
# CLASSIFICAÇÃO DE SÉRIES TEMPORAIS VIA CLASSIFICADOR DE BAYES EMPREGANDO MODELOS LINEARES DINÂMICOS

Diana Dorgam de Aguiar dos Santos

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática, da Universidade Federal do Amazonas, como parte dos requisitos necessários à obtenção do título de Mestre em Matemática

Orientador: José Raimundo Gomes Pereira

Manaus
Julho de 2016

CLASSIFICAÇÃO DE SÉRIES TEMPORAIS VIA CLASSIFICADOR DE BAYES EMPREGANDO MODELOS LINEARES DINÂMICOS

Diana Dorgam de Aguiar dos Santos

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA, DA UNIVERSIDADE FEDERAL DO AMAZONAS, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM MATEMÁTICA.

Examinada por:

Prof. Dr. José Raimundo Gomes Pereira, UFAM

Prof. Dr. José Mir Justino da Costa, UFAM

Profa. Ph.D. Cibele Queiroz da Silva , UnB

MANAUS, AM – BRASIL
JULHO DE 2016

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Aguiar, Diana Dorgam de
A282c Classificação de séries temporais via classificador de Bayes empregando modelos lineares dinâmicos / Diana Dorgam de Aguiar. 2016
64 f.: il. color; 31 cm.

Orientador: José Raimundo Gomes Pereira
Dissertação (Mestrado em Matemática Pura e Aplicada) - Universidade Federal do Amazonas.

1. Análise discriminante. 2. Classificador de Bayes. 3. Modelos lineares dinâmicos. 4. Séries Temporais. I. Pereira, José Raimundo Gomes II. Universidade Federal do Amazonas III. Título

*À minha querida Avó Magaly*
*que me ensinou a ser forte e não*
*desistir.*

# Agradecimentos

Agradeço,

À Deus .

Ao Professor José Raimundo pela confiança, incentivo, disponibilidade e pela excelente orientação.

À CAPES (coordenação de Aperfeiçoamento Pessoal de Nível Superior) pela assistência financeira.

Ao meu belíssimo marido que me apoia em todos os momentos.

Ao Professor James Dean, pelo constante ensinamento e incentivo

Ao meu pai e minha mãe pela educação que me foi dada.

Aos meus filhos, Adriana Elizabete e Arthur Lucas, por serem minha motivação para concluir esse curso.

Ao meu amigo do mestrado, Jhonata pelo apoio computacional e companhia nas muitas horas de estudo .

A todos os meus familiares e amigos que torceram e me ajudaram diretamente ou indiretamente nessa conquista

Aos professores de estatística pelos ensinamentos.

Resumo da Dissertação apresentada ao Programa de Pós-Graduação em Matemática, da Universidade Federal do Amazonas, como parte dos requisitos necessários para a obtenção do grau de Mestre em Matemática. (M.Sc.)

# CLASSIFICAÇÃO DE SÉRIES TEMPORAIS VIA CLASSIFICADOR DE BAYES EMPREGANDO MODELOS LINEARES DINÂMICOS

Diana Dorgam de Aguiar dos Santos

Julho/2016

Orientador: José Raimundo Gomes Pereira

Área de Concentração: Estatística

Na presente dissertação apresentamos uma nova abordagem para aplicações em Análise Discriminante (AD) para problemas cujas observações no conjunto de treinamento são oriundas de séries temporais, empregando o Classificador de Bayes e modelando as distribuições nas classes com o emprego de Modelos Lineares Dinâmicos. Foram realizados os desenvolvimentos teóricos necessários para a obtenção de uma forma analítica para as probabilidades a posteriori das classes. Para avaliar a abordagem proposta foram desenvolvidos estudos de simulação, tanto para avaliar as estratégias da escolha do procedimento da estimação da variância, como também, determinar as taxas de erro (TE) de classificação para compará-las com outras abordagens usuais para classificadores em AD. Foram simuladas observações de séries temporais com diferentes estruturas de separação das classes e com diferentes tamanhos para o conjunto de treinamento. A abordagem proposta também foi aplicada em dados de problemas reais, com diferentes graus de dificuldades com relação ao número de classes, tamanho das séries e o número de observações no conjunto de treinamento, sendo então comparadas suas TE com as de outros classificadores. Embora sejam necessários estudos mais completos, os resultados obtidos sugerem que a abordagem paramétrica desenvolvida se constitui em uma alternativa promissora para esta categoria de problemas em AD, com observações de séries temporais, em particular, em um contexto bastante desafiador na prática quando temos séries com tamanhos grandes com relação ao número de observações nas classes.

Abstract of Dissertation presented to Postgraduate in Mathematics, of the Federal University of Amazonas, as a partial fulfillment of the requirements for the degree of Master of Mathematics. (M.Sc.)

TIME SERIES CASSIFICATION VIA BAYES CLASSIFIER USING DYNAMIC LINEAR MODELS

Diana Dorgam de Aguiar dos Santos

July/2016

Advisor: José Raimundo Gomes Pereira

Research lines: Statistics

In this work we present a new approach for applications in Discriminant Analysis (DA) to problems whose observations in the training set are from time series, using the Bayes classifier and modeling the classes distributions in with Linear Dynamic Models. Theoretical developments were conducted to obtain an analytic form for the classe posterior probability. The simulation studies have been developed to evaluate the proposed approach, to evaluate different strategies to estimate the model variance and determine the classification error rates (ET) to compare them with other usual approaches in AD. Time series were simulated with different structures of classes separation and with different sizes for the training set. The proposed approach was also applied to data from real problems with different degrees of difficulty with respect to the classes number, the time series size and number of observations in the training set. With real data the proposed classifier was compared with other classifiers in terms of error rate. Although it is needed most complete studies, the results suggest that this parametric approach developed constitutes a promising alternative for problems in AD with time series, particularly in a challenging context when the size time series is much large than the number of observations in the classes.

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Capítulo 1

# Introdução

Os problemas abordados em Análise Discriminante (AD) são caracterizados pela observação de um conjunto de variáveis sobre os *objetos de estudo*, que possuem características ou comportamentos distintos, com o objetivo de associá-los à *classes* previamente definidas. Na abordagem desses problemas deve ser desenvolvido um procedimento para efetuar a classificação dos objetos, sendo este procedimento denominado de *classificador*. O conjunto de variáveis é denominado *vetor de características* e, na prática, dispomos de observações do mesmo para cada uma das classes. Estas observações formam o *conjunto de treinamento* para o desenvolvimento do classificador. Obtido o classificador este pode ser empregado para classificar um novo objeto cuja classe seja desconhecida.

Podemos exemplificar uma ampla gama de objetos a serem estudados em AD, tais como: indivíduos a serem associados a classes de doentes e não doentes; plantas a serem associadas a diferentes espécies; imagens digitais de tumores a serem classificados como benignos ou maligno; sinais de espectrometria de massa a serem classificados como provenientes de diferentes fontes. A AD é uma das técnicas dentro da área de reconhecimento de padrões supervisionado e para vários outros exemplos de aplicações, veja , por exemplo, Hastie et al. (2009)

Neste trabalho, propomos uma nova abordagem para problemas em AD onde os vetores de características são séries temporais, usando o conceito de classificador de Bayes e o de Modelo Linear Dinâmico (MDL) para construir uma ferramenta capaz de se auto calibrar através de um grupo de observações cujas classes são conhecidas. Como ilustração, considere o problema descrito a seguir:

O Robô Sony AIBO é um pequeno robô quadrúpede em forma de cachorro equipado com múltiplos sensores, incluindo um acelerômetro tri axial. O conjunto de observações criado por Vail and Veloso (2004) no qual medidas do acelerômetro foram registradas enquanto o robô andava em círculos em dois tipos de superfícies: cimento e carpete. Os dados obtidos para um eixo horizontal, disponíveis em Chen et al. (2015) sob o nome SonyAIBORobot Surface. Cada série temporal representa uma volta completa. Foram registradas 621 voltas, sendo 349 no cimento e 272 no carpete. O cimento é mais duro

que o carpete, o que faz com que exista mais variabilidade na superfície. Considere cada superfície como uma classe. O objetivo é identificar qual das duas superfícies o Robô esta percorrendo novamente, com base na observação da série temporal. A Figura 1.1 mostra o gráfico das séries nas duas superfícies.

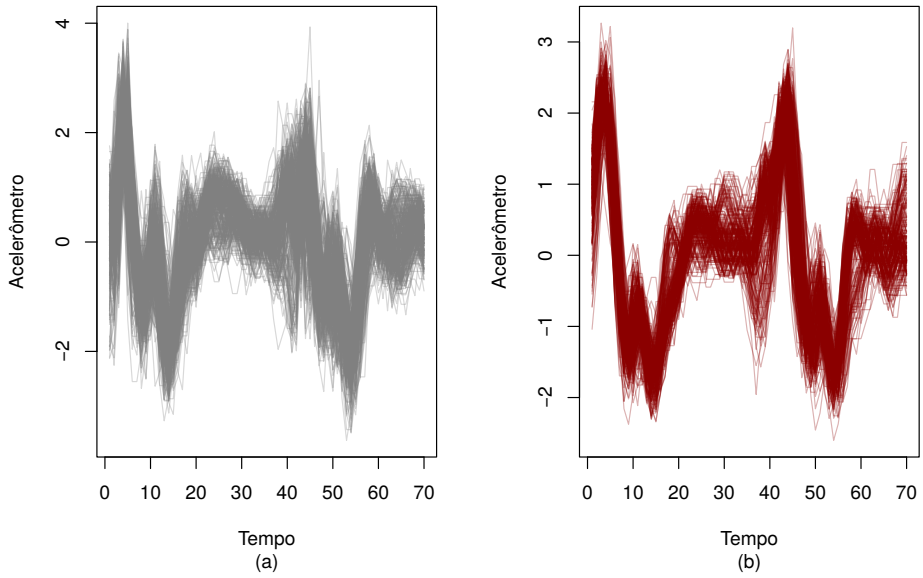


Figura 1.1: Séries do acelerômetro do Robô Sony AIBO em duas superfícies: (a) Cimento, (b) Carpete.

O problema descrito é típico dos que mencionamos, isto é, os dados observados para o vetor de características fornecem uma série temporal o que é desafiante em AD.

## 1.1 Objetivos desta Dissertação

### 1.1.1 Objetivo Geral

Construir um classificador capaz de distinguir as características de uma série temporal, considerando uma amostra de treino, pressupondo a estrutura de modelo linear dinâmico aos dados.

### 1.1.2 Objetivo Específico

Objetivo específicos:

1. Descrever a abordagem estatística para AD e o classificador de Bayes;
2. Descrever os classificadores usuais paramétricos e não paramétricos a serem empregados para comparação com o novo classificador a ser desenvolvido;

3. Apresentar as definições de MLD e sua evolução no tempo através de uma cadeia de Markov afim de modelar um classificador para séries temporais com características ajustáveis.

4. Desenvolver um classificador paramétrico partindo de suposições razoáveis, considerando que o conjunto de dados composto de séries temporais assuma a forma de MLD e fazer os devidos cálculos de evolução para determinar os parâmetros em cada estado $t$.

5. Simular em ambientes com evolução gradativa do grau de dificuldade de classificação com um número pequeno de classes e observar se o classificador construído obtém bons resultados mediante os já conhecidos.

6. Avaliar o classificador de Bayes desenvolvido, que emprega MLD, em dados reais e expor os resultados contrastados com os resultados obtidos pelos classificadores mais usuais.

## 1.2 Organização desta Dissertação

Esta dissertação está dividida em oito capítulos sendo que no primeiro apresentamos uma introdução e uma motivação com um problema real, e os objetivos deste trabalho.

No Capitulo 2 definimos nossa ferramenta principal nesta dissertação que é a análise discriminante e seus elementos. Abordamos também o classificador de Bayes, e descrevemos as principais abordagens paramétricas e não paramétricas, para implementação do classificador e, também, descrevemos o procedimento de avaliação de performance para classificadores.

No Capitulo 3 apresentamos todos os conceitos, características e definições dos modelos lineares dinâmicos (MLD) tendo em vista que vamos criar um classificador voltado para processos estocásticos com modelagem linear dinâmica.

No Capítulo 4 apresentamos o desenvolvimento da fundamentação teórica para o classificador proposto, baseado em modelo linear dinâmico.

No Capitulo 5 desenvolvemos um estudo de simulação computacional, composto por diferentes análises de classificação onde as observações são séries temporais, realizadas com intuito de compreender as características da abordagem proposta.

No Capitulo 6 apresentamos os resultados do emprego do classificador proposto CBMLD à três séries temporais reais as quais são, dados do Robô SONY AIBO, dados de tipos de café e as folhas suecas e faremos um estudo do desempenho do nosso classificador em relação à classificadores mais usuais.

No Capítulo 7 apresentamos nossas considerações sobre todos os procedimentos realizados.

# Capítulo 2

# Análise Discriminante

Neste capitulo descreveremos a modelagem estatística para AD, o classificador de Bayes, alguns classificadores mais usuais e procedimentos de avaliação dos classificadores.

## 2.1 Elementos da Análise Discriminante

A Análise Discriminante (AD), como mencionado no Capitulo 1, uma técnica de classificação dentro do campo de *Reconhecimento de Padrões Supervisionados* (RPS), aborda problemas onde devemos alocar objetos cujas classes são previamente conhecidas (ver Hastie et al. (2009), Izenman (2008) e McLachlan (2004)). Os objetos podem ser de qualquer natureza, pessoas, plantas, imagens digitais, etc., que são descritos por observações oriundas de um conjunto de variáveis. Em particular, nosso interesse são os casos onde as observações são obtidas ao longo do tempo formando uma série temporal. As classes são categorias previamente definidas onde os objetos devem ser alocados. As observações obtidas a respeito de um objeto são modeladas como um vetor aleatório $\mathbf{X}^T = (X_1, X_2, \cdots, X_t)$, onde temos que suas componentes também são variáveis aleatórias. Tal vetor $\mathbf{X}$ é denominado de *"vetor de características"*.

Considere uma variável indicadora $Z$, onde $Z = i$ indica que a classe em questão é a classe $i$, para $i \in \{1,2,3,...,M\}$, onde $M$ é o número de classes definido no problema. Para cada classe $i$, o vetor de características $\mathbf{X}$ é modelado por uma distribuição de probabilidade $f^{(i)}(.)$, uma função densidade de probabilidade ou função de probabilidade, denominada de *distribuição condicional da classe*. Nesta modelagem estatística para o problema consideramos também a probabilidade $P(Z = i) = P^{(i)}$ do objeto provir da classe $i$, para $i \in \{1,2,3,...,M\}$, denominada de *probabilidade a priori da classe*.

Suponhamos que existe uma função desconhecida $\mathfrak{F}(.)$ que associa a $\mathbf{X}$ o valor de $Z$. O nosso problema consiste em estimar essa função desconhecida ou, ainda, construir um *classificador r* que se aproxime a função $\mathfrak{F}(.)$.

**Definição 1** (Classificador)**.** *Um classificador é uma função r tal que:*

$$\begin{cases} r : \boldsymbol{X} \subset \mathbb{R}^t & \longrightarrow \{1,2,\ldots,M\} \\ \boldsymbol{x} & \longmapsto r(\boldsymbol{x}) = i \end{cases}$$

Pela definição acima, $r(\mathbf{x}) = i$ indica que um objeto com observação $\mathbf{x}$ para $\mathbf{X}$ é alocado na classe $i$.

## 2.2 Classificador de Bayes

Com as distribuições condicionais $f^{(i)}(.)$ e as probabilidades a priori $P^{(i)}$ empregamos o teorema de Bayes para obter as *probabilidades a posteriori* das classes, dadas por

$$P(Z = i|\mathbf{x}) = P(Z = i|\mathbf{X} = \mathbf{x}) = \frac{f^{(i)}(\mathbf{x})P^{(i)}}{f(\mathbf{x})}, \quad i = 1,2,3,\ldots,M, \tag{2.1}$$

onde $f(\mathbf{x}) = \sum_{i=1}^{M} f^{(i)}(\mathbf{x})P^{(i)}$ é a *densidade marginal* de $\mathbf{X}$.

A ideia é empregar elementos da Teoria da Decisão para obter o classificador de Bayes.

**Definição 2** (Função de Custo ou de Perda)**.** *Seja $\lambda$ uma função tal que:*

$$\begin{cases} \lambda : Z \times r(Y) \subset \mathbb{Z}^2 & \longrightarrow \mathbb{R} \\ (i, r(\boldsymbol{x})) & \longmapsto \lambda(i, r(\boldsymbol{x})) = \lambda(i,j) \end{cases}$$

A função de custo $\lambda$ quantifica o custo da má alocação de um objeto de classe $i$ na classe $j$, ou seja $\lambda(i,j) = e \in \mathbb{R}$, e portanto quando temos que $i = j$, $\lambda(i,j) = 0$ que significa que não houve erro na classificação.

Quando temos a informação a respeito do problema de que o custo de má alocação podem ser considerados iguais, ou quando não se pode especificar esse custo, podemos empregar a *função de perda* $0-1$ dada por:

$$\begin{cases} \lambda(i,j) = 1, & \textit{para } i \neq j \\ \lambda(i,j) = 0, & \textit{para } i = j \end{cases}$$

Note que a função de perda é uma função aplicada em uma variável aleatória, $\lambda(i,j) = \lambda(i, r(\mathbf{X}) = j)$ e portanto é uma variável aleatória.

Fixada uma função de perda, $\lambda(i,j)$, uma abordagem é desenvolver um classificador que minimize a mesma, em termos de seu valor esperado. Para esse fim, considere as definições a seguir.

**Definição 3.** *Dado um classificador r e sua função perda $\lambda(i,j)$ ;*

*a) A função Risco é a perda esperada como função de uma classe i fixada.*

$$\begin{aligned} R(r,i) &= E\{\lambda(i, r(\boldsymbol{X}) = j)|Z = i\} \\ &= \sum_{i\neq j=1}^{M} \lambda(i,j) P^{(i)}(r(\boldsymbol{X}) = j) \end{aligned}$$

*b) A função Risco Total, ou Risco de r, é a perda total esperada das variáveis aleatórias* $\boldsymbol{X}$ *para todo i.*

$$\begin{aligned} R(r) &= E\{R(r,Z)\} \\ &= \sum_{i=1}^{M} R(r,i) P^{(i)} \\ &= \sum_{i=1}^{M} \sum_{i\neq j=1}^{M} \lambda(i,j) P^{(i)}(r(\boldsymbol{X}) = j) P^{(i)} \end{aligned}$$

Para a função $0-1$ temos que o risco é da forma:

$$R(r,i) = \sum_{i\neq j=1}^{M} P(r(\mathbf{X} = j)|Z = i) \tag{2.2}$$

e

$$R(r) = \sum_{i=1}^{M} \sum_{i\neq j=1}^{M} P(r(\mathbf{X}) = j|Z = i) P^{(i)} \tag{2.3}$$

Temos de (2.2) e (2.3), vemos que $R(r;i)$ e $R(r)$ são funções das taxas de alocação. Para a função de perda $0-1$, portanto, $R(r;i)$ é a probabilidade de classificação errônea dos objetos da classe $i$ e $R(r)$ é a probabilidade total de classificação errônea do classificador $r$. A probabilidade total de classificação errônea para $r$ também recebe a denominação de erro de classificação de $r$.

Estabelecida a função de perda $\lambda(i,j)$, o objetivo é construir um classificador que minimize o risco total $R(r)$. Para esse fim, consideremos o seguinte classificador:

$$r^*(\mathbf{x}) = k \; se \; \sum_{i=1}^{M} \lambda(i,k) f^{(i)}(\mathbf{x}) P^{(i)} = min_j \sum_{i=1}^{M} \lambda(i,j) f^{(i)}(\mathbf{x}) P^{(i)} \tag{2.4}$$

*No caso de o mínimo ocorrer para mais de uma classe, o objeto é associado a qualquer uma das classes que o atingirem.*□

O teorema a seguir estabelece que o classificador definido acima minimiza o risco total.

**Teorema 1.** *Dado uma função perda* $\lambda(i,j)$ *o classificador* $r^*$ *minimiza o risco total, ou*

*seja,* $R(r^*) \leq R(r)$ *para todo classificador* $r$*.*

*Demonstração.* Vamos usar a propriedade básica da esperança aplicando na função risco total convenientemente temos,

$$R(r) = E\{E[R(r,Z)|\mathbf{X}]\} \tag{2.5}$$

$$= \int_{\mathbb{R}^d} E[R(r,Z)|\mathbf{X} = \mathbf{x}] f^{(i)}(\mathbf{x}) dx \tag{2.6}$$

Assim podemos observar que para minimizar o risco total $R(r)$ basta minimizar a esperança condicional no integrando da equação $(2.6)$.

$$E\left[\lambda(Z, r(\mathbf{x}) = j|\mathbf{X} = \mathbf{x})\right] = \sum_{i=1}^{M} \lambda(i,j) P(Z = i|\mathbf{X} = \mathbf{x}) = \sum_{i=1}^{M} \lambda(i,j) \frac{f^{(i)}(\mathbf{x}) P^{(i)}}{f(\mathbf{x})} \tag{2.7}$$

Da equação $(2.6)$ vemos que a esperança condicional será minimizada se tomarmos uma classe $Z = k$ para a qual $\sum_{i=1}^{M} \lambda(i,k) f^{(i)}(\mathbf{x}) P^{(i)}$ é um mínimo. □

Como as expressões para as probabilidades a posteriori das classes tem o mesmo denominador $f(\mathbf{x})$ (veja (2.1)), a regra em (2.4) pode ser estabelecida em termos dessas probabilidades, ou seja,

$$r^*(\mathbf{x}) = k \quad se \quad \sum_{i=1}^{M} \lambda(i,k) P(Z = i|\mathbf{x}) = min_j \left\{ \sum_{i=1}^{M} \lambda(i,j) P(Z = i|\mathbf{x}) \right\} \tag{2.8}$$

Da equação $(2.8)$, com a função de perda $0-1$ temos que:

$$\sum_{i=1}^{M} \lambda(i,k) P(Z = i|\mathbf{x}) = \sum_{k \neq i=1}^{M} P(Z = i|\mathbf{x}) = 1 - P(Z = k|\mathbf{x}), \tag{2.9}$$

então, minimizar o risco total é equivalente a selecionar a classe com maior probabilidade a posteriori. Portanto, com função de perda $0-1$ o classificador de Bayes fica forma

$$r^*(\mathbf{x}) = k \quad se \quad P(Z = k|\mathbf{x}) = max_j \{P(Z = j|\mathbf{x})\} \tag{2.10}$$

Uma a vez que $r^*$ minimiza o risco total, o valor do risco de Bayes é o menor valor que pode ser atingido por qualquer classificador e, por isso, serve como referência para comparação de classificadores. No caso da função de perda $0-1$, $e^*$ é equivalente ao erro de classificação de $r^*$.

O classificador de Bayes está bem definido teoricamente, porém, na prática as distribuições condicionais, as probabilidades a priori e, consequentemente, as probabilidades a posteriori são desconhecidas, sendo necessário estimá-las. Descrevemos a seguir algumas abordagens estatísticas mais usuais empregadas para estimar, com base no conjunto

de treinamento, estes elementos necessários para implementação prática do classificador de Bayes.

## 2.3 Abordagens Usuais para o Classificador de Bayes

### 2.3.1 Classificação com Modelos Normais

Um dos objetivos principais na classificação é estimar qual a distribuição que cada classe apresenta diferenciando-as pelos seus parâmetros. Na abordagem para o classificador de Bayes com modelos normais em cada classe a densidade $f^{(i)}(.)$ é assumida como normal multivariada com seu conjunto de parâmetros $(\mu^{(i)}, \Sigma^{(i)})$ onde a matriz de covariância é não singular também por suposição, ou seja,

$$f^{(i)}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma^{(i)}|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(x-\mu^{(i)})^T\Sigma^{-1(i)}(x-\mu^{(i)})\right\} \quad i = 1,2,\cdots,M \quad (2.11)$$

Conhecendo sua probabilidade a priori $P^{(i)}$ o modelo $(2.11)$ é usado para determinar $r^*$ (vide expressão (2.8)). Empregando a função logaritmo obtemos

$$d^{(i)Q}(\mathbf{x}) = ln\left\{f^{(i)}P^{(i)}\right\} = -\frac{d}{2}ln(2\pi) - \frac{1}{2}(\boldsymbol{x}-\mu^{(i)})^T\Sigma^{-1(i)}(\boldsymbol{x}-\mu^{(i)}) + lnP^{(i)}, \quad (2.12)$$

ficando o classificador na forma

$$r^*(\mathbf{x}) = k \quad se \quad d^{(k)Q}(\mathbf{x}) = max_j\left\{d^{(j)Q}(\mathbf{x})\right\}. \quad (2.13)$$

Os modelos normais com matrizes de covariância iguais são conhecidos por modelos normais homocedásticos e no caso contrário, para modelos normais cujas matrizes de covariância são diferentes são denominados modelos normais heterocedásticos. Sendo assim, podemos considerar algumas simplificações, como por exemplo no caso homocedástico onde $\Sigma^{(j)} = \Sigma \ \forall j$, onde podemos expandir a forma quadrática e desprezando os termos que são constantes para todas as classes, obtemos:

$$d^{(i)L}(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu^{(i)})^T\Sigma^{-1}(\mathbf{x}-\mu^{(i)}) + ln(P^{(i)}) \quad (2.14)$$

ficando o classificador da forma

$$r^*(\mathbf{x}) = k \quad se \quad d^{(k)L}(\mathbf{x}) = max_j\left\{d^{(j)L}(\mathbf{x})\right\}. \quad (2.15)$$

Com a maximização em $(2.15)$ denominamos $max_j\{d^{(i)L}(\mathbf{x})\}$ como *Análise Discriminante Linear* (ADL) e é assim denotada por $d^{(i)L}(\mathbf{x})$ ser linear em $\mathbf{x}$. Se multiplicarmos $d^{(i)L}(\mathbf{x})$ por $-2$ na equação $(2.14)$ e minimizarmos $j$ vamos obter uma forma equivalente

de:

$$d^{(i)L}(\mathbf{x}) = (\mathbf{x} - \mu^{(i)})^T \Sigma^{-1} (\mathbf{x} - \mu^{(i)}) - 2ln(P^{(i)}) \tag{2.16}$$

assim, o classificador fica na forma

$$r^*(\mathbf{x}) = k \quad se \quad d^{(k)L}(\mathbf{x}) = min_j \left\{ d^{(j)L}(\mathbf{x}) \right\}. \tag{2.17}$$

O primeiro termo no segundo membro da igualdade (2.16) é, por definição, a distância de Mahalanobis ao quadrado entre $\mathbf{x}$ e $\mu^{(i)}$. No caso de se ter as probabilidades a priori iguais para todas as classes, então para um objeto com vetor de observação $\mathbf{x}$, a regra seleciona a classe que possui a menor distância de Mahalanobis entre o vetor de médias e $\mathbf{x}$. Se a matriz de covariância $\Sigma$ for proporcional à matriz identidade, então essa proximidade pode ser calculada com a distância Euclidiana.

Para o caso onde temos a heterocedasticidade, multiplicando por $-2$ e desprezando o termo $-\frac{d}{2} ln(2\pi)$ em (2.12), temos.

$$d^{(i)Q}(\mathbf{x}) = ln|\Sigma^{(i)}| + (\mathbf{x} - \mu^{(i)})^T (\Sigma^{(i)})^{-1} (\mathbf{x} - \mu^{(i)}) - 2ln(P^{(i)}) \tag{2.18}$$

assim, obtemos o classificador na forma

$$r^*(\mathbf{x}) = k \quad se \quad d^{(k)Q}(\mathbf{x}) = min_j \left\{ d^{(j)Q}(\mathbf{x}) \right\}. \tag{2.19}$$

Logo, temos que $d^{(i)Q}(\mathbf{x})$ tem forma quadrática em $\mathbf{x}$ , e por isso, a aplicação desta forma é denominado de Análise Discriminante Quadrática (ADQ).

Ao modelar as classes com distribuições normais multivariadas, estamos admitindo que as observações do vetor de características pertencem a elipsoides no espaço d-dimensional. Tais elipsoides são centradas nos vetores de médias $\mu^{(j)}$ e suas formas são determinadas pelas matrizes de covariância $\Sigma^{(j)}$. Além disso, as regras de alocação obtidas definem as fronteiras de decisão através de hiperplanos no caso de modelos homocedásticos e heterocedásticos temos que essas fronteiras são quadráticas(veja Duda et al. (2000), Capítulo 2). Para empregar os conceitos acima é necessário que os parâmetros $(\mu^{(j)}, \Sigma^{(j)})$ e as probabilidades a priori $P^{(j)}$ sejam estimados. As estimativas são feitas a partir de observações no conjunto de treino. Em alguns casos o especialista pode dispor de informações sobre a probabilidade priori, porem não sendo esse o caso pode-se estimá-la através do estimador de máxima verosimilhança.

O desenvolvimento dos passos necessários para a determinação dos estimadores de máxima verossimilhança para os parâmetros em distribuições normais já são bastante conhecidos na literatura ( para isso, veja por exemplo, Mardia et al. (1979)).

É importante destacar que tanto na ADL e ADQ torna-se necessário estimar a matriz de covariâncias, sendo uma para a ADL e uma para cada classe na ADQ. Na forma da expressão desses classificadores emprega-se a inversa destas matrizes. Para o tipo de

problemas abordado nesta dissertação, esta matriz é de alta dimensão pois o vetor de características (a série temporal) é de alta dimensão. Esta inversão leva a um problema computacional, pois é comum termos poucas séries temporais, gerando uma matriz de covariâncias singular. Esta dificuldade inviabiliza o emprego da ADL e da ADQ em muitos problemas reais cujo vetor de observações é uma série temporal de alta dimensão.

### 2.3.2 Análise Discriminante Regularizada

Em Friedman (1989) é proposta a Análise Discriminante Regularizada (ADR) como uma combinação entre ADL e ADQ. O autor apresenta um método para regularizar as matrizes de covariância dentro das classes.

Primeiramente, são obtidas as matrizes de covariâncias estimadas $\hat{\Sigma}_k$ para cada classe e estimada a matriz de covariância combinada $\hat{\Sigma}$ dada por

$$\hat{\Sigma} = \frac{\sum_{j=1}^{M}(n_j - 1\hat{\Sigma}_j)}{n_1 + n_2 + \cdots + n_M - M} \tag{2.20}$$

Em seguida, para $\lambda \in [0,1]$, é especificada a seguinte combinação convexa

$$\hat{\Sigma}_k(\lambda) := (1-\lambda)\hat{\Sigma}_k + \lambda\hat{\Sigma}.$$

Após este passo, para $\gamma \in [0,1]$, outra combinação convexa é estabelecida

$$\hat{\Sigma}_k(\lambda,\gamma) = (1-\gamma)\hat{\Sigma}_k(\lambda) + \gamma\frac{1}{d}tr[\hat{\Sigma}_k(\lambda)]I_d, \tag{2.21}$$

onde $d$ é o número de variáveis.

Desta forma, a matriz de covariâncias estimada de cada classe é dada pela equação (2.21) obtida determinado os valores de $\lambda$ e $\gamma$ que minimizam a taxa de erro. Para os quatro valores extremos de $\gamma$ e $\lambda$ a estrutura de covariâncias estimadas se reduz aos seguintes casos especiais:

- Para $\gamma = 0$ e $\lambda = 0$: Covariância individual para cada classe ADQ.
- Para $\gamma = 0$ e $\lambda = 1$: Uma matriz de covariância comum para todas as classes ADL.
- Para $\gamma = 1$ e $\lambda = 0$: Os elementos da diagonal principal da matriz de covariância são iguais dentro de cada classe.
- Para $\gamma = 1$ e $\lambda = 1$: Similar ao caso anterior, mas as variâncias são as mesmas para todos as classes.

Algumas extensões tem sido desenvolvidas para a regularização das matrizes de covarianças das classes em AD, veja por exemplo, Dai and Yuen (2003) e Witten and Tibshi-

rani (2009). Para maiores discussões sobre a ADR veja, por exemplo, Hastie et al. (2009), Subseção 4.3.1.

### 2.3.3 Naive Bayes Normal e com Estimadores por Função Núcleo

O Classificador Naive Bayes assume que em cada classe as variáveis que compõem o vetor de características $\mathbf{X}$ são independentes. Embora esta hipótese não seja geralmente verdadeira, ela simplifica a estimativa das densidades condicionais drasticamente. Apesar destas hipóteses bastante otimistas, o classificador do Naive Bayes muitas vezes superam alternativas muito mais sofisticados sendo bastante apropriado em problemas de AD quando o número de variáveis do vetor de características $\mathbf{X}$ é muito grande, em particular, quando o tamanho da amostra é muito menor que o número de observações.(Hastie et al. (2009), Seção 6.6).

Com modelos normais, que denominamos como Naive Bayes Normal (NBN), as densidades marginais condicionais de classe individuais $f^{(j)}(\mathbf{x})$ podem ser ajustadas utilizando separadamente estimativas dos parâmetros das distribuições normais unidimensionais, ou seja,

$$f^{(j)}(\mathbf{x}) = \prod_{l=1}^{t} \frac{1}{\sqrt{2\pi}\sigma_l^{(j)}} \exp\{\frac{1}{2\sigma_l^{(j)}}(x_l - \mu_l^{(j)})^2\}, \quad j = 1,2,\cdots,M. \tag{2.22}$$

Outra abordagem bastante empregada é estimar as densidades marginais em $\mathbf{X}$ empregando estimação de densidades por função núcleo (*Kernel Density Estimation*), que denominamos Naive Bayes Kernel (NBK). Neste caso as densidades condicionais são da forma

$$\hat{f}_h^{(j)}(\mathbf{x}) = \prod_{l=1}^{t} \frac{1}{n^{(j)} h_l^{(j)}} \sum_{l=1}^{n^{(j)}} K\left(\frac{x_{li}^{(j)} - x_{li}}{h_l^{(j)}}\right), \quad j = 1,2,\cdots,M. \tag{2.23}$$

Para maiores discussões sobre estimadores por função núcleo veja, por exemplo, Izenman (2008), Capítulo 4.

Em Domingos and Pazzani (1997), os autores discutem as condições de consistência do Naive Bayes e demonstram que o classificador pode ser ótimo mesmo quando a suposição de independência é violada, considerando a função de perda $0-1$.

No artigo de Bickel and Levina (2004), os autores discutem a consistência do Naive Bayes quando o número de variáveis aumenta de forma mais rápida que o número de observações. Os autores consideram apenas duas classes modeladas com distribuições normais multivariadas. Além das deduções teóricas das condições de consistência, os autores concluem que às vezes o Naive Bayes apresenta melhor desempenho que outras modelagens que estimam a estrutura de dependência das observações.

### 2.3.4 Classificador com os $K$ Vizinhos Mais Próximos (K-NN)

Inicialmente, com todas as observações das $M$ classes no conjunto de treino, é formado um único conjunto com $n$ observações ($\Sigma_{j=1}^{M} n^{(j)} = n$). Seja $V_k(\mathbf{x})$ o volume de uma hiper esfera em torno de $\mathbf{x}$ necessária para conter um número fixo $k$ de pontos, onde vamos supor que entre os $k$ pontos, $k^{(j)}$ sejam os pontos pertencentes a classe $j$. Então definiremos o classificador pelos seus $k$ vizinhos mais próximos para a densidade condicional dessa classe dada por:

$$\hat{f}_{(KNN)}^{(j)}(\mathbf{x}) = \frac{k^{(j)}}{n^{(j)} V_k(\mathbf{x})}. \tag{2.24}$$

Para verificar que (2.24) é um estimador por função núcleo, considere que $\{\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \cdots, \mathbf{x}_{(t)}\}$ são observações do conjunto de treino na classe $j$, em ordem crescente de acordo com a distancia euclidiana entre cada observação e $\mathbf{x}$. Podemos então escrever:

$$\hat{f}_{(KNN)}^{(j)}(\mathbf{x}) = \frac{1}{n^{(j)}} \sum_{i=1}^{n^{(j)}} \frac{w(\mathbf{x}_{(i)})}{V_k(\mathbf{x})} \mathbf{I}_{\{(i)=j\}} \boldsymbol{X}_{(i)}, \tag{2.25}$$

onde $w(\mathbf{x}_{(i)}) = 1$ se $(i) \leq k$ e $w(\mathbf{x}_{(i)}) = 0$ se $(i) > k$.

Estimando as densidades a priori como $\hat{P}^{(j)} = \frac{n^{(j)}}{n}$ a densidade não condicional de $\mathbf{X}$ é estimada por

$$\hat{f}_{(KNN)}(\mathbf{x}) = \Sigma_{j=1}^{M} \left( \frac{n^{(j)}}{n} \right) \frac{k^{(j)}}{n^{(j)} V^{(k)}(\mathbf{x})}. \tag{2.26}$$

Empregando $(2.24)$ e $(2.26)$ temos que as probabilidades a posteriori são estimadas por

$$\hat{P}_{(KNN)}(Z = j|\mathbf{x}) = \frac{\left(\frac{n^{(j)}}{n}\right) \frac{k^{(j)}}{n^{(j)} V_k(\mathbf{x})}}{\frac{k}{n V_k(\mathbf{x})}} = \frac{k^{(j)}}{k}, \tag{2.27}$$

para $j \in \{1, 2, 3, \cdots, M\}$.

Empregando as probabilidades definidas na equação $(2.27)$ a regra de $K-NN$ é definida por

$$r_{(KNN)}(\mathbf{x}) = j \text{ se } k^{(j)} = max_i \quad k^{(i)}. \tag{2.28}$$

Da definição $(2.28)$, podemos observar que, para $k = 1$, a regra aloca a observação a ser classificada na classe do vizinho mais próximo em distância euclidiana, esta regra chamaremos de *regra do vizinho mais próximo* e denotaremos como $(1-NN)$.

As propriedades do classificador $K-NN$ estão bem estabelecidas na literatura, em particular, para maiores aprofundamentos sobre este classificador citamos McLachlan (2004), Seção 9.7. Em problemas de AD com observações provenientes de séries temporais, alguns autores consideram o classificador $1-NN$ como "padrão-ouro", portanto neste trabalho, será considerado como principal parâmetro de comparação (ver, por exemplo, Bagnall et al. (2012), e as referências dentro do artigo).

## 2.4 Critérios para Avaliar um Classificador

### 2.4.1 Validação Cruzada

Em problemas de classificação (como, por exemplo, predizer a classe de um objeto) é necessário avaliar como um modelo preditivo (um classificador) irá se comportar na prática, com relação a seu desempenho com novas observações cujas classes são desconhecidas. Uma abordagem usual consiste em dividir o conjunto de treinamento em duas partes: um conjunto denominado *treino* e outro *teste*. O classificador é construído utilizando apenas o conjunto de treino e a capacidade preditiva do classificador é avaliada com base no conjunto de teste. No entanto, em muitos problemas reais o conjunto de treinamento não é suficientemente grande e, ao realizar a divisão, haverá poucas observações no conjunto de treino, o que prejudica o ajuste (estimação) do classificador, como também no conjunto de teste o que não permite uma estimativa confiável da taxa de erro do classificador.

Para contornar as dificuldades mencionadas na avaliação do classificador, uma alternativa é o emprego da validação cruzada (*cross-validation*).

Podemos dividir os métodos de validação cruzada em dois tipos: exaustivos e não exaustivos. Nos métodos exaustivos, todos as formas possíveis de se particionar a amostra nos conjuntos do tipo treino e teste são considerados. Nos métodos não exaustivos, apenas parte dos possíveis conjuntos do tipo treino e teste são considerados.

Dentre os métodos de validação cruzada exaustivos listamos os seguintes:

- "Deixa $p$ de fora" (tradução livre do termo *leave-p-out*): neste tipo de validação cruzada, $p$ observações são utilizadas como conjunto de teste e as $n-p$ restantes forma o conjunto de treino. O procedimento é repetido até que todos os subconjuntos de tamanho $p$ tenham sido selecionados. Existem $\binom{n}{p}$ subconjuntos, o que torna este método computacionalmente inviável para valores grandes de $n$.

- "Deixa 1 de fora" (tradução livre do termo *leave-1-out*): é um caso particular da anterior, com $p = 1$. Sua importância está relacionada com seu custo computacional: para uma amostra de tamanho $n$, existem apenas $n$ partições da amostra para serem consideradas. Neste caso, o procedimento de estimação e classificação é repetido $n$ vezes, de igual modo ao *leave-p-out* mas com $p = 1$, portanto sempre teremos $n-1$ restante para conjunto de teste.

Dentre os métodos de validação cruzada não exaustivos listamos os seguintes:

- "Validação Cruzada com $k$ fora" (do inglês *k-fold cross-validation*): a amostra original é particionada aleatoriamente em $k$ conjuntos de tamanhos iguais. Destes $k$ conjuntos, um é utilizado como teste e os $k-1$ restantes como treino. O processo

de validação cruzada é repetido $k$ vezes, sendo que cada conjunto só pode ser utilizado como teste uma única vez. Note que se $k$ for igual ao tamanho do conjunto de treinamento, este método é denominado na literatura por *leave-one-out cross-validation*.

- Repedidas subamostras aleatórias (tradução livre do termo *Repeated random subsampling validation*): este método também é conhecido como validação cruzada de Monte Carlo. Em cada repetição, o conjunto de dados particionado ao acaso em dois subconjuntos constituindo a amostra de treino e teste. Diferente do método anterior, o número de repetições não depende do tamanho do conjunto de teste. Entretanto, como a partição é escolhida ao acaso, é possível que algumas observações nunca sejam escolhidas para fazer parte do conjunto de teste. Para evitar este tipo de situação, recomenda-se fazer muitas repetições. Em particular, este foi o método utilizado nas comparações desta dissertação.

O objetivo da validação cruzada é estimar o nível esperado de ajuste do classificador a um conjunto de dados, independentemente do conjunto que foi utilizado como treino. Em geral, para cada repetição, utiliza-se qualquer medida de ajuste tradicionalmente utilizada para avaliar classificadores. Após o termino das repetições, retiramos a média destas medidas e avaliamos o desempenho médio do classificador.

Como classificadores são construídos para minimizar a perda 0-1 (na qual perdemos uma unidade sempre que cometemos um erro de classificação), uma medida natural para avaliar um classificar é sua taxa de erro:

$$\text{Taxa de erro} = \frac{\text{Número de classificações errôneas}}{\text{Total de classificações}}.$$

Utilizando o método de validação cruzada para repetidas subamostras aleatórias, calculamos $\tau_1, \ldots, \tau_N$, onde $\tau_i$ é a taxa de erro na $i$-ésima repetição e $N$ é o número de repetições. Para cada estimador calculamos as medidas

$$\bar{\tau} = \sum_{i=1}^{N} \frac{\tau_i}{N}$$

e

$$s_\tau = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (\tau_i - \bar{\tau})^2},$$

representando a média e o desvio padrão dos erros de classificação.

### 2.4.2 Abordagem Empregada para Comparar Classificadores

Como em geral estas taxas médias tendem a ser próximas para alguns métodos neste trabalho, comparamos também os classificadores em termos da proporção de vezes que

a taxa de erro de um dado classificador é menor ou igual a de outro classificador. Deste modo, calculamos quantas vezes o Classificador *A* conseguiu uma taxa de erro igual ou menor que o Classificador *B*. A tabela abaixo ilustra como este tipo de resultado foi sumarizado.

No quadro abaixo ilustramos o procedimento descrito para comparação das taxas de erro dos classificadores.

| | | Classificador A | Classificador B |
|---|---|---|---|
| Classificador A | Igual | - | 75% |
| | Menor | - | 20% |
| | Total | - | 95% |
| Classificador B | Igual | 75% | - |
| | Menor | 10% | - |
| | Total | 85% | - |

Do quadro acima podemos fazer as seguintes inferências:

- Em 75% das vezes os classificadores tem desempenho igual
- Em 20% das vezes o desempenho do Classificador A é melhor que o B.
- A linha Total dá a porcentagem de classificações boas de um método em relação ao outro. Por exemplo, ter escolhido o Classificador A resultaria em boas classificações em 95% das vezes, enquanto que com o Classificador B obteríamos boas classificações em 85%. Neste sentido o Classificador A é mais eficiente que o B.

# Capítulo 3

# Tópicos de Modelos Lineares Dinâmicos

Neste capítulo vamos abordar alguns tópicos de modelos lineares dinâmicos (veja Harrison and West, 1999 para maiores detalhes). Discutiremos a obtenção das distribuições *a posteriori* envolvidas e alguns modelos particulares que serão utilizados neste trabalho.

## 3.1 Modelo Linear Dinâmico

Vamos adotar a notação $Y_{1:t} := (Y_1, Y_2, \ldots, Y_t)$ para uma série temporal, onde as observações são recebidas sequencialmente ao longo do tempo com $t$ observações e $Y_{0:t}$ para série temporal com informação inicial que denotaremos por $Y_0$ (tal informação inicial reflete o conhecimento do processo antes de realizar a primeira observação). De modo análogo, defina $\mathbf{Y}_{1:t} := (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_t)$, representando uma série temporal onde vetores de dimensão $m$ são observados sequencialmente e $\mathbf{Y}_0$ representa a informação inicial .

O Modelo Linear Dinâmico (MLD) é caracterizado pela seguinte quádrupla:

$$\{\boldsymbol{F}, \boldsymbol{G}, \boldsymbol{V}, \boldsymbol{W}\}_t = \{\boldsymbol{F}_t, \boldsymbol{G}_t, \boldsymbol{V}_t, \boldsymbol{W}_t\}$$

para cada tempo t, onde:

1. $\boldsymbol{F}_t$ é uma matriz com dimensões $(p \times m)$;
2. $\boldsymbol{G}_t$ é denominada matriz de evolução, com dimensões $(p \times p)$;
3. $\boldsymbol{V}_t$ é denominada matriz de variância observacional, com dimensões $(m \times m)$;
4. $\boldsymbol{W}_t$ é denominada como a matriz de variância dos estados, com dimensões $(p \times p)$;

A quádrupla define o modelo que relaciona o vetor aleatório $\mathbf{Y}_t$ com o vetor de estados $\boldsymbol{\theta}_t$ no tempo $t$ através das seguintes distribuições de probabilidade:

$$(\mathbf{Y}_t|\boldsymbol{\theta}_t) \sim N[\boldsymbol{F}_t^T\boldsymbol{\theta}_t, V_t], \tag{3.1}$$

$$(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) \sim N[\boldsymbol{G}_t\boldsymbol{\theta}_{t-1}, W_t], \tag{3.2}$$

onde assume-se implicitamente que as distribuições são condicionais a $\mathbf{Y}_{0:t-1}$, o conjunto de informações disponível antes do tempo $t$. Também pode-se representar as probabilidades acima com as seguintes equações

$$\mathbf{Y}_t = \boldsymbol{F}_t^T \boldsymbol{\theta}_t + \boldsymbol{\nu}_t,\ \boldsymbol{\nu}_t \sim N[\mathbf{0}, \boldsymbol{V}_t] \tag{3.3}$$

$$\boldsymbol{\theta}_t = \boldsymbol{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t,\ \boldsymbol{\omega}_t \sim N[\mathbf{0}, \boldsymbol{W}_t], \tag{3.4}$$

onde as sequências de erros $\boldsymbol{\nu}_t$ e $\boldsymbol{\omega}_t$ são mutuamente independentes entre si e dentro de cada série. Denominamos $\boldsymbol{\nu}_t$ como o erro observacional e $\boldsymbol{\omega}_t$ como o erro de evolução ou erro de estados. A Equação (3.3) é denominada a equação de observação do modelo e define a distribuição de $\mathbf{Y}_t$ condicionado ao vetor de estados $\boldsymbol{\theta}_t$. A independência condicional é válida aqui e os $\mathbf{Y}_t$ são independentes entre si, dado o vetor $\boldsymbol{\theta}_t$. Esta equação relaciona a variável resposta $\mathbf{Y}_t$ ao vetor de estados $\boldsymbol{\theta}_t$ através de uma regressão linear com erros que tem uma distribuição normal multivariada se a variável resposta for multivariada também.

A Equação (3.4) é denominada equação de evolução ou equação dos estados e define a evolução do vetor de estados. A evolução é dada por uma cadeia de Markov, como descrito anteriormente, de forma que dado $\boldsymbol{\theta}_{t-1}$ e os valores de $\boldsymbol{G}_t$ e $\boldsymbol{W}_t$, $\boldsymbol{\theta}_t$ é independente de $\boldsymbol{\theta}_{0:t-2}$.

Agora pode-se introduzir formalmente a definição geral do MLD:

**Definição 4.** *Para cada índice de tempo $t$, o Modelo Linear Dinâmico multivariado é definido por*

$$\begin{aligned} \textit{Equação de Observação: } \mathbf{Y}_t &= \boldsymbol{F}_t^T \boldsymbol{\theta}_t + \boldsymbol{\nu}_t,\ \boldsymbol{\nu}_t \sim N[\mathbf{0}, \boldsymbol{V}_t] \\ \textit{Equação de Sistema: } \boldsymbol{\theta}_t &= \boldsymbol{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t,\ \boldsymbol{\omega}_t \sim N[\mathbf{0}, \boldsymbol{W}_t] \\ \textit{Priori Inicial:} (\boldsymbol{\theta}_0 | Y_0) &\sim N[\boldsymbol{m}_0, \boldsymbol{C}_0], \end{aligned}$$

*onde assume-se que as sequências de erros observacionais $\boldsymbol{\nu}_t$ e de evolução $\boldsymbol{\omega}_t$ são independentes ao longo do tempo e entre si, e independentes da priori $(\boldsymbol{\theta}_0 | \boldsymbol{Y}_0)$.*

Note que o erro $\boldsymbol{\nu}_t$ é simplesmente uma perturbação aleatória no processo de medida das observações $\mathbf{Y}_t$. Esse erro de evolução $\boldsymbol{\omega}_t$, influencia no desenvolvimento do sistema ao longo do tempo. Supondo que estes erros são independentes entre si, claramente separa estas duas fontes de variação estocástica e torna mais nítido o papel que cada uma representa. Se algum componente é dado como conhecido, basta assumir que a sua respectiva variância/covariância é zero.

### 3.1.1 Modelo Polinomial de Ordem 1

O MLD mais simples é o Modelo Polinomial de Ordem 1, também denominado de *passeio aleatório*. Este modelo é usado sobretudo para previsões de curto prazo, sendo caracterizado pela quadrupla.

$$\{1, 1, V_t, W_t\},$$

sendo introduzido formalmente pela seguinte definição:

**Definição 5.** *Para cada índice de tempo t, o Modelo Linear Dinâmico de ordem 1 é definido por*

$$\begin{aligned} \textit{Equação de Observação: } Y_t &= \theta_t + \nu_t,\ \nu_t \sim N[0, V_t] \\ \textit{Equação de Sistema: } \theta_t &= \theta_{t-1} + \omega_t,\ \omega_t \sim N[0, W_t] \\ \textit{Priori Inicial: } (\theta_0|Y_0) &\sim N[m_0, C_0], \end{aligned}$$

*onde assume-se que as sequências de erros observacionais $\nu_t$ e de evolução $\omega_t$ são independentes ao longo do tempo e entre si, e independentes da priori $(\theta_0|\boldsymbol{Y}_0)$.*

### 3.1.2 Modelo Polinomial de Ordem 2

Modelos Polinomiais de Ordem 2 são usados para descrever as séries temporais que apresentam tendência linear. Estes modelos também são denominados de Modelos de Crescimento Linear (*linear growth models*) e são caracterizados pela quádrupla

$$\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, V_t, \begin{pmatrix} W_1 & 0 \\ 0 & W_2 \end{pmatrix}_t \right\}, \tag{3.5}$$

Sendo sua definição formal dada por

**Definição 6.** *Para cada índice de tempo t, o Modelo Linear Dinâmico de ordem 2 é definido por*

$$\begin{aligned} \textit{Equação de Observação: } Y_t &= \theta_{1,t} + \nu_t,\ \nu_t \sim N[0, V_t] \\ \textit{Equação de Sistema: } \theta_{1,t} &= \theta_{1,t-1} + \theta_{2,t-1} + \omega_{1,t},\ \omega_{1,t} \sim N[0, W_{1,t}] \\ \theta_{2,t} &= \theta_{2,t-1} + \omega_{2,t},\ \omega_{2,t} \sim N[0, W_{2,t}] \\ \textit{Priori Inicial: } (\boldsymbol{\theta}_0|Y_0) &\sim N[\boldsymbol{m}_0, \boldsymbol{C}_0], \end{aligned}$$

*onde*

$$\boldsymbol{\theta}_t = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}_t \tag{3.6}$$

*e onde assume-se que as sequências de erros observacionais* $\nu_t$ *e de evolução* $\omega_{1,t}$ *e* $\omega_{2,t}$ *são independentes ao longo do tempo e entre si, e independentes da priori* $(\boldsymbol{\theta}_0|Y_0)$.

Como usual, $\theta_{1,t}$ é o nível da série e $\theta_{2,t}$ representa o crescimento incremental onde podemos notar este fato observando a função de previsão $k$ passos a frente para este modelo, a qual é dada por:

$$f_t(k) = FG^{(k)}m_t = m_{1,t} + km_{2,t},$$

onde $\boldsymbol{m}_t = E(\boldsymbol{\theta}_t|Y_{0:t})$. Logo, a média a posteriori de $\theta_{1,t}$ representa o nível médio da previsão e a média de $\theta_{2,t}$ representa o crescimento incremental, ou tendência linear.

### 3.1.3 Modelo com Representação Trigonométrica

O comportamento sazonal é um padrão que se repete em intervalos regulares de tempo. Em geral, tais comportamentos podem ser descritos por funções cíclicas (ou periódicas). Dizemos que a função real $g(.)$ definida para os inteiros não negativos é cíclica se, para algum inteiro $p \geq 1$, $g(t+np) = g(t)$ para todo inteiro $t \geq 0$ e todo $n \geq 0$. Neste caso, $p$ é denominado período da função e $\theta_j = g(j+np)$, para $j = 1,\ldots,p$ são denominados *fatores sazonais*.

Modelos lineares dinâmicos com funções de previsão dadas por

$$f_t(k) = g(k),$$

são denominados Modelos Lineares Dinâmicos Sazonais. Existem duas formas tradicionais deste tipo de modelo: forma livre e representação trigonométrica. Para este último, notemos que os $p$ fatores sazonais podem ser representados por

$$\theta_j = \begin{cases} a_0 + \sum_{r=1}^{q} A_r \cos(\omega r t + \phi_r), & \text{com } q = (p+1)/2 \text{ se } p \text{ é ímpar} \\ a_0 + \sum_{r=1}^{q} A_r \cos(\omega r t + \phi_r) + a_q \cos(\pi t), & \text{com } q = p/2 \text{ se } p \text{ é par} \end{cases}$$

onde $t$ corresponde ao $j$-ésimo período sazonal com $\omega = 2\pi/p$. Cada componente da soma é denominado harmônico e os termos $A_r$ e $\phi_r$ são denominados amplitude e fase do harmônico de ordem $r$. Embora o número de harmônicos cresça em função do período $p$, na prática apenas um número reduzido de harmônicos possui efeito relevante ($A_r$ pequeno) na construção dos fatores sazonais. Podemos então definir formalmente o modelo linear dinâmico com representação trigonométrica.

**Definição 7.** *Para um ciclo sazonal de período p defina*

$$\boldsymbol{J}(r\omega) = \begin{pmatrix} \cos(\omega r t) & \sin(\omega r t) \\ -\sin(\omega r t) & \cos(\omega r t) \end{pmatrix}. \tag{3.7}$$

*O modelo linear dinâmico com representação trigonométrica considerando os h primeiros harmônicos é definido por:*

1. *Se p é ímpar:*

$$\left\{ \mathbf{1}_h \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix}, diag\{J(\omega), \ldots, J(h\omega)\}, V_t, \boldsymbol{W}_t \right\} \tag{3.8}$$

2. *Se p é par:*

$$\left\{ \begin{pmatrix} \mathbf{1}_h \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ 1 \end{pmatrix}, diag\{J(\omega), \ldots, J((h-1)\omega), -1\}, V_t, \boldsymbol{W}_t \right\} \tag{3.9}$$

Nesta dissertação utilizamos apenas os MLD's com representação trigonométrica. Pode se mostrar que, a partir destes modelos é possível obter os MLD's de forma livre (veja Harrison and West (1999) , Seção 8.6.5).

## 3.2 Filtro de Kalman para Modelos Lineares Dinâmicos

Em geral, a obtenção das distribuições condicionais relevantes não é de todo uma tarefa fácil. Os MLD's são um caso simples, onde as simplificações recursivas gerais são consideravelmente mais simples. Neste caso, usando resultados padrão a cerca das distribuições normais multivariadas, é facilmente provado que o vector aleatório $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t, \mathbf{Y}_1, \ldots, \mathbf{Y}_t)$ tem uma distribuição normal multivariado para qualquer $t \geq 1$. Isto implica que as distribuições marginais e condicionais também são normais. Uma vez que todas as distribuições relevantes são também normais, elas são completamente determinadas por seus vetores de médias e matrizes de variâncias.

A solução do problema de filtragem para MLD é dado pelo famoso filtro de Kalman. Apresentado no teorema a seguir.

**Teorema 2.** ***(Filtro de Kalman)*** *Considere um MLD. Seja*

$$(\boldsymbol{\theta}_{t-1} | \boldsymbol{Y}_{0:t-1}) \sim N[\boldsymbol{m}_{t-1}, \boldsymbol{C}_{t-1}],$$

*então valem as seguintes afirmações:*

1. *A preditiva um passo a frente da distribuição de $\boldsymbol{\theta}_t$ dado $\boldsymbol{Y}_{0:t-1}$ é Normal com*

*parâmetros*

$$\boldsymbol{a}_t = E(\boldsymbol{\theta}_t|\boldsymbol{Y}_{0:t-1}) = \boldsymbol{G}_t m_{t-1} \tag{3.10}$$

$$\boldsymbol{R}_t = Var(\boldsymbol{\theta}_t|\boldsymbol{Y}_{0:t-1}) = \boldsymbol{G}_t \boldsymbol{C}_{t-1} \boldsymbol{G}_t^T + \boldsymbol{W}_t \tag{3.11}$$

2. *A preditiva um passo a frente da distribuição de $\boldsymbol{Y}_t$ dado $\boldsymbol{Y}_{0:t-1}$ é normal com parâmetros*

$$\boldsymbol{f}_t = E(\boldsymbol{Y}_t|\boldsymbol{Y}_{0:t-1}) = \boldsymbol{F}_t \boldsymbol{a}_t, \tag{3.12}$$

$$\boldsymbol{Q}_t = Var(\boldsymbol{Y}_t|\boldsymbol{Y}_{0:t-1}) = \boldsymbol{F}_t^T \boldsymbol{R}_t \boldsymbol{F}_t + \boldsymbol{V}_t \tag{3.13}$$

3. *A distribuição de filtragem de $\boldsymbol{\theta}_t$ dado $\boldsymbol{Y}_{0:t-1}$ é normal com os parâmetros*

$$\boldsymbol{m}_t = E(\boldsymbol{\theta}|\boldsymbol{Y}_{0:t}) = \boldsymbol{a}_t + \boldsymbol{R}_t \boldsymbol{F}_t \boldsymbol{Q}_t^{-1} \boldsymbol{e}_t, \tag{3.14}$$

$$\boldsymbol{C}_t = Var(\boldsymbol{\theta}_t|\boldsymbol{Y}_{0:t}) = \boldsymbol{R}_t - \boldsymbol{R}_t \boldsymbol{F}_t \boldsymbol{Q}_t^{-1} \boldsymbol{F}_t^T \boldsymbol{R}_t \tag{3.15}$$

*onde $\boldsymbol{e}_t = \boldsymbol{Y}_t - \boldsymbol{f}_t$ é o erro de previsão.*

*Demonstração.* Ver demonstração em (Petris et al. (2009),Seção 2.7). □

O filtro de Kalman permite que calculemos a preditiva e o filtro da distribuição recursivamente, iniciando de $\boldsymbol{\theta}_0 \sim N(\boldsymbol{m}_0, \boldsymbol{C}_0)$ onde calculamos $f(\boldsymbol{\theta}_1|y_1)$, e procedendo recursivamente a medida que novos dados estejam disponíveis.

## 3.3 Variâncias Observacionais

A quádrupla $\{\boldsymbol{F}_t, \boldsymbol{G}_t, V_t, \boldsymbol{W}_t\}$, onde $\boldsymbol{F}_{t_{(1\times m)}}, \boldsymbol{G}_{t_{(p\times p)}}, \boldsymbol{W}_{t_{(p\times p)}}$ são vetores e $V_t$ um escalar, caracteriza um MLD univariado. Geralmente a variância observacional $V_t$, que é desconhecida, precisa ser estimada. Como esta variância é geralmente a principal fonte de incerteza no processo estocástico sendo modelado, foram desenvolvidos procedimentos usando o enfoque Bayesiano no caso de ser desconhecida, porém constante, isto é, $V_t = V$ para todo $t$. É mais conveniente trabalhar com a sua inversa, denominada de "*precisão*"e denotada por $\phi$, onde $\phi = 1/V$.

Segue abaixo a definição geral.

**Definição 8.** *Para cada t, o MLD univariado com aprendizagem de variância desconhe-*

*cida observacional é definido por*

$$\begin{aligned}
\textit{Equação de Observação} &: Y_t = \boldsymbol{F}_t^T \boldsymbol{\theta}_t + \nu_t,\ \nu_t \sim N[0,V] \\
\textit{Equação de Sistema} &: \boldsymbol{\theta}_t = \boldsymbol{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t,\ \boldsymbol{\omega}_t \sim N[\boldsymbol{0}, V\boldsymbol{W}_t^*] \\
\textit{Priori Inicial} &: (\boldsymbol{\theta}_0|Y_0,\phi) \sim N[\boldsymbol{m}_0, V\boldsymbol{C}_0^*] \\
&: (\phi|Y_0) \sim \textit{Gama}\left[\frac{n_0}{2}, \frac{n_0 S_0}{2}\right],
\end{aligned}$$

*onde assume-se que as sequências de erros observacionais $\nu_t$ e de evolução $\boldsymbol{\omega}_t$ são independentes ao longo do tempo e entre si, e independentes da priori $(\boldsymbol{\theta}_0|Y_0,\phi)$.*

O teorema abaixo mostra as equações de evolução. Note que as distribuições de previsão agora são *t-Student* ao invés de normais.

**Teorema 3.** *O MLD definido acima possui as seguintes distribuições condicionais*

*(a) Condicionado a V:*

$$\begin{aligned}
(\boldsymbol{\theta}_{t-1}|Y_{0:t-1},V) &\sim N[\boldsymbol{m}_{t-1}, V\boldsymbol{C}_{t-1}^*] \\
(\boldsymbol{\theta}_t|Y_{0:t-1},V) &\sim N[\boldsymbol{a}_t, V\boldsymbol{R}_t^*] \\
(Y_t|Y_{0:t-1},V) &\sim N[f_t, VQ_t^*] \\
(\boldsymbol{\theta}_t|Y_{0:t},V) &\sim N[\boldsymbol{m}_t, V\boldsymbol{C}_t^*],
\end{aligned}$$

*com*

$$\begin{aligned}
\boldsymbol{a}_t &= \boldsymbol{G}_t \boldsymbol{m}_{t-1}, & \boldsymbol{R}_t^* &= \boldsymbol{G}_t \boldsymbol{C}_{t-1}^* \boldsymbol{G}_t^T + \boldsymbol{W}_t^* \\
f_t &= \boldsymbol{F}_t^T \boldsymbol{a}_t, & Q_t^* &= 1 + \boldsymbol{F}_t^T \boldsymbol{R}_t^* \boldsymbol{F}_t \\
e_t &= Y_t - f_t, & \boldsymbol{A}_t &= \boldsymbol{R}_t^* \boldsymbol{F}_t / Q_t^* \\
\boldsymbol{m}_t &= \boldsymbol{a}_t + \boldsymbol{A}_t e_t, & \boldsymbol{C}_t^* &= \boldsymbol{R}_t^* - \boldsymbol{A}_t \boldsymbol{A}_t^T Q_t^*.
\end{aligned}$$

*(b) Para a precisão $\phi = 1/V$, temos:*

$$\begin{aligned}
(\phi|Y_{0:t-1}) &\sim \textit{Gama}\left[\frac{n_{t-1}}{2}, \frac{n_{t-1}S_{t-1}}{2}\right], \\
(\phi|Y_{0:t}) &\sim \textit{Gama}\left[\frac{n_t}{2}, \frac{n_t S_t}{2}\right],
\end{aligned}$$

*onde $n_t = n_{t-1} + 1$ e $S_t = S_{t-1} + \frac{S_{t-1}}{n_t}\left(\frac{e_t^2}{Q_t} - 1\right)$,*

*(c) Marginalizando as distribuições em relação a V, temos:*

$$\begin{aligned}(\boldsymbol{\theta}_{t-1}|Y_{0:t-1}) &\sim T_{n_{t-1}}[\boldsymbol{m}_{t-1}, \boldsymbol{C}_{t-1}],\\ (\boldsymbol{\theta}_t|Y_{0:t-1}) &\sim T_{n_{t-1}}[\boldsymbol{a}_t, \boldsymbol{R}_t],\\ (Y_t|Y_{0:t-1}) &\sim T_{n_{t-1}}[f_t, Q_t],\\ (\boldsymbol{\theta}_t|Y_{0:t}) &\sim T_{n_t}[\boldsymbol{m}_t, \boldsymbol{C}_t],\end{aligned}$$

*onde* $\boldsymbol{R}_t = S_{t-1}\boldsymbol{R}_t^*$, $Q_t = S_{t-1}Q_t^*$ *e* $\boldsymbol{C}_t = S_t\boldsymbol{C}_t^*$, *e* $T_{n_t}$ *denota a distribuição t multivariada com* $n_t$ *graus de liberdade.*

*(d) As equações de atualização são dadas abaixo, onde* $Q_t = \boldsymbol{F}_t^T\boldsymbol{R}_t\boldsymbol{F}_t + S_{t-1}$ *e* $\boldsymbol{A}_t = \boldsymbol{R}_t\boldsymbol{F}_t/Q_t$*:*

$$\boldsymbol{m}_t = \boldsymbol{a}_t + \boldsymbol{A}_t e_t,\ \boldsymbol{C}_t = \frac{S_t}{S_{t-1}}(\boldsymbol{R}_t - \boldsymbol{A}_t\boldsymbol{A}_t^T Q_t). \tag{3.16}$$

A demonstração usa conceitos da teoria da distribuição normal-gama e será omitida por usar resultados padrões (ver Harrison and West (1999)).

## 3.4 Variância da Evolução

Discutiremos nesta seção a obtenção de $\boldsymbol{W}_t$ via método de elicitação utilizando fatores de desconto e estimação via método de Bayes empírico.

Para ilustrar a ideia de fator de descontos, considere o modelo polinomial de ordem 1, onde

$$\begin{aligned}Var(\theta_t|\theta_{t-1}) &= W_t\\ Var(\theta_t|Y_{0:t-1}) &= W_t + C_{t-1}\end{aligned}$$

Note que a segunda variância pode ser interpretada como sendo a variância de $\theta_t$ quando retiramos a informação $\theta_{t-1}$. Isto gera um aumento na incerteza igual a $\boldsymbol{C}_{t-1}$. Suponha que desejamos fixar esse aumento em, por exemplo, 10%. Então,

$$1,1 = \frac{Var(\theta_t|Y_{0:t-1})}{Var(\theta_t|\theta_{t-1})} = \frac{W_t + C_{t-1}}{W_t} = 1 + \frac{C_{t-1}}{W_t}$$

o que implica em

$$W_t = \frac{1}{0,1} \times C_{t-1}.$$

Em termos gerais, expressando esse aumento em termos que $\delta \in (0,1]$, teremos

$$\frac{1}{1-\delta} = \frac{Var(\boldsymbol{\theta}_t|Y_{0:t-1})}{Var(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})} \Rightarrow W_t = \frac{1-\delta}{\delta}\boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t^T,$$

onde $\delta$ é denominado fator de desconto. Para um modelo linear dinâmico qualquer, a estratégia de descontos estabelece que

$$\boldsymbol{W}_t = \frac{1-\delta}{\delta}\boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t^T.$$

Em geral, escolhemos valores para $\delta$ dentro do intervalo $(0,8\ ,0,99)$, que implicam em pouco aumento na variância. Note que $\boldsymbol{W}_t$ não é estimada, mas sim obtida através de alguma informação dada pelo usuário (por isso, optamos utilizar o termo elicitação no lugar de estimação).

Como alternativa aos fatores de desconto, Petris et al. (2009) propõe considerar um modelo linear dinâmico com $\boldsymbol{W}_t = \boldsymbol{W}$ para todo $t \geq 1$ e estimar $\boldsymbol{W}$ através de $\hat{\boldsymbol{W}}$, onde

$$\widehat{\boldsymbol{W}} = \text{argsup} f\left(y_{0:t}|\boldsymbol{W}\right) \approx \text{argsup} \prod_{j=1}^{t} f\left(y_j|y_{0:j-1}, \boldsymbol{W}\right).$$

Como o estimador para $\boldsymbol{W}$ é obtido via maximização da distribuição preditiva, temos que este estimador foi obtido via método de Bayes empírico. Usualmente representamos $\boldsymbol{W}$ por uma matriz diagonal.

# Capítulo 4

# Classificador de Bayes para Séries Temporais Utilizando Modelos Lineares Dinâmicos

Neste capitulo, apresentamos os desenvolvimentos necessários para a determinação do classificador de Bayes com base em MLD (CBMLD), que consiste em nossa contribuição para os problemas em Análise Discriminante para séries temporais.

## 4.1 Filtro de Kalman Para Múltiplas Séries Provenientes de uma Classe

Vamos considerar que um determinado conjunto de dados associado a um problema de classificação de séries temporais. Cada observação é proveniente de uma dentre $M$ classes conhecidas. Usaremos a variável aleatória indicadora $Z$ e a série temporal $\mathbf{X}_{1:t}$, já citadas anteriormente, tal que $(\mathbf{X}_{0:t} = \mathbf{x}_{0:t}^{(j)} | Z = i) = \mathbf{x}_{0:t}^{(j,i)}, \quad i \in \{1,2,3,\ldots,M\}$ indica que a j-ésima série observada é da classe $i$ (veja o Capítulo 2, Seção 2.2). Para representar todas as séries temporais de classe $i$, no conjunto de treinamento, usaremos a notação $\mathbf{X}_{0:t}^{(i)}$. Observe que quando temos o conhecimento da classe da série temporal em questão, acrescentamos essa informação com o índice 0. A quantidade de séries observadas numa mesma classe $i \in \{1,2,3,\ldots,M\}$ será denotado como $l^{(i)} \leqslant n$ onde $n$ é a quantidade total de séries observadas no banco de dados.

É razoável supormos que para toda série temporal proveniente de uma mesma classe observamos o mesmo modelo de probabilidade, ou seja, $\mathbf{X}_{0:t}^{(i)} \sim \mathbf{M}_i$. Sendo assim, suponhamos que $\mathbf{X}_{0:t}^{(i)}$ é bem representada pelo MLD.

$$\left\{ \boldsymbol{F}_t^{(i)}, \boldsymbol{G}_t^{(i)}, \boldsymbol{V}_t^{(i)}, \boldsymbol{W}_t^{(i)} \right\} \equiv \{\boldsymbol{F}_t, \boldsymbol{G}_t, \boldsymbol{V}_t, \boldsymbol{W}_t\}^{(i)}.$$

Ou seja,

$$
\begin{aligned}
(\mathbf{X}_t^{(i)}|\boldsymbol{\theta}_t^{(i)}, \mathbf{X}_{0:t-1}^{(i)}) &\sim \mathrm{N}\left[(\mathbf{1}_{l^{(i)}} \otimes \mathbb{F}_t^{(i)T})\boldsymbol{\theta}_t^{(i)}, \boldsymbol{I}_{l^{(i)}} \otimes \boldsymbol{V}_t^{(i)}\right], \\
(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}, \mathbf{X}_{0:t-1}^{(i)}) &\sim \mathrm{N}\left[(\boldsymbol{G}_t^{(i)}\boldsymbol{\theta}_{t-1}^{(i)}, \boldsymbol{W}_t^{(i)}\right], \\
(\boldsymbol{\theta}_0^{(i)}|\mathbf{X}_0^{(i)}) &\sim \mathrm{N}\left[(\boldsymbol{m}_0^{(i)}, \boldsymbol{C}_0^{(i)}\right].
\end{aligned}
$$

Com todas estas suposições feitas é possível construir um filtro de Kalman, como mostra o teorema a seguir. Nas demonstrações desenvolvidas empregamos alguns lemas que são apresentados no Apêndice desta dissertação.

**Teorema 4.** *Para o i-ésimo modelo, e para cada $t \geq 1$, valem as seguintes afirmações:*

1. *Priori no tempo $t$:*

$$(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{X}_{0:t-1}) \sim N\left[\boldsymbol{a}_t^{(i)}, \boldsymbol{R}_t^{(i)}\right],$$

   *onde*

$$
\begin{aligned}
\boldsymbol{a}_t^{(i)} &= \boldsymbol{G}_t^{(i)}\boldsymbol{m}_{t-1}^{(i)} \\
\boldsymbol{R}_t^{(i)} &= \boldsymbol{W}_t^{(i)} + \boldsymbol{G}_t^{(i)}\boldsymbol{C}_{t-1}^{(i)}\boldsymbol{G}_t^{(i)T}
\end{aligned}
$$

2. *Previsão para o tempo $t$:*

$$(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{X}_{0:t-1}) \sim N\left[\boldsymbol{f}_t^{(i)}, \boldsymbol{Q}_t^{(i)}\right],$$

   *onde*

$$
\begin{aligned}
\boldsymbol{f}_t^{(i)} &= (\mathbf{1}_{l^{(i)}} \otimes \mathbb{F}_t^T)\boldsymbol{a}_t^{(i)} \\
\boldsymbol{Q}_t^{(i)} &= \boldsymbol{V}^{(i)} + (\mathbf{1}_{l^{(i)}} \otimes \mathbb{F}_t^T)\boldsymbol{R}_t^{(i)}(\mathbf{1}_{l^{(i)}} \otimes \mathbb{F}_t^T)^T
\end{aligned}
$$

3. *Posteriori para o tempo $t$:*

$$(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{X}_{0:t}) \sim N\left[\boldsymbol{m}_t^{(i)}, \boldsymbol{C}_t^{(i)}\right],$$

   *onde*

$$
\begin{aligned}
\boldsymbol{A}_t^{(i)} &= \boldsymbol{R}_t^{(i)}(\mathbf{1}_{l^{(i)}} \otimes \boldsymbol{F}_t^T)^T\boldsymbol{Q}_t^{(i)-1} \\
\boldsymbol{m}_t^{(i)} &= \boldsymbol{a}_t^{(i)} + \boldsymbol{A}_t^{(i)}\left[\boldsymbol{X}_t^{(i)} - \boldsymbol{f}_t^{(i)}\right] \\
\boldsymbol{C}_t^{(i)} &= \boldsymbol{R}_t^{(i)} - \boldsymbol{A}_t^{(i)}\boldsymbol{Q}_t^{(i)-1}\boldsymbol{A}_t^{(i)T}
\end{aligned}
$$

*Demonstração.* O item 1 do Teorema é verdadeiro para para $t = 1$, pois trata-se da informação inicial do modelo linear dinâmico. Vamos demonstrar o Teorema por indução. Suponha que

$$(\boldsymbol{\theta}_{t-1}^{(i)}|\mathbf{X}_{0:t-1}^{(i)}) \sim \mathrm{N}\left[\boldsymbol{m}_{t-1}^{(i)}, \boldsymbol{C}_{t-1}^{(i)}\right].$$

Então:

- Como $(\boldsymbol{\theta}_{t-1}|\mathbf{X}_{0:t-1}) \sim \mathrm{N}[\boldsymbol{m}_{t-1}, \boldsymbol{C}_{t-1}]$ e $\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \mathbf{X}_{0:t-1} \sim \mathrm{N}[\boldsymbol{G}_t\boldsymbol{\theta}_{t-1}, \boldsymbol{W}_t]$, pelo Lema (3),

$$\begin{pmatrix} \boldsymbol{\theta}_t \\ \boldsymbol{\theta}_{t-1} \end{pmatrix} \Bigg| \mathbf{X}_{0:t-1} \sim \mathrm{N}\left[\begin{pmatrix} \boldsymbol{G}_t\boldsymbol{m}_{t-1} \\ \boldsymbol{m}_{t-1} \end{pmatrix}, \begin{pmatrix} \boldsymbol{W}_t + \boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t^T & \boldsymbol{G}_t\boldsymbol{C}_{t-1} \\ \boldsymbol{C}_{t-1}\boldsymbol{G}_t^T & \boldsymbol{C}_{t-1} \end{pmatrix}\right]$$

  e o resultado do item 2 é imediato.

- Como $(\mathbf{X}_t|\boldsymbol{\theta}_t, \mathbf{X}_{0:t-1}) \sim \mathrm{N}\left[(\mathbf{1}_{l^{(i)}} \otimes \mathbb{F}_t^T)\boldsymbol{\theta}_t, \boldsymbol{I}_{l^{(i)}} \otimes \boldsymbol{V}_t\right]$ e $(\boldsymbol{\theta}_t|\mathbf{X}_{0:t-1}) \sim \mathrm{N}[\boldsymbol{a}_t, \boldsymbol{R}_t]$, pelo Lema (3)

$$\begin{pmatrix} \mathbf{X}_t \\ \boldsymbol{\theta}_t \end{pmatrix} \Bigg| \mathbf{X}_{0:t-1} \sim \mathrm{N}\left[\begin{pmatrix} \mathbb{F}_t^T\boldsymbol{a}_t \\ \boldsymbol{a}_t \end{pmatrix}, \begin{pmatrix} \boldsymbol{I}_{l^{(i)}} \otimes \boldsymbol{V}_t + \mathbb{F}_t^T\boldsymbol{R}_t\mathbb{F}_t & \mathbb{F}_t^T\boldsymbol{R}_t \\ \boldsymbol{R}_t\mathbb{F}_t & \boldsymbol{R}_t \end{pmatrix}\right]$$

  o resultado do item 3 é imediato.

- Utilizando a conjunta do item anterior e o Lema (2), é imediato que

$$(\boldsymbol{\theta}_t|\mathbf{X}_{0:t}) \sim N[\boldsymbol{m}_t, \boldsymbol{C}_t],$$

  onde

$$\begin{aligned} \boldsymbol{A}_t &= \boldsymbol{R}_t\mathbb{F}_t\boldsymbol{Q}_t^{-1} \\ \boldsymbol{m}_t &= \boldsymbol{a}_t + \boldsymbol{A}_t[\mathbf{X}_t - \boldsymbol{f}_t] \\ \boldsymbol{C}_t &= \boldsymbol{R}_t - \boldsymbol{A}_t\boldsymbol{Q}_t\boldsymbol{A}_t^T. \end{aligned}$$

□

Portanto, o problema de lidar com múltiplas séries de uma mesma classe através de um modelo dinâmico é relativamente simples, uma vez que todas as séries mantém a mesma estrutura de evolução.

## 4.2 O Classificador de Bayes utilizando Modelos Lineares Dinâmicos (CBMLD)

Consideremos uma nova série temporal $Y_{1:t}$ (uma série não classificada), vamos tomar uma amostra de treino $\mathbf{X}_{0:t} = \{\mathbf{X}_{0:t}^{(1)}, \mathbf{X}_{0:t}^{(2)}, \ldots, \mathbf{X}_{0:t}^{(M)}\}$ grande o suficiente. Então para todo $i \in \{1, \ldots, M\}$, $(Y_{1:t}|Z = i) = Y_{1:t}^{(i)}$ nos diz que $Y_{1:t} \sim \mathbf{M}_i$. Neste momento usaremos o classificador de Bayes que vai utilizar MLD para encontrar a classe $i$ que melhor classifica a série temporal $Y_{1:t}$ de classe desconhecida, ou seja, vamos classificar.

$$r_{(CBMLD)}(\mathbf{x}) = i \quad se \quad P(Z = i|y_{1:t}, \mathbf{x}_{0:t}) = max_j \quad P(Z = j|y_{1:t}, \mathbf{x}_{0:t}) \tag{4.1}$$

onde $P(Z = i|\mathbf{y}_{0:t}, \mathbf{x}_{0:t})$ representa a probabilidade de classificar a nova série, após observada, como pertencente à classe $i$, conhecendo todas as classificações das séries de treino.

Por sua vez, temos que

$$\begin{aligned} P(Z = i|y_{1:t}, \mathbf{x}_{0:t}) &= \frac{f(y_{1:t}|\mathbf{x}_{0:t}, Z = i)P(Z = i|\mathbf{x}_{0:t})}{f(y_{1:t}|\mathbf{x}_{0:t})} \\ &\propto f(y_{1:t}^{(i)}|\mathbf{x}_{0:t}, Z = i)P(Z = i|\mathbf{x}_{0:t}), \end{aligned}$$

e, supondo que $Y_{1:t}^{(i)}$ condicionado com $\mathbf{X}_{0:t}^{(i)}$ é independente de qualquer $\mathbf{X}_{0:t}^{(j)}$ com $j \neq i$, teremos

$$\begin{aligned} P(Z = i|y_{1:t}, \mathbf{x}_{0:t}^{(j)}) &\propto f(y_{1:t}^{(i)}|\mathbf{x}_{0:t}, Z = i)P(Z = i|\mathbf{x}_{0:t}) \\ &= f(y_{1:t}^{(i)}|\mathbf{x}_{0:t}^{(i)}, \mathbf{x}_{0:t}^{(j)}, Z = i)P(Z = i|\mathbf{x}_{0:t}), i \neq j \\ &= \int f(y_{1:t}^{(i)}, \boldsymbol{\theta}_{1:t}^{(i)}|\mathbf{x}_{0:t}^{(i)})d\boldsymbol{\theta}_{1:t}^{(i)}P(Z = i|\mathbf{x}_{0:t}), \quad pelo \quad Lema \quad (4) \\ &\propto \int f(y_{1:t}^{(i)}|\boldsymbol{\theta}_{1:t}^{(i)}, \mathbf{x}_{0:t}^{(i)})f(\boldsymbol{\theta}_{1:t}^{(i)}|\mathbf{x}_{0:t}^{(i)})d\boldsymbol{\theta}_{1:t}^{(i)}P(Z = i|\mathbf{x}_{0:t}) \\ &= \int f(y_{1:t}^{(i)}|\boldsymbol{\theta}_{1:t}^{(i)})f(\boldsymbol{\theta}_{1:t}^{(i)}|\mathbf{x}_{0:t}^{(i)})d\boldsymbol{\theta}_{1:t}^{(i)}P(Z = i|\mathbf{x}_{0:t}) \\ &= \int \prod_{j=1}^{t} f\left(y_j^{(i)}|\boldsymbol{\theta}_j^{(i)}\right) \times f\left(\boldsymbol{\theta}_{1:t}^{(i)}|\mathbf{x}_{0:t}^{(i)}\right) d\boldsymbol{\theta}_{1:t}^{(i)}P(Z = i|\mathbf{x}_{0:t}). \end{aligned}$$

Logo, temos um classificador baseado em MLD, bastando para isso resolver a integral:

$$\int \prod_{j=1}^{t} f\left(y_j^{(i)}|\boldsymbol{\theta}_j^{(i)}\right) \times f\left(\boldsymbol{\theta}_{1:t}^{(i)}|\mathbf{x}_{0:t}^{(i)}\right) d\boldsymbol{\theta}_{1:t}^{(i)}P(Z = i|\mathbf{x}_{0:t})$$

Vamos discutir como obter cada uma das distribuições envolvidas nesta integral e mostrar que ela tem solução analítica. Como já temos o conhecimento a respeito da distribuição

$(y_j^{(i)}|\boldsymbol{\theta}_j^{(i)})$ pois,

$$\prod_{j=1}^{t} f\left(y_j^{(i)}|\boldsymbol{\theta}_j^{(i)}\right) = \prod_{j=1}^{t}\left(\frac{1}{2\pi V_j^{(i)}}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(y_j^{(i)} - \boldsymbol{F}_j^{(i)T}\boldsymbol{\theta}_j^{(i)})^T[V_j^{(i)}]^{-1}(y_j^{(i)} - \boldsymbol{F}_j^{(i)T}\boldsymbol{\theta}_j^{(i)})\right\} \tag{4.2}$$

$$= \left(\frac{1}{2\pi}\right)^{\frac{t}{2}} \frac{1}{\prod_{j=1}^{t}\sqrt{V_j^{(i)}}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{t}(y_j^{(i)} - \boldsymbol{F}_j^{(i)T}\boldsymbol{\theta}_j^{(i)})^T[V_j^{(i)}]^{-1}(y_j^{(i)} - \boldsymbol{F}_j^{(i)T}\boldsymbol{\theta}_j^{(i)})\right\} \tag{4.3}$$

Vamos utilizar a comutatividade no produto em (4.2) para inverter a ordem dos índices de forma que, $\prod_{j=1}^{t} f\left(y_j^{(i)}|\boldsymbol{\theta}_j^{(i)}\right) = \prod_{j=t}^{1} f\left(y_j^{(i)}|\boldsymbol{\theta}_j^{(i)}\right)$, e assim podemos expor a distribuição como segue:

$$\left(y_{t:1}^{(i)}|\boldsymbol{\theta}_{t:1}^{(i)}\right) \sim N\left[\begin{pmatrix} \boldsymbol{F}_t^{(i)T}\boldsymbol{\theta}_t^{(i)} \\ \boldsymbol{F}_{t-1}^{(i)T}\boldsymbol{\theta}_{t-1}^{(i)} \\ \vdots \\ \boldsymbol{F}_1^{(i)T}\boldsymbol{\theta}_1^{(i)} \end{pmatrix}, \begin{pmatrix} V_t^{(i)} & 0 & \dots & 0 \\ 0 & V_{t-1}^{(i)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & V_1^{(i)} \end{pmatrix}\right].$$

Encontraremos agora a distribuição suavizada da conjunta $f\left(\boldsymbol{\theta}_{1:t}^{(i)}|\mathbf{x}_{0:t}^{(i)}\right)$. Para todo $t \geq 1$ definimos a matriz $\boldsymbol{E}_{t-1} = (\boldsymbol{I}_p\ \mathbf{0}_{p\times p(t-1)})$ onde $p$ é a dimensão da matriz $\boldsymbol{G}_t^{(i)}$ .

**Teorema 5.** *Para $t \geq 1$ e para a i-ésima classe, verificam-se as seguintes distribuições:*

1. *Posteriori no tempo $t-1$*

$$(\boldsymbol{\theta}_{0:t-1}^{(i)}|\boldsymbol{X}_{0:t-1}^{(i)}) \sim N\left[\mathscr{M}_{t-1}^{(i)}, \mathscr{C}_{t-1}^{(i)}\right]$$

2. *Priori no tempo t*

$$(\boldsymbol{\theta}_{0:t}^{(i)}|\boldsymbol{X}_{0:t-1}^{(i)}) \sim N\left[\mathscr{A}_t^{(i)}, \mathscr{R}_t^{(i)}\right],$$

*onde*

$$\mathscr{A}_t^{(i)} = \begin{pmatrix} \boldsymbol{G}_t\boldsymbol{E}_{t-1}\mathscr{M}_{t-1}^{(i)} \\ \mathscr{M}_{t-1} \end{pmatrix}$$

$$\mathscr{R}_t^{(i)} = \begin{pmatrix} \boldsymbol{W}_t^{(i)} + \boldsymbol{G}_t^{(i)}\boldsymbol{E}_{t-1}\mathscr{C}_{t-1}^{(i)}\boldsymbol{E}_{t-1}^T\boldsymbol{G}_t^{(i)T} & \boldsymbol{G}_t^{(i)}\boldsymbol{E}_{t-1}\mathscr{C}_{t-1}^{(i)} \\ \mathscr{C}_{t-1}^{(i)}\boldsymbol{E}_{t-1}^T\boldsymbol{G}_t^{(i)T} & \mathscr{C}_{t-1}^{(i)} \end{pmatrix}$$

3. *Com*

$$\begin{pmatrix} \boldsymbol{X}_t^{(i)} \\ \boldsymbol{\theta}_{0:t}^{(i)} \end{pmatrix} \middle| \boldsymbol{X}_{0:t-1}^{(i)} \sim N\left[\begin{pmatrix} \mathscr{F}_t^{(i)} \\ \mathscr{A}_t^{(i)} \end{pmatrix}, \begin{pmatrix} \mathscr{Q}_t^{(i)} & \mathbb{F}_t^{(i)T}\boldsymbol{E}_t\mathscr{R}_t^{(i)} \\ \mathscr{R}_t^{(i)}\boldsymbol{E}_t^T\mathbb{F}_t^{(i)} & \mathscr{R}_t^{(i)} \end{pmatrix}\right]$$

*onde*

$$\mathscr{F}_t^{(i)} = \mathbb{F}_t^{(i)T} \boldsymbol{E}_t \mathscr{A}_t^{(i)} \tag{4.4}$$

$$\mathscr{Q}_t^{(i)} = V_t^{(i)} + \mathbb{F}_t^{(i)T} \boldsymbol{E}_t \mathscr{R}_t^{(i)} \boldsymbol{E}_t^T \mathbb{F}_t^{(i)} \tag{4.5}$$

*tem-se que a posteriori da conjunta no tempo t é*

$$(\boldsymbol{\theta}_{0:t}^{(i)} | \boldsymbol{X}_{0:t}^{(i)}) \sim N\left[\mathscr{M}_t^{(i)}, \mathscr{C}_t^{(i)}\right],$$

*onde*

$$\mathscr{M}_t^{(i)} = \mathscr{A}_t^{(i)} + \mathscr{R}_t^{(i)} \boldsymbol{E}_t^T \mathbb{F}_t^{(i)} \mathscr{Q}_t^{(i)-1} \left(\boldsymbol{x}_t^{(i)} - \mathscr{F}_t^{(i)}\right)$$
$$\mathscr{C}_t^{(i)} = \mathscr{R}_t^{(i)} - \mathscr{R}_t^{(i)} \boldsymbol{E}_t \mathbb{F}_t^{(i)} \mathscr{Q}_t^{(i)-1} \mathbb{F}_t^{(i)T} \boldsymbol{E}_t \mathscr{R}_t^{(i)}$$

*Demonstração.* Demonstraremos novamente por indução fazendo

$$(\boldsymbol{\theta}_{0:t-1}^{(i)} | \mathbf{X}_{0:t-1}^{(i)}) \sim \mathrm{N}\left[\mathscr{M}_{t-1}^{(i)}, \mathscr{C}_{t-1}^{(i)}\right] \text{ hipótese de Indução}$$

Vamos usar o Lema (3) para

$$\left(\boldsymbol{\theta}_t^{(i)} | \mathbf{X}_{0:t-1}\right) \sim N\left[\mathscr{M}_{t-1}^{(i)}, \mathscr{C}_{t-1}^{(i)}\right] \tag{4.6}$$

$$\left(\boldsymbol{\theta}_t^{(i)} | \boldsymbol{\theta}_{0:t-1}^{(i)}\right) \sim \mathrm{N}\left[\boldsymbol{G}_t^{(i)} \boldsymbol{E}_{t-1} \boldsymbol{\theta}_{t-1:0}^{(i)}, \boldsymbol{W}_t^{(i)}\right] \tag{4.7}$$

encontrarmos,

$$\begin{pmatrix} \boldsymbol{\theta}_t^{(i)} \\ \boldsymbol{\theta}_{t-1:0}^{(i)} \end{pmatrix} \middle| \mathbf{X}_{0:t-1}^{(i)} \sim \mathrm{N}\left[\begin{pmatrix} \boldsymbol{G}_t^{(i)} \boldsymbol{E}_{t-1} \mathscr{M}_{t-1}^{(i)} \\ \mathscr{M}_{t-1}^{(i)} \end{pmatrix}, \begin{pmatrix} \boldsymbol{W}_t^{(i)} + \boldsymbol{G}_t^{(i)} \boldsymbol{E}_{t-1} \mathscr{C}_{t-1}^{(i)} \boldsymbol{E}_{t-1}^T \boldsymbol{G}_t^{(i)T} & \boldsymbol{G}_t^{(i)} \boldsymbol{E}_{t-1} \mathscr{C}_{t-1}^{(i)} \\ \mathscr{C}_{t-1}^{(i)T} \boldsymbol{E}_{t-1}^T \boldsymbol{G}_t^{(i)T} & \mathscr{C}_{t-1}^{(i)} \end{pmatrix}\right],$$

onde podemos chamar

$$\mathscr{A}_t^{(i)} = \begin{pmatrix} \boldsymbol{G}_t^{(i)} \boldsymbol{E}_{t-1} \mathscr{M}_{t-1}^{(i)} \\ \mathscr{M}_{t-1}^{(i)} \end{pmatrix} \tag{4.8}$$

$$\mathscr{R}_t^{(i)} = \begin{pmatrix} \boldsymbol{W}_t^{(i)} + \boldsymbol{G}_t^{(i)} \boldsymbol{E}_{t-1} \mathscr{C}_{t-1}^{(i)} \boldsymbol{E}_{t-1}^T \boldsymbol{G}_t^{(i)T} & \boldsymbol{G}_t^{(i)} \boldsymbol{E}_{t-1} \mathscr{C}_{t-1}^{(i)} \\ \mathscr{C}_{t-1}^{(i)T} \boldsymbol{E}_{t-1}^T \boldsymbol{G}_t^{(i)T} & \mathscr{C}_{t-1}^{(i)} \end{pmatrix} \tag{4.9}$$

logo,

$$(\boldsymbol{\theta}_{t:0}^{(i)} | \mathbf{X}_{0:t-1}^{(i)}) \sim \mathrm{N}(\mathscr{A}_t^{(i)}, \mathscr{R}_t^{(i)}) \tag{4.10}$$

e com equação de evolução $(\mathbf{X}_t | \boldsymbol{\theta}_{t:0}) \sim \mathrm{N}(\mathbb{F}_t^{(i)T} \boldsymbol{E}_t \boldsymbol{\theta}_{t:0}^{(i)}, V_t^{(i)})$ podemos usar o Lema (3) e

encontrar

$$\left( \begin{array}{c} \mathbf{X}_t^{(i)} \\ \boldsymbol{\theta}_{t:0}^{(i)} \end{array} \middle| \mathbf{X}_{0:t-1}^{(i)} \right) \sim \mathrm{N}\left[ \left( \begin{array}{c} \mathbb{F}_t^{(i)T} \mathscr{A}_t^{(i)} \\ \mathscr{A}_t^{(i)} \end{array} \right), \left( \begin{array}{cc} V^{(i)} + \mathbb{F}_t^{(i)} \mathscr{R}^{(i)} \mathbb{F}_t^{(i)} & \mathbb{F}_t^{(i)T} \mathscr{R}_{(i)} \\ \mathscr{R}^{(i)T} \mathbb{F}_t^{(i)} & \mathscr{R}^{(i)} \end{array} \right) \right]$$

Usando o Lema (2) para

$$(\boldsymbol{\theta}_{t:0} | \mathbf{X}_t, \mathbf{X}_{0:t-1}) = (\boldsymbol{\theta}_{t:0}^{(i)} | \mathbf{X}_{0:t})$$

Vamos obter uma distribuição Normal com parâmetros

$$\left( \mathscr{A}_t^{(i)} + \mathscr{R}_t^{(i)} \boldsymbol{E}_t^T \mathbb{F}_t^{(i)} \mathscr{Q}_t^{(i)-1} \left( \mathbf{x}_t^{(i)} - \mathscr{F}_t^{(i)} \right), \mathscr{R}_t^{(i)} - \mathscr{R}_t^{(i)} \boldsymbol{E}_t^T \mathbb{F}_t^{(i)} \mathscr{Q}_t^{(i)-1} \mathbb{F}_t^{(i)T} \boldsymbol{E}_t \mathscr{R}_t^{(i)} \right)$$

onde podemos chamar

$$\mathscr{M}_t^{(i)} = \mathscr{A}_t^{(i)} + \mathscr{R}_t^{(i)} \boldsymbol{E}_t^T \mathbb{F}_t^{(i)} \mathscr{Q}_t^{(i)-1} \left( \mathbf{x}_t^{(i)} - \mathscr{F}_t^{(i)} \right) \tag{4.11}$$

$$\mathscr{C}_t^{(i)} = \mathscr{R}_t^{(i)} - \mathscr{R}_t^{(i)} \boldsymbol{E}_t^T \mathbb{F}_t^{(i)} \mathscr{Q}_t^{(i)-1} \mathbb{F}_t^{(i)T} \boldsymbol{E}_t \mathscr{R}_t^{(i)} \tag{4.12}$$

□

O Teorema (5) apresenta a distribuição conjunta de todos os estados condicionadas à série temporal observada até o tempo $t$.

$$\begin{pmatrix} \boldsymbol{F}_t^{(i)T} \boldsymbol{\theta}_t^{(i)} \\ \boldsymbol{F}_{t-1}^{(i)T} \boldsymbol{\theta}_{t-1}^{(i)} \\ \vdots \\ \boldsymbol{F}_1^{(i)T} \boldsymbol{\theta}_1^{(i)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{F}_t^{(i)T} & 0 & 0 & \dots & 0 \\ 0 & \boldsymbol{F}_{t-1}^{(i)T} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & 0 & \dots & \boldsymbol{F}_1^{(i)T} \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta}_t^{(i)} \\ \boldsymbol{\theta}_{t-1}^{(i)} \\ \vdots \\ \boldsymbol{\theta}_1^{(i)} \end{pmatrix} = \mathrm{bdiag}(\boldsymbol{F}_{t:1}^{(i)T}) \boldsymbol{\theta}_{t:1}^{(i)} \tag{4.13}$$

onde $\mathrm{bdiag}(\boldsymbol{F}_{t:1}^{(i)T})$ é a matriz bloco diagonal e

$$\begin{pmatrix} V_t^{(i)} & 0 & \dots & 0 \\ 0 & V_{t-1}^{(i)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & V_1^{(i)} \end{pmatrix} = \mathrm{diag}(V_{t:1}^{(i)}). \tag{4.14}$$

Logo,

$$(Y_{t:1}^{(i)} | \boldsymbol{\theta}_{t:1}^{(i)}) \sim N\left( \mathrm{bdiag}(\boldsymbol{F}_{t:1}^{(i)T}) \boldsymbol{\theta}_{t:1}^{(i)}, \mathrm{diag}(V_{t:1}^{(i)}) \right) \tag{4.15}$$

Note que podemos decompor o $(\boldsymbol{\theta}_{t:1}^{(i)} | \mathbf{X}_{0:t}^{(i)})$ como

$$\left( \begin{array}{c} \boldsymbol{\theta}_{t:1}^{(i)} \\ \boldsymbol{\theta}_0^{(i)} \end{array} \middle| \mathbf{X}_{t:0}^{(i)} \right) \sim \mathrm{N}\left[ \left( \begin{array}{c} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{array} \right), \left( \begin{array}{cc} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{array} \right) \right]$$

Portanto, $\boldsymbol{\theta}_{t:1}^{(i)} | \mathbf{X}_{t:0}^{(i)} \sim N[\boldsymbol{\alpha}, \boldsymbol{A}]$, sendo necessário especificar o vetor de médias $\boldsymbol{\alpha}$ e a matriz

de covariância $\boldsymbol{A}$. Para encontrar os parâmetros desta distribuição, considere a matriz

$$\boldsymbol{D}_t = \begin{pmatrix} \mathbf{I}_{tp} & \mathbf{0}_{tp\times p} \end{pmatrix}$$

tal que $\boldsymbol{\theta}_{t:1}^{(i)} = \boldsymbol{D}_t \boldsymbol{\theta}_{t:0}^{(i)}$ assim podemos calcular seus parâmetros de modo simples com propriedades de esperança e variância

$$\boldsymbol{\alpha} = E(\boldsymbol{\theta}_{t:1}^{(i)}) = E(\boldsymbol{D}_t \boldsymbol{\theta}_{t:0}^{(i)}) = \boldsymbol{D}_t E(\boldsymbol{\theta}_{t:0}^{(i)}) = \boldsymbol{D}_t E(\boldsymbol{\theta}_{t:0}^{(i)}) = \boldsymbol{D}_t \mathscr{M}_t^{(i)} \tag{4.16}$$

$$\boldsymbol{A} = Cov(\boldsymbol{\theta}_{t:1}^{(i)}) = Cov(\boldsymbol{D}_t \boldsymbol{\theta}_{t:0}^{(i)}) = \boldsymbol{D}_t Cov(\boldsymbol{\theta}_{t:0}^{(i)} \boldsymbol{D}_t^T) = \boldsymbol{D}_t \mathscr{C}_t^{(i)} \boldsymbol{D}_t^T \tag{4.17}$$

Finalmente aplicando o Lema (3), em conjunto com a distribuição $(\boldsymbol{\theta}_{t:1}^{(i)} | \mathbf{X}_{0:t}^{(i)})$ e $(Y_{t:1}^{(i)} | \boldsymbol{\theta}_{t:1}^{(i)})$

$$\left( \begin{array}{c} Y_{t:1}^{(i)} \\ \boldsymbol{\theta}_{t:1}^{(i)} \end{array} \middle| \mathbf{X}_{t:0}^{(i)} \right)$$

com distribuição Normal

$$\mathrm{N}\left[ \begin{pmatrix} \mathrm{bdiag}(\boldsymbol{F}_{t:1}^{(i)T}) \\ \boldsymbol{D}_t \mathscr{M}_t^{(i)} \end{pmatrix}, \begin{pmatrix} \mathrm{diag}(V_{t:1}^{(i)}) + \mathrm{bdiag}(\boldsymbol{F}_{t:1}^{(i)}) \boldsymbol{D}_t \mathscr{C}_t^{(i)} \boldsymbol{D}_t^T \mathrm{bdiag}(\boldsymbol{F}_{t:1}^{(i)T})^T & \mathrm{bdiag}(\boldsymbol{F}_{t:1}^{(i)T}) \boldsymbol{D}_t \mathscr{C}_t^{(i)} \boldsymbol{D}_t^T \\ \boldsymbol{D}_t \mathscr{C}_t^{(i)T} \boldsymbol{D}_t^T \mathrm{bdiag}(\boldsymbol{F}_{t:1}^{(i)T})^T & \boldsymbol{D}_t \mathscr{C}_t^{(i)} \boldsymbol{D}_t^T \end{pmatrix} \right]$$

onde obtemos a marginal

$$\left(Y_{t:1}^{(i)} | \mathbf{X}_{t:0}^{(i)}\right) \sim \mathrm{N}\left[\mathrm{bdiag}(\boldsymbol{F}_{t:1}^{(i)T}), \mathrm{diag}(\boldsymbol{V}_{1:t}^{(i)}) + \mathrm{diag}(\boldsymbol{F}_{t:1}^{(i)}) \boldsymbol{D}_t \mathscr{C}_t^{(i)} \boldsymbol{D}_t^T \mathrm{diag}(\boldsymbol{F}_{t:1}^{(i)T})^T\right]$$

## 4.3 Lidando com $V_t$ desconhecida

O MLD univariado está completamente especificado com o conhecimento da quádrupla $\{\boldsymbol{F}_t, \boldsymbol{G}_t, V_t, \boldsymbol{W}_t\}$. Enquanto $\boldsymbol{F}_t$ e $\boldsymbol{G}_t$ são escolhidas de acordo com a estrutura das séries (como tendências e sazonalidades) as variâncias, ou as matrizes de covariâncias, são estimadas de diferentes modos. Para $\boldsymbol{W}_t$, esta pode ser elicitada via fatores de descontos ou estimada (conforme discutido na Seção (3.4) ). Nesta seção discutimos como lidar com o caso no qual $V_t$ é desconhecida.

Quando supomos que $V_t = V$ para todo $t \geq 1$, é possível obter expressões analíticas para as equações do filtro de Kalman e de suavização (ver Seção (3.3)). Os resultados da seção anterior podem ser reescritos para este caso.

**Definição 9.** *Definimos o MLD com* $V_t^{(i)} = V^{(i)} = \frac{1}{\phi^{(i)}}$ *como*

$$\{\boldsymbol{F}^{(i)}, \boldsymbol{G}^{(i)}, \frac{1}{\boldsymbol{\phi}^{(i)}}, \boldsymbol{W}^{(i)}\}_t$$

*ou seja:*

$$(\boldsymbol{X}_t^{(i)}|\boldsymbol{\theta}_t^{(i)}) \sim N\left[\mathbb{F}_t^{T(i)}\boldsymbol{\theta}_t^{(i)}, \frac{1}{\phi^{(i)}}\right] \tag{4.18}$$

$$(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{\theta}_{t-1}^{(i)}) \sim N\left[\boldsymbol{G}_t^{(i)}\boldsymbol{\theta}_{t-1}^{(i)}, \frac{\boldsymbol{W}_t^{\star(i)}}{\phi^{(i)}}\right] \tag{4.19}$$

$$(\boldsymbol{\theta}_0|\boldsymbol{X}_0^{(i)}) \sim N\left[\boldsymbol{m}_0, \frac{\boldsymbol{C}_0^{\star}}{\phi^{(i)}}\right] \tag{4.20}$$

$$(\phi|\boldsymbol{X}_0^{(i)}) \sim Gama\left[\frac{n_0}{2}, \frac{n_0 S_0}{2}\right] \tag{4.21}$$

Considerando o MLD definido acima, o filtro de Kalman é dado pelo seguinte teorema.

**Teorema 6.** *Para o i-ésimo modelo, e para cada $t \geq 1$, valem as seguintes afirmações:*

1. *Posteriori no tempo $t-1$:*

$$(\boldsymbol{\theta}_{t-1}^{(i)}|\boldsymbol{X}_{0:t-1}^{(i)}, \phi^{(i)}) \sim N\left[\boldsymbol{m}_{t-1}^{(i)}, \frac{\boldsymbol{C}_{t-1}^{\star(i)}{}^{(i)}}{\phi}\right]$$

$$(\boldsymbol{\theta}_{t-1}^{(i)}|\boldsymbol{X}_{0:t-1}^{(i)}) \sim T_{n_{t-1}}\left[\boldsymbol{m}_{t-1}^{(i)}, \boldsymbol{C}_{t-1}^{(i)}\right]$$

$$(\phi^{(i)}|\boldsymbol{X}_{0:t-1}^{(i)}) \sim Gama\left[\frac{n_{t-1}}{2}, \frac{n_{t-1}S_{t-1}}{2}\right]$$

2. *Priori no tempo $t$:*

$$(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{X}_{0:t-1}, \boldsymbol{\phi}^{(i)}) \sim N\left[\boldsymbol{a}_t^{(i)}, \frac{\boldsymbol{R}_t^{\star(i)}}{\phi}\right],$$

$$(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{X}_{0:t-1}) \sim T_{n_{t-1}}\left[\boldsymbol{a}_t^{(i)}, \boldsymbol{R}_t^{(i)}\right],$$

*onde*

$$\begin{aligned}
\boldsymbol{a}_t^{(i)} &= \boldsymbol{G}_t^{(i)}\boldsymbol{m}_{t-1}^{(i)} \\
\boldsymbol{R}_t^{\star(i)} &= \frac{1}{\phi^{(i)}}\left(\boldsymbol{W}_t^{\star(i)} + \boldsymbol{G}_t^{(i)}\boldsymbol{C}_{t-1}^{\star(i)}\boldsymbol{G}_t^{(i)T}\right) \\
\boldsymbol{R}_t^{(i)} &= S_{t-1}\boldsymbol{R}_t^{\star(i)}.
\end{aligned}$$

3. *Previsão para o tempo $t$:*

$$(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{X}_{0:t-1}^{(i)},\phi^{(i)}) \sim N\left[\boldsymbol{f}_t^{(i)}, \frac{\boldsymbol{Q}_t^{\star(i)}}{\phi^{(i)}}\right],$$
$$(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{X}_{0:t-1}^{(i)}) \sim T_{n_{t-1}}\left[\boldsymbol{f}_t^{(i)}, \boldsymbol{Q}_t^{(i)}\right],$$

*onde*

$$\begin{aligned}
\boldsymbol{f}_t^{(i)} &= \mathbb{F}_t^{T(i)}\boldsymbol{a}_t^{(i)} \\
\boldsymbol{Q}_t^{(i)\star} &= \frac{1}{\phi^{(i)}}\left[\boldsymbol{V}_t^{(i)} + \mathbb{F}_t^{T(i)}\boldsymbol{R}_t^{(i)}\mathbb{F}_t^{(i)})\right] \\
\boldsymbol{Q}_t^{(i)} &= S_{t-1}\boldsymbol{Q}_t^{\star(i)}
\end{aligned}$$

4. *Posteriori para o tempo t:*

$$(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{X}_{0:t}^{(i)},\phi^{(i)}) \sim N\left[\boldsymbol{m}_t^{(i)}, \frac{\boldsymbol{C}_t^{\star(i)}}{\phi^{(i)}}\right],$$
$$(\boldsymbol{\theta}_t^{(i)}|\boldsymbol{X}_{0:t}^{(i)}) \sim T_{n_t}\left[\boldsymbol{m}_t^{(i)}, \boldsymbol{C}_t\right],$$

*com*

$$\begin{aligned}
\boldsymbol{A}_t^{(i)} &= \boldsymbol{R}_t^{(i)}\mathbb{F}_t^{(i)}\boldsymbol{Q}_t^{(i)-1} \\
\boldsymbol{m}_t^{(i)} &= \boldsymbol{a}_t^{(i)} + \boldsymbol{A}_t^{(i)}\left[\boldsymbol{x}_t^{(i)} - \boldsymbol{f}_t^{(i)}\right] \\
\boldsymbol{C}_t^{\star(i)} &= \frac{1}{\phi^{(i)}}\left(\boldsymbol{R}_t^{(i)} - \boldsymbol{A}_t^{(i)}\boldsymbol{Q}_t^{(i)}\boldsymbol{A}_t^{(i)T}\right) \\
\boldsymbol{C}_t^{(i)} &= S_t\boldsymbol{C}_t^{\star(i)} \\
n_t &= n_{t-1} + l^{(i)} \\
S_t &= \frac{n_{t-1}S_{t-1}}{n_t} + \frac{1}{n_t}\left(\boldsymbol{x}_t^{(i)} - \boldsymbol{f}_t^{(i)}\right)^T \boldsymbol{Q}_t^{\star(i)-1}\left(\boldsymbol{x}_t^{(i)} - \boldsymbol{f}_t^{(i)}\right)
\end{aligned}$$

*Demonstração.* A demonstração utiliza relações conhecidas entre as distribuições normal e gama. Veja (Harrison and West (1999)). □

O corolário abaixo mostra que ainda temos uma forma recursiva para obter a distribuição da suavização conjunta.

**Corolário 1.** *Para $t \geq 1$ e para a i-ésima classe,*

1. *Posteriori no tempo $t-1$*

$$(\boldsymbol{\theta}_{t-1:0}^{(i)}|\boldsymbol{X}_{0:t-1}^{(i)},\boldsymbol{\phi}^{(i)}) \sim N\left[\mathscr{M}_{t-1}^{(i)},\frac{1}{\boldsymbol{\phi}^{(i)}}\mathscr{C}_{t-1}^{\star(i)}\right]$$
$$(\boldsymbol{\theta}_{t-1:0}^{(i)}|\boldsymbol{X}_{0:t-1}^{(i)}) \sim T_{n_{t-1}}\left[\mathscr{M}_{t-1}^{(i)},\mathscr{C}_{t-1}^{(i)}\right]$$
$$(\boldsymbol{\phi}^{(i)}|\boldsymbol{X}_{0:t-1}^{(i)}) \sim Gama\left[\frac{n_{t-1}}{2},\frac{n_{t-1}S_{t-1}}{2}\right]$$

2. *Priori no tempo t*

$$(\boldsymbol{\theta}_{t:0}^{(i)}|\boldsymbol{X}_{0:t-1}^{(i)},\boldsymbol{\phi}^{(i)}) \sim N\left[\mathscr{A}_{t}^{(i)},\frac{1}{\boldsymbol{\phi}^{(i)}}\mathscr{R}_{t}^{\star(i)}\right]$$
$$(\boldsymbol{\theta}_{t:0}^{(i)}|\boldsymbol{X}_{0:t-1}^{(i)}) \sim T_{n_{t-1}}\left[\mathscr{A}_{t}^{(i)},\mathscr{R}_{t}^{(i)}\right]$$

*onde*

$$\mathscr{A}_{t}^{(i)} = \begin{pmatrix} \boldsymbol{G}_t\boldsymbol{E}_{t-1}\mathscr{M}_{t-1}^{(i)} \\ \mathscr{M}_{t-1}^{(i)} \end{pmatrix}$$
$$\mathscr{R}_{t}^{\star(i)} = \frac{1}{\phi}\begin{pmatrix} \boldsymbol{W}_t^{(i)}+\boldsymbol{G}_t^{(i)}\boldsymbol{E}_{t-1}\mathscr{C}_{t-1}^{\star(i)}\boldsymbol{E}_{t-1}^{T}\boldsymbol{G}_t^{(i)T} & \boldsymbol{G}_t^{(i)}\boldsymbol{E}_{t-1}\mathscr{C}_{t-1}^{\star(i)} \\ \mathscr{C}_{t-1}^{\star(i)}\boldsymbol{E}_{t-1}^{T}\boldsymbol{G}_t^{(i)T} & \mathscr{C}_{t-1}^{\star(i)} \end{pmatrix}$$
$$\mathscr{R}_{t}^{(i)} = S_{t-1}\mathscr{R}_{t}^{\star(i)}$$

3. *Com*

$$\begin{pmatrix} \boldsymbol{X}_t^{(i)} \\ \boldsymbol{\theta}_{t:0}^{(i)} \end{pmatrix}\Bigg|\boldsymbol{X}_{0:t-1}^{(i)} \sim N\left[\begin{pmatrix} \mathscr{F}_t^{(i)} \\ \mathscr{A}_t^{(i)} \end{pmatrix},\frac{1}{\boldsymbol{\phi}^{(i)}}\begin{pmatrix} \mathscr{Q}_t^{(i)} & \mathbb{F}_t^{(i)T}\boldsymbol{E}_t\frac{1}{\phi}\mathscr{R}_t^{\star(i)} \\ \mathscr{R}_t^{\star(i)}\boldsymbol{E}_t^{T}\mathbb{F}_t^{(i)} & \mathscr{R}_t^{\star(i)} \end{pmatrix}\right]$$

*onde*

$$\mathscr{F}_t^{(i)} = \mathbb{F}_t^{(i)T}\boldsymbol{E}_t\mathscr{A}_t^{(i)} \tag{4.22}$$
$$\mathscr{Q}_t^{(i)} = \frac{1}{\boldsymbol{\phi}^{(i)}}\left(\boldsymbol{I}_{l^{(i)}}+\mathbb{F}_t^{(i)T}\boldsymbol{E}_t\mathscr{R}_t^{\star(i)}\boldsymbol{E}_t^{T}\mathbb{F}_t^{(i)}\right) \tag{4.23}$$

*tem-se que a posteriori da conjunta no tempo t é*

$$(\boldsymbol{\theta}_{t:0}^{(i)}|\boldsymbol{X}_{0:t}^{(i)}) \sim T_{n_t}\left[\mathscr{M}_{t}^{(i)},\mathscr{C}_{t}^{(i)}\right]$$

*onde*

$$\mathscr{M}_t^{(i)} = \mathscr{A}_t^{(i)} + \mathscr{R}_t^{(i)} \boldsymbol{E}_t^T \mathbb{F}_t^{(i)} \mathscr{Q}_t^{(i)-1} \left( \boldsymbol{x}_t^{(i)} - \mathscr{F}_t^{(i)} \right)$$
$$\mathscr{C}_t^{(i)} = S_t \left( \mathscr{R}_t^{\star(i)} - \mathscr{R}_t^{\star(i)} \boldsymbol{E}_t \mathbb{F}_t^{(i)} \mathscr{Q}_t^{(i)-1} \mathbb{F}_t^{(i)T} \boldsymbol{E}_t \mathscr{R}_t^{\star(i)} \right)$$

*e $n_t$ e $S_t$ são como definidos no Teorema (6).*

*Demonstração.* Basta utilizar os resultados da distribuição normal-gama em conjunto com o Teorema (6). □

# Capítulo 5

# Estudos de Simulação

Neste capítulo, apresentamos alguns estudos de simulação computacional onde geramos quatro cenários distintos, correspondentes a observações de séries temporais provenientes de diferentes MLD. Analisamos estratégias para estimar a variância nos modelos e comparar o desempenho do CBMLD com relação a outros classificadores usuais em Análise Discriminante.

## 5.1 Organização das Simulações

Antes de testar o desempenho do CBMLD em séries temporais reais, é importante adquirir conhecimento sobre sua performance no caso em que conhecemos a verdadeira estrutura da série. Isto pode ser realizado através de estudos de simulação, nos quais podemos gerar séries temporais a partir de um determinado MLD. Para $i = 1, \ldots, K$, suponha que queremos gerar $l^{(i)}$ séries temporais de comprimento $t$ segundo o modelo linear dinâmico $\{\boldsymbol{F}, \boldsymbol{G}, \boldsymbol{V}, \boldsymbol{W}\}_t^{(i)}$. Utilizamos o seguinte algoritmo para gerar as séries temporais em todas as simulações:

1. Simule $\boldsymbol{\theta}_0^{(i)} \sim \mathrm{N}[\boldsymbol{m}_0, \boldsymbol{C}_0]$.
2. Para $j$ variando de 1 a $t$, simule $\boldsymbol{\theta}_j \sim \mathrm{N}\left[\boldsymbol{G}_j^{(i)} \boldsymbol{\theta}_{j-1}^{(i)}, \boldsymbol{W}_j^{(i)}\right]$.
3. Para cada $j$ variando de 1 a $t$ gere, independentemente, $y_1, \ldots, y_{l^{(i)}} \sim \mathrm{N}\left[\boldsymbol{F}_j^{T(i)} \boldsymbol{\theta}_j^{(i)}, V_j^{(i)}\right]$

Note que em cada classe é gerada uma única sequência de parâmetros onde as séries temporais dentro de cada classe são simuladas a partir destas. Portanto, é possível que duas ou mais classes sejam provenientes do mesmo MLD, sendo que suas diferenças serão dadas pelo vetor latente de parâmetros. Esta estratégia e simulação condiz com as séries reais que temos observado na prática (para alguns exemplos, veja o Capítulo 6).

Na Seção 5.2 apresentamos dois estudos de simulação(doravante denotados por $S1$ e $S2$) realizados com finalidade de avaliar o impacto das estratégias de estimação (ou elicitação) das variâncias presentes no CBMLD. No cenário $S1$ geramos 100 séries temporais de comprimento 20 com duas classes (sendo 50 séries para cada) geradas a partir de um MLD polinomial de ordem 1. No cenário $S2$ geramos 100 séries temporais de comprimento 20 com duas classes (50 para cada) geradas a partir de um MLD trigonométrico de período 6 e 1 harmônico. Para a avaliação dos dois estudos, foi utilizada a validação cruzada com repetidas subamostras aleatórias, com 500 repetições, utilizando sempre 70% das séries temporais como conjunto de treinamento.

Na Seção 5.3 apresentamos dois estudos de simulação (doravante denotados por $S3$ e $S4$) realizados para comparar o desempenho dos classificadores discutidos nesta dissertação. No cenário $S3$ geramos 180 séries temporais de comprimento 30 com duas classes (90 séries para cada classe) geradas a partir de um MLD polinomial de ordem 1. No cenário $S4$ geramos 20 amostras de séries temporais de comprimento 30 com duas classes (10 séries para cada) geradas também a partir de MLD polinomial de ordem 1. Para a avaliação dos dois estudos, foi utilizada a validação cruzada com repetidas subamostras aleatórias, com 500 repetições, utilizando sempre 70% das séries temporais como conjunto de treinamento sendo que o número mínimo de séries temporais de uma classe no conjunto de treinamento foram 40 e 5 para $S3$ e $S4$ respectivamente (estes valores foram determinados de modo empírico, assegurando que seria possível aplicar a ADQ em $S3$ e a ADR em $S4$).

## 5.2 Comparando Estratégias para Estimar as Variâncias

Os classificadores ADQ, ADR e Naive Bayes estimam a variância de cada classe. Neste sentido, é importante que os modelos lineares dinâmicos tenham a vantagem de estimar de maneira adequada as variâncias $V_t$ e $\boldsymbol{W}_t$, (tendo como objetivo atingir boas taxas de acerto nos estudos comparativos de classificação).

Nas Seções 3.4 e 4.3 discutimos diferentes métodos para lidar com as variâncias do modelo linear dinâmico. Abaixo, listamos as três estratégias que foram avaliadas nesta dissertação:

- $(\tilde{V}, \delta)$: Nesta estratégia, $V_t$ é estimado em cada tempo pelo seu estimador de máxima verossimilhança $\tilde{V}_t$, enquanto que $\boldsymbol{W}_t$ é elicitado através de fatores de descontos.

- $(\phi, \delta)$: Nesta estratégia, $V_t$ é considerado fixado ao longo do tempo, resultando no classificador que utiliza o modelo $t$-Student. $\boldsymbol{W}_t$ é elicitado através de fatores de descontos.

- $(\tilde{V}, \boldsymbol{W})$: Nesta estratégia, $V_t$ é estimado em cada tempo pelo seu estimador de máxima verossimilhança $\tilde{V}_t$. Para $h$ modelos superpostos, $\boldsymbol{W}_t = \text{diag}\{\boldsymbol{W}_{1t}, \ldots, \boldsymbol{W}_{ht}\}$ e cada $\boldsymbol{W}_{jt} = w_j \boldsymbol{I}$. Os hiper parâmetros $w_1, \ldots, w_h$ são estimados através da maximização da distribuição preditiva.

Com o objetivo de determinar quais estratégias são adequadas, fizemos dois estudos de simulação. O seguinte algoritmo foi empregado nos estudos de simulação S1 e S2 visando determinar as taxas de erro associadas às estratégias comparadas:

**Algoritmo A**

1. Retire uma amostra aleatória simples sem reposição de 70 séries temporais para constituir a amostra de treino.

2. Utilize a amostra de treino para construir os classificadores com cada estratégia de variância.

3. Classifique as séries temporais restantes utilizando os classificadores obtidos no Passo 2.

4. Guarde a taxa de erro de classificação para cada classificador definidos pelas distintas estratégias.

No primeiro estudo de simulação computacional (S1) foram considerados os seguintes modelos polinomiais de ordem 1:

- Classe 1: $\{1, 1, 10, .1\}$ com $m_0 = 0$

- Classe 2: $\{1, 1, 10, .1\}$ com $m_0 = 1$

A Figura 5.1 mostra o gráfico de exemplos de duas classes simuladas a partir do modelo de ordem 1.

A Tabela 5.1 mostra a média e desvio padrão das taxas de erro expressas em porcentagem. Podemos notar que a estratégia que emprega o $\tilde{V}_t$, tanto como com a elicitação como a estimação de $W$, apresentou desempenho superior a estratégia que emprega $V = 1/\phi$.

Na Tabela 5.2 como descrito no Capítulo 2, apresenta a porcentagem do número de vezes (Total) que um método tem taxa de erro menor ou igual a de outro. Desta tabela, observamos que a estratégia com $(\phi, \delta)$, em termos da taxa de erro, foi superior em torno de somente 10% das vezes quando comparada com as outras estratégias.

No segundo estudo de simulação, o estudo $S2$, foram simuladas observações de séries temporais em duas classes, a partir do mesmo modelo trigonométrico onde o único diferencial entre as duas classes se baseia em $\theta$. Esta situação, em geral, é o que observamos em dados reais, onde os dados são bastante sobrepostos e, quando classificados, o MLD os classifica através da flutuação das séries observadas em torno de $F_t\theta$:

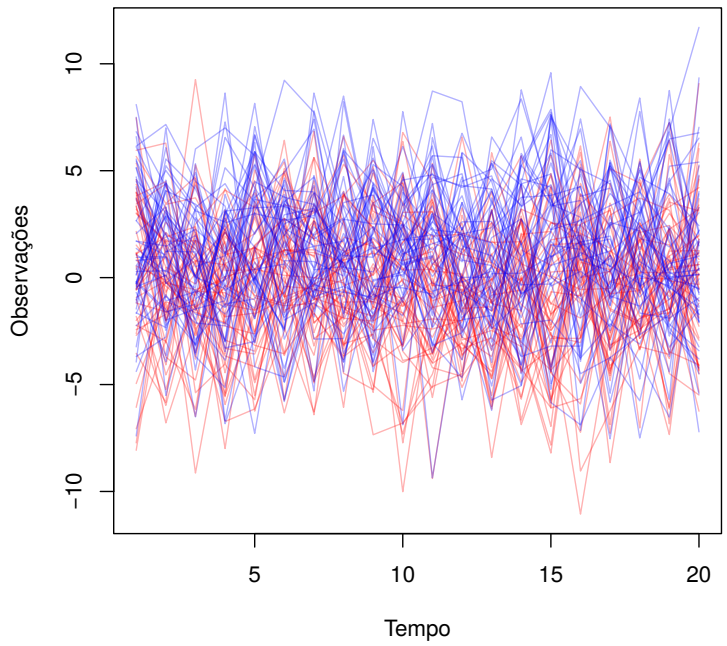


Figura 5.1: Séries simuladas com duas classes a partir do MLD polinomial de ordem 1.

| | $\tilde{V}_t$, $\delta$ | $\phi$, $\delta$ | $\tilde{V}_t$, $W$ |
|---|---|---|---|
| Média | 0,1859 | 0,4310 | 0,1882 |
| Desvio padrão | 0,06459 | 0,13448 | 0,06504 |

Tabela 5.1: Média e desvio padrão das taxas de erro (em %) com diferentes estratégias de estimação das variâncias para o cenário S1.

| | | $\tilde{V}_t$,$\delta$ | $\phi$, $\delta$ | $\tilde{V}_t$, $W$ |
|---|---|---|---|---|
| $\tilde{V}_t$, $\delta$ | Igual | - | 4,0 | 56,4 |
| | Menor | - | 90,8 | 23,4 |
| | Total | - | 94,8 | 79,8 |
| $\phi$, $\delta$ | Igual | 4,0 | - | 4,2 |
| | Menor | 5,2 | - | 6,0 |
| | Total | 9,2 | - | 10,2 |
| $\tilde{V}_t$, $W$ | Igual | 56,4 | 4,2 | - |
| | Menor | 20,2 | 89,8 | - |
| | Total | 76,6 | 94,0 | - |

Tabela 5.2: Desempenho em termos das taxas de erro (em %) de classificação com diferentes estratégias para estimação da variância para o cenário S1.

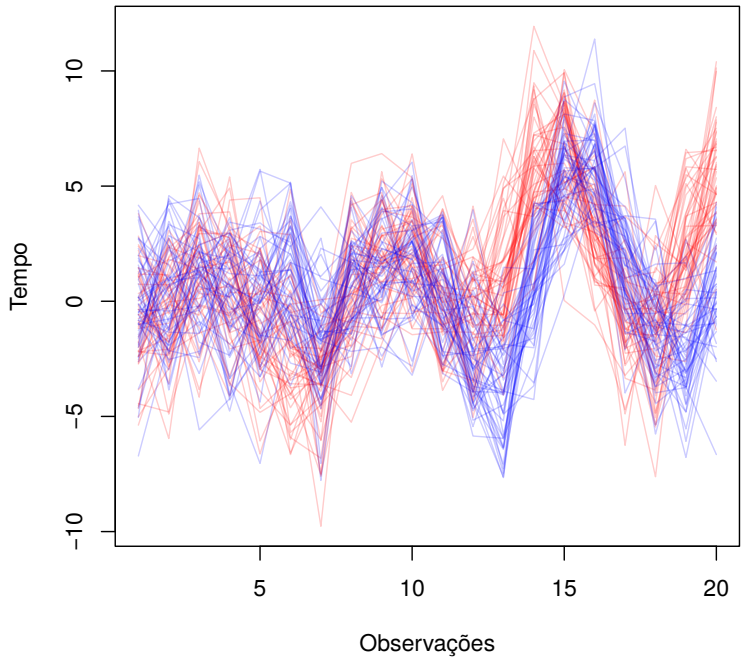


Figura 5.2: Séries simuladas com duas classes a partir do MLD trigonométrico de período 6 com um harmônico.

$$\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} \cos(\pi/3) & \sin(\pi/3) \\ -\sin(\pi/3) & \cos(\pi/3) \end{pmatrix}, 5, \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix} \right\}$$

Com este modelo trigonométrico empregado as classes serão diferenciadas pelas estruturas latentes provenientes das sequências $\theta_j$ simuladas. Desta forma, foram obtidas séries correspondentes às classes com acentuada superposição, que se constituiu numa tentativa de reproduzir o comportamento das séries observadas em problemas reais.

A Figura 5.2 apresenta exemplos das séries simuladas no segundo estudo de simulação. A Tabela 5.3 mostra a média e o desvio padrão das taxas de erro obtidas para cada estratégia. Note que a taxa média foi menor para $V_t$ estimado por $\tilde{V}_t$ e $\boldsymbol{W}$ otimizado. Através da Tabela 5.4 percebemos que os métodos com $\tilde{V}_t$ tiveram bom desempenho, sendo que neste estudo o método que otimiza $\boldsymbol{W}$ foi superior aos fatores de descontos. Antecipando um comentário sobre o emprego do CBMLD com séries oriundas de conjunto de dados reais, também percebemos que o fator de descontos apresentou performance ligeiramente inferior aos demais métodos. Portanto, decidimos empregar a estratégia $(\tilde{V}_t, W)$, que se mostrou superior nesta análise inicial, nos estudos de simulação comparando classificadores, apresentado na seção a seguir, e nas aplicações com dados reais apresentadas no Capítulo 6.

| | $\tilde{V}_t$, $\delta$ | $\phi$, $\delta$ | $\tilde{V}_t$, $W$ |
|---|---|---|---|
| Média | 0,002067 | 0,278533 | 0,000067 |
| Desvio Padrão | 0,008582189 | 0,188624573 | 0,001490712 |

Tabela 5.3: Média e desvio padrão das taxas de erro (em %) comparando diferentes estratégias de variância para o cenário S2.

| | | $\tilde{V}_t$, $\delta$ | $\phi$, $\delta$ | $\tilde{V}_t$, $W$ |
|---|---|---|---|---|
| $\tilde{V}_t$, $\delta$ | Igual | - | 0,082 | 0,944 |
| | Menor | - | 0,918 | 0,000 |
| | Total | - | 1,000 | 0,944 |
| $\phi$, $\delta$ | Igual | 0,082 | - | 0,076 |
| | Menor | 0,000 | - | 0,000 |
| | Total | 0,082 | - | 0,076 |
| $\tilde{V}_t$, $W$ | Igual | 0,944 | 0,076 | - |
| | Menor | 0,056 | 0,924 | - |
| | Total | 1,000 | 1,000 | - |

Tabela 5.4: Desempenho em termos das taxas de erro (em %) comparando diferentes estratégias de variância para o cenário S2.

## 5.3 Comparações do CBMLD com outros classificadores

Nesta seção comparamos o desempenho do classificador proposto, CBMLD, com estratégia de estimação de variância $(\tilde{V}_t, \boldsymbol{W})$, com os outros classificadores discutidos nesta dissertação. Este estudo de simulação considera o modelo $\{1, 1, 10, 1\}$, Modelo Polinomial de Ordem 1. Neste modelo, a variância das observações é maior que a dos parâmetros, algo comum na prática. Portanto, teremos séries da mesma classe distantes entre si com um comportamento estrutural pouco aparente.

Criamos dois cenários S3 e S4, como já descritos anteriormente. No primeiro, S3, o número de séries no conjunto de treinamento é maior que o comprimento das séries temporais. Neste caso, todos os métodos de classificação discutidos nesta dissertação podem ser utilizados. No segundo, S4, o número de séries no conjunto de treino é menor que o comprimento das séries e métodos de classificação como o ADL e o ADQ não podem ser utilizados.

Nesse segundo estudo de simulação foi considerado seguinte algoritmo:

**Algoritmo B**

1. Retire uma amostra aleatória simples sem reposição de $n_T$ (número de séries no conjunto de treinamento) séries temporais.

   (a) Verifique quantas séries de cada classe foram selecionadas. Se o total de séries para uma das classes for menor que $n_C$ (número de elementos no conjunto de teste) , volte para o Passo 1. Senão, prossiga para o Passo 2.

2. Utilize a amostra de treino para construir os classificadores.

3. Classifique as séries temporais restantes utilizando os classificadores obtidos no Passo 2.

4. Guarde a taxa de erro de classificação para cada classificador.

O procedimento estabelecido no Algoritmo B foi repetido 500 vezes nos estudos de simulação S3 e S4.

- No primeiro cenário, estudo S3, foram geradas 180 séries temporais de comprimento 30, sendo 90 para cada classe. Destas séries, em cada repetição do Algoritmo B, foram utilizados $n_T = 112$ (que corresponde a 70% dos dados) e $n_C = 40$ (para evitar problemas com a ADL e a ADQ).

- No segundo cenário, estudo S4, foram geradas 20 séries temporais de comprimento 30, sendo 10 de cada classe. Destas séries, em cada repetição do Algoritmo $B$, foram utilizados $n_T = 14$ (que corresponde a 70% dos dados) e $n_C = 5$. Nestes casos, os classificadores ADL e ADQ não foram utilizados.

Na Figura 5.3 apresentamos exemplos de séries simuladas nos cenários S3 e S4.

A média e o desvio padrão das taxas de erro de classificação para cada método no cenário S3 estão registradas nas Tabelas 5.5. Dos valores apresentados nesta Tabela 5.5, se consideramos intervalos de confiança com aproximação normal para a média da taxa de erro e um nível de significância de 5%, observamos que o desempenho do CBMLD foi equivalente aos dos classificadores ADR, NBN e NBK. Ainda neste cenário, os métodos ADQ e 1-NN tiveram desempenho ruim, sendo os únicos significativamente inferiores aos demais.

Na Tabela 5.6, estão descritas as proporções de vezes que cada método apresentou taxa de erro menor ou igual a de outro método. Em termos comparativos, podemos notar que o classificador ADR teve melhor desempenho contra todas as alternativas, seguido do CBMLD.

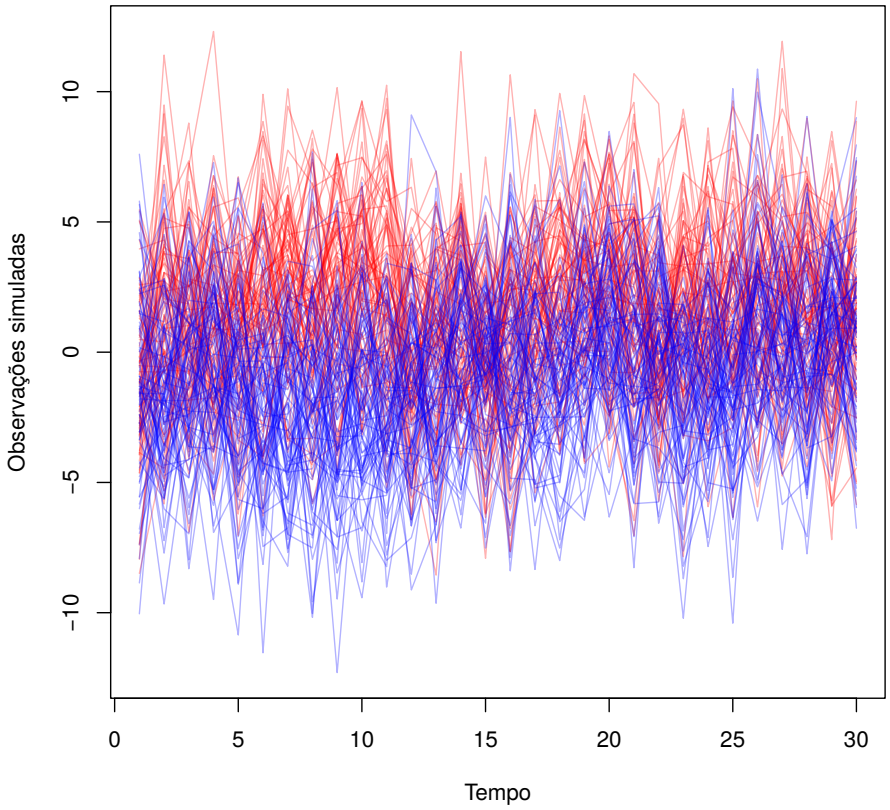


Figura 5.3: Séries simuladas com duas classes a partir de um MLD polinomial de ordem 1 para os cenários S3 e S4.

| Estatística | CBMLD | 1-NN | ADL | ADQ | ADR | NBN | NBK |
|---|---|---|---|---|---|---|---|
| Média | 0,0665 | 0,1765 | 0,0800 | 0,2435 | 0,0592 | 0,0673 | 0,0941 |
| Desvio padrão | 0,0301 | 0,0453 | 0,0346 | 0,0648 | 0,0294 | 0,0296 | 0,0348 |

Tabela 5.5: Média e desvio padrão das taxas de erro (em %) para o cenário S3.

| | | CBMLD | 1-NN | ADL | ADQ | ADR | NBN | NBK |
|---|---|---|---|---|---|---|---|---|
| CBMLD | Igual | - | 1,0 | 22,2 | 0,2 | 26,4 | 41,4 | 17,6 |
| | Menor que | - | 98,2 | 52,6 | 99,6 | 26,6 | 32,0 | 71,6 |
| | Total | - | 99,2 | 74,8 | 99,8 | 53,0 | 73,4 | 89,2 |
| 1-NN | Igual | 1,0 | - | 2,0 | 7,8 | 0,8 | 1,2 | 3,2 |
| | Menor que | 0,8 | - | 2,2 | 76,6 | 0,6 | 0,6 | 4,2 |
| | Total | 1,8 | - | 4,2 | 84,4 | 1,4 | 1,8 | 7,4 |
| ADL | Igual | 22,2 | 2,0 | - | 0,2 | 20 | 22,8 | 17,2 |
| | Menor que | 25,2 | 95,8 | - | 99,4 | 15,2 | 27,0 | 56,2 |
| | Total | 47,4 | 97,8 | - | 99,6 | 35,2 | 49,8 | 73,4 |
| ADQ | Igual | 0,2 | 7,8 | 0,2 | - | 0 | 0,4 | 0,4 |
| | Menor que | 0,2 | 15,6 | 0,4 | - | 0,2 | 0,2 | 1,2 |
| | Total | 0,4 | 23,4 | 0,6 | - | 0,2 | 0,6 | 1,6 |
| ADR | Igual | 26,4 | 0,8 | 20,0 | 0,0 | - | 29,0 | 12,4 |
| | Menor que | 47,0 | 98,6 | 64,8 | 99,8 | - | 48,8 | 77,2 |
| | Total | 73,4 | 99,4 | 84,8 | 99,8 | - | 77,8 | 89,6 |
| NBN | Igual | 41,4 | 1,2 | 22,8 | 0,4 | 29,0 | - | 17,8 |
| | Menor que | 26,6 | 98,2 | 50,2 | 99,4 | 22,2 | - | 70,4 |
| | Total | 68,0 | 99,4 | 73,0 | 99,8 | 51,2 | - | 88,2 |
| NBK | Igual | 17,6 | 3,2 | 17,2 | 0,4 | 12,4 | 17,8 | - |
| | Menor que | 10,8 | 92,6 | 26,6 | 98,4 | 10,4 | 11,8 | - |
| | Total | 28,4 | 95,8 | 43,8 | 98,8 | 22,8 | 29,6 | - |

Tabela 5.6: Desempenho em termos das taxas de erro (em %) entre os classificadores para o cenário S3.

A média e o desvio padrão das taxas de erro de classificação para cada método no cenário S4 estão registradas nas Tabelas 5.7. Dos valores apresentados nesta Tabela 5.7, temos que os classificadores CBMLD, NBN e NBK apresentaram resultados equivalentes em desempenho, entretanto os melhores resultados foram os apresentados pelos classificadores ADR, com a melhor performance, seguido do 1-NN.

Na Tabela 5.8, estão descritas as proporções de vezes que cada método apresentou taxa de erro menor ou igual a de outro método, ainda no cenário S4. Da Tabela 5.8, em concordância com os resultados da Tabela 5.7, observamos que o classificador CBMLD apresentou o desempenho equivalente ao NBN e inferior ao $1-NN$ e o ADR. Destacando-se ainda, neste cenário, o desempenho do ADR superior a todos os demais classificadores.

| | CBMLD | 1-NN | ADR | NBN | NBK |
|---|---|---|---|---|---|
| Média | 0,052666667 | 0,003333333 | 0,001333333 | 0,052000000 | 0,065333333 |
| Desvio padrão | 0,10500501 | 0,02335670 | 0,01486224 | 0,11149877 | 0,12331428 |

Tabela 5.7: Média e desvio padrão das taxas de erro (em %) para o cenário S4.

| | | CBMLD | 1-NN | ADR | NBN | NBK |
|---|---|---|---|---|---|---|
| CBMLD | Igual | - | 73,2 | 75,2 | 88,6 | 64,0 |
| | Menor | - | 02,0 | 00,2 | 05,6 | 20,2 |
| | Total | - | 75,2 | 75,4 | 94,2 | 84,2 |
| 1-NN | Igual | 73,2 | - | 97,2 | 75,8 | 70,8 |
| | Menor | 24,8 | - | 00,8 | 22,2 | 27,6 |
| | Total | 98,0 | - | 98,0 | 98,0 | 98,4 |
| ADR | Igual | 75,2 | 97,2 | - | 77,6 | 72,2 |
| | Menor | 24,6 | 02,0 | - | 22,2 | 27,6 |
| | Total | 99,8 | 99,2 | - | 99,8 | 99,8 |
| NBN | Igual | 88,6 | 75,8 | 77,6 | - | 66,4 |
| | Menor | 05,8 | 02,0 | 00,2 | - | 19,8 |
| | Total | 94,4 | 77,8 | 77,8 | - | 86,2 |
| NBK | Igual | 64,0 | 70,8 | 72,2 | 66,4 | - |
| | Menor | 15,8 | 01,6 | 00,2 | 13,8 | - |
| | Total | 79,8 | 72,4 | 72,4 | 80,2 | - |

Tabela 5.8: Desempenho em termos das taxas de erro (em %) entre os classificadores para o cenário S4.

# Capítulo 6

# Aplicações em Dados Reais

Neste Capítulo analisamos algumas series temporais disponíveis no *UCR Time Series Classification Arquive* (Chen et al. (2015)). Até o momento da confecção desta dissertação haviam 85 arquivos de séries temporais para classificação. Para cada um deles realizamos uma inspeção visual tentando identificar séries que poderiam ser ajustadas com modelos lineares dinâmicos simples (como os polinomiais e os trigonométricos). Como a análise computacional demanda bastante tempo, selecionamos alguns conjuntos de dados e destacamos três deles nesta dissertação para devidas comparações dos resultados de classificação empregando o CBMLD com os dos classificadores usuais já citados. Tais conjuntos de dados são compostos de séries temporais reais (dados do Robô SONY AIBO e espectrometria de tipos de café) e pseudo-série temporal (dados das folhas Suecas adaptados para séries temporais). Cada conjunto de dados será discutido em uma seção deste capítulo.

Ressaltamos, ainda, que o Algoritmo B, descrito na Seção 5.3, foi empregado para determinar os conjuntos de treino e conjuntos de teste nas aplicações com os conjuntos de dados reais.

## 6.1 Classificação do Solo pelo Robô SONY AIBO

Voltando ao conjunto de dados do problema de AD do robô SONY AIBO, que é um pequeno robô quadrúpede em forma de cachorro equipado com múltiplos sensores, incluindo um acelerômetro tri axial. Como já citamos, temos medidas do acelerômetro no eixo horizontal que foram registradas enquanto o robô andava em círculos em dois tipos de superfícies: cimento e carpete. Cada série temporal representa uma volta completa. Foram registradas 621 voltas, sendo 349 no cimento e 272 no carpete, todas as séries com 71 de comprimento. O cimento é mais duro que o carpete, o que faz com que exista mais variabilidade na superfície. Considerando cada superfície como uma classe, o objetivo é classificar as séries com respeito aos dois tipos de superfícies.

Para cada classe no conjunto de treino, foi ajustado um modelo linear dinâmico polinomial de primeira ordem utilizado a estratégia de variância $(\tilde{V}_t, W)$. As séries observadas com a previsão um passo à frente são mostradas na Figura 6.1.

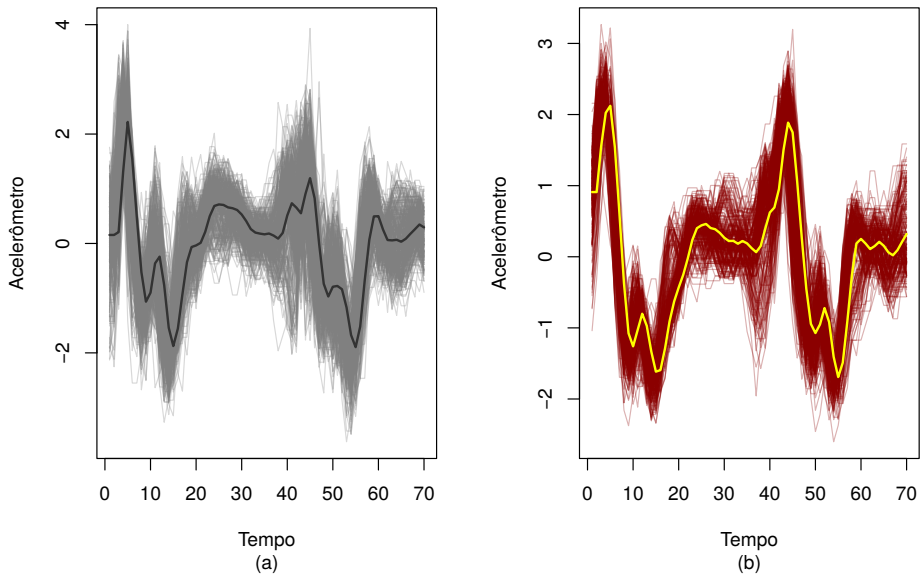


Figura 6.1: Séries do acelerômetro do Robô Sony AIBO nas duas superfícies (a) Cimento e (b) Carpete, com a previsão um passo à frente ajustado pelo MLD polinomial de primeira ordem.

Uma vez que escolhemos um modelo linear dinâmico apropriado, passamos a avaliar o desempenho dos classificadores. Realizamos um estudo retirando uma amostra aleatória de séries temporais de tamanho 140 para constituir o conjunto de treino e utilizamos as restantes como conjunto de teste. Este procedimento foi repetido 500 vezes. Na Tabela 6.1 estão apresentadas as médias e os desvios padrão das taxas de erros dos classificadores, os resultados indicam que, nosso classificador não foi o melhor, embora se considerarmos intervalos de confiança com aproximação normal para a média da taxa de erro e um nível de significância de 5% não exista diferença significativa entre os métodos se comparados 2 a 2. Para melhor ilustrar essa comparação, introduzimos o gráfico da Figura 6.2. Com relação ao ADQ, que obteve desempenho inferior a todos os demais classificadores, provavelmente seu desempenho foi prejudicado pelo pequeno número de observações por classe, o que impossibilitou uma estimação eficiente da matriz de covariâncias nas classes.

A Tabela 6.2 mostra o desempenho do CBMLD em comparação aos classificadores usuais, em termos da porcentagem de vezes que apresentou taxa de erro menor ou igual. Desta tabela observamos que o CBMLD apresentou desempenho inferior ao 1-NN, ADR e NBK, porém com desempenho superior ao ADL, ADQ e NBN.

Os gráficos da Figura 6.3, ilustram os resultados da Tabela 6.2 caso a caso somente no item Total, indicando na cor preto o quanto o primeiro classificador foi eficiente em

| | CBMLD | 1-NN | ADL | ADQ | ADR | NBN | NBK |
|---|---|---|---|---|---|---|---|
| Média | 0,03163 | 0,02577 | 0,05875 | 0,33517 | 0,01488 | 0,03189 | 0,02339 |
| Desvio Padrão | 0,012476 | 0,007846 | 0,012595 | 0,080105 | 0,012529 | 0,012554 | 0,009526 |

Tabela 6.1: Média e desvio padrão das taxas de erro (em %) dos classificadores para as séries do Robô SONY AIBO.

| | | CBMLD | 1-NN | ADL | ADQ | ADR | NBN | NBK |
|---|---|---|---|---|---|---|---|---|
| CBMLD | Igual | - | 8,0 | 1,4 | 0,0 | 2,2 | 79,6 | 6,6 |
| | Menor | - | 30,8 | 93,6 | 100,0 | 15,6 | 15,0 | 7,0 |
| | Total | - | 38,8 | 95,0 | 100,0 | 17,8 | 94,6 | 13,6 |
| 1-NN | Igual | 8,0 | - | 0,6 | 0,0 | 4,0 | 8,6 | 8,6 |
| | Menor | 61,2 | - | 99,2 | 100,0 | 19,6 | 61,4 | 34,8 |
| | Total | 69,2 | - | 99,8 | 100,0 | 23,6 | 70,0 | 43,4 |
| ADL | Igual | 1,4 | 0,6 | - | 0,0 | 0,0 | 1,0 | 0,2 |
| | Menor | 5,0 | 0,2 | - | 99,8 | 0,8 | 5,4 | 1,6 |
| | Total | 6,4 | 0,8 | - | 99,8 | 0,8 | 6,4 | 1,8 |
| ADQ | Igual | 0,0 | 0,0 | 0,0 | - | 0,0 | 0,0 | 0,0 |
| | Menor | 0,0 | 0,0 | 0,2 | - | 0,0 | 0,0 | 0,0 |
| | Total | 0,0 | 0,0 | 0,2 | - | 0,0 | 0,0 | 0,0 |
| ADR | Igual | 2,2 | 4,0 | 0,0 | 0,0 | - | 2,0 | 3,2 |
| | Menor | 82,2 | 76,4 | 99,2 | 100,0 | - | 82,6 | 76,4 |
| | Total | 84,4 | 80,4 | 99,2 | 100,0 | - | 84,6 | 79,6 |
| NBN | Igual | 79,6 | 8,6 | 1,0 | 0,0 | 2,0 | - | 5,6 |
| | Menor | 5,4 | 30,0 | 93,6 | 100,0 | 15,4 | - | 6,8 |
| | Total | 85,0 | 38,6 | 94,6 | 100,0 | 17,4 | - | 12,4 |
| NBK | Igual | 6,6 | 8,6 | 0,2 | 0,0 | 3,2 | 5,6 | - |
| | Menor | 86,4 | 56,6 | 98,2 | 100,0 | 20,4 | 87,6 | - |
| | Total | 93,0 | 65,2 | 98,4 | 100,0 | 23,6 | 93,2 | - |

Tabela 6.2: Desempenho em termos das taxas de erro (em %) entre os classificadores para as séries do Robô SONY AIBO.

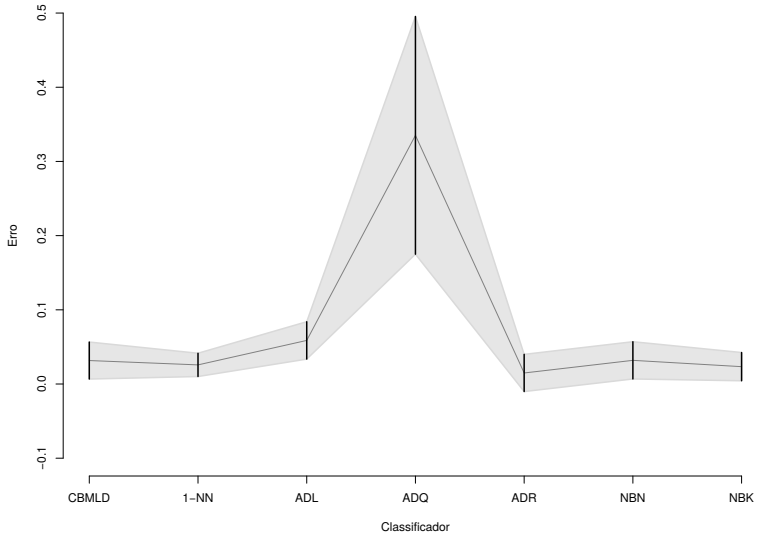


Figura 6.2: Comparação de Intervalos de Confiança das taxas de erro dos classificadores para as séries do Robô Sony AIBO.

relação ao segundo classificador representado na cor cinza.

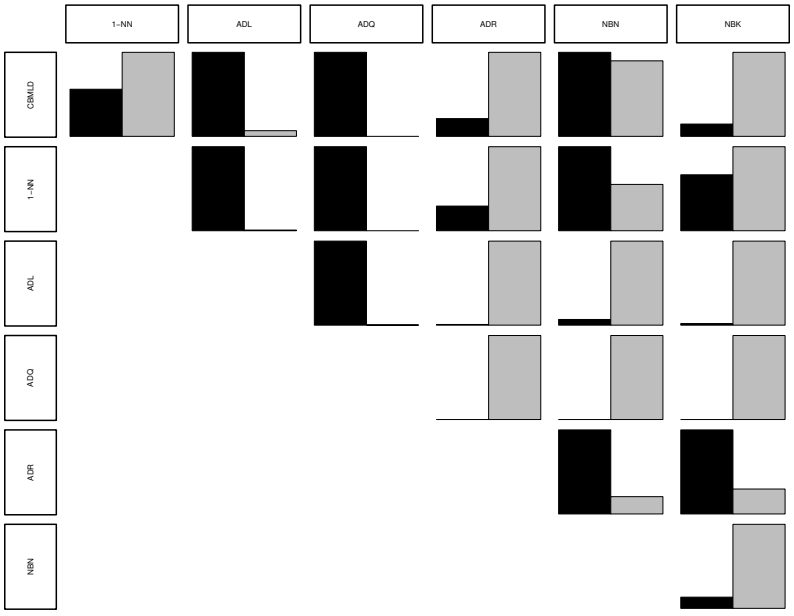


Figura 6.3: Comparação das taxas de erro (em %) entre os classificadores para as séries do Robô SONY AIBO.

## 6.2 Classificação de Tipos de Café

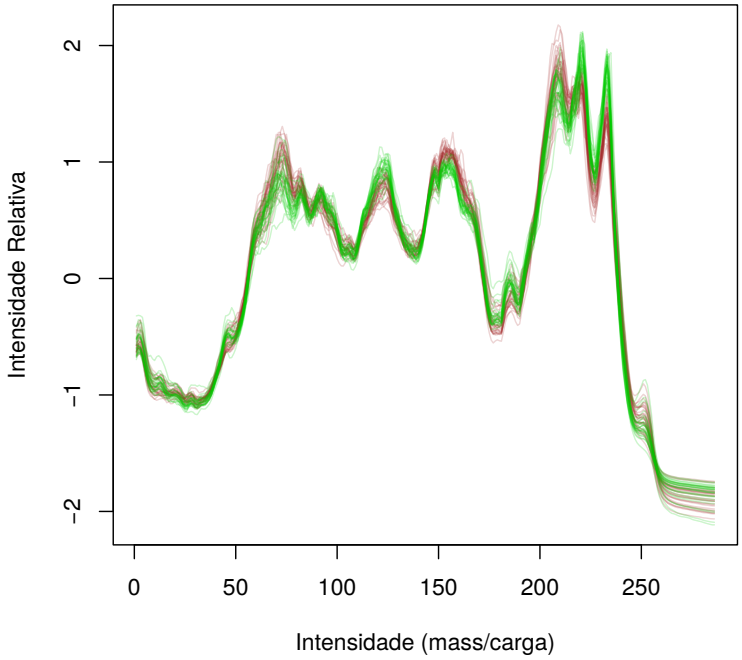


Figura 6.4: Espectro de massa das amostras de café, *Canephora* (marrom) e *Arabica* (verde)

As duas principais espécies de café cultivadas no mundo são a *Arábica* e a *Canephora*. Elas são diferentes em sabor, meio de cultivo e valor comercial, sendo a *Arábica* mais cara que a *Canephora*, embora esta última seja menos suscetível a doenças. Cinquenta e seis amostras de café desidratadas e congeladas foram analisadas através de espectrometria de massa, 29 da espécie *Canephora* e 27 da *Arábica*, todas as séries com 236 de comprimento.

A espectrometria de massa é uma técnica na qual moléculas de uma amostra são convertidas em íons em forma gasosa, e que são separados de acordo a razão de sua massa por sua carga. O resultado final é o espectro de massa - um gráfico que mostra a abundância de cada intensidade (massa/carga). Fazendo $y_t$ como sendo a abundância (também denominada intensidade relativa) observada na intensidade $t$, obtemos um espectro de massa como uma série temporal.

Para uma amostra de treino consistindo de 70% das séries originais ajustamos novamente um modelo polinomial de ordem 1. Em seguida, retiramos ao acaso uma nova amostra de treino e classificamos as restantes. Repetimos isto 500 vezes. Como o tamanho das séries é maior que o número de séries disponíveis, os classificados ADL e ADQ não foram considerados. Além disso, descartamos amostras de treino que tivessem menos

| Estatística | CBMLD | 1-NN | ADR | NBN | NBK |
|---|---|---|---|---|---|
| Média | 0,0000 | 0,02138 | 0,5153 | 0,0552 | 0,0782 |
| Desvio padrão | 0,0000 | 0,03380 | 0,07169 | 0,05463 | 0,0604 |

Tabela 6.3: Média e desvio padrão das taxas de erro (em %) dos classificadores para as séries dos tipos de café.

| | | CBMLD | 1-NN | ADR | NBN | NBK |
|---|---|---|---|---|---|---|
| CBMLD | Igual | - | 64,4 | 0,0 | 34,0 | 20,0 |
| | Menor | - | 35,6 | 100,0 | 66,0 | 80,0 |
| | Total | - | 100,0 | 100,0 | 100,0 | 100,0 |
| 1-NN | Igual | 64,4 | - | 0,0 | 36,2 | 26,6 |
| | Menor | 0,0 | - | 100,0 | 54,6 | 71,2 |
| | Total | 64,4 | - | 100,0 | 90,8 | 97,8 |
| ADR | Igual | 0,0 | 0,0 | - | 0,0 | 0,0 |
| | Menor | 0,0 | 0,0 | - | 0,0 | 0,0 |
| | Total | 0,0 | 0,0 | - | 0,0 | 0,0 |
| NBN | Igual | 34,0 | 36,2 | 0,0 | - | 41,2 |
| | Menor | 0,0 | 9,2 | 100,0 | - | 44,2 |
| | Total | 34,0 | 45,4 | 100,0 | - | 85,4 |
| NBK | Igual | 20,0 | 26,6 | 0,0 | 41,2 | - |
| | Menor | 0,0 | 2,2 | 100,0 | 14,6 | - |
| | Total | 20,0 | 28,8 | 100,0 | 55,8 | - |

Tabela 6.4: Desempenho em termos das taxas de erro (em %) entre os classificadores para as séries dos tipos de café.

de 10 séries em alguma das classes.

Os resultados da classificação dos tipos de café estão sumarizados nas Tabelas 6.3 e 6.4. Na Tabela 6.3, representada graficamente na Figura 6.5, verifica-se que o método CBMLD não apresentou erros de classificação, enquanto que o ADR apresentou o desempenho inferior aos demais classificadores. O classificador 1-NN obteve o segundo melhor desempenho, superior ao NBK e NBN, e estes com desempenhos próximos entre si. Os valores na Tabela 6.4 mostram que o o classificador 1-NN em 64,4% das vezes apresentou taxa de erro igual ao do CBMLD, ou seja, sem erro de classificação. Por outro lado, o ADR apresentou erro de classificação em todas as repetições. Na Figura 6.6, temos a representação gráfica do total do desempenho dos classificadores, onde o primeiro classificador é representado na cor preto r o segundo na cor cinza.

Os resultados obtidos com o CBMLD, neste problema de classificação dos tipos de café empregando dados de espectrometria de massa, nos indicam a relevância desta proposta de classificador. Esta afirmação se justifica uma vez que, como mencionado, o classificador 1-NN é considerado como o "padrão ouro"na literatura sobre classificação de séries temporais, e no entanto temos aqui um caso onde o classificador proposto é superior ao 1-NN.

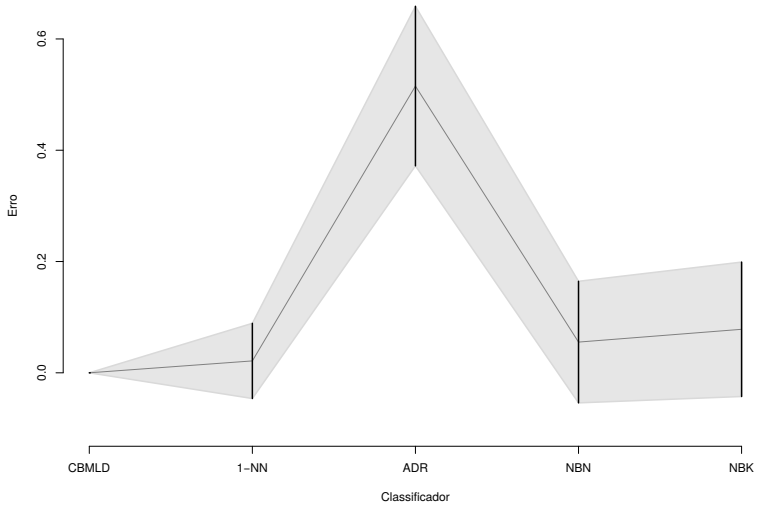


Figura 6.5: Comparação de Intervalos de Confiança das taxas de erro dos classificadores para as séries dos tipos de café.

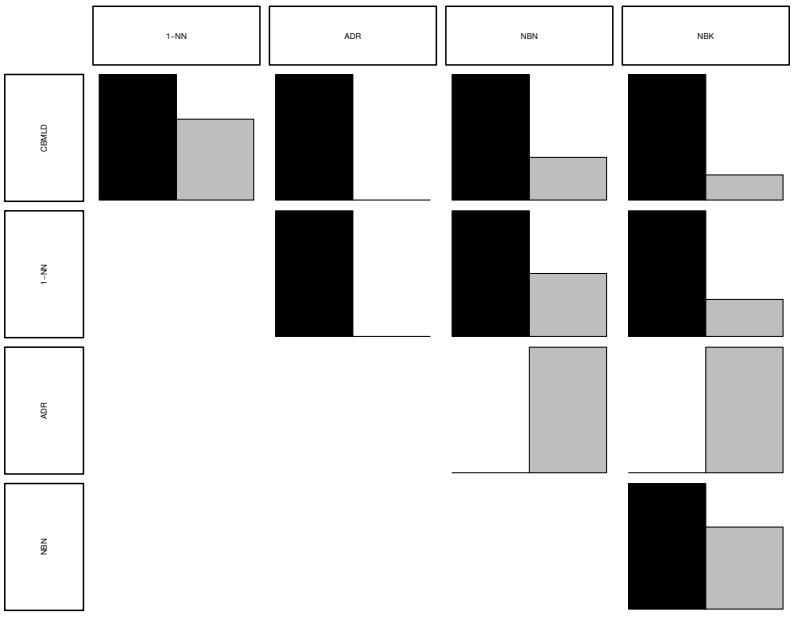


Figura 6.6: Comparação das taxas de erro (em %) entre os classificadores para as séries dos tipos de café.

## 6.3 Classificação de Folhas Suecas

Analisamos o conjunto de dados que denominamos por "Folhas Suecas"(do original, *Swedish Leaf*, em Chen et al. (2015)). Este conjunto de dados é composto de 1.125 imagens de folhas suecas divididas em 15 classes. Cada imagem foi convertida em uma 'pseudo série temporal' de comprimento 128, onde $y_t$ é a distância do $t$-ésimo ponto da borda da folha até o centroide da mesma. Na Figura 6.7, as etapas (a), (b) e (c) ilustram a construção das pseudos séries temporais através das medidas das distâncias euclidianas do centroide da folha até as suas bordas. Na Figura 6.8, é apresentada as séries temporais obtidas para as 15 classes.

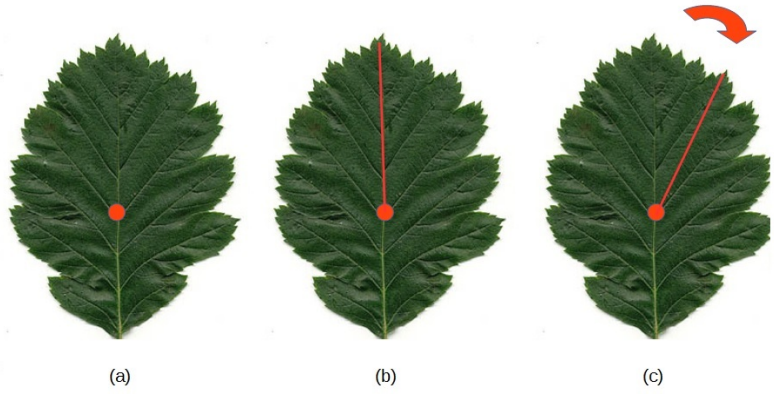


Figura 6.7: Etapas para obtenção das pseudo séries temporais para as folhas suecas.

Como nas aplicações anteriores, 70% das amostras das séries temporais foram selecionadas para compor o conjunto de treino, se cada classe tivesse pelo menos 35 séries temporais, construímos os classificadores e classificamos as séries restantes. Este procedimento foi realizado 300 vezes.

Após algumas análises para uma amostra de treino, identificamos um período sazonal igual a 128 e escolhemos os harmônicos para os modelos lineares dinâmicos trigonométricos:

- Classe 1: harmônicos 1, 2 e 3
- Classe 2: harmônicos 2, 3, 4, 5 e 6
- Classe 3: harmônicos 1, 2 e 3
- Classe 4: harmônicos 2 e 3
- Classe 5: harmônico 2 e 3
- Classe 6: harmônico 2

| Estatística | CBMLD | 1-NN | ADR | NBN | NBK |
|---|---|---|---|---|---|
| Média | 0,1698 | 0,1956 | 0,1744 | 0,1700 | 0,1529 |
| Desvio padrão | 0,01830 | 0,02073 | 0,02903 | 0,01839 | 0,01844 |

Tabela 6.5: Média e desvio padrão das taxas de erro (em %) dos classificadores para os tipos de folhas suecas.

- Classe 7: harmônico 2
- Classe 8: harmônicos 2 e 3
- Classe 9: harmônicos 1, 2 e 3
- Classe 10: harmônico 2
- Classe 11: harmônico 2
- Classe 12: harmônicos 1, 2, 3, 4, 5, 6 e 7
- Classe 13: harmônicos 2 e 3
- Classe 14: harmônico 2
- Classe 15: harmônicos 2 e 3.

A Tabela 6.5 mostra a média e o desvio padrão das taxas de erro para cada classificador e estão graficamente representados na Figura 6.9. Desta tabela, observa-se que as médias das taxas de erro são muito próximas (sem diferença significativa!), embora o método CBMLD tenha apresentado uma média ($0,1698\%$) inferior apenas ao do NBK ($0,1529\%$).

Na Tabela 6.6, observa-se a superioridade do CBMLD, principalmente, com relação ao 1-NN e ao ADR, uma vez que apresentou taxa de erro menor que a destes métodos em $82,94\%$ e $50,5\%$ das vezes nas repetições, respectivamente. Comparando com o NBK, que apresentou a menor média de taxa de erro, o CBMLD ainda apresentou taxa de erro menor em $9,36\%$ das vezes. Os resultados desta Tabela, no quesito total, estão ilustrados graficamente na Figura 6.10.

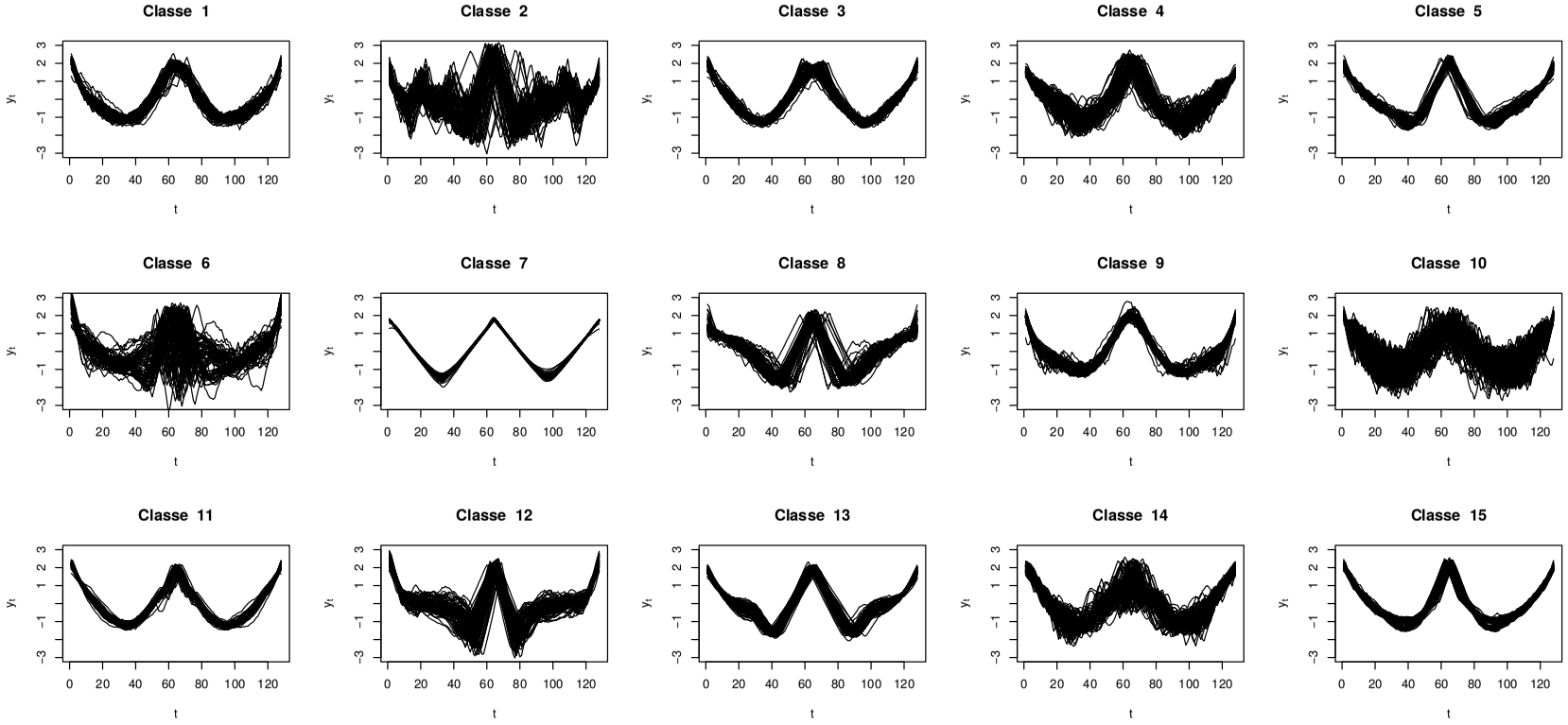


Figura 6.8: Pseudo séries temporais obtidas para os 15 tipos de folhas suecas.

| | | CBMLD | 1-NN | ADR | NBN | NBK |
|---|---|---|---|---|---|---|
| CBMLD | Igual | - | 2,34 | 3,01 | 51,50 | 3,01 |
| | Menor | - | 82,94 | 50,50 | 26,08 | 9,36 |
| | Total | - | 85,28 | 53,51 | 77,59 | 12,37 |
| 1-NN | Igual | 2,34 | - | 2,34 | 02,67 | 0,66 |
| | Menor | 14,71 | - | 22,40 | 14,04 | 2,67 |
| | Total | 17,05 | - | 24,74 | 16,72 | 3,34 |
| ADR | Igual | 03,01 | 2,34 | - | 5,01 | 2,67 |
| | Menor | 46,48 | 75,25 | - | 45,81 | 22,74 |
| | Total | 49,49 | 77,59 | - | 50,83 | 25,41 |
| NBN | Igual | 51,50 | 2,67 | 5,01 | - | 5,01 |
| | Menor | 22,40 | 83,27 | 49,16 | - | 8,69 |
| | Total | 73,91 | 85,95 | 54,18 | - | 13,71 |
| NBK | Igual | 03,01 | 0,66 | 2,67 | 5,01 | - |
| | Menor | 87,62 | 96,65 | 74,58 | 86,28 | - |
| | Total | 90,63 | 97,32 | 77,25 | 91,30 | - |

Tabela 6.6: Desempenho em termos das taxas de erro (em %) entre os classificadores para as séries dos tipos de folhas suecas

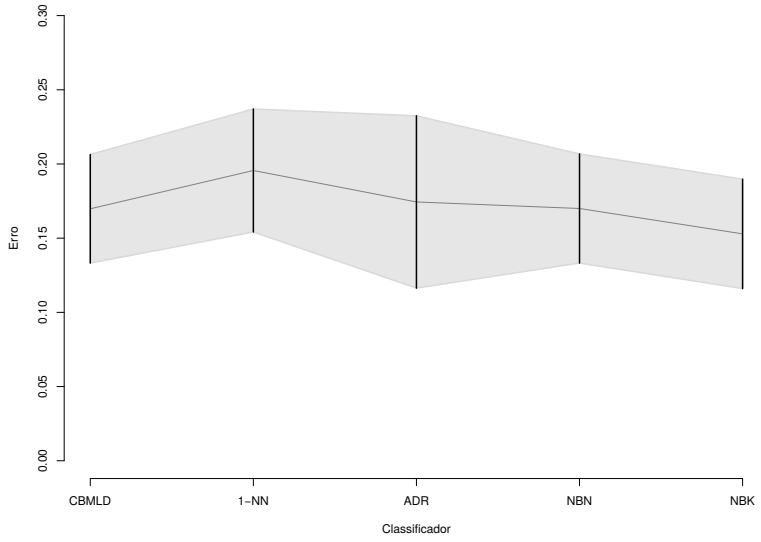


Figura 6.9: Comparação de Intervalos de Confiança das taxas de erro dos classificadores para as séries dos tipos de folhas suecas.


Figura 6.10: Comparação das taxas de erro (em %) entre os classificadores para as séries dos tipos de folhas suecas.

# Capítulo 7

# Considerações Finais

Neste trabalho, apresentamos uma nova abordagem para Análise Discriminante (AD) de séries temporais, propondo uma versão para classificador de Bayes empregando Modelos Lineares Dinâmicos, que denotamos por CBMLD. Nos estudos realizados, de simulação computacional e aplicações em conjuntos de dados reais, o classificador proposto apresentou um bom desempenho em comparação com os classificadores mais usuais, paramétricos e não paramétricos, como a Análise Discriminante Linear (ADL), Análise Discriminante Quadrática (ADQ), Análise Discriminante Regularizada (ADR), Naive Bayes com Distribuição Normal (NBN), Naive Bayes com Estimadores por Função Núcleo (NBK) e o Classificador por Vizinho mais Próximo (1-NN).

Realizamos estudos de simulação com modelos simples que são úteis na prática. Tais estudos, embora não exaustivos, sugerem que o CBMLD se configura como uma proposta eficiente de classificador, desde que seja utilizado o Modelo Linear Dinâmico adequado. Nestes estudos de simulação realizados, estabelecemos comparações entre algumas estratégias de estimação da variância ou matriz de covariâncias: $V_t$ estimado em cada estante $t$, e $\mathbf{W}$ elicitado através de fator de desconto $(\tilde{V}, \delta)$; $V_t$ considerado fixado e $\mathbf{W}$ elicitado através de fator de desconto $(\phi, \delta)$; $V_t$ estimado em cada estante $t$, e $\mathbf{W}_t$ considerando $h$ modelos superpostos $(\tilde{V}, \boldsymbol{W})$. Os estudos de simulação realizados com estas estratégias, analisando as taxas de erro do classificador, indicaram a estratégia $(\tilde{V}, \boldsymbol{W})$ como a que apresentou melhores resultados.

Como desvantagem para o CBMLD, podemos citar o custo computacional na fase de ajuste (treinamento do classificador) empregando a validação cruzada. No processo de validação cruzada o classificador é estimado tantas vezes quanto o número de observações do conjunto de treinamento, isto significa que a matriz $\mathbf{W}$ deve ser estimada em todas estas repetições o que representa um elevado custo computacional. No entanto, ajustado os parâmetros do modelo no CBMLD, este classificador pode ser empregado na prática para diversos problemas cujas observações sejam oriundas de séries temporais.

De modo geral, considerando as séries temporais reais analisadas neste trabalho, podemos afirmar que o CBMLD se mostrou competitivo frente aos resultados dos classifi-

cadores $1-NN$, *ADR*, *NBN* e *NBK*, e superior aos métodos *ADL* e *ADQ*. É importante notar que tanto o $1-NN$ e o *NBK*, que são reconhecidos na literatura como classificadores eficientes para problemas em AD, estes métodos são não paramétricos, cujo emprego na classificação de novas observações exige sempre todas as observações do conjunto de treinamento, o que limita suas aplicações em situações que exijam classificadores para serem empregados em tempo real. Desta forma, a competitividade demonstrada pelo CBMLD, o credencia como uma alternativa para esta classe de problemas.

Existem certas abordagens, estratégias e conjecturas que propomos como trabalhos futuros, tais como:

- **Estudos de outras estratégias para estimação de variância**. Em particular, destacamos:

  1. O uso de métodos do tipo MCMC para estimação de $\boldsymbol{W}^{(i)}$.
  2. Estudar estruturas alternativas de regularização para $\boldsymbol{W}_t^{(i)}$ (como, por exemplo, estruturas semelhantes as do método ADR).

- **Predição da classe sem observar a série toda**. Por exemplo, no caso do Robô SONY AIBO, a classificação é realizada após o robô completar uma volta. Contudo, é mais interessante detectar o tipo de superfície em tempo real.

- **Estruturas mais complexas**. Muitas séries apresentadas em Chen et al. (2015) possuem comportamentos mais complexos e seria interessante ter uma noção da performance do CBMLD para todas elas.

# Capítulo 8

# Apêndice

**Lema 1.** *Transformações lineares: se A é uma matriz $r \times p$ e b é um vetor r-dimensional então*

$$y = Ax + b \sim N[A\mu + b, A\Sigma A^T]$$

**Lema 2.** *Distribuições Marginais: se o vetor $\boldsymbol{x}$ é dividido em 2 blocos $\boldsymbol{x}_1$ contendo os primeiros r componentes de x e $\boldsymbol{x}_2$ contendo os outros $p-r$ componentes então procedendo a mesma partição em $\mu$ e $\Sigma$ na forma,*

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} e\ \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

*obtém-se que $x_i \sim N[\mu_i, \Sigma_{ii}],\ i = 1,2$*

**Lema 3.** *Reconstrução da Conjunta: Se $\boldsymbol{x}_1|\boldsymbol{x}_2 \sim N[\mu_1 - B_1(x_2 - \mu_2), B_2]$ e $\boldsymbol{x}_2 \sim N[\mu_2, \varepsilon_{22}]$ então*

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix} \sim N[\mu, \Sigma]$$
$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} e\ \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

*onde $\Sigma_{11} = B_2 + B_1\Sigma_{22}B_1^T$ e $\Sigma_{21}^T = \Sigma_{12} = B_1\Sigma_{22}$.*

**Lema 4.** *Para A,B e C eventos aleatórios em $(\omega.\mathbf{A}, P)$ tal que $A \perp (C|B)$, então*

$$\begin{aligned} P(A|B \cap C) &= \frac{P(A \cap B \cap C)}{P(B \cap C)} \\ &= \frac{P(A \cap C|B)P(B)}{P(B \cap C)} \\ &= \frac{P(A|B)P(C|B)P(B)}{P(B \cap C)} \\ &\propto P(A|B)P(C|B)P(B) \end{aligned}$$

**Definição 10.** *(Matrizes Particionadas) Uma matriz está na forma particionada ou na forma de blocos quando está é expressa através de suas submatrizes.*

*Por exemplo, considere a seguinte partição da matriz $A_{m\times n}$, nas submatrizes ou blocos* $\left\{\boldsymbol{A}_{11}^T, \boldsymbol{A}_{12}^T, \boldsymbol{A}_{21}^T, \boldsymbol{A}_{22}^T\right\}$

$$\boldsymbol{A}_{m\times n} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{pmatrix} \tag{8.1}$$

$$= \begin{pmatrix} \boldsymbol{A}_{11}^T & \boldsymbol{A}_{12}^T \\ \boldsymbol{A}_{21}^T & \boldsymbol{A}_{22}^T \end{pmatrix}. \tag{8.2}$$

*Onde,*

$$\boldsymbol{A}_{11}^T = (a_{11})$$

$$(\boldsymbol{A}_{12})^T = \begin{pmatrix} a_{12} \\ a_{13} \\ \vdots \\ a_{1n} \end{pmatrix}^T$$

$$\boldsymbol{A}_{21}^T = \begin{pmatrix} a_{21} \\ a_{31} \\ \vdots \\ a_{m1} \end{pmatrix}$$

$$\boldsymbol{A}_{22}^T = \begin{pmatrix} a_{22} & a_{23} & \dots & a_{2n} \\ a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m2} & a_{m3} & \dots & a_{mn} \end{pmatrix}$$

*Podemos assim utilizar a partição de matrizes de forma conveniente denotando cada um dos seus blocos de acordo com a necessidade do problema.*

# Referências Bibliográficas


A. Bagnall, L. M. Davis, J. Hills, and J. Lines. Transformation based ensembles for time series classification. In *SDM*, volume 12, pages 307–318. SIAM, 2012.

P. J. Bickel and E. Levina. Some theory for Fisher's linear discriminant function,'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, pages 989–1010, 2004.

Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista. The ucr time series classification archive, July 2015. `www.cs.ucr.edu/~eamonn/time_series_data/`.

D.-Q. Dai and P. C. Yuen. Regularized discriminant analysis and its application to face recognition. *Pattern Recognition*, 36(3):845–847, 2003.

P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, 29:103–130, 1997.

R. O. Duda, P. E. Hart, and S. D. G. *Pattern Classification. Second Edition*. Wiley-Interscience, 2000.

J. H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

J. Harrison and M. West. *Bayesian Forecasting & Dynamic Models*. Springer, 1999.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction. Second Edition*. Springer, 2009.

A. J. Izenman. *Modern Multivariate Statistical Techniques. Regression, Classification, and Manifold Learning*. Springer, 2008.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

G. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, 2004.

G. Petris, S. Petrone, and P. Campagnoli. *Dynamic Linear Models with R*. Springer Science & Business Media, 2009.

D. Vail and M. Veloso. Learning from accelerometer data on a legged robot. In *Proceedings of the 5th IFAC/EURON Symposium on Intelligent Autonomous Vehicles*, 2004.

D. M. Witten and R. Tibshirani. Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):615–636, 2009.