que estão de acordo com os resultados da experiência. Assim, a Mecânica Estatística acaba por elucidar a fenomenologia da Termodinâmica a partir da sua abordagem microscópica dos objetos que constituem os sistemas em estudo.

Na abordagem fenomenológica, uma forma de conhecer e entender a realidade consiste em obter informações de um sistema a partir de medidas ou dados — designados indistintamente como *observáveis* nesta tese — ordenados segundo os instantes em que foram obtidos. A evolução temporal do observável decorre da dinâmica do sistema. Este conjunto ordenado é chamado de *série temporal*. Exemplos de observáveis são grandezas físicas medidas num experimento, valores das variáveis de um sistema dinâmico obtidos por integração numérica e índices do mercado financeiro.

É razoável admitir que a dinâmica do sistema que evolui no tempo deixa sua assinatura na série temporal. Esta hipótese abre a perspectiva de obter informações válidas sobre todo o sistema a partir de um único observável conhecido, o que é aplicável em diversas áreas do conhecimento humano. O presente trabalho explora o conteúdo de uma série temporal  $X_{TS}$  formada por  $S$  quantidades escalares.

$$X_{TS} = \{X(0\Delta t), X(1\Delta t), X(2\Delta t), \dots, X((S-1)\Delta t)\} \quad (2)$$

Na construção da série, os observáveis são ordenados segundo múltiplos inteiros de um intervalo de tempo  $\Delta t$ . A dinâmica responsável pela evolução temporal na série não é sequer considerada; parte-se do princípio que é desconhecida.

O *sistema dinâmico* que se pretende investigar por meio da série temporal pode ser concebido como um conjunto de  $N$  equações diferenciais nas variáveis  $(X_1, X_2, \dots, X_N)$ :

$$\begin{aligned} \frac{dX_1}{dt} &= F_1(X_1, X_2, \dots, X_N) \\ \frac{dX_2}{dt} &= F_2(X_1, X_2, \dots, X_N) \\ &\vdots \\ \frac{dX_N}{dt} &= F_N(X_1, X_2, \dots, X_N), \end{aligned} \quad (3)$$

onde o tempo  $t$  é a única variável independente e as funções  $F_i$  possuem primeiras derivadas parciais contínuas.

**Definição 1.** *Sejam  $|a\rangle$  e  $|b\rangle$  dois vetores quaisquer pertencentes ao espaço  $E$  e  $\alpha$  um número real. Um **espaço euclidiano**  $E$  é um espaço no qual está definido o produto interno  $(|a\rangle, |b\rangle)$  com as seguintes propriedades:*

$$1a. (|a\rangle_1 + |a\rangle_2, |b\rangle) = (|a\rangle_1, |b\rangle) + (|a\rangle_2, |b\rangle);$$

$$1b. (\alpha |a\rangle, |b\rangle) = \alpha (|a\rangle, |b\rangle);$$

$$1c. (|a\rangle, |b\rangle) = (|b\rangle, |a\rangle);$$

$$1d. (|a\rangle, |a\rangle) \geq 0.$$

A reta real  $\mathbb{R}$ , o plano  $\mathbb{R}^2$  e o espaço tridimensional ordinário  $\mathbb{R}^3$  são os espaços euclidianos munidos de coordenadas mais comumente utilizados.

O *espaço de fase* das variáveis dinâmicas do conjunto de equações diferenciais (3) é um espaço euclidiano com coordenadas  $X_1, X_2, \dots, X_N$ . A evolução temporal do *estado* do sistema pode ser visualizado como uma partícula em movimento, descrevendo uma *trajetória* no espaço de fase. O conceito de espaço de fase ganhou importância crescente no cenário da Física a partir da sua ampla utilização por eminentes matemáticos e físicos como Poincaré — no seu tratamento das soluções de equações diferenciais —, Gibbs — no seu desenvolvimento da mecânica estatística — e Birkhoff — no seu tratado sobre sistemas dinâmicos (LORENZ, 1963).

Admitindo que o espaço de fase original é inacessível, assume-se a alternativa de reconstruir este espaço a partir de vetores num espaço euclidiano de dimensão  $d_E$ , com componentes escolhidos entre os observáveis da série temporal. A técnica empregada utiliza atrasos  $T$  entre os observáveis, chamados na literatura de *delay time* ou *lag* (KANTZ; SCHREIBER, 2003).

$$|x_r(t)\rangle \doteq \begin{bmatrix} x(r(d_E T)\Delta t) \\ \vdots \\ x(r(2T)\Delta t) \\ x(rT\Delta t) \end{bmatrix} \quad (4)$$

Os dados ou sinais da série temporal escolhidos são os componentes do  $r$ -ésimo vetor reconstruído  $|x_r(t)\rangle$ , que posiciona um estado no espaço reconstruído (PACKARD et al., 1980; RUELLE, 1989).

Conhecer o tipo de relação entre o espaço de fase original e o reconstruído é uma questão naturalmente colocada nesta discussão. A *Topologia Diferencial* incorpora os elementos adequados para a investigação desta relação entre os dois espaços e também da consistência apresentada pelo método da reconstrução do espaço de estados.

## 1.1 Mergulhos

O acesso ao conteúdo da evolução temporal das variáveis dinâmicas do sistema é realizado a partir dos vetores  $|x_r\rangle$ . Para que a dinâmica original seja preservada no espaço formado por estes vetores, existe a necessidade de que a reconstrução seja propriamente realizada. A base matemática que trata desta crucial questão tem seu fundamento em aplicações suaves chamadas de *mergulhos*, que fazem parte do contexto da Topologia Diferencial. Com o propósito de apresentar formalmente um mergulho, é útil partir de um conjunto de definições e também de alguns teoremas com consequências importantes para o esquema da reconstrução.

**Definição 2.** Um **homeomorfismo** é uma aplicação contínua, inversível e cuja inversa também é contínua.

**Definição 3.** Um **difeomorfismo** é uma aplicação inversível, diferenciável e cuja inversa é diferenciável também.

Quando existe a intenção de recuperar aspectos globais a partir de uma abordagem local dos espaços, convém raciocinar em termos da *compacidade* — uma condição necessária — das topologias envolvidas. Uma caracterização possível de um espaço compacto é realizada através de um teorema formulado a partir da *interseção finita*, cuja prova pode ser encontrada na obra de Kelley (1975).

**Definição 4.** Uma família  $\mathcal{B}$  de conjuntos tem a propriedade **interseção finita** se e somente se a interseção dos membros de cada subfamília finita de  $\mathcal{B}$  não é um conjunto vazio.

**Teorema 1.** Um espaço topológico é **compacto** se e somente se cada família de conjuntos fechados que possui a propriedade de interseção finita tem um conjunto interseção não vazio.

A Topologia Diferencial estuda as *variedades* diferenciáveis e *mapas*, com a tarefa de descobrir e analisar as propriedades globais das variedades (HIRSCH, 1976). Exemplos de *variedades compactas* são o círculo, a esfera  $m$ -dimensional e o toro. O termo mapa é empregado para designar uma aplicação.

**Definição 5.** Uma **variedade**  $\mathcal{M}$  de dimensão  $m$  é um espaço topológico localmente homeomórfico ao  $\mathbb{R}^m$ ; para cada ponto  $x \in \mathcal{M}$ , existe uma vizinhança de  $x$  localmente euclidiana. A variedade compacta é um espaço topológico compacto.

Para ilustrar a aplicação das variedades na Física, o conteúdo inteiro da Relatividade Geral pode ser colocado num espaço-tempo como uma variedade sobre a qual é definida a métrica de Lorentz  $g_{ab}$ . A curvatura de  $g_{ab}$  está relacionada à distribuição de matéria no espaço-tempo pela *Equação de Einstein* (WALD, 1984).

O sistema dinâmico definido pelas equações diferenciais (3) pode ser inserido no domínio das variedades. Significa a descrição de como um estado sofre a transformação para outro estado com o transcorrer do tempo. Tecnicamente, um sistema dinâmico corresponde a uma aplicação suave de números reais (*sistemas contínuos*) ou inteiros (*sistemas discretos*) em outro objeto matemático<sup>1</sup>. Usualmente as variedades fazem parte do contexto destas aplicações (WEISSTEIN, 2016).

**Definição 6.** *A aplicação  $\mathcal{G} : \mathcal{M} \mapsto \mathcal{N}$  é um **mergulho** se  $\mathcal{G}$  é uma função diferenciável entre variedades diferenciáveis, com derivada injetiva em todos os pontos, que mapeia  $\mathcal{M}$  homeomorficamente sobre sua imagem em  $\mathcal{N}$ .*

Um espaço  $\mathcal{M}$  é mergulhado em outro espaço  $\mathcal{N}$  quando as propriedades de sua imagem em  $\mathcal{N}$  são idênticas àquelas verificadas em  $\mathcal{M}$ . Neste sentido, é esclarecedora a própria construção dos números: os racionais são mergulhados nos reais, assim como os inteiros são mergulhados nos racionais. Um mergulho é a representação de um objeto topológico num determinado espaço, de maneira que as propriedades algébricas ou a sua conexidade sejam preservadas. O mergulho de um *campo vetorial*, por exemplo, preserva as propriedades de adição e produto; os *conjuntos abertos* assim permanecem em espaços topológicos mergulhados e a *conexidade* também é preservada no mergulho de *grafos* (INSALL MATT; ROWLAND; WEISSTEIN, 2015).

## 1.2 O Mergulho de Whitney

Os mapas correspondentes aos sistemas dinâmicos agem sobre variedades. No intuito de manipular o conteúdo dinâmico que pode existir em uma série temporal, a partir da reconstrução do espaço de fase, o primeiro passo é avaliar se há alguma garantia de que estas variedades sejam mergulhadas no espaço reconstruído. O segundo consiste na construção de mapas a partir dos observáveis ordenados no tempo que sejam mergulhos no espaço obtido pela reconstrução.

Whitney (1936) trata o problema do mergulho de variedades diferenciáveis de classe  $C^r$ , ou seja, de variedades que possuem derivadas parciais contínuas de ordem  $r$ , numa perspectiva puramente analítica e disponibiliza um ferramental matemático para a abordagem das variedades diferenciáveis e seus mapeamentos.

**Teorema 2.** *Qualquer variedade  $\mathcal{M}$  de dimensão  $m$  e classe  $C^r$  ( $r \geq 1$ ) é  $C^r$ -homeomórfica a uma variedade analítica no espaço euclidiano  $\mathbb{R}^{2m+1}$ .*

---

<sup>1</sup> Em seu sentido mais amplo, Sistemas Dinâmicos envolvem: equações diferenciais, topologia diferencial, topologia geral e teoria ergódica. O espaço de fase de um sistema dinâmico — no *stricto sensu* — é tratado como uma variedade diferenciável (AOKI, 1994).

O teorema acima é o primeiro dos dois teoremas fundamentais para todos os resultados do artigo de Whitney (1936), e responde de forma conclusiva que uma variedade  $\mathcal{M}$ , de dimensão  $m$ , pode ser mergulhada num espaço euclidiano com a dimensão  $2m + 1$ .

Tendo em vista as aplicações, é um aspecto favorável que as variedades sejam mapeadas em espaços euclidianos devido a vasta experiência acumulada na manipulação de vetores em espaços  $n$ -dimensionais  $\mathbb{R}^n$ . Uma partícula sem spin, por exemplo, pode ser estudada pela Mecânica Quântica no ambiente da variedade  $\mathcal{M}$ , de dimensão  $m$ , mergulhada no espaço euclidiano  $\mathbb{R}^n$  (SCHUSTER; JAFFE, 2003).

### 1.3 Teoremas de Takens

Esta etapa do trabalho tem por finalidade investigar se o mapeamento dos componentes dos vetores reconstruídos de acordo com (4) representa, ou não, um mergulho no espaço  $\mathbb{R}^{2m+1}$ . O observável contido na série temporal (2) é tratado como uma função suave  $y : M \mapsto \mathbb{R}$ , onde  $M$  é uma variedade compacta. O problema de obter informação a respeito do sistema dinâmico original e sua variedade  $M$ , com a evolução temporal  $\varphi_t$  do **observável**  $y$ , ou seja,  $t \mapsto y(\varphi_t(x))$  com  $x \in M$ , é resolvido por Takens (1981) ao provar três teoremas:

**Teorema 3.** *Seja  $M$  uma variedade compacta de dimensão  $m$ ,  $x \in M$ ,  $\varphi$  um difeomorfismo  $\varphi : M \mapsto M$  e  $y$  uma função  $y : M \mapsto \mathbb{R}$ . Para cada par genérico  $(\varphi, y)$  de classe  $C^r$  ( $r \geq 2$ ), o mapa  $\phi_{(\varphi, y)} : M \mapsto \mathbb{R}^{2m+1}$ , definido por*

$$\phi_{(\varphi, y)}(x) = \left( y(x), y(\varphi(x)), \dots, y(\varphi^{2m}(x)) \right),$$

*é um mergulho.*

**Teorema 4.** *Seja  $M$  uma variedade compacta de dimensão  $m$ ,  $x \in M$ ,  $X$  um campo vetorial,  $\varphi_t$  o fluxo de  $X$  e  $y$  uma função  $y : M \mapsto \mathbb{R}$ . Para cada par genérico  $(X, y)$  de classe  $C^r$  ( $r \geq 2$ ), o mapa  $\phi_{(X, y)} : M \mapsto \mathbb{R}^{2m+1}$ , definido por*

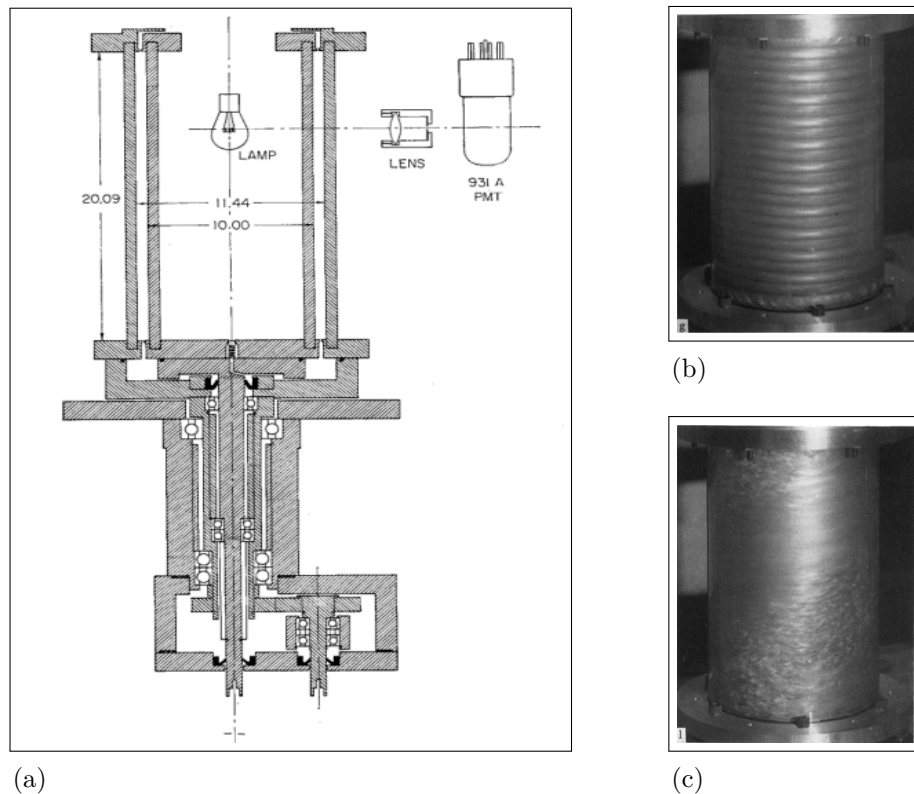
$$\phi_{(X, y)}(x) = \left( y(x), y(\varphi_1(x)), \dots, y(\varphi_{2m}(x)) \right),$$

*é um mergulho.*

**Teorema 5.** *Seja  $M$  uma variedade compacta de dimensão  $m$ ,  $x \in M$ ,  $X$  um campo vetorial,  $\varphi_t$  o fluxo de  $X$  e  $y$  uma função  $y : M \mapsto \mathbb{R}$ . Para cada par genérico  $(X, y)$  de classe  $C^r$  ( $r \geq 2m + 1$ ), o mapa  $\tilde{\phi}_{(X, y)} : M \mapsto \mathbb{R}^{2m+1}$ , definido por*

$$\tilde{\phi}_{(X, y)}(x) = \left( y(x), \left. \frac{d}{dt} y(\varphi_t(x)) \right|_{t=0}, \dots, \left. \frac{d^{2m}}{dt^{2m}} y(\varphi_t(x)) \right|_{t=0} \right),$$

Figura 4 - Fluxo circular de Couette



Legenda: (a) Seção longitudinal de um aparato experimental para o estudo do fluxo de Couette. (b) Fluxo laminar com dois harmônicos. (c) Turbulência.

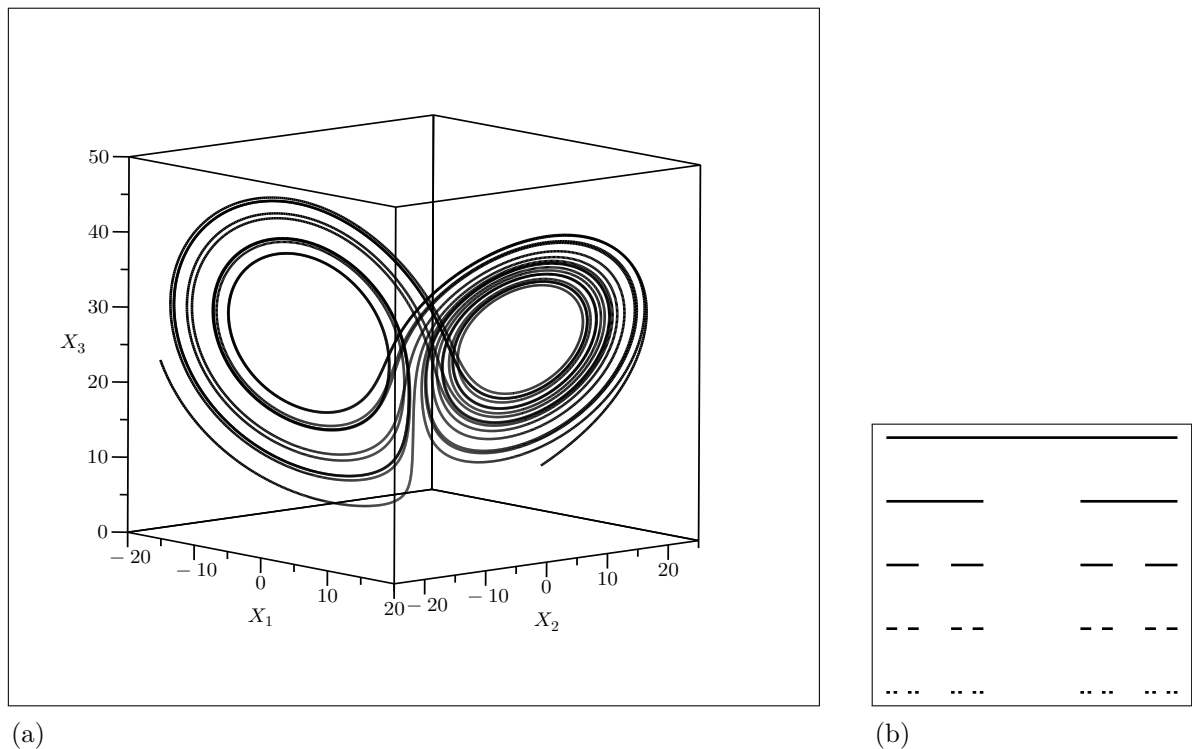
Fonte: COLES, 1965, p. 386.

*é um mergulho.*

A relação entre a dimensão  $m$  da variedade sobre a qual age o sistema dinâmico e a dimensão do espaço reconstruído  $d_E$  — dada por  $d_E = 2m + 1$  — é a garantia de que o mapa obtido pelo método do lag seja um mergulho no espaço euclidiano da reconstrução. Esta é uma **condição suficiente** para o propósito de preservar a dinâmica do sistema contida na série temporal. Todavia, não foi provado que esta condição seja **necessária**. Este é um ponto importante, porque bons resultados podem ser obtidos em **espaços de baixa dimensionalidade**.

Nas situações reais em que a dimensão  $m$  — correspondente ao sistema em estudo — é desconhecida, não se pode determinar qual deve ser a dimensão mínima  $d_E$  a ser usada na reconstrução do espaço de estados. A iniciativa de aumentar indiscriminadamente a dimensão do espaço requer cautela. Espaços reconstruídos com um número de dimensões maior que o necessário têm implicações imediatas no esforço computacional para a implementação de algoritmos. Um número muito grande de coordenadas implica também no aumento do número de parâmetros a serem ajustados nas funções responsáveis pela

Figura 5 - Atrator estranho de Lorenz e Conjunto de Cantor



Legenda: (a) Atrator de Lorenz com condições iniciais  $X_1(0) = X_2(0) = X_3(0) = 10$  e parâmetros  $\sigma = 10$ ,  $r = 28$ ,  $b = 8/3$ . (b) Todos os níveis de construção do Conjunto de Cantor são autossimilares.

Fonte: O autor, 2017.

predição. O resultado pode ser um ajuste muito bom para os observáveis utilizados na minimização; mas, ao mesmo tempo, pouco acurado no instante de interesse da previsão.

Os parâmetros necessários à reconstrução dos vetores de estado são a dimensão de mergulho  $d_E$  e o *delay time*  $T$ . Como ambos estão conjugados no vetor  $|x_r\rangle$  pela multiplicação  $d_E T$  (ver Equação (4)), a qualidade da reconstrução é definida por este produto. A própria *tentativa e erro* — devido ao poder computacional disponível nos dias atuais — mostra-se adequada para a escolha conjunta dos dois parâmetros (KANTZ; SCHREIBER, 2003).

#### 1.4 Atratores estranhos

Os Teoremas de Takens, apresentados na Seção 1.3, têm como motivação física a *turbulência* observada no fluido entre dois cilindros em rotação, observada no *Experimento de Taylor-Couette* (TAKENS, 1981). A Figura 4a mostra um dispositivo e seu esquema

ótico para o estudo de fluidos entre dois cilindros de acrílico em rotação. Nesta experiência, diferentes estados do fluxo circular de Couette são investigados (COLES, 1965). As diferentes características do escoamento dependem das velocidades angulares dos cilindros de acrílico (ver Figuras 4b e 4c).

Matematicamente, regimes diferenciados de escoamento são identificados na evolução temporal das velocidades de um fluido incompressível viscoso na análise da *Equação de Navier-Stokes*. Desprezando os efeitos térmicos, esta equação balanceia a derivada da velocidade  $v$  com relação ao tempo  $t$  com um determinado campo e pode ser colocada na forma

$$\frac{dv}{dt} = X_{\mu}(v), \quad (5)$$

na qual o campo vetorial  $X_{\mu}$  depende do parâmetro  $\mu$ .

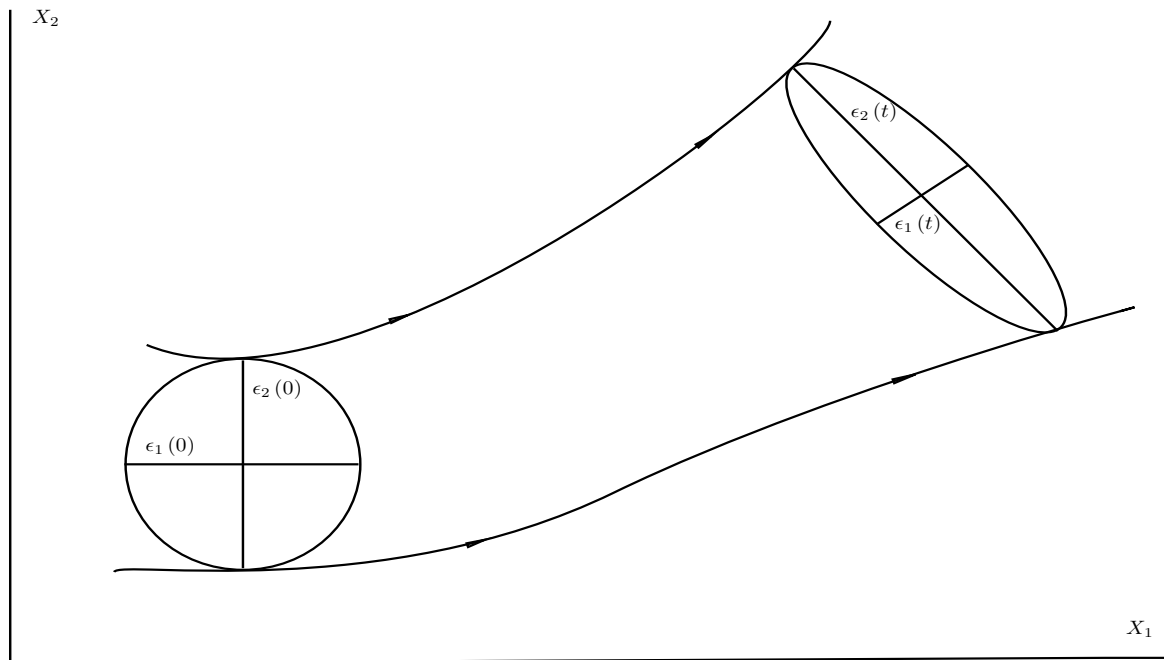
Ruelle e Takens (1971) estudam soluções da Equação de Navier-Stokes e, fazendo suposições muito gerais a respeito do campo  $X_{\mu}$ , verificam que o movimento do fluido se torna *caótico* quando  $\mu$  assume valores suficientemente altos. Este parâmetro  $\mu$  pode representar, por exemplo, o número de *Reynolds* ou o número de *Rayleigh*. O espaço de fase desses sistemas dinâmicos possuem regiões chamadas de *atratores* e *bacias de atração*, nas quais as trajetórias manifestam habitualmente uma dependência sensível das condições iniciais. Além de tais regiões, os *atratores estranhos* e os *atratores caóticos* também são precisamente definidos no presente trabalho. Estas quatro definições foram extraídas do artigo de Grebogi et al. (1984).

**Definição 7.** *Seja um espaço de fase que contém uma órbita que evolui no tempo  $t$ . Um **atrator** é um conjunto compacto contido neste espaço que é limitado pela órbita quando  $t \rightarrow +\infty$ , para quase todas as condições iniciais de sua vizinhança.*

**Definição 8.** *A **bacia de atração** de um atrator é o fecho do conjunto das condições iniciais que se aproxima do atrator quando o tempo transcorrido tende para infinito.*

No estudo qualitativo das equações diferenciais realizado por Ruelle e Takens (1971), o atrator com a estrutura geométrica local definida pelo produto de um *Conjunto de Cantor* por uma variedade bidimensional é considerado “estranho”. O processo repetitivo de dividir um segmento unitário em três partes e, depois, remover a parte do meio, tem como resultado o Conjunto de Cantor. A Figura 5b mostra quatro níveis desta construção. Como cada nível corresponde a uma réplica do outro, numa determinada escala, esta estrutura é *autossimilar*. Quando a ordem da construção aumenta indefinidamente, o comprimento de cada segmento tende a zero e o número de segmentos tende para infinito.

Figura 6 - Divergência de trajetórias com proximidade infinitesimal



Legenda: Contração e expansão no disco da perturbação inicial no plano de fase  $X_1X_2$ .

Fonte: O autor, 2017.

Objetos geométricos que, em uma ou mais dimensões, apresentam a característica de um Conjunto de Cantor são tipificados como *fractais* (GRASSBERGER; PROCACCIA, 1983; GREBOGI et al., 1984).

**Definição 9.** Um **atrator estranho** é um atrator que não é um conjunto finito de pontos e que também não é diferenciável por partes.

Quando um atrator é um fractal no espaço de fase, ele possui uma propriedade geométrica chamada *estranheza*. Um exemplo de atrator estranho no espaço de fase tridimensional é o emblemático Atrator de Lorenz (ver Figura 5a).

## 1.5 Séries temporais e o caos

Os fluidos em movimento, no regime da turbulência, apresentam uma complexidade comumente associada ao *caos*, que admite uma interpretação em termos de atratores estranhos. Existe uma ligação direta entre a evolução temporal dos sistemas que apresentam comportamento caótico e os observáveis armazenados numa série temporal. Kantz e Schreiber (2003) avaliam que a conexão mais direta entre a *teoria do caos* e o mundo real se dá através da análise de séries temporais, obtidas de sistemas reais, em termos da dinâmica não linear.

A complexidade detectada geometricamente através da estranheza tem atributos dinâmicos que se manifestam nas *órbitas* do atrator. Os efeitos mais contundentes dizem respeito ao afastamento que duas órbitas correspondentes a condições iniciais infinitesimalmente próximas sofrem com o transcorrer do tempo. Nos atratores estranhos, duas órbitas inicialmente localizadas numa pequena região do espaço de fase frequentemente divergem exponencialmente no tempo<sup>2</sup>. A dinâmica caótica corresponde a uma *dependência sensível* em relação às condições iniciais. O cálculo dos *expoentes de Lyapunov* quantifica a divergência exponencial entre duas órbitas.

No espaço de fase de dimensão  $N$ , cada expoente  $\lambda_j$  ( $j = 1, 2, \dots, N$ ) corresponde à evolução temporal do eixo principal  $\epsilon_j(t)$  de um hiperelipsoide que tem a forma de uma hiperesfera de diâmetro  $\epsilon_j(0)$  no instante inicial. O expoente de Lyapunov é o limite do logaritmo da razão entre o eixo principal e o respectivo diâmetro, por um longo tempo  $t$ :

$$\lambda_j = \lim_{t \rightarrow \infty} \frac{1}{t} \ln \left( \frac{\epsilon_j(t)}{\epsilon_j(0)} \right). \quad (6)$$

Os eixos principais da hiperesfera sofrem diferentes deformações ao longo do tempo, tanto de contração quanto de expansão. Isto implica, por exemplo, que uma esfera num espaço de estados tridimensional assume rapidamente a forma de um elipsoide. Uma ilustração deste comportamento no plano de fase — definido pelas variáveis dinâmicas  $X_1$  e  $X_2$  — é dada na Figura 6. Expoentes de Lyapunov positivos indicam expansão de um eixo e a contração está associada a valores negativos destes expoentes.

**Definição 10.** *Um atrator é **caótico** quando possui, para duas órbitas típicas e infinitesimalmente próximas num instante inicial, pelo menos um expoente de Lyapunov positivo.*

A direção dos eixos presentes na Equação (6) varia continuamente e em geral de uma forma complicada no atrator. Então, o cálculo de um determinado expoente de Lyapunov associado a uma orientação bem definida do hiperelipsoide no espaço de fase não faz sentido (WOLF et al., 1985).

Todavia, o conteúdo do *Teorema Ergódico Multiplicativo de Oseledec* assegura a existência dos expoentes de Lyapunov. Este teorema garante que o espectro de autovalores de uma matriz  $\mathbf{O}_{\text{SL}}$  é o mesmo para todas as condições iniciais. Isto é, o conjunto dos autovalores de  $\mathbf{O}_{\text{SL}}$  é **invariante** (OSELEDEC, 1969). Cada expoente de Lyapunov  $\lambda_j$ , correspondente a perturbação inicial unitária  $\delta|X_0\rangle_j / \|\delta|X_0\rangle_j\|$ , é definido pelo limite

---

<sup>2</sup> Estranheza de um atrator não implica necessariamente na existência de um expoente de Lyapunov positivo para as órbitas desse atrator. Existem atratores que possuem a estranheza geométrica mas não apresentam dinâmica caótica. Osciladores do tipo de pêndulos amortecidos e forçados em frequências diferentes — experimentalmente realizáveis — podem exibir estranheza sem serem caóticos (GREBOGI et al., 1984)

(OTT, 2002):

$$\lambda_j = \lim_{n \rightarrow \infty} \frac{1}{n(\Delta t)} \ln \left\| \mathbf{O}_{\text{SL}} \cdot \frac{\delta |X_0\rangle_j}{\|\delta |X_0\rangle_j\|} \right\|. \quad (7)$$

## 1.6 O problema inverso das séries temporais

Ao tratar da predição em *séries temporais caóticas*, Casdagli (1989) define que o **problema padrão** em sistemas dinâmicos é obter — a partir de um mapa não linear conhecido — o comportamento assintótico das iterações. O problema inverso, por sua vez, é construir um mapa não linear que se aproxime de uma sequência de iterações dada. Este mapa desempenha o papel de um modelo preditivo.

**Definição 11.** *O problema inverso das séries temporais é obter conhecimento a respeito de um sistema dinâmico a partir de um conjunto de observáveis ordenados no tempo.*

Especificamente, neste trabalho, a aquisição de conhecimento se dá através da predição e da caracterização da série temporal no ambiente da Computação Algébrica.

No Atrator de Lorenz exemplificado na Seção 1.4, o caminho percorrido é o mesmo do problema padrão, ou seja: a solução numérica do sistema formado pelas equações de Lorenz (8) — com as condições iniciais e parâmetros dados na Figura 5a — gera os valores de uma série temporal associada a uma determinada variável  $X_i$  ( $i = 1, 2, 3$ ).

$$\begin{aligned} \frac{dX_1}{dt} &= -\sigma X_1 + \sigma X_2 \equiv F_1 \\ \frac{dX_2}{dt} &= X_1 X_3 + r X_1 - X_2 \equiv F_2 \\ \frac{dX_3}{dt} &= X_1 X_2 - b X_3 \equiv F_3 \end{aligned} \quad (8)$$

Nestas equações:  $\sigma$ ,  $r$  e  $b$  são parâmetros dimensionais. O tempo  $t$  é a única variável independente e as funções  $F_1$ ,  $F_2$  e  $F_3$  possuem primeiras derivadas parciais contínuas.

Tendo em vista a abordagem inversa das séries temporais, *experimentos numéricos* são adequados para o teste da teoria e das rotinas computacionais. Dados de experimentos deste tipo — que exigem integração numérica — são empregados nos testes da previsibilidade e da caracterização dinâmica realizados nesta tese.

## 2 A APROXIMAÇÃO GLOBAL

Um mapeamento entre diferentes estados do sistema é o ponto de partida para tratar da previsibilidade e também realizar o diagnóstico da série quanto a presença de caos ou de uma aleatoriedade desprovida de qualquer dinâmica determinística. O mapa em questão se traduz numa função das coordenadas do espaço de fase reconstruído.

### 2.1 Previsão pela aproximação global

Sendo  $|x(t)\rangle$  um vetor de estado correspondente ao instante  $t$ , uma função  $f_\tau$  relaciona este vetor com o estado  $|x(t + \tau\Delta t)\rangle$  (FARMER; SIDOROWICH, 1987). O instante futuro é calculado pela adição do instante  $t$  com o produto de um inteiro  $\tau$  pelo intervalo de tempo  $\Delta t$ . Ou seja, uma data futura corresponde a  $t + \tau\Delta t$ . O parâmetro  $\tau$  especifica a ordem de entrada na série. Tomando como exemplo o fechamento diário do mercado de ações, a previsão para o índice três dias após uma determinada data requer  $\tau = 3$ . A proposta é encontrar funções  $\mathcal{P}_\tau$  que sejam aproximações satisfatórias de  $f_\tau$ .

$$|x(t + \tau\Delta t)\rangle = f_\tau(|x(t)\rangle) \quad (9)$$

**Definição 12.** A **Aproximação Global** é um método para a estimativa de um observável a partir de uma série temporal através da forma padrão de um preditor. A **Aproximação Global Polinomial** utiliza um polinômio como preditor <sup>3</sup>.

Face a definição apresentada, um preditor também pode ser designado como um *mapa global*.

Considerando que o último observável conhecido da série temporal é  $x(r (d_E T) \Delta t)$  (4), o *valor previsto*  $\bar{x}_{1(r+\tau)}$  resulta da aplicação do preditor  $\mathcal{P}_\tau(|x_r\rangle)$ . Trata-se da aproximação obtida para o primeiro componente do vetor de estado  $|x(t + \tau\Delta t)\rangle$ :

$$x((r + \tau) (d_E T) \Delta t) = \bar{x}_{1(r+\tau)} + \epsilon_{r+\tau}, \quad (10)$$

onde  $\epsilon_{r+\tau}$  é o *erro da predição*.

Depende da escolha do preditor e da determinação de seus parâmetros o sucesso da previsão. Aqui, cabe iniciar uma discussão a respeito da forma funcional mais apropriada. No exemplo dado na Introdução, os nove parâmetros do polinômio (1) foram determinados

---

<sup>3</sup> Nas técnicas globais de aproximação, as funções que desempenham o papel de preditores possuem uma forma padrão (CASDAGLI, 1989).

Tabela 1 - Exemplos de funções  $\phi_i(|x_r\rangle)$ 

$\phi_1$	$\phi_2$	$\phi_3$	$\phi_4$	$\phi_5$	$\phi_6$	$\phi_7$	$\phi_8$	$\phi_9$
$x_{1r}$	$x_{2r}$	$x_{3r}$	$x_{1r}^2$	$x_{1r}x_{2r}$	$x_{1r}x_{3r}$	$x_{2r}^2$	$x_{2r}x_{3r}$	$x_{3r}^2$

Legenda: As funções são as mesmas do preditor (1).

Fonte: O autor, 2017.

numa parte da série temporal que resultou num erro de apenas +0.23% para o Índice Dow Jones. A procura pela combinação ideal entre a forma da função preditiva, o número de parâmetros de ajuste e a extensão de observáveis a serem utilizados prescinde da tentativa e erro. Se os tempos de execução no computador forem pequenos, este método é viável.

No âmbito da Computação Algébrica, são diversas as possibilidades na seleção de funções. Existem, entretanto, conveniências matemáticas e computacionais na limitação da forma destas funções. Outro ponto de suma importância é o método de minimização a ser aplicado. Caso sejam escolhidos preditores com termos que não são lineares nos parâmetros de ajuste — do tipo

$$\mathcal{P}_\tau(|x_r\rangle) = c_1x_{1r} + c_2x_{2r} + c_3x_{3r} + c_4\sin(c_5x_{1r}) + c_6\sin(c_7x_{2r}) + c_8\sin(c_9x_{3r}) \quad (11)$$

— será necessário um método iterativo de minimização para encontrar os valores dos parâmetros  $c_5$ ,  $c_7$  e  $c_9$ <sup>4</sup>.

Logo, os preditores convenientes são as combinações lineares de funções dos vetores reconstruídos. Sendo o  $i$ -ésimo parâmetro  $c_i$  e a  $i$ -ésima função  $\phi_i(|x_r\rangle)$  — a Tabela 1 apresenta exemplos destas funções —, a forma funcional dos preditores com  $n$  parâmetros a serem determinados fica restrita a

$$\mathcal{P}_\tau(|x_r\rangle) = \sum_{i=1}^n c_i \phi_i(|x_r\rangle). \quad (12)$$

Levando em conta os  $M$  observáveis escolhidos para o mapeamento, o melhor estimador — não tendencioso — para o *desvio num tempo de predição*  $\sigma_\tau$  é calculado com base na variância amostral (BEVINGTON; ROBINSON, 2003). Este tipo de medida é a mais utilizada para o erro esperado nas previsões em séries temporais (KANTZ; SCHREIBER, 2003).

$$\sigma_\tau = \sqrt{\frac{\sum_{\dot{r}=1}^M \left( x_{1\dot{r}} - \sum_{i=1}^l c_i \phi_i(|x_{\dot{r}-\tau}\rangle) \right)^2}{M-1}} \quad (13)$$

<sup>4</sup> De fato, todos os *métodos de otimização não lineares* são iterativos. Os algoritmos clássicos de Levenberg-Marquardt e de Powell's Dog Leg são empregados quando funções não lineares nos parâmetros de ajuste estão presentes (MADSEN; NIELSEN; TINGLEFF, 2004. E-book).

## 2.2 Determinação do preditor

Nesta modelagem, os coeficientes do preditor resultam da determinação do vetor de parâmetros  $|c\rangle$ . Um conjunto de  $M$  vetores reconstruídos  $\{|x_{\dot{r}}\rangle\}$  toma parte na minimização da *função custo*  $\mathcal{F}(|c\rangle)$ . Somente a classe restrita de funções especificada em (12) é admitida no processo.

$$\mathcal{F}(|c\rangle) = \sum_{\dot{r}=1}^M (\epsilon_{\dot{r}})^2 = \sum_{\dot{r}=1}^M \left( x_{1\dot{r}} - \sum_{i=1}^l c_i \phi_i(|x_{\dot{r}-\tau}\rangle) \right)^2 \quad (14)$$

Os *resíduos*  $\epsilon_{\dot{r}-\tau}$  neste método são computados com base no primeiro componente do vetor  $|x_{\dot{r}}\rangle$ , sendo análogos aos erros da predição  $\epsilon_{r+\tau}$  (10).

$$\epsilon_{\dot{r}-\tau} = x(\dot{r}(d_E T) \Delta t) - x((\dot{r} - \tau)(d_E T) \Delta t) \equiv x_{1\dot{r}} - x_{1(\dot{r}-\tau)} \quad (15)$$

Assumindo que os resíduos (15) seguem uma *distribuição gaussiana*, a probabilidade para a ocorrência do resíduo  $\epsilon_{(\dot{r}-\tau)j}$  tem uma distribuição  $\Theta_j(\epsilon_{(\dot{r}-\tau)j})$  que corresponde a

$$\Theta_j = \kappa \exp\left(-\frac{\epsilon_{(\dot{r}-\tau)j}^2}{2\sigma^2}\right), \quad (16)$$

com as constantes  $\kappa$  e  $\sigma$ .

Tratando-se de densidades de probabilidades, o *método da máxima verossimilhança* atende ao propósito de estimar parâmetros (FISHER, 1922). No caso específico do preditor (12), o vetor de parâmetros  $|c\rangle$  maximiza o produto das densidades de probabilidades dos  $M$  resíduos  $\epsilon_{(\dot{r}-\tau)j}$  — que correspondem ao conjunto de vetores reconstruídos  $\{|x_{\dot{r}}\rangle\}$  empregado no ajuste. É imediato concluir que o logaritmo do produto é também maximizado pelo mesmo vetor  $|c\rangle$ .

$$0 = \frac{\partial}{\partial c_k} [\ln(\Theta_1 \Theta_2 \cdots \Theta_M)] \quad (17)$$

$$0 > \frac{\partial^2}{\partial c_k^2} [\ln(\Theta_1 \Theta_2 \cdots \Theta_M)] \quad (18)$$

Substituindo a densidade de probabilidade (16) nas relações (17) e (18) — que impõe as condições de máximo requeridas pela implementação do método de Fisher (1922) — são encontrados os vínculos entre os resíduos:

$$0 = \frac{\partial}{\partial c_k} \left[ - \left( \epsilon_{(\dot{r}-\tau)1}^2 + \epsilon_{(\dot{r}-\tau)2}^2 + \cdots + \epsilon_{(\dot{r}-\tau)M}^2 \right) \right]$$

$$0 > \frac{\partial^2}{\partial c_k^2} \left[ - \left( \epsilon_{(\dot{r}-\tau)1}^2 + \epsilon_{(\dot{r}-\tau)2}^2 + \cdots + \epsilon_{(\dot{r}-\tau)M}^2 \right) \right].$$

Uma simples multiplicação — da equação e da inequação escritas acima — por  $-1$ , conduz à quantidade de interesse na obtenção do vetor  $|c\rangle$ . A soma entre colchetes tem que ser um mínimo pelas relações (19) e (20) e é idêntica àquela presente na função custo  $\mathcal{F}(|c\rangle)$  (14). Trata-se de um típico **problema de mínimos quadrados**.

$$0 = \frac{\partial}{\partial c_k} \left[ \epsilon_{(\dot{r}-\tau)1}^2 + \epsilon_{(\dot{r}-\tau)2}^2 + \cdots + \epsilon_{(\dot{r}-\tau)M}^2 \right] \quad (19)$$

$$0 < \frac{\partial^2}{\partial c_k^2} \left[ \epsilon_{(\dot{r}-\tau)1}^2 + \epsilon_{(\dot{r}-\tau)2}^2 + \cdots + \epsilon_{(\dot{r}-\tau)M}^2 \right] \quad (20)$$

Iniciando pela derivação da função custo em relação ao parâmetro  $c_1$ , a equação

$$\begin{aligned} 0 &= \frac{\partial \mathcal{F}(|c\rangle)}{\partial c_1} = \frac{1}{2} \frac{\partial \mathcal{F}(|c\rangle)}{\partial c_1} \\ &= \frac{1}{2} \cdot 2 \cdot \sum_{\dot{r}=1}^M \left[ \left( x_{1\dot{r}} - \sum_{i=1}^l c_i \phi_i(|x_{\dot{r}-\tau}\rangle) \right) \phi_1(|x_{\dot{r}-\tau}\rangle) \right] \\ &= \sum_{\dot{r}=1}^M \left( x_{1\dot{r}} - \sum_{i=1}^l c_i \phi_i(|x_{\dot{r}-\tau}\rangle) \right) \phi_1(|x_{\dot{r}-\tau}\rangle) \\ &= \sum_{\dot{r}=1}^M x_{1\dot{r}} \phi_1(|x_{\dot{r}-\tau}\rangle) - \sum_{\dot{r}=1}^M \sum_{i=1}^l c_i \phi_i(|x_{\dot{r}-\tau}\rangle) \phi_1(|x_{\dot{r}-\tau}\rangle) \\ &= \sum_{\dot{r}=1}^M x_{1\dot{r}} \phi_1(|x_{\dot{r}-\tau}\rangle) - \sum_{i=1}^l c_i \sum_{\dot{r}=1}^M \phi_i(|x_{\dot{r}-\tau}\rangle) \phi_1(|x_{\dot{r}-\tau}\rangle) \end{aligned}$$

é construída. Então,

$$\sum_{i=1}^l c_i \left( \sum_{\dot{r}=1}^M \phi_i(|x_{\dot{r}-\tau}\rangle) \phi_1(|x_{\dot{r}-\tau}\rangle) \right) = \sum_{\dot{r}=1}^M x_{1\dot{r}} \phi_1(|x_{\dot{r}-\tau}\rangle) . \quad (21)$$

Com o intuito explicitar a relação (21), a função preditiva (1) é tomada como exemplo:

$$\begin{aligned} &c_1 \sum_{\dot{r}=1}^M x_{1(\dot{r}-\tau)} x_{1(\dot{r}-\tau)} + c_2 \sum_{\dot{r}=1}^M x_{2(\dot{r}-\tau)} x_{1(\dot{r}-\tau)} + c_3 \sum_{\dot{r}=1}^M x_{3(\dot{r}-\tau)} x_{1(\dot{r}-\tau)} \\ &+ c_4 \sum_{\dot{r}=1}^M x_{1(\dot{r}-\tau)}^2 x_{1(\dot{r}-\tau)} + c_5 \sum_{\dot{r}=1}^M x_{1(\dot{r}-\tau)} x_{2(\dot{r}-\tau)} x_{1(\dot{r}-\tau)} + c_6 \sum_{\dot{r}=1}^M x_{1(\dot{r}-\tau)} x_{3(\dot{r}-\tau)} x_{1(\dot{r}-\tau)} \\ &+ c_7 \sum_{\dot{r}=1}^M x_{2(\dot{r}-\tau)}^2 x_{1(\dot{r}-\tau)} + c_8 \sum_{\dot{r}=1}^M x_{2(\dot{r}-\tau)} x_{3(\dot{r}-\tau)} x_{1(\dot{r}-\tau)} + c_9 \sum_{\dot{r}=1}^M x_{3(\dot{r}-\tau)}^2 x_{1(\dot{r}-\tau)} \\ &= \sum_{\dot{r}=1}^M x_{1\dot{r}} x_{1(\dot{r}-\tau)} . \end{aligned}$$

Prosseguindo com o mesmo procedimento de derivação e de manipulação algébrica que conduziu à Equação (21), são encontradas  $l$  equações correspondentes aos parâmetros de ajuste. O sistema de equações

$$\left\{ \begin{array}{l} \sum_{i=1}^l c_i \left( \sum_{\dot{r}=1}^M \phi_i(|x_{\dot{r}-\tau}\rangle) \phi_1(|x_{\dot{r}-\tau}\rangle) \right) = \sum_{\dot{r}=1}^M x_{1\dot{r}} \phi_1(|x_{\dot{r}-\tau}\rangle) \\ \sum_{i=1}^l c_i \left( \sum_{\dot{r}=1}^M \phi_i(|x_{\dot{r}-\tau}\rangle) \phi_2(|x_{\dot{r}-\tau}\rangle) \right) = \sum_{\dot{r}=1}^M x_{1\dot{r}} \phi_2(|x_{\dot{r}-\tau}\rangle) \\ \vdots \\ \sum_{i=1}^l c_i \left( \sum_{\dot{r}=1}^M \phi_i(|x_{\dot{r}-\tau}\rangle) \phi_l(|x_{\dot{r}-\tau}\rangle) \right) = \sum_{\dot{r}=1}^M x_{1\dot{r}} \phi_l(|x_{\dot{r}-\tau}\rangle) \end{array} \right. \quad (22)$$

tem como solução o conjunto  $\{c_1, c_2, \dots, c_l\}$ . Definindo

$$\alpha_{pq} \equiv \sum_{\dot{r}=1}^M \phi_p(|x_{\dot{r}-\tau}\rangle) \phi_q(|x_{\dot{r}-\tau}\rangle)$$

$$\beta_p \equiv \sum_{\dot{r}=1}^M x_{1\dot{r}} \phi_p(|x_{\dot{r}-\tau}\rangle),$$

o sistema (22) é colocado na forma matricial e o problema de determinar o preditor fica reduzido à solução da equação

$$\begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1l} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{l1} & \alpha_{l2} & \dots & \alpha_{ll} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_l \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_l \end{bmatrix}. \quad (23)$$

O desenvolvimento desta seção conduz a um formato de solução próprio da **Álgebra Linear**. Determinar o preditor significa encontrar a solução de mínimos quadrados  $|c\rangle$  para um sistema linear do tipo  $\hat{\alpha}|c\rangle = |\beta\rangle$ . Esta solução minimiza  $\|\hat{\alpha}|c\rangle - |\beta\rangle\|^2$  (ROBINSON, 2006).

### 2.3 Distribuição dos resíduos

No processo que conduz ao método dos mínimos quadrados, a distribuição dos resíduos (16) é normal. Assumir a forma gaussiana no ajuste, contudo, pode não estar de

acordo com a combinação real entre a parte da série temporal e o preditor selecionados. Em situações como esta, outros métodos de minimização poderiam ser propostos.

Para fazer frente a medidas realizadas a partir de observações astronômicas, Branham R. L. (1982) apresenta um estudo comparativo entre os métodos das *médias*, de *Chebyshev* e da *soma mínima*. A motivação é que o método dos mínimos quadrados pode não ser o mais adequado se os erros de observação não estiverem normalmente distribuídos. Neste caso, não há a garantia do *Teorema de Gauss-Markov* para que a variância estimada pelos mínimos quadrados seja mínima. Na conclusão do trabalho, o método da soma mínima foi o único que se mostrou competitivo com o tradicional método dos mínimos quadrados, merecendo maior atenção por parte dos astrônomos. Os outros dois métodos mostraram-se inadequados para o tratamento dos dados observacionais.

**Teorema 6. Teorema de Gauss-Markov** (SCHEFFÉ, 1959; SILVEY, 1975). *Seja um vetor  $|\beta\rangle$  de  $p$  dimensões definido pela Equação  $|\beta\rangle = \hat{\alpha}|c\rangle + |\epsilon\rangle$ , onde  $\hat{\alpha}$  é uma matriz conhecida  $p \times l$  de posto  $l$ ,  $|c\rangle$  é um vetor desconhecido com  $l$  dimensões e  $|\epsilon\rangle$  é o **vetor erro** com valor esperado  $E(|\epsilon\rangle) = |0\rangle$  e variância  $\hat{\sigma}^2(|\epsilon\rangle) = \sigma^2\hat{I}$ , onde  $\sigma^2$  é desconhecida e  $\hat{I}$  é a matriz identidade; isto é, os componentes de  $|\epsilon\rangle$  **possuem a mesma variância  $\sigma^2$  e não são correlacionados**.*

*Seja  $|\tilde{c}\rangle$  o único estimador de mínimos quadrados de  $|c\rangle$  e  $\Omega = \xi|c\rangle$  uma função linear paramétrica. Então  $\xi|\tilde{c}\rangle$  é um estimador não tendencioso de  $\Omega$  e, se  $\tilde{\Omega}$  é qualquer outro estimador linear não tendencioso de  $\Omega$ , temos que*

$$\hat{\sigma}^2(\xi|\tilde{c}\rangle) \leq \hat{\sigma}^2(\tilde{\Omega}).$$

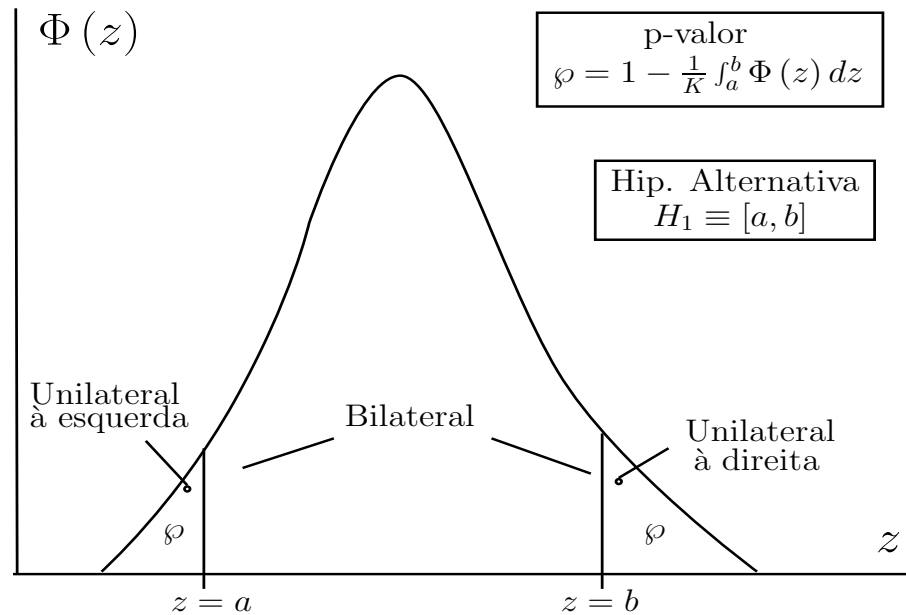
Com o objetivo de obter uma aproximação verossímil na solução de mínimos quadrados (23), a estratégia a ser seguida é avaliar se convém, ou não, assumir a distribuição dos resíduos num determinado mapeamento como gaussiana. Um *teste de normalidade da distribuição* permite a tomada de decisão a partir de evidências estatísticas.

## 2.4 Teste de hipóteses e o p-valor

Na Estatística, frequentemente se extrai informações sobre um conjunto inteiro, a *população*, através da investigação de um pequeno subconjunto deste, a *amostra*. Como consequência, conclusões erradas podem ser tiradas a respeito da população. Na metodologia que envolve a tomada de uma *decisão estatística*, geralmente são formuladas hipóteses sobre as populações.

**Definição 13.** A **Hipótese Nula**  $H_0$  é a afirmação feita sobre a população.

Figura 7 - Teste de hipóteses e cálculo do p-valor



Legenda: Teste unilateral:  $a = -\infty$  e  $b$  finito — à direita;  $a = \text{finito}$  e  $b = +\infty$  — à esquerda. Teste bilateral:  $a$  e  $b$  finitos.

Fonte: O autor, 2017.

**Definição 14.** Uma proposição divergente da Hipótese Nula constitui uma **Hipótese Alternativa**  $H_1$ .

Para aceitar uma hipótese, ou rejeitá-la, são aplicados *testes de hipóteses*. Quando uma hipótese nula que se aproxima bem da realidade da população é rejeitada no teste em favor da hipótese alternativa — ou seja, se **ocorre rejeição indevida** de  $H_0$  em favor de uma aproximação pior  $H_1$  — o erro cometido é do *Tipo I*. Se **não ocorre a devida rejeição** de  $H_0$  em favor de  $H_1$  — na situação em que a hipótese nula é menos representativa que a alternativa — então a decisão estatística implica num erro do *Tipo II*.

A hipótese nula habitualmente é formulada com relação a um parâmetro da população. Por exemplo, um valor especificado para a média. Este parâmetro está associado a uma estatística  $z$  — que pode ser a *variável normal padronizada*. Cabe ressaltar que a quantidade a ser manipulada no teste de hipóteses não é o parâmetro populacional, mas a sua estatística correspondente. Se a afirmação da hipótese alternativa for superior ao valor especificado pela hipótese nula para  $z$ , então trata-se de um *teste unilateral à direita*. No outro sentido, o do *teste unilateral à esquerda*, a estatística tem um valor menor do que o estabelecido para a hipótese nula. A proposição de que a estatística é apenas diferente — com um valor de  $z$  maior ou menor — corresponde ao *teste bilateral*.

Existe uma quantidade chamada de *p-valor* que permite avaliar a intensidade da aproximação ou do afastamento dos observáveis em relação à hipótese. No critério de

avaliação do teste de hipóteses, o p-valor precisa estar num intervalo considerado aceitável.

**Definição 15.** *Seja a Hipótese Nula  $H_0$  e a Hipótese Alternativa  $H_1$  especificada pelo intervalo  $[a, b]$ . No teste de hipóteses para a estatística  $z$ , com distribuição de probabilidade  $\Phi(z)$ , o **p-valor**  $\wp$  é determinado por*

$$\wp = 1 - \frac{1}{K} \int_a^b \Phi(z) dz, \quad (24)$$

com a constante de normalização  $K = \int_{-\infty}^{+\infty} \Phi(z) dz$ .

Nos testes unilaterais, um dos limites do intervalo  $[a, b]$  tem magnitude infinita. O teste unilateral à direita corresponde a  $a = -\infty$  e o unilateral à esquerda é realizado fazendo  $b = +\infty$ . Valores finitos para  $a$  e  $b$  caracterizam os testes bilaterais. A Figura 7 sumariza o teste de hipóteses e o respectivo cálculo do p-valor nestes termos.

Com a intenção de interpretar o significado do p-valor, uma distribuição de probabilidade idealizada traz o esclarecimento desejado. Supondo que existe completa certeza de que a estatística  $z$  — associada a um determinado parâmetro populacional — tem o valor  $z_0$ , a distribuição de probabilidade corresponde à função delta de Dirac  $\delta(z - z_0)$ . Escolhendo um intervalo para a hipótese alternativa com  $b < z_0$  ou, por outro lado,  $a > z_0$ , o cálculo

$$\begin{aligned} \wp &= 1 - \frac{1}{K} \int_a^b \delta(z - z_0) dz = 1 - \frac{1}{\int_{-\infty}^{+\infty} \delta(z - z_0) dz} \int_a^b \delta(z - z_0) dz \\ &= 1 - 0 = 1 \end{aligned}$$

tem como resultado um p-valor igual a um. Rejeitar a hipótese nula em favor da hipótese alternativa significa estar completamente errado, porque  $z_0 \notin [a, b]$ , por definição. Tal decisão estatística implicaria num erro do Tipo I. Este caso ilustra que a atitude apropriada com um alto p-valor para o teste — a partir da amostra — é a aceitação da hipótese concebida a respeito da população.

Ampliando o intervalo anterior de forma que  $z_0 \in [a', b']$ , chega-se ao p-valor do extremo oposto:

$$\begin{aligned} \wp &= 1 - \frac{1}{K} \int_{a'}^{b'} \delta(z - z_0) dz = 1 - \frac{1}{\int_{-\infty}^{+\infty} \delta(z - z_0) dz} \int_{a'}^{b'} \delta(z - z_0) dz \\ &= 1 - 1 = 0. \end{aligned}$$

Neste exemplo, a hipótese nula deve ser rejeitada em favor do intervalo alternativo para os valores da estatística  $z$ . Um p-valor baixo indica que existe evidência estatística para a rejeição da hipótese nula em favor da alternativa. A Tabela 2 apresenta o panorama do teste de hipóteses desta discussão.

**Definição 16.** *O **nível de significância**  $\alpha$  é o limite inferior do p-valor, fixado previamente para o teste de hipóteses, para a aceitação da hipótese nula  $H_0$ .*

Tabela 2 - Decisão estatística e p-valor

p-valor $\wp$	decisão estatística
alto	aceitar $H_0$
baixo	rejeitar $H_0$ em favor de $H_1$

Legenda: Critério de aceitação de  $H_0$ :  $\wp \geq \alpha$ .

Fonte: O autor, 2017.

A quantificação do p-valor dá uma ideia de quão fortemente os observáveis da amostra podem se opor à hipótese feita sobre a população. Está plenamente adequada aos problemas científicos. Um valor fixo do nível de significância  $\alpha$  é inevitável na decisão pela aceitação ou rejeição de uma hipótese (LEHMANN, 2005).

## 2.5 Teste de Shapiro-Wilk

Nos testes de normalidade de uma distribuição, formula-se a hipótese de que a população segue a distribuição normal. Dentre os testes disponíveis, a opção pelo de Shapiro-Wilk tem a justificativa da sua performance superior em relação aos de Kolmogorov-Smirnov, Lilliefors e Anderson-Darling na comparação de Razali e Wah (2011).

Considerando uma amostra ordenada de uma população  $(y_1, y_2, \dots, y_n) — y_1 < y_2 < \dots < y_n$  — com média  $\bar{y}$ , determina-se a estatística  $W$  do teste de Shapiro-Wilk pela razão entre uma combinação apropriada dos elementos desta amostra e a habitual estimativa da variância (SHAPIRO; WILK, 1965).

$$W = \frac{(\sum_{i=1}^n a_i y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (25)$$

O numerador corresponde ao produto da melhor estimativa *não tendenciosa* da variância amostral por  $n-1$ . Cada um dos coeficientes  $a_i$  pode ser encontrado exatamente. Todavia, considerável esforço computacional e um uso exagerado de memória são necessários nas manipulações matriciais exigidas no cálculo. Existem fórmulas que fornecem resultados que se aproximam satisfatoriamente dos coeficientes exatos (ROYSTON, 1992).

Para o cálculo do p-valor obtém-se a variável  $w$  pela normalização da estatística  $W$ . Na composição da variável normal padronizada são feitas estimativas — que dependem do tamanho da amostra — para a média  $\mu$  e para o desvio padrão  $\sigma$ . Royston (1992) apresenta as fórmulas aproximativas e encontra o p-valor  $\wp = 0.018$  para o caso em que  $w = -1.3779$ ,  $\mu = -1.6198$  e  $\sigma = 0.11544$ . Relacionado a este exemplo, há um ponto a favor da aproximação para os coeficientes  $a_i$  no mesmo trabalho: o resultado da

estatística  $W$  calculado pelos coeficientes exatos e os aproximados coincidem até a quarta casa decimal.

$$b = \frac{w - \mu}{\sigma} = \frac{-1.3779 - (-1.6198)}{0.11544} = 2.095$$

$$\wp = 1 - \frac{1}{K} \int_a^b \Phi(z) dz = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{2.095} e^{-\frac{z^2}{2}} dz = 0.0180854970$$

Considerando os níveis de significância  $\alpha = 0.01$  e  $\alpha = 0.05$ , o p-valor  $\wp = 0.018$  implicaria na aceitação da hipótese nula no nível mais baixo (0.01), e na sua rejeição no mais alto (0.05).

Voltando ao problema dos resíduos na determinação do preditor, convém fazer uma consideração sobre o que se pode esperar do teste de normalidade. O conjunto dos resíduos calculados pode ser considerado como uma amostra retirada de uma determinada população. Se pelo teste de Shapiro-Wilk houver evidência estatística de que esta população segue uma distribuição normal, então a forma gaussiana (16) é uma aproximação válida para a distribuição dos resíduos no mapeamento global. Nesse contexto, as hipóteses do teste no método da aproximação global são formuladas:

$H_0^{GA} \equiv$  Amostra retirada de uma população que segue uma distribuição normal;

$H_1^{GA} \equiv$  Amostra retirada de uma população que não segue uma distribuição normal.<sup>5</sup>

## 2.6 Nível de confiança na predição

Encontrar preditores cuja distribuição dos resíduos no mapeamento seja bem aproximada por uma função gaussiana está no escopo do método desenvolvido neste trabalho. Do ponto de vista prático, existe a possibilidade de modificação da forma do preditor — a mudança do grau do polinômio de dois para três, por exemplo — ou da extensão da parte da série que fará parte do *global fitting*, caso o resultado do teste de normalidade seja negativo. Mais uma vez, operações deste tipo são facilitadas no ambiente da Computação Algébrica até que o resultado do teste seja positivo. Com esta motivação, uma formulação precisa para o nível de confiança mostra-se adequada.

**Definição 17.** *O nível de confiança para um tempo de predição  $\mathcal{L}_\tau$  é o resultado do cálculo*

$$\mathcal{L}_\tau = 1 - \frac{2}{\sqrt{2\pi}} \int_{\epsilon_{r+\tau}/\sigma_\tau}^{\infty} \exp\left\{-\frac{\epsilon_{r-\tau}^2}{2\sigma_\tau^2}\right\} d(\epsilon_{r-\tau}/\sigma_\tau), \quad (26)$$

---

<sup>5</sup> o rótulo “GA” abrevia *global approach*

em que o desvio  $\sigma_\tau$  (13) corresponde à raiz quadrada da variância amostral.

Se os resíduos no mapeamento forem aceitos no teste de Shapiro-Wilk, então existe uma forte razão para se esperar que o valor verdadeiro da série temporal — que é desconhecido, por hipótese — esteja na faixa entre  $x_{1P(r+\tau)} - \sigma_\tau$  e  $x_{1P(r+\tau)} + \sigma_\tau$  com uma confiança de 68.3%, de 95.5% para  $\epsilon_{r+\tau} = 2\sigma_\tau$  e 99.7% quando  $\epsilon_{r+\tau} = 3\sigma_\tau$ . Calculando:

$$1 - \frac{2}{\sqrt{2\pi}} \int_1^\infty \exp\left\{-\frac{\epsilon_{r-\tau}^2}{2\sigma_\tau^2}\right\} d(\epsilon_{r-\tau}/\sigma_\tau) \cong 0.683,$$

$$1 - \frac{2}{\sqrt{2\pi}} \int_2^\infty \exp\left\{-\frac{\epsilon_{r-\tau}^2}{2\sigma_\tau^2}\right\} d(\epsilon_{r-\tau}/\sigma_\tau) \cong 0.955,$$

$$1 - \frac{2}{\sqrt{2\pi}} \int_3^\infty \exp\left\{-\frac{\epsilon_{r-\tau}^2}{2\sigma_\tau^2}\right\} d(\epsilon_{r-\tau}/\sigma_\tau) \cong 0.997.$$

Estes cálculos expressam exatamente o conteúdo da **Regra dos Três Sigmas** (PUKELSHEIM, 1994). A magnitude do *erro tolerável na predição* pode ser estabelecida como sendo de  $3\sigma_\tau$ , com significado idêntico ao habitualmente adotado na Física Experimental.

### 3 CARACTERIZAÇÃO DINÂMICA

A evolução temporal dos desvios calculados com a aproximação global informa a natureza dinâmica do sistema subjacente à série temporal. O método apresentado neste capítulo resulta numa alternativa ao cálculo dos expoentes de Lyapunov a partir do conjunto de observáveis. Comportamentos típicos deixam sua impressão em gráficos que permitem diagnosticar séries como caóticas, periódicas ou randômicas.

#### 3.1 Expoentes de Lyapunov a partir de séries temporais

Quando um sistema apresenta uma dinâmica caótica, pelo menos um autovalor da matriz de Oseledec deve ser positivo. Sendo assim, basta determinar o maior expoente de Lyapunov  $\lambda_{max}$  para atender ao propósito da caracterização dinâmica. Caso seja positivo, o sistema pode ser considerado caótico de acordo com a **Definição 10** — devida a Grebogi et al. (1984) — da Seção 1.5.

Existem algoritmos para a determinação de  $\lambda_{max}$ . No trabalho pioneiro desta proposta, Wolf et al. (1985) assumem a existência de uma divergência exponencial entre órbitas infinitesimalmente próximas e o cálculo baseia-se nesta hipótese. Infelizmente, a sua aplicação requer muito cuidado e facilmente pode conduzir a resultados errados. No algoritmo de Sano e Sawada (1985), uma performance satisfatória pode ser obtida quando os observáveis permitem uma boa aproximação da dinâmica do sistema. Testes da divergência exponencial de trajetórias próximas são realizadas com o método de Rosenstein, Collins e Luca (1993). Kantz (1994) também disponibiliza um algoritmo com esta característica. Os resultados dos testes permitem decidir quando realmente faz sentido o cálculo dos expoentes de Lyapunov a partir de um conjunto de observáveis ordenados no tempo (KANTZ; SCHREIBER, 2003).

Algoritmos que determinam os demais expoentes foram aplicados por Sano e Sawada (1985), além de Eckmann et al. (1986); seu desempenho depende fortemente da qualidade dos observáveis (KANTZ; SCHREIBER, 2003). Com relação a iniciativas deste tipo, uma discussão é crucial. O conjunto dos expoentes de Lyapunov para o sistema dinâmico (3) — definido por  $N$  equações diferenciais — corresponde ao espectro de autovalores  $\lambda_i (i = 1, 2, \dots, N)$  (7) da matriz de Oseledec. Se o espaço de reconstrução possuir uma dimensão  $d_E \neq N$ , então a matriz correspondente possuirá um número de autovalores incompatível com a dimensão real do sistema.

Kantz e Schreiber (2003) descrevem uma aplicação do seu algoritmo que ilustra muito bem o problema. Trata-se do espectro de Lyapunov determinado a partir de uma série temporal formada pelas intensidades de um *laser de Ressonância Magnética Nuclear*

Tabela 3 - Expoentes de Lyapunov espúrios

$m$	$k$	$\lambda_{\text{dubious}}$	$\lambda_+$	$\lambda_{\text{dubious}}$			$\lambda_-$		
3	20		0.32	-0.40			-1.13		
	40		0.30	-0.51			-1.21		
	160		2.28	-0.68			-1.31		
4	20		0.34	-0.03	-0.49		-1.08		
	40		0.31	-0.01	-0.52		-1.12		
	160		0.29	-0.03	-0.69		-1.35		
5	20		0.36	0.16	-0.20	-0.57		-1.11	
	40		0.35	0.14	-0.21	-0.59		-1.14	
	160		0.31	0.13	-0.35	-0.77		-1.34	
6	20	0.39	0.24	-0.02	-0.26	-0.58		-1.09	
	40	0.41	0.25	-0.02	-0.27	-0.64		-1.18	
	160	0.38	0.25	-0.17	-0.44	-0.83		-1.34	
7	20	0.42	0.27	0.08	-0.09	-0.28	-0.57		-1.06
	40	0.45	0.27	0.11	-0.12	-0.33	-0.65		-1.16
	160	0.40	0.25	-0.02	-0.24	-0.50	-0.85		-1.38
3	global <sup>a</sup>		0.272	-0.64			-1.31		
3	global <sup>b</sup>		0.273	-0.64			-1.15		

Legenda: O espaço reconstruído para o laser NMR tem dimensão  $m$  e são usadas  $k$  medidas de intensidade no processo de *data fitting*. Os expoentes da região sombreada são claramente espúrios.

Fonte: KANTZ, SCHREIBER, 2003, p. 209.

— laser NMR —, operado de maneira que as amplitudes apresentem uma evolução temporal complexa. A Tabela 3 mostra os expoentes de Lyapunov para diferentes dimensões de reconstrução — designadas por  $m$  na tabela. O sistema que dá origem aos observáveis admite uma modelagem plenamente satisfatória com três equações de primeira ordem, ou seja, pode ser colocado na forma do sistema (3) fazendo  $N = 3$ . São considerados, a priori, *expoentes espúrios* aqueles que estão na região sombreada, num espaço com dimensão mínima igual a cinco. Alguns outros valores apresentados também precisam ser investigados mais profundamente.

Embora existam critérios para a interpretação dos expoentes como espúrios ou não, o problema continua a ser de difícil solução (PARLITZ, 1992). O fato é que as especificidades requeridas neste tipo de cálculo estão longe de serem triviais. Na procura por um caminho, cabe questionar se vale a pena prosseguir com este tipo de iniciativa. Por outro lado, a construção de um novo método que reconheça a natureza da série — caótica ou não — parece ser uma alternativa viável no escopo da **Análise de Séries Temporais**.

Tabela 4 - Séries temporais de naturezas distintas

caracterização	$\mathcal{A}_\tau$	$\mathcal{D}_\tau$
caótica	decrecente	crescente
randômica	sempre baixa	sempre baixo
periódica	sempre alta	sempre baixo

Legenda: As quantidades estatísticas acurácia  $\mathcal{A}_\tau$  e desvio relativo  $\mathcal{D}_\tau$  são responsáveis pela caracterização dinâmica das séries temporais.

Fonte: O autor, 2017.

### 3.2 A caracterização dinâmica pela Aproximação Global

Os desvios num tempo de predição  $\sigma_\tau$  — calculados para um determinado preditor — decorrem da natureza do sistema dinâmico que gera os observáveis da série temporal. Casdagli (1989) propõe uma lei de escala baseada no maior expoente de Lyapunov:

$$\sigma_\tau = \sigma_1 e^{\tau \Delta t \lambda_{max}} . \quad (27)$$

Se os desvios  $\sigma_\tau$  entre o valor real  $|x(t + \tau \Delta t)\rangle$  e o resultado da predição — dado pelo valor de  $\mathcal{P}_\tau(|x\rangle)$  — crescem exponencialmente para  $\tau \rightarrow +\infty$ , então o sistema é caótico. Por outro lado, na ausência de caos não existe nenhuma razão para haja crescimento dos desvios com o transcorrer do tempo, como é o caso das séries temporais completamente aleatórias e também das séries periódicas. Entretanto, a acurácia na predição em sistemas periódicos é muito maior que a obtida em séries aleatórias.

Neste trabalho, duas quantidades estatísticas são construídas para caracterizar o comportamento dinâmico da série temporal. A razão entre a média dos valores absolutos e o erro admissível na predição recebe o nome de *acurácia para um tempo de predição*, denotada por  $\mathcal{A}_\tau$ .

$$\mathcal{A}_\tau = \frac{\sum_{\hat{r}=1}^{M-1} |x_{1\hat{r}}|}{3(M-1)\sigma_\tau} \quad (28)$$

Os módulos acima são os primeiros componentes dos vetores que participam do mapeamento global.

O outro quantificador criado — chamado de *desvio relativo para um tempo de predição*, designado por  $\mathcal{D}_\tau$  — compara o desvio  $\sigma_\tau$  calculado para o preditor  $\mathcal{P}_\tau(|x_r\rangle)$

Tabela 5 - Critérios de qualificação para os preditores

critério	descrição
<b>QP<sub>1</sub></b>	O <i>global fitting</i> é aceitável se $\sigma_\tau < 0.003(3\eta_1 + 1)  x_{1\tau} _{\max}$ .
<b>QP<sub>2</sub></b>	O <i>global fitting</i> é aceitável se $\eta_2 < 0.317N$ .
<b>QP<sub>3</sub></b>	O <i>global fitting</i> é aceitável se $\eta_3 < 0.045N$ .
<b>QP<sub>4</sub></b>	O <i>global fitting</i> é aceitável se $\eta_4 = 0$ .
<b>QP<sub>5</sub></b>	O <i>global fitting</i> é aceitável se $\eta_5 < 0.317N(\eta_1 + 1)$ .

Legenda: Os critérios acima são complementares. Caso um tipo de *outlier* não esteja no escopo de um determinado **QP**, ele pode ainda ser detectado por qualquer um dos quatro critérios remanescentes.

Fonte: O autor, 2017.

com aquele que corresponde a  $\tau = 1$ .

$$\mathcal{D}_\tau = \frac{\sigma_\tau}{\sigma_1} \quad (29)$$

Se os tempos de predição são longos, os sistemas caóticos tornar-se-ão imprevisíveis. Por outro lado, nas séries aleatórias a acurácia necessariamente será baixa; não há dependência do valor assumido pelo parâmetro  $\tau$ . Nisto reside a diferença com as séries caóticas, que permitem previsões acuradas para  $\tau = 1$ .

Nos sistemas periódicos, os desvios relativos  $\mathcal{D}_\tau$  são baixos todo o tempo, da mesma maneira que nos randômicos. A grande distinção entre os dois tipos de dinâmica está justamente nos valores de  $\mathcal{A}_\tau$ . Para qualquer tempo de predição permanecerá alta nas séries periódicas. A Tabela 4 resume a discussão em termos do comportamento esperado para as quantidades estatísticas.

### 3.3 Qualificação dos preditores

Os valores obtidos pela aplicação da função preditiva devem estar de acordo com os observáveis da série temporal. Então, alguns critérios para o controle de qualidade do mapeamento global constituem um guia útil no método de caracterização dinâmica. A Tabela 5 mostra os requisitos que os preditores precisam atender. Cada critério de qualificação tem o propósito de detectar uma modalidade de *outlier*. Existe uma complementariedade no conjunto de critérios. Assim, se um determinado tipo de *outlier* não for detectável pelo critério **QP<sub>1</sub>**, ele pode ser detectado com **QP<sub>2</sub>** e assim por diante.

Na aplicação dos critérios, os parâmetros do conjunto  $\{\eta_1, \eta_2, \eta_3, \eta_4, \eta_5\}$  (ver Tabela

Tabela 6 - Parâmetros para a qualificação dos preditores

parâmetro	quantidade especificada pelo parâmetro
$\eta_1$	nível de detecção de <i>outliers</i>
$\eta_2$	número de valores absolutos $ \epsilon_{\hat{r}} $ maiores que $\sum_{r=1}^N  x_{1\hat{r}}  / 3N$
$\eta_3$	número de valores absolutos $ \epsilon_{\hat{r}} $ maiores que $\sum_{r=1}^N  x_{1\hat{r}}  / N$
$\eta_4$	número de valores absolutos $ \epsilon_{\hat{r}} $ maiores que $ x_{1\hat{r}} _{\max}$
$\eta_5$	número de valores absolutos $ \epsilon_{\hat{r}} $ maiores que $\sigma_\tau$

Legenda: Os  $N$  observáveis contabilizam os vetores reconstruídos que participam e também aqueles que ficam de fora do mapeamento global.

Fonte: O autor, 2017.

6) são quantificadores para a detecção de *outliers* (em fase de elaboração)<sup>6</sup>. Uma escolha do parâmetro  $\eta_1$  define o nível de exigência para o processo de *global fitting*. Se  $\eta_1 = 0$ , somente desvios  $\sigma_\tau < 0.003 \times |x_{1\hat{r}}|_{\max}$  são admissíveis. A Regra dos Três Sigmas é empregada como um guia para a qualificação dos preditores (PUKELSHEIM, 1994).

É importante ressaltar que os vetores que não participam do processo de minimização — cujos primeiros componentes são os observáveis contidos na parte da série selecionada para o mapeamento — são também levados em conta nesta estatística. Em outras palavras, todo o intervalo da série temporal é considerado na avaliação.

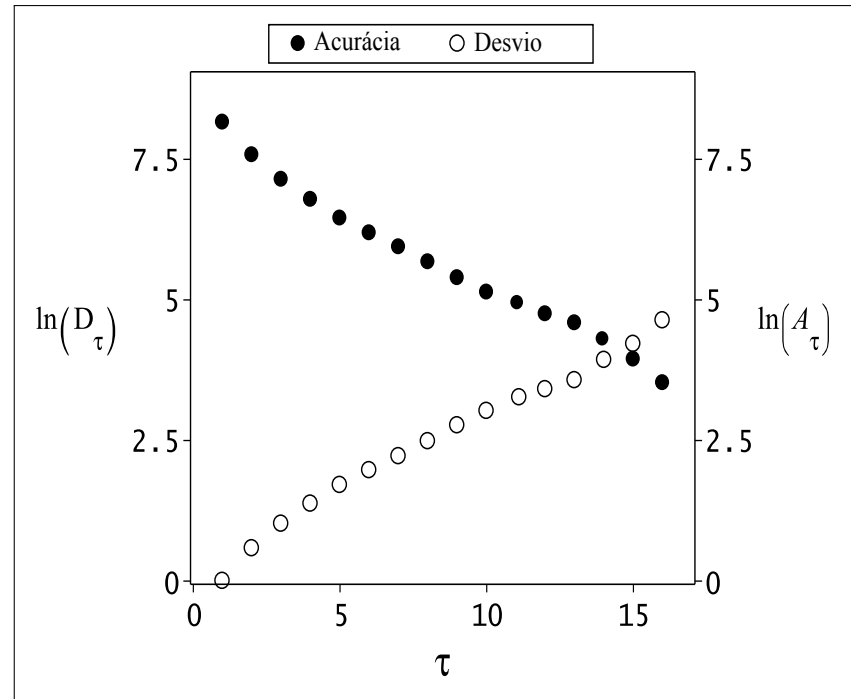
### 3.4 Diagramas Acurácia-Desvio e Acurácia-Desvio Logarítmico

A característica dinâmica de uma série temporal se manifesta em duas representações gráficas especialmente criadas para este fim, chamadas de *Diagrama Acurácia-Desvio* e *Diagrama Acurácia-Desvio Logarítmico*. Neste tipo de diagrama, uma “bola fechada” • representa o valor de  $\mathcal{A}_\tau$  — ou então  $\ln(\mathcal{A}_\tau)$  — e uma “bola aberta” ○ está associada a outra quantidade  $\mathcal{D}_\tau$  — ou ao seu logaritmo. O eixo das abcissas corresponde ao parâmetro  $\tau$ , utilizado nos cálculos de  $\mathcal{A}_\tau$  e  $\mathcal{D}_\tau$ .

Curvas crescentes para  $\mathcal{A}_\tau$  — ou  $\ln(\mathcal{A}_\tau)$  — e, ao mesmo tempo, decrescentes para  $\mathcal{D}_\tau$  — ou  $\ln(\mathcal{D}_\tau)$  —, caracterizam o comportamento caótico. A Figura 8 mostra um diagrama deste tipo em escala logarítmica para uma série temporal formada pela variável

<sup>6</sup> ALVES, P.; DUARTE, L.; da MOTA, L. Dynamical characterization of a time series by the polynomial global approach. Submetido à publicação em 2017.

Figura 8 - Diagrama Acurácia-Desvio Logarítmico



Legenda: A série temporal corresponde à variável  $X$  do Sistema de Lorenz.

Fonte: O autor, 2017.

$X$  do Sistema de Lorenz (8).

Se a série for periódica, a expectativa é a formação de uma *banda de acurácia* na parte superior do diagrama e uma *banda de desvio* na parte inferior. No caso de séries randômicas estas duas bandas deverão ser localizadas na parte inferior do diagrama. Os três tipos de possível caracterização dinâmica estão identificados na Tabela 7.

### 3.5 Um novo quantificador para o caos

As curvas dos Diagramas Acurácia-Desvio (em fase de elaboração)<sup>7</sup> informam o tipo da dinâmica responsável pela evolução do observável contido na série temporal. Então, este tipo de representação gráfica é adequado também para a extração de uma informação quantitativa sobre a natureza do sistema em estudo. A manipulação da quantidade estatística  $\mathcal{A}_\tau$  permite a construção de um quantificador que atenda a este propósito.

Em um diagrama para uma série caótica — como aquele representado na Figura

<sup>7</sup> ALVES, P.; DUARTE, L.; da MOTA, L. Dynamical characterization of a time series by the polynomial global approach. Submetido à publicação em 2017.

Tabela 7 - Caracterização dinâmica com o Diagrama Acurácia-Desvio Logarítmico

caracterização	$\ln(\mathcal{A}_\tau)$	$\ln(\mathcal{D}_\tau)$	$\ln(\mathcal{A}_1)$	$\lambda_{dyn}$
caótica	curva decrescente	curva crescente	$+\infty$	$> 0$
randômica	banda inferior	banda inferior	0	0
periódica	banda superior	banda inferior	$+\infty$	0

Legenda: A escala logarítmica facilita o diagnóstico da série temporal. Os valores de referência acima correspondem a mapeamentos globais ideais.

Fonte: O autor, 2017.

8 —, a curva  $\ln(\mathcal{A}_\tau)$  sugere um decaimento exponencial do tipo

$$\ln(\mathcal{A}_\tau) = \{\ln(\mathcal{A}_1)\} e^{-\acute{\lambda}\tau}. \quad (30)$$

A constante  $\acute{\lambda}$  deve ser positiva se o observável apresenta uma evolução caótica.

Tanto  $\ln(\mathcal{A}_1)$  quanto  $\acute{\lambda}$  descrevem aspectos que marcam a natureza da dinâmica. De acordo com a discussão a respeito dos típicos Diagramas Acurácia-Desvio Logarítmicos, existe a expectativa que os valores de  $\ln(\mathcal{A}_1)$  manifestem características de periodicidade ou aleatoriedade (no caso da série temporal não apresentar determinismo). Por outro lado, nos dois casos, os valores de  $\acute{\lambda}$  devem ser próximos de zero.

Neste estágio da construção, uma boa estratégia consiste no desacoplamento destas duas quantidades. No lugar de determinar uma constante do tipo de  $\acute{\lambda}$ , é mais conveniente fazer a troca

$$\acute{\lambda} \rightarrow \lambda_{dyn}$$

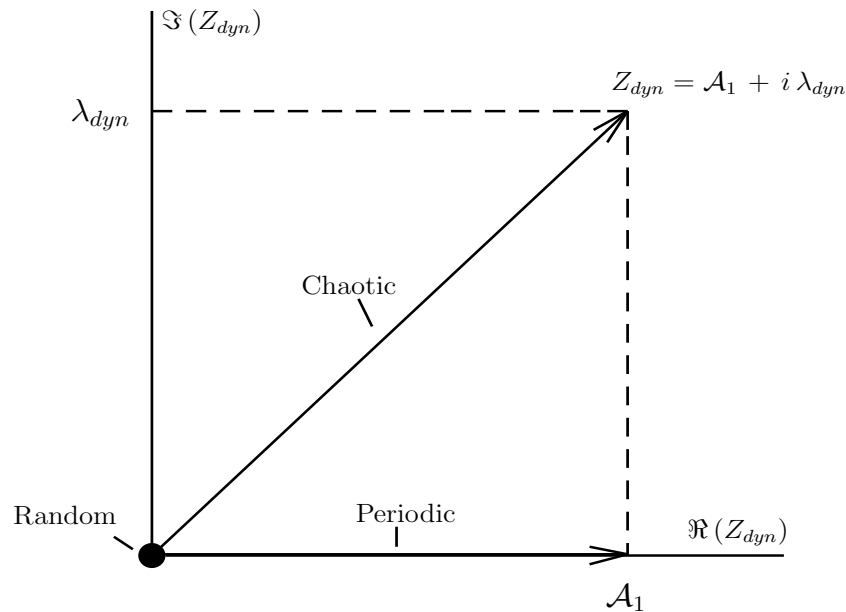
porque a Equação (30) pode resultar num ajuste ruim para a curva  $\ln(\mathcal{A}_\tau)$ . Definindo  $\lambda_{dyn}$  como uma média,

$$\lambda_{dyn} = \sum_{\tau=1}^{\zeta} \frac{\mathcal{A}_\tau - \mathcal{A}_{\tau+1}}{\zeta} \quad (31)$$

o tipo de evolução temporal pode ser reconhecido mais adequadamente. Esta quantidade não depende da forma funcional de  $\mathcal{A}_\tau$  e leva em conta as inevitáveis flutuações estatísticas presentes no mapeamento global. Para preditores ideais, as magnitudes de interesse apresentam os valores da Tabela 7.

Não existe dependência entre  $\mathcal{A}_1$  e  $\lambda_{dyn}$ . Logo, ambos admitem uma representação conjunta nos eixos ortogonais de um plano cartesiano. Lançando mão da conveniência

Figura 9 - Quantificador de caos no plano complexo



Legenda: As dinâmicas típicas estão de acordo com mapeamentos globais realísticos.

Fonte: O autor, 2017.

dos números complexos, o novo quantificador de caos assume a forma

$$Z_{dyn} = \mathcal{A}_1 + i \lambda_{dyn}, \quad (32)$$

onde  $i$  representa a unidade imaginária (em fase de elaboração)<sup>8</sup>.

As dinâmicas típicas possuem uma esclarecedora representação no Diagrama de Argand. Na Figura 9, a magnitude de cada grandeza corresponde à esperada em processos realísticos de *global fitting*. Nestes casos, os valores para  $\mathcal{A}_1$  são finitos e  $\lambda_{dyn}$  se aproxima de zero para séries periódicas e randômicas.

<sup>8</sup> ALVES, P.; DUARTE, L.; da MOTA, L. Dynamical characterization of a time series by the polynomial global approach. Submetido à publicação em 2017.

## 4 ROTINAS COMPUTACIONAIS

A reconstrução dos vetores de estado, a determinação do vetor de parâmetros  $|c\rangle$ , a predição e o desvio são determinados pelas rotinas computacionais — executadas no ambiente Maple — dos pacotes `TimeS` e `LinMapTS`. O mesmo programa que calcula  $\sigma_\tau$ , aplica o teste de Shapiro-Wilk para a distribuição dos resíduos. Acurácia, baixo tempo de execução e um nível de confiança bem estabelecido são os atributos que estão na mira dos procedimentos computacionais desenvolvidos. Cada etapa da programação tem como objetivo contemplar estas três qualidades. Os Diagramas Acurácia-Desvio e Acurácia-Desvio Logarítmico, bem como o quantificador de caos  $Z_{dyn}$ , são disponibilizados pelo programa `DynCharTS`.

### 4.1 Pacote `LinMapTS`

Partindo de uma lista de vetores reconstruídos, o procedimento `LinGfiTS` seleciona qual parte da série temporal será empregada no mapeamento global e constrói as matrizes da Equação (23). No *output* do programa, a solução da equação matricial é incorporada na forma final de utilização da função preditiva (ALVES; DUARTE; da MOTA, 2016a).

- `[> LinGfiTS (argumentos);`
- Argumentos requeridos na ordem apresentada:
  - Lista de vetores reconstruídos — atribuída como  $V$  neste trabalho.
  - O vetor que possui, como seu primeiro componente, o último valor conhecido da série temporal — atribuído como *final* neste trabalho.
- Argumentos opcionais:
  - `Degree = <inteiro>` — grau do polinômio selecionado como preditor. O *default* é 2.
  - `Func = <expressão>` — função do vetor  $d$ -dimensional  $|X\rangle \doteq (X_1, X_2, \dots, X_d)$ , que contém os  $l$  componentes do vetor  $|c\rangle$  a serem determinados no processo de minimização. Prevalece sobre o argumento `Degree`, caso ambos sejam utilizados no mesmo comando.
  - `Level = <inteiro>` — nível de mapeamento. Seleciona a parte da série temporal a ser empregada no ajuste de dados. O *default* é 5.
  - `PT = <inteiro>` — ordem de entrada na série para o mapeamento e a predição. Corresponde ao parâmetro  $\tau$  que define o tempo de predição. O *default* é 1.

- `Analysis = 1`. Este *input* argumento é necessário para que a rotina realize uma análise gráfica do mapa global.

- [`>`] `LinGfiTS (V, final, <opcionais>);`

Faz parte da logística de programação assegurar que todos observáveis da série temporal — que correspondem a um determinado *nível de mapeamento* — participem do *data fitting* sem repetição. Cabe esclarecer que esta participação seletiva se dá nos componentes de cada vetor reconstruído. O resíduo emprega no seu cálculo o primeiro componente do vetor no espaço  $d$ -dimensional (15).

Determinado o preditor, o programa `ConfITS` calcula os resíduos de todos os pontos empregados no mapeamento global. Além de serem usados diretamente na determinação do desvio  $\sigma_\tau$  — o *output* do comando —, o conjunto ordenado dos resíduos compõe a amostra a ser submetida ao teste de normalidade.

- [`>`] `ConfITS (argumentos);`

- Argumentos requeridos na ordem apresentada:

- O mapa global — atribuído como *map* neste trabalho.
- Lista de vetores reconstruídos — atribuída como *V* neste trabalho.
- O vetor que possui, como seu primeiro componente, o último valor conhecido da série temporal — atribuído como *final* neste trabalho.

- Argumentos opcionais:

- `Level = <inteiro>` — nível de mapeamento. Seleciona a parte da série temporal a ser empregada no ajuste de dados. O *default* é 5.
- `PT = <inteiro>` — ordem de entrada na série para o mapeamento e a predição. Corresponde ao parâmetro  $\tau$  que define o tempo de predição. O *default* é 1.
- `Analysis = 1`. Este *input* é necessário para que a rotina realize a análise gráfica da distribuição dos resíduos e aplique o teste de normalidade.

- [`>`] `ConfITS (map, V, final, <opcionais>);`

O procedimento computacional imprime os valores da estatística  $W$ , o p-valor calculado e o resultado do teste de normalidade. Na mesma sequência, a rotina constrói uma curva normal padronizada sobreposta ao histograma da distribuição dos resíduos. Neste processo, um aspecto gráfico merece destaque: o critério para a *binagem* do histograma adota integralmente a regra criada por SCOTT (1979). Sendo a distribuição normal uma aproximação satisfatória para os resíduos apurados, a largura do *bin* resultará num histograma compatível com a curva gaussiana impressa. A formulação da *Regra de Scott* tem justamente o propósito de atender amostras de observáveis com distribuição normal.

## 4.2 Sobre o pacote TimeS

Existem três comandos do pacote TimeS que estão diretamente conectados com as rotinas LinGfiTS e ConfiTS, tanto na execução quanto na pedição (CARLI; DUARTE; da MOTA, 2014a). A lista de vetores reconstruídos é gerada pelo comando VecTS, enquanto que o procedimento ForecasTS aplica o mapa global num observável da série temporal e imprime o resultado da previsão. O comando GfiTS determina um preditor polinomial igual ao obtido com o programa LinGfiTS, mas com diferenças substanciais de performance. Está listado aqui porque o aspecto computacional destas duas diferentes rotinas — que podem gerar o mesmo *output* — tem enorme relevância nesta linha de pesquisa.

Pela estrutura do pacote, cada comando produz o *input* necessário ao aplicado na sequência. Mas não é necessário que todos sejam executados, porque em cada rotina estão incorporados todos os argumentos necessários à execução das tarefas realizadas pelo comando anterior. Então, o comando VecTS produz o *input* para GfiTS e este disponibiliza o mapa a ser empregado no procedimento ForecasTS. Ou, alternativamente, o programa ForecasTS pode ser executado diretamente a partir dos mesmos argumentos das rotinas VecTS e GfiTS.

- [`>` VecTS (argumentos);
- Argumentos
  - `DataFile` = `<string>` (opcional) — nome do arquivo da série temporal.
  - `TimeLag` = `<integer>` (opcional) — o parâmetro  $T$  que determina o lag da reconstrução (4). O *default* é 6.
  - `Dim` = `<integer>` (opcional) — a dimensão do espaço reconstruído. O *default* é 3.
  - `Data` = `<list[integer]>` (opcional) — uma lista com a série temporal.
- [`>` GfiTS (argumentos);
- Argumentos:
  - `DataFile` = `<string>` (opcional) — nome do arquivo da série temporal.
  - `TimeLag` = `<integer>` (opcional) — o parâmetro  $T$  que determina o lag da reconstrução (4). O *default* é 6.
  - `Dim` = `<integer>` (opcional) — a dimensão do espaço reconstruído. O *default* é 3.
  - `Data` = `<list[integer]>` (opcional) — uma lista com a série temporal.
  - `Vects` = `<list[lists]>` (opcional) — a lista de vetores.

- `Final` = `<integer>` (opcional) — o vetor que possui, como seu primeiro componente, o último valor conhecido da série temporal. O *default* é o vetor de ordem mais alta na lista selecionada para o mapeamento global.
  - `HP` = `<integer>` (opcional) — um número associado com a quantidade de vetores empregados no *data fitting*. O *default* é 4.
  - `Degree` = `<integer>` (opcional) — o grau do preditor polinomial. O *default* é 2.
  - `Nmaps` = `<integer>` (opcional) — número de mapas usados na determinação do preditor. O *default* é 1.
  - `PT` = `<inteiro>` — ordem de entrada na série para o mapeamento e a predição. Corresponde ao parâmetro  $\tau$  que define o tempo de predição. O *default* é 1.
- [`>`] `ForecasTS` (argumentos);
    - Argumentos:
    - `DataFile` = `<string>` (opcional) — nome do arquivo da série temporal.
    - `TimeLag` = `<integer>` (opcional) — o parâmetro  $T$  que determina o lag da reconstrução (4). O *default* é 6.
    - `Dim` = `<integer>` (opcional) — a dimensão do espaço reconstruído. O *default* é 3.
    - `Data` = `<list[integer]>` (opcional) — uma lista com a série temporal.
    - `Vects` = `<list[lists]>` (opcional) — a lista de vetores.
    - `Final` = `<integer>` (opcional) — o vetor que possui, como seu primeiro componente, o último valor conhecido da série temporal. O *default* é o vetor de ordem mais alta na lista selecionada para o mapeamento global.
    - `HP` = `<integer>` (opcional) — um número associado com a quantidade de vetores empregados no *data fitting*. O *default* é 4.
    - `Degree` = `<integer>` (opcional) — o grau do preditor polinomial. O *default* é 2.
    - `Nmaps` = `<integer>` (opcional) — número de mapas usados na determinação do preditor. O *default* é 1.
    - `PT` = `<inteiro>` — ordem de entrada na série para o mapeamento e a predição. Corresponde ao parâmetro  $\tau$  que define o tempo de predição. O *default* é 1.
    - `Map` = `<expressão>` (opcional) — o mapa global a ser empregado na predição.
    - `Position` = `<integer>` (opcional) — a posição do observável na série temporal imediatamente anterior ao que se espera prever. O *default* é o número de elementos da lista correspondente ao argumento `Data`.

Tabela 8 - Rotinas do pacote `TimeS`

comando	<i>output</i>
TS	lista com os observáveis da série temporal
VecTS	vetores no espaço reconstruído
GfiTS	mapa global polinomial
ForecasTS	predição baseada no mapa global
IforecasTS	aperfeiçoamento <i>one-step</i> da predição
NIforecasTS	aperfeiçoamento <i>N-step</i> da predição
AnalysTS	análises da série temporal
GrafiTS	resultados das análises em forma de gráficos

Legenda: Cada comando pode ser executado com os resultados dos anteriores ou então de maneira independente dos demais.

Fonte: O autor, 2017.

O resultado da previsão pode ser melhorado pela aplicação de um algoritmo que proporciona um *ganho de informação*. O procedimento `IforecasTS` calcula o termo a ser adicionado ao resultado obtido com o mapa global, que pode aumentar substancialmente a acurácia da predição (CARLI; DUARTE; da MOTA, 2014b). Também estão disponíveis neste pacote procedimentos de análise com opções gráficas. A Tabela 8 apresenta uma síntese de todos os comandos do pacote `TimeS`.

### 4.3 O algoritmo para o aperfeiçoamento da previsão global revisitado

No sistema dinâmico  $\mathcal{S}$  — definido pelas equações diferenciais (3) —, a derivada temporal de cada variável  $X_i$  é uma função  $F_i$  das  $N$  variáveis dinâmicas<sup>9</sup>.

$$\frac{dX_i}{dt} = F_i(X_1, X_2, \dots, X_N) \doteq F_i(|X\rangle) \quad (33)$$

A evolução de um estado  $|X_P\rangle$  para outro  $|X_{P+1}\rangle$ , num intervalo de tempo  $\delta t$ , corresponde a  $N$  mapeamentos

$$X_{i(P+1)} = G_i(|X_P\rangle, \delta t) \quad i = 1, 2, \dots, N \quad (34)$$

<sup>9</sup> Esta seção refaz o caminho percorrido por Carli, Duarte e da Mota (2014b) na construção do algoritmo que melhora a previsão pela aproximação global e fornece a base necessária à sua extensão para diferentes tempos de predição, ou seja,  $\tau = 1, 2, 3, \dots$  (ALVES; DUARTE; da MOTA, 2016d).

entre os pontos  $P$  e  $P + 1$  do espaço de fase original. Cada função  $F_i(|X_P\rangle)$  se relaciona com  $G_i(|X_P\rangle, \delta t)$  pela derivada parcial  $F_i(|X_P\rangle) = \left. \frac{\partial G_i}{\partial t} \right|_{t=0}$ .

Pela **Teoria de Lie**, um mapa deste tipo tem as propriedades de um **grupo** (BLUMAN, 2002; ARNOLD, 1973). Trata-se de numa transformação que fornece uma solução exponencial para o sistema de equações diferenciais (3).

$$G_i(|X\rangle, t) = \sum_{k=0}^{+\infty} \frac{t^k}{k!} \left[ \sum_{j=1}^N F_j \frac{\partial}{\partial X_j} \right]^k X_i \quad (35)$$

O operador  $\left[ \sum_{j=1}^N F_j \frac{\partial}{\partial X_j} \right]^k$  é o gerador infinitesimal do grupo de Lie. O mapeamento (34) corresponde à expansão

$$X_{i(P+1)} = G_i(|X_P\rangle, \delta t) = \sum_{k=0}^{+\infty} \frac{(\delta t)^k}{k!} \left[ \sum_{j=1}^N F_j \frac{\partial}{\partial X_j} \right]^k X_{i(P)} \quad (36)$$

Contudo, num mapa real  $\tilde{X}_{i(P+1)}$ , não se pode evitar o truncamento da série em alguma ordem  $\eta$ :

$$\tilde{X}_{i(P+1)} = \tilde{G}_i(|X_P\rangle, \delta t) = \sum_{k=0}^{\eta} \frac{(\delta t)^k}{k!} \left[ \sum_{j=1}^N F_j \frac{\partial}{\partial X_j} \right]^k X_{i(P)} \quad (37)$$

A expectativa é que  $\tilde{X}_{i(P+1)} \rightarrow X_{i(P+1)}$  quando  $\delta t \rightarrow 0$ . Algumas funções relacionadas aos erros de truncamento são importantes para a construção do algoritmo responsável pelo ganho de informação.

$$\epsilon_i(|X_P\rangle) = \sum_{k=\eta+1}^{+\infty} \frac{(\delta t)^k}{k!} \left[ \sum_{j=1}^N F_j \frac{\partial}{\partial X_j} \right]^k X_{i(P)} \equiv X_{i(P+1)} - \tilde{X}_{i(P+1)} \quad (38)$$

$$\delta \epsilon_i(|X_P\rangle) = \sum_{j=1}^N \frac{\partial \epsilon_i(|X_P\rangle)}{\partial X_j} \delta X_j + O(\delta X_j^2) \equiv \epsilon_i(|X_{P+1}\rangle) - \epsilon_i(|X_P\rangle) \quad (39)$$

⋮

$$\delta^k \epsilon_i(|X_P\rangle) = \sum_{j=1}^N \frac{\partial \delta^{k-1} \epsilon_i(|X_P\rangle)}{\partial X_j} \delta X_j + O(\delta X_j^2) \equiv \delta^{k-1} \epsilon_i(|X_{P+1}\rangle) - \delta^{k-1} \epsilon_i(|X_P\rangle) \quad (40)$$

Nas equações acima,  $k \geq 2$ . Considerando o intervalo de tempo infinitesimal  $\delta t \rightarrow 0$ , a equação à esquerda de (40) permite concluir que

$$\lim_{\delta t \rightarrow 0} \frac{\delta^k \epsilon}{\delta^{k-1} \epsilon} = 0, \quad (41)$$

para qualquer inteiro positivo  $k$ .

Antes de atacar propriamente o problema de aperfeiçoar a predição obtida na téc-

nica da aproximação global, convém raciocinar sobre os erros cometidos pelo truncamento em (37). O espaço de fase original não pode ser acessado a partir de uma série temporal. As próprias equações diferenciais (3) são desconhecidas. No esquema da reconstrução pelo método do *delay time*, um sistema dinâmico  $\mathcal{S}_{\mathcal{R}}$  equivale topologicamente ao original  $\mathcal{S}$  se o requisito dimensional do Primeiro Teorema de Takens (ver Seção 1.3) for atendido. Considerações de natureza operacional precisam ser feitas neste cenário.

O mapa obtido pelo truncamento da série (37) também não pode ser calculado. Numa condição ideal, o preditor  $\mathcal{P}_1(|X_P\rangle)$  se aproximaria do mapa  $\tilde{G}_i(|X_P\rangle, \delta t)$  com absoluta precisão. Na prática, felizmente, o resultado  $\bar{X}_{1(P+1)}$  (10) pode se aproximar satisfatoriamente daquele que seria obtido caso fosse possível trabalhar no sistema dinâmico original  $\mathcal{S}$  ou com um mapa global perfeitamente acurado no sistema  $\mathcal{S}_{\mathcal{R}}$ . Um facilitador neste processo é o método para o controle de qualidade apresentado na Seção 3.3.

Levando em consideração que o vetor  $|c\rangle$  é determinado num processo finito a partir da série temporal — ou de uma parte desta —, as quantidades infinitesimais  $\delta t$  e  $\delta \epsilon$  acima dão lugar às magnitudes finitas  $\Delta t$  e  $\Delta \epsilon$ .

$$\delta t \mapsto \Delta t$$

$$\delta \epsilon \mapsto \Delta \epsilon$$

A desigualdade  $\Delta^{\kappa+1}\epsilon \ll \Delta^{\kappa}\epsilon$  pode não ser válida para qualquer inteiro  $k$ . Todavia, determinar uma ordem  $\kappa$  que corresponda à aproximação

$$\Delta^{\kappa+1}\epsilon \approx \Delta^{\kappa}\epsilon, \quad (42)$$

constitui uma tarefa exequível.

A partir de uma lista com os resultados da aplicação do preditor  $\mathcal{P}_1(|X_P\rangle)$  nos  $a$  estados reconstruídos que antecedem o observável  $X_{i(P+1)}$  — o objeto de interesse na predição —

$$[\mathcal{P}_1(|X_{P-a}\rangle), \mathcal{P}_1(|X_{P-a+1}\rangle), \dots, \mathcal{P}_1(|X_P\rangle)] = [\bar{X}_{i(P-a)}, \bar{X}_{i(P-a+1)}, \dots, \bar{X}_{i(P)}] \quad (43)$$

são definidas as funções  $\Delta^k \epsilon_i(|X\rangle)$ , análogas a  $\delta^k \epsilon_i(|X\rangle)$ :

$$\Delta^0 \epsilon_i(|X_J\rangle) \equiv X_{i(J)} - \bar{X}_{i(J)} \quad (44)$$

$$\Delta^1 \epsilon_i(|X_J\rangle) \equiv \Delta^0 \epsilon_i(|X_J\rangle) - \Delta^0 \epsilon_i(|X_{J-1}\rangle) \quad (45)$$

⋮

$$\Delta^k \epsilon_i(|X_J\rangle) \equiv \Delta^{k-1} \epsilon_i(|X_J\rangle) - \Delta^{k-1} \epsilon_i(|X_{J-1}\rangle), \quad (46)$$

onde  $k \geq 2$  e  $J = P - a, P - a + 1, \dots, P$ . Com as funções acima, o inteiro  $\kappa$  em (42) pode ser determinado. Neste processo, é importante ter em mente que, a partir de uma

certa ordem  $\acute{k}$ ,  $\Delta^{\acute{k}-1}\epsilon$  começa a divergir.

A função  $\Delta^0\epsilon_i(|X_{P+1}\rangle)$  corresponde ao erro da predição  $\epsilon_{P+1}$ . Então, o conteúdo da Equação (10) pode ser reescrito no contexto da presente seção como

$$X_{i(P+1)} = \bar{X}_{i(P+1)} + \Delta^0\epsilon_i(|X_{P+1}\rangle). \quad (47)$$

Parte deste erro pode ser conhecida levando em conta a função (45), fazendo  $J \mapsto P+1$ .

$$\Delta^0\epsilon_i(|X_{P+1}\rangle) = \Delta^0\epsilon_i(|X_P\rangle) + \Delta^1\epsilon_i(|X_{P+1}\rangle) \quad (48)$$

Se  $P$  e  $P+1$  são próximos o suficiente, de modo que

$$\Delta^1\epsilon_i(|X_{P+1}\rangle) \ll \Delta^0\epsilon_i(|X_P\rangle),$$

um novo termo desconhecido é obtido pela substituição da relação (48) em (47):

$$X_{i(P+1)} = \bar{X}_{i(P+1)} + \Delta^0\epsilon_i(|X_P\rangle) + \Delta^1\epsilon_i(|X_{P+1}\rangle). \quad (49)$$

O **ganho de informação** ocorre porque o termo  $\Delta^1\epsilon_i(|X_{P+1}\rangle)$  — desconhecido em (49) — tem magnitude muito menor que  $\Delta^0\epsilon_i(|X_{P+1}\rangle)$  — desconhecido em (47). Então,  $\Delta^0\epsilon_i(|X_P\rangle)$  funciona como uma correção para  $\bar{X}_{i(P+1)}$ . Prosseguindo, a partir da relação

$$\Delta^1\epsilon_i(|X_{P+1}\rangle) = \Delta^1\epsilon_i(|X_P\rangle) + \Delta^2\epsilon_i(|X_{P+1}\rangle),$$

com  $P$  e  $P+1$  suficientemente próximos na condição

$$\Delta^2\epsilon_i(|X_{P+1}\rangle) \ll \Delta^1\epsilon_i(|X_P\rangle),$$

é possível obter uma correção de segunda ordem para  $\bar{X}_{i(P+1)}$ . Desta forma — enquanto a desigualdade

$$\Delta^{k+1}\epsilon_i(|X_{P+1}\rangle) \ll \Delta^k\epsilon_i(|X_P\rangle)$$

for aplicável —, correções de ordem superior podem ser incorporadas para a melhoria da previsão pela técnica global.

$$X_{i(P+1)} = \bar{X}_{i(P+1)} + \Delta^0\epsilon_i(|X_P\rangle) + \Delta^1\epsilon_i(|X_P\rangle) + \cdots + \Delta^k\epsilon_i(|X_P\rangle) + \Delta^{k+1}\epsilon_i(|X_{P+1}\rangle) \quad (50)$$

Finalizando, o algoritmo que aperfeiçoa o resultado da predição obtido de um mapa global deve ser capaz de identificar inteiro  $\kappa$  em que aproximação começa a falhar, ou seja, de detectar as magnitudes que atendem à condição  $\Delta^{\kappa+1}\epsilon_i \approx \Delta^\kappa\epsilon_i$ . A Tabela 9 apresenta

Tabela 9 - Os passos do algoritmo que aperfeiçoa a previsão global

passo	descrição
1	A rotina fixa $n = 10$ ;
2	A rotina calcula os módulos $ \Delta^k \epsilon_i $ até $k = n$ para o ponto $P$ ;
3	Para cada ordem $k$ , a rotina verifica se $ \Delta^k \epsilon_i  <  \Delta^{k+1} \epsilon_i $ ;
4	Se $ \Delta^k \epsilon_i  >  \Delta^{k+1} \epsilon_i $ , a rotina fixa $n = n + 10$ e retorna ao passo 2. Caso contrário, calcula $X_{i(P+1)}$ (51).

Legenda: O resultado do algoritmo corresponde ao *output* do comando `IforecastS` do pacote `TimeS` (ver Seção 4.2).

Fonte: O autor, 2017.

o método desenvolvido para encontrar a ordem da aproximação  $\kappa$  e os demais passos do algoritmo. Negligenciando o último termo em (50), o resultado final da predição é calculado por

$$X_{i(P+1)} \cong \bar{X}_{i(P+1)} + \Delta^0 \epsilon_i(|X_P\rangle) + \Delta^1 \epsilon_i(|X_P\rangle) + \dots + \Delta^\kappa \epsilon_i(|X_P\rangle). \quad (51)$$

#### 4.4 A extensão do algoritmo para diferentes tempos de predição

Algumas modificações na construção discutida na seção anterior permitem a extensão do algoritmo para diferentes valores do parâmetro  $\tau$  (ALVES; DUARTE; da MOTA, 2016d). Pela aplicação  $\delta t \mapsto \tau \delta t'$ , os mapeamentos (36) e (37) devem ser substituídos — na nova modelagem — por

$$X_{i(P+\tau)} = G'_i(|X_P\rangle, \tau \delta t') = \sum_{k=0}^{+\infty} \frac{(\tau \delta t')^k}{k!} \left[ \sum_{j=1}^N F_j \frac{\partial}{\partial X_j} \right]^k X_{i(P)} \quad (52)$$

e

$$\tilde{X}_{i(P+\tau)} = \tilde{G}'_i(|X_P\rangle, \tau \delta t') = \sum_{k=0}^{\eta} \frac{(\tau \delta t')^k}{k!} \left[ \sum_{j=1}^N F_j \frac{\partial}{\partial X_j} \right]^k X_{i(P)}. \quad (53)$$

Também neste caso, há a expectativa de que  $\tilde{X}_{i(P+\tau)} \rightarrow X_{i(P+\tau)}$  quando  $\tau\delta t' \rightarrow 0$ .

$$\epsilon'_i(|X_P\rangle) \equiv X_{i(P+\tau)} - \tilde{X}_{i(P+\tau)} \quad (54)$$

$$\delta\epsilon'_i(|X_P\rangle) \equiv \epsilon'_i(|X_{P+\tau}\rangle) - \epsilon'_i(|X_P\rangle) \quad (55)$$

⋮

$$\delta^k\epsilon'_i(|X_P\rangle) \equiv \delta^{k-1}\epsilon'_i(|X_{P+\tau}\rangle) - \delta^{k-1}\epsilon'_i(|X_P\rangle) \quad (k \geq 2) \quad (56)$$

A mesma justificativa que conduz ao limite (41) permite concluir — a partir das funções acima, para qualquer inteiro positivo  $k$  — que

$$\lim_{\tau\delta t' \rightarrow 0} \frac{\delta^k\epsilon'}{\delta^{k-1}\epsilon'} = 0. \quad (57)$$

No mapeamento do ponto  $P$  em  $P+\tau$ , um aspecto que merece atenção diz respeito aos resultados da aplicação do preditor  $\mathcal{P}_\tau(|X_P\rangle)$ . Os  $a$  estados reconstruídos para o aumento de acurácia na predição do observável  $X_{i(P+\tau)}$ , passam a ser separados por  $\tau$ . A lista a ser utilizada para o cálculo das funções  $\Delta^k\epsilon'_i(|X\rangle)$  — análogas às funções  $\delta^k\epsilon'_i(|X_P\rangle)$  — é dada por

$$[\mathcal{P}_\tau(|X_{P-\tau a}\rangle), \mathcal{P}_\tau(|X_{P-\tau a+\tau}\rangle), \dots, \mathcal{P}_\tau(|X_P\rangle)] = [\bar{X}_{i(P-\tau a)}, \bar{X}_{i(P-\tau a+\tau)}, \dots, \bar{X}_{i(P)}]. \quad (58)$$

Nesta proposta de extensão do algoritmo responsável pelo ganho de informação, as funções de interesse passam a ser definidas como

$$\Delta^0\epsilon'_i(|X_J\rangle) \equiv X_{i(J)} - \bar{X}_{i(J)} \quad (59)$$

$$\Delta^1\epsilon'_i(|X_J\rangle) \equiv \Delta^0\epsilon'_i(|X_J\rangle) - \Delta^0\epsilon'_i(|X_{J-\tau}\rangle) \quad (60)$$

⋮

$$\Delta^k\epsilon'_i(|X_J\rangle) \equiv \Delta^{k-1}\epsilon'_i(|X_J\rangle) - \Delta^{k-1}\epsilon'_i(|X_{J-\tau}\rangle), \quad (61)$$

onde  $k \geq 2$  e  $J = P - \tau a, P - \tau a + \tau, \dots, P$ .

A argumentação que conduz ao ganho de informação do método é exatamente a mesma que foi apresentada na seção anterior. Usando as aplicações

$$P + 1 \mapsto P + \tau$$

$$\epsilon_i \mapsto \epsilon'_i$$

na aproximação (51), obtém-se a forma final do algoritmo. A Tabela 10 mostra a adap-

Tabela 10 - Os passos do algoritmo que aperfeiçoa a previsão global estendido

passo	descrição
1	A rotina fixa $n = 10$ ;
2	A rotina calcula os módulos $ \Delta^k \epsilon'_i $ até $k = n$ para o ponto $P$ ;
3	Para cada ordem $k$ , a rotina verifica se $ \Delta^k \epsilon'_i  <  \Delta^{k+1} \epsilon'_i $ ;
4	Se $ \Delta^k \epsilon'_i  >  \Delta^{k+1} \epsilon'_i $ , a rotina fixa $n = n + 10$ e retorna ao passo 2. Caso contrário, calcula $X_{i(P+\tau)}$ (62).

Legenda: Fazendo  $\tau = 1$  e  $\epsilon'_i \mapsto \epsilon_i$ , os passos se tornam idênticos aos descritos na Tabela 9.

Fonte: O autor, 2017.

tação necessária nos passos da rotina.

$$X_{i(P+\tau)} \cong \bar{X}_{i(P+\tau)} + \Delta^0 \epsilon'_i(|X_P\rangle) + \Delta^1 \epsilon'_i(|X_P\rangle) + \cdots + \Delta^\kappa \epsilon'_i(|X_P\rangle) \quad (62)$$

#### 4.5 A nova versão do pacote TimeS

Ampliar a capacidade de previsão e análise constitui o objetivo principal da segunda versão do pacote TimeS. A inclusão do argumento opcional PT nos comandos GfiTS, ForecasTS e IforecasTS introduz a liberdade de escolha do parâmetro  $\tau$  nos programas (ALVES; DUARTE; da MOTA, 2016d).

O uso dos comandos segue essencialmente a proposta da primeira versão. Todavia, a logística de programação na rotina GfiTS precisou ser atualizada para que pudesse ser executada nas versões mais atuais do *software* Maple. O procedimento interno *gerpoly* — responsável pela geração dos polinômios no mapeamento global — deixou de funcionar corretamente a partir do Maple 17. Foi excluído do pacote e os polinômios passaram a ser gerados internamente na própria rotina GfiTS. Outra modificação realizada foi no argumento opcional que seleciona o intervalo da série temporal que deve ser empregado no *global fitting*. O argumento IniPoint foi substituído pelo Final em todas as rotinas nesta atualização. A motivação para esta alteração foi relacionar mais facilmente o mapa global com o observável de interesse na predição. No argumento opcional Poptions — que realiza análises da previsibilidade —, o erro percentual exibido foi substituído pelo próprio erro cometido.

Na rotina NIforecasTS, a previsão de um valor com ordem superior à primeira, já podia ser feito desde a primeira versão. Neste sentido, este comando desempenha

uma tarefa semelhante ao novo procedimento `IforecasTS`. Contudo, existe uma diferença substancial entre as duas rotinas. O procedimento `NIforecasTS` emprega o mapa que corresponde ao parâmetro  $\tau = 1$  — ou seja, correspondente ao passo  $N = 1$  — e faz a extensão para os passos  $N = 2$ ,  $N = 3$  e assim por diante. Por outro lado, na nova rotina `IforecasTS`, o aperfeiçoamento da previsão global já emprega o mapa global obtido com o parâmetro  $\tau$  correspondente — isto é, o comando carrega argumentos como `PT=2` e `PT=3`. Devido a esta diferença de concepções, a programação das rotinas `NIforecasTS`, `AnalysTS` e `GrafiTS` não foi modificada nesta segunda versão do pacote.

Convém, neste estágio da explicação, fazer uma comparação entre duas predições para o mesmo valor a partir dos procedimentos tratados no parágrafo anterior. A mesma série utilizada no trabalho original é utilizada neste exemplo (CARLI; DUARTE; da MOTA, 2014b). Corresponde aos valores — obtidos por integração numérica — da variável dinâmica  $X$  do Sistema de Lorenz (LORENZ, 1963), lida a partir do arquivo `'ts37.txt'`. O valor da série a ser previsto — supostamente desconhecido — é o `dat[110]`. Os comandos apresentados a seguir cobrem todas as etapas do método: da reconstrução dos vetores de estado até o aperfeiçoamento da previsão global.

```
[> V := VecTS (DataFile='ts37.txt'):
```

```
[> Mapag[2] := GfiTS(Vects=V, Final=108, PT=2);
```

```
[13,14,15,16,17,18,31,32,33,34,35,36,49,50,51,52,53,54,67,
68,69,70,71,72,85,86,87,88,89,90,103,104,105,106,107,108]
```

$$0.1164988046 X_1 X_2 - 0.02676222359 X_1 X_3 - 0.06491900291 X_2^2 + \\ 0.04190306144 X_2 X_3 - 0.009500679617 X_3^2 - 0.01709136487 X_1^2 + 1.503387694 X_1 - \\ 0.1930132013 X_2 - 0.02137643664 X_3$$

```
[> dat[110]; ForecastTS(Vects=V, Map=Mapag[2], Position=108, PT=2);
```

```
6.230624134
```

```
6.346599433
```

```
[>dat[110]; IforecasTS(Vects=V,Map=Mapag[2],Position=108,OK=5,PT=2);
```

```
6.230624134
```

```
6.213680535
```

Neste caso, a predição corresponde ao observável duas ordens à frente do último observável `dat [108]` utilizado no *global fitting* e considerado conhecido. O argumento `PT=2` especifica o parâmetro  $\tau = 2$ . Prosseguindo com a comparação, o *prompt* e o *output* do comando `NIforecastTS` são apresentados abaixo:

```
[>NIforecastTS(Vects =V,Map=Mapag[1],Position=108,Nsteps=2) [2] [1];
6.299945113.
```

O argumento `Nsteps=2` indica que  $\tau = 2$  e o preditor `Mapag[1]` acima corresponde ao parâmetro  $\tau = 1$ .

No resultado da comparação, a predição com o comando `IforecastTS` é mais acurada — 6.214 contra 6.300, sendo 6.231 o valor verdadeiro — que o resultado do procedimento `NIforecastTS`. Além de ampliar as possibilidades de análise do pacote `TimeS`, as funcionalidades da nova versão se traduzem também numa alternativa para o aumento da acurácia na predição.

Finalizando a apresentação das rotinas computacionais, é oportuno realizar uma análise na aplicação acima. Para que as rotinas realizem esta tarefa, é necessária a inclusão do argumento opcional `Poptions`.

```
[>IforecastTS(Vects=V,Map=Mapag[2],Position=108,OK=5,PT=2,Poptions=1);
```

```
Optimal k :,-----, 5
Value of the standard prediction :,-----, 6.346601601
Value of the improved prediction :,-----, 6.213684819
Real value of the time series :,-----, 6.230624134
Error of the standard prediction :,-----, 0.115977467
Error of the improved prediction :,-----, 0.016939315
6.213684819
```

## 4.6 O programa DynCharTS

No procedimento computacional para a caracterização dinâmica, o foco da rotina é o cálculo da acurácia (28) e do desvio (29). Ocorre que, neste processo, um número considerável de preditores precisa ser determinado simultaneamente. Se a caracterização emprega o maior tempo de predição  $\tau_{\max} \Delta t$ , então a rotina precisa resolver  $\tau_{\max}$  sistemas lineares com a forma (22) e calcular  $\tau_{\max}$  desvios do tipo (13). O esforço computacional aumenta, ainda mais, se o número de vetores reconstruídos  $M$  — em cada mapeamento global — for grande. Como o pacote `LinMapTS` requer um baixo tempo de execução, o seu núcleo foi incorporado no programa `DynCharTS`.

O único *input* necessário para o programa é a série temporal. Existem duas possibilidades para a entrada dos observáveis. O argumento opcional `DataFile` habilita a leitura destes a partir de um arquivo no formato ASCII. Uma lista de observáveis também pode ser informada diretamente com o argumento `Data`.

Para a detecção — gráfica e analítica — da presença de *outliers* em cada *global fitting*, é necessário incluir a opção `Analysis= 1` no comando. Na ausência deste argumento, não se pode assegurar a qualidade dos preditores e o método pode dar resultados errados. Para cada valor do parâmetro  $\tau$ , a rotina imprime um gráfico com o resultado do mapeamento global.

Com o argumento `PT`, o pesquisador pode especificar quantos parâmetros  $\tau$  serão empregados na caracterização dinâmica e também a escala do eixo horizontal dos diagramas impressos.

Apenas preditores polinomiais são admitidos no programa. Se um conjunto de mapeamentos globais apresentar qualidade insatisfatória, outro grau do polinômio pode ser selecionado com o argumento opcional `Degree`.

O programa imprime os dois tipos de diagramas do método de caracterização e o seu *output* é o quantificador de caos  $Z_{dyn}$ . A seguir, são apresentados os argumentos — com a descrição correspondente — do programa `DynCharTS`.

- [`>`] `DynCharTS (argumentos)`;
- Argumentos:
  - `DataFile = <string>` - (opcional). Este argumento informa o nome do arquivo no formato ASCII que contém a série temporal.
  - `Data = <list[integer]>` - (opcional). Este argumento lista os observáveis contidos na série temporal.
  - `Dim = <integer>` - (opcional). Este argumento especifica a dimensão do espaço de reconstrução.

- `TimeLag = <integer>` - (opcional). Este argumento especifica o *delay time* no esquema da reconstrução.
- `Degree = <integer>` - (opcional). Este argumento especifica o grau do preditor polinomial. O *default* é 3.
- `Level <integer>` - (opcional). Este argumento seleciona o intervalo da série temporal para o mapeamento global. Pelo *default*, o intervalo abrange a série inteira.
- `PT = <integer>` - (opcional). Este argumento especifica o máximo valor do parâmetro  $\tau$ . O *default* é 20.
- `Analysis = 1` - (opcional). Este argumento é necessário para a aplicação do procedimento de qualificação dos preditores.
- `OutLier = <integer>` - (opcional). Este argumento especifica o valor do parâmetro  $\eta_1$  (ver Tabela 6). O *default* é 1.
- `Final = <integer>` - (opcional). Este argumento especifica o último vetor empregado no mapeamento global. O *default* é aquele que contém o último dado da série temporal.
- `Mean = <integer>` - (opcional). Este argumento especifica o número de termos empregados no cálculo da média  $\lambda_{dyn}$  (31). O *default* é 11.

#### 4.7 Sobre os resultados do programa DynCharTS

Como o desvio  $\sigma_\tau$  (13) está presente no cálculo das magnitudes  $\mathcal{A}_\tau$  (28) e  $\mathcal{D}_\tau$  (29), os resíduos nos mapeamentos globais afetam os dois diagramas e o quantificador de caos. Embora isto não implique em prejuízo para a detecção do caos, algumas considerações são importantes para escalar a natureza dos resultados obtidos com o programa DynCharTS.

Acima de tudo, o grau selecionado para os preditores polinomiais deve resultar em mapeamentos livres de uma presença significativa de *outliers*. Então, uma forte recomendação consiste na inclusão do argumento `Analysis= 1` no comando DynCharTS.

Uma mesma série temporal pode apresentar diferentes diagramas e quantificadores  $Z_{dyn}$  se o método empregar graus distintos para os polinômios. Devido a esta realidade, as coordenadas verticais nos Diagramas Acurácia-Desvio Logarítmicos e as grandezas  $\{\mathcal{A}_1, \lambda_{dyn}\}$  (32) podem variar. Todavia, a assinatura da dinâmica ainda estará presente nos resultados.

Existe um recurso no programa que avalia se o cálculo de  $Z_{dyn}$  é aplicável ou não. Nas séries temporais caóticas, as curvas de  $\mathcal{A}_\tau$  ou de  $\ln(\mathcal{A}_\tau)$  apresentam decaimento. A

rotina, por sua vez, contabiliza o número de termos — denotados por  $\hat{\zeta}$  — que experimentam crescimento no lugar de decaimento.

Devido às flutuações estatísticas, até num sistema caótico, alguns pares  $(\mathcal{A}_{\tau+1}, \mathcal{A}_{\tau})$  podem satisfazer a desigualdade

$$\mathcal{A}_{\tau+1} - \mathcal{A}_{\tau} > 0.$$

No sentido de caracterizar a dinâmica como caótica, considera-se o resultado aceitável se  $\hat{\zeta} < \zeta/10$  (ver Equação 31). Quando este critério não é atendido — como ocorre nas séries periódica e randômica da Seção 5.9 — a rotina imprime a mensagem

*“The quantifier of chaos is not applicable for this Time Series.”*

Um último aspecto, com respeito ao argumento opcional **Mean**, merece uma consideração. Escolhas de parâmetros com grande magnitude devem ser feitas com cuidado. Além de ser desnecessário para a detectar a presença do caos, um alto valor de  $\zeta$  (31) pode mascarar a natureza caótica porque a parte imaginária de  $Z_{dyn}$  se aproximaria necessariamente de zero. Para uma manipulação útil deste argumento, é recomendável uma inspeção prévia dos diagramas que a própria rotina imprime.

#### 4.8 Problemas convencionais de *data fitting*

A atividade experimental prescinde da informação disponibilizada por dados e medidas. No desenvolvimento da Ciência propriamente dita, a análise de erros experimentais se constituiu num guia importantíssimo para o entendimento dos fenômenos da Natureza. A partir de um conjunto de medidas e suas incertezas, o *data fitting* estabelece a comparação entre um modelo matemático e a experiência, desempenhando um papel central no teste de um modelo em Física.

Numa experiência, o cientista usualmente precisa planejar os próximos passos de seu trabalho. Então, um ajuste preliminar de dados durante a execução do experimento constitui uma prática altamente recomendada (SQUIRES, 2001). Terminada uma determinada etapa, a tomada de mais alguma(s) medida(s) é seguida por um novo ajuste e assim por diante.

A *modelagem* para um conjunto de observáveis segue o mesmo método que um físico experimental costuma aplicar. Se uma lei matemática descreve um fenômeno, então a curva ajustada assume a função de um *preditor*. Outra importante categoria de aplicação dos métodos de *data fitting* trata dos procedimentos que envolvem a *calibração* de instrumentos (NIST/SEMATECH, 2016).

O conteúdo deste trabalho nasceu do esforço em aperfeiçoar a predição a partir de uma série temporal. No esquema da reconstrução, a Aproximação Global pode ser vista

como um processo sofisticado de *data fitting*, recebendo também a denominação de *global fitting* (CARLI; DUARTE; da MOTA, 2014b). Existe uma significativa quantidade de material desenvolvida — no campo da análise de séries temporais — a partir de ideias que vão ao encontro das demandas de um ajuste de dados convencional. No aspecto computacional, a logística de programação tem plena compatibilidade com as tarefas de encontrar parâmetros e realizar análises em ajustes de curvas (ALVES; DUARTE; da MOTA, 2016a).

Na proposta de disponibilizar um algoritmo de alta performance no *data fitting*, foi desenvolvido o programa `LinFit` (em fase de elaboração)<sup>10</sup>. Este conjuga uma solução de mínimos quadrados de um sistema linear de equações — perfeitamente análogo ao (22) — com análises do ajuste no ambiente da computação simbólica. Trata-se de uma transposição do pacote `LinMapTS` para uma categoria semelhante de problemas, que difere apenas na sua formatação. Com o objetivo de flexibilizar a aplicação do programa, as rotinas oferecem uma gama considerável de opções analíticas e gráficas. Algumas quantidades estatísticas e o *controle de qualidade do data fitting* seguem a inspiração da qualificação dos preditores e da acurácia  $\mathcal{A}_\tau$  introduzidos ao longo do texto. Toda a construção e também as instruções necessárias para a utilização do programa estão descritas nos Apêndices A, B, C, D, E, F.

---

<sup>10</sup> ALVES, P.; DUARTE, L.; da MOTA, L. A high-performance method for data fitting and its analysis. Submetido à publicação em 2017.

## 5 RESULTADOS E DISCUSSÕES

Na estrutura deste capítulo, cada resultado é apresentado em conjunto com a sua discussão. As aplicações propostas tem os objetivos de testar as rotinas e de exemplificar o uso dos programas. A evolução temporal de observáveis com dinâmicas típicas — e previamente conhecidas — validam o método de caracterização. Séries temporais afetadas pela complexidade do mercado financeiro e da atividade solar são investigadas neste contexto.

### 5.1 Aplicação do pacote LinMapTS

Na aproximação global polinomial, os comandos `GfiTS` e `LinGfiTS` devem gerar o mesmo mapa global. A partir da série temporal utilizada para ilustrar os comandos `IforecasTS` e `NIforecasTS` na Seção 4.5 — representada graficamente na Figura 10a —, a reconstrução dos vetores com o *time lag* igual a seis, no espaço euclidiano tridimensional, é realizada pelo comando `VecTS`.

```
[> V1 := VecTS (DataFile='ts37.txt', TimeLag=6, Dim=3):
```

Para um determinado intervalo da série temporal, os preditores são obtidos no *worksheet* Maple com os seus respectivos argumentos:

```
[> Map1:= GfiTS (Vects=V1, Degree=2, Final=236, PT=1);
```

```
[141, 142, 143, 144, 145, 146, 159, 160, 161, 162, 163, 164, 177, 178, 179, 180, 181, 182,
195, 196, 197, 198, 199, 200, 213, 214, 215, 216, 217, 218, 231, 232, 233, 234, 235, 236]
```

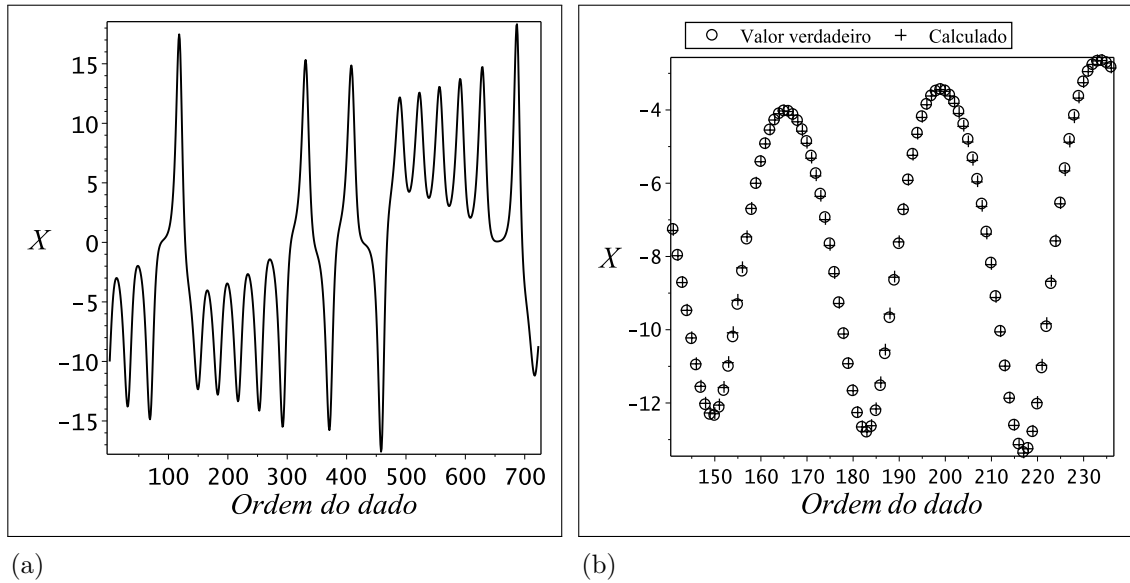
$$\begin{aligned} \text{Map1} := & 1.310847436 X_1 - 0.2238324059 X_2 + 0.01260642766 X_3 \\ & - 0.001533559052 X_1^2 + 0.03134153352 X_1 X_2 - 0.004000218087 X_1 X_3 \\ & - 0.01101483408 X_2^2 - 0.002050958127 X_2 X_3 + 0.00006832853081 X_3^2 \end{aligned}$$

```
[> final := 236;
```

```
[> LinMap1 := LinGfiTS (V1, final, Degree=2, Level=5, PT=1);
```

```
[141, 142, 143, 144, 145, 146, 159, 160, 161, 162, 163, 164, 177, 178, 179, 180, 181, 182,
195, 196, 197, 198, 199, 200, 213, 214, 215, 216, 217, 218, 231, 232, 233, 234, 235, 236]
```

$$\begin{aligned} \text{LinMap1} := & 1.310885114 X_1 - 0.2239015600 X_2 + 0.01261609078 X_3 \\ & - 0.001531089863 X_1^2 + 0.03133713828 X_1 X_2 - 0.003997791057 X_1 X_3 \\ & - 0.01101522745 X_2^2 - 0.00205353663 X_2 X_3 + 0.00006823234 X_3^2. \end{aligned}$$

Figura 10 - Variável dinâmica  $X$  do Sistema de Lorenz

Legenda: (a) Série temporal gerada por integração numérica. (b) Intervalo correspondente aos mapas `Map1` e `LinMap1`. A barra de erro — cujo comprimento é o desvio  $\sigma_1 \cong 0.013$  — não pode ser visualizada na escala do gráfico.

Fonte: O autor, 2017.

O resultado impresso mostra que, para uma lista idêntica de vetores reconstruídos, os polinômios são essencialmente os mesmos. As diferenças nos coeficientes são devidas aos *truncamentos* realizados pelas rotinas. O resultado das predições

```
[> V1[237][1] - ForecastTS(Vects=V, Map=Map1, Position=236);
0.037211619
```

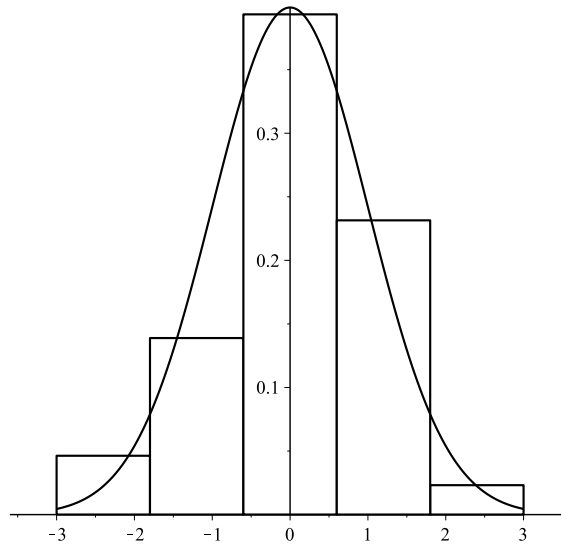
```
[> V1[237][1] - ForecastTS(Vects=V, Map=LinMap1, Position=236);
0.037213863
```

são praticamente iguais. Os erros para o dado de ordem 237 — que é o primeiro componente do vetor reconstruído `V1[237]`, isto é, `V1[237][1]` — apresentam diferenças somente a partir da sexta casa decimal.

Aplicando também o procedimento `ConfITS` para o intervalo da série temporal acima, o segundo preditor obtido pela técnica global é utilizado no cálculo do desvio  $\sigma_1$  e no teste de Shapiro-Wilk:

```
[> sigma1 := ConfITS(LinMap1,V1,final,Level=5,PT=1,Analysis=1);
```

```
[141,142,143,144,145,146,159,160,161,162,163,164,177,178,179,180,181,182,
195,196,197,198,199,200,213,214,215,216,217,218,231,232,233,234,235,236]
```



### Shapiro and Wilk's W-Test for Normality

-----  
Null Hypothesis:

Sample drawn from a population that follows a normal distribution

Alt. Hypothesis:

Sample drawn from population that does not follow a normal distribution

Sample size:                   36  
Computed statistic:        0.982286  
Computed pvalue:           0.868912

Result: [Accepted]

There is no statistical evidence against the null hypothesis

$$\sigma_1 := 0.01302910469.$$

Este desvio no gráfico da Figura 10b, nem chega a ser perceptível. São computados, nesta representação, os resultados do preditor em todos os dados que correspondem ao intervalo da série temporal selecionado. Todavia, apenas os vetores listados na impressão feita pela rotina são utilizados no mapeamento global. Vale a pena lembrar que esta estratégia impede a repetição de dados no *global fitting*.

O alto p-valor calculado, igual a 0.868912, é uma forte evidência a favor da hipótese de normalidade da distribuição dos resíduos. O histograma construído pelo procedimento *ConfITS* está em pleno acordo com o resultado do teste de Shapiro-Wilk. A razão entre o erro da predição e o desvio calculado pela rotina (13) — em torno de 2.856, na sequência do *prompt* do *software* Maple reproduzido a seguir — é aceitável, se for levada em conta a faixa de  $3\sigma_\tau$  (ver Seção 2.6).

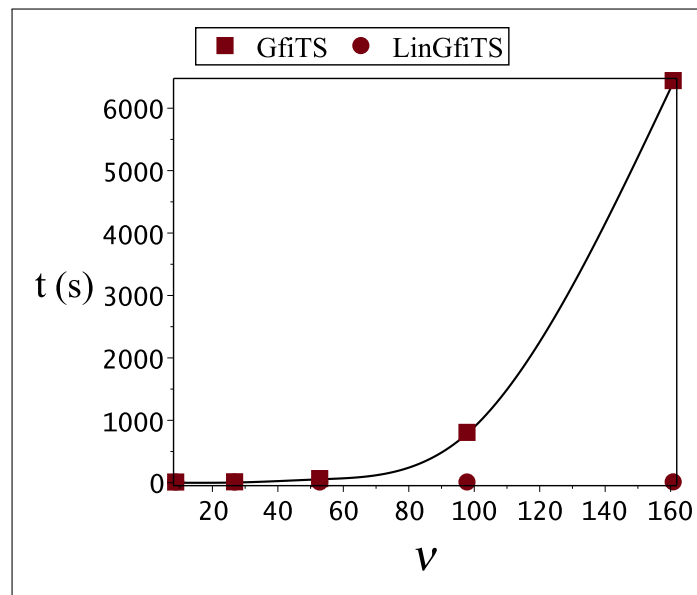
Tabela 11 - Medida do tempo de execução das rotinas GfiTS e LinGfiTS

grau do polinômio	número de vetores	graus de liberdade $\nu$	GfiTS	LinGfiTS
1	12	9	0.187 s	0.015 s
2	36	27	3.681 s	0.031 s
3	72	53	55.161 s	0.156 s
4	132	98	795.355 s	0.764 s
5	216	161	6430.969 s	2.854 s

Legenda: As medidas de tempo foram obtidas com um processador Pentium (R)  
Dual-Core 2.60 GHz.

Fonte: O autor, 2017.

Figura 11 - Número de graus de liberdade e eficiência das rotinas GfiTS e LinGfiTS



Legenda: A rotina LinGfiTS tem uma sensibilidade consideravelmente menor em relação ao número de graus de liberdade  $\nu$ .

Fonte: O autor, 2017.

```
[> (V1[237][1] - ForecastTS(Vects=V1, Map=LinMap1, Position=236))/sigma1;
2.856121881
```

Para uma razão com este valor, a priori, existe confiança ao nível de 99.6% de que o valor verdadeiro esteja no intervalo entre  $x_{1P(r+\tau)} - 2.86\sigma_\tau$  e  $x_{1P(r+\tau)} + 2.86\sigma_\tau$ . É o que resulta da aplicação da fórmula (80):

$$\mathcal{L}_1 = 1 - \frac{2}{\sqrt{2\pi}} \int_{2.86}^{\infty} \exp\left\{-\frac{\epsilon_{r-\tau}^2}{2\sigma_\tau^2}\right\} d(\epsilon_{r-\tau}/\sigma_\tau) = 0.996.$$

## 5.2 Comparação dos tempos de execução das rotinas GfiTS e LinGfiTS

Quando o programa GfiTS é empregado para obter um preditor polinomial com um número considerável de parâmetros de ajuste, o tempo necessário para o mapeamento global pode ser considerado longo. Todavia, realizar análises com baixo custo computacional constitui uma das propostas do pacote TimeS (CARLI; DUARTE; da MOTA, 2014b). O desenvolvimento da rotina LinGfiTS tem, como motivação primeira, a redução do tempo necessário para o cálculo dos coeficientes de um preditor. No procedimento GfiTS, o caso de uma função com muitos parâmetros de ajuste torna-se crítico. Outro aspecto importante diz respeito a análise de intervalos extensos da série temporal. O ponto crucial responsável pelo aumento de eficiência é o algoritmo — responsável pela minimização — desenvolvido e incorporado ao programa LinGfiTS.

```
[> time(GfiTS (Vects=V1, Degree=2, Final=236, PT=1));
3.681
```

```
[> time(LinGfiTS (V1, final, Degree=2, Level=5, PT=1));
0.031
```

O resultado destas medidas de tempo — realizadas no ambiente Maple — é dado em segundos.

O *número de graus de liberdade*  $\nu$  é empregado para comparar as eficiências das rotinas que realizam o mapeamento global. Tal quantidade corresponde à diferença entre o número de vetores usados no *data fitting* e o número de parâmetros ajustados na função preditiva (BEVINGTON; ROBINSON, 2003). No exemplo dos mapas Map1 ou LinMap1 da Seção 5.1, são 36 os vetores reconstruídos e 9 os coeficientes do polinômio de grau dois no espaço tridimensional, o que resulta em  $\nu = 36 - 9 = 27$  graus de liberdade.

As medidas de tempo mostradas na Tabela 11 e a respectiva comparação gráfica da Figura 11 mostram que existe uma brutal redução no tempo de execução da rotina `LinGfiTS` em relação ao requerido pelo comando `GfiTS`. Este enorme ganho na eficiência é consequência da substituição de uma rotina do *software* Maple chamada `minimize` — que realiza a minimização no procedimento `GfiTS` — por um algoritmo que constrói e determina os parâmetros diretamente da equação matricial (23), no programa `LinGfiTS`.

### 5.3 Aperfeiçoamento na Aproximação Global Polinomial

Um preditor mais acurado pode ser obtido pelo uso de um polinômio de grau mais elevado. Vale destacar que resultados desastrosos na predição também são obtidos pelo aumento indiscriminado do número de parâmetros de ajuste. O equilíbrio entre o número de graus de liberdade e a forma funcional da predição tem uma verificação imediata pelo valor do desvio  $\sigma_\tau$  (13) — o *output* do comando `ConfITS`.

No mapa global `LinMap1`, impresso na Seção 5.1, o nível de mapeamento e o último observável da série empregado no ajuste são especificados pelos argumentos `Level=5` e `final=236`. A Figura 12a mostra o resultado da aplicação deste preditor para a série inteira. Existem, nitidamente, discrepâncias entre o resultado da aplicação do mapa e o valor verdadeiro em vários pontos — principalmente na faixa de observáveis de ordem superior a 236.

Com o intuito de obter um mapa global mais acurado para a série temporal inteira, a estratégia a ser seguida, neste caso, é aumentar o grau do polinômio e escolher um nível de mapeamento que leve em conta todos os observáveis da série temporal.

$$\mathcal{P}(|x_r|) = c_1x_{1r} + c_2x_{2r} + c_3x_{3r} + \dots + c_{55}x_{3r}^5 \quad (63)$$

Considerando que o polinômio de quinto grau no espaço tridimensional (63) possui 55 componentes e que 240 vetores que participam do *data fitting*, o número de graus de liberdade é 185 — pois  $240 - 55 = 185$ . O *output* correspondente ao mapeamento da série inteira — intervalo selecionado pelo argumento `Level=39` — resulta da aplicação dos comandos reproduzidos a seguir.

```
[> V1 := VecTS (DataFile='ts37.txt', TimeLag=6, Dim=3):
```

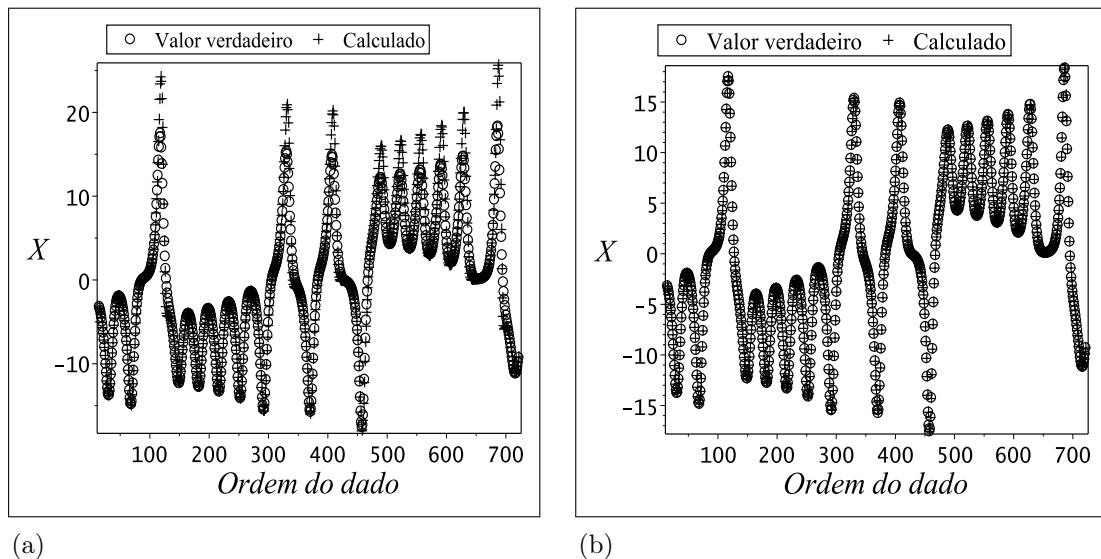
```
[> final := 722:
```

```
[> LinMap2 := LinGfiTS (V1, final, Degree=5, Level=39, PT=1);
```

[15, 16, 17, 18, 19, 20, 33, 34, 35, 36, 37, 38, 51, 52, 53, 54, 55, 56, 69, 70, 71, 72, 73, 74, 87, 88, 89, 90, 91, 92, 105, 106, 107, 108, 109, 110, 123, 124, 125, 126, 127, 128, 141, 142, 143, 144, 145, 146, 159, 160, 161, 162, 163, 164, 177, 178, 179, 180, 181, 182, 195, 196, 197, 198, 199, 200, 213, 214, 215, 216, 217, 218, 231, 232, 233, 234, 235, 236, 249, 250, 251, 252, 253, 254, 267, 268, 269, 270, 271, 272, 285, 286, 287, 288, 289, 290, 303, 304, 305, 306, 307, 308, 321, 322, 323, 324, 325, 326, 339, 340, 341, 342, 343, 344, 357, 358, 359, 360, 361, 362, 375, 376, 377, 378, 379, 380, 393, 394, 395, 396, 397, 398, 411, 412, 413, 414, 415, 416, 429, 430, 431, 432, 433, 434, 447, 448, 449, 450, 451, 452, 465, 466, 467, 468, 469, 470, 483, 484, 485, 486, 487, 488, 501, 502, 503, 504, 505, 506, 519, 520, 521, 522, 523, 524, 537, 538, 539, 540, 541, 542, 555, 556, 557, 558, 559, 560, 573, 574, 575, 576, 577, 578, 591, 592, 593, 594, 595, 596, 609, 610, 611, 612, 613, 614, 627, 628, 629, 630, 631, 632, 645, 646, 647, 648, 649, 650, 663, 664, 665, 666, 667, 668, 681, 682, 683, 684, 685, 686, 699, 700, 701, 702, 703, 704, 717, 718, 719, 720, 721, 722]

$$\begin{aligned}
LinMap2 := & -0.0000600514052 X_1^2 X_2 X_3 - 0.00000641264175 X_1 X_2 X_3^3 \\
& -0.00004214395253 X_1^3 X_2 X_3 + 0.001447344520 X_1 X_2 X_3 \\
& +0.0000025085972 X_1^2 X_2 X_3^2 + 0.0000332479287 X_1^2 X_2^2 X_3 \\
& -0.00001009951401 X_1 X_2^3 X_3 + 0.0001147408728 X_1 X_2^2 X_3 \\
& +0.000001423638047 X_1^4 X_2 + 0.00001096339200 X_1^4 X_3 \\
& +0.0000076962891 X_1^3 X_2^2 + 0.0000095483628 X_1^3 X_3^2 \\
& -0.00001273685136 X_1^2 X_2^3 + 0.00000026956042 X_1^2 X_3^3 \\
& -0.0000072740231 X_1 X_2^4 + 0.00000150427907 X_1 X_3^4 \\
& -0.00000263242158 X_2^4 X_3 + 0.00001841464019 X_2^3 X_3^2 \\
& +0.00001116121846 X_2^2 X_3^3 + 0.00000371432080 X_2 X_3^4 \\
& +0.001802084193 X_1^2 - 0.01060143548 X_1 X_2 + 0.005604747413 X_1 X_3 \\
& +0.01391977937 X_2^2 - 0.01264694194 X_2 X_3 + 0.001668551168 X_3^2 \\
& -0.005461828112 X_1^2 X_2 - 0.001220698474 X_1^2 X_3 \\
& +0.007880003353 X_1 X_2^2 + 0.002650073097 X_1 X_3^2 \\
& -0.002180278582 X_2^2 X_3 - 0.004300482492 X_2 X_3^2 \\
& -0.00000069142151 X_1^3 X_2 + 0.00000543185810 X_1^3 X_3 \\
& +0.00002788889508 X_1^2 X_2^2 + 0.00003053220324 X_1^2 X_3^2 \\
& -0.00004386510504 X_1 X_2^3 + 0.00002293134682 X_1 X_3^3 \\
& -0.00006035172667 X_2^3 X_3 + 0.00005730839100 X_2^2 X_3^2 \\
& -0.00001920240350 X_2 X_3^3 + 1.361113994 X_1 \\
& -0.4639975694 X_2 + 0.05876569915 X_3 + 0.000659550426 X_1^3 \\
& -0.000404392460 X_2^3 + 0.001199988282 X_3^3 \\
& -0.000000377207113 X_1^4 + 0.000001264800674 X_2^4 \\
& +0.000007154220429 X_3^4 - 0.00000103407107 X_1^5 \\
& +0.00000230569077 X_2^5 + 0.0000007643101649 X_3^5
\end{aligned}$$

Figura 12 - Mapeamentos distintos para o Sistema de Lorenz



Legenda: (a) Intervalo correspondente ao ajuste:  $[141, 236]$ . (b) *Data fitting* realizado para a série inteira. Este tipo de análise gráfica pode ser disponibilizada com a inclusão do argumento `Analysis = 1` no comando `LinGfiTS`.

Fonte: O autor, 2017.

O erro de cada predição para o observável de ordem 723 — supostamente desconhecido — e o desvio esperado, são determinados no próprio ambiente Maple:

```
[> V1[723][1] - ForecastTS(Vects=V1, Map=LinMap1, Position=722);
```

```
-0.121391908
```

```
[> V1[723][1] - ForecastTS(Vects=V1, Map=LinMap2, Position=722);
```

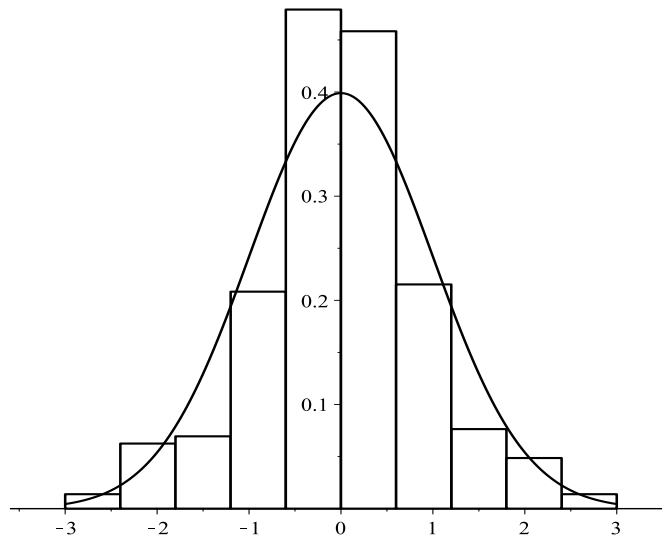
```
-0.003730711
```

```
[> sigma2:=ConfiTS(LinMap2, V1, final, Level=39, PT=1);
```

```
 $\sigma 2 := 0.003322854986 .$ 
```

No *output* acima, verifica-se um aumento significativo da acurácia com o preditor `LinMap2`. O desvio  $\sigma 2 \cong 0.0033$  está em pleno acordo com o erro de predição  $\epsilon_{722+1} \cong -0.0037$ . Neste caso, a distribuição dos resíduos — visualizada graficamente na Figura 13 — é aceita como gaussiana no teste de Shapiro-Wilk. Vale ressaltar que, se o mapa `LinMap2` fosse gerado pela rotina `GfiTS`, o tempo de execução seria consideravelmente maior (ver Tabela 11 e Figura 11).

Figura 13 - Distribuição dos resíduos no mapa LinMap2



Legenda: O teste de normalidade tem resultado positivo.

Fonte: O autor, 2017.

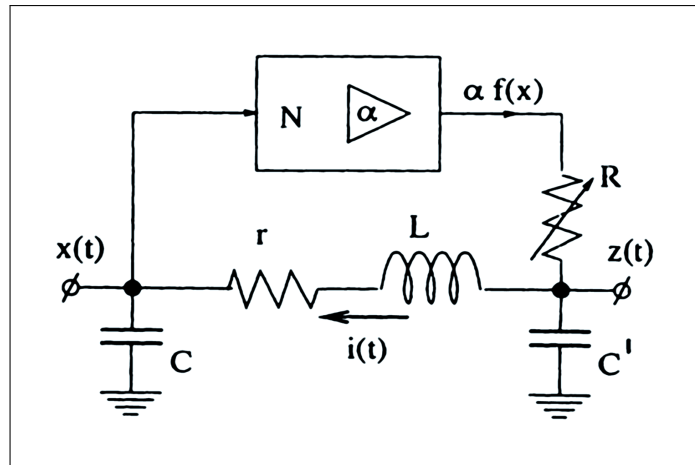
#### 5.4 Série experimental caótica

A série temporal representada graficamente na Figura 3a resulta das medidas de voltagem realizadas em um circuito idealizado por Brown, Rulkov e Tracy (1994). O esquema correspondente está representado na Figura 14. Duas demandas deste trabalho são atendidas a partir deste conjunto de medidas elétricas. Uma delas é o teste das rotinas computacionais. A outra diz respeito à escolha dos parâmetros de reconstrução.

Além da forma da função escolhida como preditiva e do número de graus de liberdade no mapeamento global, a qualidade do preditor depende também da dimensão do espaço e do *time lag*. Valores exagerados para os parâmetros de reconstrução devem ser evitados. Uma quantidade insuficiente de dimensões implica que a dinâmica original deixa de ser preservada na reconstrução. Por outro lado — de acordo com a argumentação dos dois últimos parágrafos da Seção 1.3 — a elevação desmedida do número de coordenadas também não dá a garantia de um mapeamento satisfatório. Com relação ao *delay T* (4), é necessário que exista *correlação* entre os observáveis para que os vetores sejam representativos do estado do sistema. Então, espaçamentos altos demais são inadequados. No outro extremo, observáveis da série muito próximos são pouco informativos com respeito à evolução temporal.

Métodos clássicos para a escolha da dimensão dos vetores de estado e do *time lag* são empregados, respectivamente, com a determinação da quantidade de *falsos vizinhos próximos* e o cálculo da *informação mútua média* (SCHELTER, 2006). Ambos estão dis-

Figura 14 - Esquema de geração do sinal caótico



Legenda: A voltagem é medida no capacitor  $C$ .

Fonte: BROWN, RULKOV, TRACY, 1994, p. 3791.

poníveis para implementação computacional (HEGGER; KANTZ; SCHREIBER, 1999).

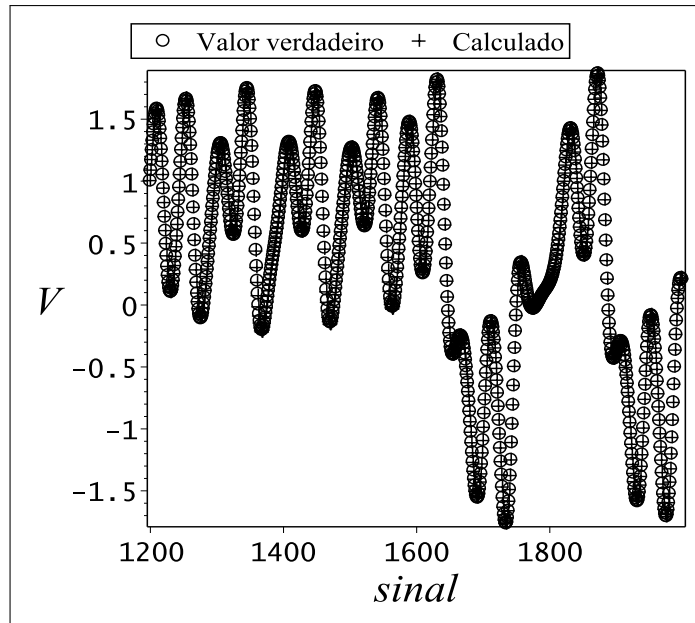
Os estados próximos no espaço com uma dimensão devem assim permanecer quando um novo espaço for reconstruído com uma dimensão maior. Aumenta-se progressivamente o número de dimensões até que se chegue a um nível aceitável de estados que mudam drasticamente sua proximidade. A estatística empregada no cômputo dos falsos vizinhos avalia os pontos que se afastam fortemente neste processo.

A mútua informação média é uma medida da informação que um observável da série temporal possui a respeito do outro. Seu cálculo é baseado numa quantidade da Teoria da Informação chamada de *entropia de Shannon* — uma generalização da entropia termodinâmica que caracteriza a informação armazenada em termos de distribuições probabilísticas (SHANNON, 1948). De fato, a teoria da informação fornece os elementos para uma importante abordagem das séries temporais. O conjunto de observáveis pode ser tratado como uma mensagem transmitida (KANTZ; SCHREIBER, 2003).

Na série temporal contendo medidas de voltagem, o percentual de falsos vizinhos torna-se aceitável num espaço reconstruído com apenas três dimensões (ver Figura 17a). Brown, Rulkov e Tracy (1994) também consideram que o *delay time* ótimo corresponde ao primeiro mínimo da informação mútua. Trata-se do critério recomendado por Fraser e Swinney (1986). De acordo com o gráfico na Figura 17b, o valor do *lag* considerado ideal pelos autores é  $T = 10$ .

Partindo do arquivo `circuit.txt`, que contém as medidas de voltagem do sinal caótico — disponibilizadas por McSharry (2014) —, os vetores são reconstruídos num espaço de cinco dimensões e *delay* igual a dez pela rotina `VecTS`. Levando em conta a análise dos parâmetros de reconstrução reproduzida graficamente na Figura 17, existem fortes razões para considerar que a imersão do sistema original no espaço de fase reconstruído

Figura 15 - Mapa global do sinal elétrico caótico



Legenda: Existe uma nítida sobreposição dos valores calculados com a voltagem verdadeira  $V$ .

Fonte: O autor, 2017.

será plenamente satisfatória.

```
[> V3 := VecTS (DataFile='circuit.txt', TimeLag=10, Dim=5):
[> final := 1999:
[> LinMap3 := LinGfiTS (V3, final, Degree=3, Level=38, PT=1):
[> sigma3 := ConfiTS(LinMap3, V3, final, Level=38, PT=1);
```

$$\sigma_3 := 0.001536774328$$

```
[> V3[2000][1] - ForecastTS(Vects=V3, Map=LinMap3, Position=1999);
```

$$-0.0019464679$$

A concordância da voltagem verdadeira  $V$  com a calculada pelo mapa `LinMap3` — visível no gráfico da Figura 15 — indica que a dinâmica responsável pela evolução temporal é preservada no esquema da reconstrução. Esta evidência está de acordo com a escolha criteriosa da dimensão dos vetores  $d_E = 5$  e do *time delay*  $T = 10$ . Nesta série experimental, o valor verdadeiro `V3 [2000] [1]` difere da voltagem calculada em menos de duas vezes o desvio  $\sigma_3$ .

## 5.5 O desvio numa série randômica

Em um sistema que não apresenta nenhuma dinâmica determinística, a aproximação global não tem como ser acurada. Desta forma, o desvio calculado deve estar de acordo com as previsões completamente erradas numa série temporal formada de números aleatórios. Uma instrutiva comparação visual entre a evolução caótica e a absolutamente aleatória é proporcionada nos dois gráficos da Figura 3.

Com o propósito de obter uma série temporal randômica — com dados obtidos do mundo real —, podem ser coletados os números sorteados numa loteria. No Brasil, a loteria mais famosa e que paga milhões de reais aos vencedores do primeiro prêmio é a Mega-Sena. Podem ser sorteadas dezenas entre 01 e 60. A série randômica, a ser usada para os testes, foi obtida pela coleção das primeiras bolas sorteadas ao longo de toda a história da loteria Mega-Sena (CEF, 2014). Parte desta série temporal está mostrada na Figura 3b.

O preditor escolhido corresponde a um polinômio de grau quatro, num espaço reconstruído de três dimensões. A partir do arquivo ‘mega1.txt’ — que armazena a série temporal —, as rotinas são empregadas de maneira idêntica aos casos anteriores:

```
[> V4 := VecTS (DataFile='mega1.txt', TimeLag=6, Dim=3):
[> final := 1633:
[> LinMap4 := LinGfiTS (V4, final, Degree=4, Level=32, PT=1):
[> sigma4 := ConfiTS(LinMap4, V4, final, Level=32, PT=1);
```

$$\sigma_4 := 15.36455808.$$

Nesta aplicação, a distribuição dos resíduos pode ser considerada gaussiana. O histograma e a curva normal padronizada — mostrados na Figura 16a — reproduzem a impressão gerada a partir do comando `ConfiTS`. O desvio apurado  $\sigma_4$  tem uma magnitude de cerca de um quarto de todo o espectro de dezenas que podem ser sorteadas. O gráfico da Figura 16b apresenta este desvio como a barra de erro para a aplicação do mapa `LinMap4`, nos pontos do mapeamento global.

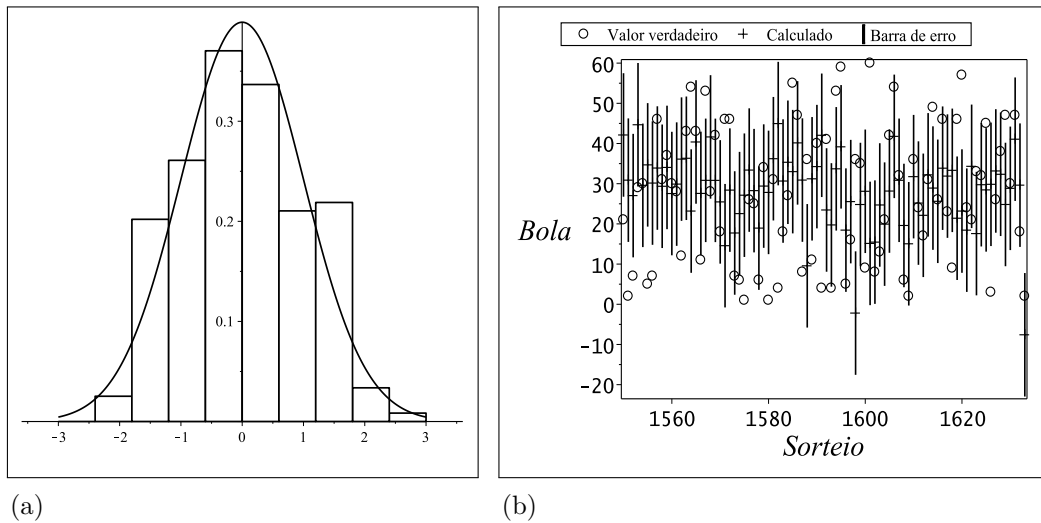
```
[> V4[1634][1] - ForecastTS(Vects=V4, Map=LinMap4, Position=1633);
```

$$11.31694096$$

O nível de confiança no resultado da previsão calculada pelo comando `ForecastTS` é de

$$\mathcal{L}_1 = 1 - \frac{2}{\sqrt{2\pi}} \int_{11.3/15.4}^{\infty} \exp\left\{-\frac{\epsilon_{\hat{r}-\tau}^2}{2\sigma_{\tau}^2}\right\} d(\epsilon_{\hat{r}-\tau}/\sigma_{\tau}) \cong 0.537.$$

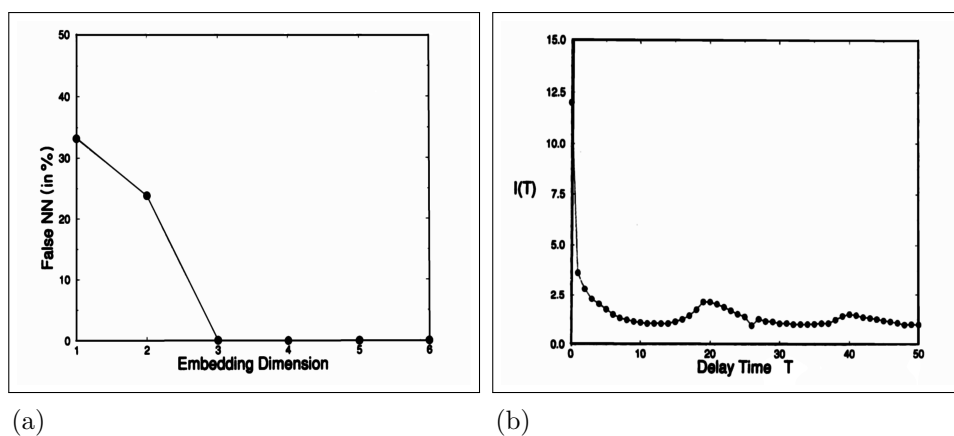
Figura 16 - Desvio na predição das dezenas sorteadas na Mega-Sena



Legenda: (a) Os resíduos possuem uma distribuição aceita como normal no teste de Shapiro-Wilk. (b) As longas barras de erro são compatíveis com a predição dos resultados de uma loteria.

Fonte: O autor, 2017.

Figura 17 - Falsos vizinhos próximos e mútua informação média



Legenda: Determinação dos parâmetros de imersão para o circuito caótico. (a) O percentual de falsos vizinhos próximos torna-se desprezível a partir de  $d_E = 3$ . (b) O delay time ótimo é  $T = 10$ .

Fonte: BROWN, RULKOV, TRACY, 1994, p. 3792.

O confronto do valor previsto com o calculado resulta num erro de predição em torno de 11 dezenas. Considerando que o erro tolerável de 45 neste caso, o resultado está dentro do esperado para um sistema absolutamente randômico. Em outras palavras, o método da **Aproximação Global** reflete a realidade de que qualquer bola pode ser sorteada na loteria.

## 5.6 Viabilidade dos preditores não polinomiais

Uma das funcionalidades do programa LinGfiTS permite que os preditores possuam qualquer forma funcional nas coordenadas utilizadas no esquema da reconstrução (ALVES; DUARTE; da MOTA, 2017a). Este recurso abre a perspectiva para a obtenção de mapas mais acurados para a previsão com a técnica global. A viabilidade da escolha da função preditiva — pelo argumento opcional **Func** — será demonstrada com a mesma série temporal que vem sendo explorada desde a Seção 4.5. O conjunto de observáveis utilizados também está armazenado no arquivo ‘ts37.txt’ e o restante do procedimento é idêntico ao que vem sendo empregado.

O mapa LinMap1 — apresentado na Seção 5.1 — corresponde ao preditor polinomial (1), que passa a ser identificado neste parágrafo como  $\mathcal{P}_{pol}(|x_r\rangle)$ .

$$\mathcal{P}_{pol}(|x_r\rangle) = c_1 x_{1r} + c_2 x_{2r} + c_3 x_{3r} + c_4 x_{1r}^2 + c_5 x_{1r} x_{2r} + c_6 x_{1r} x_{3r} + c_7 x_{2r}^2 + c_8 x_{2r} x_{3r} + c_9 x_{3r}^2 \quad (64)$$

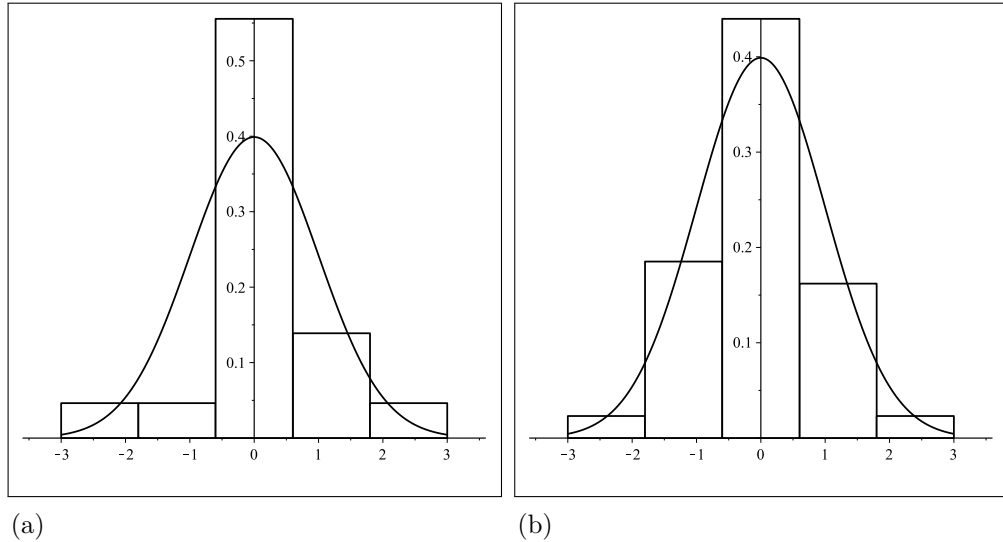
Os preditores  $\mathcal{P}_{pot}$  e  $\mathcal{P}_{log}$  são propostos para demonstrar que a escolha da forma funcional tem implicações diretas no mapeamento global e no resultado da predição.

$$\begin{aligned} \mathcal{P}_{pot}(|x_r\rangle) &= c_1 \frac{x_{1r}}{|x_{1r}|} |x_{1r}|^{0.9} + c_2 \frac{x_{1r}}{|x_{1r}|} |x_{1r}|^{1.1} + c_3 \frac{x_{2r}}{|x_{2r}|} |x_{2r}|^{0.9} + c_4 \frac{x_{2r}}{|x_{2r}|} |x_{2r}|^{1.1} \\ &+ c_5 \frac{x_{3r}}{|x_{3r}|} |x_{3r}|^{0.9} + c_6 \frac{x_{3r}}{|x_{3r}|} |x_{3r}|^{1.1} + c_7 \frac{x_{1r} x_{2r} x_{3r}}{|x_{1r} x_{2r} x_{3r}|} |x_{1r} x_{2r} x_{3r}|^{0.9} \\ &+ c_8 \frac{x_{1r} x_{2r} x_{3r}}{|x_{1r} x_{2r} x_{3r}|} |x_{1r} x_{2r} x_{3r}|^{1.1} \end{aligned} \quad (65)$$

$$\begin{aligned} \mathcal{P}_{log}(|x_r\rangle) &= c_1 x_{1r} + c_2 x_{2r} + c_3 x_{3r} + c_4 \ln \left( 1 + \frac{1}{10} \cos(x_{1r}) \right) + c_5 \ln \left( 1 + \frac{1}{10} \cos(x_{2r}) \right) \\ &+ c_6 \ln \left( 1 + \frac{1}{10} \cos(x_{3r}) \right) + c_7 \ln \left( 1 + \frac{1}{10} \sin(x_{1r}) \right) \\ &+ c_8 \ln \left( 1 + \frac{1}{10} \sin(x_{2r}) \right) + c_9 \ln \left( 1 + \frac{1}{10} \sin(x_{3r}) \right) \end{aligned} \quad (66)$$

Na aplicação das três funções preditivas para o mesmo intervalo da série temporal

Figura 18 - Distribuição dos resíduos com os preditores  $\mathcal{P}_{pol}(|x_r|)$  e  $\mathcal{P}_{log}(|x_r|)$



Legenda: (a) Distribuição dos resíduos com o preditor  $\mathcal{P}_{pol}(|x_r|)$ . (b) Gráfico plenamente compatível com o p-valor 0.4102, calculado para o mapa  $\mathcal{P}_{log}(|x_r|)$ .

Fonte: O autor, 2017.

Tabela 12 - Comparação entre preditores

preditor	erro	desvio	parâmetros	p-valor	regra dos $3\sigma$
$\mathcal{P}_{pol}( x_r )$	0.749	0.207	9	0.0016	não atende
$\mathcal{P}_{pot}( x_r )$	0.244	0.175	8	0.2021	atende
$\mathcal{P}_{log}( x_r )$	0.169	0.276	9	0.4102	atende

Legenda: Os resultados são obtidos pelas rotinas **LinGfiTS** e **ConfiTS** com os mesmos vetores reconstruídos. Cada p-valor acima foi calculado pela mesma rotina que aplica o teste de Shapiro-Wilk.

Fonte: O autor, 2017.

— definido pelo argumento `Level=5` — e adotando `V[722]` [1] como o último observável conhecido, os vetores que participam do mapeamento global são

[627, 628, 629, 630, 631, 632, 645, 646, 647, 648, 649, 650, 663, 664, 665, 666, 667, 668, 681, 682, 683, 684, 685, 686, 699, 700, 701, 702, 703, 704, 717, 718, 719, 720, 721, 722].

A lista acima é impressa na tela pelas rotinas `LinGfiTS` e `ConfITS`. Os resultados da previsão e do desvio para o observável de ordem 723, bem como o p-valor calculado no teste de Shapiro-Wilk, são mostrados na Tabela 12. Neste caso, os dois mapas não polinomiais são mais acurados que o polinomial. A função das variáveis de reconstrução que resulta numa previsão mais próxima do valor verdadeiro é a estabelecida em  $\mathcal{P}_{log}(|x_r\rangle)$  (66). Todavia, o preditor  $\mathcal{P}_{pot}(|x_r\rangle)$  (65) apresenta um desvio calculado menor que o anterior —  $\sigma_{pot} = 0.244$  contra  $\sigma_{log} = 0.276$ . Podemos dizer que ambas as alternativas não polinomiais para o mapeamento global são competitivas (ALVES; DUARTE; da MOTA, 2017b).

Um outro aspecto relevante diz respeito à relação entre o p-valor e o confronto com a Regra dos Três Sigmas. Novamente, o preditor polinomial — cuja hipótese de normalidade da distribuição dos resíduos é rejeitada no teste de Shapiro-Wilk — apresenta baixo p-valor e um erro na predição maior que o triplo do desvio calculado. Por outro lado, tanto  $\mathcal{P}_{log}(|x_r\rangle)$  como  $\mathcal{P}_{pot}(|x_r\rangle)$  estão associados a p-valores aceitáveis e resultam em predições com erros entre  $\sigma$  e  $2\sigma$ . A Figura 18 mostra os gráficos obtidos com a rotina `ConfITS` para o melhor (ver Figura 18b) e o pior desempenho (ver Figura 18a) neste quesito. O histograma para os desvios obtidos com o mapa polinomial — baseado em  $\mathcal{P}_{pol}(|x_r\rangle)$  — está incompatível com uma distribuição gaussiana dos resíduos, porém de acordo com o baixo p-valor 0.0016 apresentado na Tabela 12 (ALVES; DUARTE; da MOTA, 2017b).

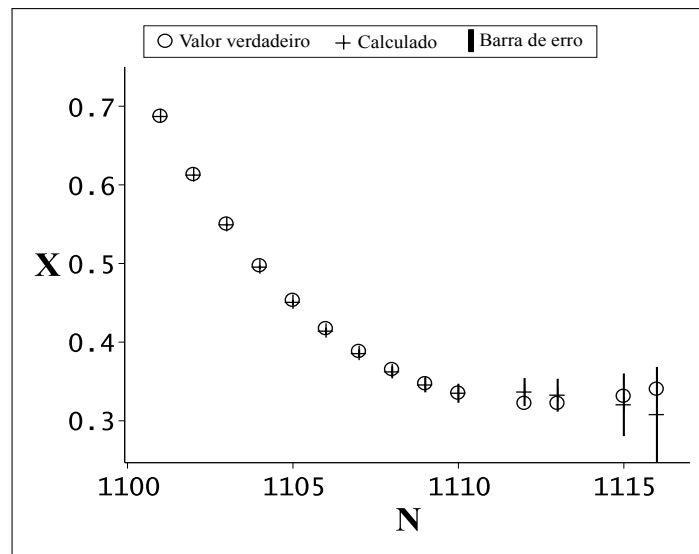
## 5.7 Variável dinâmica $X$ do Sistema de Lorenz

Para testar o método de caracterização dinâmica (ver Capítulo 3), a forma funcional selecionada consiste no polinômio de quinto grau no espaço de reconstrução tridimensional. Alguns dos 55 termos deste modelo preditivo são

$$\mathcal{P}_\tau(x_{1r}, x_{2r}, x_{3r}) = c_1x_{1r} + c_2x_{2r} + c_3x_{3r} + \dots + c_{55}x_{3r}^5. \quad (67)$$

A série para o Sistema de Lorenz — cujo Diagrama Acurácia-Desvio Logarítmico já foi apresentado na Figura 8 — foi gerada com a rotina `lorenz` do pacote `Tisean` e armazenada no arquivo `LorenzX.txt` (HEGGER; KANTZ; SCHREIBER, 1999). Existem

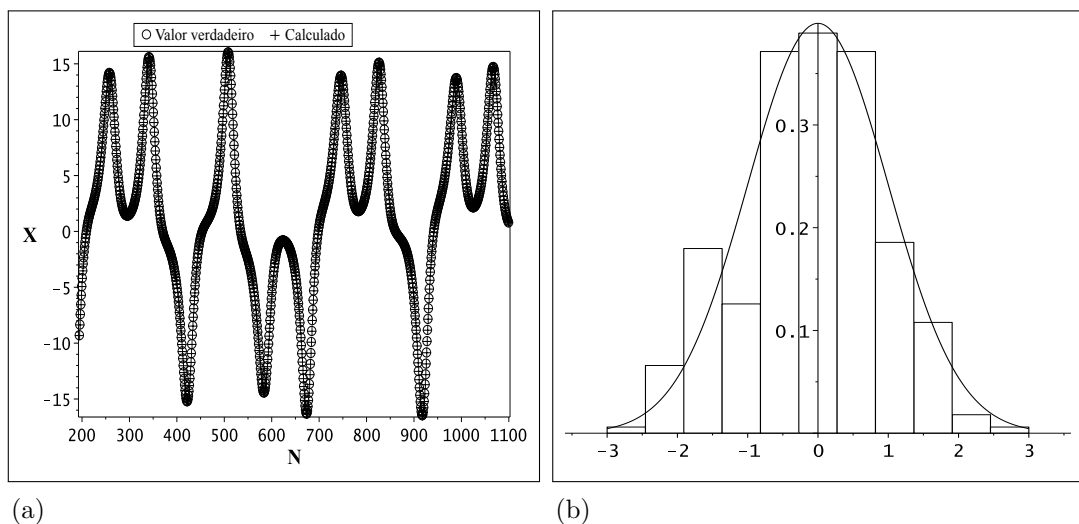
Figura 19 - Desvios numa série temporal caótica



Legenda: O crescimento das barras de erro manifesta o caráter caótico do Sistema de Lorenz.

Fonte: O autor, 2017.

Figura 20 - Preditor polinomial aplicado na série temporal LorenzX



Legenda: (a) Variável dinâmica  $X$  versus a ordem  $N$  do observável. O mapa global obtido com a forma funcional (63) apresenta uma descrição da série temporal inteiramente satisfatória com o parâmetro  $\tau = 1$ . (b) A curva gaussiana revela-se adequada para descrever a distribuição dos resíduos no *global fitting*.

Fonte: O autor, 2017.

1200 observáveis nesta série que cobrem 12 unidades de tempo  $u_t$ . Os expoentes de Lyapunov — calculados pela própria rotina — são 0.965908289,  $-0.0467140898$  e  $-13.9133472$  em  $u_t^{-1}$ . O expoente de Lyapunov positivo implica que o sistema é caótico (ver Definição 10).

O gráfico da Figura 20a mostra o resultado da aplicação do preditor (67) com o parâmetro  $\tau = 1$ . O teste de normalidade com o mapa gerado apresenta resultado positivo. De fato, o histograma e a curva normal padronizada (ver Figura 20b) permitem uma visualização convincente de que a distribuição dos resíduos pode ser bem aproximada por uma curva gaussiana.

O pacote `LinMapTS` neste caso específico foi evocado — no ambiente Maple — da mesma maneira que foi feita na Seção 5.1.

```
[> V5 := VecTS (DataFile='LorenzX.txt', TimeLag=6, Dim=3):
[> final := 1100:
[> LinMap5 := LinGfiTS (V5,final,Degree=3,Level=50,PT=1,Analysis=1);
[> sigma5 := ConfiTS(LinMap5,V5,final,Level=50,PT=1);
```

Para a determinação dos coeficientes do preditor e o cálculo do desvio, a rotina empregou 306 vetores do espaço reconstruído. No intuito de reforçar a explicação a respeito do modo de escolha dos vetores que participam do mapeamento global e da qualificação dos preditores, uma parte da lista de vetores impressa pelas rotinas `LinGfiTS`, `ConfiTS` e `DynCharTS` é reproduzida abaixo.

```
[195, 196, 197, 198, 199, 200, 213, 214, 215, 216, 217, 218, 231,
 232, 233,234, 235, 236, ..., 1095, 1096, 1097, 1098, 1099, 1100]
```

Tomando como exemplo os observáveis de ordem 207 e 201 na série, estes participam da minimização como componentes do vetor  $V[213]$ . Isto é, são incluídos no processo como  $V[213][3]$  e  $V[213][2]$ . Entretanto, os vetores  $V[207]$  e  $V[201]$  não tomam parte no *global fitting*. Por outro lado, todos os 906 observáveis do intervalo — ou seja,  $1100 - 195 + 1 = 906$  — são contabilizados na detecção de *outliers* com o programa `DynCharTS`.

Quanto à caracterização propriamente dita, a natureza caótica se manifesta claramente no Diagrama Acurácia-Desvio Logarítmico da Figura 8. As escalas logarítmicas nos dois eixos do gráfico reforçam o decaimento de  $\mathcal{A}_\tau$  e o crescimento de  $\mathcal{D}_\tau$ . O diagrama está em pleno acordo com o comportamento esperado para estas quantidades estatísticas (ver Tabela 7).

Outra análise esclarecedora do tipo de dinâmica impressa na série temporal diz respeito à evolução dos desvios nas predições de acordo com o parâmetro  $\tau$ . O gráfico da Figura 19 corresponde ao mesmo intervalo da série do Diagrama Acurácia-Desvio Logarítmico. O último observável utilizado no mapeamento global ocupa a ordem 1100.

Então, a predição para o observável de ordem 1109, por exemplo, empregou o parâmetro  $\tau = 9$ . O crescimento das barras de erro — que pode ser visualizado no gráfico — confirma o diagnóstico de uma dinâmica caótica para esta série.

## 5.8 Série experimental caótica revisitada

É instrutivo verificar a performance do método de caracterização num conjunto de dados experimentais. No circuito esquematizado na Figura 14, a voltagem medida no capacitor **C** atende muito bem a esta proposta, uma vez que a sua dinâmica é reconhecidamente caótica (BROWN; RULKOV; TRACY, 1994). Além de ilustrar a aplicabilidade dos diagramas no contexto da experiência, este conjunto de medidas constitui a segunda série caótica do teste.

Num primeiro exemplo da caracterização dinâmica com o programa DynCharTS, apenas parte da lista dos vetores do mapeamento global e o resultado da qualificação dos preditores correspondente ao parâmetro  $\tau = 1$  — no total são 16 gráficos, de acordo com o argumento PT=16 — são reproduzidos a seguir.

```
Z[dyn]:=DynCharTS(DataFile='circuit.txt',Degree=5,PT=16,Analysis=1);
```

*List of reconstructed vectors in global fitting*

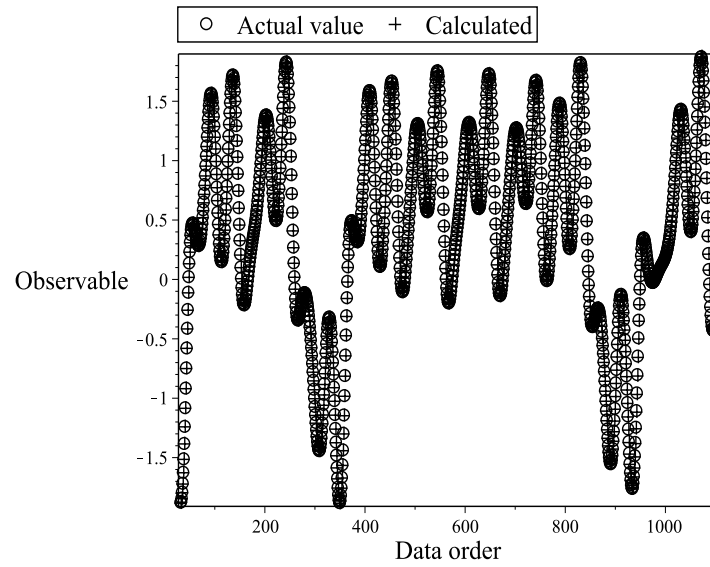
[43,44,45,46,47,48,61,62,63,64,65,66,79, . . . ,1159,1160,1161,1162,1163,  
1164,1177,1178, 1179,1180,1181,1182, 1195,1196,1197,1198, 1199, 1200]

*The predictor for the global fittings has the functional form*

$$\begin{aligned}
& a_{47}X_1^2 + a_{16}X_2^5 + a_{36}X_3^4 + a_1X_1^5 + a_{50}X_2^2 + a_{53}X_1 + a_{43}X_2^3 + a_{22}X_1^4 + a_{54}X_2 + \\
& a_{37}X_1^3 + a_{46}X_3^3 + a_{52}X_3^2 + a_{32}X_2^4 + a_{21}X_3^5 + a_{55}X_3 + a_9X_3^2X_1^2X_2 + a_{26}X_3X_1^2X_2 + \\
& a_{35}X_3^3X_2 + a_{29}X_3X_1X_2^2 + a_{39}X_1^2X_3 + a_{48}X_1X_2 + a_{24}X_3X_1^3 + a_{40}X_1X_2^2 + \\
& a_{13}X_3^2X_1X_2^2 + a_{30}X_3^2X_1X_2 + a_5X_2X_3X_1^3 + a_{27}X_3^2X_1^2 + a_{23}X_2X_1^3 + a_{42}X_1X_3^2 + \\
& a_{34}X_3^2X_2^2 + a_{12}X_2^3X_3X_1 + a_{33}X_3X_2^3 + a_{51}X_2X_3 + a_{18}X_3^2X_2^3 + a_{10}X_3^3X_1^2 + \\
& a_{44}X_2^2X_3 + a_{41}X_1X_2X_3 + a_{25}X_2^2X_1^2 + a_8X_2^2X_3X_1^2 + a_6X_3^2X_1^3 + a_{14}X_3^3X_1X_2 + \\
& a_{20}X_3^4X_2 + a_{15}X_3^4X_1 + a_{17}X_2^4X_3 + a_4X_2^2X_1^3 + a_3X_1^4X_3 + a_2X_1^4X_2 + a_{45}X_2X_3^2 + \\
& a_{19}X_3^3X_2^2 + a_{11}X_2^4X_1 + a_{49}X_1X_3 + a_7X_2^3X_1^2 + a_{38}X_1^2X_2 + a_{31}X_3^3X_1 + a_{28}X_2^3X_1
\end{aligned}$$

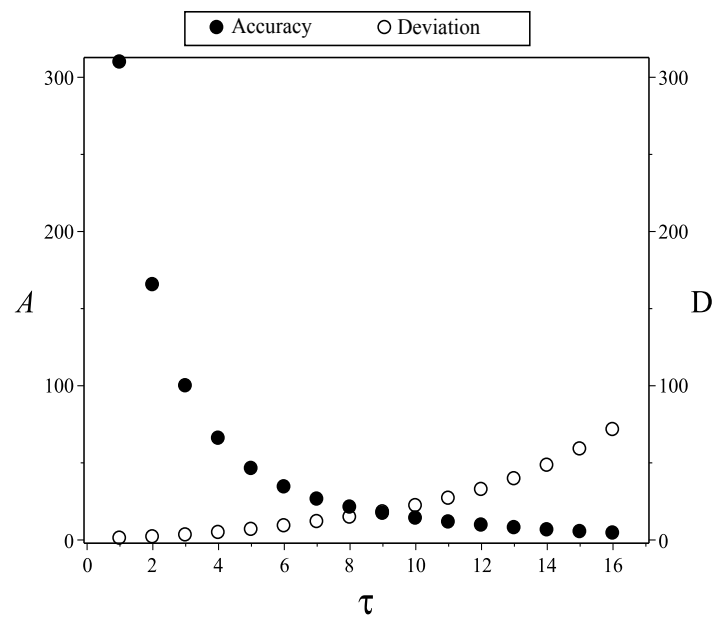
*Analysis for the parameter*

$$\tau = 1$$

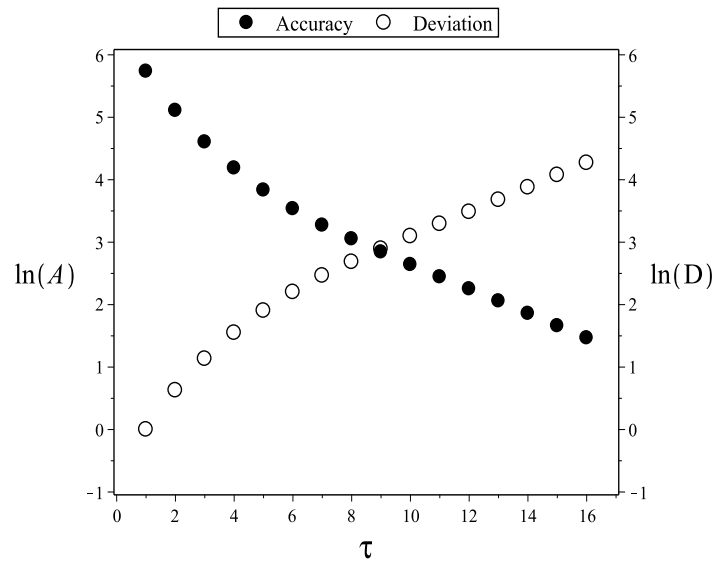


*The global fitting meets all the criteria of qualification.*

*Diagram Accuracy-Deviation:*



*Logarithmic Diagram Accuracy-Deviation:*



$$Z_{dyn} := 319.8242146 + 16.69686340 I$$

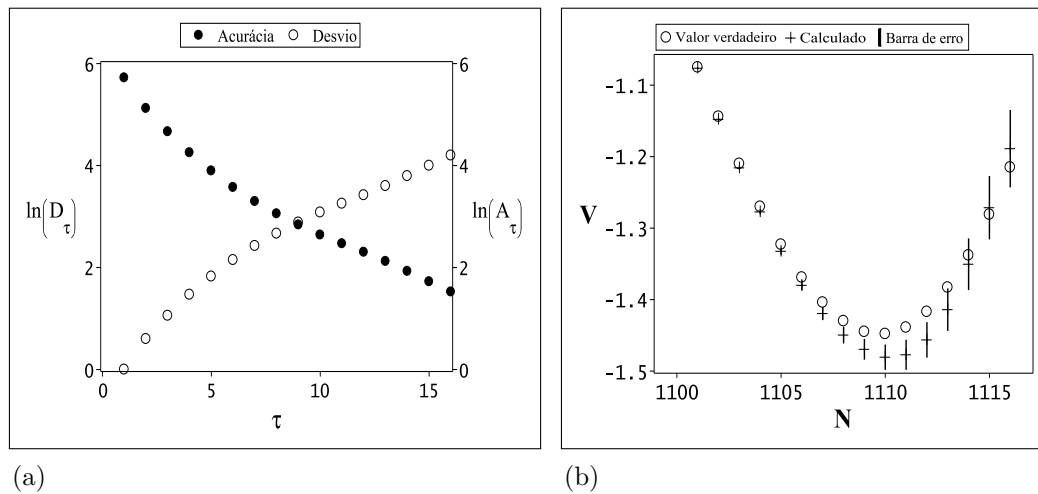
O procedimento computacional — bem como a escolha da ordem dos observáveis — permanece idêntico ao adotado no Sistema de Lorenz. A análise das curvas de acurácia e de desvio no gráfico da Figura 21a revelam que, com o aumento do tempo de predição  $\tau\Delta t$ , a previsibilidade se torna cada vez mais difícil. Nesta série experimental, o caos deixa seu registro gráfico no Diagrama Acurácia-Desvio Logarítmico, assim como no teste com o Sistema de Lorenz (ver Figura 8) — no qual a série temporal foi gerada por integração numérica.

Os erros da predição aumentam de acordo com o incremento do parâmetro  $\tau$ , assim como ocorre no caso anterior. Um ponto a favor da teoria neste quesito é confronto do valor verdadeiro com a barra de erro nos gráficos das Figuras 19 e 21b; visualmente é possível observar que em nenhuma predição o erro admissível de  $3\sigma_\tau$  foi excedido. Vale destacar que, nestas duas series caóticas, a presença de *outliers* no mapeamento global foi diminuta.

## 5.9 Sinal periódico e série randômica

Duas dinâmicas de evolução temporal diametralmente opostas são adequadas para testar a sensibilidade do método. Um sinal elétrico típico constitui o modelo usado para gerar uma série temporal periódica e, a priori, completamente previsível. No outro extremo, a coleção de resultados de uma loteria fornece uma série desprovida de qualquer determinismo. As duas séries também possuem, como aquelas empregadas anteriormente, 1200 observáveis.

Figura 21 - Caracterização da série caótica experimental



Legenda: (a) As curvas de acurácia e de desvio indicam que a voltagem medida no circuito esquematizado na Figura 14 tem uma evolução temporal caótica. (b) Gráfico da medida de voltagem  $V$  em função da sua ordem  $N$  na série temporal. Assim como na série temporal gerada para o Sistema de Lorenz da Figura 19, as barras de erro apresentam um crescimento compatível com a presença de caos.

Fonte: O autor, 2017.

A função

$$g(t) = 3 \sin(4\pi t) + 7 \cos(3\pi t) \quad (68)$$

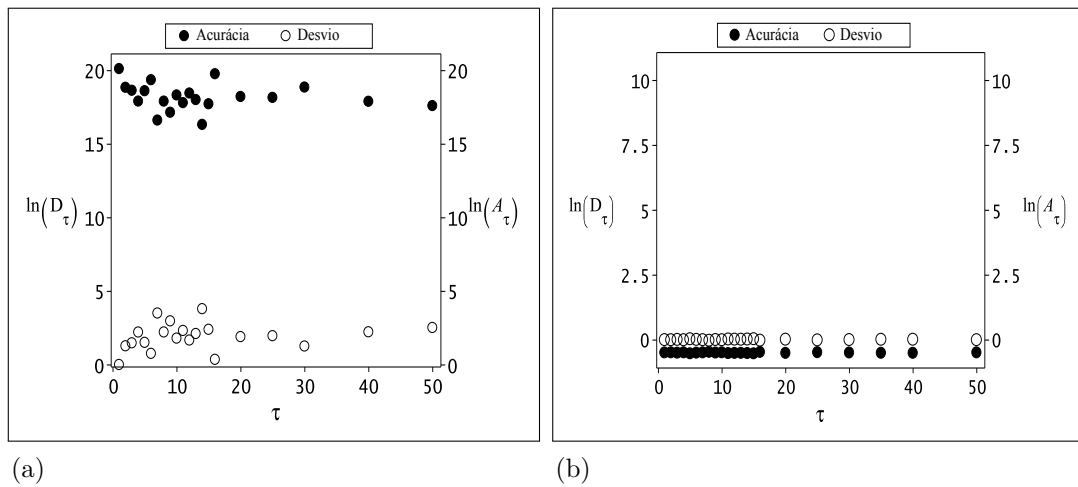
— cobrindo 24 unidades de tempo  $u_t$  — foi usada para gerar a série periódica (MANDAL, 2007). A mesma série aleatória formada pelas primeiras dezenas da loteria Mega-Sena — já explorada na Seção 5.5 — atende ao propósito do teste nesta seção.

O procedimento computacional é idêntico<sup>11</sup> ao aplicado nas séries de teste caóticas (ver Seções 5.7 e 5.8). Os diagramas correspondentes na Figura 22 apresentam bandas bem definidas de acurácia e de desvio. No sistema periódico, verifica-se um distanciamento das ‘bolas fechadas’ — na região superior do gráfico — das ‘bolas abertas’, que se mantém na região inferior da Figura 22a. De fato, no caso de regularidade e previsibilidade extremas, não se espera queda significativa de acurácia e nem aumento de desvios quando os tempos de predição  $\tau\Delta t$  se tornam maiores.

Já no diagrama da Figura 22b, a banda de acurácia permanece alinhada com a de desvio. Ambas possuem ordenadas próximas de zero. Tanto esta série randômica, quanto a periódica, apresentam desvios que não manifestam crescimento na escala logarítmica.

<sup>11</sup> Os preditores adotados nas séries de Lorenz, do circuito caótico, periódica e randômica — empregadas no teste da caracterização dinâmica — possuem a mesma forma funcional (67).

Figura 22 - Testes com as séries periódica e aleatória



Legenda: (a) Sinal periódico  $g(t) = 3 \sin(4\pi t) + 7 \cos(3\pi t)$ . Bandas de alta acurácia e de desvio num mesmo patamar estão de acordo com as predições acuradas em séries temporais periódicas. (b) Neste diagrama, a banda de baixa acurácia corresponde à ausência de uma dinâmica determinística em séries randômicas.

Fonte: O autor, 2017.

Não há nenhum motivo para se esperar que a previsibilidade numa loteria caia com o transcorrer do tempo. Assim como são grandes os desvios na predição da primeira dezena em um sorteio, estes serão também enormes se a primeira bola for sorteada dez semanas depois.

## 5.10 O Número de Manchas Solares e o Índice Dow Jones

Nos diagramas para o Sistema de Lorenz e o Circuito caótico (ver Figuras 8 e 21), a assinatura do caos foi obtida pelo simultâneo decaimento da acurácia e crescimento do desvio relativo. A análise de observáveis extraídos do mundo real tem a perspectiva de avançar no entendimento da dinâmica dos fenômenos complexos. Retomando os exemplos dados na Introdução deste trabalho, a proposta desta seção consiste no estudo de séries temporais formadas pelo Número de Manchas Solares e pelo Índice Dow Jones.

Para análise, são tomadas as médias mensais do Número de Manchas Solares — fornecidas por Hathaway (2015) — e os índices financeiros de toda a história do mercado de ações Dow Jones — disponibilizados em Indices (2015).

$$\mathcal{P}_\tau(x_{1r}, x_{2r}, x_{3r}, x_{4r}) = c_1 x_{1r} + c_2 x_{2r} + c_3 x_{3r} + c_4 x_{4r} + \dots + c_{14} x_{4r}^2 \quad (69)$$

Em ambos os casos, o esquema da reconstrução tem sua base num espaço euclidiano

Tabela 13 - Qualificação dos preditores para as séries temporais caóticas

QP	CS	Parâmetro $\tau$															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	LX	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	CC	×	×	×	×	×	×	×	×	×	×						
	DJ	×	×	×	×	×	×	×	×								
	SN																
2	LX	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	CC	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	DJ	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	SN	×	×	×	×												
3	LX	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	CC	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	DJ	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	SN	×	×	×	×	×	×	×	×	×	×	×	×				
4	LX	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	CC	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	DJ	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	SN	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
5	LX	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	CC	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	DJ	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	SN	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×

Legenda: **QP**≡Critério de qualificação para o preditor. O símbolo × indica que o mapa global atende ao critério. **CS**≡Série temporal caótica, **LX**≡Variável dinâmica  $X$  para o Sistema de Lorenz, **CC**≡Circuito caótico, **SN**≡Número de Manchas Solares e **DJ**≡Índice Dow Jones.

Fonte: O autor, 2017.

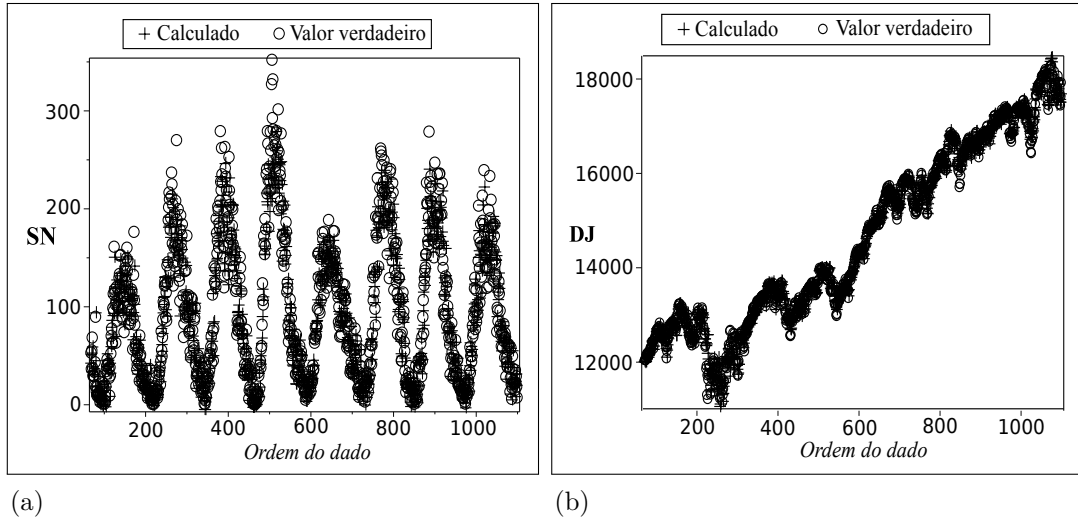
Tabela 14 - Resultados para o quantificador de caos  $Z_{dyn}$ 

CS	Preditores	$d_E$	Quantidade estatística	
			$\mathcal{A}_1$	$\lambda_{dyn}$
LX	$deg = 5$	3	3500.06	307.639
CC	$deg = 5$	3	319.82	28.279
DJ	$deg = 3$	4	46.66	2.599
SN	$deg = 2$	4	1.27	0.049

Legenda: Cada espaço de reconstrução tem dimensão  $d_E$  e o grau do preditor polinomial é designado por  $deg$ .

Fonte: O autor, 2017.

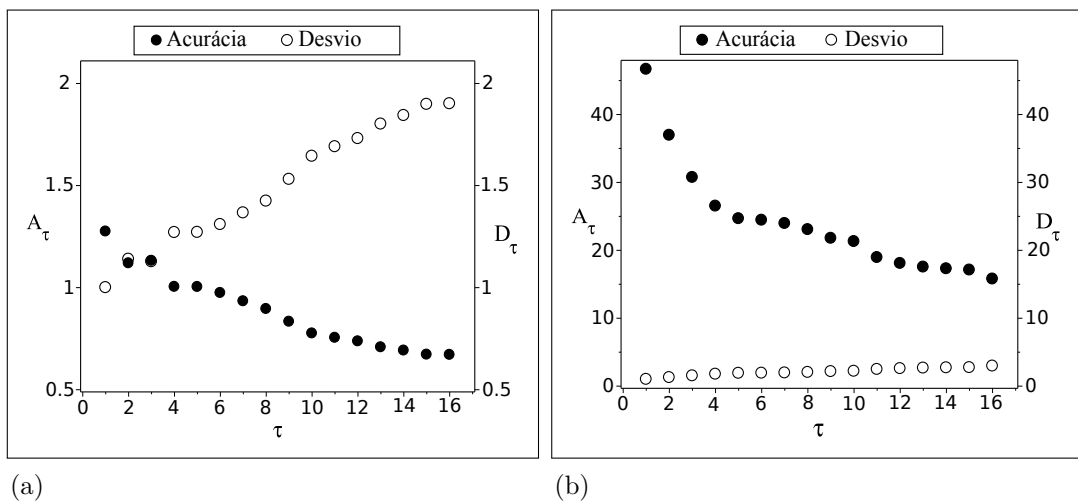
Figura 23 - Mapeamento global para os estudos de caso ( $\tau = 1$ )



Legenda: (a) Número de Manchas Solares  $SN$ . (b) Índice Dow Jones  $DJ$ .

Fonte: O autor, 2017.

Figura 24 - Diagramas Acurácia-Desvio para os estudos de caso



Legenda: (a) Número de Manchas Solares. (b) Índice Dow Jones.

Fonte: O autor, 2017.

de quatro dimensões. Os preditores escolhidos são polinômios de segundo grau (69) — para o Número de Manchas Solares — e de terceiro grau (70) — para o Índice Dow Jones.

$$\mathcal{P}_\tau(x_{1r}, x_{2r}, x_{3r}, x_{4r}) = c_1x_{1r} + c_2x_{2r} + c_3x_{3r} + c_4x_{4r} + \dots + c_{34}x_{4r}^3 \quad (70)$$

A Figura 23 apresenta o resultado da aplicação dos preditores para as duas séries com a seleção do parâmetro  $\tau = 1$ . Os mapeamentos globais para estas duas séries temporais do mundo real registram uma presença de *outliers* mais significativa que as duas séries caóticas já discutidas neste capítulo. A Tabela 13 mostra uma comparação entre os resultados da qualificação de preditores para as séries temporais deste trabalho que manifestam evolução caótica.

A partir dos diagramas da Figura 24, pode-se concluir que o Índice Dow Jones e o Número de Manchas Solares possuem sua origem em sistemas onde existe a presença de caos. O comportamento qualitativo do crescimento das curvas de desvio e decaimento da acurácia está presente nas duas representações gráficas. Contudo, alguns aspectos a respeito do emprego da **Aproximação Global Polinomial** — no caso do sistema astrofísico — requerem considerações adicionais.

As predições para o menor tempo de predição — especificadas pelo parâmetro  $\tau = 1$  — já apresentam baixa acurácia. Mesmo na reconstrução em um espaço de quatro dimensões, o processo de *global fitting* não permite a obtenção de baixos valores para o desvio  $\sigma_\tau$ . A própria natureza dos dados impõe uma séria limitação na capacidade preditiva do método. Esta dificuldade em obter mapas acurados para a série tem um efeito direto sobre a magnitude do quantificador de caos  $Z_{dyn}$ .

O mapa global de melhor qualidade neste trabalho foi gerado para a série temporal formada pela variável  $X$  do Sistema de Lorenz (8). A Tabela 14 apresenta as quantidades estatísticas calculadas para as quatro séries caóticas discutidas neste capítulo. De fato, o caso mais difícil para a caracterização da dinâmica corresponde à série temporal formada pelo Número de Manchas Solares porque os valores de  $\mathcal{A}_1$  e  $\lambda_{dyn}$  são os mais baixos dos casos em estudo.

Foi necessário reduzir a escala logarítmica do diagrama da Figura 24a para que as curvas de acurácia e de desvio pudessem ser avaliadas. Por outro lado, existe, neste conjunto de observáveis, uma informação sobre a dinâmica do sistema que não deve ser negligenciada. Apesar de serem pouco expressivas as curvas de acurácia e desvio, traços da fenomenologia associada ainda podem ser identificados. A partir do Diagrama Acurácia-Desvio obtido e do valor calculado para o quantificador  $Z_{dyn}$ , *a atividade solar parece ser caótica* (PETROVAY, 2010).

## CONCLUSÃO

A Aproximação Global atende aos propósitos da predição e da caracterização dinâmica na perspectiva de um problema inverso. O primeiro Teorema de Takens garante que o espaço de fase de um sistema dinâmico pode ser mergulhado num espaço euclidiano pelo método do *delay time*. Ou seja, a dinâmica original é preservada no esquema da reconstrução. Esta solidez matemática sustenta a técnica aproximativa desenvolvida.

No desenvolvimento da teoria, a restrição na forma da função custo resulta na seleção de preditores que se mostram altamente convenientes do ponto de vista matemático e computacional. Solucionar uma simples equação matricial significa determinar os parâmetros do preditor. A forma padrão da função preditiva pode ser empregada para diferentes tempos de predição. Este é um aspecto de grande utilidade, porque permite diagnosticar o comportamento dinâmico da série pela evolução dos erros na previsão.

O nível de confiança na predição é bem estabelecido desde que os resíduos do ajuste de dados no mapeamento global admitam uma distribuição normal. O teste de normalidade de Shapiro-Wilk cumpre bem a tarefa de avaliar se a forma gaussiana assumida para os resíduos é adequada, ou não, aos observáveis extraídos da série temporal. Sendo positivo o resultado do teste, há evidência estatística suficiente para considerar que o método dos mínimos quadrados — empregado na determinação do preditor — apresenta a melhor performance que pode ser obtida. Sendo assim, a questão da otimização tem um desfecho plenamente satisfatório, não sendo necessário buscar outros métodos de minimização diferente dos mínimos quadrados; basta modificar o modelo do preditor quando necessário.

Na nova versão do pacote `TimeS`, o mapeamento e o algoritmo que aperfeiçoa o resultado da aplicação do mapa global passam a admitir diferentes tempos de predição. O tempo necessário para o mapeamento global com o pacote `LinMapTS` foi reduzido drasticamente em relação ao requerido pela rotina `GfiTS` do pacote `TimeS`. Este ganho de eficiência permite o uso do pacote para séries com grandes quantidades de observáveis e parâmetros de ajuste, com baixo esforço computacional. A praticidade da manipulação algébrica e os baixos tempos de execução facilitam a obtenção de preditores mais acurados.

Funções preditivas diferentes das polinomiais se constituem numa alternativa viável na Análise de Séries Temporais. O exemplo do preditor baseado em termos de potência apresentou um desempenho superior ao polinomial, utilizando o mesmo número de graus de liberdade. Também neste aspecto, o pacote `LinMapTS` coloca à disposição do pesquisador um maior poder de análise e predição.

Na análise dos resíduos com o programa `ConfITS`, ficou demonstrado o acordo entre o erro da predição e o desvio calculado. Foi verificada também a correspondência

da estatística calculada com os gráficos impressos pela rotina. O exemplo apresentado para o resultado negativo no teste revelou também uma histograma incompatível com a distribuição normal dos resíduos.

As rotinas desenvolvidas mostraram-se adequadas à manipulação da série temporal obtida de um experimento, quando aplicadas ao sinal elétrico caótico. Neste exemplo, também foi verificada a consistência dos métodos de escolha dos parâmetros de reconstrução, notadamente a mútua informação média e os falsos vizinhos próximos. Outra categoria de conjunto de observáveis extraído do mundo real — a das séries randômicas — foi empregada para verificar a sensibilidade do procedimento `ConfITS`. Os desvios calculados pela rotina estão de acordo com ausência de uma dinâmica nas dezenas sorteadas numa loteria.

Na caracterização da série temporal, as quantidades estatísticas  $\mathcal{A}_\tau$  e  $\mathcal{D}_\tau$  revelam o tipo de dinâmica responsável pela evolução temporal nos Diagramas Acurácia-Desvio e Acurácia-Desvio Logarítmico. O método desenvolvido apresenta um novo quantificador para o caos que contorna o problema da identificação dos expoentes espúrios de Lyapunov e evita as dificuldades do cálculo destes expoentes no esquema da reconstrução. Os testes realizados nas séries de diferentes naturezas — caótica, periódica e randômica — comprovaram que a dinâmica subjacente da série temporal deixa sua assinatura nos diagramas e se manifesta na magnitude de  $Z_{dyn}$ . Nas duas séries caóticas em que o crescimento das barras de erro foi investigado graficamente, houve concordância com a caracterização dinâmica obtida com diagramas e quantificadores de caos  $Z_{dyn}$ .

No estudo dos índices do mercado financeiro, as curvas de acurácia e de desvio sustentam que, no intervalo da série temporal sob análise, a dinâmica é caótica. Um procedimento que se mostrou útil no estudo deste caso foi o da qualificação dos preditores empregado pelo programa `DynCharTS`, que permitiu o controle da qualidade do mapeamento global para diversos tempos de predição.

Não foi possível obter preditores que apresentassem baixa frequência de *outliers* — segundo todos os critérios de qualificação — para o caso da série temporal formada pelo Número de Manchas Solares. Como consequência, foi necessário usar uma escala mais reduzida no Diagrama Acurácia-Desvio e o quantificador de caos apresentou a magnitude mais baixa entre os demais sistemas caóticos. Apesar dessa condição indesejável — devida à qualidade dos dados disponíveis — ainda foi possível reconhecer o caos neste sistema astrofísico.

Os resultados obtidos confirmam que a teoria e as rotinas computacionais apresentadas nesta tese são adequadas tanto para a predição quanto para a caracterização dinâmica de uma série temporal. Logo, o objetivo inicial — de obter conhecimento a respeito de um sistema dinâmico a partir de um conjunto de observáveis ordenados no tempo — foi atingido; e por uma metodologia que permite a investigação de fenômenos não lineares em diversas áreas do conhecimento.

## REFERÊNCIAS

ALVES, P.; DUARTE, L.; da MOTA, L. *Computer Physics Communications Program Library*. Elsevier, 2016. Disponível em: <[http://cpc.cs.qub.ac.uk/summaries/AFAJ\\_v1\\_0.html](http://cpc.cs.qub.ac.uk/summaries/AFAJ_v1_0.html)>. Acesso em: 20 dez. 2016.

\_\_\_\_\_. *Computer Physics Communications Program Library*. Elsevier, 2016. Disponível em: <<https://data.mendeley.com/datasets/nhtmjc8yp8/1>>. Acesso em: 21 dez. 2016.

\_\_\_\_\_. Improvement in global forecast for chaotic time series. *Computer Physics Communications*, North-Holland, v. 207, p. 325 – 340, 2016.

\_\_\_\_\_. A new method for improved global mapping forecast. *Computer Physics Communications*, North-Holland, v. 207, p. 539 – 541, 2016.

\_\_\_\_\_. *Computer Physics Communications Program Library*. Elsevier, 2017. Disponível em: <<https://data.mendeley.com/datasets/pnhy9zymrp/1>>. Acesso em: 21 mar. 2017.

\_\_\_\_\_. Alternative predictors in chaotic time series. *Computer Physics Communications*, North-Holland, v. 215, p. 265 – 268, 2017.

\_\_\_\_\_. Dynamical characterization of a time series by the polynomial global approach. *Computer Physics Communications*, submetido à publicação em 2017.

\_\_\_\_\_. A high-performance method for data fitting and its analysis. *Computer Physics Communications*, submetido à publicação em 2017.

AOKI, N. *Topological theory of dynamical systems : recent advances*. Amsterdam: North-Holland, 1994.

ARNOLD, V. I. *Ordinary differential equations*. Cambridge: MIT Press, 1973.

BEVINGTON, P.; ROBINSON, D. *Data reduction and error analysis for the physical sciences*. 3rd. ed. New York: McGraw-Hill, 2003.

BLUMAN, G. *Symmetry and integration methods for differential equations*. New York: Springer, 2002.

BRANHAM R. L., J. Alternatives to least squares. *Astronomical Journal*, Madison, v. 87, p. 928–937, 1982.

BROWN, R.; RULKOV, N. F.; TRACY, E. R. Modeling and synchronizing chaotic systems from time-series data. *Phys. Rev. E*, American Physical Society, v. 49, p. 3784–3800, May 1994.

CARLI, H.; DUARTE, L.; da MOTA, L. *Computer Physics Communications Program Library*. Elsevier, 2014. Disponível em: <[http://cpc.cs.qub.ac.uk/summaries/AERW\\_v1\\_0.html](http://cpc.cs.qub.ac.uk/summaries/AERW_v1_0.html)>. Acesso em: 11 mar. 2015.

\_\_\_\_\_. A maple package for improved global mapping forecast. *Computer Physics Communications*, North-Holland, v. 185, n. 3, p. 1115 – 1129, 2014.

CASDAGLI, M. Nonlinear prediction of chaotic time series. *Physica D: Nonlinear Phenomena*, Amsterdam, v. 35, n. 3, p. 335 – 356, may 1989.

CEF. *Mega-sena - download de todos os resultados*. Caixa Econômica Federal. 2014. Divulga resultados de loterias. Disponível em: <<http://www1.caixa.gov.br/loterias/loterias/megasena/download.aspl>>. Acesso em: 13 set. 2014.

COLES, D. Transition in circular couette flow. *Journal of Fluid Mechanics*, Cambridge, v. 21, p. 385–425, mar 1965.

HATHAWAY, D. H. *The Sunspot Cycle*. NASA. 2015. Disponibiliza conteúdo sobre o Ciclo das Manchas Solares. Disponível em: <<http://solarscience.msfc.nasa.gov/SunspotCycle.shtml>>. Acesso em: 13 nov. 2015.

DEVORE, J. *Probability and statistics for engineering and the sciences*. Australia: Cengage Learning, 2016.

ECKMANN, J. P. et al. Liapunov exponents from time series. *Phys. Rev. A*, American Physical Society, v. 34, p. 4971–4979, Dec 1986.

FARMER, J. D.; SIDOROWICH, J. J. Predicting chaotic time series. *Phys. Rev. Lett.*, American Physical Society, v. 59, p. 845–848, Aug 1987.

FISHER, R. A. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, The Royal Society, London, v. 222, n. 594-604, p. 309–368, 1922.

FRASER, A. M.; SWINNEY, H. L. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A*, American Physical Society, v. 33, p. 1134–1140, Feb 1986.

FRIIS-CHRISTENSEN, E.; LASSEN, K. Length of the solar cycle: An indicator of solar activity closely associated with climate. *Science*, Washington, v. 254, n. 5032, p. 698–700, 1991.

- GALILEI, G. *Istoria e dimostrazioni intorno alle macchie solari e loro accidenti*. In Roma: Appresso G. Mascardi, 1613.
- GEDDES, K. O. *Algorithms for computer algebra*. Boston: Kluwer Academic, 1992.
- GRASSBERGER, P.; PROCACCIA, I. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, Amsterdam, v. 9, n. 1–2, p. 189 – 208, oct 1983.
- GREBOGI, C. et al. Strange attractors that are not chaotic. *Physica D: Nonlinear Phenomena*, Amsterdam, v. 13, n. 1–2, p. 261 – 268, aug 1984.
- HEGGER, R.; KANTZ, H.; SCHREIBER, T. Practical implementation of nonlinear time series methods: The tisean package. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, New York, v. 9, n. 2, p. 413–435, jun 1999.
- HEGGER, R.; KANTZ, H.; SCHREIBER, T. *Nonlinear Time Series Analysis*. 2016. Disponível em: <[http://www.mpipks-dresden.mpg.de/~tisean/Tisean\\_3.0.1/index.html](http://www.mpipks-dresden.mpg.de/~tisean/Tisean_3.0.1/index.html)>. Acesso em: 21 jan. 2016.
- HIRSCH, M. *Differential topology*. New York: Springer-Verlag, 1976. v. 33.
- INDICES, S. D. J. *Dow Jones Industrial Average: Index Data*. 2015. Disponível em: <<http://www.djaverages.com/?go=industrial-index-data>>. Acesso em: 29 jun. 2015.
- INSALL MATT; ROWLAND, T.; WEISSTEIN, E. W. “*Embedding*”. *From MathWorld—A Wolfram Web Resource*. 2015. Disponível em: <<http://mathworld.wolfram.com/Embedding.html>>. Acesso em: 11 abr. 2015.
- JING, Y.; CHI, Y.-J. Effects of twin-screw extrusion on soluble dietary fibre and physicochemical properties of soybean residue. *Food Chemistry*, Amsterdam, v. 138, n. 2–3, p. 884 – 889, 2013.
- KANTZ, H. A robust method to estimate the maximal lyapunov exponent of a time series. *Physics Letters A*, Amsterdam, v. 185, n. 1, p. 77 – 87, 1994.
- KANTZ, H.; SCHREIBER, T. *Nonlinear Time Series Analysis*. 2. ed. New York: Cambridge University Press, 2003.
- KELLER, J. B. Inverse problems. *The American Mathematical Monthly*, Mathematical Association of America, v. 83, n. 2, p. 107–118, 1976.
- KELLEY, J. *General Topology*. New York: Springer-Verlag, 1975. v. 27.
- LEHMANN, E. L. *Testing statistical hypotheses*. New York: Springer, 2005.

- LORENZ, E. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, Washington, v. 20, n. 2, p. 130–141, jan 1963.
- MADSEN, K.; NIELSEN, H. B.; TINGLEFF, O. *Methods for non-linear least squares problems*. 2nd. ed. Kongens Lyngby: Technical University of Denmark, 2004. E-book.
- MANDAL, M. *Continuous and discrete time signals and systems*. Cambridge, UK New York: Cambridge University Press, 2007.
- MANTEGNA, R. *An introduction to econophysics : correlations and complexity in finance*. Cambridge, UK New York: Cambridge University Press, 2000.
- MCSHARRY, P. *Nonlinear Dynamics and Chaos Workshop*. University of Oxford, 2014. Disponível em: <<http://people.maths.ox.ac.uk/mcsharry/lectures/ndc/ndcworkshop.shtml>>. Acesso em: 08 set. 2014.
- NASA. Solar Dynamics Observatory. 2014. Divulga imagens de observações do Sol. Disponível em: <<https://www.nasa.gov/content/goddard/giant-january-sunspots/#.VmqInUorKM->>. Acesso em: 11 dez. 2015.
- NIST/SEMATECH. *e-Handbook of Statistical Methods*. 2016. Este site apresenta métodos estatísticos aplicados a problemas científicos e tecnológicos. Disponível em: <<http://www.itl.nist.gov/div898/handbook/pmd/section6/pmd64.htm>>. Acesso em: 05 jul. 2016.
- OSELEDEC, V. I. *Tran Moscow Math Soc, Vol 19-1968*. USA: American Mathematical Soc., 1969.
- OTT, E. *Chaos in dynamical systems*. Cambridge, U.K. New York: Cambridge University Press, 2002.
- PACKARD, N. H. et al. Geometry from a time series. *Phys. Rev. Lett.*, American Physical Society, v. 45, p. 712–716, sep 1980.
- PARLITZ, U. Identification of true and spurious lyapunov exponents from time series. *International Journal of Bifurcation and Chaos*, [S.l.], v. 02, n. 01, p. 155–165, 1992.
- PETROVAY, K. Solar cycle prediction. *Living Reviews in Solar Physics*, [S.l.], v. 7, n. 2, 2010.
- PRILEPKO, A. I. *Methods for solving inverse problems in mathematical physics*. New York: Marcel Dekker, 2000.
- PUKELSHEIM, F. The three sigma rule. *The American Statistician*, American Statistical Association, v. 48, n. 2, p. 88–91, 1994.

- RAZALI, N. M.; WAH, Y. B. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, [S.l.], v. 2, n. 1, p. 21–33, 2011.
- ROBINSON, D. *A course in linear algebra with applications*. 2nd. ed. Singapore River Edge, NJ: World Scientific, 2006.
- ROSENSTEIN, M. T.; COLLINS, J. J.; LUCA, C. J. D. A practical method for calculating largest lyapunov exponents from small data sets. *Physica D: Nonlinear Phenomena*, Amsterdam, v. 65, n. 1, p. 117 – 134, 1993.
- ROYSTON, P. Approximating the shapiro-wilk w-test for non-normality. *Statistics and Computing*, New York, v. 2, n. 3, p. 117–119, 1992.
- RUELLE, D. *Chaotic evolution and strange attractors: The statistical analysis of time series for deterministic nonlinear systems*. Cambridge: Cambridge University Press, 1989.
- RUELLE, D.; TAKENS, F. On the nature of turbulence. *Communications in Mathematical Physics*, Springer, Verlag, v. 23, n. 4, p. 343–344, dez 1971.
- SANO, M.; SAWADA, Y. Measurement of the lyapunov spectrum from a chaotic time series. *Phys. Rev. Lett.*, American Physical Society, v. 55, p. 1082–1085, Sep 1985.
- SCHEFFÉ, H. *The analysis of variance*. New York: Wiley, 1959.
- SCHELTER, B. *Handbook of time series analysis : recent theoretical developments and applications*. Weinheim: Wiley-VCH, 2006.
- SCHUSTER, P.; JAFFE, R. Quantum mechanics on manifolds embedded in euclidean space. *Annals of Physics*, [S.l.], v. 307, n. 1, p. 132 – 143, sep 2003.
- SCOTT, D. W. On optimal and data-based histograms. *Biometrika*, London, v. 66, n. 3, p. 605–610, 1979.
- SHANNON, C. E. A mathematical theory of communication. *Bell System Technical Journal*, New York, v. 27, p. 379–423, 1948.
- SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, London, v. 52, n. 3-4, p. 591–611, 1965.
- SILVEY, S. D. *Statistical inference*. New York: Chapman and Hall Wiley, 1975.
- SQUIRES, G. L. *Practical physics*. Cambridge: Cambridge University Press, 2001.

- TAKENS, F. Detecting strange attractors in turbulence. In: RAND, D.; YOUNG, L.-S. (Ed.). *Dynamical Systems and Turbulence, Warwick 1980*. Berlin: Springer Berlin Heidelberg, 1981, (Lecture Notes in Mathematics, v. 898). p. 366–381.
- WALD, R. *General relativity*. Chicago: University of Chicago Press, 1984.
- WEISSTEIN, E. W. “*Dynamical System*”. *From MathWorld—A Wolfram Web Resource*. 2016. Disponível em: <<http://mathworld.wolfram.com/DynamicalSystem.html>>. Acesso em: 21 fev. 2016.
- WHITNEY, H. Differentiable manifolds. *Annals of Mathematics*, Princeton, v. 37, n. 3, p. 645–680, jul 1936.
- WOLF, A. et al. Determining lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena*, Amsterdam, v. 16, n. 3, p. 285 – 317, jul 1985.
- YULE, G. U. On a method of investigating periodicities in disturbed series, with special reference to wolfer’s sunspot numbers. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, The Royal Society, London, v. 226, n. 636-646, p. 267–298, 1927.

## APÊNDICE A – Classe restrita de funções nos problemas de *data fitting*

Em um ajuste de dados para  $N(m + 1)$  observáveis, cada conjunto de *observáveis independentes* —  $(X_{1r}, X_{2r}, \dots, X_{mr}) \doteq |X_r\rangle$  — está conjugado com o *observável dependente*  $Y_r$ . A *função de estado*  $\phi_i(|X\rangle)$  tem o valor  $\acute{\phi}_i(|X_r\rangle)$  para o vetor de estado  $|X_r\rangle$ , ordenado pelo índice  $r$ . O problema a ser atacado consiste em encontrar uma função de  $|X\rangle$  que seja uma aproximação satisfatória de  $Y_r$  para todos os  $N$  pares  $(Y_r, |X_r\rangle)$ .

O objeto matemático do processo de otimização é o vetor de  $n$  parâmetros  $(a_1, a_2, \dots, a_n) \doteq |a\rangle$ . No *modelo de ajuste*  $\mathcal{M}_L(|a\rangle, |X\rangle)$  — a letra  $L$  em geral dá lugar a um rótulo que identifica o modelo —, a restrição

$$\mathcal{M}_R(|a\rangle, |X\rangle) = \sum_{i=1}^n a_i \phi_i(|X\rangle) \quad (71)$$

tem o propósito de evitar o uso de técnicas iterativas para a minimização — como requerem os métodos de Levenberg-Marquardt ou de Powell's Dog Leg (MADSEN; NIELSEN; TINGLEFF, 2004. E-book). Somente funções lineares nos parâmetros de ajuste poderão integrar a *função custo*  $\mathcal{F}(|a\rangle)$ .

$$\mathcal{F}(|a\rangle) = \sum_{r=1}^N \left( Y_r - \sum_{i=1}^n a_i \acute{\phi}_i(|X_r\rangle) \right)^2 \quad (72)$$

Para aplicar o método dos mínimos quadrados na determinação do vetor de parâmetros, devem ser calculadas  $n$  derivadas parciais nulas

$$\frac{1}{2} \frac{\partial \mathcal{F}}{\partial a_j} = 0 = \acute{\phi}_j(|X_r\rangle) \sum_{r=1}^N \left( Y_r - \sum_{i=1}^n a_i \acute{\phi}_i(|X_r\rangle) \right) \quad (73)$$

na minimização da função custo  $\mathcal{F}(|a\rangle)$ . Após algumas manipulações algébricas nas  $n$  derivadas parciais — do mesmo tipo de (73) —, as equações que resultam deste procedimento são agrupadas no sistema

$$\begin{cases} \sum_{i=1}^n a_i \left( \sum_{r=1}^N \acute{\phi}_i(|X_r\rangle) \acute{\phi}_1(|X_r\rangle) \right) = \sum_{r=1}^N Y_r \acute{\phi}_1(|X_r\rangle) \\ \sum_{i=1}^n a_i \left( \sum_{r=1}^N \acute{\phi}_i(|X_r\rangle) \acute{\phi}_2(|X_r\rangle) \right) = \sum_{r=1}^N Y_r \acute{\phi}_2(|X_r\rangle) \\ \vdots \\ \sum_{i=1}^n a_i \left( \sum_{r=1}^N \acute{\phi}_i(|X_r\rangle) \acute{\phi}_n(|X_r\rangle) \right) = \sum_{r=1}^N Y_r \acute{\phi}_n(|X_r\rangle) \end{cases} \quad (74)$$

Definindo os componentes

$$A_{kl} \equiv \sum_{r=1}^N \hat{\phi}_k(|X_r\rangle) \hat{\phi}_l(|X_r\rangle)$$

$$b_k \equiv \sum_{r=1}^N Y_r \hat{\phi}_k(|X_r\rangle)$$

para a matriz  $\hat{A}$  e o vetor  $|b\rangle$ , o conteúdo inteiro do sistema linear (74) pode ser escrito na forma

$$\hat{A} |a\rangle = |b\rangle . \quad (75)$$

O vetor  $|a\rangle$  — isto é, o conjunto dos  $n$  parâmetros de interesse — é uma *solução de mínimos quadrados* do sistema linear (74) no contexto da Álgebra Linear (ROBINSON, 2006).

As incertezas das medidas realizadas durante a experiência requerem uma atitude cautelosa na aplicação da equação matricial (75). Ocorre que os diferentes conjuntos de erros — associados aos seus respectivos observáveis independentes — precisam ser levados em conta na determinação do vetor de parâmetros  $|a\rangle$ . Se as incertezas não forem iguais, então existe uma forte razão para o tratamento diferenciado entre os diversos pares  $(Y_r, |X_r\rangle)$ .

Devido a razões práticas, os  $m$  erros  $\sigma_{X_j}$  dos observáveis independentes são usualmente transferidos para as incertezas dos observáveis dependentes. Bevington e Robinson (2003) apresentam o método a ser seguido para esta transferência. A partir de um ajuste preliminar realizado com um algoritmo que resolve a Equação (75), o vetor designado como  $|\tilde{a}\rangle$  tem como componentes os  $n$  coeficientes preliminares  $\tilde{a}_i$ . Sendo o erro provisório para o observável  $Y_r$  denotado por  $\tilde{\sigma}_r$  — no modelo preliminar de ajuste  $\tilde{f}(|\tilde{a}\rangle, |X\rangle)$  —, a incerteza final de interesse  $\sigma_r$  é calculada por

$$\sigma_r = \sqrt{\tilde{\sigma}_r^2 + \sum_{j=1}^m \left( \frac{\partial \tilde{f}}{\partial X_j} \sigma_{X_j} \right)_r^2} . \quad (76)$$

A presença do índice  $r$  indica que os componentes do vetor de estado  $|X_r\rangle$  são aplicados no cálculo da quantidade entre parêntesis. O resultado líquido da transferência de incertezas pode ser traduzido dizendo que o efeito do conjunto de todas as incertezas está embutido no erro  $\sigma_r$  do observável dependente  $Y_r$ .

Num experimento pode ocorrer que as medidas sejam afetadas por incertezas diferentes. É natural atribuir pesos maiores aos observáveis que possuem erros de menor magnitude. Sendo assim, o peso de cada ponto — num ajuste de curvas, por exemplo — admite uma mensuração adequada se for calculado como o inverso da incerteza do

observável. Para operacionalizar esta concepção, a *função custo com peso*  $\mathcal{F}_w$  inclui o fator peso  $w$ , calculado como  $w_r = 1/\sigma_r$  (BEVINGTON; ROBINSON, 2003).

$$\mathcal{F}_w(|a\rangle) = \sum_{r=1}^N w_r^2 \left( Y_r - \sum_{i=1}^n a_i \phi_i(|X_r\rangle) \right)^2 \quad (77)$$

A partir das aplicações

$$Y_{wr} \rightarrow w_r^2 Y_r$$

$$a_{wi} \rightarrow w_r^2 a_i$$

na função (77) acima, o *problema de mínimos quadrados com peso* assume a mesma forma matricial de (75).

$$\hat{A}_w |a_w\rangle = |b_w\rangle \quad (78)$$

As quantidades ponderadas  $\{\hat{A}_w, |a_w\rangle, |b_w\rangle\}$  tomam o lugar de  $\{\hat{A}, |a\rangle, |b\rangle\}$ .

## APÊNDICE B – Abordagem analítica do ajuste de dados

O mapeamento dos vetores de estado  $|X_r\rangle$  em um conjunto observáveis dependentes  $Y_r$  pode ser analisado sob diversos ângulos. Além de determinar a *função de ajuste*  $f_L(|X\rangle)$ , a ideia de oferecer um espectro de opções de análise tem por finalidade flexibilizar a sua aplicação nas tarefas científicas ou tecnológicas. Um determinado resultado do ajuste de dados pode ser adequado numa aplicação, mas inaceitável em outra. A decisão de aceitar ou não um processo de *data fitting* se dá com base num conjunto de análises que sejam relevantes para o pesquisador.

### B.1 Distribuição de resíduos

Pelo Teorema de Gauss-Markov, nenhum método de minimização é superior aos mínimos quadrados se os resíduos apresentarem uma distribuição normal (SILVEY, 1975). No sentido de dispensar a procura por um outro método de otimização, uma avaliação estatística funciona como o guia para a tomada de decisão em aceitar o vetor de parâmetros  $|a\rangle$  — ou  $|a_w\rangle$ , se for o caso — calculado pelos algoritmos (75) ou (78). É importante ter em mente que, para um dado conjunto de observáveis e um específico modelo de ajuste, nada impede que a otimização pela *soma mínima*, por exemplo, possa apresentar variância menor do que a solução de mínimos quadrados.

Num teste de normalidade, a hipótese de que uma população segue uma distribuição normal pode ser aceita ou rejeitada a partir de uma amostra retirada desta população. A estratégia consiste em tratar o conjunto de erros no ajuste de dados como a amostra do teste e estender a inferência sobre a população inteira para o próprio conjunto de resíduos. O raciocínio empregado consiste em admitir que, se o teste é aplicável a uma população inteira — que neste caso é imaginária —, certamente será uma decisão estatística adequada para a própria amostra.

Entre os testes disponíveis, o teste de Shapiro-Wilk apresenta a melhor performance na comparação de Razali e Wah (2011). Além disto, está disponível para implementação imediata em sistemas de computação algébrica a partir de um comando. Neste teste, a hipótese nula tem a seguinte formulação: *A amostra é retirada de uma população que segue a distribuição normal*. Um valor mínimo para o *p-valor* — por exemplo, 0.05 — precisa ser especificado para a aceitação da hipótese nula. A estatística do teste de Shapiro-Wilk  $W$  é submetida a um processo de normalização e o resultado final  $\mathbf{w}$  precisa atingir o *p-valor* especificado para que a hipótese nula seja aceita (ROYSTON, 1992).

## B.2 Desvio, nível de confiança e acurácia

Considerando a função de ajuste  $f_L(|X\rangle)$  como um preditor, uma estimativa do erro na determinação de uma grandeza certamente tem grande interesse numa predição. Na expectativa de que a função gaussiana seja uma aproximação satisfatória para a distribuição dos resíduos, a variância amostral constitui um *estimador não tendencioso* para o erro da predição (BEVINGTON; ROBINSON, 2003). Esta medida estatística de dispersão é denotada por  $\sigma_L$  e não deve ser confundida com a incerteza total  $\sigma_r$  para o observável  $Y_r$  (76).

$$\sigma_L = \sqrt{\frac{\sum_{r=1}^N \left( Y_r - \sum_{i=1}^n a_i \phi_i(|X_r\rangle) \right)^2}{N-1}} \quad (79)$$

A classe restrita de funções (71) já foi incluída nesta fórmula.

Sendo positivo o resultado no teste de Shapiro-Wilk, existe a expectativa de que o valor verdadeiro do observável  $Y_r$  esteja no intervalo entre  $f_L(|X_r\rangle) - 3\sigma_L$  e  $f_L(|X_r\rangle) + 3\sigma_L$  com 99.7% de confiança. O *nível de confiança no data fitting*  $\mathcal{L}_L$  é definido por

$$\mathcal{L}_L = 1 - \frac{2}{\sqrt{2\pi}} \int_{\epsilon_r/\sigma_L}^{\infty} \exp\left\{-\frac{\epsilon_r^2}{2\sigma_L^2}\right\} d(\epsilon_r/\sigma_L) . \quad (80)$$

O  $r$ -ésimo resíduo recebe a notação  $\epsilon_r$ . Então, um erro na predição igual a  $\epsilon_r = 2\sigma_L$  corresponde a um nível de confiança em torno de 0.955 e assim por diante. Esta definição expressa o conteúdo da *Regra dos Três Sigmas* (PUKELSHEIM, 1994).

A capacidade de análise neste método é crucial. Uma quantidade estatística especialmente concebida para atender a esta demanda é a *acurácia no data fitting*, que recebe a notação  $\mathcal{A}_L$ . Trata-se do logaritmo de uma razão entre a média dos valores absolutos dos observáveis dependentes e o erro admissível na predição, estabelecido neste trabalho como  $3\sigma_L$ .

$$\mathcal{A}_L = \ln\left(\frac{\sum_{r=1}^N |Y_r|}{3N\sigma_L}\right) \quad (81)$$

## B.3 Matrizes de covariância e de correlação

Quando existe uma relação dinâmica entre dois observáveis  $\hat{X}_i$  and  $\hat{X}_j$ , ambos estão *correlacionados*. Uma medida que reflete como as variações em um observável têm conexões com as alterações em outro é a *covariância amostral* (BEVINGTON; ROBINSON, 2003). Se for calculada uma outra quantidade — designada *correlação entre dois*

observáveis — como

$$R_{\hat{X}_i \hat{X}_j} = \frac{N \sum_{r=1}^N \frac{\hat{X}_{ir} \hat{X}_{jr}}{\sigma_r^2} - \sum_{r=1}^N \frac{\hat{X}_{ir}}{\sigma_r} \sum_{r=1}^N \frac{\hat{X}_{jr}}{\sigma_r}}{\left[ N \sum_{r=1}^N \left( \frac{\hat{X}_{ir}}{\sigma_r} \right)^2 - \left( \sum_{r=1}^N \frac{\hat{X}_{ir}}{\sigma_r} \right)^2 \right]^{\frac{1}{2}} - \left[ N \sum_{r=1}^N \left( \frac{\hat{X}_{jr}}{\sigma_r} \right)^2 - \left( \sum_{r=1}^N \frac{\hat{X}_{jr}}{\sigma_r} \right)^2 \right]^{\frac{1}{2}}}, \quad (82)$$

então o resultado  $|R_{\hat{X}_i \hat{X}_j}| = 1$  corresponde a uma correlação perfeita, enquanto que a ausência de correlação implica que  $R_{\hat{X}_i \hat{X}_j} = 0$ . Este coeficiente tem maior conveniência que a *covariância amostral* porque o intervalo para  $R_{\hat{X}_i \hat{X}_j}$  está limitado a  $[-1, 1]$ . A existência de uma completa relação entre duas magnitudes, por exemplo, requer um módulo unitário para este coeficiente. Todavia, não existe nenhuma prova que uma correlação unitária implique a existência de uma relação profunda entre dois observáveis; a condição é necessária — porém insuficiente.

Tomando um conjunto de  $m + 1$  observáveis  $\{Y, X_1, \dots, X_m\}$ , podem ser determinados um total de  $(m + 1) \times (m + 1)$  coeficientes do tipo de (82). A *matriz de correlação*  $\hat{R}$  coleciona todos os coeficientes.

$$\hat{R} = \begin{bmatrix} R_{YY} & R_{YX_1} & \dots & R_{YX_m} \\ R_{X_1Y} & R_{X_1X_1} & \dots & R_{X_1X_m} \\ \vdots & \vdots & \ddots & \vdots \\ R_{X_mY} & R_{X_mX_1} & \dots & R_{X_mX_m} \end{bmatrix} \quad (83)$$

Outra matriz de capital interesse trata a função  $f_L(|X\rangle)$  como um observável independente. Em uma condição ideal, a imagem de uma função de ajuste é idêntica ao conjunto de observáveis dependentes. Como consequência, a correlação entre os valores preditos e os observáveis verdadeiros obrigatoriamente será **igual a um** neste caso hipotético. A *matriz de correlação do data fitting*  $\hat{R}_L$  tem o *coeficiente de correção linear do data fitting* como um elemento fora da diagonal —  $R_{Y f_L(|X\rangle)}$  ou  $R_{f_L(|X\rangle) Y}$ . A matriz  $\hat{R}_L$  é simétrica, então  $R_{Y f_L(|X\rangle)} = R_{f_L(|X\rangle) Y}$ . Ambos os coeficientes são designados por  $R_L$ .

$$\hat{R}_L = \begin{bmatrix} R_{YY} & R_{Y f_L(|X\rangle)} \\ R_{f_L(|X\rangle) Y} & R_{f_L(|X\rangle) f_L(|X\rangle)} \end{bmatrix} \quad (84)$$

Sendo ideal o *data fitting*, a matriz acima será unitária.

Grandezas físicas podem ser determinadas a partir de parâmetros ajustados. As fórmulas para a propagação de erros são ferramentas matemáticas para a estimativa das incertezas destas quantidades. As variâncias e as covariâncias dos parâmetros são ingredientes neste tipo de cálculo. Para os parâmetros  $a_i$  and  $a_j$ , a covariância é dada por

$$\sigma_{a_i a_j} = \sum_{r=1}^N \left[ \sigma_r^2 \frac{\partial a_i}{\partial Y_r} \frac{\partial a_j}{\partial Y_r} \right]. \quad (85)$$

Cada covariância é um componente da *matriz erro*  $\hat{\epsilon}$ . Pode ser mostrado que esta matriz corresponde ao inverso de  $\hat{A}_w$  (78) (BEVINGTON; ROBINSON, 2003).

$$\hat{\epsilon} = \hat{A}_w^{-1} = \begin{bmatrix} \sigma_{a_1 a_1} & \sigma_{a_1 a_2} & \cdots & \sigma_{a_1 a_m} \\ \sigma_{a_2 a_1} & \sigma_{a_2 a_2} & \cdots & \sigma_{a_2 a_m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{a_m a_1} & \sigma_{a_m a_2} & \cdots & \sigma_{a_m a_m} \end{bmatrix} \quad (86)$$

Logo, basta inverter a matriz  $\hat{A}_w$  para determinar a variância  $\sigma_{a_i a_i}$  e a covariância  $\sigma_{a_i a_j}$ .

## APÊNDICE C – Análise gráfica

Com o propósito de “trabalhar os resultados em tempo real”, uma rápida análise gráfica — após cada passo — disponibiliza um valioso *auxílio visual* no curso do experimento. Gráficos deste tipo podem resultar na descoberta de relações empíricas entre observáveis (SQUIRES, 2001).

### C.1 Gráficos de dispersão

No primeiro estágio da análise gráfica, os valores verdadeiros — representados pelo símbolo  $\circ$  — e os valores calculados do observável dependente — representados por  $+$  — são comparados no mesmo gráfico de dispersão. A ideia é construir um gráfico para cada observável independente. O eixo vertical fica reservado ao observável dependente e o horizontal ao independente. Um gráfico adicional — para uma avaliação preliminar da acurácia do ajuste — trata a ordem do dado como um observável independente. Dessa forma, para cada conjunto de  $m$  observáveis independentes estarão disponíveis, simultaneamente,  $m + 1$  gráficos de dispersão para análise.

A análise do ajuste de dados da expansão térmica do cobre ilustra a aplicação deste procedimento gráfico. Os dados — fornecidos pelo cientista Thomas Hahn — e o modelo de ajuste foram obtidos do *website* da agência NIST do *U.S. Department of Commerce* (NIST/SEMATECH, 2016). O modelo empregado tem o nome de “*Cubic/Cubic Rational Function Model*”, onde o observável dependente é o coeficiente de expansão térmica e o observável independente é a temperatura absoluta. Este modelo de função para o *data fitting* assume a forma

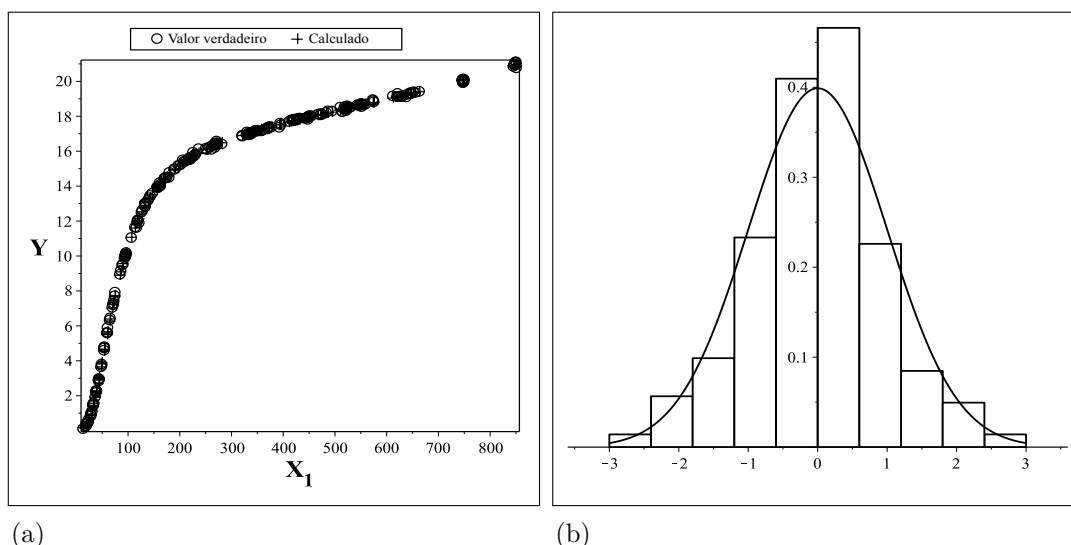
$$\mathcal{M}_{CC}(|a\rangle, |X\rangle) = \frac{a_1 + a_2 X_1 + a_3 X_1^2 + a_4 X_1^3}{1 + a_5 X_1 + a_6 X_1^2 + a_7 X_1^3}. \quad (87)$$

A classe restrita de funções (71) não está habilitada determinar os parâmetros de ajuste acima. Uma alternativa viável ao ajuste não linear será apresentada no Apêndice F. Contudo, apenas um procedimento de análise — aquele que tem base na matriz erro  $\hat{\epsilon}$  (86) — fica impossibilitado nesta aplicação. A substituição dos coeficientes disponibilizados pelo NIST/SEMATECH (2016) — no modelo (87) — especifica completamente a função de ajuste  $f_{CC}(|X\rangle)$ .

$$f_{CC}(|X\rangle) = \frac{1.07913 - 0.122801 X_1 + 0.00408837 X_1^2 - 0.00000142848 X_1^3}{1 - 0.00576111 X_1 + 0.000240629 X_1^2 - 0.000000123254 X_1^3} \quad (88)$$

Os dois tipos de gráficos de dispersão na análise do ajuste realizado no caso da expansão

Figura 25 - Expansão térmica do cobre



Legenda: (a) O algoritmo para os valores calculados (símbolo  $\circ$ ) é  $f_{CC}(|X\rangle)$  (88). (b) Os resíduos no ajuste estão compatíveis com uma distribuição gaussiana.

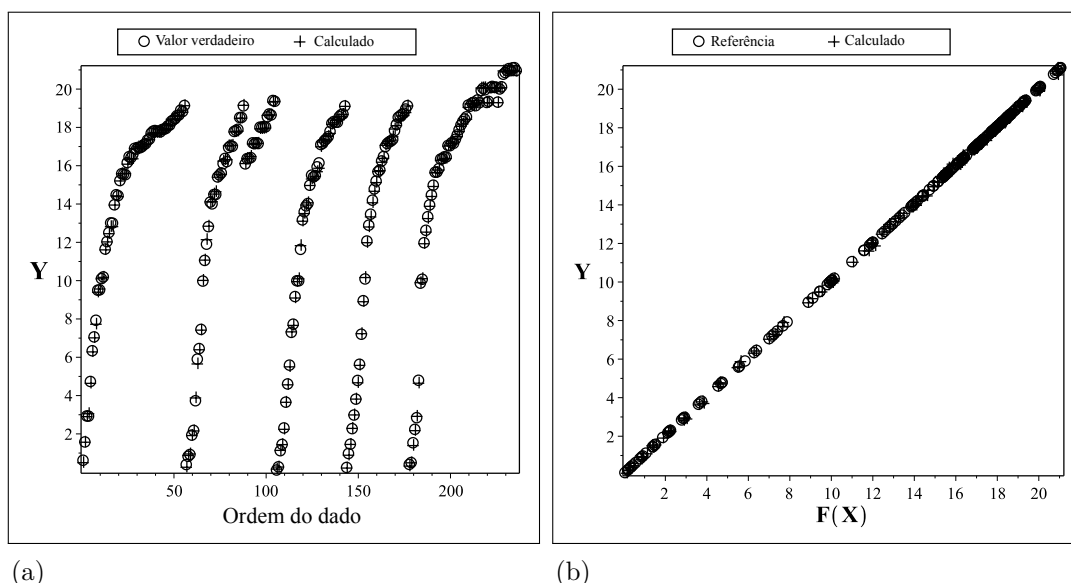
Fonte: O autor, 2017.

térmica do cobre são mostrados nas Figuras 25a e 26a. A sobreposição entre o resultado do ajuste e o valor verdadeiro do observável dependente revela a ótima qualidade do *data fitting* realizado com o modelo  $\mathcal{M}_{CC}(|a\rangle, |X\rangle)$  (87).

## C.2 Análise dos resíduos

Havendo interesse na forma da distribuição dos resíduos durante o ajuste de dados, uma visualização preliminar por meio de histogramas tem sua utilidade. Em diversos casos, este auxílio visual pode ser até suficiente. Um recurso disponibilizado neste espectro de opções gráficas consiste na impressão de um histograma conjugado com uma curva normal padronizada. Tomando o exemplo do ajuste apresentado na Seção C.1, a Figura 25b ilustra a construção de um histograma deste tipo. De fato, a distribuição gaussiana se mostra completamente adequada no exemplo da expansão térmica do cobre. O emprego da Regra de Scott resulta na *binagem* plenamente satisfatória deste histograma (SCOTT, 1979).

Figura 26 - Gráficos adicionais para a expansão térmica do cobre.



Legenda: (a) O modelo de ajuste em ambos os gráficos é  $\mathcal{M}_{CC}(|a|, |X|)$ . A superposição dos valores dos observáveis dependentes com os preditos está de acordo com o gráfico da Figura 25a. (b) A coincidência entre os símbolos  $\circ$  e  $+$  mostra que o *data fitting* realizado pelo NIST/SEMATECH (2016) apresenta absoluta conformidade com os dados experimentais.

Fonte: O autor, 2017.

### C.3 Acurácia da função de ajuste

A investigação da acurácia de uma função de ajuste como (88) também admite o mesmo procedimento da Seção C.1, no qual o efeito de cada observável independente conta com o monitoramento de um gráfico de dispersão. Com relação ao acompanhamento simples da acurácia no ajuste, o interesse principal passa a ser no resultado do conjunto de observáveis independentes e não mais em um observável específico.

Um único gráfico de dispersão disponibiliza os resultados do *data fitting*. A curva principal formada pelos símbolos  $+$  relaciona o observável dependente  $Y_r$  e o valor da função de ajuste  $f_L(|X_r|)$ . Se houver acordo entre a predição e o valor verdadeiro, os pontos estarão perfeitamente alinhados. A inclinação da reta formada num ajuste de dados ideal será igual a um. Na Figura 26b conclui-se que o *data fitting* no caso da expansão térmica do cobre apresenta alta acurácia. Para auxiliar a identificação de desvios significativos — quando estiverem presentes no ajuste de dados — uma linha reta formada pelos símbolos  $\circ$  acompanha a curva dos resultados. Num *data fitting* de alta acurácia, as cruces  $+$  deverão coincidir com os símbolos  $\circ$ .

## APÊNDICE D – O programa LinFit

O procedimento computacional desenvolvido admite três possibilidades de aplicação no ambiente Maple. A primeira se limita a encontrar o vetor de parâmetros  $|a\rangle$  — ou  $|a_w\rangle$  — na classe restrita de funções (71). Se o pesquisador pretende analisar o *data fitting*, ele precisará adicionar opções de análise no comando `LinFit`. A terceira possibilidade consiste na análise de funções com coeficientes já determinados, como foi feito com a função de ajuste (88).

### D.1 Argumentos para o ajuste de dados

Na programação das rotinas, a única informação necessária para a determinação dos parâmetros de ajuste é o conjunto dos  $N(m + 1)$  observáveis. Todos os outros argumentos são opcionais neste caso. Para a execução da rotina, o *input* mínimo se traduz numa lista de  $(m + 1)$  listas, uma para cada observável.

O argumento `Data` faculta a entrada de dados durante a realização da experiência, no sentido de facilitar o “monitoramento em tempo real” dos dados e o planejamento de uma próxima tomada de medidas. Alternativamente, o argumento opcional `DataFile` habilita a leitura dos dados diretamente de um arquivo no formato ASCII. As listas ou as colunas no arquivo devem obedecer a uma ordem específica para que os coeficientes da função de ajuste sejam determinados corretamente. A primeira lista ou coluna está destinada ao observável dependente  $Y_r$ , seguida pelos observáveis  $X_{1r}$  na segunda coluna ou lista e assim por diante. Os dados referentes à incerteza total  $\sigma_r$  (76) — num ajuste de dados com peso — são lidos na última coluna do arquivo ou então na última lista do argumento `Data`.

O modelo de ajuste  $\mathcal{M}_L(|a\rangle, |X\rangle)$  — somente a forma restrita (71) é admitida — pode ser informado diretamente pelo argumento opcional `Func` ou na forma polinomial com o argumento `Degree`. Se nenhuma destas duas opções for evocada, o default do programa automaticamente seleciona o argumento `Degree=1`.

### D.2 Opções de análise

Existem dez possibilidades para análise disponíveis no programa. Uma lista de opções seleciona as rotinas de interesse pelo argumento `Analysis`. Para a escolha das rotinas 1, 2 e 5, por exemplo, a lista `Analysis=[1,2,5]` deve ser incluída no comando `LinFit`. Tanto as quantidades discutidas na abordagem analítica quanto os gráficos disponíveis

Tabela 15 - Opções de análise disponíveis no programa LinFit

opção	descrição
1	A rotina imprime os gráficos de dispersão.
2	A rotina imprime a distribuição de resíduos.
3	A rotina imprime o gráfico para avaliação da função de ajuste.
4	A rotina aplica o teste de Shapiro-Wilk.
5	A rotina determina e imprime a matriz de correlação $\hat{R}$ (83).
6	A rotina determina e imprime a matriz erro $\hat{\epsilon}$ (86).
7	A rotina determina e imprime a matriz de correlação do <i>data fitting</i> $\hat{R}_L$ (84).
8	A rotina determina e imprime o coeficiente de correlação linear do <i>data fitting</i> $R_L$ (84).
9	A rotina determina e imprime o desvio $\sigma_L$ (79).
10	A rotina determine a acurácia $\mathcal{A}_L$ (81), aplica os critérios de qualificação do <i>data fitting</i> (ver Tabela 16) e imprime os resultados.

Legenda: Os números correspondentes às opções selecionadas para análise devem ser incluídos na lista correspondente do argumento opcional **Analysis**.

Fonte: O autor, 2017.

Tabela 16 - Critérios para qualificação do *data fitting*

critério	descrição
$QC_1$	O <i>data fitting</i> é aceitável se $\sigma_L < 0.003(3\eta_1 + 1) Y_r _{\max}$ .
$QC_2$	O <i>data fitting</i> é aceitável se $\eta_2 < 0.317N$ .
$QC_3$	O <i>data fitting</i> é aceitável se $\eta_3 < 0.045N$ .
$QC_4$	O <i>data fitting</i> é aceitável se $\eta_4 = 0$ .
$QC_5$	O <i>data fitting</i> é aceitável se $\eta_5 < 0.317N(\eta_1 + 1)$ .
$QC_6$	O <i>data fitting</i> é aceitável se $\eta_6 < 0.045N(\eta_1 + 1)$ .
$QC_7$	O <i>data fitting</i> é aceitável se $\eta_7 < 0.003N(\eta_1 + 1)$ .

Legenda: Os sete critérios acima são aplicados nos  $N$  pares  $(Y_r, |X_r|)$ .

Fonte: O autor, 2017.

são identificadas por um número que precisa ser incluído no comando.

Todas as opções de análise estão descritas na Tabela 15. *Critérios de qualificação para o data fitting* são aplicados pela última opção — o argumento `Analysis=[... , 10]` — simultaneamente com o cálculo da acurácia  $\mathcal{A}_L$  (81). Este conjunto de critérios consiste num mecanismo de controle de qualidade que quantifica a presença de *outliers*. Se os resíduos no ajuste apresentarem magnitude incompatível com o padrão estabelecido na Tabela 16, a função de ajuste deve ser rejeitada.

No plano para a qualificação do ajuste de dados, diferentes quantidades são agrupadas para investigar a natureza dos resíduos. Cada uma contempla um aspecto que não está no escopo de detecção das demais. A reunião destas quantidades forma a base para avaliar se a presença de *outliers* é significativa ou não. Nestes cálculos, os sete parâmetros  $\{\eta_1, \eta_2, \dots, \eta_7\}$  (ver Tabela 17) são as ferramentas para a detecção de *outliers*. Em face da necessidade de atender aos diferentes objetivos do *data fitting* ou a conjunto de observáveis distintos, o primeiro parâmetro  $\eta_1 \geq 0$  — cuja selecção se dá pelo argumento `Level` — é o responsável pela flexibilização do controle de qualidade. Na condição de maior rigor no ajuste de dados, o *nível de detecção de outliers*  $\eta_1 = 0$  implica que o maior desvio aceitável  $\sigma_L$  tem magnitude menor que  $0.003 \times |Y_r|_{\max}$  — e os resíduos no ajuste atendem a Regra dos Três Sigmas (PUKELSHEIM, 1994).

### D.3 Outros argumentos do comando `LinFit`

Para o correto funcionamento de todas as rotinas, o número de observáveis dependentes deve ser informado pelo argumento `Dim` se o vetor de estado  $|X\rangle$  tem dimensão maior que um.

Barras de erro são incluídas com os argumentos `ErrorBar=1` ou `ErrorBar=2`. Na primeira escolha, a rotina lê a incerteza total  $\sigma_r$  (76) contida na última lista dos dados de entrada — informadas pelos argumentos `Data` ou `DataFile`. Mas, se o pesquisador optar por realizar uma estimativa da incerteza de cada coeficiente ajustado  $a_i$  sem que os erros atribuídos a cada observável estejam disponíveis, a opção `ErrorBar=2` informa que a incerteza para o erro da predição  $\sigma_L$  (79) deve ser empregada no lugar do erro total  $\sigma_r$ . Esta substituição é uma alternativa razoável quando o tratamento rigoroso das incertezas não tem como ser realizado.

Embora o programa `LinFit` não esteja aparelhado para determinar o vetor  $|a\rangle$  se o modelo proposto para o ajuste for não linear, a análise do *data fitting* a partir de uma função do tipo de  $f_{CC}$  (88) pode ser realizada. O argumento `FitFunc=1` habilita este recurso. Como as equações (78) e (75) não comportam ajustes não lineares, a matriz erro  $\hat{\epsilon}$  (86) não pode ser determinada neste caso.

A inclusão do argumento `OutPut=1` no comando `LinFit` determina que o *output*

Tabela 17 - Parâmetros para a detecção de *outliers*

parâmetro	descrição
$\eta_1$	Este parâmetro especifica o nível de detecção de <i>outliers</i> .
$\eta_2$	Este parâmetro calcula o número de resíduos absolutos $ \epsilon_r $ maiores que $\sum_{r=1}^N  Y_r  / 3N$ .
$\eta_3$	Este parâmetro calcula o número de resíduos absolutos $ \epsilon_r $ maiores que $\sum_{r=1}^N  Y_r  / N$ .
$\eta_4$	Este parâmetro calcula o número de resíduos absolutos $ \epsilon_r $ maiores que $ Y_r _{\max}$ .
$\eta_5$	Este parâmetro calcula o número de resíduos absolutos $ \epsilon_r $ maiores que $\sigma_L$ .
$\eta_6$	Este parâmetro calcula o número de resíduos absolutos $ \epsilon_r $ maiores que $2\sigma_L$ .
$\eta_7$	Este parâmetro calcula o número de resíduos absolutos $ \epsilon_r $ maiores que $3\sigma_L$ .

Legenda: Os parâmetros acima são calculados levando em conta os  $N$  pares  $(Y_r, |X_r)$ .

Fonte: O autor, 2017.

do programa será uma lista com a função de ajuste e todas as quantidades descritas no Apêndice B. Pelo *default*, o programa disponibiliza apenas o resultado do *data fitting*.

• `[> LinFit(Argumentos);`

• Argumentos:

–Data = `<list[integer]>` - (opcional)

–DataFile = `<string>` - (opcional)

–Dim = `<integer>` - (opcional) Este argumento especifica a dimensão do vetor de estado  $|X\rangle$ . O *default* é 1.

–Func = `<expression>` - (opcional) Este argumento especifica o modelo de ajuste  $\mathcal{M}_L(|a\rangle, |X\rangle)$ .

–Degree = `<integer>` - (opcional) Este argumento especifica o grau do polinômio selecionado para o modelo de ajuste. O *default* é 1.

–Analysis = `<list[integer]>` - (opcional) Esta lista contém as opções selecionadas para análises.

–ErrorBar = `<integer>` - (opcional) Este argumento é necessário para que seja realizado o *data fitting* com peso (78).

- FitFunc = 1 - (opcional) Este argumento habilita as opções de análise sem que seja realizado o ajuste de dados.
- OutPut = 1 - (opcional) Este argumento inclui as quantidades estatísticas que devem ser incluídas no *output*.
- Level = <numeric> - (opcional) Este argumento especifica o valor do parâmetro  $\eta_1$ .
- SigDig = <integer> - (opcional) Este argumento especifica o número de algarismos significativos do *output*.

## APÊNDICE E – Aplicações do programa LinFit

Parte das quantidades estatísticas calculadas pelo programa `LinFit` apresenta numerosas aplicações na literatura, com valores de referência suficientes para testar as rotinas desenvolvidas. Os conjuntos de observáveis possuem suas fontes em áreas distintas do conhecimento. Por outro lado, as funcionalidades específicas do programa aperfeiçoam a qualidade e o poder de análise do *data fitting*.

### E.1 Parâmetros de ajuste

Além de fornecer o material para testar o desempenho da rotina, as aplicações a seguir ilustram o uso do comando `LinFit`. O primeiro conjunto de observáveis pertence ao campo da Química dos Alimentos. Na tecnologia de extrusão de alimentos, o resíduo de soja — observável dependente  $Y_r$ , em % — tem uma relação com a temperatura de extrusão — observável independente  $X_{1r}$ , em graus *Celsius* —, a umidade do alimento — observável independente  $X_{2r}$ , em % — e a velocidade de rotação — observável independente  $X_{3r}$ , em *rpm*. O modelo de ajuste — proposto por Jing e Chi (2013) e aplicado por Devore (2016) — para o *data fitting* tem a forma funcional

$$\begin{aligned} \mathcal{M}_{soy}(|a\rangle, |X\rangle) = & a_1 + a_2 X_1 + a_3 X_2 + a_4 X_3 + a_5 X_1 X_2 + a_6 X_1 X_3 \\ & + a_7 X_2 X_3 + a_8 X_1^2 + a_9 X_2^2 + a_{10} X_3^2. \end{aligned} \quad (89)$$

Os valores de referência para o ajuste de dados pelo método dos mínimos quadrados de acordo com Devore (2016) são:  $a_1 = -131.61$ ,  $a_2 = 1.6875$ ,  $a_3 = 0.77688$ ,  $a_4 = 0.79788$ ,  $a_5 = -0.0008750$ ,  $a_6 = 0.0006000$ ,  $a_7 = -0.0006188$ ,  $a_8 = -0.027000$ ,  $a_9 = -0.0027563$  e  $a_{10} = -0.0020375$ . O comando no ambiente Maple para este *data fitting* e a função de ajuste  $f_{soy}(|X\rangle)$  está reproduzido abaixo.

```
[> f[soy] := LinFit (DataFile='soy.txt', Degree=2, Dim=3, SigDig=6);
```

$$\begin{aligned} f_{soy} := & -0.0270000 X_1^2 - 0.000874995 X_1 X_2 + 0.000600005 X_1 X_3 \\ & -0.00275625 X_2^2 - 0.000618745 X_2 X_3 - 0.00203749 X_3^2 \\ & + 1.68750 X_1 + 0.776874 X_2 + 0.797872 X_3 - 131.614 \end{aligned}$$

Devido a presença do argumento `SigDig=6`, os coeficientes foram arredondados para seis algarismos significativos e coincidem com os valores de referência.

O ajuste de dados por um polinômio em espaços de dimensões mais altas é realizado de maneira muito prática. Se, por exemplo, houver a necessidade de determinar os

coeficientes de um polinômio de terceiro grau para com nove observáveis independentes — isto é, 220 parâmetros para serem determinados — basta incluir os argumentos `Degree=3` e `Dim=9`.

## E.2 Matriz erro

Em uma situação experimental típica, a voltagem de um termopar depende da temperatura da junção. Usando dados da excelente publicação de Bevington e Robinson (2003), as medidas e os erros foram armazenados no arquivo de três colunas ‘`therm.txt`’. Na primeira, o observável dependente é a voltagem medida em *mV*; a temperatura em graus *Celsius* é o observável independente na segunda. As barras de erro, em *mV*, são alocadas na última coluna.

Com o modelo de ajuste

$$\mathcal{M}_{deg=2}(|a\rangle, |X\rangle) = a_1 + a_2 X_1 + a_3 X_1^2, \quad (90)$$

os parâmetros de referência e sua matriz erro disponibilizados por Bevington e Robinson (2003) são

$$|a\rangle_{ref} = \begin{bmatrix} -0.918 \\ 0.0377 \\ 0.000055 \end{bmatrix} \quad e \quad (91)$$

$$\hat{\epsilon}_{ref} = \begin{bmatrix} 8.907 \times 10^{-5} & -3.473 \times 10^{-5} & 2.823 \times 10^{-7} \\ -3.473 \times 10^{-5} & 1.913 \times 10^{-6} & -1.783 \times 10^{-8} \\ 2.823 \times 10^{-7} & -1.783 \times 10^{-8} & 1.783 \times 10^{-10} \end{bmatrix}. \quad (92)$$

Com o comando referente ao *data fitting* — que inclui a opção de análise corresponde à matriz erro, ou seja, `Analysis=[6]` —,

```
[> h:= a[1]+a[2]*X[1]+a[3]*(X[1])^2:
```

```
[> f[therm]:=LinFit(DataFile='therm.txt',Func=h,ErrorBar=1,Analysis=[6]);
```

Tabela 18 - Comparação de grandezas estatísticas e detecção de *outliers*

grau	quantidade estatística			critério $QC$						
	$R_{therm}$	$\sigma_{therm}$	$\mathcal{A}_{therm}$	1	2	3	4	5	6	7
1	0.99849	0.0737	1.852		×	×	×	×	×	×
2	0.99908	0.0576	2.098		×	×	×	×	×	×
3	0.99913	0.0558	2.130		×	×	×	×	×	×
4	0.99919	0.0540	2.164		×	×	×	×	×	×
5	0.99933	0.0489	2.261		×	×	×	×	×	×
6	0.99959	0.0382	2.509	×	×	×	×	×	×	×

Legenda: Ajuste de dados para a expansão térmica do cobre com polinômios de graus distintos.

Fonte: O autor, 2017.

o *output*

$$\begin{bmatrix} 0.0008907396951 & -0.00003472614342 & 0.0000002823263693 \\ -0.00003472614342 & 0.000001912984041 & -0.00000001783113911 \\ 0.0000002823263693 & -0.00000001783113911 & 0.0000000001783113911 \end{bmatrix}$$

$$f_{therm} := 0.00005490088859 X_1^2 + 0.03765432673 X_1 - 0.9181038961$$

tem total equivalência com os valores da literatura reproduzidos em (91) e (92).

### E.3 Funcionalidades específicas

Modelos alternativos para o ajuste de dados podem ter como resultado coeficientes mais acurados. A avaliação da performance de cada função de ajuste tem a base das quantidades estatísticas descritas no Apêndice B. O exemplo das medidas de voltagem do termopar em função da temperatura — explorado na seção anterior — admite bons ajustes a partir de modelos polinomiais distintos.

Variando o grau do polinômio de um a seis, as rotinas de análise do programa `LinFit` calcularam as quantidades apresentadas na Tabela 18. No procedimento de detecção de *outliers*, o parâmetro  $\eta_1 = 1$  — argumento `Level=1`, pelo *default* — foi utilizado em todas as análises. O polinômio do quinto grau foi obtido com o comando reproduzido a seguir.

```
[> LinFit(DataFile='therm.txt',Degree=5,ErrorBar=1,Analysis=[8,9,10]):
```

Pelos valores apresentados na Tabela 18, fica claro que o aumento no grau do polinômio tem como consequência a melhoria da qualidade do ajuste. Em outras palavras, as quantidades estatísticas  $\sigma_{therm}$  apresentam valores mais baixos, enquanto que o coeficiente  $R_{therm}$  e a acurácia  $\mathcal{A}_{therm}$  crescem para os polinômios de ordem mais alta. Os *outliers* são irrelevantes ao nível de  $\eta_1 = 1$  para o polinômio de sexto grau.

## APÊNDICE F – Computação Algébrica no *data fitting*

Problemas difíceis de ajuste de dados requerem uma função adequada do vetor de estado  $|X\rangle$ . Um aspecto favorável no programa `LinFit` está relacionado à praticidade de manipulação disponibilizada com o emprego da Computação Algébrica. A modelagem da função proposta para o ajuste de dados facilita a procura de uma forma funcional — na forma restrita (71) — que atenda aos critérios de qualificação listados na Tabela 16.

### F.1 Comparação entre diferentes modelos lineares

O *data fitting* da expansão térmica do cobre já foi utilizado como um estudo de caso no Apêndice C. No presente contexto, diferentes formas funcionais são empregadas para o mesmo conjunto de dados — totalizando 236 pares de observáveis  $(Y_r, |X_r\rangle)$ . Os modelos a serem trabalhados possuem ou sete ou quinze parâmetros a serem ajustados. Isto significa que o número de graus de liberdade pode ser igual a  $236 - 7 = 229$  ou então a  $236 - 15 = 221$ .

$$\mathcal{M}_{deg=6}(|a\rangle, |X\rangle) = a_1 + a_2 X_1 + \cdots + a_7 X_1^6 \quad (93)$$

$$\begin{aligned} \mathcal{M}_{pot=7}(|a\rangle, |X\rangle) = & a_1 X_1^{0.1} + a_2 X_1^{0.7} + a_3 X_1^{-0.3} + a_4 X_1^{1.1} \\ & + a_5 X_1^{-0.7} + a_6 X_1^{1.5} + a_7 X_1^{-1.1} \end{aligned} \quad (94)$$

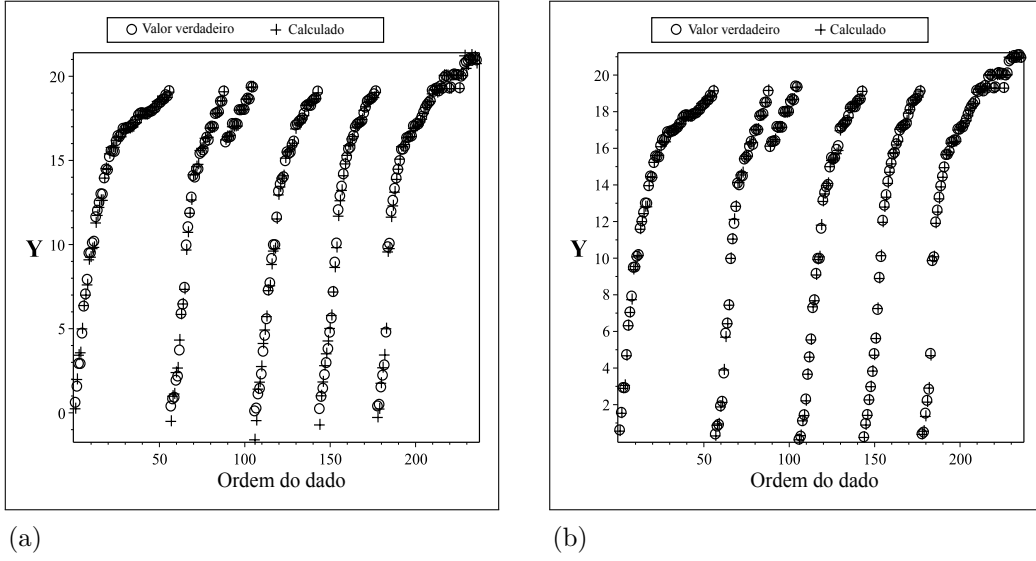
Tabela 19 - Data fitting com diferentes modelos de ajuste

modelo	quantidade estatística			critério $\mathcal{QC}$						
	$R_L$	$\sigma_L$	$\mathcal{A}_L$	1	2	3	4	5	6	7
$\mathcal{M}_{CC}$ (87)	0.99990	0.0808	4.072	×	×	×	×	×	×	×
$\mathcal{M}_{deg=6}$ (93)	0.99882	0.2791	2.832		×	×	×	×	×	
$\mathcal{M}_{pot=7}$ (94)	0.99984	0.1040	3.819	×	×	×	×	×	×	×
$\mathcal{M}_{deg=14}$ (95)	0.99904	0.2533	2.929		×	×	×	×	×	
$\mathcal{M}_{pot=15}$ (96)	0.99991	0.0780	4.106	×	×	×	×	×	×	×

Legenda: O mesmo conjunto de dados da expansão térmica do cobre é empregado em todas as aplicações acima.

Fonte: O autor, 2017.

Figura 27 - Comparação entre dois diferentes modelos de ajuste quanto ao seu desempenho



Legenda: (a) Predição com o modelo  $\mathcal{M}_{deg=14}$ . (b) Predição com o modelo  $\mathcal{M}_{pot=15}$ .

Fonte: O autor, 2017.

$$\mathcal{M}_{deg=14}(|a\rangle, |X\rangle) = a_1 + a_2 X_1 + \cdots + a_{15} X_1^{14} \quad (95)$$

$$\begin{aligned} \mathcal{M}_{pot=15}(|a\rangle, |X\rangle) &= a_1 X_1^{0.1} + a_2 X_1^{0.7} + a_3 X_1^{-0.3} + a_4 X_1^{1.1} \\ &+ a_5 X_1^{-0.7} + a_6 X_1^{1.5} + a_7 X_1^{-1.1} + a_8 X_1^{1.9} \\ &+ a_9 X_1^{-1.5} + a_{10} X_1^{2.3} + a_{11} X_1^{-1.9} \\ &+ a_{12} X_1^{2.7} + a_{13} X_1^{-2.3} + a_{14} X_1^{3.1} + a_{15} X_1^{-2.7} \end{aligned} \quad (96)$$

Em conjunto com o preditor  $f_{CC}$  (88), os quatro modelos acima constituem o objeto de estudo deste Apêndice. Mais uma vez, as quantidades estatísticas  $R_L$ ,  $\sigma_L$  and  $\mathcal{A}_L$  formam a base para a comparação dos resultados. Os gráficos da Figura 27 também ilustram a diferença de desempenho entre dois modelos de ajuste. A Tabela 19 apresenta as magnitudes para as análises e os critérios de qualificação satisfeitos em cada ajuste de dados.

Na comparação desta seção, os resultados obtidos a partir de diferentes formas funcionais — com o mesmo número de parâmetros a serem ajustados — são confrontados. Não há influência do número de graus de liberdade nesta fase. Os resultados das quantidades estatísticas na Tabela 19 mostram que os modelos  $\mathcal{M}_{pot=7}$  e  $\mathcal{M}_{pot=15}$  são superiores às formas polinomiais  $\mathcal{M}_{deg=6}$  e  $\mathcal{M}_{deg=14}$ , respectivamente. Com respeito à qualidade do ajuste, somente os preditores baseados em potências racionais — ou seja,

aqueles que possuem termos do tipo de  $X_1^{0.1}$ ,  $X_1^{0.7}$  e assim por diante — satisfazem a todos os critérios de qualificação. Visualmente, nos gráficos das Figuras 27b e 27a, o número de “dardos no alvo” a partir do modelo  $\mathcal{M}_{pot=15}$  é maior do que o obtido com a proposta de ajuste polinomial  $\mathcal{M}_{deg=14}$ .

## F.2 Alternativa aos modelos não lineares

No foco principal desta discussão está a alternativa aos métodos não lineares que são usualmente necessários para que o ajuste de dados seja satisfatório. O melhor *data fitting* com o programa `LinFit` foi obtido a partir do modelo  $\mathcal{M}_{pot=15}$ . A Tabela 19 apresenta as quantidades estatísticas que permitem a comparação com o desempenho da função  $f_{CC}(|X\rangle)$  (88). Como os coeficientes de correlação linear  $R_L$  de ambos são próximos — 0.99991 contra 0.99990 —, o gráfico da Figura 27b não apresenta uma diferença perceptível em relação ao da Figura 26a.

O ajuste com o preditor linear baseado em potências racionais apresenta valores para o desvio e a acurácia —  $\sigma_{pot=15} = 0.0780$  e  $\mathcal{A}_{pot=15} = 4.106$  — ligeiramente mais favoráveis que a regressão não linear —  $\sigma_{CC} = 0.0808$  and  $\mathcal{A}_{CC} = 4.072$ . Contudo, foi necessário aumentar o número de parâmetros de ajuste para o ganho de acurácia neste exemplo. Com relação a este ponto, o número de graus de liberdade empregado no *data fitting* merece algumas considerações.

A medida de dispersão estatística  $\sigma_L$  (79) não leva em conta a dimensão do vetor  $|a\rangle$ . Trata-se de uma importante estimativa do erro em uma predição e que também desempenha um papel relevante na abordagem analítica e na análise gráfica. Na construção de uma medida de dispersão que considere o número de graus de liberdade, o número de parâmetros  $l$  pode ser colocado no lugar de  $-1$  na fórmula (79). Então, a quantidade assim definida — designada *desvio reduzido*  $\sigma_L^\nu$  — depende apenas dos dados e do número de graus de liberdade. De fato, Bevington e Robinson (2003) consideram o quadrado desta quantidade como a melhor estimativa da variância amostral.

$$\sigma_L^\nu = \sqrt{\frac{\sum_{r=1}^N \left( Y_r - \sum_{i=1}^n a_i \phi_i(|X_r\rangle) \right)^2}{N-l}} = \sqrt{\frac{N-1}{N-l}} \sigma_L \quad (97)$$

A partir da fórmula anterior, a *acurácia reduzida* é definida como

$$\mathcal{A}_L^\nu = \ln \left( \sum_{r=1}^N \frac{|Y_r|}{3N\sigma_L^r} \right) = \mathcal{A}_L + \ln \left( \sqrt{\frac{N-l}{N-1}} \right). \quad (98)$$

As medidas estatísticas  $\sigma_L^\nu$  e  $\mathcal{A}_L^\nu$  são trivialmente determinadas com o *output* do programa

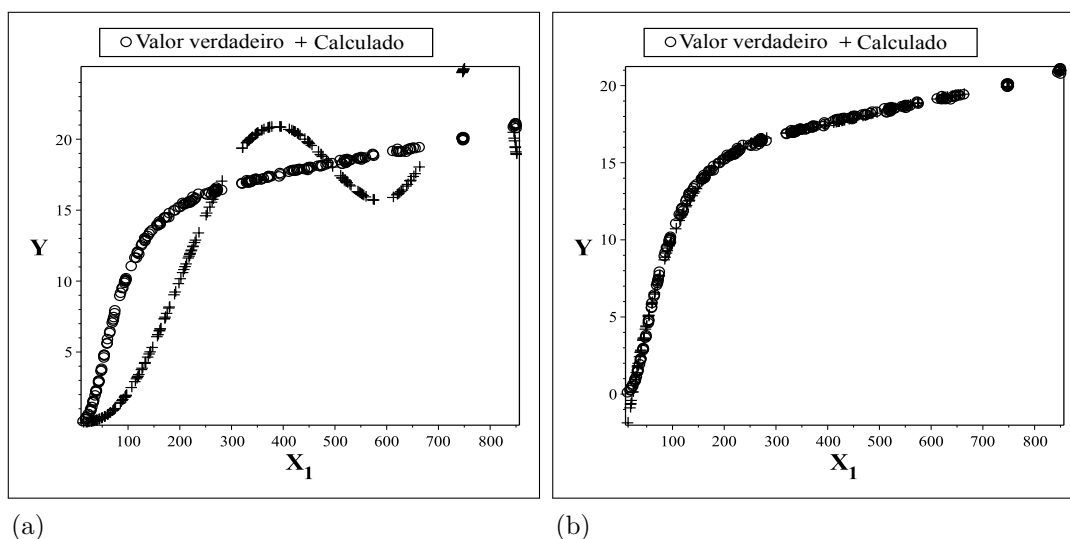
Tabela 20 - Comparação entre o desempenho dos comandos do pacote *Statistics* e do programa *LinFit*

Modelo	Comando	<i>Runtime</i> (s)	Quantidade estatística		
			$R_L$	$\sigma_L$	$\mathcal{A}_L$
$\mathcal{M}_{deg=6}$	<i>PolynomialFit</i>	0.05	0.89191	4.3704	0.081
	<i>NonlinearFit</i>	0.25	0.99140	0.7963	1.784
	<i>LinFit</i>	0.09	0.99882	0.2791	2.832
$\mathcal{M}_{pot=7}$	<i>LinearFit</i>	0.25	0.99984	0.1037	3.822
	<i>NonlinearFit</i>	0.17	0.99984	0.1037	3.822
	<i>LinFit</i>	0.87	0.99984	0.1040	3.819
$\mathcal{M}_{deg=14}$	<i>PolynomialFit</i>	0.06	0.48365	12.337	-0.957
	<i>NonlinearFit</i>	1.78	0.11075	$\approx 10^{17}$	-39.17
	<i>LinFit</i>	0.22	0.99904	0.2533	2.929
$\mathcal{M}_{pot=15}$	<i>LinearFit</i>	0.41	0.99990	0.0835	4.038
	<i>NonlinearFit</i>	1.08	0.99991	0.0782	4.104
	<i>LinFit</i>	1.87	0.99991	0.0780	4.106

Legenda: As medidas de tempo foram obtidas com um processador Pentium (R) Dual-Core 2.60 GHz.

Fonte: O autor, 2017.

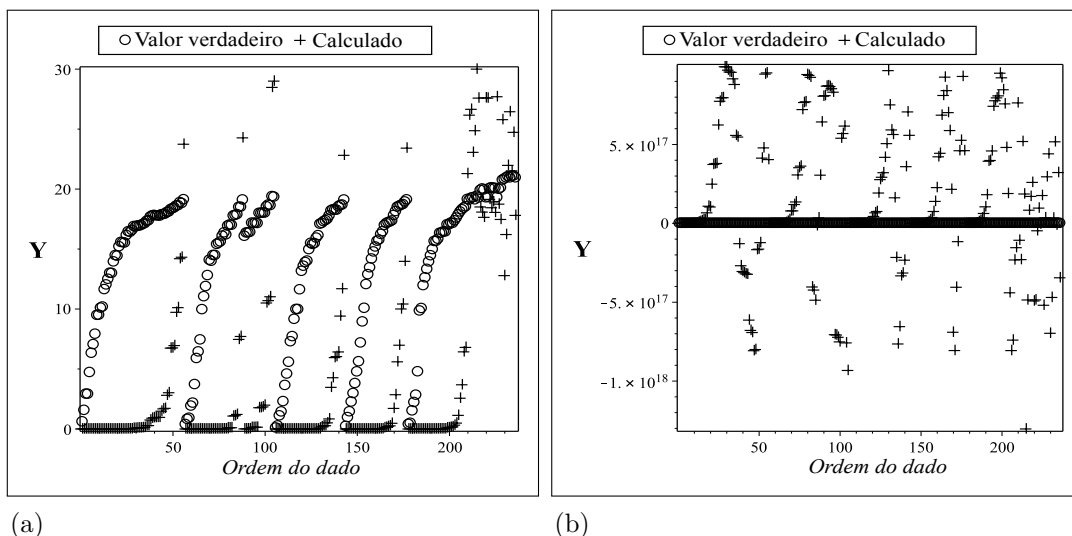
Figura 28 - Comparação entre o comando *PolynomialFit* e programa *LinFit*



Legenda: (a) Ajuste com o comando *PolynomialFit*. A função de ajuste obtida — a partir da temperatura — não é aceitável como um preditor para o coeficiente de expansão térmica do cobre. (b) *Data fitting* com o programa *LinFit*. Praticamente toda a faixa de temperaturas pode ser empregada no cálculo do coeficiente de expansão térmica com este ajuste polinomial.

Fonte: O autor, 2017.

Figura 29 - Desempenho dos comandos PolynomialFit e NonlinearFit



Legenda: (a) Resultado do comando `PolynomialFit`. (b) Comando `NonlinearFit`. As duas rotinas produzem funções de ajuste inverossímeis. Um confronto direto do desempenho entre estes comandos e o programa `LinFit` pode ser feito a partir das análises gráficas 29a, 29b, 27a.

Fonte: O autor, 2017.

`LinFit` — a partir da evocação dos argumentos opcionais `Analysis=[9,10]` e `OutPut=1` — pelas fórmulas acima.

Calculando a acurácia reduzida e o desvio reduzido com os valores das quantidades estatísticas  $\sigma_L$  and  $\mathcal{A}_L$ , os resultados

$$\sigma_{pot=15}^\nu = 0.0804, \quad \sigma_{CC}^\nu = 0.0819, \quad \mathcal{A}_{pot=15}^\nu = 4.076 \quad \text{e} \quad \mathcal{A}_{CC}^\nu = 4.058$$

mostram que o modelo linear  $\mathcal{M}_{pot=15}$  continua sendo superior ao não linear  $\mathcal{M}_{CC}$ .

### F.3 Programa `LinFit` e comandos do pacote `Statistics`

As rotinas de análises disponíveis no programa `LinFit` podem ser conjugadas com os comandos do pacote `Statistics` — da distribuição oficial do *software* Maple. Então, um *output*  $f(x)$  da rotina `NonlinearFit` pode integrar o comando

```
[>F:=subs(x=X[1],NonlinearFit(f,X,Y,x)):
[>LinFit(Data=[Y,X],Func=F,Analysis=[1,3,8,9,10],FitFunc=1);
```

para análise. A substituição  $x \rightarrow X[1]$  acima é necessária porque as variáveis globais do programa `LinFit` possuem as formas  $a[i]$  e  $X[i]$ .

Os modelos de ajuste (93), (94), (95), (96) — para o exemplo da expansão térmica do cobre — são empregados numa comparação entre o desempenho dos comandos do pacote `Statistics` e o programa `LinFit`. Nos resultados apresentados na Tabela 20, os tempos de execução no computador possuem uma dependência da forma funcional empregada e as rotinas podem ser consideradas competitivas.

Do ponto de vista da qualidade do ajuste, o resultado prático dos comandos com as funções construídas com potências racionais — isto é, (94) e (96) — é essencialmente o mesmo. Não existe, neste caso, prevalência efetiva de uma rotina em relação a outra.

Por outro lado, nos modelos polinomiais (93) e (95), os ajustes realizados com os comandos do pacote `Statistics` são inaceitáveis. No *data fitting* com polinômios de grau seis, os comandos `PolynomialFit` e `NonlinearFit` apresentam valores mais desfavoráveis para as quantidades estatísticas  $R_L$ ,  $\sigma_L$ ,  $\mathcal{A}_L$  do que o programa `LinFit` (ver Tabela 20). Na comparação direta dos gráficos 28a e 28b, a rotina do pacote `Statistics` não permite uma predição aceitável para a expansão térmica do cobre a partir de uma determinada temperatura. Já o programa `LinFit` resulta uma boa aproximação entre os valores verdadeiros e calculados em quase toda a faixa de temperaturas do ajuste.

No caso do polinômio de mais alta ordem — ou seja, o modelo de ajuste  $\mathcal{M}_{deg=14}$  (95) — os comandos `PolynomialFit` e `NonlinearFit` chegam a apresentar valores negativos para a acurácia  $\mathcal{A}_L$  (respectivamente:  $-0.957$  e  $-39.17$ , de acordo com a Tabela 20). Este resultado desastroso se manifesta também graficamente. Um confronto direto — na mesma modalidade de análise — entre o programa `LinFit` e os comandos concorrentes pode ser feito através das Figuras 29a, 29b, 27a.

## ÍNDICE DE ASSUNTOS

### Computação Algébrica

- análise de séries temporais, 12
- caracterização da série temporal, 26
- manipulação simbólica, 12
- reconstrução no ambiente Maple, 13

### Definição

- Aproximação Global, 27
- Aproximação Global Polinomial, 27
- atrator, 23
- atrator caótico, 25
- atrator estranho, 24
- bacia de atração, 23
- desvio num tempo de predição, 28
- difeomorfismo, 18
- erro tolerável na predição, 37
- espaço de fase, 17
- espaço euclidiano, 17
- expoente de Lyapunov, 25
- forma funcional dos preditores, 28
- função custo, 29
- homeomorfismo, 18
- interseção finita, 18
- mergulho, 19
- nível de confiança para um tempo de predição, 36
- nível de significância  $\alpha$ , 34
- problema inverso das séries temporais, 26
- quantificado de caos  $Z_{dyn}$ , 46
- resíduos, 29
- variedade, 18
- vetor reconstruído, 17

### Equações

- determinação do preditor, 31
- Lorenz, 26
- Navier-Stokes, 23
- sistemas dinâmicos, 16

### Estatística

- acurácia no data fitting, 102
- acurácia para um tempo de predição, 40
- acurácia reduzida, 121
- amostra, 32
- desvio reduzido, 119

- desvio relativo para um tempo de predição, 40
- distribuição gaussiana, 29
- entropia de Shannon, 73
- Erro Tipo I, 33
- Erro Tipo II, 33
- Hipótese Alternativa, 33
- Hipótese Nula, 32
- nível de significância  $\alpha$ , 36
- p-valor, 32, 79
- população, 32
- Regra de Scott, 48
- regra de Scott, 106
- Regra dos Três Sigmas, 37, 79, 102, 110
- teste bilateral, 33
- teste de hipóteses, 32
- teste de Shapiro-Wilk, 35, 65, 79
- teste unilateral, 33
- testes de Kolmogorov-Smirnov, Lilliefors e Anderson-Darling, 35
- variável normal padronizada, 33, 35
- variância amostral, 28

### Fenômenos complexos

- índice do mercado de ações, 11
- atrator estranho de Lorenz, 23
- autossimilar, 23
- circuito caótico, 15
- fluxo circular de Couette, 21
- fractais, 23
- manchas solares, 10
- série randômica, 15

### Método

- Chebyshev, 32
- critérios de qualificação para o *data fitting*, 110
- delay time ou lag, 17, 21, 22, 49, 53, 64, 72–74, 90
- Diagrama Acurácia-Desvio, 42
- Diagrama Acurácia-Desvio Logarítmico, 42
- falsos vizinhos próximos, 72
- ganho de informação, 54, 56
- global fitting, 57
- informação mútua média, 72

Levenberg-Marquardt, 28, 98  
 mínimos quadrados, 13, 28, 30, 32, 63,  
 90, 98–101, 113  
 máxima verossimilhança, 14, 29  
 médias, 32  
 Powell's Dog Leg, 28, 98  
 problema direto, 10, 12  
 qualificação dos preditores, 41, 42, 63,  
 81, 82  
 Reconstrução do Espaço de Fase, 11  
 solução de mínimos quadrados, 31  
 soma mínima, 32  
 tentativa e erro, 22, 28

### Rotinas computacionais

AnalysTS, 51, 57  
 ConfiTS, 48, 49, 64–66, 69, 75, 79, 81,  
 90, 91  
 ForecastTS, 13, 49, 51, 57, 64, 75  
 GfiTS, 49, 51, 57, 64  
 GrafiTS, 51, 57  
 IforecastTS, 51, 57, 64  
 LinFit, 63, 108, 113, 117, 122  
 LinGfiTS, 13, 47, 49, 64, 68, 69, 77, 79,  
 81  
 NiforecastTS, 51, 57, 64  
 TS, 51  
 VecTS, 13, 49, 51, 57, 64, 73  
 gerpoly, 57  
 lorenz, 79  
 comando NonlinearFit, 122  
 comando PolynomialFit, 122  
 pacote LinMapTS, 47, 63, 64, 81, 90  
 pacote Statistics, 122  
 pacote TimeS, 47, 49, 51, 57, 59, 68  
 pacote Tisean, 79  
 programa DynCharTS, 47, 60, 61, 82, 91

### Séries temporais

Dow Jones, 13, 86  
 LorenzX, 79  
 Mega-Sena, 15, 85  
 Número de Manchas Solares, 86  
 série temporal periódica, 84  
 sinal elétrico caótico, 15

### Teorema

de Gauss-Markov, 32  
 de Takens, 20

de Whitney, 19  
 do espaço compacto, 18  
 ergódico multiplicativo de Oseledec, 25

### Teoria

da elasticidade, 10  
 da Informação, 73  
 das Equações Diferenciais, 12  
 de Lie, 52  
 de primeiros princípios, 16  
 do Caos, 14, 24  
 dos potenciais, 10  
 dos problemas inversos, 10  
 dos Sistemas Dinâmicos, 12  
 ergódica, 19  
 teorias fenomenológicas, 16