Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação - Doutorado em Estatística

Tese de Doutorado

Inferência em Alguns Modelos de Processos Estocasticamente Perturbados

Aluno: Wecsley Otero Prates[1]

Belo Horizonte, Junho de 2016

[1] Universidade Federal de Minas Gerais

Inferência em Alguns Modelos de Processos Estocasticamente Perturbados

Wecsley Otero Prates

Tese de Doutorado apresentada ao Departamento de Estatística da

Universidade Federal de Minas Gerais

Programa: Pós-Graduação em Estatística
Orientadora: Prof[a]. Dr[a] Denise Duarte[2]
Co-Orientador: Prof. Dr. Sokol Ndreca[2]

Belo Horizonte, Maio de 2016

[2]Universidade Federal de Minas Gerais

Esta tese foi realizada no Instituto de Ciências Exatas do Departamento de Estatística da Universidade Federal de Minas Gerais, sob a orientação da Professora Doutora Denise Duarte e pela co-orientação do Professor Doutor Sokol Ndreca e contou nos primeiros 3 anos com o financiamento da Bolsa REUNI.

Agradecimentos

A preocupação, o medo e as incertezas são assuntos complexos e estão associados com tantas de nossas ansiedades e problemas que é impossível separá-los completamente.

Agradeço primeiramente a Deus que me deu força, coragem e persistência para cansar as adversidades que encontrei pelo caminho.

Agradeço a Profª e amiga Denise Duarte pela sua orientação. A sua generosidade e entusiasmo foram grandes incentivos nestes anos.

Agradeço ao Profº e Co-orientador Sokol Ndreca pelo seu apoio e suas idéias para o desenvolvimento deste trabalho. Sou grato aos colegas Paulo Cerqueira, Rodrigo Citton, Luis Gustavo, Rodolfo Lorenzutti e Silvio Souza, que posso chamá-los de amigos e aprendi de cada um deles lições de responsabilidade e companherismo, que muitas vezes me ajudaram com esta tese na parte computacional.

Aos meus Professores que dos conhecimentos que me foram passados, o mais importante é que faríamos muitas coisas se não as julgássem-mos muitas vezes impossíveis. Aos amigos que conquistei nessa jornada e que foram importantes para mais essa etapa.

Agradeço também a minha esposa Carolina Mulek Prates pelo apoio, paciência e por sempre estar ao meu lado nos momentos de dificuldades.

E um grande agradecimento à minha família que me deu amor, apoio e me encorajaram a realizar mais um sonho.

Resumo

Em um modelo de processos estocasticamente perturbados as observações do processo original podem sofrer perturbações, em cada instante de tempo, por um ruído aleatorio. Dessa forma, o processo observado pode não ser mais uma amostra do processo original.

Nesta tese apresentamos metodologias para fazer estimacão dos parâmetros de alguns modelos estocasticamente perturbados tendo como base os modelos propostos por [7] e [12]. Assumimos que o processo original, oculto, é uma cadeia de Markov de alcance variável. Essa classe de processos permite muitas aplicações por ser parcimoniosa em relacão ao número de parâmetros e também bastante maleável, englobando a classe das cadeias de Markov de ordem fixa.

Propomos uma adaptação no algoritmo de Baum-Welch e um estimador BIC bootstrap para os parâmetros dos modelos analisados, cuja convergência foi demonstrada, e através de simulações, mostramos que a metodologia proposta é capaz de recuperar muito bem a verdadeira árvore de contextoss de uma cadeia de Markov com alcande variável estocasticamente perturbada, assim como as probabilidades de transição associadas a essa árvore, dentro de um intervalo de níveis de perturbação. Também conseguimos recuperar o grau de perturbação qualquer que tenha sido.

Propomos uma modificação no algoritmo de Viterbi para encontrar a sequência oculta mais provável de uma cadeia de Markov com alcande variável estocasticamente perturbada.

Apresentamos um critério de seleção de modelos para identificar o modelo mais adequado, dada uma amostra observada, dentre os analisados nessa tese.

Aplicamos a metodologia proposta a um banco de dados de registros de atividade de neurônios de um grupo de corujas em um experimento controlado em laboratório. Os dados foram codificados em 2 estados, disparo e repouso, e o nosso objetivo é identificar a existência de diferentes padrões de comportamentos dessa atividade neuronal, de acordo com a lei de probabilidades estimada para o processo, em relação ao tipo de estímulo visual a que o grupo de corujas foi submetido.

Palavras-chave: Processos perturbados, Cadeias de Markov de Alcance Variável, árvore de contextos, Algoritmo BIC, Bootstrap, Algoritmo de Baum-Welch.

Abstract

In a model of stochastically disturbed processes each observation of the original process can be disturbed at any moment of time by a random noise. Thus the observed process could not be a sample of the original process.

In this thesis we present a methodology in order to estimate the parameters of some disturbed stochastically models based on the models proposed by [7] and [12]. We assume that the original hidden process is a variable length Markov chain. This class processes allows many applications since it is parsimonious in relation to the number of parameters and also quite malleable, including the class of fixed-order Markov chains. We propose an adaptation in the Baum-Welch algorithm and a bootstrap Bayesian Information Criterion as a way to estimate the parameters of the models analyzed, whose convergence was shown, and show through simulations that the proposed methodology is able to recover very well the real context tree of a stochastically disturbed variable length Markov chain as well as the transition probabilities associated with the tree, within a reasonable range of disturbance levels. We also able to recover the degree of disturbance whatever it has been.

We propose a modification to the Viterbi algorithm to find the most appropriate hidden sequence of a stochastically disturbed variable length Markov chain.

We present a model selection criterion to identify the most appropriate model given the observed sample among those analyzed in this thesis.

We apply the proposed methodology to a database of neurons activity records of a group of owls in a controlled laboratory experiment. Data were coded in two states, spike and rest. Our goal is to identify the existence of different patterns of behavior that neuronal activity according to the estimated probability for the process in relation to the type of visual stimulus that the group of owls was submitted.

Keywords: Disturbed Process, Variable Length Memory Chains, Context tree, Boootstrap, Bic algorithm, Baum-Welch algorithm.

Índice

Lista de Símbolos

$\boldsymbol{X}$ - VLMC Oculta
$\boldsymbol{Y}$ - VLMC Oculta
$\boldsymbol{Z}$ - Processo Perturbado
$\boldsymbol{X^*}$ - Cadeia de Markov de Ordem k
$\boldsymbol{Z^*}$ - Processo Perturbado Transformado
$\boldsymbol{\xi}$ - Sequência de Variáveis Aleatórias
ϵ - Parâmetro de Perturbação
a_t - Estado Oculto de $\boldsymbol{X}$ no Tempo t
c_t - Estado Oculto de $\boldsymbol{Y}$ no Tempo t
b_t - Valor da Variável $\boldsymbol{\xi}$ no Tempo t
z_t - Símbolo Observado de $\boldsymbol{Z}$ no Tempo t
ω - Contexto
ν - Contexto
υ - Contexto
$\mathcal{T}$ - Árvore de contextos
$\mathcal{T}_k$ - Árvore de contextos $k - full$
$\mathcal{T}|_k$ - Árvore de contextos Truncada na Ordem k
$\hat{\mathcal{T}}$ - Árvore de contextos Estimada
$\hat{\mathcal{T}}_k$ - Árvore de contextos $k - full$ Estimada
$\hat{\mathcal{T}}|_k$ - Árvore de contextos Truncada na Ordem k Estimada
X - Amostra de $\boldsymbol{X}$
X^* - Amostra de $\boldsymbol{X^*}$
Z - Amostra de $\boldsymbol{Z}$
Z^* - Amostra do Processo Perturbado Transformado $\boldsymbol{Z^*}$
$\hat{X}$ - Amostra Bootstrap de $\boldsymbol{X}^*$
$\mathbf{A}$ - Matriz de Transição de $\boldsymbol{X}$
$\mathbf{A}^*$ - Matriz de Transição de $\boldsymbol{X^*}$
$\hat{\mathbf{A}}^*$ - Matriz de Transição Estimada de $\boldsymbol{X^*}$
$\mathbf{B}$ - Distribuição de Emissão entre $\boldsymbol{X}$ e $\boldsymbol{Z}$
$\mathbf{B}^*$ - Distribuição de Emissão entre $\boldsymbol{X^*}$ e $\boldsymbol{Z^*}$
$\hat{\mathbf{B}}^*$ - Distribuição de Emissão Estimada entre $\boldsymbol{X^*}$ e $\boldsymbol{Z^*}$
$p(a|\omega)$ - Probabilidade de Transição de $\boldsymbol{X}$
$\hat{p}(a|\omega)$ - Probabilidade de Transição Estimada de $\boldsymbol{X}$
$p^*(\omega|\nu)$ - Probabilidade de Transição de $\boldsymbol{X^*}$
$\hat{p}^*(\omega|\nu)$ - Probabilidade de Transição Estimada de $\boldsymbol{X^*}$
$p(a|\nu)$ - Probabilidade de Transição de $\boldsymbol{Y}$
$b_\omega(z_t)$ - Elemento de $\mathbf{B}$
$b_\omega(z_t)^*$ - Elemento de $\mathbf{B}^*$
$\hat{b}_\omega(z_t)^*$ - Elemento Estimado de $\hat{\mathbf{B}}^*$
$\boldsymbol{\pi}$ - Distribuição Inicial de $\boldsymbol{X}$
$\boldsymbol{\pi}^*$ - Distribuição Inicial de $\boldsymbol{X^*}$
$(\boldsymbol{Z}, \boldsymbol{X})$ - Cadeia de Markov de Alcance Variável Oculta
$\boldsymbol{\lambda}$ - Vetor de Parâmetros de $(\boldsymbol{Z}, \boldsymbol{X})$
$(\boldsymbol{Z^*}, \boldsymbol{X^*})$ - Cadeia de Markov Oculta
$\boldsymbol{\lambda^*}$ - Vetor de Parâmetros de $(\boldsymbol{Z^*}, \boldsymbol{X^*})$

Capítulo 1

Introdução

Esta tese aborda a questão de inferir se uma amostra observada foi, de fato, gerada por um determinado processo estocástico ou se essa amostra foi perturbada por um ruído aleatório. No caso onde o processo original que pode ter sofrido a perturbação é uma cadeia de Markov, esse modelo é bastante conhecido na literatura como Modelo de Markov Oculto, HMM[1] introduzido em 1966 por [2] e tem uma grande quantidade de trabalhos dedicados a esse tipo de modelagem devido a sua importância e aplicações, tais como em machine learning, genética, reconhecimento de voz, etc ([20] e [21]).

Analisaremos esse problema considerando que o processo oculto original pertence a uma grande classe de processos onde a ordem de dependência no passado não é fixa, o que não acontece em uma cadeia de Markov. A questão que queremos responder é: Dada uma amostra de símbolos observados de um processo estocástico é possível saber se amostra está ou não perturbada por algum ruído aleatório? Através dessa amostra perturbada, é possível mensurar o grau de perturbação dessa amostra? E ainda descobrir a verdadeira fonte da qual os dados foram gerados, antes de terem sido perturbados? É possível recuperar a lei original dos dados para qualquer que seja o grau de perturbação?

Modelos com tais características são chamados na literatura de Modelos de Markov Oculto de Alcance Variável (VLHMM). Os VLHMM[2] apareceram pela primeira vez, segundo [11], na análise do movimento corporal humano, como pode ser visto em [20] e [21]. Em [21], o autor analisa o movimento 3D através da rotação de 19 grandes articulações do corpo humano, e [20] em seguida usa uma representação VLHMM em que a cadeia de Markov de alcance variável (VLMC)[3] oculta é a pose no tempo n e os dados observados são as posições do corpo dadas pelas rotações 3D dos 19 pontos principais. Eles argumentam que VLHMM é superior em eficiência e precisão na modelagem multivariada em séries temporais com alta variedade dinâmica.

Existem alguns trabalhos anteriores que analisam esses modelos com perturbação do ponto de vista teórico, e que tomamos como ponto de partida. Em [7], os autores descrevem um processo estocástico perturbado como sendo uma função da fonte original e de um ruído aleatório. Eles supõem que a fonte original é uma cadeia com ordem infinita, assumindo valores em um alfabeto binário e que pode sofrer perturbações por um ruído aleatório Bernoulli independente da fonte original. Em [12], os autores consideram que a fonte original é uma VLMC, onde cada símbolo é multiplicado por um ruído aleatório Bernoulli, também independente da fonte original. Eles chamaram esse modelo com Modelo perturbado Inflacionado de Zeros. Nesse segundo trabalho também é considerado o caso em que a amostra observada pode ter sido gerada de uma mistura de processos com ordem variável.

Em ambos os trabalhos os autores mostraram que se o ruído aleatório Bernoulli for pequeno, então a amostra perturbada pode ser usada para estimar a matriz de transição do processo original. Eles mostraram que a diferença entre as probabilidades de transição do processo perturbado e do processo original é limitado por uma constante c, em que c é uma função linear do ruído aleatório Bernoulli (mais detalhes em [7] e [12]). Porém, se o ruído aleatório não for pequeno suficiente, então a aproximação das probabilidades de transição oculta pelas probabilidades de transição estimadas do processo perturbado não será satisfatória, uma vez que, segundo os autores, a medida que

[1] HMM é a sigla em inglês para Hidden Markov Model
[2] VLHMM é a sigla em inglês para Variable Length Hidden Markov Model
[3] VLMC é a sigla em inglês para Variable Length Markov Chain

o ruído aleatório aumenta, a constante c que limita a diferença entre as probabilidades de transição verdadeiras e estimadas pelo processo contamindo também aumenta. Portanto é crucial estimar o parâmetro de perturbação, a fim de saber se tal aproximação pode ser aplicada ou não. Mas, os autores não abordaram esse problema de estimação dos parâmetros do modelo.

Um importante resultado em estimação de parâmetros para uma classe de modelos perturbados é apresentado em [11]. A classe de modelos discutida nesse artigo é bastante abrangente, uma vez que permite que o ruído aleatório seja proveniente de uma variedade maior de distribuições, mas é restritiva em relação a distribuição condicional entre o processo observado e o original, ou seja, apenas o último símbolo oculto no passado do processo oculto é considerado nas distribuições condicionais, enquanto que nessa tese consideramos que essa dependência pode ser um contexto. O autor propõe um estimador baseado em uma função de verossimilhança penalizada, assim como no Critério de Informação Bayesiana (BIC)[4] proposto em [9], mas de acordo com o próprio autor, os resultados empíricos mostraram que o algoritmo com a penalização do BIC é mais eficiente do que o proposto no artigo. O autor mostra que o estimador proposto, com essa outra penalização, é fortemente consistente.

Nessa tese apresentamos estimadores consistentes para a árvore de contextos, associada a VLMC oculta, e para o parâmetro de perturbação dos processos perturbados como descritos em [7] e [12]. A simplicidade desses modelos nos permite aplicar um algoritmo EM para obter os estimadores.

Além disso, apresentamos um estudo de sensibilidade dos estimadores para verificar o comportamento dos estimadores propostos na medida em que o nível de perturbação aumenta. Nosso objetivo com essa análise de sensibilidade é saber se existe um intervalo de níveis de perturbação em que o procedimento de estimação é mais eficiente.

Apresentamos também um critério de seleção a fim de escolher, entre modelos perturbados discutidos, qual é o mais apropriado para uma dada amostra perturbada.

Como aplicação da nossa metodologia a dados reais realizamos uma análise de um banco de dados muito interessante que nos foi gentilmente cedidos pelo Laboratório de Neurofisiologia da Visão da UFMG, coordenado pelo Dr Jerome Baron. Nesse banco de dados corujas são submetidas a estimúlos visuais em um experimento controlado e as respostas de neurônios a esses estímulos foram medidas. Essas respostas neuronais são chamadas de "spikes"que podem ser consideradas como "disparos"dos neurônios. Devido ao fato de esses disparos poderem ser erroneamente medidos, por razões técnicas, consideramos que a sequência de disparos dos neurônios observados no tempo pode ser modelada como um processo estocástico que pode ter sofrido uma perturbação por um ruído aleatório. Os resultados obtidos com a metodologia aqui proposta são bem interessantes e coerentes com o que se esperava encontrar.

Esta tese está organizada da seguinte maneira: no capítulo 2 apresentamos as notações básicas e algumas definições preliminares de metodologias conhecidas como o algoritmo de Baum-Welch e Verossimilhança perfilada, faremos revisões dos modelos já propostos por [7] e [12]. No capítulo 3 são apresentados os modelos propostos nessa tese e alguns resultados encontrados para os modelos propostos. No capítulo 4 é mostrado um dos modelos propostos por [12], no qual mostramos alguns resultados e apresentamos uma proposta de estimação dos parâmetros para esse modelo. No capítulo 5 são apresentados as propostas de algoritmos e estimadores para os modelos em questão, juntamente com uma versão do algoritmo de Viterbi para VLMC. No capítulo 6 apresentamos um estudo de simulação e sensibilidade para o ruído aleatório para alguns modelos em questão. No capítulo 7 apresentamos dois critérios de seleção de modelos, afim de decidir qual de dois modelos estudados é o mais adequado para modelar uma sequência de símbolos, dada uma amostra observada. No capítulo 8 apresentamos uma aplicação. No capítulo 9 apresentamos conclusões gerais a respeito da tese realizada. No capítulo 10 abordamos limitações dos modelos abordados e trabalhos futuros. E no Apêncice apresentamos as provas relacionadas aos resultados encontrados no capítulo 3 e 4 e apresentamos uma breve definição sobre estimação dos parâmetros de um modelo usando Verossimilhança Perfilada, que é uma das metodologias de estimação dos parâmetros do modelo proposto por [12] apresentado no capítulo 4.

[4] BIC é uma sigla em inglês para denotar Bayesian Information Criteria

Capítulo 2

Notações e Definições

Considere o alfabeto discreto finito $E = \{0, 1, ..., N-1\}$ com cardinalidade de $|E| = N$. Dados dois inteiros $m, n \in \mathbb{Z}$, com $m \leq n$, usaremos a notação ω_m^n para denotar a sequência $(\omega_m, ..., \omega_n)$ de símbolos em E, e seja $E^{l(\omega_m^n)}$ o conjunto que contém tais sequências onde $l(\omega_m^n) = |n - m + 1|$ é o comprimento da sequência ω_m^n. Uma sequência vazia é denotada por $\emptyset$ e $l(\emptyset) = 0$.

A concanetação das sequências ω e υ consiste dos símbolos de ω seguidos pelos símbolos de ν. Dadas duas sequências ω e υ, tal que $l(\omega) < \infty$, denotamos por $\upsilon\omega$ a sequência de comprimento $l(\upsilon) + l(\omega)$ obtida pela concatenação dessas duas sequências. A concatenação pode ser extendida para o caso quando as sequências são semi-infinitas $\upsilon = ...\omega_{-2}\omega_{-1}$.

Dizemos que a sequência ν é um sufixo da sequência ω se existe uma sub-sequência η, com $l(\eta) \geq 1$, tal que $\omega = \eta\nu$ e denotamos $\nu \preceq \omega$, e se ν é um sufixo próprio de ω escrevemos $\nu \prec \omega$.

2.1 Cadeia de Markov Oculta com Alcance Variável

Considere $\boldsymbol{X} = \{X_t\}_{t\in\mathbb{Z}}$ um processo estacionário ergódico no alfabeto discreto E. Dada uma sequência $\omega \in E^\infty$ e um símbolo $a \in E$, denotamos

$$p(a|\omega) := P(X_0 = a | X_{-1} = \omega_{-1}, X_{-2} = \omega_{-2}, ...)$$

como as probabilidades de transição do processo $\boldsymbol{X}$. E para uma sequência finita $\omega \in E^j$, denotamos

$$p(\omega) := P(X_{-j}^{-1} = \omega).$$

Definição 2.1. *Uma sequência finita $\omega \in \cup_{j=1}^{\infty} E^j$ é um contexto de $\boldsymbol{X}$ se satisfaz:*

i) *Para toda sequência semi-infinita $x_{-\infty}^{-1}$ com ω como sendo um sufixo,*

$$P\left(X_0 = a | X_{-\infty}^{-1} = x_{-\infty}^{-1}\right) = p(a|\omega) > 0 \tag{2.1}$$

para todo $a \in E$.

ii) *Nenhum sufixo próprio de ω satisfaz (2.1).*

Um contexto infinito é uma sequência semi-infinita $\omega_{-\infty}^{-1}$ tal que nenhum sufixo $\omega_{-j}^{-1}, j = 1, 2, ...$ é um contexto.

Definição 2.2. *O conjunto $\mathcal{T}$ de contextos é chamado árvore de contextos se nenhum $\omega_1 \in \mathcal{T}$ é um sufixo próprio de algum outro $\omega_2 \in \mathcal{T}$. Devido á condição ii) a árvore de contextos é chamada de irredutível.*

Cada contexto $\omega \in \mathcal{T}$ pode ser visto como um caminho de uma folha até a raiz (veja Figura 2.1). Os galhos da árvore $\mathcal{T}$ são identificados pelos contextos (finito ou infinito) $\omega \in \mathcal{T}$, a raiz é o contexto vazio $\emptyset$.

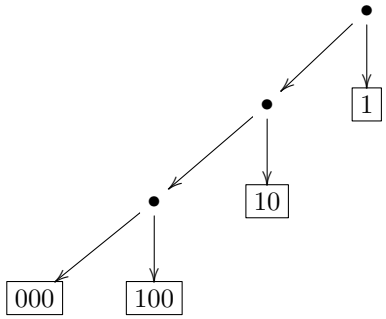


Figura 2.1. *árvore de contextos $\mathcal{T}$ com $k = 3$*

A Figura 2.1 mostra uma árvore de contextos de ordem 3 que assume valores em um alfabeto $E = \{0, 1\}$.

Definição 2.3. *Uma árvovre $\mathcal{T}$ é completa se cada nó interno tem $|E|$ galhos (ou filhos).*

Definição 2.4. *Uma árvore $\mathcal{T}$ é chamada L-full se $l(\omega) = L, \forall \omega \in \mathcal{T}$.*

Denotamos a profundidade da árvore $\mathcal{T}$, $d(\mathcal{T}) := \max\{l(\omega) : \omega \in \mathcal{T}\}$
Um processo estocástico estacionário $\boldsymbol{X}$ em E é uma VLMC compatível com o par $(\mathcal{T}, p(a|\omega))$ se satisfaz a Definição 2.1.

Definição 2.5. *Dado um inteiro k, definimos a árvore truncada $\mathcal{T}\big|_k$ de ordem k por*

$$\mathcal{T}\big|_k := \{\omega \in \mathcal{T} : \ l(\omega) \leq k\} \bigcup \{\omega : l(\omega) = k \ \ e \ \omega \prec v, \ para \ algum \ \ v \in \mathcal{T}\}.$$

Dada uma amostra de estados x_1^T de uma VLMC $\boldsymbol{X}$, seja $N_T(\omega, a)$ o número de ocorrências da sequência $\omega \in \cup_{j=1}^{k} E^j$ seguida pelo símbolo $a \in \ E$ na amostra x_1^T e seja $d(T) = O(\log T)$,

$$N_T(\omega, a) = \left|\left\{i : d(T) < i \leq T, x_{i-l(\omega)}^{i-1} = \omega, x_i = a\right\}\right|$$

e o número de ocorrências de ω em x_1^T é

$$N_T(\omega) = \left|\left\{i : d(T) < i \leq m, x_{i-l(\omega)}^{i-1} = \omega\right\}\right|$$

uma árvore de contextos viável é tal que $d(\mathcal{T}) \leq d(T)$, $N_T(\omega) \geq 1$ para todo $\omega \in \mathcal{T}$ e ω' com $N_T(\omega') \geq 1$ sufixo de algum $\omega \in \mathcal{T}$. O conjunto de árvores viáveis é denotado por $\mathcal{F}_{(x_1^T, d(T))}$.

Definição 2.6. *O Critério de Informação Bayesiana (BIC) para uma árvore viável é*

$$BIC_{\mathcal{T}} = -logML_{\mathcal{T}}(x_1^T) + \frac{(|E| - 1)|\mathcal{T}|}{2} logT, \tag{2.2}$$

onde $ML_{\mathcal{T}}(x_1^T) = \prod_{\omega \in \tau : N_T(a\omega) \geq 1} \prod_{a \in E} \left(\frac{N_T(\omega, a)}{N_T(\omega)}\right)^{N_T(\omega, a)}$

O Teorema principal provado em [9] (Teorema 2.6) é enunciado a seguir.

Teorema 2.1. *Seja x_1^T uma amostra de uma VLMC $\boldsymbol{X}$. Para $d(\mathcal{T}) < \infty$, o estimador BIC de $\mathcal{T}$ definido por*

$$\hat{\mathcal{T}}_{BIC}\left(x_1^T\right) = \arg \min_{\mathcal{T} \in \mathcal{F}_{(x_1^T, d(T))}} BIC_{\mathcal{T}}(x_1^T), \tag{2.3}$$

satisfaz

$$\hat{\mathcal{T}}_{BIC}\left(x_1^T\right) = \mathcal{T}$$

, quase certamente quando $T \to \infty$.

No caso geral, tem-se que

$$\hat{\mathcal{T}}|_{k_{BIC}}\left(x_1^T\right) = \mathcal{T}|_k,$$

quase certamente quando $T \to \infty$.

Definição 2.7. *Uma Cadeia de Markov Oculta de Alcance Variável (VLHMM) é um processo estocástico bivariado* $(\boldsymbol{X}, \boldsymbol{Z})$ *tal que:*

1) N *é o número de estados da VLMC oculta* $\boldsymbol{X}$*, com árvore* $\mathcal{T}$*;*

2) M *é o número de estados do processo observável* $\boldsymbol{Z}$ *com espaço de estados* O*;*

3) $\mathbb{A}$ *é a matriz das probabilidades de transição do processo oculto* $\boldsymbol{X}$ *definida por* $p(a|\omega)$, $\forall\ a \in E, \forall \omega \in \mathcal{T}$, *onde* a *é um estado da VLMC oculta* $\boldsymbol{X}$*;*

4) $\mathbb{B}$ *(Distribuição de Emissão) é o vetor das distribuição das probabilidades condicionais para algum símbolo do processo observável dado o contexto* ω *do processo oculto, definida por* $P(Z_t = z|X_{(t-l(\omega))+1}^t = \omega)$, $\forall \omega \in \mathcal{T}$, $\forall z \in O$*;*

5) π *é o vetor com a distribuição inical do processo oculto, definido por* $\pi(\omega) = P(X_1^{l(\omega)} = \omega)$, $\forall \omega \in \mathcal{T}$.

Observação 2.1. *Se o processo oculto* $\boldsymbol{X}$ *for markoviano e se tivermos* $P(Z_t = k|X_{(t-l(\omega))+1}^t = \omega) = P(Z_t = k|X_t = j)$*, ou seja, a distribuição de emissão perde memória de todo o contexto, então nesse caso, temos um caso particular de um VLHMM bem conhecido na literatura que são os modelos de Markov ocultos (HMM).*

2.2 Algoritmo de Baum-Welch

Dada uma sequência de observações de tamanho $T \in \mathbb{N}$, $Z = (z_1, z_2, ..., z_T)$, o algoritmo Expectation-Maximization de Baum-Welch [17] é usado para estimar o vetor de parâmetros de um HMM, dado por $\Theta = (\mathbb{A}, \mathbb{B}, \pi)$, onde $\mathbb{A} = \{p_{ij}\} = \{P(X_t = j|X_{t-1} = i)\}$, com $\boldsymbol{X}$ sendo uma cadeia de Markov assumindo valores em E, $\mathbb{B} = \{b_j(z_t)\} = \{P(Z_t = z_t|X_t = j)\}$ e $\pi = \{\pi_i\} = \{P(X_1 = i)\}$, para todo $i, j = 1, ..., N$ e todo $t \in \mathbb{Z}$.

Considere a variável $\rho_t(i, j)$, como sendo

$$\begin{aligned}
\rho_t(i,j) &= P(X_t = i, X_{t+1} = j|\boldsymbol{Z}, \Theta) \\
&= \frac{P(X_t = i, X_{t+1} = j, \boldsymbol{Z}|\Theta)}{P(\boldsymbol{Z}|\Theta)} \\
&= \frac{\alpha_t(i)a_{ij}b_j(z_{t+1})\beta_{t+1}(j)}{\sum_{k=1}^{N}\sum_{l=1}^{N}\alpha_t(k)a_{kl}b_l(z_{t+1})\beta_{t+1}(l)}
\end{aligned}$$

onde $\alpha_t(i)$ e $\beta_t(i)$ podem ser calculados usando os procedimentos forward e backward descritos a seguir, respectivamente,

$$\alpha_t(i) = P(z_1, z_2, \ldots, z_t, X_t = i|\Theta), \quad \beta_t(i) = P(z_{t+1}, z_{t+2}, \ldots, z_T|X_t = i, \Theta)$$

Seja $\gamma_t(i)$,

$$\gamma_t(i) = \sum_{j=1}^{N}\rho_t(i,j)$$

Somando $\gamma_t(i)$ em t obtemos o número esperado de transições do estado $i = 1, ..., N$, $\sum_{t=1}^{T-1}\gamma_t(i)$. Do mesmo modo, obtemos o número esperado de transição do estado i para o estado $j = 1, ..., N$, $\sum_{t=1}^{T-1}\rho_t(i,j)$. O vetor de parâmetros

Θ pode ser atualizado da seguinte forma

$$\begin{aligned}\tilde{\pi}_i &= \gamma_1(i) \\ \tilde{a}_{ij} &= \frac{\sum_{t=1}^{T-1}\rho_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \\ \tilde{b}_i(k) &= \frac{\sum_{t=1}^{T}1_{\{z_t=k\}}\gamma_t(i)}{\sum_{t=1}^{T}\gamma_t(i)}\end{aligned}$$

onde

$$1_{\{z_t=k\}} = \begin{cases}1, & \text{se } z_t = k \\ 0, & \text{caso contrário}\end{cases}$$

A prova da convergência desse algoritmo EM é apresentada em [2].

2.3 Revisão de Alguns Modelos de Perturbação Estocástica

Nesta seção apresentaremos modelos apresentados em [7] e [12] que são a base desta tese. Os autores consideram que a amostra observável está perturbada por algum tipo de ruído. Estes modelos de perturbação estocástica são interessantes porque podem ser usados para aproximar muitos fenômenos em que a variável em estudo é binária e pode ser lida com erro.

Em [7] os autores apresentam um modelo onde a cadeia de ordem infinita é estocasticamente perturbada por um ruído Bernoulli. Eles consideram $\boldsymbol{X}$ como uma cadeia estocástica binária de ordem infinita e $\boldsymbol{\xi}$ como uma sequência de variáveis aleatórias independentes Bernoulli tal que $P(\xi_t = 0) = 1 - \epsilon$ e independente de $\boldsymbol{X}$. Considerando

$$a \oplus b = a + b \ (mod\ 2), \quad a, b \in \{0, 1\}.$$

Para cada tempo t o valor do processo pode, aleatoriamente e independentemente, mudar com probabilidade fixa. O processo perturbado $\boldsymbol{Z}$ é definido por

$$Z_t = X_t \oplus \xi_t \quad t \in \mathbb{Z}. \tag{2.4}$$

Os autores demonstraram que a diferença entre as probabilidades de transição do processo original e o processo perturbado é limitado por uma constante, que é uma função crescente do parâmetro de ruído ϵ. Portanto, se este parâmetro de ruído é pequeno o suficiente, então é possível utilizar as estimativas das probabilidades de transição do processo perturbado como uma boa aproximação das probabilidades de transição de processo original.

Outro resultado apresentado em [7] é que, para uma amostra finita $z_1, z_2, ..., z_n$ do processo perturbado, a probabilidade da árvore de contextos estimada truncada na ordem k, $\hat{\mathcal{T}}|_k$, ser diferente da árvore de contextos verdadeira truncada na ordem k, $\mathcal{T}|_k$ decresce exponencialmente como função do tamanho da amostra e do parâmetro de ruído. Dessa maneira os autores obtêm um resultado de consistência forte, em que para uma amostra infinita $z_1, z_2, ...$ existe um $\bar{n}$ tal que para todo $n \geq \bar{n}$ $\hat{\mathcal{T}}|_k = \mathcal{T}|_k$, quase certamente, desde que algumas condições sejam satisfeitas (mais detalhes em [7]).

Em [12] os autores apresentam uma perturbação estocástica em que $\boldsymbol{X}$ é uma VLMC e o processo perturbado $\boldsymbol{Z}$ é gerado da seguinte maneira

$$Z_t = X_t \cdot \xi_t, \tag{2.5}$$

onde $\boldsymbol{\xi}$ é uma variável aleatória Bernoulli independente de $\boldsymbol{X}$, com $P(\xi_t = 1) = 1 - \epsilon$, em que ϵ é o parâmetro de ruído.

Assim como em [7], eles também mostram que se o parâmetro de ruído for pequeno suficiente, então as probabilidades de transição do processo original podem ser bem aproximadas por aquelas do processo perturbado, pois é provado que a diferença entre essas duas probabilidades são limitadas por uma constante que cresce com o ruído aleatório ϵ.

Para esse modelo, os autores também concluíram que para uma amostra finita Z_1^n do processo perturbado, a probabilidade da árvore de contextos estimada truncada na ordem k, $\hat{\mathcal{T}}|_k$, ser diferente da árvore de contextos

verdadeira truncada na ordem k, $\mathcal{T}\big|_k$, decresce exponencialmente com o tamanho da amostra e com o parâmetro de ruído. E assim como em [7], os autores obtém um resultado de consistência forte, em que para uma amostra infinita $z_1, z_2, ...$ existe um $\bar{n}$ tal que para todo $n \geq \bar{n}$ $\hat{\mathcal{T}}\big|_k = \mathcal{T}\big|_k$, quase certamente, desde que algumas condições sejam satisfeitas (mais detalhes em [12]).

Outro modelo considerado em [12] é uma mistura de duas cadeias de Markov de alcance variável (VLMC) independentes $\boldsymbol{X}$ e $\boldsymbol{Y}$ assumindo valores em um alfabeto finito $E = \{0, 1, ..., N-1\}$. Os autores consideraram $\boldsymbol{\xi}$ uma v.a Bernoulli, independente de $\boldsymbol{X}$ e $\boldsymbol{Y}$, com $P(\xi_t = 1) = 1 - \epsilon$, onde ϵ é um parâmetro de ruído conhecido e fixo em $(0, 1)$. Definiram um modelo perturbado dado por:

$$\boldsymbol{Z}_t = \begin{cases} \boldsymbol{X}_t, & \text{se} \quad \boldsymbol{\xi}_t = 1 \\ \boldsymbol{Y}_t, & \text{se} \quad \boldsymbol{\xi}_t = 0. \end{cases} \tag{2.6}$$

E também para esse modelo os autores concluíram que para uma amostra finita Z_1^n do processo perturbado, a probabilidade da árvore de contextos estimada truncada na ordem k, $\hat{\mathcal{T}}\big|_k$, ser diferente da árvore de contextos verdadeira truncada na ordem k, $\mathcal{T}\big|_k$, decresce exponencialmente com tamanho da amostra e com o parâmetro de ruído. E obtiveram também resultado de consistência forte, em que para uma amostra infinita Z_1^∞ existe um $\bar{n}$ tal que para todo $n \geq \bar{n}$ $\hat{\mathcal{T}}\big|_k = \mathcal{T}\big|_k$, quase certamente, desde que algumas condições sejam satisfeitas (mais detalhes em [12]).

Algumas perguntas sobre inferência para estes modelos permanecem sem resposta e serão abordados nesta tese, como por exemplo, se houve algum tipo de perturbação aplicada no processo oculto $\boldsymbol{X}$ e se essa perturbação é ou não pequena. Também estamos interessados em estimar os parâmetros do modelo, inclusive no caso em que o parâmetro de perturbação não for pequeno o suficiente para usar a amostra para estimar as probabilidades de transição do processo oculto $\boldsymbol{X}$.

Capítulo 3

Modelos de Perturbação Propostos

Nosso principal objetivo nesta tese é propor metodologias para estimar os parâmetros de uma classe de modelos a partir dos modelos propostos por [7] e [12], e realizar uma análise cuidadosa dos resultados seguindo o esquema de perturbação proposta por esses autores.

A seguir iremos propor os modelos de perturbação estocástica que analizaremos nesta tese. Consideramos $\boldsymbol{X}$ uma VLMC, como na Definição 2.1, assumindo valores em um alfabeto discreto $E = \{0, 1, ..., N-1\}, N \in \mathbb{N}$ e $\boldsymbol{\xi} = \{\xi_t\}_{t\in\mathbb{Z}}$ como sendo uma sequência de variáveis aleatórias com $P(\xi_t = i) = \epsilon_i$ tal que $\sum_{i=0}^{N-1} \epsilon_i = 1$, independente de $\boldsymbol{X}$.

Seguindo de perto os modelos apresentados em [7] e [12] consideramos os modelos de perturbação estocásticas detalhados nas próximas seções.

3.1 Modelo de Perturbação Tipo Soma

Em um Modelo de Perturbação Tipo Soma, que denotaremos resumidamente por TSCM[1], o processo perturbado $\boldsymbol{Z}$ é definido da seguinte maneira

$$Z_t = X_t \oplus \xi_t \ (mod|E|), \tag{3.1}$$

onde $\mathbf{X}$ é uma VLMC, com árvore $\mathcal{T}$ associada, e não uma cadeia de ordem infinita como em [7]. Observamos que o **TSCM** dado em (3.1) é um processo bivariado $(\boldsymbol{Z}, \boldsymbol{X})$ com vetor de parâmetros $\boldsymbol{\lambda_S} = (\mathbf{A_S}, \mathbf{B_S}, \boldsymbol{\pi_S})$, onde

$$\mathbf{A_S} = \{p(a|\omega)\} = P\left(X_0 = a \left| X_{-l(\omega)}^{-1} = \omega\right.\right), \forall\ a \in\ E, \forall\ \omega \in \mathcal{T},$$

são as probabilidades de transição do processo oculto $\boldsymbol{X}$,

$$\mathbf{B_S} = \{b_\omega(z_t)\} = \left\{P\left(Z_t = z_t \left| X_{t-l(\omega)+1}^{t} = \omega\right.\right)\right\}, \forall\omega \in E^{l(\omega)}, \forall z_t \in\ E,$$

é a distribuição de probabilidade do símbolo observado dada a sequência oculta do processo original (Distribuição de Emissão).

$$\boldsymbol{\pi_S} = \{\pi_\omega\} = \left\{P\left(X_{-l(\omega)}^{-1} = \omega\right)\right\}, \forall\ \omega \in \mathcal{T},$$

é a distribuição estacionária do contexto ω do processo original $\boldsymbol{X}$.

Seja $\boldsymbol{Z}$ um processo perturbado assumindo valores em um alfabeto discreto E, e seja $\boldsymbol{\lambda_S}$ o vetor de parâmetros do processo bivariado $(\boldsymbol{Z}, \boldsymbol{X})$. A função de verossimilhança $\mathbb{L}(\boldsymbol{\lambda_S}|Z)$ do conjunto dos valores do vetor $\boldsymbol{\lambda_S}$, dada uma amostra perturbada de símbolos observáveis Z de tamanho $T \in \mathbb{N}$, é definida por

$$\mathbb{L}(\boldsymbol{\lambda_S}|Z) = P(Z|\boldsymbol{\lambda_S}).$$

Considerando o modelo TSCM, temos que

[1] TSCM é a sigla em inglês para Type Sum Contaminated Model

Proposição 3.1.1. *Seja* $\boldsymbol{Z}$ *um processo perturbado definido em* ***TSCM****.*

i) *Para todo* $z_0, a_0, b_0 \in E$ *e todo* $\omega \in \mathcal{T}$ *a distribuição de emisssão* $\mathbf{B_S}$ *pode ser escrita como*

$$P\left(Z_0 = z_0 | X_{-l(\omega)+1}^0 = \omega\right) = P\left(Z_0 = z_0 | X_0 = a_0\right) = P\left(\xi_0 = b_0\right) I_{\{z_0 = a_0 \oplus b_0\}}. \tag{3.2}$$

ii) *As probabilidades de transição do processo* $\boldsymbol{Z}$ *truncado em alguma ordem* $k \in \mathbb{N}$, $\forall\ z_0 \in\ E$ *e* $\forall\ z_{-k}^{-1} \in\ E^k$, *pode ser escrita como :*

$$P\left(Z_0 = z_0 | Z_{-k}^{-1} = z_{-k}^{-1}\right) = \frac{\sum\limits_{\substack{a_t, b_t \in E: \\ -k \le t \le 0}} \prod\limits_{t=-k}^{0} P(\xi_t = b_t) P\left(\bigcap\limits_{t=-k}^{0} \{X_t = a_t\}\right) \prod\limits_{t=-k}^{0} I_{\{z_t = a_t \oplus b_t\}}}{\sum\limits_{\substack{a_t, b_t \in E: \\ -k \le t \le -1}} \prod\limits_{t=-k}^{-1} P(\xi_t = b_t) P\left(\bigcap\limits_{t=-k}^{-1} \{X_t = a_t\}\right) \prod\limits_{t=-k}^{-1} I_{\{z_t = a_t \oplus b_t\}}}. \tag{3.3}$$

Prova: A prova desse resultado é apresentada no Apêndice.

Considere uma amostra de tamanho $T \in \mathbb{N}$ do processo perturbado $\boldsymbol{Z}$, tal que $l(\omega) \le T, \forall \omega \in \mathcal{T}$, e $k = \max\{l(\omega) : \omega \in \mathcal{T}\}$, então se $d(\mathcal{T}) < \infty$ a função de verossimilhança $\mathbb{L}(\boldsymbol{\lambda_S}|Z)$ do processo perturbado $\boldsymbol{Z}$ pode ser escrito como:

$$\sum_{\substack{a_t, b_t \in E: \\ -k \le t \le T-k-1}} \prod_{t=-k}^{T-k-1} [P(\xi_t = b_t)] \left[P\left(X_{-k}^{-1} = a_{-k}^{-1}\right) \prod_{t=0}^{T-k-1} P\left(X_t = a_t | X_{t-l(\omega)}^{t-1} = a_{t-l(\omega)}^{t-1}\right)\right] \prod_{t=-k}^{T-k-1} I_{\{z_t = a_t \oplus b_t\}}, \tag{3.4}$$

e se a VLMC $\boldsymbol{X}$ com árvore de contextos $\mathcal{T}$ tal que $d(\mathcal{T}) = \infty$, então para algum $L \in \mathbb{N}$, tal que $L < T$, a função de verossimilhança $\mathbb{L}(\boldsymbol{\lambda_S}|Z)$ para o processo perturbado $\boldsymbol{Z}$ pode ser escrito como:

$$\sum_{\substack{a_t, b_t \in E: \\ -L \le t \le T-L-1}} \prod_{t=-L}^{T-L-1} [P(\xi_t = b_t)] \left[P\left(X_{-L}^{-1} = a_{-L}^{-1}\right) \prod_{t=0}^{T-L} P\left(X_t = a_t | X_{t-L}^{t-1} = a_{t-L}^{t-1}\right)\right] \prod_{t=-L}^{T-L-1} I_{\{z_t = a_t \oplus b_t\}}. \tag{3.5}$$

Observação 3.1.1. *Observamos que o TSCM é um VLHMM. Em consequência do modelo, a distribuição de emissão depende apenas do último símbolo do contexto, em vez de todo o contexto. Este fato nos permite propor algumas adaptações no algoritmo de Baum-Welch, originalmente para HMM, para estimar o vetor de parâmetros do VLHMM,* $\boldsymbol{\lambda_S}$*. Explicaremos a metodologia na próxima seção.*

Como ilustração do cálculo da verossimilhança, apresentamos o exemplo a seguir considerando um caso particular.

Exemplo 3.1.1. *Seja* $\boldsymbol{X}$ *uma cadeia de Markov assumindo valores em* $E = \{0, 1\}$ *e seja* $\boldsymbol{\xi}$ *uma sequência de v.a i.i.d com distribuição Bernoulli de parâmetro* $\epsilon \in (0, 1)$*. Considere uma amostra do processo perturbado de tamanho* $T = 2$ *com valores* $Z_{-1} = 0, Z_0 = 0$*. Portanto,*

$$\mathbb{L}(\boldsymbol{\lambda_S}|Z_{-1} = 0, Z_0 = 0) = P\left(\left[\{X_{-1} = 0, \xi_{-1} = 0\} \bigcup \{X_{-1} = 1, \xi_{-1} = 1\}\right] \bigcap \left[\{X_0 = 0, \xi_0 = 0\} \bigcup \{X_0 = 1, \xi_0 = 1\}\right]\right)$$

Observamos que
$\{X_t = a_i, \xi_t = b_j\} \bigcap \{X_t = a_k, \xi_t = b_m\} = \emptyset$*, para todo* $k \neq i, m \neq j$*. Portanto,*

$$\begin{aligned}
\mathbb{L}(\boldsymbol{\lambda_S}|Z_{-1} = 0, Z_0 = 0) &= P\left(X_{-1} = 0, \xi_{-1} = 0, X_0 = 0, \xi_0 = 0\right) + P\left(X_{-1} = 0, \xi_{-1} = 0, X_0 = 1, \xi_0 = 1\right) + \\
&+ P\left(X_{-1} = 1, \xi_{-1} = 1, X_0 = 0, \xi_0 = 0\right) + P\left(X_{-1} = 1, \xi_{-1} = 1, X_0 = 1, \xi_0 = 1\right) \\
&= P(\xi_0 = 0)P(\xi_{-1} = 0)P(X_{-1} = 0, X_0 = 0) + P(\xi_0 = 1)P(\xi_{-1} = 0)P(X_{-1} = 0, X_0 = 1) + \\
&+ P(\xi_0 = 0)P(\xi_{-1} = 1)P(X_{-1} = 1, X_0 = 0) + P(\xi_0 = 1)P(\xi_{-1} = 1)P(X_{-1} = 1, X_0 = 1) \\
&= P(\xi_0 = 0)P(\xi_{-1} = 0)P(X_0 = 0|X_{-1} = 0)P(X_{-1} = 0) + P(\xi_0 = 1)P(\xi_{-1} = 0) \times \\
&\times P(X_0 = 0|X_{-1} = 1)P(X_{-1} = 1) + P(\xi_0 = 0)P(\xi_{-1} = 1)P(X_0 = 0|X_{-1} = 1)P(X_{-1} = 1) + \\
&+ P(\xi_0 = 1)P(\xi_{-1} = 1)P(X_0 = 1|X_{-1} = 1)P(X_{-1} = 1)
\end{aligned}$$

3.2 Modelo de Perturbação Tipo Produto

Em um Modelo de Perturbação Tipo Produto, que resumidamente denotaremos por TPCM[2], o processo perturbado $\boldsymbol{Z}$ é definido da seguinte maneira

$$Z_t = X_t.\xi_t, \tag{3.6}$$

onde $\mathbf{X}$ é uma VLMC. Observamos que o TPCM é também um processo bivariado $(\boldsymbol{Z}, \boldsymbol{X})$ com vetor de parâmetros $\boldsymbol{\lambda_P} = (\mathbf{A_P}, \mathbf{B_P}, \boldsymbol{\pi_P})$, onde

$$\mathbf{A_P} = \{p(a|\omega)\} = P\left(X_0 = a \left| X_{-l(\omega)}^{-1} = \omega\right.\right), \forall\ a \in\ E, \forall\ \omega \in \mathcal{T},$$

são as probabilidades de transição do processo oculto $\boldsymbol{X}$,

$$\mathbf{B_P} = \{b_\omega(z_t)\} = \left\{P\left(Z_t = z_t \left| X_{t-l(\omega)+1}^{t} = \omega\right.\right)\right\}, \forall\ \omega \in\ E^{l(\omega)}, \forall z_t \in\ E,$$

é a distribuição de probabilidade do símbolo observado dado uma sequência oculta do processo original (Distribuição de Emissão). E

$$\boldsymbol{\pi}_P = \{\pi_\omega\} = \left\{P\left(X_{-j}^{-1} = \omega\right)\right\}, \forall\ \omega \in \mathcal{T}.$$

é a probabilidade estacionária do contexto ω do processo original $\boldsymbol{X}$.

Considerando o TPCM, temos que

Proposição 3.2.1. *Seja $\boldsymbol{Z}$ definido como em um TPCM.*

i) *Para todo $z_0, a_0, b_0 \in\ E$ e todo $\omega \in \mathcal{T}$ a distribuição de emissão $\mathbf{B}_P$ pode ser escrita como*

$$P\left(Z_0 = z_0 | X_{-l(\omega)+1}^{0} = \omega\right) = P\left(Z_0 = z_0 | X_0 = a_0\right) = P\left(\xi_0 = b_0\right) I_{\{z_0 = a_0.b_0\}}. \tag{3.7}$$

ii) *As probabilidades de transição do processo perturbado $\boldsymbol{Z}$ truncado em alguma ordem $k \in \mathbb{N}$, $\forall\ z_0 \in\ E$ e $\forall\ z_{-k}^{-1} \in\ E^k$, podem ser escritas como :*

$$P\left(Z_0 = z_0 | Z_{-k}^{-1} = z_{-k}^{-1}\right) = \frac{\displaystyle\sum_{\substack{a_t, b_t \in E: \\ -k \le t \le 0}} \prod_{t=-k}^{0} P(\xi_t = b_t) P\left(\bigcap_{t=-k}^{0} \{X_t = a_t\}\right) \prod_{t=-k}^{0} I_{\{z_t = a_t.b_t\}}}{\displaystyle\sum_{\substack{a_t, b_t \in E: \\ -k \le t \le -1}} \prod_{t=-k}^{-1} P(\xi_t = b_t) P\left(\bigcap_{t=-k}^{-1} \{X_t = a_t\}\right) \prod_{t=-k}^{-1} I_{\{z_t = a_t.b_t\}}}. \tag{3.8}$$

Prova: A prova desse resultado é apresentada no Apêndice.

Considere uma amostra de tamanho $T \in \mathbb{N}$ do processo perturbado $\boldsymbol{Z}$, tal que $l(\omega) \le T, \forall \omega \in \mathcal{T}$, e $k = \max\{l(\omega) : \omega \in \mathcal{T}\}$, então se $d(\mathcal{T}) < \infty$ a função de verossimilhança $\mathbb{L}(\boldsymbol{\lambda_S}|Z)$ do processo perturbado $\boldsymbol{Z}$ pode ser escrito como:

$$\sum_{\substack{a_t, b_t \in E: \\ -k \le t \le T-k-1}} \prod_{t=-k}^{T-k-1} [P(\xi_t = b_t)] \left[P\left(X_{-k}^{-1} = a_{-k}^{-1}\right) \prod_{t=0}^{T-k-1} P\left(X_t = a_t | X_{t-l(\omega)}^{t-1} = a_{t-l(\omega)}^{t-1}\right)\right] \prod_{t=-k}^{T-k-1} I_{\{z_t = a_t \cdot b_t\}}, \tag{3.9}$$

e se a VLMC $\boldsymbol{X}$ com árvore de contextos $\mathcal{T}$ tal que $d(\mathcal{T}) = \infty$, então para algum $L \in \mathbb{N}$, tal que $L < T$, a função de verossimilhança $\mathbb{L}(\boldsymbol{\lambda_S}|Z)$ para o processo perturbado $\boldsymbol{Z}$ pode ser escrito como:

$$\sum_{\substack{a_t, b_t \in E: \\ -L \le t \le T-L-1}} \prod_{t=-L}^{T-L-1} [P(\xi_t = b_t)] \left[P\left(X_{-L}^{-1} = a_{-L}^{-1}\right) \prod_{t=0}^{T-L} P\left(X_t = a_t | X_{t-L}^{t-1} = a_{t-L}^{t-1}\right)\right] \prod_{t=-L}^{T-L-1} I_{\{z_t = a_t \cdot b_t\}}. \tag{3.10}$$

Observação 3.2.1. *Observe que o TPCM é também um VLHMM e também por consequência do modelo, a distribuição de emissão depende somente do último símbolo do contexto e não de todo o contexto.*

[2] TPCM é a sigla em inglês para Type Product Contaminated Model

Capítulo 4

Modelo de Perturbação Tipo Mistura

No modelo dado pela equação (2.6), proposto por [12], que aqui chamaremos de TMCM[1], observamos que o processo perturbado $\boldsymbol{Z}$ assume valores no mesmo alfabeto E que os processos $\boldsymbol{X}$ e $\boldsymbol{Y}$. Para esse caso, considere que os processos $\boldsymbol{X}$ e $\boldsymbol{Y}$ são independentes, e $\boldsymbol{\xi} = \{\xi_t\}_{t\in\mathbb{Z}}$ como sendo uma sequência de variáveis aleatórias Bernoulli, com $P(\xi_t = 1) = 1 - \epsilon$ com $\epsilon \in (0, 1)$, independente de $\boldsymbol{X}$ e $\boldsymbol{Y}$.

A estrutura de dependência do processo perturbado para uma amostra finita de tamanho $T \in \mathbb{N}$ é ilustrado na Figura 4.1.

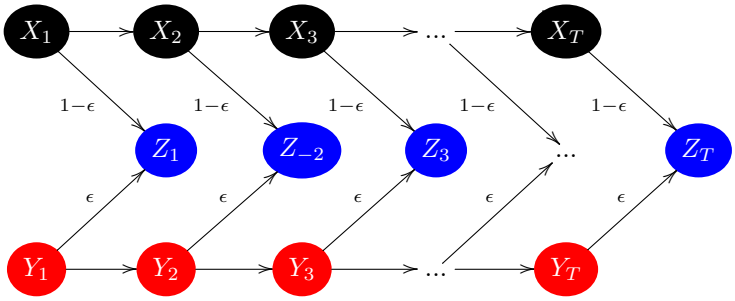


Figura 4.1. *Esquema do modelo TMCM*

A primeira sequência de círculos da Figura 4.1 representa a sequência dos estados ocultos em cada tempo t do processo $\boldsymbol{X}$ e a terceira sequencia representa a do processo $\boldsymbol{Y}$. As setas indicam a probabilidade de o processo $\boldsymbol{Z}$ assumir ou o símbolo do processo $\boldsymbol{X}$ ou o símbolo do processo $\boldsymbol{Y}$ no tempo t. A segunda sequência de círculos representa a sequência dos símbolos observados do processo $\boldsymbol{Z}$, que depende do resultado de cada estado atual ou do processo $\boldsymbol{X}$ ou do processo $\boldsymbol{Y}$.

Observamos que o modelo TMCM tem uma estrutura diferente dos modelos TSCM e TPCM e não é um VLHMM. Os elementos desse modelo são :

1) As probabilidades de transição do processo oculto $\boldsymbol{X}$, definidas por $\mathbf{A}^X_{\boldsymbol{M}} = \{p(a|\omega)\}$, $\forall\, a \in E$, $\forall\, \omega \in\ \mathcal{T}^X$,

$$p(a|\omega) = P\left(X_0 = a \,\middle|\, X_{-l(\omega)}^{-1} = \omega\right).$$

onde $\mathcal{T}^X$ é a árvore de contextos do processo oculto $\boldsymbol{X}$.

2) As probabilidades de transição do processo oculto $\boldsymbol{Y}$, definidas por $\mathbf{A}^Y_{\boldsymbol{M}} = \{p(c|\nu)\}$, $\forall\, c \in E$, $\forall\, \nu \in\ \mathcal{T}^Y$,

$$p(a|\nu) = P\left(Y_0 = c \,\middle|\, Y_{-l(\nu)}^{-1} = \nu\right).$$

onde $\mathcal{T}^Y$ é a árvore de contextos do processo oculto $\boldsymbol{Y}$.

[1] TMCM é a sigla em inglês para Type Mixture Contaminated Model

3) A probabilidade do ruído aleatório Bernoulli $\epsilon \in (0,1)$ é definida por

$$P(\xi_t = 1) = 1 - \epsilon, \forall t$$

Usaremos a notação compacta $\boldsymbol{\lambda_M} = (\mathbf{A}_{\boldsymbol{M}}^X, \mathbf{A}_{\boldsymbol{M}}^Y, \epsilon)$ para indicar o conjunto de parâmetros completo do modelo.

Para o processo estocasticamente perturbado $\boldsymbol{Z}$ proposto por [12], nós encontramos os seguintes resultados:

Proposição 4.1. *Seja $\boldsymbol{Z}$ um processo estocástico perturbado de acordo com o modelo TMCM, então*

i) As probabilidades de transição do processo $\boldsymbol{Z}$ truncadas em alguma ordem $k \in \mathbb{N}$, $\forall\ z_t, a_t, c_t \in E$, $b_t = \{0,1\}$ e $\forall\ z_{-k}^{-1} \in E^k$, são dadas por:

$$P\left(Z_0 = z_0 | Z_{-k}^{-1} = z_{-k}^{-1}\right) = \frac{\sum\limits_{\substack{a_t,b_t,c_t:\\ -k\leq t\leq 0}} \left\{ \prod\limits_{t=-k}^{0} P(\xi_t = b_t) P\left(\bigcap\limits_{-k\leq t\leq 0} \{X_t = a_t\} \right) P\left(\bigcap\limits_{-k\leq t\leq 0} \{Y_t = c_t\} \right) \right\} \prod\limits_{t=-k}^{0} I_{\{z_t = b_t.a_t + (1-b_t)c_t\}}}{\sum\limits_{\substack{a_t,b_t,c_t:\\ -k\leq t\leq -1}} \left\{ \prod\limits_{t=-k}^{-1} P(\xi_t = b_t) P\left(\bigcap\limits_{-k\leq t\leq -1} \{X_t = a_t\} \right) P\left(\bigcap\limits_{-k\leq t\leq 1} \{Y_t = c_t\} \right) \right\} \prod\limits_{t=-k}^{-1} I_{\{z_t = b_t.a_t + (1-b_t)c_t\}}} \tag{4.1}$$

Prova no Apêndice.

Dada uma amostra de tamanho T do processo perturbado $\boldsymbol{Z}$, tal que $l(\omega) \leq T, \forall \omega \in \mathcal{T}^X$, $l(\nu) \leq T, \forall \nu \in \mathcal{T}^Y$, e para $k = \max\{(l(\omega), l(\nu)) : \omega \in \mathcal{T}^X, \nu \in \mathcal{T}^Y$, então a função de verossimilhança $\mathbb{L}(\boldsymbol{\lambda_M}|Z)$ do processo perturbado $\boldsymbol{Z}$ truncado em k pode ser escrita como:

$$\mathbb{L}(\boldsymbol{\lambda_M}|Z) = \sum_{\substack{a_t,b_t,c_t:\\ -k\leq t\leq T-k-1}} \left\{ \prod_{t=-k}^{T-k-1} P(\xi_t = b_t) P\left(\bigcap_{0\leq t\leq T-k-1} \{X_t = a_t\} \right) P\left(\bigcap_{0\leq t\leq T-k-1} \{Y_t = c_t\} \right) \right\} \prod_{t=-k}^{T-k-1} I_{\{z_t = b_t \cdot a_t + (1-b_t)c_t\}} \tag{4.2}$$

4.1 Estimação via Verossimilhança Perfilada para o modelo TMCM

Seja $\boldsymbol{\lambda_M}$ o conjunto de parâmetros do modelo TMCM e considere a verossimilhança do processo perturbado $\boldsymbol{Z}$, dada pela equação (4.2). O objetivo é estimar $\boldsymbol{\lambda_M}$ que maximiza a verossimilhança dado a amostra Z. Uma possibilidade é utilizar a verossimilhança perfilada (ver Apêndice) para estimar o vetor de parâmetros $\boldsymbol{\lambda_M}$. O procedimento é descrito a seguir.

Para cada $\epsilon \in (0, 0.5)$ fixo, utilizamos as probabilidades de transição do processo $\boldsymbol{Z}$, truncado em alguma ordem $k \in \mathbb{N}$, como valor inical das probabilidades de transição do processo oculto $\boldsymbol{X}$, pois teremos $P(\xi_t = 1) \geq 0.5$. Isso significa que a grande maioria dos símbolos observados do processo perturbado $\boldsymbol{Z}$ é oriundo do processo oculto $\boldsymbol{X}$, logo é razoável utilizar as probabilidades de transição do processo $\boldsymbol{Z}$ como valor inicial das probabilidades de transição do processo oculto $\boldsymbol{X}$. Sendo assim, fixado ϵ e a matriz $\mathbf{A}_{\boldsymbol{M}}^X$, o vetor de parâmetros completo $\boldsymbol{\lambda_M}$, se restringe apenas a matriz de transição do processo oculto $\boldsymbol{Y}$, dada por $\mathbf{A}_{\boldsymbol{M}}^Y$.

Similarmente, para cada $\epsilon \in (0.5, 1)$ fixo, tal que tenhamos $P(\xi_t = 1) < 0.5$, utilizamos as probabilidades de transição do processo $\boldsymbol{Z}$, truncado em alguma ordem $k \in \mathbb{N}$, como valor inical das probabilidades de transição do processo oculto $\boldsymbol{Y}$. Pois nesse caso, a maioria dos símbolos observados do processo perturbado $\boldsymbol{Z}$ é oriundo do processo oculto $\boldsymbol{Y}$. Portanto, fixado ϵ e a matriz $\mathbf{A}_{\boldsymbol{M}}^Y$, o vetor de parâmetros completo $\boldsymbol{\lambda_M}$, se resume apenas na matriz de transição do processo oculto $\boldsymbol{X}$, dada por $\mathbf{A}_{\boldsymbol{M}}^X$. Chamaremos, aqui nesta Seção, $\mathbf{A}_{\boldsymbol{M}}^X$ e $\mathbf{A}_{\boldsymbol{M}}^Y$ de vetor de parâmetros de perturbação.

Vamos considerar primeiro o caso em que $d(\mathcal{T}^X) < \infty$ e $d(\mathcal{T}^Y) < \infty$. Considere o modelo TMCM onde $\boldsymbol{X}$ está associado a uma árvore finita $\mathcal{T}^X$ e $\boldsymbol{Y}$ a uma árvore finita $\mathcal{T}^Y$. Como não sabemos quais são os contextos $\omega \in \mathcal{T}^X$ e quais os contextos $\nu \in \mathcal{T}^Y$, e nem os comprimentos desses contextos, então nesse caso, para estimar as árvores de contextos $\mathcal{T}^X$ e $\mathcal{T}^Y$ dos processos ocultos $\boldsymbol{X}$ e $\boldsymbol{Y}$, respectivamente, iremos primeiramente estimar as árvores de contextos $k - full\ \mathcal{T}_k^X$ e $\mathcal{T}_k^Y$, para algum $k \in \mathbb{N}$.

Portanto, temos o vetor de parâmetros completo $\boldsymbol{\lambda^*} = \left\{\mathbf{A}_{\boldsymbol{M}}^X, \mathbf{A}_{\boldsymbol{M}}^Y, \epsilon\right\}$, onde $\mathbf{A}_{\boldsymbol{M}}^X$ é o conjunto formado pelas probabilidades de transição do processo $\boldsymbol{X}$ de ordem k, $\mathbf{A}_{\boldsymbol{M}}^Y$ é o conjunto contendo as probabilidades de transição do processo $\boldsymbol{Y}$, também de ordem k e ϵ é o parâmetro de ruído.

Considere $\boldsymbol{\lambda}_{\boldsymbol{M}}^{\boldsymbol{P}}$ o conjunto de parâmetros de interesse. Portanto, para cada $\epsilon \in (0, 0.5)$, fixo, tal que $P(\xi_t = 1) \geq 0.5$, definimos a função de verossimilhança perfilada para o vetor de parâmetros de interesse $\boldsymbol{\lambda}_{\boldsymbol{M}}^{\boldsymbol{P}}$, dada uma amostra Z de tamanho T do processo perturbado $\boldsymbol{Z}$, truncado em alguma ordem $k \in \mathbb{N}$, como sendo

$$\mathbb{L}_P(\boldsymbol{\lambda}_{\boldsymbol{M}}^{\boldsymbol{P}}|Z) = \max_{\mathbf{A}_{\boldsymbol{M}}^Y} \mathbb{L}(\boldsymbol{\lambda}_{\boldsymbol{M}}^{\boldsymbol{P}}, \mathbf{A}_{\boldsymbol{M}}^Y|Z) \tag{4.3}$$

onde $\mathbb{L}(\boldsymbol{\lambda}_{\boldsymbol{M}}^{\boldsymbol{P}}, \mathbf{A}_{\boldsymbol{M}}^Y|Z)$ é definida pela verossimilhança completa, equação (4.2), fixado ϵ e $\mathbf{A}_{\boldsymbol{M}}^X$.

Assim, para cada $\epsilon \in (0, 0.5)$ fixo, podemos calcular o valor da função de máxima verossimilhança, avaliada nas várias estimativas do vetor de parâmetros $\hat{\mathbf{A}}_{\boldsymbol{M}}^X$ e $\hat{\mathbf{A}}_{\boldsymbol{M}}^Y$, ou seja, podemos calcular $\mathbb{L}(\hat{\mathbf{A}}_{\boldsymbol{M}}^X, \hat{\mathbf{A}}_{\boldsymbol{M}}^Y, \epsilon|Z)$ dada pela equação (4.2).

E para o caso quando se tem $P(\xi_t = 1) < 0.5$, definimos a função de verossimilhança perfilada para o vetor de parâmetros de interesse $\boldsymbol{\lambda}_{\boldsymbol{M}}^{\boldsymbol{P}}$, dada uma amostra Z de tamanho T, do processo perturbado $\boldsymbol{Z}$, truncado em alguma ordem $k \in \mathbb{N}$, como sendo

$$\mathbb{L}_P(\boldsymbol{\lambda}_{\boldsymbol{M}}^{\boldsymbol{P}}|Z) = \max_{\mathbf{A}_{\boldsymbol{M}}^X} \mathbb{L}(\boldsymbol{\lambda}_{\boldsymbol{M}}^{\boldsymbol{P}}, \mathbf{A}_{\boldsymbol{M}}^X|Z) \tag{4.4}$$

onde $\mathbb{L}(\boldsymbol{\lambda}_{\boldsymbol{M}}^{\boldsymbol{P}}, \mathbf{A}_{\boldsymbol{M}}^X|Z)$ é definida pela verossimilhança completa, equação (4.2), fixado ϵ e $\mathbf{A}_{\boldsymbol{M}}^Y$.

E nesse caso, para cada $\epsilon \in (0, 0.5)$ fixo, podemos calcular o valor da função de máxima verossimilhança, avaliada nas várias estimativas do vetor de parâmetros $\hat{\mathbf{A}}_{\boldsymbol{M}}^X$ e $\hat{\mathbf{A}}_{\boldsymbol{M}}^Y$, ou seja, podemos calcular $\mathbb{L}(\hat{\mathbf{A}}_{\boldsymbol{M}}^X, \hat{\mathbf{A}}_{\boldsymbol{M}}^Y, \epsilon|Z)$ dada pela equação (4.2).

Desse modo, para todo $\epsilon \in (0, 1)$ o estimador do vetor de parâmetros completo $\boldsymbol{\lambda^*}$, do modelo TMCM, será dado por

$$\hat{\boldsymbol{\lambda}}^* = \arg \max_{\hat{\mathbf{A}}_{\boldsymbol{M}}^X, \hat{\mathbf{A}}_{\boldsymbol{M}}^Y, \epsilon} \left[\mathbb{L}_P(\boldsymbol{\lambda}_{\boldsymbol{M}}^{\boldsymbol{P}}|Z)\right]$$

Feito isso, dado o vetor de parâmetros completo $\hat{\boldsymbol{\lambda}}^*$, aplicamos o procedimento de estimação da árvore de contextos como descrito na Seção 5.1, a fim de obter as árvores estimadas $\hat{\mathcal{T}}^X$ e $\hat{\mathcal{T}}^Y$.

No caso em que $\boldsymbol{X}$ e $\boldsymbol{Y}$ têm árvores de contextos de comprimento infinito, é possível estimar somente as árvores $\mathcal{T}^X\big|_k$ e $\mathcal{T}^Y\big|_k$, onde $k \in \mathbb{N}$ é tão gande quanto possível, dada uma amostra de tamanho T. Aplicamos a mesma metodologia para árvores finitas, isto é, estimamos as árvores $k - full$ e então aplicamos o procedimento de estimação da árvore de contextos, a fim de obter as árvores estimadas $\hat{\mathcal{T}}^X$ e $\hat{\mathcal{T}}^Y$.

Capítulo 5

Algoritmos e Estimadores Propostos para os Modelos TSCM e TPCM

Enfatizamos que, em um procedimento de estimação, a principal diferença entre um VLHMM e um HMM é que em um HMM os estados processo original são conhecidos, enquanto que em um VLHMM os contextos são desconhecidos. Então temos primeiramente que conhecer quais são os contextos que pertencem a árvore $\mathcal{T}$ associada ao VLMC $\boldsymbol{X}$, a fim de estimar os parâmetros do modelo. Esse fato torna o processo de estimação muito mais complexo. Nossa proposta para contornar essa dificuldade é composta por duas partes. Na primeira parte, estimamos a árvore $k-full$, aqui denotada por $\mathcal{T}_k$, dada as observações, ou seja, estimamos a cadeia de Markov de ordem k, com k tão grande quanto possível, mas fixo. Na segunda parte, aplicamos um procedimento de poda dos galhos a fim de obter a árvore estimada $\hat{\mathcal{T}}$ de $\mathcal{T}$.

Vamos considerar primeiramente $d(\mathcal{T}) < \infty$. Considere o VLHMM $(\boldsymbol{X}, \boldsymbol{Z})$, onde $\boldsymbol{X}$ tem uma árvore finita $\mathcal{T}$, e seja k o comprimento do maior contexto, $k = \max\{l(\omega) : \omega \in \mathcal{T}\}$. Seja $\boldsymbol{X^*} = \{X_r^*\}_{r\in N}$ uma cadeia de Markov de ordem k, com árvore $k-full$ $\mathcal{T}_k$, assumindo valores em E^k tal que

$$\boldsymbol{X^*}_r := \boldsymbol{X}^r_{(r+k)-1}, r = 1, ..., (T-k)+1.$$

As probabilidades de transição de $\boldsymbol{X^*}$ são dadas por $\boldsymbol{A^*} = \{p^*(\omega|\nu)\}, \forall \omega, \nu \in E^k$ e com distribuição inicial $\boldsymbol{\pi^*} = \{P(X_1^* = \omega)\}, \forall \omega \in E^k$.

Similarmente, definimos um novo processo observável $\boldsymbol{Z}^* = Z^*_{r\in N}$ com valores em E^k como

$$\boldsymbol{Z}_r^* = \boldsymbol{Z}^r_{(r+k)-1}, r = 1, ..., (T-k)+1.$$

Considerando o TSCM, a distribuição de emissão do processo bivariado $(\boldsymbol{Y}\ , \boldsymbol{Z}^*)$, definida por $\boldsymbol{B^*} = \{b_\omega(\upsilon)\}$ é

$$\begin{aligned} P(Z_r^* = \upsilon | X_r^* = \omega) &= P(Z_{-k}^{-1} = z_{-k}^{-1} | X_{-k}^{-1} = x_{-k}^{-1}) \\ &= \prod_{t=-k}^{-1} P(Z_t = z_t | X_t = x_t) = \prod_{t=-k}^{-1} P(\xi_t = b_t) I_{\{z_t = x_t \oplus b_t\}} \end{aligned}$$

E para o regime TPCM, a distribuição de emissão $\boldsymbol{B^*} = \{b_\omega(\upsilon)\}$ é dada por

$$\begin{aligned} P(Z_r^* = \upsilon | X_r^* = \omega) &= P(Z_{-k}^{-1} = z_{-k}^{-1} | X_{-k}^{-1} = x_{-k}^{-1}) \\ &= \prod_{t=-k}^{-1} P(Z_t = z_t | X_t = x_t) = \prod_{t=-k}^{-1} P(\xi_t = b_t) I_{\{z_t = x_t . b_t\}} \end{aligned}$$

Desse modo, o VLHMM $(\boldsymbol{X}, \boldsymbol{Z})$ pode ser visto como sendo um HMM $(\boldsymbol{Y}, \boldsymbol{Z^*})$ com vetor de parâmetros $\boldsymbol{\lambda^*} = (\boldsymbol{A^*}, \boldsymbol{B^*}, \boldsymbol{\pi^*})$ e podemos aplicar o algoritmo de Baum-Welch a fim de estimar os parâmetros do modelo, como será detalhado posteriormente.

Como exemplo, suponha que $\boldsymbol{X}$ é uma VLMC que assume valores em um alfabeto $E = \{0, 1\}$ e $k = \max\{l(\omega) : \omega \in \mathcal{T}\} = 2$. Dada uma amostra $X = \{0, 0, 1, 0, 1, 1, 0, 1, ..., 0, 1\}$, então uma amostra associada a cadeia de Markov $\boldsymbol{X^*}$ de ordem $k = 2$ será dada por

$$X^* = \{00, 01, 10, 01, 11, 10, 01, ..., 01\}.$$

E para uma amostra observada $Z = \{0, 0, 1, 1, 0, 0, 1, ..., 1, 0\}$ do processo perturbado $\boldsymbol{Z}$, temos que a nova amostra observada do processo $\boldsymbol{Z^*}$ será dada por

$$Z^* = \{00, 01, 11, 10, 00, 01, ..., 10\}.$$

No caso em que $\boldsymbol{X}$ tem árvore de contextos de comrimento infinito, ou seja, $d(\mathcal{T}) = \infty$, e temos uma amostra finita, é possivel estimar somente a árvore truncada $\mathcal{T}\big|_k$, onde $k \in \mathbb{N}$ é tão grande quando possível, dada uma amostra de tamanho T. Aplicamos a mesma metodologia proposta para árvores finitas, ou seja, primeiro estimamos a árvore $k - full$, através do algoritmo de Baum-Welch e então aplicamos a procedimento de poda para obter a árvore estimada $\hat{\mathcal{T}}\big|_k$ de $\mathcal{T}\big|_k$.

Em ambos os casos, ou seja, quando a árvore de contextos $\mathcal{T}$ do processo oculto $\boldsymbol{X}$ tem comprimento finito ou finito, não estimamos os parâmetros da VLHMM original $(\boldsymbol{Z}, \boldsymbol{X})$, mas ao invés disso, estimamos o vetor de parâmetros do HMM $(\boldsymbol{X^*}, \boldsymbol{Z^*})$ dado por $\boldsymbol{\lambda^*} = (\boldsymbol{A^*}, \boldsymbol{B^*}, \boldsymbol{\pi^*})$, associado a árvore $\mathcal{T}_k$ (caso finito) ou árvore truncada $\mathcal{T}\big|_k$ (caso infinito).

Note que se $\mathcal{T}$ é finita, então a ordem inicial da árvore $k-full$ é $k \geq l(\omega)$ para todo $\omega \in \mathcal{T}$. Se $\mathcal{T}$ é infinita, então $k \leq l(\omega)$ para todo $\omega \in \mathcal{T}$. Mas, como esta informação sobre a ordem da verdadeira árvore não está disponível, em geral, não levaremos isso em conta na metodologia proposta.

Como visto, o objetivo de transformar o VLHMM ($\boldsymbol{X}$, $\boldsymbol{Z}$) em um HMM ($\boldsymbol{X^*}$, $\boldsymbol{Z^*}$) é utilizar o algoritmo de Baum-Welch para estimar o vetor de parâmetros $\boldsymbol{\lambda^*}$. Porém, como pode ser visto em [14], o algoritmo de Baum-Welch é um algoritmo EM e portanto existe a possibilidade da convergência para um máximo local da função de verossimilhança.

Nossa proposta para evitar um máximo local é utilizar vários valores iniciais distintos da distribuição de emissão $\boldsymbol{B^*}$, deixando o valor de ϵ_i percorrer todo o espaço paramétrico, ou seja, para cada $\epsilon \in (0, 1)$ temos uma distribuição de emissão $\boldsymbol{B^*}$ diferente. E para as probabilidades de transição da árvore $k-full$ de $\boldsymbol{X^*}$, usamos como valor inicial a matriz de transição empírica do processo $\boldsymbol{Z}^*$ truncado na ordem k. Então, para cada valor do parâmetro de ruído que aparece na distribuição de emissão $\boldsymbol{B^*}$, e utilizando o algortimo de Baum-Welch, obtemos uma estimativa $\tilde{\boldsymbol{\lambda}}^*$ do vetor de parâmetros $\boldsymbol{\lambda}^*$.

Nossa proposta para estimar $\boldsymbol{\lambda^*}$ é selecionar o vetor que maximiza a verossimilhança $\mathbb{L}(\boldsymbol{\lambda^*}|Z^*)$, dada uma amostra observada $Z^* = {z^*_{r=1}}^{(T-k)+1}$, ou seja

$$\hat{\boldsymbol{\lambda}}^* = \arg\max_{\tilde{\boldsymbol{\lambda}}^* \in \boldsymbol{\Lambda}} \mathbb{L}\left(\tilde{\boldsymbol{\lambda}}^* | Z^*\right), \tag{5.1}$$

onde $\boldsymbol{\Lambda}$ é o conjunto de estimativas $\tilde{\boldsymbol{\lambda}}^*$ de $\boldsymbol{\lambda}^*$. Cada uma das estimativas $\tilde{\boldsymbol{\lambda}}^*$ é um estimador de máxima verossimilhança (EMV) obtido através do algoritmo de Baum-Welch com distintos valores iniciais para $\boldsymbol{\lambda}^*$, em que estes distintos valores iniciais são dados pelos distintos valores do ruído de perturbação que aparece na distribuição de emissão $\boldsymbol{B^*}$.

Desse modo, dada uma amostra observada Z^* do processo $\boldsymbol{Z^*}$, os procedimentos forward e backward, avaliados na amostra Z^*, são descritos a seguir. Defina

$$\gamma_r(\omega) = P(X^*_r = \omega | Z^*, \boldsymbol{\lambda}^*), \tag{5.2}$$

como sendo a probabilidade de estar no contexto $\omega \in E^{l(\omega)}$, no tempo r, dada a sequência de observação Z^* e o vetor de $\boldsymbol{\lambda}^*$. A variável forward é dada por

$$\alpha_r(\omega) = P\left(z^*_1, ..., z^*_r, X^*_r = \omega \middle| \boldsymbol{\lambda}^*\right),$$

e, por indução

$$\alpha_1(\omega) = \pi_\omega b_\omega(z_1^*), \forall \omega \in E^k,$$

$$\alpha_{r+1}(\omega) = \left[\sum_{\nu \in E^k} \alpha_r(\omega) p^*(\omega|\nu)\right] b_\omega\left(z_{r+1}^*\right), \forall \omega \in E^k, \quad 2 \leq r \leq (T-k)+1.$$

Similarmente, a variável backward é definida por

$$\beta_r(\omega) = P\left(z_{r+1}^*, z_{r+2}^*, ..., z_{(T-k)+1} | X_r^* = \omega, \boldsymbol{\lambda}^*\right),$$

e, por indução segue, para $r = (T-k)+2, (T-k)+3, ..., 1$

$$\beta_{(T-k)+1}(\omega) = 1, \forall \omega \in E^k,$$

$$\beta_r(\omega) = \sum_{\nu \in E^k} p^*(\omega|\nu) b_\omega\left(z_{r+1}^*\right) \beta_{r+1}(\omega), \forall \omega \in E^k.$$

Dadas as variáveis $\alpha_r(\omega)$ e $\beta_r(\omega)$, a equação (5.2) pode ser expressa em termos das variáveis forward e backward, ou seja

$$\gamma_r(\omega) = \frac{\alpha_r(\omega)\beta_r(\omega)}{\sum_{\omega \in E^k} \alpha_r(\omega)\beta_r(\omega)}, \tag{5.3}$$

Para descrever o procedimento de reestimação dos parâmetros do HMM ($\boldsymbol{X}^*$, $\boldsymbol{Z}^*$), defina $\delta_r(\omega, \nu)$ como sendo a probabilidade de estar no contexto ω no tempo r e no contexto ν no tempo $r+1$, dado o vetor $\boldsymbol{\lambda}^*$ e a sequência de observação Z^*, ou seja

$$\delta_r(\omega, \nu) = P\left(X_r^* = \omega, X_{r+1}^* = \nu | Z^*, \boldsymbol{\lambda}^*\right).$$

Usando as variáveis forward e backward, podemos escrever δ_r como

$$\delta_r(\omega, \nu) = \frac{\alpha_r(\omega) p^*(\omega|\nu) b_\omega\left(z_{r+1}^*\right) \beta_{r+1}(\omega)}{\sum_{\omega \in E^k} \sum_{\nu \in E^k} \alpha_r(\omega) p^*(\omega|\nu) b_\omega\left(z_{r+1}^*\right) \beta_{r+1}(\omega)}, \ \forall \omega, \nu \in E^k.$$

Portanto, o vetor de parâmetros pode ser atualizado da seguinte maneira:

$$\hat{\boldsymbol{\pi}}^* = \{\hat{\pi}_\omega^*\}_{\omega \in E^k} = \gamma_1(\omega),$$

$$\hat{\mathbf{A}}^* = \{\hat{p}^*(\omega|\nu)\}_{\omega,\nu \in E^k} = \frac{\sum_{r=1}^{T+k-1} \delta_r(\omega, \nu)}{\sum_{r=1}^{T+k-1} \gamma_r(\omega)}, \tag{5.4}$$

$$\hat{\mathbf{B}}^* = \left\{\hat{b}_\omega(\nu)\right\}_{\omega,\nu \in E^k} = \frac{\sum_{r=1}^{T+k} I_{\{z_r^* = \nu\}} \gamma_r(\omega)}{\sum_{r=1}^{T+k} \gamma_r(\omega)}, \tag{5.5}$$

onde

$$I_{\{z_r^* = \nu\}} = \begin{cases} 1 & \text{se } z_r^* = \nu, \\ 0 & \text{caso contrário.} \end{cases}$$

Uma alternativa para estimar $\boldsymbol{\lambda}^*$ é escolher $\tilde{\boldsymbol{\lambda}}^* \in \Lambda$ que minimiza a divergência de Kullback-Liebler entre o verdadeiro processo observável $\boldsymbol{Z^*}$ e o processo observável estimado $\hat{\boldsymbol{Z}}^*$, dado $\tilde{\boldsymbol{\lambda}}^*$. O procedimento é descrido a seguir.

Dado $\tilde{\boldsymbol{\lambda}}^*$, usamos a função $\nu(i,\nu') = \nu$, que adiciona o estado $i \in E$ ao contexto ν' de comprimento $l(\nu') = l(\nu)-1$, para todo $\nu \in \hat{\mathcal{T}}_k$ (para $d(\mathcal{T}) < \infty$) ou $\nu \in \hat{\mathcal{T}}\big|_k$ (para $d(\mathcal{T}) = \infty$). Então, as estimativas das probabilidades de transição de ordem k do processo oculto $\boldsymbol{X}$ dadas pela equação (5.4) podem ser escritas como:

$$\hat{\mathbf{A}}^* = \left\{\hat{p}^*(a|\omega(i,\omega') = \omega) = \hat{p}^*(\nu(\nu',j) = \nu|\omega(i,\omega') = \omega) I_{\{\omega'=\nu'\}} I_{\{i=j\}}\right\}, \tag{5.6}$$

$\forall i,j \in E, \quad \forall \omega', \nu' \in E^{k-1}$.

Dessa maneira, geramos uma amostra do processo oculto estimado $\hat{\boldsymbol{X}}$ truncado na ordem k e através de uma amostra de $\hat{\xi}_t$ associada as estimativas do ruído aleatório $\hat{\epsilon}_i$, encontradas através da equação (5.5), aplicamos o regime de perturbação (**TSCM** ou **TPCM**) como se segue

$$\hat{Z}_t = \hat{X}_t \oplus \hat{\xi}_t, \tag{5.7}$$

ou

$$\hat{Z}_t = \hat{X}_t \cdot \hat{\xi}_t. \tag{5.8}$$

Depois disso, podemos comparar a lei do processo $\hat{\boldsymbol{Z}}$ truncado na ordem k com a lei do processo observado $\boldsymbol{Z}$ através da divergência de Kullback-Leibler definida por

$$D_{\mathrm{KL}}(p_z|p_{\hat{z}}) = \sum_i p(z_i|z_{i-k}) \ \log \frac{p(z_i|z_{i-k})}{p\left(\hat{z}_i|\hat{z}_{i-k}\right)}, \ \forall\ i = 1, ..., T. \tag{5.9}$$

Observação 5.1. *A equivalência entre os dois métodos de estimação é apresentada no Apêndice .*

5.1 Procedimento de Estimação da árvore de contextos

Na segunda parte do procedimento de estimação queremos estimar o verdadeiro vetor de parâmetros $\boldsymbol{\lambda}$, uma vez que temos as estimativas de $\boldsymbol{\lambda^*}$. Com essa finalidade realizamos um processo de poda dos galhos da árvore estimada usando uma adaptação do Critério de Informação Bayesiana (BIC) proposta por [9], que é explicada nesta seção. Sob algumas poucas condições, em [9] os autores mostraram que o BIC fornece um estimador consistente para uma VLMC quando a amostra vem de uma VLMC.

Como temos a árvore de contextos $k-full$ estimada, $\hat{\mathcal{T}}_k$ (para $d(\mathcal{T} < \infty)$) e $\hat{\mathcal{T}}\big|_k$ (para $d(\mathcal{T}) = \infty$), obtida via algoritmo Baum-Welch, queremos agora estimar a árvore de contextos $\mathcal{T}$ (se $d(\mathcal{T}) < \infty$) ou $\mathcal{T}\big|_k$ (se $d(\mathcal{T}) = \infty$) que é um subconjunto dos galhos de $\hat{\mathcal{T}}\big|_k$. Para isso, aplicamos o procedimento de estimação dos galhos baseado na verossimilhança da amostra gerada da matriz das probabilidades de transição estimada $\hat{\mathbf{A}}$.

O algoritmo BIC proposto por [9] para estimar a árvore de contextos $\mathcal{T}$, utiliza a amostra verdadeira da VLMC $\boldsymbol{X}$. No nosso caso, não dispomos da verdadeira amostra X da VLMC oculta $\boldsymbol{X}$. Então, a nossa proposta é aplicar uma versão bootstrap do algoritmo BIC substituindo a amostra da verdadeira VLMC pela amostra bootstrap $\hat{X} := \hat{x}_1, \ldots, \hat{x}_m, m = O(T)$ gerada através da matriz de transição estimada $\hat{\mathbf{A}}^*$ da cadeia de Markov $\boldsymbol{X^*}$ de ordem k que foi estimada através do algoritmo de Baum-Welch.

Seguindo [9], seja $N_m^{\hat{X}}(\omega, a)$ o número de ocorrências da sequência $\omega \in \cup_{j=1}^k E^j$ seguido pelo símbolo $a \in E$ na amostra bootstrap $\hat{X}$ e $D(m) = o(\log m)$,

$$N_m^{\hat{X}}(\omega, a) = \left|\left\{i : D(m) < i \leq m, \hat{x}_{i-l(\omega)}^{i-1} = \omega, \hat{x}_i = a\right\}\right|,$$

e o número de ocorrências de ω em $\hat{X}$ é dado por

$$N_m^{\hat{X}}(\omega) = \left|\left\{i : D(m) < i \leq m, x_{i-l(\omega)}^{i-1} = \omega\right\}\right|.$$

Dada uma amostra $\hat{X}$, uma árvore de contextos viável é tal que $d(\hat{\mathcal{T}}_k) \leq D(m)$, $N_m^{\hat{X}}(\omega) \geq 1$ para todo $\omega \in \hat{\mathcal{T}}_k$ e ω' com $N_m^{\hat{X}}(\omega') \geq 1$ sufixo de algum $\omega \in \hat{\mathcal{T}}_k$.

Seja $\mathcal{F}(\hat{X}, D(m))$ uma família de árvores de contextos viáveis. Considerando uma amostra de $\hat{\mathcal{T}}_k \in \mathcal{F}(\hat{X}, D(m))$ definimos a função de máxima verossimilhança bootstrap por

$$L_{\hat{\boldsymbol{\lambda}}^*}(\hat{X}) = \prod_{\omega \in \hat{\mathcal{T}}_k} \tilde{P}_{L,\omega}(\hat{X}), \tag{5.10}$$

onde $L_{\hat{\boldsymbol{\lambda}}_k^*}(\hat{X})$ é a função de máxima verossimilhança da amostra $\hat{X}$ e

$$\tilde{P}_{L,\omega}(\hat{X}) = \begin{cases} \prod_{a \in E} \left(\dfrac{N_m^{\hat{X}}(\omega, a)}{N_m^{\hat{X}}(\omega)} \right)^{N_m^{\hat{X}}(\omega,a)} & \text{se} \quad N_m^{\hat{X}}(\omega) \geq 1 \\ 1 & \text{se} \quad N_m^{\hat{X}}(\omega) = 0. \end{cases} \tag{5.11}$$

E o estimador $\hat{\mathcal{T}}_{BIC}\left(\hat{X}\right)$ pode ser representado por

$$\hat{\mathcal{T}}_{BIC}\left(\hat{X}\right) = \arg \max_{\hat{\mathcal{T}} \in \mathcal{F}(\hat{X},\, D(m))} \prod_{\omega \in \hat{\mathcal{T}}} \tilde{P}_\omega\left(\hat{X}\right), \tag{5.12}$$

onde

$$\tilde{P}_\omega\left(\hat{X}\right) = m^{-\frac{|E|-1}{2}} \tilde{P}_{L,\omega}(\hat{X}). \tag{5.13}$$

Como temos uma amostra bootstrap do processo markoviano $\boldsymbol{X}^*$, precisamos mostrar que a árvore estimada $\hat{\mathcal{T}}$ é uma boa estimativa para a árvore verdadeira $\mathcal{T}$. Portanto, para isso, apresentamos a seguinte definição e proposição, que é o crucial para a prova da consistência do nosso estimador BIC bootstrap. Primeiramente, apresentamos o procedimento de poda da árvore de contextos $k - full$.

Definição 5.1.1. *Dada uma amostra $\hat{X}$, seja S_d o conjunto de todos os contextos de tamanho máximo $d = D(m)$ e tal que $N_m^{\hat{X}}(\omega) \geq 1$. Para cada sequência $\omega \in S_d$ com $N_m(\omega) \geq 1$, definimos recursivamente, a partir das folhas da árvore $d - full$ $\hat{\mathcal{T}}_d$, o valor*

$$V_\omega^d(\hat{X}) = \begin{cases} \max \left\{ \tilde{P}_\omega(\hat{X}), \prod_{a \in E : N_T^{\hat{X}}(a\omega) \geq 1} V_{a\omega}^d(\hat{X}) \right\} & se \quad 0 \leq l(\omega) < d \\ \tilde{P}_\omega(\hat{X}) & se \qquad l(\omega) = d. \end{cases}$$

e a função indicadora

$$\mathcal{X}_\omega^d(\hat{X}) = \begin{cases} 1 \quad se \quad 0 \leq l(\omega) < d, \prod_{a \in E : N_T^{\hat{X}}(a\omega) \geq 1} V_{a\omega}^d(\hat{X}) > \tilde{P}_\omega(\hat{X}) \\ 0 \quad se \quad 0 \leq l(\omega) < d, \prod_{a \in E : N_T^{\hat{X}}(a\omega) \geq 1} V_{a\omega}^d(\hat{X}) \leq \tilde{P}_\omega(\hat{X}) \\ 0 \quad se \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad l(\omega) = d. \end{cases}$$

Para cada $\omega \in S_d$ o estimador BIC bootstrap $\hat{\mathcal{T}}$ é o conjunto dos contextos $\nu \succeq \omega$ tal que
$\hat{\mathcal{T}} := \left\{ \nu \in S_d : \mathcal{X}_\nu^d(\hat{X}) = 0,\ \mathcal{X}_\upsilon^d(\hat{X}) = 1,\ \forall \omega \preceq \upsilon \preceq \nu, \text{ se } \mathcal{X}_\omega^d(\hat{X}) = 1, \text{ e igual a } \{\omega\} \text{ se } \mathcal{X}_\omega^d(\hat{X}) = 0 \right\}$.

Proposição 5.1.1. *Seja $\hat{\mathbf{A}}^*$ um Estimador de Máxima Verossimilhança (EMV) da matriz das probabilidades de transição do processo markoviano $\boldsymbol{X}^*$, com lei $\hat{Q}$. E seja $\hat{X}$ uma amostra bootstrap de tamanho $m = O(T)$ vinda de $\hat{Q}$ fixada. Então, condicionalmente em $\hat{\mathbf{A}}^*$, para quase toda realização do processo $\boldsymbol{Z}$,*

i) $\frac{N_m^{\hat{X}}(\omega,a)}{m} \longrightarrow \hat{Q}(\omega a)$ *quase certamente quando* $m \to \infty$;
ii) $\frac{N_m^{\hat{X}}(\omega)}{m} \longrightarrow \hat{Q}(\omega)$ *quase certamente quando* $m \to \infty$;
iii) $\frac{\hat{Q}(\omega a)}{\hat{Q}(\omega)} \longrightarrow p(a|\omega)$ *quase certamente quando* $m \to \infty$.

Prova no Apêndice.

Agora, podemos apresentar o principal resultado desta tese.

Teorema 5.1.1. *Seja* $\hat{X}$ *uma amostra de tamanho* $m = O(T)$ *vinda de* $\hat{Q}$ *fixa. Para* $d(\mathcal{T}) < \infty$, *o estimador BIC bootstrap de* $\mathcal{T}$, *dada pela equação (5.12) é definido*

$$\hat{\mathcal{T}}_{BIC}\left(\hat{X}\right) = \arg \min_{\mathcal{T} \in \mathcal{F}(\hat{X},\, D(m))} BIC_{\mathcal{T}}(\hat{X}), \tag{5.14}$$

onde $BIC_{\mathcal{T}}(\hat{X})$ *é definda pela equação (2.3), mas agora utilizando a amostra bootstrap* $\hat{X}$. *Então*

$$\hat{\mathcal{T}}_{BIC}\left(\hat{X}\right) = \mathcal{T}$$

quase certamente quando $m \to \infty$.

No caso geral, temos que

$$\hat{\mathcal{T}}_{BIC}\left(\hat{X}\right)|_k = \mathcal{T}|_k,$$

quase certamente quando $m \to \infty$.

Prova no Apêndice.

5.2 Algoritmo Viterbi Modificado Para os Modelos Propostos

Uma vez estimados os contextos de uma VLMC oculta $\boldsymbol{X}$ através do algoritmo BIC bootstrap, podemos agora resolver os mesmos problemas que são abordados para o caso de um HMM ([17]). Um dos problemas abordados em um HMM é como obter a sequência X de estados ocultos que melhor explica a sequência de símbolos observados Z.

No nosso caso, o processo oculto não é uma cadeia de Markov e sim uma VLMC. Sendo assim, propomos uma versão adaptada do algoritmo de Viterbi, ver [17] para estimar a sequência mais provável de estados ocultos da VLMC $\boldsymbol{X}$. A seguir apresentaremos essa proposta de modificação do algoritmo de Viterbi para um VLHMM. O procedimento a seguir é válido tanto para o TSCM quanto para o TPCM. E também é válido tanto para o caso em que se tem uma VLMC com árvore de contextos $\mathcal{T}$ de comprimento finito ou infinito. Logo, apresentaremos apenas para o caso em que se tem um TSCM e para o caso de uma VLMC com árvore de contextos de comprimento finito.

Dada uma amostra observável z_1^T, de tamanho $T \in \mathbb{N}$, a fim de encontrar a sequência x_1^T mais provável de estados do processo oculto $\boldsymbol{X}$, definimos algumas variáveis auxiliares,

$$\begin{aligned}
\zeta_t(\omega) &= \max_{x_1^{t-l(\omega)}} \left[P\left(x_1^{t-l(\omega)}, x_{t-l(\omega)+1}^{t} = \omega, z_1^t \middle| \hat{\boldsymbol{\lambda}}_{\boldsymbol{S}} \right) \right], \\
\psi_t(\omega) &= \arg \max_{\substack{j \in E, \omega' \in E^{l(\omega')}: \\ \omega((j,\omega'))=\omega}} \left[\zeta_{t-1}(\omega) p(j|\omega) \right],
\end{aligned}$$

Por indução temos que

$$\zeta_{t+1}(\omega) = \max_{\substack{j \in E, \omega' \in E^{l(\omega')}: \\ \omega((j,\omega'))=\omega}} [\zeta_t(\omega) p(j|\omega)] \sum_{i \in E} b_j(z_{t+1}) I_{\{z_{t+1}=j \oplus i\}}.$$

Para $L \leq T$, seja $L = \max\left\{ l(\omega) : \omega \in \hat{\mathcal{T}} \right\}$, a computação das variáveis auxiliares é descrita a seguir:

1: Inicialização

$$\begin{aligned}\zeta_L(\omega) &= \pi_\omega b_\omega(z_L),\ \forall \omega \in \hat{\mathcal{T}},\\ \psi_1(\omega) &= 0.\end{aligned}$$

2: Recurssão

$$\begin{aligned}\zeta_t(\omega) &= \max_{\substack{j\in E,\omega'\in\ E^{l(\omega')}:\\ \omega((j,\omega'))=\omega}} [\zeta_{t-1}(\omega)p(j|\omega)] \sum_{i\in E} b_j(z_t) I_{\{z_t=j\oplus i\}},\ L\le t\le T, \omega\in\hat{\mathcal{T}},\\ \psi_t(\omega) &= \arg \max_{\substack{j\in E,\omega'\in\ E^{l(\omega')}\\ \omega((j,\omega'))=\omega}} [\zeta_{t-1}(\omega)p(j|\omega)],\ L\le t\le T, \omega\in\hat{\mathcal{T}}.\end{aligned}$$

3: Término

$$X_{T-l(\omega)+1}^{T} = \arg \max_{\substack{j\in E,\omega'\in\ E^{l(\omega')}:\\ \omega((j,\omega'))=\omega}} [\zeta_T(\omega)]$$

4: Sequência de Estados

$$X_{t-l(\omega)+1}^{t} = \psi_{t+1}(X_{t+l(\omega)}^{t+1}), t = T-l(\omega), T-l(\omega)-1, ..., l(\omega).$$

Assim, teremos uma estimação da sequência mais provável do processo oculto $\boldsymbol{X}$. A diferença entre esse procedimento e o Viterbi para um HMM é que neste procedimento existe a possibilidade, recursivamente, de se encontrar contextos ao longo da sequência oculta de tamanho T, ao invés de ir encontrando símbolo por símbolo como é feito no algoritmo original.

Capítulo 6

Simulação e Análise de Sensibilidade do Ruído Aleatório

Neste capítulo apresentaremos os resultados de algumas simulações com o objetivo de avaliar a metodologia proposta nesta tese. Nessas simulações estamos interessados em avaliar o impacto na estimação dos parâmetros do processo estocástico oculto $\boldsymbol{X}$ à medida em que aumentamos o grau da perturbação ϵ, tanto para o TSCM quanto para o TPCM. Analisando também o impacto nessas estimativas à medida em que aumentamos o tamanho da amostra.

Nas simulações usamos amostras de tamanho $T = 5000, 10.000$ e 30.000 com 100 repetições de Monte Carlo. Os verdadeiros valores do parâmetro de perturbação ϵ variaram de 0.01 até 0.99 com amplitude de 0.01. Para permitir fazer simulações com um refinamento tão grande no espaço paramétrico do ruído aleatório decidimos utilizar um alfabeto binário para diminuir o tempo das simulações, mas não há nenhuma restrição na metodologia quanto a usar alfabetos maiores.

Para as simulações, primeiramente foi gerada uma amostra de tamanho T de uma VLMC verdadeira $\boldsymbol{X}$ com matriz de transição fixa e conhecida, definimos um valor fixo e verdadeiro do parâmetro de ruído ϵ e geramos uma amostra da variável aleatória Bernoulli com o parâmetro ϵ definido e aplicamos o regime de perturbação TSCM e TPCM. Após isso, através da amostra perturbada Z, encontramos a matriz de transição $\mathbf{A}^*$ da cadeia de Markov $\boldsymbol{X}^*$ de ordem $k \in \mathbb{N}$ e encontramos a distribuição de emissão $\mathbf{B}^*$ para o valor definido do parâmetro de ruído ϵ.

Em seguida aplicamos o procedimento para recuperar a matriz de transição e a árvore de contextos $\mathcal{T}$ da VLMC oculta $\boldsymbol{X}$. Então, de acordo com a metodologia proposta, foi utilizado o algoritmo de Baum-Welch (algoritmo 1), que utiliza os procedimentos forward e backward e estimamos $\hat{\mathbf{A}}^*$ e $\hat{\mathbf{B}}^*$. Após isso, geramos uma amostra $\hat{X}$ de $\hat{\mathbf{A}}^*$ e utilizamos o algortimo BIC bootstrap (algoritmo 2) para estimar a árvore de contextos $\mathcal{T}$ da VLMC oculta $\boldsymbol{X}$.

Este capítulo é organizado da seguinte forma: apresentamos dois cenários de simulação com árvores com estruturas bem diferentes quanto ao número e disposição dos galhos. Para cada cenário aplicamos dois modelos de perturbação, TSCM e TPCM, e avaliamos as estimativas dos parâmetros dos modelos na medida em que aumentamos o grau de perturbação da amostra.

6.1 Primeiro Cenário: Modelo TSCM

Para essa primeira simulação escolhemos valores das probabilidades de transição as mais variadas possíveis, afim de verificar se haveria diferentes comportamentos nas estimativas dessas probabilidades.

Utilizamos uma VLMC $\boldsymbol{X}$ de ordem $k = 3$ com árvore de contextos $\mathcal{T}$ como mostra a Figura 6.1.

Algorithm 1 Computação do Vetor de Parâmetros $\boldsymbol{\lambda}^*$

Entrada: Valor inicial da Matriz das probabilidades de transição do processo $\boldsymbol{X^*}$, dada pela matriz $\mathbf{A}^*$, utilizando a amostra Z^*, Valor inicial da distribuição inicial $\boldsymbol{\pi^*}$ e Valor inicial da distribuição de emissão dada pela matriz $\mathbf{B}^*$, que utiliza o valor do parâmetro de ruído ϵ

Valor inicial de $\boldsymbol{\lambda}_0^* = \{\mathbf{A}^*, \mathbf{B}^*, \boldsymbol{\pi^*}\}$

$\Gamma =$ limite do número de iterações

$\eta > 0$ limite da melhoria de $P\left(Z^*|\boldsymbol{\lambda}^*\right)$

Inicialização:

$\hat{P} = P\left(Z^*|\boldsymbol{\lambda}_0^*\right)$

Repita

$P = \hat{P}$

Passo E

for $1 \leq r \leq T+k$ **do**

  for $\omega \in \{0,1\}^k$ **do**

$$\gamma_r(\omega) = \frac{\alpha_r(\omega)\beta_r(\omega)}{\sum_{\omega \in E^k} \alpha_r(\omega)\,\beta_r(\omega)}$$

$$\delta_r(\omega,\nu) = \frac{\alpha_r(\omega)p(\omega|\nu)b_\omega\left(z_{r+1}^*\right)\beta_{r+1}(\omega)}{\sum_{\omega,\nu \in E^k} \alpha_r(\omega)\,p(\omega|\nu)\,b_\omega\left(z_{r+1}^*\right)\beta_{r+1}(\omega)}$$

  end for

end for

Passo M

for $\omega \in E^k$ **do**

$$\hat{A}^* = \hat{a}(\omega|\nu) = \frac{\sum_{r=1}^{T+k-1} \delta_1(\omega,\nu)}{\sum_{r=1}^{T+k-1} \gamma_r(\omega)}$$

$$\hat{\mathbf{B}}^* = \hat{b}_\omega(\nu) = \frac{\sum_{r=1}^{T+k} I_{\{z_r^*=\nu\}}\gamma_r(\nu)}{\sum_{r=1}^{T+k} \gamma_r(\nu)}$$

end for

Retorne: $\hat{P} = P\left(Z^*|\hat{\boldsymbol{\lambda}}_1^*\right)$

até $\left(|\hat{P} - P| < \eta\right)$

Algorithm 2 Computação para estimar $\mathcal{T}$

Entrada: Amostra $\hat{X} = \{\hat{x}_1, ..., \hat{x}_m\}$: Gerada a partir da Matriz das probabilidades de transição estimadas de ordem k, $\hat{\mathbf{A}}^*$, encontradas através do algortimo 1.
Seja S_d o conjunto de todos os contextos de tamanho máximo $d = \log(m)$.

1: **for** $l(\omega) = d$ **do**
Calcule a variável
$V_\omega^d = \tilde{P}_\omega(\hat{X})$ dada pela equação (5.13)
E atribua o valor 0 a função
$\mathcal{X}_\omega^d(\hat{X})$
2: **end for**
3: **for** $0 \leq l(\omega) < d$ **do**
recursivamente calcule as variáveis

$$V_\omega^d(\hat{X}) = \max\left\{\tilde{P}_\omega(\hat{X}), \prod_{a\in A: N_m^{\hat{X}}(a\omega)\geq 1} V_{a\omega}^d(\hat{X})\right\}$$

E atribua o valor 1 a função
$\mathcal{X}_\omega^d(\hat{X})$, se $\prod_{a\in A: N_m^{\hat{X}}(a\omega)\geq 1} V_{a\omega}^d(\hat{X}) > \tilde{P}_\omega(\hat{X})$
E atribua o valor 0 a função
$\mathcal{X}_\omega^d(\hat{X})$, se $\prod_{a\in A: N_m^{\hat{X}}(a\omega)\geq 1} V_{a\omega}^d(\hat{X}) \leq \tilde{P}_\omega(\hat{X})$
4: **end for**
5: **for** $\omega \in S_d$ **do**
6: **for** $\nu \succeq \omega$ **do**
$\hat{\mathcal{T}} := \Big\{\nu \in S_d : \mathcal{X}_\nu^d(\hat{X}) = 0,\ \mathcal{X}_\upsilon^d(\hat{X}) = 1,\ \forall \omega \preceq \upsilon \preceq \nu$, se $\mathcal{X}_\omega^d(\hat{X}) = 1$, e igual a $\{\omega\}$ se $\mathcal{X}_\omega^d(\hat{X}) = 0\Big\}$.
7: **end for**
8: **end for**

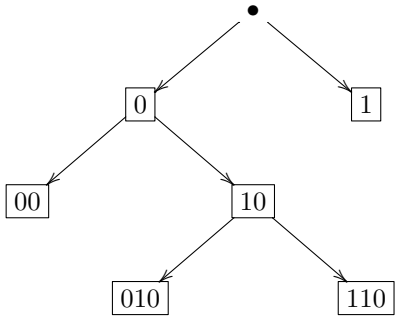


Figura 6.1. *Árvore $\mathcal{T}$ de um processo estocástico $\boldsymbol{X}$ de ordem $k = 3$.*

A matriz das probabilidades de transição verdadeira associada a árvore $\mathcal{T}$ do processo original $\boldsymbol{X}$ é dada pela Tabela 6.1.

Tabela 6.1. *Matriz de Transição do processo $\boldsymbol{X}$*

ω	$p(0\|\omega)$	$p(1\|\omega)$
010	*0.05*	*0.95*
110	*0.87*	*0.13*
00	*0.27*	*0.73*
1	*0.38*	*0.62*

A Tabela 6.2 mostra a estimativa do valor médio e do desvio padrão das 100 repetição de Monte Carlo, para diferentes valores do grau de perturbação ϵ, utilizando os modelos analisados TSCM e TPCM para diferentes tamanhos de amostra.

Vale salientar que as caselas em branco, para o modelo TPCM, informam que não conseguimos estimar o parâmetro de ruído ϵ, pois a amostra não foi suficientemente grande para o caso analisado. Isto é, para uma perturbação muito grande, as probabilidades de transição observadas (que são os valores iniciais utilizados no algoritmo de Baum-Welch) estão cada vez mais próximas de zero, o que prejudica o processo de estimação do parâmetro de ruído, mesmo para amostras grandes. Ou seja, se a pertutubação for muito grande, é preciso ter uma amostra bastante grande para poder conseguir estimar o parâmetro de perturbação.

Tabela 6.2. *Estimativas de alguns valores do parâmetro de rúido usando TSCM e TPCM*

N=5.000			*N=10.000*		*N=30.000*	
Ruido	*Estimativa*		*Estmativa*		*Estimativa*	
Real	*TSCM*	*TPCM*	*TSCM*	*TPCM*	*TSCM*	*TPCM*
0.01	*0.028± 0.016*	*0.029± 0.017*	*0.019± 0.011*	*0.020± 0.013*	*0.015± 0.008*	*0.017± 0.009*
0.05	*0.062± 0.015*	*0.064± 0.018*	*0.055± 0.012*	*0.058± 0.013*	*0.046± 0.008*	*0.054± 0.009*
0.25	*0.261± 0.017*	*0.259± 0.016*	*0.256± 0.013*	*0.253± 0.012*	*0.245± 0.007*	*0.246± 0.008*
0.45	*0.441± 0.015*	*0.462± 0.015*	*0.457± 0.012*	*0.443± 0.011*	*0.454± 0.008*	*0.455± 0.009*
0.55	*0.541± 0.016*	*0.558± 0.015*	*0.557± 0.011*	*0.544± 0.012*	*0.553± 0.006*	*0.556± 0.007*
0.75	*0.738± 0.018*	-	*0.742± 0.013*	*0.758± 0.014*	*0.753± 0.006*	*0.746± 0.007*
0.95	*0.943± 0.015*	-	*0.954± 0.012*	-	*0.947± 0.007*	-
0.99	*0.983± 0.013*	-	*0.986± 0.011*	-	*0.992± 0.006*	-

Podemos perceber que quando o verdadeiro valor do ruído é muito pequeno, 1% por exemplo, uma amostra de tamanho 5000 não é suficiente para fornecer boas estimativas, uma vez que a amostra perturbada teria apenas por volta de 50 valores trocados. Mas mesmo assim o verdadeiro valor do parâmetro de ruído está contido no intervalo estimado. Percebemos também que a variabilidade das estimativas diminui á medida que amostra aumenta e o intervalo estimado fica bem menor. Para amostras maiores, 10000 e 30000, temos estimativas pontuais acuradas dos parâmetros.

Observamos que á medida que a perturbação aumenta a estimativa pontual do parâmetro de ruído fica cada vez mais próxima do valor verdadeiro, mesmo para amostras pequenas, e que a variabilidade diminui á medida que o tamanho da amostra aumenta, concluindo que a estimativa é cada vez mais precisa.

A Tabela 6.3 mostra o valor médio e o desvio padrão das estimativas das probabilidades de transição nas 100 repetição de Monte Carlo, para um ruído igual a $\epsilon = 0.01$ e as probabilidades de transição verdadeiras. Percebemos que as estimativas das probabilidades de transição continuam próximas das probabilidades verdadeiras. O que era de se esperar, uma vez que houve poucas mudanças de símbolos, então teríamos que ter as probabilidades estimadas próximas das verdadeiras. E notamos também que á medida que o tamanho da amostra aumenta, as estimativas ficam cada vez mais próximas das verdadeiras e com menor variabilidade.

Tabela 6.3. *Matriz de Transição Estimada do TSCM para um ruído* $\epsilon = 0.01$

	N=5.000		*N=10.000*		*N=30.000*	
ω	$\hat{p}(0\|\omega)$	$p(0\|\omega)$	$\hat{p}(0\|\omega)$	$p(0\|\omega)$	$\hat{p}(0\|\omega)$	$p(0\|\omega)$
010	*0.041± 0.027*	*0.05*	*0.060± 0.016*	*0.05*	*0.046± 0.010*	*0.05*
110	*0.884± 0.026*	*0.87*	*0.880± 0.018*	*0.87*	*0.874± 0.009*	*0.87*
00	*0.260± 0.027*	*0.27*	*0.261± 0.019*	*0.27*	*0.274± 0.009*	*0.27*
1	*0.361± 0.026*	*0.38*	*0.369± 0.018*	*0.38*	*0.374± 0.011*	*0.38*

A Figura 6.2 mostra o comportamento das estimativas das probabilidades de transição $\hat{p}(0|00) = 0.27$ e $\hat{p}(0|110) = 0.87$, para o parâmetro de perturbação igual a 1% para amostras de tamanho 5.000 e 10.000. O comportamento das demais probabilidades de transição, nesse caso, foram bem próximas. A Figura 6.2 mostra também uma evidência de normalidade no comportamento das estimativas das probabilidades de transição á medida que se aumenta o tamanho da amostra.

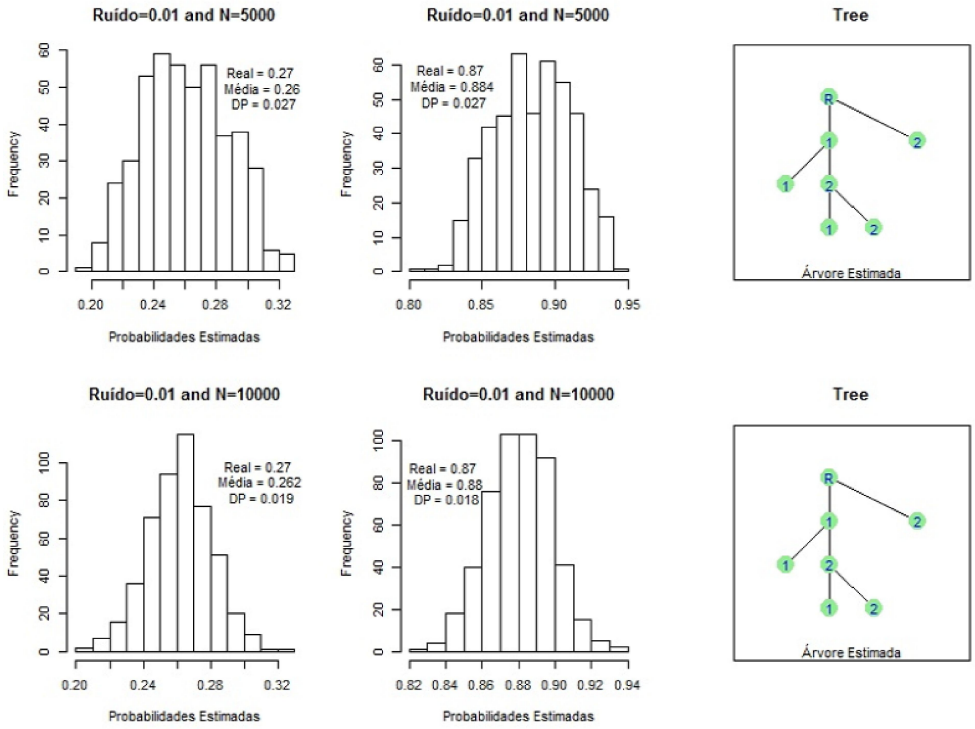


Figura 6.2. *Histograma das Probabilidades de Transição e Gráfico da Árvore Estimada*

E evidenciamos que a estimativa da árvore oculta através do algoritmo BIC bootstrap coincide com a verdadeira

árvore de contextos, tanto para uma amostra de tamanho 5.000 quanto para o caso em que se tem uma amostra de tamanho 10.000.

Para um grau de perturbação de $\epsilon = 0.05$, percebemos as mesmas evidências de normalidade no comportamento das estimativas das probabilidades de transição á medida que se aumenta o tamanho da amostra. E notamos, pela Tabela 6.4, que mostra a média e os desvios padrão das estimativas das probabilidades de transição, que conseguimos ótimas estimativas das probabilidades de transição do processo oculto $\boldsymbol{X}$, conseguindo também estimar a verdadeira árvore de contextos, mesmo para o caso em que a amostra é pequena.

Tabela 6.4. *Matriz de Transição Estimada do TSCM para um ruído* $\epsilon = 0.05$

N=5.000			*N=10.000*		*N=30.000*	
ω	$\hat{p}(0\|\omega)$	$p(0\|\omega)$	$\hat{p}(0\|\omega)$	$p(0\|\omega)$	$\hat{p}(0\|\omega)$	$p(0\|\omega)$
010	*0.087± 0.029*	*0.05*	*0.076± 0.020*	*0.05*	*0.068± 0.015*	*0.05*
110	*0.891± 0.028*	*0.87*	*0.885± 0.021*	*0.87*	*0.862± 0.014*	*0.87*
00	*0.284± 0.026*	*0.27*	*0.279± 0.022*	*0.27*	*0.275± 0.013*	*0.27*
1	*0.337± 0.029*	*0.38*	*0.350± 0.021*	*0.38*	*0.362± 0.013*	*0.38*

A Figura 6.3 mostra o comportamento das estimativas das probabilidades de transição $\hat{p}(0|00) = 0.27$ e $\hat{p}(0|110) = 0.87$, para diferentes valores do parâmetro de perturbação ϵ para amostras de tamanho 5.000 e 10.000. Podemos observar claramente o impacto nas estimativas das probabilidades de transição. Observamos que fora do intervalo de 40% e 60%, conseguimos ótimas estimativas das probabilidades de transição verdadeiras e, nota-se também uma diminuição na variabilidade das estimativas com o aumento no tamanho da amostra.

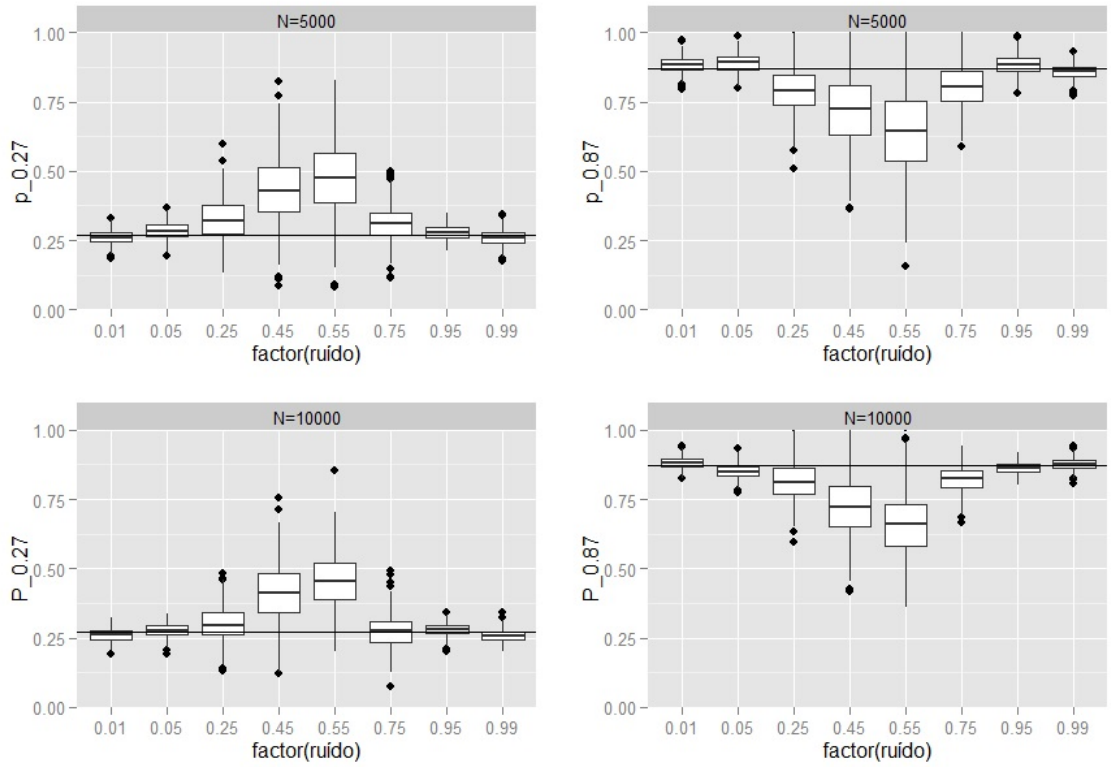


Figura 6.3. *Gráfico das Probabilidades de Transição em função do ruído*

Vale salientar que á medida que o ruído se aproxima de 50%, apesar de o parâmetro de ruído ser bem estimado, as estimativas das probabilidades de transição tendem a se aproximar de 50%. E á medida que o ruído se afasta do

valor de 50% temos boas estimativas para as probabilidades de transição. Uma observação importante a respeito dos resultados diz respeito a estimação da árvore de contextos quando o ruído aumenta. Quando se tem um ruído no intervalo de 40% a 60% o algoritmo BIC bootstrap estima um modelo independente, ou seja, uma árvore apenas com a raiz. Isso se deve ao fato de que todas as estimativas das probabilidades de transição ficam em torno de 50%, tornando assim impraticável a estimação correta dos parâmetros do modelo e consequentemente da árvore de contextos.

Portanto, se o valor do ruído estimado estiver fora do intervalo de 40% a 60%, teremos boas estimativas das probabilidades de transição e consequentemente da árvore de contextos do verdadeiro processo. Caso o ruído estimado esteja entre 40% a 60%, podemos concluir que não teremos uma boa estimação das probabilidades de transição, portanto, não teremos estimação da árvore de contextos verdadeira.

6.2 Primeiro Cenário: Modelo TPCM

Avaliando agora a simulação para o caso em que utilizamos o modelo TPCM, usando a mesma matriz de transição dada pela Tabela 6.1. Percebemos pelos resultados mostrados na Tabela 6.5, que as estimativas das probabilidades de transição ficam bastante próximas das verdadeiras probabilidades, para o caso em que se tem um grau de perturbação de $\epsilon = 0.01$, se tornando cada vez mais precisas com o aumento da amostra.

Tabela 6.5. *Matriz de Transição Estimada do TPCM para um ruído* $\epsilon = 0.01$

	N=5.000		*N=10.000*		*N=30.000*	
ω	$\hat{p}(0\|\omega)$	$p(0\|\omega)$	$\hat{p}(0\|\omega)$	$p(0\|\omega)$	$\hat{p}(0\|\omega)$	$p(0\|\omega)$
010	*0.044± 0.026*	*0.05*	*0.062± 0.015*	*0.05*	*0.055± 0.010*	*0.05*
110	*0.863± 0.027*	*0.87*	*0.882± 0.018*	*0.87*	*0.871± 0.008*	*0.87*
00	*0.261± 0.027*	*0.27*	*0.264± 0.019*	*0.27*	*0.277± 0.008*	*0.27*
1	*0.362± 0.026*	*0.38*	*0.371± 0.018*	*0.38*	*0.376± 0.011*	*0.38*

Para o caso em que se tem um grau de perturbação igual a 5%, percebemos, através da Tabela 6.6, que as estimativas das probabilidades de transição usando o TPCM também estão próximas das probabilidades verdadeiras e ficam mais precisas á medida que aumentamos o aumento da amostra.

Tabela 6.6. *Matriz de Transição Estimada do TPCM para um ruído* $\epsilon = 0.05$

	N=5.000		*N=10.000*		*N=30.000*	
ω	$\hat{p}(0\|\omega)$	$p(0\|\omega)$	$\hat{p}(0\|\omega)$	$p(0\|\omega)$	$\hat{p}(0\|\omega)$	$p(0\|\omega)$
010	*0.089± 0.029*	*0.05*	*0.078± 0.022*	*0.05*	*0.066± 0.016*	*0.05*
110	*0.862± 0.028*	*0.87*	*0.881± 0.021*	*0.87*	*0.867± 0.016*	*0.87*
00	*0.257± 0.029*	*0.27*	*0.281± 0.021*	*0.27*	*0.278± 0.014*	*0.27*
1	*0.341± 0.028*	*0.38*	*0.353± 0.022*	*0.38*	*0.364± 0.014*	*0.38*

Notamos que, mesmo para uma amostra pequena, tamanho 5000, conseguimos fazer estimação das probabilidade de transição para valores do ruído abaixo de 55%. Para uma amostra de tamanho 10.000 conseguimos fazer estimações das probabilidades de transição até o caso em que o ruído é no máximo 75%. Porém, na estimação da árvore de contextos, para os casos em que o ruído está acima de 40%, não foi possível estimar a verdadeira árvore de contextos da VLMC oculta, uma vez que o algoritmo BIC bootstrap estimou um modelo independente, mesmo para amostras grandes, visto que, como observamos no gráfico, as probabilidades de transição ficam cada vez mais próximas de 50%.

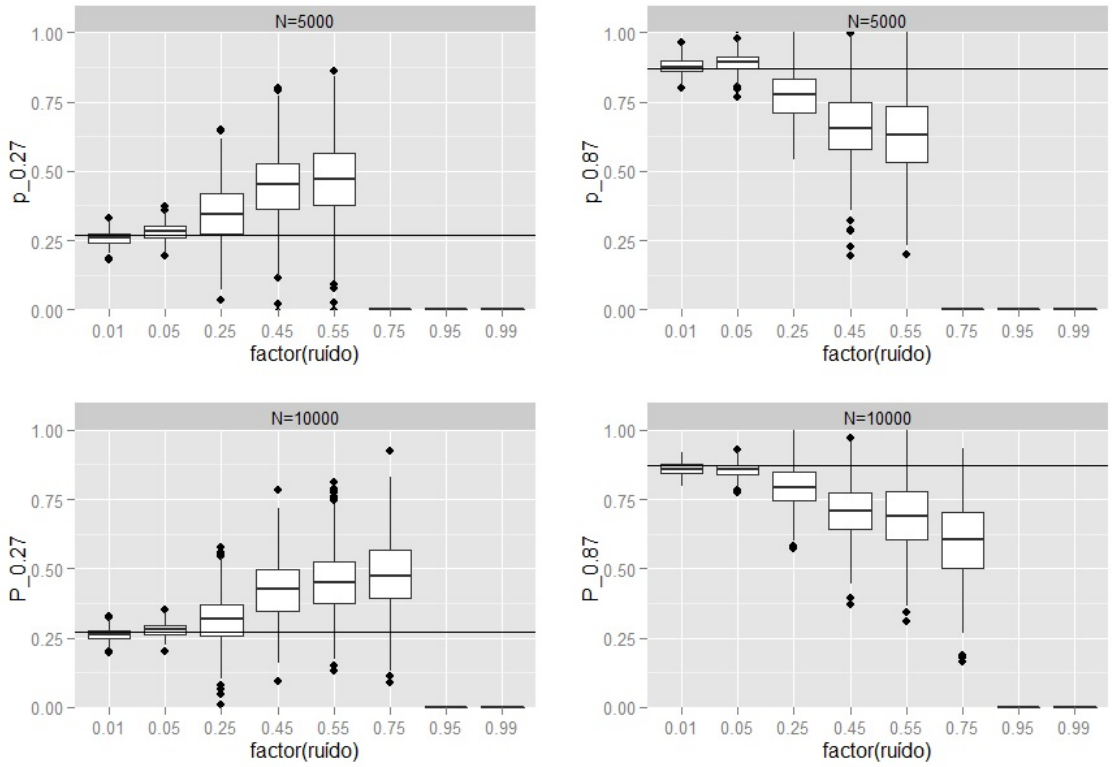


Figura 6.4. *Gráfico das Probabilidades de Transição em função do ruído*

No entanto, em todos os outros casos, conseguimos estimar os parâmetros do modelo TPMC e o algoritmo BIC bootstrap conseguiu encontrar a verdadeira árvore de contextos $\mathcal{T}$.

6.3 Segundo Cenário: Modelo TSCM

Para a segunda simulação utilizamos uma matriz de transição associada ao processo $\boldsymbol{X}$, que é mostrada na Tabela 6.7.

Tabela 6.7. *Matriz de Transição do processo* $\boldsymbol{X}$

ω	$P(0\|\omega)$	$P(1\|\omega)$
0000	*0.10*	*0.90*
1000	*0.50*	*0.50*
100	*0.83*	*0.17*
10	*0.25*	*0.75*
1	*0.25*	*0.75*

A árvore de contextos $\mathcal{T}$ associada a $\boldsymbol{X}$ é apresentada na figura 6.5 (ordem $k = 4$).

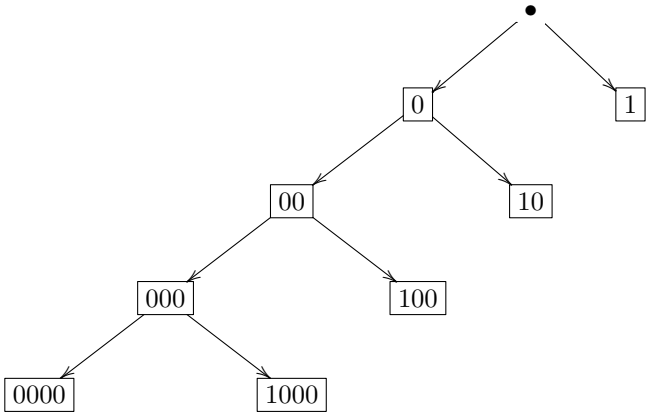


Figura 6.5. *Árvore $\mathcal{T}$ do processo estocástico $\boldsymbol{X}$ de ordem $k = 4$.*

A Tabela 6.8 apresenta o valor médio e o desvio padrão das estimativas das probabilidades de transição das 100 repetições de Monte Carlo e o valor verdadeiro das probabilidades de transição. Não foram colocados os resultados para o caso em que se tem uma amostra de tamanho 5.000, pois esse tamanho de amostra não foi suficiente para estimar a verdadeira árvore. Isso se deve à estrutura mais complexa dessa árvore e à ordem maior.

Tabela 6.8. *Matriz de Transição Estimada do TSCM para um ruído $\epsilon = 0.01$*

	N=10.000		*N=30.000*	
ω	$\hat{p}(0\|\omega)$	$p(0\|\omega)$	$\hat{p}(0\|\omega)$	$p(0\|\omega)$
0000	*0.132± 0.019*	*0.10*	*0.112± 0.012*	*0.10*
1000	*0.532± 0.018*	*0.50*	*0.515± 0.011*	*0.50*
100	*0.838± 0.015*	*0.83*	*0.825± 0.009*	*0.83*
10	*0.258± 0.016*	*0.25*	*0.246± 0.011*	*0.25*
1	*0.243± 0.018*	*0.25*	*0.253± 0.011*	*0.25*

A Figura 6.6 mostra o comportamento das estimativas das probabilidades de transição $\hat{p}(0|10) = 0.25$ e $\hat{p}(0|100) = 0.83$, para diferentes valores do parâmetro de perturbação ϵ, para amostras de tamanho 10.000 e 30.000.

De acordo com a Figura 6.6, assim como na primeira simulação, concluímos que as estimativas das probabilidades de transição estão próximas das verdadeiras e com menor variabiliadde para valores do ruído fora do intervalo de 40% a 60%.

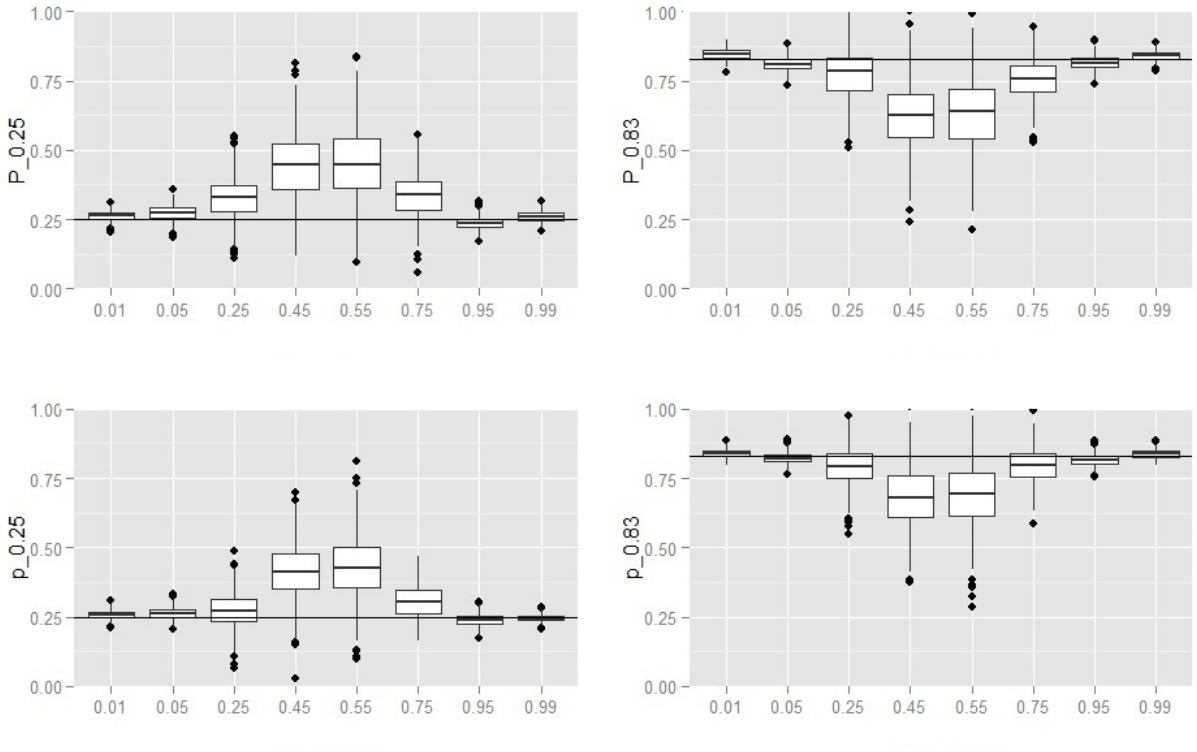


Figura 6.6. *Gráfico das Probabilidades de Transição em função do ruído*

Em relação a variabilidade das estimativas, notamos que existe um intervalo onde a variabilidade também aumenta, para tamanho de amostra fixo, mas diminui com o tamanho da amostra.

6.4 Segundo Cenário: Modelo TPCM

Para o modelo TPCM, podemos verificar, através da Figura 6.7 o mesmo comportamento apresentado no primeiro cenário para amostras de tamanho 10.000 e 30.000. Verificamos que quando se mantém o ruído fixo, a variabilidade das estimativas das probabilidades de transição diminui com o tamanho da amostra, mas quando se tem o tamanho da amostra fixa, a variabilidade aumenta com o aumento do ruído.

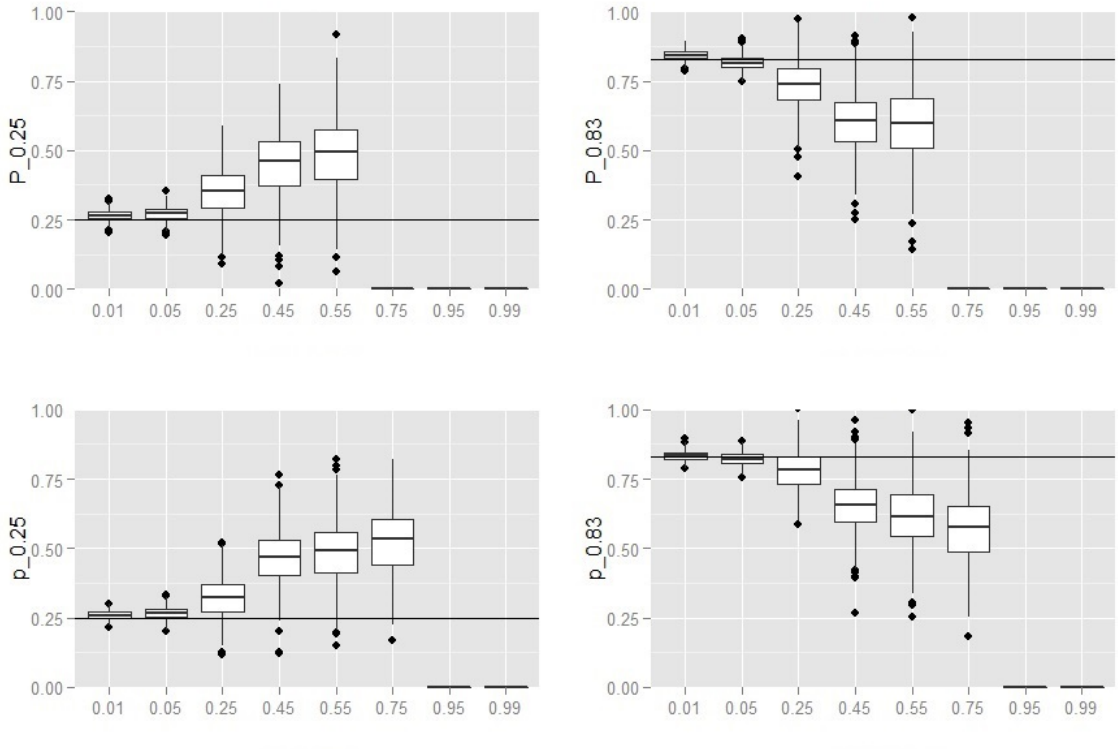


Figura 6.7. *Gráfico das Probabilidades de Transição em função do ruído*

Portanto, quando utilizamos o modelo TSCM, em ambas simulações, tivemos o mesmo comportamento nas estimações das probabilidades de transição e do parâmetro de ruído estimado. E quando utilizamos o modelo TPCM podemos observar o mesmo comportamento no impacto das probabilidades de transição estimadas em relação ao parâmetro de ruído.

Capítulo 7

Critério de Seleção de Modelos: TSCM ou TPCM

A escolha do modelo apropriado, do ponto de vista estatístico, é um tópico extremamente importante na análise de dados. Na situação que abordamos nesta tese, a pergunta que queremos responder é a seguinte:

Dada uma amostra do processo perturbado Z, qual dos modelo estudados aqui é mais adequado para fazer a estimação dos parâmetros do processo oculto X e do parâmetro de perturbação?

A nossa proposta para responder a essa pergunta é a seguinte: primeiro fazemos a estimação dos parâmetros do modelo, seja TSCM ou TPCM e, verificamos, através da divergência de Kullback-Leibler ou da máxima verossimilihança, qual o modelo mais adequado, dada a amostra perturbada Z.

Usando a divergência de KL como critério de seleção, primeiro utilizamos o vetor estimado $\hat{\boldsymbol{\lambda}}^*$ para gerar uma nova amostra perturbada $\hat{Z}$ através das equações

$$\hat{Z}_t = \hat{X}_t \oplus \hat{\xi}_t, \tag{7.1}$$

quando utilizado o modelo TCSM, e

$$\hat{Z}_t = \hat{X}_t \cdot \hat{\xi}_t, \tag{7.2}$$

para o modelo TPCM respectivamente.

Desse modo, utilizamos a amostra observada Z e as novas amostras perturbadas $\hat{Z}_1$, $\hat{Z}_2$ geradas pelos modelos TSCM e TPCM, respectivamente, e encontramos a divergência de Kullback-Leibler entre a lei da amostra observada Z e a estimada com cada um dos modelos, como a seguir

$$D_{\mathrm{KL}}(p(Z)|p(\hat{Z})_j) = \sum_i p(z_i|z_{i-k}) \ \log \frac{p(z_i|z_{i-k})}{\hat{p}\,(\hat{z}_i|\hat{z}_{i-k})}, \ \forall\ i = 1, ..., T. \tag{7.3}$$

Assim, escolhermos aquele modelo M que nos fornecer menor divergência de Kullback-Leibler, isto é

$$M = \arg\min_j \left(D_{\mathrm{KL}}(p(Z)|p(\hat{Z}_j) \right).$$

Para usar o critério de máxima verossimilhança, dado o vetor de parâmetros estimado $\hat{\boldsymbol{\lambda}}^*$ usado para encontrar a árvore de contextos $\hat{\mathcal{T}}$ da VLMC oculta $\boldsymbol{X}$, escolhemos aquele modelo M, $M = TSCM$ ou $TPCM$, tal que a função de verossimilhança $\mathbb{L}_{\hat{\mathcal{T}}_M}(\hat{X})$ seja máxima.

Dada uma árvore de contextos estimada $\hat{\mathcal{T}}$ (para o caso em que $d(\mathcal{T} < \infty)$) ou $\hat{\mathcal{T}}\big|_k$ (para o caso em que $d(\mathcal{T} = \infty)$), assumindo valores em um alfabeto E e dada uma amostra $\hat{X}$ com ordem de tamanho $m = O(T)$, gerada através da árvore estimada $\hat{\mathcal{T}}$ ou $\hat{\mathcal{T}}\big|_k$, a função de verossimilhança $\mathbb{L}_{\hat{\mathcal{T}}_M}(\hat{X})$ é definida por

$$\mathbb{L}_{\hat{\mathcal{T}}_M}(\hat{X}) = P(\hat{X}_k^1 = a_k^1) \prod_{\omega \in \hat{\mathcal{T}}} \prod_{u \in\ E} \hat{p}(u|\omega)^{N_T^{\hat{X}}(\omega u)}$$

em que $k = \max\left\{l(\omega) : \omega \in \hat{\mathcal{T}}\right\}$ e $N_T(\omega u) = \sum_{t=k}^{T} I\left\{a_t^{t-l(\omega)} = \omega u\right\}$, onde $\hat{p}(u|\omega)$ são as probabilidades de transição estimadas.

7.1 Simulação 1: Modelo TSCM como verdadeiro

Para verificar a eficácia do critério de seleção (usando a distância de Kullback-Leibler) foram realizados dois estudos de simulação diferentes. No primeiro estudo fixamos o modelo TSCM como sendo o verdadeiro modelo gerador da amostra perturbada Z. Em seguida estimamos os parâmetros utilizando os dois modelos propostos e verificamos se o critério de seleção consegue indentificar qual é o verdadeiro modelo. No segundo estudo de simulação fixamos o modelo TPCM como sendo o verdadeiro modelo gerador da amostra perturbada Z e também estimamos os parâmetros através dos modelos propostos e verificamos a proporção de acertos do critério de seleção.

Para essas simulações, foram utilizadas amostras de tamanho $T = 5000, 10.000, 30.000$ e 50.000 com 500 repetições de Monte Carlo. Para o ruído de perturbação foi utilizada uma sequência de variáveis aleatórias Bernoulli independentes e idênticamente distribuídas $\boldsymbol{\xi}$, independente do processo $\boldsymbol{X}$ cujo parâmetro de ruído ϵ variou de 0.01 até 0.99 com amplitude de 0.01.

Utilizamos como o processo oculto uma VLMC $\boldsymbol{X}$ binária de ordem $k = 3$ com árvore de contextos $\mathcal{T}$ dada na na Figura 7.1.

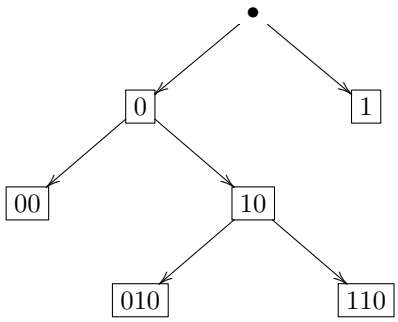


Figura 7.1. *Árvore verdadeira $\mathcal{T}$ do processo $\boldsymbol{X}$*

A matriz das probabilidades de transição verdadeira associada ao processo $\boldsymbol{X}$ é apresntada na Tabela 7.1.

Tabela 7.1. *Matriz de Transição de $\boldsymbol{X}$*

ω	$P(0\|\omega)$	$P(1\|\omega)$
010	*0.05*	*0.95*
110	*0.87*	*0.13*
00	*0.27*	*0.73*
1	*0.38*	*0.62*

No primeiro estudo de simulação verificamos, através da Figura 7.2, que à medida que o grau de perturbação aumenta, o critério de seleção através da distância de Kullback-Leibler consegue selecionar de maneira perfeita o verdadeiro modelo gerador de uma dada amostra perturbada. Observamos também que para um grau de perturbação abaixo de 20% o critério já é bastante eficiente em nos dizer de qual modelo a amostra perturbada Z foi gerada. O que é bastante razoável de se pensar, uma vez que o modelo TPCM é inflacionado de zeros (nesse exemplo em que o alfabeto é binário), então a amostra gerada se tornará cada vez mais composta por símbolos zeros á medida em que se aumenta o parâmetro de perturbação, portanto, é esperado que seja gerada uma amostra perturbada $\hat{Z}$ bem diferente da perturbada original Z.

E na medida em que o tamanho da amostra aumenta essa eficência em selecionar o verdadeiro modelo é evidente. Ou seja, dada uma amostra perturbada, podemos estimar os parâmetros do VLHMM através do algoritmo de Baum-Welch, com qualquer um dos dois modelos estudados, e depois utilizar o critério de seleção de modelos para escolher o melhor entre eles.

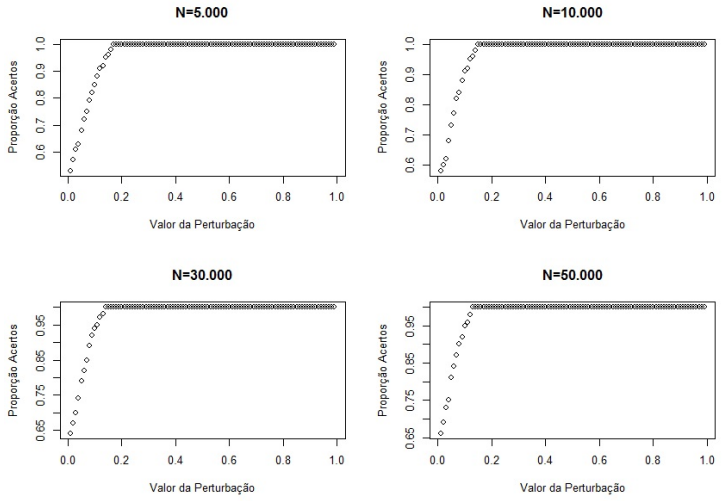


Figura 7.2. *Proporção de acertos do modelo TCSM através da divergência de KL*

7.2 Simulação 2: Modelo TPCM como verdadeiro

Quando o modelo verdadeiro é o TPCM, percebemos que o critério de seleção também identifica o verdadeiro modelo (ver Figura 7.3). Notamos que quando o parâmetro de ruído passa dos 20% já conseguimos identificar perfeitamente de onde os dados são provenientes.

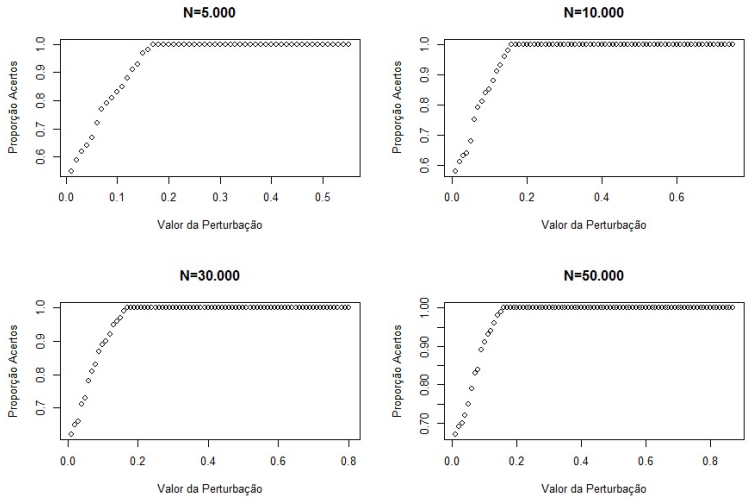


Figura 7.3. *Proporção de acertos do modelo TPCM através da divergência de KL*

Capítulo 8

Aplicação

Os dados e as informações dessa aplicação foram gentilmente cedidos pelo *Laboratório de Neurofisiologia da Visão* da UFMG, coordenado pelo Dr Jerome Baron e fazem parte da tese de doutorado apresentada ao Programa de Pós-Graduação em Ciências Biológicas - Fisiologia e Farmacologia do Instituto de Ciências Biológicas, da Universidade Federal de Minas Gerais, pela aluna Claudiana Souza Amorim.

Os animais utilizados neste estudo foram corujas buraqueiras (Athene cunicularia) obtidas por doação do Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis (IBAMA).

Uma câmara de registro de baixo peso foi implantada cirurgicamente sobre a área craniana de acesso à região de interesse. Essa mesma câmara foi utilizada para fixar a cabeça das aves durante os registros. Os registros foram feitos em corujas acordadas com restrição de movimentos. Inicialmente o animal foi submetido a um período de adaptação aos ambiente e câmara de registros, utilizando reforço positivo.

A atividade neuronal estudada foi obtida a partir do registro de potenciais de ação de um pequeno grupo de neurônios localizados ao redor do eletrodo. Foram a todo, 39 neurônios localizados ao redor do eletrodo. Os estímulos eram apresentados em grades senoidais (com barras pretas e brancas) variando em 16 direções de movimento das barras, iniciando da direção 0^o (as barras se movem para a direita) e os passos são de $22{,}5^o$, como pode ser visto pela Figura 8.1. Assim a última direção é $337{,}5^o$. Porém, essas 16 condições são apresentadas de forma aleatória, isto é, pode começar e terminar de qualquer uma das 16. Cada estímulo foi apresentado 10 vezes em ordem pseudo-aleatória durante 2 ou 4 s, precedido e seguido da apresentação do fundo de tela durante 1 e 2 s respectivamente.

Os registros dos potencias de ação dos neurônios isolados foram submetidos a um procedimento conhecido como spike sorting, usado para separar os potencias de ação de células individuais com base nas diferenças das formas de ondas apresentadas. Formas provenientes de uma mesma célula tendem a ser semelhantes e por isso tendem a ocupar posições próximas em um espaço paramétrico, formando aglomerados bem definidos. Essas semelhanças são definidas quanto às características das formas de onda, como a amplitude do pico, vale e largura. Portanto, devido ao fato de esses spikes poderem ser, por razões técnicas, erroneamente medidos consideramos que a sequência de spikes dos neurônios observados no tempo pode ser modelada como um processo estocástico que pode ter sofrido uma perturbação por um ruído aleatório e a ordem de dependência no passado pode não ser fixa. Portanto, usamos o VLHMM como modelo para esse banco de dados.

O banco de dados analisado nessa tese é formado pelos tempos onde aconteceram os potenciais de ação de células individuais (spikes). Assim, para cada um dos 16 estímulos, repetidos 10 vezes, temos 39 neurônios. O intervalo de tempo de observação foi de 4 segundos para cada estímulo em cada um dos neurônios em cada repetição. Sendo que destes 4 segundos, o primeiro e o último segundo são de repouso, ou seja, sem apresentação de estímulo. E entre 1 e 3 segundos, foram apresentados os estímulos.

Temos então uma matriz de tamanho 160×39, onde cada linha da matriz é um vetor contendo os tempos de observações dos spikes. A fim de fazer a aplicação dos modelos e metodologias apresentados nesta tese, categorizamos os tempos de observação em um espaço binário, onde o valor 1 foi atribuído aquele tempo onde ocorreu o spike e o valor 0 ao tempos onde não aconteceram os spikes.

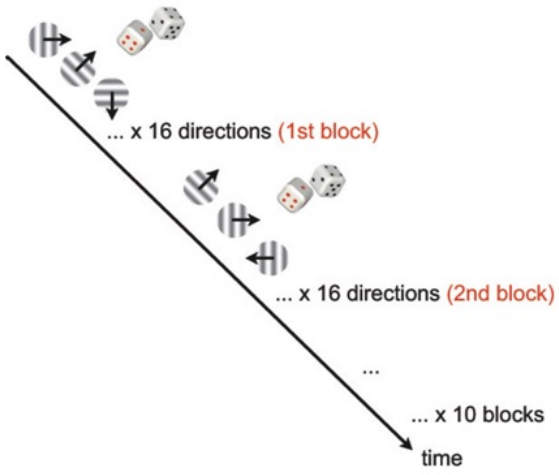


Figura 8.1. *Protocolo de Direção dos Estímulos*

Cada spike tem uma duração de 1.2 ms nesse banco. Assim, o intervalo de tempo entre 1 e 3 segundos (intervalo da apresentação dos estímulos), foi dividido por 1.2 ms. Dessa maneira, criamos um intervalo de tempos ,$t \pm 0.006$, de observação do spike, em que t é o tempo do pico da observação de um spike. Portanto, cada intervalo de observação do spike é formado por uma sequência de 12 símbolos iguais a 1, e fora desse intervalo até o próximo intervalo de spikes temos uma sequência de símbolos observados iguais a 0, como pode ser observado na Figura 6.2.

$$\overbrace{000000000000}^{0}\overbrace{000000000000}^{0}\underbrace{111111111111}_{1}\overbrace{000000000000}^{0}$$

Figura 8.2. *Procedimento de construção do banco de dados*

Logo, para cada sequência de 12 símbolos iguais a 0, atribuímos um único símbolo igual a 0, e para cada sequência de 12 símbolos iguais a 1, um símbolo igual a 1.

O interesse da modelagem está em saber se existe diferença de comportamento da sequencia de spikes nos neurônios quando submetidos a distintos estímulos. Sendo assim, assumimos uma mesma lei de probabilidades para todos os 39 neurônios quando em repouso (não submetidos a estímulos) em relação ao tempo entre os spikes. Por sugestão da equipe do Laboratório escolhemos comparar dois estímulos, o 13 e o 7, nos quais se esperava encontrar padrões distintos de resposta dos spikes. Utilizamos ambos modelos, TSCM e TPCM, e aplicamos nosso critério de seleção de modelos para decidir qual deles era o mais adequado ao banco de dados.

A seguir apresentamos os resultados obtidos usando o modelo TSCM

Tabela 8.1. *Matriz de Transição do Estimulo 13 e 07 usando o TSCM*

Estímulo 13			*Estímulo 07*		
ω	$\hat{p}(0\|\omega)$	$\hat{p}(1\|\omega)$	ω	$\hat{p}(0\|\omega)$	$\hat{p}(1\|\omega)$
1	*0.78*	*0.22*	*1*	*0.79*	*0.21*
110	*0.63*	*0.37*	*0110*	*0.69*	*0.31*
1010	*0.59*	*0.41*	*1110*	*0.46*	*0.54*
0010	*0.83*	*0.17*	*1010*	*0.62*	*0.38*
1100	*0.67*	*0.33*	*0010*	*0.84*	*0.16*
0100	*0.81*	*0.19*	*1100*	*0.62*	*0.38*
1000	*0.81*	*0.19*	*0100*	*0.80*	*0.20*
0000	*0.96*	*0.04*	*1000*	*0.83*	*0.17*
-	-	-	*0000*	*0.96*	*0.04*

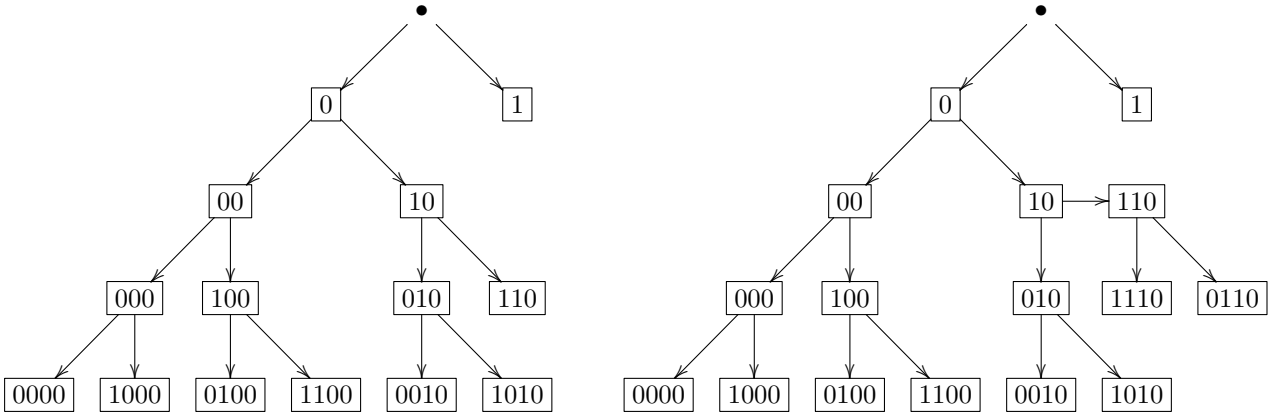


Figura 8.3. *Árvores de Contexto Estimada dos Estímulos usando o TSCM*

O parâmetro de ruído estimado para o estímulo 13 usando o TSCM foi de $\hat{\epsilon} = 0.01011$. Significando que a capacidade do sistema captar o spike, condicionado ao fato de que a coruja teve um spike é de 0.98989. Para o estímulo 07, usando o TSCM, o parâmetro de ruído estimado foi de $\hat{\epsilon} = 0.01085$, indicando que dado que se teve um spike, o sistema tem aproximadamente $1,01\%$ de chance de não identificar o spike. Apesar das simulações mostrarem que quando se estima um parâmetro de ruído pequeno, possivelmente esse valor está superestimado, ou seja, tanto para o estímulo 13 e 07 possivelmente se tem um valor de perturbação abaixo de $1,01\%$. Por outro lado, vimos também através das simulações, se o parâmetro de ruído estimado for pequeno, as probabilidades de transição do verdadeiro processo será bem estimado.

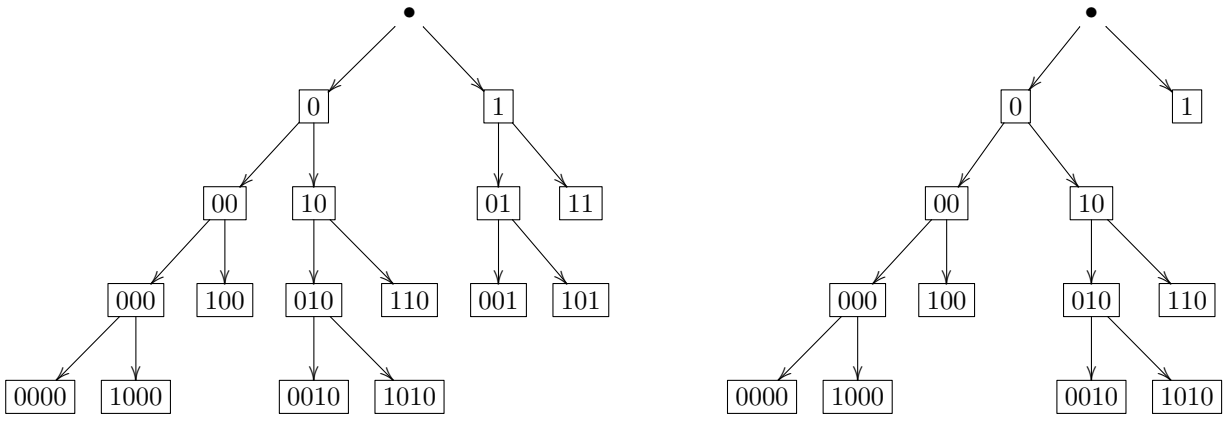
Tabela 8.2. *Matriz de Transição Estimada para o Estímulo 13 e 07 usando o TPCM*

Estímulo 13			*Estímulo 07*		
ω	$\hat{p}(0\|\omega)$	$\hat{p}(1\|\omega)$	ω	$\hat{p}(0\|\omega)$	$\hat{p}(1\|\omega)$
11	*0.38*	*0.62*	*1*	*0.80*	*0.20*
101	*0.64*	*0.36*	*110*	*0.73*	*0.27*
001	*0.85*	*0.15*	*100*	*0.78*	*0.22*
110	*0.59*	*0.41*	*1010*	*0.61*	*0.39*
100	*0.78*	*0.21*	*0010*	*0.82*	*0.18*
1010	*0.64*	*0.36*	*1000*	*0.82*	*0.18*
0010	*0.80*	*0.20*	*0000*	*0.96*	*0.04*
1000	*0.78*	*0.21*	-	-	-
0000	*0.95*	*0.05*	-	-	-

Depois refizemos as estimações usando o modelo TPCM. Podemos ver, através da Figura 8.4, uma diferença maior entre as árvores estimadas. Isso pode nos levar a crer que o comportamento de atividade neuronal das corujas é diferente dependendo do tipo de estímulo a que ela é submetida. As árvores de contextos estimadas para os dois estímulos usando o TPCM apresentam alguns galhos a mais para o estímulo 13. A Tabela 8.2 mostra as probabilidades de transição dos 2 estímulos.

O parâmetro de ruído estimado para o estímulo 13 usando o TPCM foi de $\hat{\epsilon} = 0.00997$. O que significa que a probabilidade do sistema captar o spike, dado que a coruja teve um spike é de 0.99003. Enquanto que para o estímulo 07, usando o TPCM, o parâmetro de ruído estimado foi de $\hat{\epsilon} = 0.01204626$, indicando que condicionado ao fato de ser um spike, o sistema tem somente aproximadamente $1,2\%$ de chance de não identificar o spike.

Fizemos então o critério de seleção de modelos para saber qual tipo de regime de perturbação a amostra seria proveniente, se do regime TSCM ou TPCM.



Estímulo 13 Estímulo 7

Figura 8.4. *Árvores de Contexto Estimada dos Estímulos usando o TPCM*

Após a estimação das probabilidades de transição e da árvore de contextos dos estímulos 13 e 07 utilizando os dois regimes de perturbação, fizemos 500 repetições de Monte Carlo em cada caso. E através da divergência de Kullback-Leibler, observamos que em $89,6\%$ das vezes, o critério selecionou o regime de perturbação TPCM ao regime TSCM quando comparado o estímulo 13. E quando considerado o estímulo 07 o critério mostrou que existe $87,3\%$ das vezes o regime de perturbação dos dados também é do TPCM.

Ou seja, tanto para o estímulo 13 e 07 o critério de seleção de modelos através da divergência de KL mostrou que é mais verossímil que os dados sejam provenientes de uma perturbação TPCM.

Capítulo 9

Conclusão

Nesta tese definimos alguns modelos estocasticamente perturbados tendo como base os modelos definidos por [7] e [12]. A partir desses modelos propusemos algumas extensões e propusemos metodologias para fazer inferência para os parâmetros de tais modelos.

Conseguimos mostrar que, através das metodologias propostas, é possível recuperar a verdadeira árvore de contextos de uma VLMC estocasticamente perturbada e saber o grau de tal perturbação, a depender do grau de perturbação e do regime de perturbação associado. Propusemos um estimador BIC bootstrap, cuja convergência forte foi demonstrada, para as probabilidades de transição da VLMC oculta.

Mostramos, através de simulações, que para amostras acima de 10000 observações a precisão das estimativas é bastante satisfatória em um intervalo razoável com estimativas pontuais dos parâmetros próximas dos valores verdadeiros e com variância pequena que diminui com o aumento da amostra.

Quando temos uma VLMC binária e perturbada de acordo como o modelo TPCM com ruído Bernoulli, mostramos que, dependendo do tamanho da amostra, existe um valor limite que o parâmetro do ruído pode assumir no qual nos permite fazer a estimação das probabilidades de transição e consequentemente da verdadeira árvore de contextos. Mas que, quando se tem um ruído abaixo de 40% conseguimos fazer boas estimativas do ruído aleatório, das probabilidades de transição e da árvore de contextos mesmo para amostras pequenas (5000). A partir de um ruído acima desse valor a amostra perturbada vai se tornando cada vez mais inflacionada de zeros, no caso do alfabeto binário, se tornando cada vez mais difícil a recuperação da verdadeira lei de formação da VLMC oculta.

Apesar das simulações terem sido realizadas com apenas ruídos Bernoulli, a metodologia pode ser aplicada a qualquer tipo de distribuição de emissão, assumindo valores em qualquer alfabeto discreto, assim como a VLMC oculta pode assumir valores em qualquer alfabeto discreto.

Conseguimos, através do critério de seleção de modelos, identificar entre os analisados nesta tese, qual seria o mais provável pelo qual os dados sofreram (ou não) alguma perturbação. E na aplicação conseguimos identificar a existência de diferentes comportamentos na atividade neuronal de corujas em relação ao tipo de estímulo visual a que foram submetidas e com isso temos ferramenta para comparar os estímulos através das leis de formação de cada um.

Propomos também uma modificação no algoritmo de Viterbi para encontar a sequência oculta mais provável de uma VLMC que sofreu algum tipo perturbação.

Capítulo 10

Limitações da Pesquisa e Sugestões para Trabalhos Futuros

Os modelos definidos nesta tese e a metodologia proposta se mostraram, nas simulações, capazes de recuperar as estimativas das probabilidades de transição em relação ao grau do parâmetro de ruído. Porém, não foi mostrado, matematicamente, o motivo pelo qual, tanto as probabilidades de transição do processo perturbado $\boldsymbol{Z}$, quanto as probabilidades de transição estimadas de um processo binário oculto $\boldsymbol{X}$ ficam em torno de 50%, quando se tem um rúido Bernoulli. Acreditamos que se avaliarmos a entropia do processo perturbado $\boldsymbol{Z}$ obteremos respostas de tal comportamento.

Como trabalho futuro pretendemos implementar um algoritmo para o modelo TMCM, mostrando a convergência de estimadores e algoritmos. Pretendemos também desenvolver um critério de seleção de modelos que leve em conta os três modelos em questão.

Acreditamos que esta tese serve como referencial de pesquisas futuras para generalização de qualquer tipo de perturbação e qualquer tipo de distribuição de emissão (contínuo e discreto), uma vez que mostramos que podemos utilizar uma amostra da matriz de transição das probabilidades estimadas, desde que sejam estimadores de máxima verossimilhança, para estimar a árvore de contextos da VLMC oculta através o algorimto BIC bootstrap proposto.

Capítulo 11

Apêndice

Neste Apêndice são apresentadas as provas dos resultados propostos nessa tese.

Demonstração. **Proposição 3.1.1**

Seja $\boldsymbol{Z}$ um processo perturbado de acordo com o TSCM. **i)** Para todo $z_0, a_0, b_0 \in E$ e todo $\omega \in \mathcal{T}$

$$P\left(Z_0 = z_0 | X^0_{-l(\omega)+1} = \omega\right) = P(Z_0 = z_0 | X_0 = a_0),$$

para algum $\omega = a^0_{-l(\omega)+1} \in \mathcal{T}$, com $l(\omega)$ temos que

$$P\left(Z_0 = z_0 | X^0_{-l(\omega)+1} = \omega\right) = \frac{P\left(Z_0 = z_0, X_0 = a_0, ..., X_{-l(\omega)+1} = a_{-l(\omega)+1}\right)}{P\left(X_0 = a_0, ..., X_{-l(\omega)+1} = a_{-l(\omega)+1}\right)}.$$

O evento $\{Z_0 = z_0\}$ pode ser escrito em termos de $\boldsymbol{X}$ e $\boldsymbol{\xi}$, de acordo com o TSCM, como

$$\{Z_0 = z_0\} = \bigcup_{\substack{x_0, b_0 = 0: \\ z_0 = x_0 \oplus b_0}}^{|E|-1} \{X_0 = x_0, \xi_0 = b_0\}.$$

Portanto,

$$P\left(Z_0 = z_0 | X^0_{-l(\omega)+1} = \omega\right) = \frac{P\left(\bigcup_{\substack{x_0, b_0 = 0: \\ z_0 = x_0 \oplus b_0}}^{|E|-1} \{X_0 = x_0, \xi_0 = b_0\}, X_0 = a_0, ..., X_{-l(\omega)+1} = a_{-l(\omega)+1}\right)}{P\left(X_0 = a_0, X_{-1} = a_{-1}, ..., X_{-l(\omega)+1} = a_{-l(\omega)+1}\right)}.$$

Note que $\{X_0 = x_0, X_0 = a_0\}$ são conjuntos vazios se $x_0 \neq a_0$, então

$$P\left(Z_0 = z_0 | X^0_{-l(\omega)+1} = \omega\right) = \frac{P\left(X_0 = a_0, \xi_0 = b_0, ..., X_{-l(\omega)+1} = a_{-l(\omega)+1}\right) I_{\{z_0 = a_0 \oplus b_0\}}}{P\left(X_0 = a_0, \xi_0 = b_0, ..., X_{-l(\omega)+1} = a_{-l(\omega)+1}\right)}.$$

Então, pela independência de $\boldsymbol{X}$ e $\boldsymbol{\xi}$ temos que

$$\begin{aligned} P\left(Z_0 = z_0 | X^0_{-l(\omega)+1} = \omega\right) &= \frac{P\left(\xi_0 = b_0\right) P\left(X_0 = a_0, ..., X_{-l(\omega)+1} = a_{-l(\omega)+1}\right) I_{\{z_0 = a_0 \oplus b_0\}}}{P\left(X_0 = a_0, ..., X_{-l(\omega)+1} = a_{-l(\omega)+1}\right)} \\ &= P\left(\xi_0 = b_0\right) I_{\{z_0 = a_0 \oplus b_0\}}. \end{aligned} \tag{11.1}$$

Por outro lado, temos que

$$P\left(Z_0 = z|X_0 = a_0\right) = \frac{P\left(\bigcup\limits_{\substack{x_0,b_0=0:\\ z_0=x_0\oplus b_0}}^{|E|-1}\left\{X_0 = x_0, \xi_0 = b_0\right\}, X_0 = a_0\right)}{P\left(X_0 = a_0\right)}.$$

Como os eventos $\{X_0 = x_0, X_0 = a_0\}$ são vazios para todo $x_0 \neq a_0$, então, temos somente os eventos $\{X_0 = a_0\}$. Note que os eventos $\{X_0 = a, \xi_0 = b_0\}$ são mutuamente exclusivos e $\boldsymbol{X}$ é independente de $\boldsymbol{\xi}$, portanto

$$\begin{aligned} P\left(Z_0 = z_0|X_0 = a_0\right) &= \frac{P\left(X_0 = a_0, \xi_0 = b_0\right) I_{\{z_0=a_0\oplus b_0\}}}{P\left(X_0 = a_0\right)} \\ &= P\left(\xi_0 = b_0\right) I_{\{z_0=a_0\oplus b_0\}}. \end{aligned} \tag{11.2}$$

Isso conclui a proba do item **i)**.

ii) Queremos mostra que as probabilidades de transição do processo observado $\boldsymbol{Z}$, truncado em alguma ordem $k \in \mathbb{N}$, $\forall\ z_0 \in E$ e $\forall z_{-k}^{-1} \in E^k$, são:

$$P\left(Z_0 = z_0|Z_{-k}^{-1} = z_{-k}^{-1}\right) = \frac{\sum\limits_{\substack{a_t,b_t\in E:\\ -k\leq t\leq 0}} \prod\limits_{t=-k}^{0} P(\xi_t = b_t) P\left(\bigcap\limits_{t=-k}^{0}\left\{X_t = a_t\right\}\right) I_{\{z_0=a_0\oplus b_0\}} \prod\limits_{t=-k}^{-1} I_{\{z_t=a_t\oplus b_t\}}}{\sum\limits_{\substack{a_t,b_t\in E:\\ -k\leq t\leq -1}} \prod\limits_{t=-k}^{-1} P(\xi_t = b_t) P\left(\bigcap\limits_{t=-k}^{-1}\left\{X_t = a_t\right\}\right) \prod\limits_{t=-k}^{-1} I_{\{z_t=a_t\oplus b_t\}}}.$$

Para $\boldsymbol{Z}$ perturbado de acordo com o TSCM, truncado na ordem k, $P\left(Z_0 = z_0|Z_{-k}^{-1} = z_{-k}^{-1}\right)$ pode ser escrito como

$$P\left(Z_0 = z_0|Z_{-k}^{-1} = z_{-k}^{-1}\right) = \frac{P\left(Z_0 = z_0, Z_{-1} = z_{-1}, ..., Z_{-k} = z_{-k}\right)}{P\left(Z_{-1} = z_{-1}, ..., Z_{-k} = z_{-k}\right)}.$$

Como no item **i)** na Proposição 3.1.1 podemos escrever os eventos $\{Z_t = z_t\}$ em termos de $\boldsymbol{X}$ e $\boldsymbol{\xi}$,

$$P\left(Z_0 = z_0|Z_{-k}^{-1} = z_{-k}^{-1}\right) = \frac{P\left(\bigcap\limits_{-k\leq t\leq 0}\left[\bigcup\limits_{\substack{a_t,b_t=0:\\ z_t=a_t\oplus b_t}}^{|E|-1}\left\{X_t = a_t, \xi_t = b_t\right\}\right]\right)}{P\left(\bigcap\limits_{-k\leq t\leq -1}\left[\bigcup\limits_{\substack{a_t,b_t=0:\\ z_t=a_t\oplus b_t}}^{|E|-1}\left\{X_t = a_t, \xi_t = b_t\right\}\right]\right)}.$$

Pela propriedade distributiva $A \cap \{B \cup C\} = \{A \cap B\} \cup \{A \cap C\}$, temos que

$$P\left(Z_0 = z_0|Z_{-k}^{-1} = z_{-k}^{-1}\right) = \frac{P\left(\bigcup\limits_{\substack{a_t,b_t=0:\\ z_t=a_t\oplus b_t}}^{|E|-1}\left[\bigcap\limits_{-k\leq t\leq 0}\left\{X_t = a_t, \xi_t = b_t\right\}\right]\right)}{P\left(\bigcup\limits_{\substack{a_t,b_t=0:\\ z_t=a_t\oplus b_t}}^{|E|-1}\left[\bigcap\limits_{-k\leq t\leq -1}\left\{X_t = a_t, \xi_t = b_t\right\}\right]\right)}.$$

Como $\{X_t = a_t, \xi_t = b_t\}$ são mutuamente exclusivos,

$$P\left(Z_0 = z_0 | Z_{-k}^{-1} = z_{-k}^{-1}\right) = \frac{\sum_{a_t,b_t=0}^{|E|-1} P\left(\bigcap_{-k\leq t\leq 0} \{X_t = a_t, \xi_t = b_t\}\right) \prod_{t=-k}^{0} I_{\{z_t = a_t \oplus b_t\}}}{\sum_{a_t,b_t=0}^{|E|-1} P\left(\bigcap_{-k\leq t\leq -1} \{X_t = a_t, \xi_t = b_t\}\right) \prod_{t=-k}^{-1} I_{\{z_t = a_t \oplus b_t\}}}.$$

Finalmente a afirmação segue pela independência de $\boldsymbol{X}$ e $\boldsymbol{\xi}$. □

Demonstração. **Proposição 3.2.1** As provas dos itens **i), ii)** são análogas para provar a Proposição 3.1.1, mas trocando somente a função indicadora $a \oplus b$ por $a \cdot b$, $\forall a, b \in E$. □

Demonstração. **Proposição 5.1.1** Seja $\hat{\mathbf{A}}^*$ um EMV da matriz das probabilidades de transição do processo markoviano $\boldsymbol{X}^*$, com lei $\hat{Q}$, e seja $\hat{X}$ uma amostra bootstrap de tamanho $m = O(T)$ vinda de $\hat{Q}$ fixa, para quase toda realização do processo $\boldsymbol{Z}$,

i) Podemos escrever

$$\frac{N_m^{\hat{X}}(\omega, a)}{m} = \frac{\sum_{t=k}^{m} 1\left\{\hat{x}_t^{t+k} = \omega, \hat{x}_{t+k+1} = a\right\}}{m}$$

Então, a variável aleatória $\frac{N_m^{\hat{X}}(\omega,a)}{m}$ condicionalmente em $\hat{\boldsymbol{A}}^*$, converge quase certamente para

$$E(1_{\left\{\hat{x}_t^{t+k} = \omega, \hat{x}_{t+k+1} = a\right\}} | \hat{\boldsymbol{A}}^*) = \hat{Q}(\omega a), \text{ quando } m \longrightarrow \infty$$

pelo Teorema Ergódico, onde $\hat{Q}(\omega a)$ é a medida da sequência ωa dada $\hat{\boldsymbol{A}}^*$.

ii) Analogamente como no item **i)** temos que

$$\frac{N_m^{\hat{X}}(\omega)}{m} \longrightarrow \hat{Q}(\omega), \text{ quase certamente quando } m \longrightarrow \infty.$$

iii) Dos itens **i)** e **ii)** temos que

$$\frac{N_m^{\hat{X}}(\omega,a)}{N_m^{\hat{X}}(\omega)} \longrightarrow \hat{p}(a|\omega) = p(a|\omega), \text{ quase certamente quando } m \longrightarrow \infty.$$

Note que $\hat{p}(a|\omega)$ é um EMV das probabilidades de transição da cadeia de Markov oculta $\boldsymbol{X^*}$. Então, para cada $\omega \in E^k$ e $a \in E$ e para quase toda realização do processo $\boldsymbol{Z}$, temos que $\hat{p}(a|\omega) \longrightarrow p(a|\omega)$ quase certamente quando $m = O(T) \to \infty$. A prova da convergência do EMV dos parâmetros de um HMM é apresentada em [16]. □

Demonstração. Prova do **Teorema** 5.1.1.

Proposição 5.1.1 aplicada aos Lemmas 3.1, 3.2 e Proposições 4.3 e 4.4 apresetnadas em [9] implica na convergência do Teorema 5.1.1. □

Demonstração. **Equivalência entre a log-verossimilhança e a Divergência de Kullback-Leibler** Seja $\boldsymbol{Z}$ ser o processo perturbado verdadeiro de acordo como os modelos TSCM ou TPCM com distribuição empírica $p(.)$, e seja $\hat{\boldsymbol{Z}}$ com distribuição empírica $\hat{p}(.)$ ser o processo perturbado estimado como definido através da equação (5.7 ou 5.8).

Seja $\tilde{\boldsymbol{\lambda}}^*$ ser um vetor de estimativas dos parâmetros do HMM ($\boldsymbol{Z}^*$,$\boldsymbol{Y}$). Então um estimador de $\boldsymbol{\lambda}^*$ é dado por

$$\hat{\boldsymbol{\lambda}}^* = \arg\min_{\tilde{\boldsymbol{\lambda}}^*} D_{\mathrm{KL}}(p|\hat{p}) \tag{11.3}$$

Vamos mostrar a seguinte equivalência

$$\arg\max_{\tilde{\boldsymbol{\lambda}}^* \in \boldsymbol{\Lambda}} \mathbb{L}\left(\tilde{\boldsymbol{\lambda}}^* \,\middle|\, Z^*\right) = \arg\min_{\tilde{\boldsymbol{\lambda}}^* \in \boldsymbol{\Lambda}} D_{\mathrm{KL}}(p|\hat{p})$$

Como o vetor de parâmetros $\tilde{\boldsymbol{\lambda}}^*$ é estimado usando o algorimto EM de Baum-Welch algorithm, logo é um estimador de máxima verossimilhança dada uma amostra de tamanho T do processo perturbado $\boldsymbol{Z}$, portanto temos que

$$\arg\max_{\tilde{\boldsymbol{\lambda}}^*} \mathbb{L}\left(\tilde{\boldsymbol{\lambda}}^* \middle| Z\right) = \arg\max_{\tilde{\boldsymbol{\lambda}}^*} \prod_i \hat{p}\left(z_i|z_{i-k}, \tilde{\boldsymbol{\lambda}}^*\right) = \arg\min_{\tilde{\boldsymbol{\lambda}}^*} -\frac{1}{T}\sum_i \log\left[\hat{p}\left(z_i|z_{i-k}, \tilde{\boldsymbol{\lambda}}^*\right)\right]$$

Temos que pelo teorema ergódico, para um vetor de parâmetros $\tilde{\boldsymbol{\lambda}}^*$ fixado

$$-\frac{1}{T}\sum_i \log\left[\hat{p}\left(z_i|z_{i-k}, ..., \tilde{\boldsymbol{\lambda}}^*\right)\right] \overset{q.c}{\to} \mathbb{E}_{\tilde{\boldsymbol{\lambda}}^*}\left[-\log\left(\hat{p}\left(z_i|z_{i-k}, \tilde{\boldsymbol{\lambda}}^*\right)\right)\right] \tag{11.4}$$

Temos que $\mathbb{E}_{\tilde{\boldsymbol{\lambda}}^*}\left[-\log\left(\hat{p}\left(z_i|z_{i-k}, \tilde{\boldsymbol{\lambda}}^*\right)\right)\right]$ pode ser escrita como sendo

$$\mathbb{E}_{\tilde{\boldsymbol{\lambda}}^*}\left[-\log\left(\hat{p}\left(z_i|z_{i-k}, \tilde{\boldsymbol{\lambda}}^*\right)\right)\right] + \mathbb{E}\left[-\log\left(p\left(z_i|z_{i-k},\right)\right)\right] - \mathbb{E}\left[-\log\left(p\left(z_i|z_{i-k},\right)\right)\right]$$

ou seja,

$$\begin{aligned}\mathbb{E}_{\tilde{\boldsymbol{\lambda}}^*}\left[-\log\left(\hat{p}\left(z_i|z_{i-k}, \tilde{\boldsymbol{\lambda}}^*\right)\right)\right] &= \mathbb{E}_{\tilde{\boldsymbol{\lambda}}^*}\left[\log\frac{p\left(z_i|z_{i-k}\right)}{\hat{p}\left(z_i|z_{i-k}, \tilde{\boldsymbol{\lambda}}^*\right)}\right] - \\ &- \mathbb{E}\left[\log\left(p\left(z_i|z_{i-k}\right)\right)\right]\end{aligned}$$

Então

$$\mathbb{E}_{\tilde{\boldsymbol{\lambda}}^*}\left[-\log\left(\hat{p}\left(z_i|z_{i-k}, \tilde{\boldsymbol{\lambda}}^*\right)\right)\right] = D_{\mathrm{KL}}(p|\hat{p}) - \mathbb{E}\left[\log\left(p\left(z_i|z_{i-k}\right)\right)\right]$$

e pela equação (11.4) temos então que

$$-\frac{1}{T}\sum_i \log\left[\hat{p}\left(z_i|z_{i-k}, \tilde{\boldsymbol{\lambda}}^*\right)\right] \overset{q.c}{\to} D_{\mathrm{KL}}(p\|\hat{p}) - \mathbb{E}\left[\log\left(p\left(z_i|z_{i-k}\right)\right)\right]$$

onde o segundo termo não é função de $\tilde{\boldsymbol{\lambda}}^*$.

Portanto, maximizar a verossimilhança ou minimizar a divergência de Kullback-Leibler divergence conduz a estimadores equivalentes quando o tamanho da amostra tende ao infinito [1]. □

Demonstração. **Proposição 4.1**

i) Queremos mostrar que as probabilidades de transição $P\left(Z_0 = z_0|Z_{-k}^{-1} = z_{-k}^{-1}\right)$ do processo $\boldsymbol{Z}$ truncadas em alguma ordem $k \in \mathbb{N}$, $\forall\ z_t, a_t, c_t \in E$, $b_t = \{0, 1\}$ e $\forall\ z_{-k}^{-1} \in\ E^k$, são dadas por:

$$P\left(Z_0 = z_0|Z_{-k}^{-1} = z_{-k}^{-1}\right) = \frac{\displaystyle\sum_{\substack{a_t, b_t, c_t: \\ -k\le t\le 0}} \left\{\prod_{t=-k}^{0} P(\xi_t = b_t) P\left(\bigcap_{-k\le t\le 0}\{X_t = a_t\}\right) P\left(\bigcap_{-k\le t\le 0}\{Y_t = c_t\}\right)\right\} \prod_{t=-k}^{0} I_{\{z_t = b_t.a_t + (1-b_t)c_t\}}}{\displaystyle\sum_{\substack{a_t, b_t, c_t: \\ -k\le t\le -1}} \left\{\prod_{t=-k}^{-1} P(\xi_t = b_t) P\left(\bigcap_{-k\le t\le -1}\{X_t = a_t\}\right) P\left(\bigcap_{-k\le t\le 1}\{Y_t = c_t\}\right)\right\} \prod_{t=-k}^{-1} I_{\{z_t = b_t.a_t + (1-b_t)c_t\}}} \tag{11.5}$$

Considere um processo perturbado $\boldsymbol{Z}$ de acordo como o modelo TMCM, truncado na order k, então usando o mesmo raciocínio, como usado para a demosntração da Proposição 3.1.1, para escrever os eventos $\{Z_t = z_t\}$ em

função das variáveis $\boldsymbol{X}$, $\boldsymbol{Y}$ e ξ, temos que para todo $t \in \mathbb{Z}$

$$P\left(Z_0 = z_0 | Z_{-k}^{-1} = z_{-k}^{-1}\right) = \frac{P\left(\bigcap\limits_{-k\leq t\leq 0}\left[\bigcup\limits_{\substack{z_t,a_t,c_t,b_t:\\ z_t=b_t.a_t+(1-b_t)c_t}} \{X_t = a_t, Y_t = c_t, \xi_t = b_t\}\right]\right)}{P\left(\bigcap\limits_{-k\leq t\leq -1}\left[\bigcup\limits_{\substack{a_t,c_t,b_t:\\ z_t=b_t.a_t+(1-b_t)c_t:}} \{X_t = a_t, Y_t = c_t, \xi_t = b_t\}\right]\right)}$$

usando a propriedade distribuitiva das operações de conjuntos, temos que

$$P\left(Z_0 = z_0 | Z_{-k}^{-1} = z_{-k}^{-1}\right) = \frac{P\left(\bigcup\limits_{\substack{z_t,a_t,c_t,b_t:\\ z_t=b_t.a_t+(1-b_t)c_t}}\left[\bigcap\limits_{-k\leq t\leq 0} \{X_t = a_t, Y_t = c_t, \xi_t = b_t\}\right]\right)}{P\left(\bigcup\limits_{\substack{z_t,a_t,c_t,b_t:\\ z_t=b_t.a_t+(1-b_t)c_t}}\left[\bigcap\limits_{-k\leq t\leq -1} \{X_t = a_t, Y_t = c_t, \xi_t = b_t\}\right]\right)}$$

Como os eventos $\{X_t = a_t, Y_t = c_t, \xi_t = b_t\}$ são mutuamente exclusivos, temos que

$$P\left(Z_0 = z_0 | Z_{-k}^{-1} = z_{-k}^{-1}\right) = \frac{\sum\limits_{z_t,a_t,c_t,b_t:} P\left(\bigcap\limits_{-k\leq t\leq 0} \{X_t = a_t, Y_t = c_t, \xi_t = b_t\}\right) \prod\limits_{t=-k}^{0} I_{\{z_t=b_t.a_t+(1-b_t)c_t\}}}{\sum\limits_{z_t,a_t,c_t,b_t:} P\left(\bigcap\limits_{-k\leq t\leq 1} \{X_t = a_t, Y_t = c_t, \xi_t = b_t\}\right) \prod\limits_{t=-k}^{-1} I_{\{z_t=b_t.a_t+(1-b_t)c_t\}}}$$

e pela independência entre $\boldsymbol{X}$,$\boldsymbol{Y}$ e ξ, temos que

$$P\left(Z_0 = z_0 | Z_{-k}^{-1} = z_{-k}^{-1}\right) = \frac{\sum\limits_{\substack{a_t,b_t,c_t:\\ -k\leq t\leq 0}} \left\{\prod\limits_{t=-k}^{0} P(\xi_t = b_t) P\left(\bigcap\limits_{-k\leq t\leq 0} \{X_t = a_t\}\right) P\left(\bigcap\limits_{-k\leq t\leq 0} \{Y_t = c_t\}\right)\right\} \prod\limits_{t=-k}^{0} I_{\{z_t=b_t.a_t+(1-b_t)c_t\}}}{\sum\limits_{\substack{a_t,b_t,c_t:\\ -k\leq t\leq -1}} \left\{\prod\limits_{t=-k}^{-1} P(\xi_t = b_t) P\left(\bigcap\limits_{-k\leq t\leq -1} \{X_t = a_t\}\right) P\left(\bigcap\limits_{-k\leq t\leq 1} \{Y_t = c_t\}\right)\right\} \prod\limits_{t=-k}^{-1} I_{\{z_t=b_t.a_t+(1-b_t)c_t\}}}$$

o que prova o item **i)**

ii) Queremos mostrar que dada uma amostra de tamanho T do processo perturbado $\boldsymbol{Z}$, e o tamanho da amostra é tal que $l(\omega) \leq T, \forall \omega \in \mathcal{T}_X$, $l(\nu) \leq T, \forall \nu \in \mathcal{T}_Y$, e para $k = \max\{(l(\omega), l(\nu)) : \omega \in \mathcal{T}_X,\ \nu \in \mathcal{T}_Y$, então a função de verossimilhança $\mathbb{L}(\boldsymbol{\lambda_M}|Z)$ do processo perturbado $\boldsymbol{Z}$ pode ser escrita como:

$$\mathbb{L}(\boldsymbol{\lambda_M}|Z) = \sum_{\substack{a_t,b_t,c_t:\\ -k\leq t\leq T}} \left\{\prod_{t=-k}^{T} P(\xi_t = b_t) P\left(\bigcap_{-k\leq t\leq T} \{X_t = a_t\}\right) P\left(\bigcap_{-k\leq t\leq 1} \{Y_t = c_t\}\right)\right\} \prod_{t=-k}^{T} I_{\{z_t=b_t.a_t+(1-b_t)c_t\}} \tag{11.6}$$

A prova é a construção do item **ii)**, pois a função de verossimilhança $\mathbb{L}(\boldsymbol{\lambda}|Z_T^{-1} = \nu_T^{-1})$ é justamente o numerador da probabilidade $P\left(Z_0 = \nu_0 | Z_{-k}^{-1} = \nu_{-k}^{-1}\right)$ acrescentando o tempo $t = 0$ para uma amostra de tamanho T.

□

11.1 Verossimilhança Perfilada

Em um determinado modelo estatístico podemos estar interessados somente em parte do vetor de parâmetros e não no vetor completo ϑ. Especificamente, se o vetor de parâmetros completo ϑ pode ser decomposto como $\vartheta = (\varphi, \varsigma)$ e nos interessa estimar e inferir acerca de valores de φ, chamaremos φ de vetor de parâmetros de interesse, e ao vetor ς de parâmetros de perturbação. Em situações como esta é possível, por diferentes metodologias, construir uma função que dependa somente de φ e que possamos utilizar para realizar inferências acerca de φ. Estas funções são conhecidas como funções de pseudo-verossimilhança.

Diversas destas funções têm sido consideradas na literatura e muitos esforços dedicados a uma delas, a função de verossimilhança perfilada, [13].

Definição 11.1. *Define-se o logaritmo da função de verossimilhança perfilada para φ como sendo*

$$l_p(\varphi) = \max_{\varsigma} l(\varphi, \varsigma) \tag{11.7}$$

sendo que o máximo é obtido em todo o espaço paramétrico do modelo avaliado, fixando um valor de φ.

Observamos que o processo de maximização ao qual se faz referência na definição anterior é realizado quando obtemos $\hat{\varsigma}(\varphi)$. Desta forma a função de verossimilhança perfilada pode ser definida como

$$l_p(\varphi) = l(\varphi, \hat{\varsigma}(\varphi))$$

Temos que os máximos das funções $l_p(\varphi)$ e $l(\vartheta)$ (verossimilhança aplicada ao vetor de parâmetros completo) coincidem, ou seja, suponhamos que $\hat{\varphi}$ maximiza $l_p(\varphi)$. Temos então

$$l_p(\hat{\varphi}) \geq l_p(\varphi) \geq l(\varphi, \varsigma)$$

e dado que $\hat{\vartheta} = (\hat{\varphi}, \hat{\varsigma})$ é tal que

$$l\left(\hat{\varphi}, \hat{\varsigma}\right) = \max_{\varphi, \varsigma} l\left(\varphi, \varsigma\right)$$

então

$$l_p(\hat{\varphi}) \geq l\left(\hat{\varphi}, \hat{\varsigma}\right)$$

Por outro lado, como $\hat{\vartheta}$ é o máximo absoluto de $l(\vartheta)$ no espaço paramétrico do modelo

$$l\left(\hat{\varphi}, \hat{\varsigma}\right) \geq l_p(\hat{\varphi})$$

já que $\hat{\varphi}$ é o máximo em um subespaço do espaço paramétrico do modelo. Desta forma, obtemos que os pontos $l_p(\hat{\varphi})$ e $l\left(\hat{\varphi}, \hat{\varsigma}\right)$ coincidem.

Bibliografia

[1] Ali,S. M. e Silvey,S. D. (1966). *A general class of coefficients of divergence of one distribution from another.* J. Royal Statist. Soc. B 28 131-142.

[2] Baum, Leonard. E.; Petrie, Ted. (1966). *Statistical Inference for Probabilistic Functions of Finite State Markov Chains.* The Annals of Mathematical Statistics. vol. 37 (6), pp. 15541563.

[3] Brooke, M.; Hanley, S.; Laughlin, S. (1999) *The scaling of eye size with body mass in birds.* Proceedings of the Royal Society of London Series B-Biological Sciences, v. 266, n. 1417, pp. 405412.

[4] Bühlmann, P., A. J. Wyner, A. J. (1999) *Variable length Markov chains*, Ann. Statist., vol. 27, pp. 480513.

[5] Cappé,Olivier., Moulines, Eric., Rydén,Tobias. (2009). *Inference in Hidden Markov Models.*

[6] Amorim, Claudiana de Souza., Baron, Jerome. (2016). *Estudo da seletividade neuronal à orientação e freqüência espacial no wulst visual da coruja suindara (Tyto alba): dinâmica de surgimento e separabilidade interdimensional.* Tese de Doutarado, Ciências Biológicas - Fisiologia e Farmacologia do Instituto de Ciências Biológicas, da Universidade Federal de Minas Gerais.

[7] Collet, Pierre., Galves, Antonio., Leonardi, Florencia. (2008) *Random perturbations of stochastic processes with unbounded variable length memory.* Eletronic Journal of Probability., vol. 13, pp. 1345-1361.

[8] Csiszár, Imre., P. Shields. (2000) *The consistency of the BIC Markov order estimator.* Ann. Statist., vol. 28, pp. 16011619

[9] Csiszár, Imre., Talata, Zsolt. (2006) *Context tree estimation for not necessarily nite memory processes, via BIC and MDL.* IEEE Trans. Inform. Theory, 52(3).

[10] Dempster, A. P., Laird, N.M. e Rubin, D.B. (1977). *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society, B, 39, 1-22.

[11] Dumont, Thierry. (2014) *Context Tree Estimation in Variable Length Hidden Markov Models.* IEEE Trans. Inform. Theory, Vol. 60, NO. 6.

[12] Garcia, Nancy. L. and Moreira, Lucas. (2014). *Stochastically Perturbed Chains of Variable Memory.* arXiv:1305.5747v1 [math.PR].

[13] McCullagh, P. e Tibshirani, R. (1990). *A simple method for the adjustment of profile likelihoods.* Journal of the Royal Statistical Society, 52(2), 325344.

[14] McLachlan, Geoffrey and Krishnan, Thriyambakam. *The EM Algorithm and Extensions.* John Wiley Sons, New York, 1996

[15] Greene, William H. *Econometric Analysis.* 5th ed. Upper Saddle River, NJ: Prentice Hall.

[16] Leroux, Brian G. *Maximum-likelihood estimation for hidden Markov models.* Stochastic Processes and their Applications 40 (1992) 127-143

[17] Rabiner, R. Lawrence. (1989) *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*. Proceedings of the IEEE., vol. 77., N^o 2.

[18] Rissanen, Jorma.(1983) *A universal data compression system*. IEEE Trans. Inform.Theory, 29(5).

[19] Ron, Dana., Singer, Yoram., Tishby, Naftali. (1996) *The Power of Amnesia: Learning Probabilistic Automata with Variable Memory Length*. Mach. Learn., 25, 117149.

[20] Yi., Wang, Lizhu., Zhou, Jianyoung., Wang, Jianhua., Feng and Zhi-qiang., Liu. *Mining complex time-series by learning Markovian models*. 6th ICDM, 2005, pp. 11361140.

[21] Yi Wang. *The variable-length hidden Markov model and its applications on sequential data mining*. Dept. Comput. Sci., Rensselaer Polytech. Inst., Troy, NY, USA, Tech. Rep., 2005.