Pedro Grünauer Kassab

*Filtragem e Identificação em Sistemas Lineares
Sujeitos a Saltos Markovianos com Modo de
Operação Não Observado*

São Paulo

2010

Pedro Grünauer Kassab

*Filtragem e Identificação em Sistemas Lineares
Sujeitos a Saltos Markovianos com Modo de
Operação Não Observado*

Dissertação apresentada à Escola Politécnica
da Universidade de São Paulo para obtenção
do título de Mestre em Engenharia Elétrica

Área de Concentração:
Engenharia de Sistemas

Orientador:
Prof. Dr. Oswaldo Luiz do Valle Costa

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO

São Paulo

2010

Resumo

Este trabalho propõe uma metodologia de identificação para sistemas lineares sujeitos a saltos markovianos. Dada uma sequência de observações ruidosas da variável de estados, busca-se estimá-la juntamente com os parâmetros (desconhecidos) que descrevem o sistema dinâmico no espaço de estados. Como é bem conhecido, a filtragem ótima nesta classe de sistemas tem requisitos computacionais exponencialmente crescentes em função do tamanho da amostra, e torna-se inviável na prática. Recorre-se, portanto, a um algoritmo sub-ótimo de filtragem, cujos resultados são utilizados na identificação por máxima verossimilhança segundo a metodologia apresentada. Simulações realizadas mostram boa convergência.

Palavras-chave: Filtragem estocástica, identificação de sistemas, sistemas lineares variantes no tempo, cadeias de Markov.

Abstract

This paper proposes a methodology for the identification of Markov-jump linear systems. Given a sequence of noisy observations of the state variable, our objective is to estimate it along with the (unknown) parameters that drive the system in the state-space. As it is well known, the optimal filtering in this class of systems requires exponentially increasing computing power, in proportion to the sample size, and is not feasible in practice. We resort, therefore, to a sub-optimal algorithm, whose results are used for a maximum likelihood identification according to the methodology presented here. Simulations show a good convergence.

Keywords: Stochastic filtering, systems identification, time-varying linear systems, Markov chains.

Lista de Figuras

1	Representação das classes de modelos descritas.	p. 10
2	Representação de sistema sujeito a saltos, com transições que seguem uma cadeia de Markov.	p. 12
3	Diagrama de um modelo de Markov oculto.	p. 15
4	Relação entre os modelos apresentados. Em cinza, o foco principal deste texto.	p. 16
5	Diagrama de blocos do algoritmo GPB2	p. 31
6	Diagrama de blocos do algoritmo IMM	p. 32
7	Comparação entre o desempenho dos algoritmos GPB, IMM e Linear. .	p. 33
8	Histograma de erros de estimação para o algoritmo GPB	p. 34
9	Histograma de erros de estimação para o algoritmo IMM	p. 34
10	Histograma de erros de estimação para o algoritmo Linear	p. 35
11	Condições iniciais e valores de convergência para os parâmetros $\mathbf{A}_1$ e $\mathbf{A}_2$	p. 39
12	Condições iniciais e valores de convergência para os parâmetros $\mathbf{F}_1$ e $\mathbf{F}_2$	p. 40
13	Condições iniciais e valores de convergência para os parâmetros p_{11} e p_{22}	p. 41
14	Convergência dos parâmetros $\mathbf{A}_1$ e $\mathbf{A}_2$, Caso I	p. 45
15	Convergência dos parâmetros $\mathbf{A}_1$ e $\mathbf{A}_2$, Caso II	p. 45
16	Convergência dos parâmetros $\mathbf{F}_1$ e $\mathbf{F}_2$, Caso I	p. 46
17	Convergência dos parâmetros $\mathbf{F}_1$ e $\mathbf{F}_2$, Caso II	p. 46
18	Convergência dos parâmetros p_{11} e p_{22} , Caso I	p. 47
19	Convergência dos parâmetros p_{11} e p_{22} , Caso II	p. 47
20	Convergência do parâmetro $\mathbf{A}_1$	p. 49

21	Convergência do parâmetro $\mathbf{A}_2$	p. 49
22	Convergência do parâmetro $\mathbf{F}_1$	p. 50
23	Convergência do parâmetro $\mathbf{F}_2$	p. 50
24	Convergência do parâmetro p_{11}	p. 51
25	Convergência do parâmetro p_{22}	p. 51
26	Gráfico quantil-quantil para o parâmetro $\mathbf{A}_1$	p. 52
27	Gráfico quantil-quantil para o $\mathbf{A}_2$	p. 52
28	Gráfico quantil-quantil para o $\mathbf{F}_1$	p. 53
29	Gráfico quantil-quantil para o $\mathbf{F}_2$	p. 53
30	Gráfico quantil-quantil para o p_{11}	p. 54
31	Gráfico quantil-quantil para o p_{22}	p. 54

Sumário

1	Introdução	p. 9
1.1	Modelos sujeitos a saltos	p. 10
1.2	Sistemas lineares sujeitos a saltos markovianos	p. 13
1.3	Modelos de Markov ocultos	p. 14
1.4	Especificação do objeto de estudo	p. 15
1.5	Principais soluções a obter para a classe de SLSM	p. 16
1.6	Objetivos	p. 17
1.7	Estrutura do texto	p. 17
1.8	Notação	p. 18
2	Revisão bibliográfica	p. 19
3	Algoritmos de filtragem	p. 22
3.1	Definições	p. 22
3.2	Especificação estocástica dos SLSM	p. 23
3.3	Algoritmos de Filtragem	p. 24
3.3.1	Filtragem ótima	p. 24
3.3.2	Algoritmo GPB2	p. 27
3.4	Algoritmo IMM	p. 30
3.5	Algoritmo linear	p. 32
3.6	Comparação entre os algoritmos de filtragem	p. 32
4	Metodologia para estimação de parâmetros	p. 36

4.1	Algoritmo	p. 36
4.2	Resultados experimentais	p. 37
4.3	Comentários	p. 42
5	Simulações e resultados	p. 43
5.1	Desempenho da estimação e convergência	p. 43
5.1.1	Ensaio com amostra longitudinal	p. 43
5.1.2	Ensaio em seção transversal	p. 44
6	Conclusões	p. 55
6.1	Agradecimentos	p. 56
	Apêndice A – Fundamentos teóricos	p. 57
A.1	Definições	p. 57
A.1.1	Espaços de probabilidade	p. 57
A.1.2	Processos estocásticos e cadeias de Markov	p. 58
A.1.3	Valor esperado condicional dado um campo- σ	p. 59
A.1.4	Mudanças de medida e a derivada de Radon-Nikodym	p. 60
A.1.5	Kernels de transição e produtos projetivos	p. 61
A.1.6	Ergodicidade	p. 63
A.1.7	Filtragens	p. 63
A.2	Estimadores de máxima verossimilhança	p. 64
A.2.1	Função verossimilhança	p. 65
A.2.2	Método de máxima verossimilhança e suas propriedades	p. 66
A.2.3	Condições de consistência, normalidade assintótica e eficiência	p. 68
A.2.3.1	Consistência	p. 68
A.2.3.2	Normalidade assintótica	p. 69
A.2.3.3	Fronteira inferior de Cramér-Rao	p. 70

A.2.3.4	Eficiência	p. 70
A.3	Filtro bayesiano não-linear	p. 71
A.3.1	Metodologia geral	p. 71
A.3.2	Forma explícita do filtro bayesiano não-linear	p. 73
Referências		p. 75

1 *Introdução*

A elaboração de modelos matemáticos de sistemas reais tem como princípio aproximar o comportamento das variáveis julgadas relevantes dentro de um sistema. Esta descrição permite, sobretudo, que se explique o fenômeno observado à luz do modelo e que se extrapole o seu funcionamento a situações não verificadas empiricamente. Para cada sistema, no entanto, é possível que se construa uma infinidade de modelos, descrevendo-o com variados graus de fidelidade, de complexidade matemática e de adequação a situações específicas. Com efeito, verifica-se que não há, na maioria das situações práticas, um critério objetivo que determine a escolha do melhor entre os possíveis modelos, de forma que a experiência e o senso prático têm papel determinante.

É possível, inclusive, que se conclua que um determinado sistema a ser modelado matematicamente possa ser representado de forma mais simples ou mais fiel à realidade se descrito por uma sucessão temporal de diferentes modelos, ao invés de um único modelo que tentasse explicar seu comportamento em todas as situações. Esta técnica é utilizada em muitos casos na literatura, alcançando grande aplicabilidade em diversas áreas.

Existem, no entanto, tratamentos bastante diversos do tema, que merecem ser considerados e explicados. Este estudo se propõe a introduzir o paradigma de alternância de modelos, cujas aplicações são comumente conhecidas como *sistemas sujeitos a saltos*. Esta denominação explicita o fato de que o comportamento do sistema exhibe descontinuidades marcantes, que são descritas como uma transição entre modelos matemáticos. É evidente que esta descrição será bastante apropriada a sistemas em que é patente a existência deste tipo de descontinuidades e a sistemas em que se souber a priori da existência da possibilidade destas.

1.1 Modelos sujeitos a saltos

Optando por um modelo em tempo discreto, considere-se, para efeito de ilustração do conceito, o caso mais geral possível. Suponha-se que as variáveis de estado (não observadas) de um determinado sistema estejam reunidas em um vetor $\mathbf{x}$, indexado temporalmente da forma $\mathbf{x}(t)$, para cada instante t . A evolução temporal do vetor $\mathbf{x}$ é regida por uma família de funções f (denominadas $f_1, f_2, f_3, \dots$): o valor da variável $\mathbf{x}$ será obtido, em cada instante, pela aplicação de alguma função f_n à variável de estado do instante precedente. De forma análoga, a variável de saída (a única observável) chamada de $\mathbf{y}$, dependerá de $\mathbf{x}$ segundo alguma função g_n pertencente à família de funções $g_1, g_2, g_3, \dots$ que descreve a relação estados-saídas do modelo. Este encontra-se representado na figura 1.

Figura 1: Representação das classes de modelos descritas.

Note-se que a especificação anterior permite que, a cada instante, as funções f e g possam ser diferentes de todas as demais. É bastante evidente que um modelo com esta forma apresenta uma flexibilidade excessiva para as situações comumente encontradas. Dado que modelos são construídos e validados com base em observações empíricas, permitir um grau tão extenso de liberdade em sua evolução temporal torná-lo-ia algo simultaneamente pouco significativo e de pouca utilidade. A falta de significância deriva do fato de que a observação única de cada função não permitiria qualquer inferência sobre sua forma, enquanto que a baixa utilidade decorre do fato de que a observação do comportamento passado não resultaria em ganho de informação sobre os instantes futuros.

É necessário, portanto, que se acrescentem restrições que resolvam estas dificuldades, a saber: é necessário restringir a quantidade de possíveis funções de transição a um número finito, assim como é necessário especificar alguma forma de recorrência (possivelmente aleatória) para sua sucessão. Ressalte-se que é importante, do ponto de vista prático, que as funções de transição sejam relativamente poucas e suficientemente recorrentes

com respeito ao número de observações, de forma que haja um número significativo de realizações observadas em cada f_n .

A solução usualmente encontrada na literatura, e que se encontra bem difundida, é limitar o número de possíveis funções f e g a um número M natural finito. Este número se supõe conhecido *a priori* em algumas aplicações, e é estimado em outras. Além desta restrição, impõe-se em muitos casos que a sucessão temporal entre funções f (e, conjuntamente, g) ocorra segundo uma cadeia de Markov. Isto significa dizer que os elementos das famílias $f_1, f_2, \dots, f_M$ e $g_1, g_2, \dots, g_M$, de funções que recebem os índices $1, 2, \dots, M$, sucedem-se de forma aleatória seguindo um processo com a propriedade de Markov (dependência apenas do estado presente).

Esta progressão se realizaria da seguinte forma: crie-se um processo estocástico (em geral, não observável) $\{\theta(t)\}$, em tempo discreto, que pode assumir os valores $\theta(t) = 1, 2, \dots, M$. Diz-se que $\theta(t)$ é uma cadeia de Markov se a distribuição de probabilidade de $\theta(k+1)$ depender apenas de $\theta(k)$, sendo condicionalmente independente de todos os valores do processo $\theta(t)$, para $t < k$. Além disto, esta cadeia será chamada *homogênea* se as probabilidades de transição entre quaisquer dois estados forem constantes para qualquer instante t .

Sendo $\{\theta(t)\}$ uma cadeia de Markov homogênea, utiliza-se então o valor assumido por esta cadeia a cada instante t para denotar o índice da função f e da função g vigente naquele instante. Se, por exemplo, $\theta(t) = i$ ter-se-á, no instante t , as funções $f_i(\mathbf{x})$ e $g_i(\mathbf{x})$ ¹ determinando a evolução da variável de estado $\mathbf{x}$. Para explicitar esta dependência com relação a uma cadeia de Markov que seleciona um elemento dentro de uma família de funções, utiliza-se comumente a notação $f_{\theta(t)}(\mathbf{x})$. A figura 2 exhibe uma representação esquemática destas relações.

Postular um desenvolvimento segundo uma cadeia de Markov homogênea não é, contudo, uma escolha sem consequências. Com efeito, ao se supor esta característica, afirma-se que a sucessão de modelos ocorre por um processo *exógeno*. Isto significa que este processo é externo ao modelo no espaço de estados, que não o explica. É fácil observar que as probabilidades de transição $\mathbf{P}$ independem de todos os $\mathbf{x}(n)$ e $\mathbf{y}(n)$. Desta forma, é importante que se verifique a consistência da premissa de exogeneidade com o que se espera do modelo.

Nos numerosos exemplos encontrados na literatura, apresentados no segundo capítulo,

¹Omite-se, onde não parece haver prejuízo à clareza, a dependência temporal da variável de estado $\mathbf{x}$. O intuito é simplificar a notação.

Figura 2: Representação de sistema sujeito a saltos, com transições que seguem uma cadeia de Markov.

concluiu-se que a premissa de exogeneidade era uma simplificação aceitável. Em algumas aplicações (como reconhecimento de escrita, de voz e sequenciamento do DNA), a exogeneidade é quase imediata. Em outras (controle de falhas, modelagem de volatilidade financeira), esta decorre de um compromisso entre a generalidade e a tratabilidade do modelo.

A característica de cadeia de Markov que se impôs em $\{\theta(t)\}$, por sua vez, é pouco restritiva. Embora possa parecer uma hipótese forte, é possível contorná-la transformando um modelo de Markov de ordem superior (em que haja dependência de um número finito de instantes passados, além do atual) em uma cadeia de Markov simples equivalente. Logo, a despeito de poder trazer complicações para ordens elevadas, que resultariam em matrizes de grande dimensão no modelo equivalente, esta é uma premissa de menos consequência do que as de homogeneidade e exogeneidade.

Uma segunda etapa de simplificações pode agora ser empreendida, com a finalidade de tornar aplicáveis a este caso resultados de áreas do conhecimento bem desenvolvidas, dotadas de literatura profusa e ferramental extenso. Esta é a justificativa para a adoção de uma premissa de *linearidade* a ser aplicada às funções f e g . Esta premissa, embora muitas vezes patentemente inverossímil, é utilizada com sucesso em um grande número de instâncias como uma aproximação da realidade. Este postulado permite, por outro lado, que se traduzam a este contexto importantes resultados que são essenciais à tratabilidade matemática do objeto de estudo. Da mesma forma, postula-se uma *invariância no tempo* para cada função f e g , individualmente.²

Admitindo-se a linearidade, é possível então elaborar uma representação sob a forma

²A

de um *sistema linear no espaço de estados*. Nesta forma, exprimem-se duas grandezas como funções lineares da variável de estado $\mathbf{x}$ no instante atual: a saída observada $\mathbf{y}$ no instante atual e a variável de estado $\mathbf{x}$ no instante seguinte. Estas funções, em geral, não são determinísticas.

1.2 Sistemas lineares sujeitos a saltos markovianos

Com base nesta representação como um sistema linear no espaço de estados, constroi-se a classe de sistemas denominados *sistemas lineares sujeitos a saltos markovianos* (chamados, em inglês, Markov-jump linear systems). Esta premissa de linearidade se aplica, evidentemente, a cada função f e g representada da figura 2. Desta forma, ter-se-ia uma representação no espaço de estados associada a *cada* par $(f_1, g_1), (f_2, g_2), \dots, (f_M, g_M)$.

Considere-se que uma representação no espaço de estados se faça da seguinte forma, sem prejuízo à posterior adição de outros termos às expressões:

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{F}\mathbf{v}(t) \quad (1.1a)$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{G}\mathbf{w}(t) \quad (1.1b)$$

em que $\mathbf{A}$, $\mathbf{C}$, $\mathbf{F}$ e $\mathbf{G}$ são transformações lineares de dimensões apropriadas. Os vetores $\mathbf{v}(t)$ e $\mathbf{w}(t)$ são variáveis aleatórias. A primeira linha de (1.1) chama-se *equação de estados*, e descreve a evolução temporal da variável de estados $\mathbf{x}(t)$. A segunda denomina-se *equação de saídas*, e rege a relação entre a variável de estados $\mathbf{x}(t)$ e a variável de saída $\mathbf{y}(t)$.

É evidente que, para o caso representado na figura 2, a função f comporá a forma funcional da equação de estados, e a função g terá seu lugar correspondente na equação de saídas. Portanto, para cada par (f_i, g_i) , constroi-se um sistema linear

$$\mathbf{x}(t+1) = f_i[\mathbf{x}(t)] = \mathbf{A}_i\mathbf{x}(t) + \mathbf{F}_i\mathbf{v}(t) \quad (1.2a)$$

$$\mathbf{y}(t) = g_i[\mathbf{x}(t)] = \mathbf{C}_i\mathbf{x}(t) + \mathbf{G}_i\mathbf{w}(t) \quad (1.2b)$$

correspondente à descrição linear no espaço de estados apresentada na expressão (1.1).

A representação (1.2) explicita o fato de que é possível ainda introduzir mais uma simplificação estrutural. Como se pode observar, a referência explícita às funções f_i e g_i é supérflua: já que todas as funções de cada família têm a mesma forma linear, bastaria associar a cada estado da cadeia de Markov um parâmetro composto por quatro matrizes

$(\mathbf{A}_i, \mathbf{C}_i, \mathbf{F}_i, \mathbf{G}_i)$. Substitui-se, desta forma, a dependência de $\mathbf{x}$ de uma função arbitrária, f_i , pela dependência (muito mais tratável) de um conjunto de parâmetros dentro de uma forma funcional fixa. Com isto, pode-se substituir a forma da equação (1.2) por

$$\mathbf{x}(t+1) = \mathbf{A}_{\theta(t)}\mathbf{x}(t) + \mathbf{F}_{\theta(t)}\mathbf{v}(t) \quad (1.3a)$$

$$\mathbf{y}(t) = \mathbf{C}_{\theta(t)}\mathbf{x}(t) + \mathbf{G}_{\theta(t)}\mathbf{w}(t) \quad (1.3b)$$

que ilustra a dependência dos parâmetros $(\mathbf{A}, \mathbf{C}, \mathbf{F}, \mathbf{G})$ da cadeia de Markov $\{\theta(t)\}$. Conforme o valor do processo estocástico $\{\theta(t)\}$ no instante t , escolhe-se um conjunto $(\mathbf{A}_i, \mathbf{C}_i, \mathbf{F}_i, \mathbf{G}_i)$ dos possíveis $(\mathbf{A}_1, \mathbf{C}_1, \mathbf{F}_1, \mathbf{G}_1), (\mathbf{A}_2, \mathbf{C}_2, \mathbf{F}_2, \mathbf{G}_2), \dots, (\mathbf{A}_M, \mathbf{C}_M, \mathbf{F}_M, \mathbf{G}_M)$. Denomine-se, então, esta coleção

$$\eta \triangleq \{(\mathbf{A}_1, \mathbf{C}_1, \mathbf{F}_1, \mathbf{G}_1), (\mathbf{A}_2, \mathbf{C}_2, \mathbf{F}_2, \mathbf{G}_2), \dots, (\mathbf{A}_M, \mathbf{C}_M, \mathbf{F}_M, \mathbf{G}_M)\} \quad (1.4)$$

que não se assume, em geral, que seja conhecida *a priori*.

A forma apresentada em (1.3) é, portanto, a forma mais simples da classe de sistemas lineares sujeitos a saltos markovianos (SLSM). Esta classe de sistemas possui grande generalidade, e já existem na literatura resultados importantes pertinentes a sua aplicação.

1.3 Modelos de Markov ocultos

Os modelos de Markov ocultos (MMO), conhecidos na literatura anglófona como *Hidden Markov Models* (ou HMM), são casos particulares dos sistemas lineares sujeitos a saltos markovianos. A principal simplificação incorporada a estes modelos é a eliminação da variável de estados $\mathbf{x}(t)$, que precisa ser propagada no caso geral dos sistemas lineares. Esta simplificação, apesar de implicar uma perda de generalidade não desprezível, é responsável por uma grande simplificação da análise, levando a modelos cuja solução de filtragem ótima é bem conhecida e não tem requisitos de memória exponencialmente crescentes, o que se observa no modelo geral (apresentado nas seções precedentes).

Este modelo, que é representado na figura 3, é a versão estocástica de uma máquina de estados. Em cada possível valor da cadeia de Markov $\theta(t)$ em um determinado instante t , a variável de saída $\mathbf{y}(t)$ será uma variável aleatória, cuja função densidade de probabilidade é $g_{\theta(t)}$. Esta função é um elemento da família de funções $g_1, \dots, g_M$, correspondentes a cada possível estado $1, \dots, M$ da cadeia de Markov $\{\theta(t)\}$. A variável aleatória $\mathbf{y}(t)$ é independente, portanto, de todos os valores precedentes de $\mathbf{y}$ e de θ .

O diagrama da figura 4 explicita a relação entre os modelos mencionados. Note-se que, entre os modelos apresentados, apenas nos MMO a variável $\mathbf{y}$ é um processo de Markov (ou seja, seu valor atual independe condicionalmente de todos os valores da variável, com exceção do imediatamente precedente ao instante considerado). Os demais, apesar de funções de uma cadeia de Markov, perdem esta característica devido à propagação temporal de uma variável de estados $\mathbf{x}(t)$.

Esta classe de modelos apresenta uma literatura bastante copiosa, e encontra frequentemente aplicação prática.

Figura 3: Diagrama de um modelo de Markov oculto.

1.4 Especificação do objeto de estudo

Apresentadas as principais formulações que utilizam a alternância de modelos segundo uma cadeia de Markov não observada, pode-se agora delimitar com maior precisão o objeto de estudo. A figura 4 classifica, em ordem de generalidade, os modelos apresentados anteriormente (contidos nas figuras 1, 2 e 3). A formulação escolhida como foco, a classe dos *modelos lineares sujeitos a saltos markovianos*, é mais geral do que a classe dos modelos de Markov ocultos, e mais restrita do que a classe dos modelos sujeitos a saltos markovianos.

Reconhecendo-se os desafios apresentados pela classe eleita para análise, como a complexidade do algoritmo de filtragem ótima, justifica-se a escolha do tema - em detrimento dos modelos de Markov ocultos - exatamente pelo fato de a classe dos sistemas lineares sujeitos a saltos markovianos encontrar atualmente aplicação prática menos difundida e, conseqüentemente, maior potencial de expansão.

E, embora tenha sido objeto de estudo de um bom número de publicações recentes de grande qualidade, não representa uma técnica teoricamente e empiricamente tão consolidada quanto os modelos de Markov ocultos. Especialmente no que tange a identificação

de modelos pertencentes à classe de sistemas lineares sujeitos a saltos markovianos, muito trabalho resta a ser feito até que se atinja uma solução satisfatória.

Figura 4: Relação entre os modelos apresentados. Em cinza, o foco principal deste texto.

1.5 Principais soluções a obter para a classe de SLSM

Ao se escolher a representação de sistema linear sujeito a saltos markovianos para um determinado sistema, é imprescindível que se encontrem as soluções de um número de problemas, sem o que a formulação de um modelo sob esta forma teria pouca utilidade. É necessário encontrar:

- um estimador $\hat{\theta}(t)$ para $\theta(t)$, o estado da cadeia de Markov no instante t ;
- um estimador $\hat{\mathbf{x}}(t)$ para $\mathbf{x}(t)$, o valor da variável de estado $\mathbf{x}$ no instante t ;
- um estimador $\hat{\eta}$ para $\eta = \{(\mathbf{A}_1, \mathbf{C}_1, \mathbf{F}_1, \mathbf{G}_1), (\mathbf{A}_2, \mathbf{C}_2, \mathbf{F}_2, \mathbf{G}_2), \dots, (\mathbf{A}_M, \mathbf{C}_M, \mathbf{F}_M, \mathbf{G}_M)\}$, o conjunto de parâmetros do sistema no espaço de estados, associados a cada possível estado da cadeia de Markov.

Os problemas de obtenção dos estimadores das variáveis $\hat{\theta}(t)$ e $\hat{\mathbf{x}}(t)$ são conhecidos como problemas de *filtragem*. Isto decorre do fato de que a variável observada é apenas $\mathbf{y}(t)$, de onde se tenta estimar as variáveis não observadas $\theta(t)$ e $\mathbf{x}(t)$, de forma análoga à filtragem de um sinal ruidoso recebido por uma antena. Para referir-se ao cálculo de $\hat{\eta}$, costuma-se empregar o termo *estimação de parâmetros*. Estes estimadores serão apresentados com mais detalhes neste texto, e suas propriedades serão discutidas.

Encerra-se, assim, a conceituação inicial do problema. Como se pode verificar, o escopo potencial de um trabalho sobre o tema é bastante abrangente. Com efeito, ao

se alterarem algumas das premissas sobre as quais se constroem os modelos, criam-se diferentes classes de sistemas cujo tratamento individual poderia ocupar toda a extensão de um estudo. Faz-se necessário, portanto, delimitar tanto o objeto de estudo quanto a profundidade pretendida da elaboração deste.

1.6 Objetivos

O objetivo deste texto é apresentar técnicas de identificação para sistemas lineares sujeitos a saltos markovianos com modo de operação não observável e parâmetros desconhecidos. Espera-se sintetizar um tratado de viés primariamente prático, que apresente uma introdução a estes modelos, resuma os principais desenvolvimentos apresentados na literatura e proponha uma metodologia de identificação.

O resultado concreto que se espera atingir é a apresentação de algoritmos que levem a soluções aproximadas dos problemas de filtragem e estimação de parâmetros nos sistemas lineares sujeitos a saltos markovianos com modo de operação não observado e parâmetros desconhecidos.

Por tratar-se de um problema bastante complexo, para o qual ainda não há soluções satisfatórias na literatura, o que se pretende apresentar neste estudo é uma metodologia prática de filtragem e estimação de parâmetros desconhecidos do sistema por meio do método de máxima verossimilhança e algoritmos sub-ótimos. Devido à complexidade teórica inerente, uma prova de convergência está fora do escopo deste trabalho, mas a convergência dos parâmetros será evidenciada em simulações numéricas ilustrativas.

1.7 Estrutura do texto

O primeiro capítulo dedicou-se a introduzir e motivar o estudo dos sistemas lineares sujeitos a saltos markovianos. O segundo capítulo contém a revisão bibliográfica das principais fontes para a elaboração deste estudo. O terceiro capítulo introduz o algoritmo ótimo e as principais aproximações dele derivadas. No quarto capítulo, apresentam-se resultados experimentais e sua análise. As conclusões encontram-se no quinto capítulo.

O apêndice contém aspectos teóricos sobre processos estocásticos, filtros e estimadores de máxima verossimilhança.

1.8 Notação

Utilizam-se neste texto, salvo menção explícita em contrário, letras latinas maiúsculas (A, B) para denotar conjuntos. Com letras maiúsculas em negrito (**C**, **D**), representam-se transformações lineares $\mathbb{R}^m \rightarrow \mathbb{R}^n$, onde $\mathbb{R}^m$ e $\mathbb{R}^n$ são espaços vetoriais sobre o corpo de escalares $\mathbb{R}$. As letras minúsculas em negrito (**x**, **y**) fazem referência aos vetores, elementos de $\mathbb{R}^p$. Os termos em fonte Fraktur ($\mathfrak{F}$, $\mathfrak{S}$) fazem referência a σ -campos.

2 *Revisão bibliográfica*

Esta classe de modelos tem sido objeto de um grande número de estudos, entre os quais é possível destacar (1), (2), (3) e (4) como textos de referência. Estes modelos foram aplicados com sucesso a diversos casos de interesse prático, especialmente em aplicações como o rastreamento de alvos em manobra ((1), (5)), controle de sistemas sujeitos a falhas ((6), (7)), modelagem de variáveis econômicas sujeitas a alterações conjunturais ((8), (9)), além de outros sistemas de natureza análoga.

O problema de identificação nesta classe de sistemas é evidentemente um aspecto fundamental para a aplicabilidade prática destes modelos. Com efeito, esta questão foi analisada em algumas publicações como (10), (11), (12) e (13), mas constitui ainda uma questão em aberto. Isto se deve ao fato de o problema de identificação ser indissociável do problema de filtragem, que está sujeito a complicações computacionais importantes, como se verá na seção 3.3.

Em poucas palavras, basta constatar que a estimação de parâmetros por máxima verossimilhança (o método mais comumente utilizado) depende do cálculo das estimativas ótimas das variáveis não observadas. Para realizar uma estimação paramétrica, compara-se a distribuição do erro de estimação com a função densidade presumida dos processos estocásticos associados ao sistema. No entanto, é bem conhecido que a classe de sistemas lineares sujeitos a saltos markovianos sofre de um problema de dimensionalidade no que diz respeito a sua filtragem ótima. Apesar de ter muitos aspectos em comum com sistemas lineares invariantes no tempo e com modelos de Markov ocultos (Hidden Markov Models), os SLSM não possuem filtros ótimos de dimensão fixa, mesmo no caso linear-gaussiano. Os requisitos de processamento e memória para a filtragem ótima dos SLSM crescem exponencialmente com o tempo.

Isto levou ao desenvolvimento de um grande número de estimadores sub-ótimos, com distintas razões entre custo computacional e desvio em relação ao ótimo. Os algoritmos mais amplamente utilizados são o IMM (14), o GPB (1), e o algoritmo linear (15). Tam-

bém pode-se mencionar o uso de filtros polinomiais para o caso não-gaussiano, em (16). Mais recentemente, utilizam-se também algoritmos baseados em simulações, como os filtros de partículas ((17) e (6)). Todos estes algoritmos apresentam desvios em relação ao algoritmo ótimo, e a variância do erro de estimação não atinge assintoticamente a fronteira de Cramér-Rao. Para alguns destes algoritmos (IMM e GPB), não há sequer uma demonstração formal na literatura de que se trata de estimadores não-viesados, sendo os mesmos apenas truncamentos do algoritmo ótimo.

É evidente, portanto, que se trata de um problema bastante complexo, para o qual ainda não há soluções satisfatórias na literatura que sejam de conhecimento dos autores. O que se pretende apresentar neste estudo é uma metodologia prática de estimação dos parâmetros desconhecidos do sistema por meio do método de máxima verossimilhança, com base em uma filtragem realizada por estimadores sub-ótimos. Devido a esta complexidade teórica inerente, uma prova de convergência está fora do escopo deste artigo, mas a convergência dos parâmetros será evidenciada em simulações numéricas ilustrativas.

Em (10), utiliza-se o algoritmo ótimo de filtragem em uma janela da sequência observada como aproximação para o filtro ótimo. Como este método consome recursos computacionais exponencialmente crescentes com o tamanho da amostra, só é possível realizar a estimação com base em um número reduzido de observações - quatro, no estudo original. Busca-se, em (10), estimar os parâmetros com base nesta filtragem sub-ótima no caso em que os parâmetros pertencem a um conjunto finito e discreto de possibilidades, conhecidas ex ante. Em (18), (19) e (20), apresentam-se soluções para o problema de estimação das probabilidades de transição entre modos, dado que sejam conhecidos os demais parâmetros dinâmicos do sistema, sem a restrição de que pertençam a um número finito de possibilidades conhecidas.

Neste trabalho, pretende-se aplicar o estimador de máxima verossimilhança a diversos parâmetros do sistema no espaço de estados e às probabilidades de transição, simultaneamente, sem estabelecer um conjunto universo finito para as possibilidades. O método proposto aqui pretende utilizar um filtro sub-ótimo aplicado à totalidade da série observada (sem truncamentos) ao invés de utilizar uma estimação ótima em um pequeno subconjunto da amostra.

As conclusões demonstrarão que, dado que a aproximação do filtro sub-ótimo seja razoavelmente próxima, é possível obter boas estimativas dos parâmetros, de forma que se permitiria um desempenho adequado de um sistema de rastreamento e controle associado ao filtro adaptativo. Além disto, a utilização do algoritmo sub-ótimo permite

que se considere a amostra completa na estimação, o que se traduz em uma convergência consistente nos casos experimentais analisados. Isto abre espaço para a aplicação desta classe de modelos em muitos casos de interesse prático em que não se conheçam a priori os parâmetros que regem o sistema em questão, ou em casos em que estes possam variar no tempo e exijam uma estimação adaptativa.

3 Algoritmos de filtragem

3.1 Definições

Considere-se um espaço de probabilidades $(\Omega, \mathfrak{F}, P)$.

Defina-se I , um conjunto de índices, da forma $I = \{i : i \in \mathbb{N}, 1 \leq i \leq T\}$, para algum T . Sejam $(\mathbf{x}(t))_{t \in I \subset \mathbb{N}}$, $(\mathbf{y}(t))_{t \in I \subset \mathbb{N}}$, $(\theta(t))_{t \in I \subset \mathbb{N}}$ processos estocásticos tais que $\mathbf{x}(t) \in \mathcal{X} \subset \mathbb{R}^p$, $\mathbf{y}(t) \in \mathcal{W} \subset \mathbb{R}^q$ (todos com norma $\|\cdot\|_k$ limitada para todo k) e $\theta(t) \in \mathcal{M} \subset \mathbb{N}^+$ para todo $t \in I$. Seja, ainda, $\mathbf{z}(t) \triangleq (\mathbf{x}(t), \theta(t))$.

Os processos $(\mathbf{x}(t))$ e $(\mathbf{w}(t))$ se supõem independentes, com $(\mathbf{w}(t))$ i.i.d., seguindo uma distribuição normal com variância unitária e média nula. Seja $\theta(t)$ uma cadeia de Markov em tempo discreto com espaço de estados $\mathcal{M}$ e matriz de transição $\mathbf{P}$.

Os vetores $\mathbf{x}(t) \in \mathcal{X}$, $\mathbf{y}(t) \in \mathcal{W}$ e o escalar $\theta(t) \in \mathcal{M}$ estão definidos em um conjunto amostral $\Omega = \mathcal{X} \times \mathcal{W} \times \mathcal{M}$, com um σ -campo associado $\mathfrak{F} = \mathcal{B}(\mathcal{X}) \vee \mathcal{B}(\mathcal{W}) \vee \mathcal{P}(\mathcal{M})$.¹ Desta forma, os mapeamentos $\mathbf{x} : (\Omega, \mathfrak{F}) \rightarrow (\mathcal{X}, \mathcal{B}(\mathcal{X}))$, $\mathbf{y} : (\Omega, \mathfrak{F}) \rightarrow (\mathcal{W}, \mathcal{B}(\mathcal{W}))$ e $\theta : (\Omega, \mathfrak{F}) \rightarrow (\mathcal{M}, \mathcal{P}(\mathcal{M}))$ são variáveis aleatórias definidas em Ω .

Sejam $P_{\mathbf{x}}$, $P_{\mathbf{y}}$ e P_{θ} as medidas de probabilidade induzidas por $\mathbf{x}$, $\mathbf{y}$ e θ , respectivamente. A medida de probabilidade do espaço mensurável $(\Omega, \mathfrak{F})$ será dada por $P \triangleq P_{\mathbf{x}} \times P_{\mathbf{w}} \times P_{\theta}$.

Defina-se $(\mathcal{Y}'_t)_{t \in I}$, uma sequência de filtragens do processo $(\mathbf{z}(t))$, decorrentes do conhecimento das realizações observadas de $\mathbf{y}(t)$. Segue que

$$\mathcal{Y}'_t = \sigma(\mathbf{y}(1 : t)).$$

Seja $\mathcal{Y}_t$ a filtragem completa associada, definida como

$$\mathcal{Y}_t = \mathcal{Y}'_t \vee \mathcal{N}, \tag{3.1}$$

¹ $\mathcal{P}(\cdot)$ indica o conjunto potência de um conjunto discreto e $\mathcal{B}(\cdot)$ denota o conjunto de Borel associado a um conjunto contínuo. Vide, por exemplo, (21)

onde $\mathcal{N}$ representa o subconjunto P -nulo de Ω .

3.2 Especificação estocástica dos SLSM

A classe de sistemas lineares sujeitos a saltos markovianos em tempo discreto é descrita pelas seguintes equações de diferenças:

$$\mathbf{x}(t) = \mathbf{A}_{\theta(t)}\mathbf{x}(t-1) + \mathbf{F}_{\theta(t)}\mathbf{v}(t) \quad (3.2a)$$

$$\mathbf{y}(t) = \mathbf{C}_{\theta(t)}\mathbf{x}(t) + \mathbf{G}_{\theta(t)}\mathbf{w}(t) \quad (3.2b)$$

em que $\mathbf{A}_i$, $\mathbf{C}_i$, $\mathbf{F}_i$ e $\mathbf{G}_i$ são matrizes de dimensões apropriadas. Sejam $\mathbf{A}_i$ e $\mathbf{G}_i$ matrizes quadradas.

Na forma funcional exposta em (3.2), explicita-se a dependência das matrizes $\mathbf{A}$, $\mathbf{C}$, $\mathbf{F}$ e $\mathbf{G}$ (que descrevem a dinâmica do sistema no espaço de estados) com relação ao processo estocástico $(\theta(t))$. Deste modo, associa-se a cada possível $\theta(t) = i$ uma quadra de matrizes $\mathbf{A}_i$, $\mathbf{C}_i$, $\mathbf{F}_i$ e $\mathbf{G}_i$, que regulará tanto a equação de estados quanto a equação de saídas. Este é, portanto, um caso de sistema linear a parâmetros variantes no tempo. A variação temporal destes parâmetros, por sua vez, está condicionada pelo processo $(\theta(t))$, governado pela matriz de transição $\mathbf{P}$.

Considerando-se que $\mathbf{w}(t)$ e $\mathbf{v}(t)$ têm distribuições conhecidas, chame-se esta distribuição de ψ , tal que

$$\psi \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

onde $\mathbf{I}$ é a matriz identidade de dimensão adequada.

Conforme o valor do processo estocástico $(\theta(t))$ no instante t , escolhe-se um conjunto $(\mathbf{A}_i, \mathbf{C}_i, \mathbf{F}_i, \mathbf{G}_i)$ dos possíveis $(\mathbf{A}_1, \mathbf{C}_1, \mathbf{F}_1, \mathbf{G}_1)$, $(\mathbf{A}_2, \mathbf{C}_2, \mathbf{F}_2, \mathbf{G}_2)$, $\dots$, $(\mathbf{A}_M, \mathbf{C}_M, \mathbf{F}_M, \mathbf{G}_M)$. Denomine-se, então, esta coleção de parâmetros (à qual se soma também a matriz de transição $\mathbf{P}$, que também não é necessariamente conhecida)

$$\begin{aligned} \eta \triangleq & \{ (\mathbf{A}_1, \mathbf{C}_1, \mathbf{F}_1, \mathbf{G}_1), \\ & , (\mathbf{A}_2, \mathbf{C}_2, \mathbf{F}_2, \mathbf{G}_2), \dots, \\ & , (\mathbf{A}_M, \mathbf{C}_M, \mathbf{F}_M, \mathbf{G}_M); \mathbf{P} \} \end{aligned}$$

que não se assume, em geral, que seja conhecida *a priori*, e constitui o objeto de estimação. A realização simultânea da filtragem sobre $\hat{\mathbf{x}}(t)$ e $\theta(t)$ consiste em solucionar um problema de estimação conjunta entre as variáveis de estado desconhecidas, tomando como dados

os parâmetros $\bar{\eta}$ e as observações $\mathbf{y}(0), \mathbf{y}(1), \dots, \mathbf{y}(T)$.

3.3 Algoritmos de Filtragem

O problema de filtragem diz respeito à estimação de processos estocásticos não observados, com base no conhecimento de um outro processo que guarda alguma forma de correlação com aquele. Contextualizando-se o sistema definido em (3.2) à luz da teoria de filtros, chame-se $\mathbf{x}$ sinal; $\mathbf{w}$, ruído; $\mathbf{y}$, processo observado. O problema se define pela otimização do valor esperado de alguma função (chamada função-objetivo) do erro de estimação, sendo a mais comum a média quadrática dos erros. É nesta acepção que se entende aqui o problema de filtragem.

Considera-se, primeiramente, o problema de filtragem, isoladamente. Para isto, supõe-se que os parâmetros η sejam conhecidos, e que a cadeia de Markov θ seja observável diretamente. Os resultados desta análise serão utilizados na sequência para deduzir a estimação de parâmetros.

3.3.1 Filtragem ótima

Considerando-se o sistema definido na expressão (3.2), suponha-se, inicialmente, que $\theta(t)$ é previamente conhecido para todo t , e que η é conhecido *a priori*. Desta forma, se supõem conhecidos todos os parâmetros dinâmicos do sistema, para cada t . Trata-se, como mencionado, de um sistema linear no espaço de estados sujeito a um ruído de medida e um processo de inovação normalmente distribuídos. É bem conhecido que o filtro ótimo para este caso, no sentido do erro médio quadrático de estimação, é o filtro de Kalman a parâmetros variantes no tempo.

Seja

$$\hat{\mathbf{x}}_{p|q} \triangleq \mathbb{E}[\mathbf{x}(p)|\mathcal{Y}_q; \theta(1:q); \eta] \quad (3.3)$$

$$\mathbf{Q}_{p|q} \triangleq \mathbb{E}[(\hat{\mathbf{x}}_{p|q} - \mathbf{x}(p))(\hat{\mathbf{x}}_{p|q} - \mathbf{x}(p))'] \quad (3.4)$$

a estimativa da variável de estados $\mathbf{x}$ em $t = p$, dadas as observações de $\mathbf{y}$ realizadas em $t = 0, \dots, q$, condicionando-se também ao conhecimento (que se presumiu) de θ e η . A matriz $\mathbf{Q}_{p|q}$ é a matriz de covariância do erro de estimação associado à estimativa $\hat{\mathbf{x}}_{p|q}$. Iniciando-se com uma estimativa inicial arbitrária para $\hat{\mathbf{x}}_{0|0}$ e $\mathbf{Q}_{0|0}$, iteram-se as equações

para $k = 1, 2, \dots$:

$$\mathbf{Q}_{k|k-1} = \mathbf{A}_k \mathbf{Q}_{k-1|k-1} \mathbf{A}_k' + \mathbf{F}_k \mathbf{F}_k' \quad (3.5)$$

$$\mathbf{K}_k = \mathbf{Q}_{k|k-1} \mathbf{C}_k' (\mathbf{C}_k \mathbf{Q}_{k|k-1} \mathbf{C}_k' + \mathbf{G}_k \mathbf{G}_k')^{-1} \quad (3.6)$$

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{A}_k \hat{\mathbf{x}}_{k-1|k-1} \quad (3.7)$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k (\mathbf{y}(k) - \mathbf{C}_k \hat{\mathbf{x}}_{k|k-1}) \quad (3.8)$$

$$\mathbf{Q}_{k|k} = (\mathbf{I} - \mathbf{K}_k \mathbf{C}_k) \mathbf{Q}_{k|k-1} \quad (3.9)$$

onde $\mathbf{K}_k$ é uma matriz auxiliar, normalmente chamada ganho de Kalman. É importante notar, também, que

$$\begin{aligned} \mathbb{E} [\mathbf{y}(k) | \mathcal{Y}_{k-1}, \theta(1:k); \eta] &= \\ &= \mathbb{E} [\mathbf{C}_k \mathbf{x}(k) + \mathbf{G}_k \mathbf{w}(k) | \mathcal{Y}_{k-1}, \theta(1:k); \eta] \\ &= \mathbf{C}_k \hat{\mathbf{x}}_{k|k-1} \end{aligned}$$

onde a última passagem segue da definição (3.3), e se garante pela propriedade de otimalidade do filtro de Kalman (vide, por exemplo (22)). Além disto,

$$\begin{aligned} \mathbb{E} [\mathbf{y}(k) \mathbf{y}(k)' | \mathcal{Y}_{k-1}, \theta(1:k); \eta] &= \\ &= \mathbb{E} [(\mathbf{C}_k \mathbf{x}(k) + \mathbf{G}_k \mathbf{w}(k)) \cdot \\ &\quad \cdot (\mathbf{C}_k \mathbf{x}(k) + \mathbf{G}_k \mathbf{w}(k))' | \mathcal{Y}_{k-1}, \theta(1:k); \eta] \\ &= \mathbf{C}_k \mathbf{Q}_{k|k-1} \mathbf{C}_k' + \mathbf{G}_k \mathbf{G}_k' \\ &\triangleq \mathbf{\Sigma}_k \end{aligned}$$

Portanto,

$$\mathbf{y}(k) - \mathbf{C}_k \hat{\mathbf{x}}_{k|k-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_k)$$

e

$$\mathbf{\Sigma}_k^{-\frac{1}{2}} (\mathbf{y}(k) - \mathbf{C}_k \hat{\mathbf{x}}_{k|k-1}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) = \psi \quad (3.10)$$

onde $\mathbf{\Sigma}_k^{\frac{1}{2}}$ representa o fator de Cholesky da matriz $\mathbf{\Sigma}$ de variância a priori de $\mathbf{y}(k)$.

É evidente que a equação (3.10) só é válida para o caso hipotético em que a cadeia θ é observável. No entanto, pode-se utilizar esta expressão para auferir o quão provável é que uma determinada sequência seja a verdadeira. Para a sequência verdadeira, o conjunto dos $\mathbf{\Sigma}_k^{-1/2} (\mathbf{y}(k) - \mathbf{C}_k \hat{\mathbf{x}}_{k|k-1})$ é distribuído como uma normal padrão.

A partir daqui, deixa-se de supor que a sequência θ é conhecida, mas se supõe ainda que os parâmetros η são conhecidos (isto será mantido até o término desta seção 3.3).

Defina-se

$$\mathcal{H}_i(k) \triangleq \{\theta(1) = \omega_1^i, \theta(2) = \omega_2^i, \dots, \theta(t) = \omega_k^i\} \quad (3.11)$$

a sequência de realizações de θ , onde o subscrito i denota que esta é a i -ésima entre as possíveis M^k sequências de tamanho k que se poderiam formar, dado que em cada posição há M possibilidades. É interessante criar um mapeamento que defina exatamente a qual sucessão de modos corresponderá esta sequência i . Associe-se ao r -ésimo número da sequência $\omega_1^i, \dots, \omega_r^i, \dots, \omega_t^i$ na equação (3.11) o r -ésimo algarismo do numeral $[i]_M$ (ou seja, do número i escrito na base M).²

Defina-se, inicialmente $\hat{\mathbf{x}}_{k|k}^j$ como a estimativa sobre $\mathbf{x}(k)$ produzida pelo j -ésimo filtro de Kalman a parâmetros variantes. Pelo teorema de Bayes, obtém-se que a probabilidade de que uma sequência de modos de operação $\mathcal{H}_j(k)$ seja a correta é igual a

$$P[\mathcal{H}_j(k) | \mathbf{y}(1:k); \eta] = \frac{P[\mathbf{y}(1:k) | \mathcal{H}_j(k); \eta] P[\mathcal{H}_j(k) | \eta]}{P[\mathbf{y}(1:k) | \eta]}$$

utilizando a lei das probabilidades totais ao denominador, obtém-se

$$P[\mathcal{H}_j(k) | \mathbf{y}(1:k); \eta] = \frac{P[\mathbf{y}(1:k) | \mathcal{H}_j(k); \eta] P[\mathcal{H}_j(k) | \eta]}{\sum_{n=0}^{M^k-1} P[\mathbf{y}(1:k) | \mathcal{H}_n(k); \eta]} \quad (3.12)$$

onde

$$\begin{aligned} P[\mathbf{y}(1:k) | \mathcal{H}_n(k); \eta] &= \\ &= \prod_{k=1}^t \psi \left(\Sigma_k^{-\frac{1}{2}} (\mathbf{y}(k) - \mathbf{C}_k \hat{\mathbf{x}}_{k|k-1}^n) \right) \end{aligned} \quad (3.13)$$

para todo histórico n . O valor é obtido da equação (3.10), dos valores obtidos das expressões (3.5)-(3.9). O fator da direita do numerador da expressão (3.12) é a probabilidade conjunta *a priori* da sequência, dada por

$$P[\mathcal{H}_j(k) | \eta] = \theta_0 \cdot p_{\theta_0, \omega_1^i} \cdot p_{\omega_1^i, \omega_2^i} \cdot \dots \cdot p_{\omega_{k-1}^i, \omega_k^i} \quad (3.14)$$

onde $p_{i,j}$ é o elemento (i, j) da matriz de transição $\mathbf{P}$.

Note-se que há três condições iniciais que precisariam ser determinadas: θ_0 , $\hat{\mathbf{x}}_{0|0}$ e $\mathbf{Q}_{0|0}$. Por simplicidade, assume-se aqui a ergodicidade, supondo que os modos naturais do sistema decairão com suficiente velocidade, e atribui-se a estas um valor arbitrário. Mais rigorosamente, estas condições precisariam ser estimadas juntamente com η , o que não será feito aqui. O modo de fazê-lo, no entanto, é idêntico ao procedimento que se

²Por exemplo, se $M = 2$ e $t = 4$, o histórico de índice 10, entre os 0 a 15 possíveis, será a sequência 0110, ou $[10]_2$, onde cada algarismo se associa ao modo vigente em um período entre $t = 1$ e $t = 4$.

apresentará para os demais parâmetros.

Calculou-se, desta forma, a probabilidade de que cada histórico $\mathcal{H}_j(t)$ seja o que realmente se produziu, entre os M^t , possíveis. Isto conclui a estimação do processo $(\theta(t))$. Pode-se agora utilizar a lei das probabilidades totais para encontrar as estimativas ótimas para a variável de estados $\mathbf{x}(t)$. Dado que nas expressões (3.5)-(3.9) foram calculadas as estimativas ótimas condicionais a cada um dos modos, e que a expressão (3.13) calcula a probabilidade de que cada um destes modos esteja em operação, o próximo passo é combinar ambos os resultados. Seja $\hat{\mathbf{x}}_{k|k}^o$ a estimativa do filtro ótimo, que se busca. Segue que

$$\begin{aligned}\hat{\mathbf{x}}_{k|k}^o &= \mathbb{E}[\mathbf{x}(k)|\mathcal{Y}_k; \eta] \\ &= \sum_{n=0}^{M^k-1} \mathbb{E}[\mathbf{x}(k)|\mathcal{H}_n(k), \mathcal{Y}_k; \eta] \mathbb{P}[\mathcal{H}_n(k)|\mathcal{Y}_k; \eta] \\ &= \sum_{n=0}^{M^k-1} \hat{\mathbf{x}}_{k|k}^n \mathbb{P}[\mathcal{H}_n(k)|\mathcal{Y}_k; \eta]\end{aligned}$$

que é a estimativa ótima que se buscava.

A aplicação deste algoritmo ótimo não é factível, evidentemente. Apresenta-se na literatura um conjunto de possíveis soluções a esta dificuldade:

- Utilizar uma filtragem não-linear sub-ótima;
- Aproximar o filtro não-linear ótimo pelo filtro linear ótimo;
- Introduzir simplificações adicionais ao modelo que permitam eliminar a dependência serial.

Conforme justificou-se em 1.4, não será considerada a opção de simplificar o modelo. Resta, portanto, trabalhar com as duas possibilidades restantes.

3.3.2 Algoritmo GPB2

O algoritmo GPB2 (1) apresenta uma alternativa de menor custo computacional ao algoritmo ótimo, implicando evidentemente uma perda de desempenho. Se o algoritmo ótimo corresponde a um banco de M^t filtros em paralelo, de maneira análoga, o algoritmo GPB2 requer M^2 filtros em paralelo. Vide diagrama de blocos na figura 5, retirada de

(5). O algoritmo consiste, essencialmente, em um truncamento da forma

$$\begin{aligned}
 \hat{\mathbf{x}}_{k|k}^o &= \sum_{n=0}^{M^k-1} \hat{\mathbf{x}}_{k|k}^n P[\mathcal{H}_n(k)|\mathcal{Y}_k; \eta] \\
 &\approx \sum_{\theta(k-1)=0}^{M-1} \sum_{\theta(k)=0}^{M-1} E[\mathbf{x}(k)|\theta(k), \theta(k-1); \mathbf{y}(k), \mathbf{y}(k-1); \eta] \cdot \\
 &\quad \cdot P[\theta(k), \theta(k-1)|\mathbf{y}(k), \mathbf{y}(k-1); \eta] \\
 &\triangleq \tilde{\mathbf{x}}_{k|k}^o \quad (3.15)
 \end{aligned}$$

onde cada $\{\theta(k), \theta(k-1)\}$ são os M^2 históricos truncados, considerando-se apenas os dois instantes mais recentes, $k = t$ e $k = t-1$. Portanto, todos os históricos que diferem apenas em valores anteriores são agregados juntamente com todas as demais sequências que tenham em comum os mesmos resultados da cadeia de Markov nos dois últimos instantes. Para realizar o cálculo expresso na fórmula (3.15), iteram-se as expressões (3.5)-(3.9) utilizando-se as observações $\mathbf{y}(k), \mathbf{y}(k-1)$. Note-se, no entanto, que é necessário fornecer valores para $\hat{\mathbf{x}}_{k-2|k-2}$ e $\mathbf{Q}_{k-2|k-2}$. Estes valores devem ser obtidos durante uma fase de truncamento da filtragem ótima.

Sejam $\tilde{\mathbf{x}}_{k|k}^j$, para $j = 1, \dots, M$ e $k = 0, \dots, t$, aproximações pseudobayesianas de $\mathbf{x}(k)$, dadas as observações $\mathbf{y}(1 : k)$. Este algoritmo propagará um número de variáveis de estado com ordem de grandeza de M entre iterações – ao contrário do algoritmo ótimo, que propaga um número da ordem de M^k variáveis³. A essência do algoritmo sub-ótimo consiste em efetuar, a cada iteração, uma redução de complexidade. Em $t = k$, propagando-se para a próxima iteração M probabilidades associadas aos possíveis valores de $\theta(k)$ e M vetores de valores esperados condicionais de $\mathbf{x}$ e suas respectivas matrizes de covariância, obtêm-se M^2 possibilidades para $\{\theta(k), \theta(k+1)\}$, M^2 vetores de valores esperados condicionais de $\mathbf{x}$ e M^2 matrizes de covariância (para cada possibilidade de $\{\theta(k), \theta(k+1)\}$). Estes M^2 valores esperados e covariâncias serão reduzidos, a cada iteração a M valores esperados e covariâncias. Idealmente, isto deve ser feito de forma a perder o mínimo possível de informação sobre função densidade de probabilidade das variáveis de estado⁴. No entanto, não é adequado para o desempenho do algoritmo que se acrescente uma complexidade excessiva a esta redução, de forma que se utilizará aqui o procedimento pseudobayesiano simples.

³Para obter a probabilidade de cada um dos M^{k+1} históricos em $t = k+1$, é necessário propagar para a próxima iteração as M^k probabilidades modais, os M^k vetores de valores esperados condicionais de $\mathbf{x}$ e suas respectivas matrizes de covariância em $t = k$.

⁴Vide (23) para uma discussão sobre redução de misturas gaussianas baseada na divergência de Kullback-Leibler.

Chame-se $\tilde{\mathbf{x}}_{k|k}^1, \dots, \tilde{\mathbf{x}}_{k|k}^M$ as M estimativas e $\tilde{\mathbf{Q}}_{k|k}^1, \dots, \tilde{\mathbf{Q}}_{k|k}^M$ as M matrizes de covariância armazenadas para o instante k . Atribua-se, inicialmente

$$\begin{aligned}\tilde{\mathbf{x}}_{k|k}^1 &= \dots = \tilde{\mathbf{x}}_{k|k}^M := \hat{\mathbf{x}}_{0|0} \\ \tilde{\mathbf{Q}}_{k|k}^1 &= \dots = \tilde{\mathbf{Q}}_{k|k}^M := \mathbf{Q}_{0|0}\end{aligned}$$

onde as estimativas iniciais $\hat{\mathbf{x}}_{0|0}$, $\mathbf{Q}_{0|0}$ são arbitrárias (conforme mencionado na seção precedente) seja também uma condição inicial θ_0 arbitrária.

Ittere-se para os instantes $k = 2, 3, \dots, t$:

1. Ittere-se para os filtros $j = 0, 1, \dots, M^2 - 1$

I Encontrar a sequência de modos correspondente a j , utilizando a expressão (3.11) para encontrar ω_1^j e ω_2^j ;

II Realizar a aquisição de $\mathbf{y}(k-1)$ e $\mathbf{y}(k)$. Utilizar as expressões (3.5)-(3.9), começando-se com $\tilde{\mathbf{x}}_{k-2|k-2}^j$ e $\tilde{\mathbf{Q}}_{k-2|k-2}^j$. Utilizam-se os parâmetros dinâmicos correspondentes aos ω_1^j e ω_2^j encontrados no item anterior. Calcular $\tilde{\mathbf{x}}_{k-1|k-1}^j$, $\tilde{\mathbf{x}}_{k|k}^j, \tilde{\mathbf{Q}}_{k-1|k-1}^j, \tilde{\mathbf{Q}}_{k|k}^j$

III Calcular a verossimilhança associada à passagem do j -ésimo filtro, utilizando a expressão utilizada em (3.13). Para k , seja $\Sigma = \tilde{\mathbf{Q}}_{k|k}^j$. Atribua-se

$$\mathcal{L}_j(k) := \psi \left(\Sigma^{-\frac{1}{2}} \left(\mathbf{y}(k) - \mathbf{C}_{\omega_2} \tilde{\mathbf{x}}_{k|k-1}^j \right) \right) \quad (3.16)$$

IV (Opcional) Utilizar um passo do *smoother* RTS⁵ para calcular $\tilde{\mathbf{x}}_{k-1|k}^j$. Parte-se dos valores encontrados no passo anterior para $\tilde{\mathbf{x}}_{k|k}^j$, $\tilde{\mathbf{Q}}_{k|k-1}^j$ e $\tilde{\mathbf{Q}}_{k|k}^j$, e calcula-se:

$$\mathbf{K}_{k-1}^b = \tilde{\mathbf{Q}}_{k-1|k-1}^j \mathbf{F}_{\omega_1}' \left(\tilde{\mathbf{Q}}_{k|k-1}^j \right)^{-1} \quad (3.17)$$

$$\tilde{\mathbf{Q}}_{k-1|k}^j = \tilde{\mathbf{Q}}_{k-1|k-1}^j - \mathbf{K}_{k-1}^b \left(\tilde{\mathbf{Q}}_{k|k-1}^j - \tilde{\mathbf{Q}}_{k|k}^j \right) \left(\mathbf{K}_{k-1}^b \right)' \quad (3.18)$$

$$\tilde{\mathbf{x}}_{k|k}^j = \tilde{\mathbf{x}}_{k+1|k}^j + \mathbf{K}_{k-1}^b \left(\tilde{\mathbf{x}}_{k|k}^j - \tilde{\mathbf{x}}_{k+1|k}^j \right) \quad (3.19)$$

Este passo melhora o desempenho do rastreamento, mas não interfere na função verossimilhança. Deixa-se como opcional.

2. Calcular a probabilidade de cada modo, de forma análoga à equação (3.13). Para

⁵Rauch-Tung-Striebel (ver, por exemplo, (22))

$j = 1, \dots, M$:

$$\begin{aligned} P[\theta(k) = j | \mathbf{y}(k), \mathbf{y}(k-1); \eta] &= \\ &= \frac{\mathcal{L}_j(k) P[\theta(k) = j | \eta]}{\sum_{n=1}^M \mathcal{L}_n(k) P[\theta(k) = n | \eta]} \end{aligned} \quad (3.20)$$

onde as estimativas $P[\theta(k) = j | \eta]$ são as estimativas *a priori* definidas na expressão (3.14).

3. Calcular a verossimilhança total ponderada para o instante k , atribuindo-se:

$$\mathcal{L}(k) := \sum_{n=1}^M \mathcal{L}_n(k) P[\theta(n) = j] \quad (3.21)$$

4. Prosseguir para o próximo ciclo, utilizando $\tilde{\mathbf{x}}_{k|k}^j$ e $\tilde{\mathbf{Q}}_{k-1|k-1}^j$ como condições iniciais, para $j = 1, \dots, M$.

Finalmente, calcula-se a verossimilhança total do modelo. Dado que está é uma função implícita dos parâmetros η , defina-se

$$\mathcal{L}'(\eta) \triangleq \prod_{k=2}^t \mathcal{L}(k) \quad (3.22)$$

e

$$\mathcal{L}(\eta) \triangleq \sum_{k=2}^t \log \mathcal{L}(k) \quad (3.23)$$

a log-verossimilhança associada. Isto conclui o algoritmo. Na próxima seção, busca-se realizar a otimização de $\mathcal{L}_\eta$.

3.4 Algoritmo IMM

Ao analisar o algoritmo GPB, é natural observar que existe uma certa incongruência no modo como se geram e armazenam as variáveis de estado. De fato, basta supor um sistema em que a probabilidade de transição seja grande para concluir que existem formas mais eficientes de propagar as $M^{(K-1)}$ estimativas.

No filtro IMM, armazenam-se como a i -ésima variável de estado a estimativa a priori para esta variável, dado que seja i o modo de operação em $t+1$. Isto implica gerar hipóteses compostas a partir das probabilidades a posteriori descritas na seção anterior, ponderando cada hipótese pela probabilidade de transição entre o modo correspondente

Figura 5: Diagrama de blocos do algoritmo GPB2

a esta em t e o modo i em $t + 1$.

O conceito fundamental do algoritmo IMM é propagar M variáveis de estado $\tilde{x}_i(t)$. Dado que se dispõe de apenas M filtros, cada um associado a um modo de operação, realiza-se uma etapa de interação entre as variáveis de saída obtidas na etapa de $t - 1$. Obtém-se, em $t - 1$, as M variáveis de estado associadas a cada um M filtros de forma que cada variável $x^{(i)}(t - 1)$ de estado corresponda à melhor estimativa de $x(t - 1)$ dado que o i -ésimo modo de operação (associado ao i -ésimo filtro) esteja vigente em t . Estas variáveis $x^{(i)}(t - 1)$ são calculadas pela ponderação das estimativas $\tilde{x}_i(t - 1)$, realizada por coeficientes, chamados de μ .

Conforme demonstrado em (14) e (1), os coeficientes de ponderação, chamados $\mu_{i|j}$ são dados por

$$\begin{aligned} \mu_{i|j}(k-1) &\triangleq \text{P}[\theta(k-1) = i | \theta(k) = j, \mathcal{Y}_{k-1}] \\ &= \frac{\text{P}[\theta(k) = j | \theta(k-1) = i, \mathcal{Y}_{k-1}] \text{P}[\theta(k-1) = i | \mathcal{Y}_{k-1}]}{\sum_{n=1}^M \text{P}[\theta(k) = j | \theta(k-1) = n, \mathcal{Y}_{k-1}] \text{P}[\theta(k-1) = n | \mathcal{Y}_{k-1}]} \end{aligned}$$

de forma que as variáveis de estado a utilizar como entradas para cada um dos M filtros em paralelo é dada por

$$x^{(j)}(t-1) = \sum_{i=1}^M M \tilde{x}_i(t-1) \mu_{i|j}(k-1)$$

À exceção desta etapa, demais etapas da filtragem são idênticas ao apresentado no algoritmo GPB2. Vide diagrama de blocos na figura 6, retirada de (5).

Figura 6: Diagrama de blocos do algoritmo IMM

3.5 Algoritmo linear

Uma crítica frequente aos algoritmos GPB e IMM é o fato de estes não estarem fundamentados em nenhum critério objetivo de otimalidade, sendo apenas aproximações com grau de confiabilidade desconhecido. Não se garante, por exemplo, que o valor esperado dos erros seja nulo sempre.

É fato que os algoritmos GPB e IMM apresentam bons resultados na grande maioria dos casos, mas o fato de o algoritmo linear estar fundamentado em um critério demonstrável (vide (15), (24), (2)) constitui uma vantagem significativa.

Além disto, o algoritmo linear possui baixo custo computacional, e obedece a uma equação algébrica de Riccati que pode ser calculada previamente (como é o caso no filtro de Kalman).

3.6 Comparação entre os algoritmos de filtragem

Efetuuou-se uma comparação entre os três algoritmos, para o caso escalar, escolhendo-se os parâmetros: $A_1 = 0,9$, $A_2 = -0,8$, $C_1 = 1,2$, $C_2 = 0,8$, $F_1 = \sqrt{1,2}$, $F_2 = \sqrt{0,8}$,

$G_1 = \sqrt{0,3}$, $G_2 = \sqrt{0,2}$, $p_{11} = 0,8$ e $p_{22} = 0,7$. O ruído de medida e o processo de inovação são, neste caso, normalmente distribuídos com variância unitária.

O resultado exibido na figura 7 demonstra um excerto representativo da série histórica de $T = 3000$ pontos gerados para o SLSM em questão. Note-se que o algoritmo GPB2 segue possui um erro de estimação menor com alguma distância, enquanto os algoritmos Linear e IMM atingem desempenhos comparáveis. Este experimento foi realizado, nos três casos, com parâmetros conhecidos, dado que o objetivo era simplesmente verificar o desempenho dos filtros em questão.

As figuras 8, 8 e 8 exibem as características estatísticas do erro de estimação. Devido ao seu desempenho, escolheu-se o algoritmo GPB2 para realizar o ensaio de convergência dos parâmetros do modelo.

Figura 7: Comparação entre o desempenho dos algoritmos GPB, IMM e Linear.

Figura 8: Histograma de erros de estimação para o algoritmo GPB

Figura 9: Histograma de erros de estimação para o algoritmo IMM

Figura 10: Histograma de erros de estimação para o algoritmo Linear

4 Metodologia para estimação de parâmetros

4.1 Algoritmo

Os algoritmos apresentados na seção 3.3 permitem o cálculo das estimativas condicionais aos parâmetros η . O propósito deste estudo é, no entanto, justamente a estimação destes parâmetros que se supuseram conhecidos até este momento. Com efeito, é bastante vantajosa a abordagem de se estudar inicialmente a verossimilhança condicional.

O problema consiste em encontrar o argumento $\bar{\eta}$ que minimiza a função custo $\mathcal{L}_\eta$, tal que

$$\bar{\eta} = \arg \max_{\bar{\eta}} [\mathcal{L}(\eta)]$$

É evidente que esta otimização é um processo bastante complicado, devido à complexidade da função custo J . Pode-se, no entanto, realizar uma maximização numérica. Para isto, é suficiente que se possa calcular o valor da função J em cada ponto - o que foi descrito no capítulo precedente.

Embora possa haver uma grande diversidade de métodos para realizar esta otimização, utilizou-se nos experimentos o algoritmo BFGS (Broyden-Fletcher-Goldfarb-Shanno), que se caracteriza por ser um método *quasi*-Newton, cuja matriz Hessiana é aproximada pelos sucessivos gradientes. O método utilizado pode ser sintetizado no seguinte algoritmo.

I Escolhe-se uma estimativa inicial para os parâmetros desconhecidos. Estes são reunidos em um vetor $\mathbf{u}_0$. Crie-se um mapeamento $\mathcal{M}$ que realiza esta transformação, e seja $\mathcal{M}^{-1}$ o mapeamento inverso. Seja, portanto, $\bar{\eta}_i = \mathcal{M}^{-1}(\mathbf{u}_i)$. Por simplicidade, defina-se igualmente $\mathcal{L}(\mathbf{u}) \equiv \mathcal{L}(\mathcal{M}^{-1}(\mathbf{u}))$. O mapeamento $\mathcal{M}$ transforma o conjunto formado pelas matrizes $\mathbf{A}_1, \dots, \mathbf{A}_M$, $\mathbf{C}_1, \dots, \mathbf{C}_M$, $\mathbf{F}_1, \dots, \mathbf{F}_M$, $\mathbf{G}_1, \dots, \mathbf{G}_M$ e $\mathbf{P}$ em um vetor real $\mathbf{u}$. Isto é necessário porque a função verossimilhança é calculada com respeito ao conjunto de matrizes, ao passo que o algoritmo de otimização numérica

precisa de que os parâmetros estejam dispostos em forma vetorial.

II Seja $\mathbf{u}_0 \triangleq \mathcal{M}^{-1}(\bar{\eta}_0)$ uma estimativa inicial. Pela expressão (3.23), calcule-se $\mathcal{L}(\mathbf{u}_0)$, a log-verossimilhança de $\bar{\eta}_0$, e seu respectivo gradiente $\nabla \mathcal{L}(\mathbf{u}_0)$, que deve ser calculado numericamente. Sejam $\mathbf{B}_i$ matrizes quadradas, para todo i . Defina-se $\mathbf{B}_0 = \mathbf{I}$, a matriz identidade apropriada.

III Até que se atinja um ponto crítico, itere em passos $i = 1, 2, \dots$:

(a) Realize-se uma busca linear na direção $\mathbf{d} = -\mathbf{B}_i \nabla \mathcal{L}(\mathbf{u}_i)$, até que se encontre um fator α tal que o ponto $\mathbf{u}_{i+1} = \alpha \mathbf{d}$ satisfaça a $\mathcal{L}(\mathbf{u}_{i+1}) < \mathcal{L}(\mathbf{u}_i)$.

(b) Verifique-se a consistência dos parâmetros para determinar se o ponto $\mathbf{u}_{i+1}$ é válido. As variâncias devem ser positivas, e as linhas da matriz de transição devem somar um. Se o ponto for inválido, retorna-se ao passo anterior e realiza-se nova busca linear.

(c) Atribua-se

$$\mathbf{r} := \nabla \mathcal{L}(\mathbf{u}_{i+1}) - \nabla \mathcal{L}(\mathbf{u}_i) \quad (4.1)$$

(d) Atribua-se $\mathbf{s} := \alpha \mathbf{d}$. Calcule-se a nova aproximação da matriz hessiana, dada por

$$\mathbf{B}_{i+1} = \mathbf{B}_i + \frac{\mathbf{r}\mathbf{r}'}{\mathbf{r}'\mathbf{s}} - \frac{\mathbf{B}_i\mathbf{s}(\mathbf{B}_i\mathbf{s})'}{\mathbf{s}'\mathbf{B}_i\mathbf{s}} \quad (4.2)$$

Ao se atingir o ponto ótimo, cessa-se a otimização. A inversa da matriz hessiana final é utilizada na obtenção de intervalos de confiança para as estimativas. Os elementos da diagonal principal da inversa desta matriz hessiana corresponderão à estimativa da variância dos valores obtidos.

4.2 Resultados experimentais

Neste trabalho, utilizam-se métodos de gradiente para realizar as maximizações de verossimilhança. É importante, neste caso, verificar a adequação dos algoritmos escolhidos para solucionar o problema em questão. A função verossimilhança em questão é bastante complexa, e apresenta não-linearidades que poderiam afetar de forma significativa o desempenho da otimização numérica. Para quantificar o desempenho do algoritmo proposto e estabelecer sua sensibilidade à escolha de condições iniciais, realizou-se um experimento em que se simularam 400 condições diferentes para $\mathbf{u}_0$, retiradas (pseudo)aleatoriamente de uma distribuição uniforme. A série observada $y_0, y_1, \dots, y_T$ foi mantida constante,

com o objetivo de verificar a consistência dos valores otimizados. A tabela 1 apresenta os valores utilizados para gerar a série observada. As probabilidades de transição utilizadas foram $p_{11} = 0,8$ e $p_{22} = 0,7$.

A amostra foi gerada com $T = 400$ observações. Note-se que o tamanho estendido da amostra é importante para que o desvio esperado entre os parâmetros obtidos pela otimização e os parâmetros reais seja pequeno. Embora nesta seção ainda não se queira verificar a convergência deste aspecto em particular, pode-se observar graficamente a proximidade do ponto ótimo da função verossimilhança com os valores reais utilizados para gerar a série.

Verifique-se, por ora, se o algoritmo proposto anteriormente converge para um mesmo ponto, independentemente das condições iniciais. Conforme dito anteriormente, utilizou-se uma distribuição uniforme para simular condições iniciais a partir das quais foi realizada a otimização. Os parâmetros desta distribuição encontram-se na tabela 2. Realizaram-se $n = 400$ simulações com estes parâmetros. Estabeleceu-se um limite de 100 iterações, e uma tolerância da ordem de 10^{-2} para a otimalidade. Nas figuras 11, 12 e 13, representam-se com quadrados as condições iniciais consideradas. Com círculos, os valores finais do algoritmo.

Verificou-se que:

- Dos $n = 400$ valores considerados para as condições iniciais, 335 convergiram, dentro de 100 iterações, para uma região em que $\hat{A}_1 = 0,865 \pm 0,050$, $\hat{A}_2 = -0,775 \pm 0,050$, $\hat{F}_1 = 0,930 \pm 0,050$, $\hat{F}_2 = 1,010 \pm 0,050$, $\hat{p}_{11} = 0,812 \pm 0,050$ e $\hat{p}_{22} = 0,675 \pm 0,050$. Isto corresponde a 83,75% do total das amostras;
- Da totalidade das amostras consideradas houve 39 (ou 9,75% do total) casos em que os valores convergiram para $\hat{A}_1 = 0,57 \pm 0,10$, $\hat{A}_2 = -0,57 \pm 0,10$, $\hat{F}_1 = 1,71 \pm 0,10$, $\hat{F}_2 = 1,27 \pm 0,10$, $\hat{p}_{11} = 0,82 \pm 0,10$ e $\hat{p}_{22} = 0,82 \pm 0,10$. Estes casos, bastante interessantes, constituem uma *solução degenerada*, em que não há distinção entre os modos, tornando-se estes equivalentes (do ponto de vista dos parâmetros dinâmicos) e equiprováveis. Note-se, no entanto, que as variâncias dos modos são distintas. ;
- Houve, por fim, 26 casos em que foi excedido o limite de 100 iterações sem que se alcançasse um ponto ótimo. Verifica-se que, caso este limite de iterações seja ampliado, o algoritmo convergirá para um dos dois pontos críticos mencionados nos itens anteriores.

Tabela 1: Parâmetros reais utilizados para gerar a série observada

Modelo	$\mathbf{A}$	$\mathbf{F}$	$\mathbf{C}$	$\mathbf{G}$	p_{ii_0}
1	0,9	$\sqrt{1,2}$	1	$\sqrt{0,3}$	0,8
2	-0,8	$\sqrt{0,8}$	1	$\sqrt{0,2}$	0,7

Tabela 2: Condições iniciais para a otimização: máximos e mínimos da distribuição uniforme

Modelo	$\mathbf{A}_0$	$\mathbf{F}_0$	p_{ii_0}
Máximo	1	2	1
Mínimo	-1	0	0

Figura 11: Condições iniciais e valores de convergência para os parâmetros $\mathbf{A}_1$ e $\mathbf{A}_2$

Figura 12: Condições iniciais e valores de convergência para os parâmetros F_1 e F_2

Figura 13: Condições iniciais e valores de convergência para os parâmetros p_{11} e p_{22}

4.3 Comentários

Conclui-se que, embora haja a possibilidade de que se atinjam soluções degeneradas localmente ótimas, o algoritmo é bastante robusto. Estes pontos degenerados são de fácil detecção, já que correspondem ao caso em que não há distinção entre os modos. Caso se atinja uma tal solução, pode-se reiniciar o algoritmo com uma nova condição inicial aleatória. Verificou-se, portanto, a robustez do algoritmo de otimização escolhido, dado que se observou sua convergência para um grande número de condições iniciais.

Resta, agora, estabelecer se estes pontos críticos para os quais o algoritmo de otimização converge são os valores que foram originalmente utilizados para gerar numericamente a série empregada como objeto da otimização - ou seja, cumpre estabelecer a consistência deste estimador. Pode-se observar, nas figuras 11, 12 e 13, que os pontos de convergência são bastante próximos do que se esperaria. É necessário verificar, agora:

- Conforme se aumenta o tamanho T da amostra, como se comporta essa diferença? Pode-se observar uma convergência (ao menos aparentemente) monotônica ao valor originalmente utilizado para gerar a série observada?
- Para diferentes amostras aleatórias, com os mesmos parâmetros e um determinado tamanho T , como se distribuem estes desvios?

Verifica-se a seguir a resposta a estas questões.

5 *Simulações e resultados*

5.1 Desempenho da estimação e convergência

5.1.1 Ensaio com amostra longitudinal

Realizaram-se dois ensaios, considerando-se em cada um amostras progressivas entre $T = 1$ e $T = 3000$ pontos. Em ambos, considerou-se o caso escalar, com ruídos normalmente distribuídos, e uma cadeia de Markov com dois estados possíveis. Supôs-se que os parâmetros $\mathbf{C}$ e $\mathbf{G}$, referentes à equação de observação, eram bem conhecidos. O exercício consiste, portanto, em estimar simultaneamente os parâmetros A_1 , A_2 , F_1 , F_2 , p_{11} e p_{22} . Realizou-se uma otimização concomitante por máxima verossimilhança dos seis parâmetros. Utilizou-se a matriz de transição

$$\mathbf{P} = \begin{bmatrix} 0,9 & 0,1 \\ 0,2 & 0,8 \end{bmatrix}$$

e os demais parâmetros conforme expostos na Tabela 3.

Os dois ensaios foram realizados com o mesmo conjunto de parâmetros, variando-se a potência do erro de medida, para que se verificasse a robustez da identificação com respeito aos desvios do processo de filtragem. O primeiro experimento foi realizado tomando-se a potência do processo estocástico do ruído de medida ($\mathbf{w}(t)$) como um quarto do sinal que se buscava identificar - o processo de inovação $\mathbf{v}(t)$. O segundo experimento foi obtido igualando-se a potência de ambos, de forma a introduzir uma incerteza de estimação adicional.

Utilizaram-se como estimativas iniciais $A_1(0) = 1$, $A_2(0) = -1$, $F_1(0) = 1$, $F_2(0) = 1$, $p_{11}(0) = 0,5$ e $p_{22}(0) = 0,5$.

As figuras 14, 16 e 18 mostram a convergência dos parâmetros A , F e P no Caso I, em que esta relação sinal ruído (também chamada signal-to-noise ratio, ou SNR) é de 4. As figuras 15, 17 e 19 exibem o Caso II, em que $SNR = 1$. Exibem-se também os intervalos

Tabela 3: Números considerados nos experimentos

Experimento	Modelo	A	F	C	G	Relação Sinal/Ruído
I	1	0,9	$\sqrt{1,2}$	1	$\sqrt{0,3}$	4
	2	-0,8	$\sqrt{0,8}$	1	$\sqrt{0,2}$	
II	1	0,9	$\sqrt{1,2}$	1	$\sqrt{1,2}$	1
	2	-0,8	$\sqrt{0,8}$	1	$\sqrt{0,8}$	

de 95% de confiança para as estimativas.

Nas figuras 14, 15, 16, 17, 18 e 19, as linhas tracejadas representam os parâmetros originais utilizados para gerar a série de dados observada. Pode-se concluir que a convergência das estimativas para os valores reais é bastante satisfatória, mesmo sendo bastante grande o número de parâmetros que se otimizam simultaneamente.

Observa-se que os parâmetros **A** e **P** convergem de forma aparentemente não viesada, persistindo uma pequena variabilidade em torno dos valores originalmente utilizados para gerar a série de dados. O parâmetro **F** exibe aparentemente um viés, que decorre não-otimalidade do filtro utilizado. Especialmente no Caso II, em que os erros de estimação são bastante elevados (devido à potência do erro de medida), verifica-se que há interferência na estimação das variâncias. Ainda assim, percebe-se uma clara convergência dos parâmetros.

5.1.2 Ensaio em seção transversal

Busca-se, em seguida, avaliar a distribuição das estimativas em um ponto fixo no tempo, realizando-se simulações com um conjunto de amostras geradas aleatoriamente. Utilizaram-se os parâmetros que constam na tabela 4 e as probabilidades de transição $p_{11} = 0,8$ e $p_{22} = 0,7$. Considerando $R = 300$ simulações, com $T = 200$ pontos em cada amostra, observaram-se os histogramas exibidos nas figuras 20, 21, 22, 23, 24, 25.

A tabela 5 representa as médias de conjunto e alguns parâmetros estatísticos da distribuição obtida para cada um dos parâmetros que participam do processo de estimação. Note-se que as médias convergem visivelmente para os valores reais da distribuição. A presença de uma assimetria e curtose denota, no entanto, a presença de alguns possíveis pontos espúrios, provavelmente decorrentes da inexatidão do processo numérico de estimação. A coluna *Estatística p* da tabela 5 representa a estatística computada por um teste de Kolmogorov-Smirnov contra a distribuição normal. Embora se suspeite que haja pontos espúrios, em nenhum dos casos foi possível rejeitar, com 5% de confiabilidade, a normalidade da distribuição por este teste.

Figura 14: Convergência dos parâmetros A_1 e A_2 , Caso I

Figura 15: Convergência dos parâmetros A_1 e A_2 , Caso II

Figura 16: Convergência dos parâmetros F_1 e F_2 , Caso I

Figura 17: Convergência dos parâmetros F_1 e F_2 , Caso II

Figura 18: Convergência dos parâmetros p_{11} e p_{22} , Caso I

Figura 19: Convergência dos parâmetros p_{11} e p_{22} , Caso II

Os gráficos 26, 27, 28, 29, 30 e 31 apoiam a tese de que os erros são normalmente distribuídos, mas com a presença de poucos pontos espúrios isolados. Os percentis das distribuições dos parâmetros se aproximam muito da normal padrão, como se pode observar.

Tabela 4: Números considerados nos experimentos

Modelo	A	F	C	G
1	0,9	$\sqrt{1,2}$	1	$\sqrt{0,3}$
2	-0,8	$\sqrt{0,8}$	1	$\sqrt{0,2}$

Tabela 5: Resultados do conjunto de experimentos

Parâmetro	Modelo	Média de conjunto	Variância	Assimetria	Curtose	Estatística p	Teste KS
A	1	0,8889	0,0048	-0,9210	5,5374	0,0780	NR
	2	-0,7885	0,0077	1,1578	8,6051	0,2667	NR
F	1	1,0970	0,0136	-0,5389	4,7908	0,7035	NR
	2	0,8882	0,0327	-0,6561	6,7662	0,1765	NR
p_{11}	Ambos	0,7990	0,0033	-0,3592	3,1453	0,6462	NR
p_{22}	Ambos	0,6934	0,0065	-0,6372	3,4472	0,2037	NR

Figura 20: Convergência do parâmetro A_1

Figura 21: Convergência do parâmetro A_2

Figura 22: Convergência do parâmetro F_1 Figura 23: Convergência do parâmetro F_2

Figura 24: Convergência do parâmetro p_{11}

Figura 25: Convergência do parâmetro p_{22}

Figura 26: Gráfico quantil-quantil para o parâmetro A_1

Figura 27: Gráfico quantil-quantil para o A_2

Figura 28: Gráfico quantil-quantil para o F_1

Figura 29: Gráfico quantil-quantil para o F_2

Figura 30: Gráfico quantil-quantil para o p_{11}

Figura 31: Gráfico quantil-quantil para o p_{22}

6 *Conclusões*

Com as simulações realizadas, torna-se evidente a factibilidade da estimação de parâmetros nos sistemas lineares sujeitos a saltos markovianos. Embora se estivesse calibrando seis parâmetros simultâneos, e em uma amostra relativamente reduzida, os resultados obtidos podem ser considerados bastante satisfatórios. A técnica de maximização numérica da verossimilhança apresentou bom desempenho, e representa uma alternativa viável para a aplicação prática desta classe de modelos a parâmetros variáveis no tempo.

No entanto, algumas questões foram abordadas ao longo do trabalho, e restam como possíveis temas para investigações futuras. Por exemplo,

- É possível evitar os *outliers* no processo de otimização, e aproximar mais a distribuição das estimativas de uma normal padrão?
- É possível demonstrar formalmente a convergência e a normalidade assintótica das estimativas? Pode-se quantificar a eficiência do estimador e a ordem de grandeza da potência do erro em excesso ao estimador ótimo?
- Embora se tenha conseguido obter um estimador empiricamente consistente para os parâmetros, a técnica de otimização numérica implica que, a cada nova observação acrescentada à amostra, é necessário refazer a otimização. O processo é adequado para aplicações em que tempo de resposta não é crítico, e o crescimento do tempo para a otimização em função do tamanho da amostra pode ser contornado com um janelamento dos dados. Seria, no entanto, conveniente, obter uma lei de recursão que possibilitasse a atualização das estimativas ao se acrescentarem novas observações.

No cômputo final, pode-se afirmar que os objetivos deste trabalho foram plenamente atingidos. Apresentaram-se técnicas de filtragem e estimação de parâmetros para a classe dos sistemas lineares sujeitos a saltos, e verificou-se empiricamente o desempenho dos estimadores.

6.1 Agradecimentos

O autor agradece à FAPESP pelo apoio dispensado à execução desta pesquisa (Processo 08/51594 – 0).

APÊNDICE A – Fundamentos teóricos

Este capítulo tem como objetivo apresentar os fundamentos que poderão ser utilizados no decurso do desenvolvimento deste estudo. Os conceitos relacionados a espaços de probabilidade e teoria da medida foram, em grande parte, baseados na exposição encontrada em (21).

Os fundamentos de processos estocásticos e filtragem aqui apresentados devem-se principalmente a (25) e a (26). Utilizaram-se também argumentos encontrados em (27), especialmente na seção que trata de mudanças de medida.

A.1 Definições

A.1.1 Espaços de probabilidade

Define-se um espaço de probabilidade como uma tripla $(\Omega, \mathfrak{F}, P)$, onde Ω se denomina conjunto universo. Seja $\mathfrak{F}$ um campo- σ (ou álgebra- σ) de subconjuntos de Ω , definido como uma coleção de subconjuntos de Ω tal que $\Omega \in \mathfrak{F}$, $\mathfrak{F}$ fechado com respeito às operações de união contável e complemento. Seja, salvo menção contrária, $\mathfrak{F} = \mathcal{B}(\Omega)$, onde $\mathcal{B}(\Omega)$ é a coleção de conjuntos de Borel de Ω . Seja P uma medida de probabilidade, definida como uma função de conjunto correspondente a uma medida de Lebesgue-Stieltjes, com $P(\Omega) = 1$.

Um conjunto $A \subset \Omega$ mensurável com respeito a P é denominado um evento.

Dado um espaço mensurável $(\Omega, \mathfrak{F})$, chama-se variável aleatória a um mapeamento $\mathbf{x} : (\Omega, \mathfrak{F}) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}^n))$ mensurável no sentido de Borel. Seja $P_{\mathbf{x}}$ a medida de probabilidade induzida por $\mathbf{x}$, dada por

$$P_{\mathbf{x}}(B) = P\omega : \mathbf{x}(\omega) \in B, B \in \mathcal{B}(\mathbb{R}^n).$$

Informalmente, dizer que uma variável aleatória $\mathbf{z}$ é mensurável em um campo- σ $\sigma(\mathbf{x})$

equivale a dizer que $\mathbf{z} = f(\mathbf{x})$ (na realidade, seria necessário impor que f é uma função mensurável no sentido de Borel).

Para uma variável aleatória $\mathbf{x}$, define-se sua distribuição de probabilidade cumulativa como a função $F = F_{\mathbf{x}} : \mathbb{R}^n \rightarrow [0, 1]$ dada por

$$F(\xi) = P\omega : x_i(\omega) \leq \xi_i \text{ para } i = 0, 1, \dots, n.$$

Ainda com relação à variável aleatória $\mathbf{x}$, define-se sua função densidade de probabilidade como uma função f , mensurável no sentido de Borel, tal que

$$P_{\mathbf{x}}(B) = \int_B f(x)dx \text{ para cada } B \in \mathcal{B}(\mathbb{R}^n).$$

sendo dx a medida de Lebesgue em $\mathbb{R}$. A função f tem características de uma derivada. Com efeito, mostra-se em A.1.4 que ela corresponde à derivada de Radon-Nikodym da medida P com respeito à medida de Lebesgue.

Seja, além disto, $\sigma(\mathbf{x})$ o campo- σ induzido pela variável aleatória $\mathbf{x} : (\Omega, \mathfrak{F}) \rightarrow (\Omega', \mathfrak{F}')$, dado por

$$\sigma(\mathbf{x}) = \mathbf{x}^{-1}(\mathfrak{F}')$$

onde $\mathbf{x}^{-1}$ é a preimagem de $\mathbf{x}(\mathfrak{F}')$. Em particular, para um vetor aleatório $\mathbf{x} \in \mathbb{R}^n$, $\sigma(\mathbf{x})$ consiste na coleção de todos os conjuntos $\mathbf{x} \in B$, com $B \in \mathcal{B}(\mathbb{R}^n)$.

Defina-se, por fim, $\mathcal{M}(\Omega)$ o espaço de Hilbert formado pelas medidas σ -finitas sobre Ω e $\mathcal{P}(\Omega)$ o subespaço de $\mathcal{M}(\Omega)$ formado pelas medidas de probabilidade sobre Ω (ou seja, $p \in \mathcal{M}(\Omega)$ tal que $p(\Omega) = 1$).

A.1.2 Processos estocásticos e cadeias de Markov

Em um espaço $(\Omega, \mathfrak{F}, P)$, um processo estocástico é uma família de variáveis aleatórias $(\mathbf{x}_t)_{t \in T}$, onde T é um conjunto indexador. Em tempo discreto, toma-se $T = t \in \mathbb{N} : t < t_f$, para algum t_f .

Um processo estocástico $(\mathbf{x}_t)_{t \in T}$ pode também ser considerado como uma função de $t \in T$ e $\omega \in \Omega$ e, portanto, como um mapeamento $\mathbf{x} : (\Omega \times \mathbb{N}, \mathfrak{F} \times \mathcal{B}(\mathbb{N})) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}^n))$, onde o operador $\times$ é o produto cartesiano entre conjuntos. No caso de campos, o operador em $\mathfrak{F} \times \mathfrak{G}$ resulta o menor campo- σ que contém $\mathfrak{F}$ e $\mathfrak{G}$. A consistência da medida de probabilidade P para qualquer subconjunto de T (contínuo ou discreto) é garantida pelo teorema da extensão de Kolmogorov.

Como uma extensão natural da definição apresentada para as variáveis aleatórias, define-se também o campo- σ induzido por um processo estocástico. Pode-se interpretar $\sigma\mathbf{x}(1), \dots, \mathbf{x}(n)$ como a totalidade da informação gerada pelo conhecimento de $\mathbf{x}(1), \dots, \mathbf{x}(n)$.

Uma cadeia de Markov é um processo estocástico dotado de um conjunto finito (ou contavelmente infinito) $S \subset \mathbb{N}$, chamado espaço de estados. Seja $\mathbf{P}$ uma matriz estocástica, isto é, $\mathbf{P} = [p_{ij}]$, $i, j \in S$ tal que $p_{ij} > 0$ para todo $i, j \in S$ e $\sum_j p_{ij} = 1$. A matriz estocástica $\mathbf{P}$ associada a uma cadeia de Markov é denominada matriz de transição.

Se $\theta(t)$ é uma cadeia de Markov em tempo discreto, considere-se uma distribuição inicial em que $P\theta(0) = i = p_i$, $i \in S$. O processo $(\theta(t))$ realizará transições em $t = 1, 2, 3, \dots$ segundo o seguinte princípio: dado que se conheça que $\theta(k) = i$, então independentemente dos valores assumidos por $\theta(t)$ antes de $t = k$, sabe-se que a probabilidade de que $\theta(k+1) = j$ vale p_{ij} . Portanto, para uma sequência de valores $i_0, i_1, \dots, i_n$ resulta que

$$P\theta(0) = i_0, \theta(1) = i_1, \dots, \theta(n) = i_n = p_{i_0 i_1} \cdots p_{i_{n-1} i_n}.$$

As cadeias de Markov têm propriedades importantes, que, no entanto, não serão enunciadas neste estudo. Em caso de dúvidas, estas podem ser facilmente encontradas em textos elementares de probabilidade.

A.1.3 Valor esperado condicional dado um campo- σ

É importante, para o restante do desenvolvimento, definir o conceito de esperança condicional. Se $\mathbf{y}$ é uma variável aleatória em $(\Omega, \mathfrak{F}, P)$ e $\mathbf{x} : (\Omega, \mathfrak{F}) \rightarrow (\Omega', \mathfrak{F}')$, o conceito de probabilidade condicional de $\mathbf{y}$ dado $\mathbf{x}$ é expresso como a função $E(\mathbf{y}|\mathbf{x}) : (\Omega', \mathfrak{F}') \rightarrow (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ que satisfaz a

$$\int_{\mathbf{x} \in A} \mathbf{y} dP = \int_A E(\mathbf{y}|\mathbf{x} = \bar{\mathbf{x}}) dP_{\mathbf{x}}(\bar{\mathbf{x}}) \text{ para cada } A \in (F)'.$$

com a propriedade de que todas as funções que satisfazem a esta relação são idênticas, exceto possivelmente em um conjunto de medida zero (com relação a $P_{\mathbf{x}}$, a medida de probabilidade induzida por $\mathbf{x}$).

É útil, além disto, definir o conceito mais geral de probabilidade condicional dado um campo- σ . Em $(\Omega, \mathfrak{F}, P)$ seja $(S) \subset \mathfrak{F}$ um campo- σ . A esperança matemática de $\mathbf{y}$ dado (S) será a função $E(\mathbf{y}|(S)) : (\Omega, \mathfrak{F}) \rightarrow (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ tal que

$$\int_C \mathbf{y} dP = \int_C E(\mathbf{y}|(S)) dP \text{ para cada } C \in (S). \quad (\text{A.1})$$

Note-se que o conceito de probabilidade condicional está estreitamente relacionado. Para um $\omega \in \Omega$, verifica-se que

$$P(\omega \in A) = E(I_A(\omega))$$

onde $I_A(\omega)$ é a função indicador do conjunto A . Evidentemente, esta relação vale da mesma maneira para as probabilidades condicionais.

A.1.4 Mudanças de medida e a derivada de Radon-Nikodym

Em $(\Omega, \mathfrak{F})$ sejam λ e μ duas medidas σ -contínuas, tais que λ seja absolutamente contínua com respeito a μ (ou seja, $\mu(A) > 0 \Leftrightarrow \lambda(A) > 0$ para todo $A \in \mathfrak{F}$). O teorema de Radon-Nikodym (vide (21, pag. 65)) garante que exista uma função mensurável no sentido de Borel tal que

$$\lambda(A) = \int_A \Lambda d\mu \text{ para todo } A \in \mathfrak{F}. \quad (\text{A.2})$$

Além disto, se houver outra função Θ que satisfaça a esta igualdade, $\Lambda = \Theta$ quase sempre.

Quando duas medidas de probabilidade são absolutamente contínuas reciprocamente, diz-se que são equivalentes. Isto significa que $\mu(A) > 0 \Leftrightarrow \lambda(A) > 0$ para todo $A \in \mathfrak{F}$.

Note-se que, com mais de uma medida definida em um espaço, é necessário explicitar, em muitos casos, com relação a qual medida se realiza uma operação. Em especial, define-se

$$E_P(\mathbf{x}) = \int_{\Omega} \mathbf{x} dP$$

onde se assume que P seja uma medida de probabilidade, como o operador esperança relacionado à medida P .

A função Λ é frequentemente representada como

$$\Lambda = \frac{d\lambda}{d\mu}$$

chamada derivada de Radon-Nikodym. Com efeito, esta função tem propriedades que justificam esta denominação. Em particular, é útil sua formulação como uma mudança de medida na integração.

Se h é uma função integrável no sentido de Lebesgue com relação à medida μ (assu-

minho, novamente, λ absolutamente contínua com respeito a μ),

$$\int_A h d\mu = \int_A h \left(\frac{d\mu}{d\lambda} \right) d\lambda \text{ para qualquer } A \in \mathfrak{F}.$$

Quando $\mu = \mu(x)$ e $d\lambda = dx$, para $x \in \mathbb{R}$ é a medida de Lebesgue sobre $\mathbb{R}$, $\mu'(x) = d\mu/dx$ é chamada função densidade de x com respeito à medida μ .

Ao encontrar mudanças de medida com propriedades desejáveis, é possível obter resultados importantes, como se verá mais adiante.

É útil, por fim, definir a derivada de Radon-Nikodym condicional a um campo- σ $\mathfrak{S} \subset \mathfrak{F}$. Utiliza-se a notação (ver (4))

$$\Lambda = \left(\frac{d\lambda}{d\mu} \right) |_{\mathfrak{S}}$$

para descrever este caso. A regra de mudança de medida no caso condicional torna-se

$$E_\lambda(\mathbf{x}|\mathfrak{S}) = \frac{E_\mu(\Lambda \mathbf{x}|\mathfrak{S})}{E_\mu(\Lambda|\mathfrak{S})}.$$

A.1.5 Kernels de transição e produtos projetivos

Esta seção segue o argumento de (26, p. 258).

Em um espaço mensurável $(\Omega, \mathfrak{F})$ define-se um átomo de Ω como o conjunto $A(\omega) \subset \Omega$ tal que

$$A(\omega) \triangleq \bigcap \{B : B \in \mathfrak{F}, \Omega \in B\}$$

ou seja, $A(\omega)$ é o menor subconjunto de Ω que contém ω .

Um kernel de transição em $(\Omega, \mathfrak{F})$ é uma função definida em $\mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$ tal que, para todo $t \in \mathbb{N}$ e todo $x \in \mathbb{R}^d$

$$K_t(\mathbf{x}, A) = P(\mathbf{x}(t+1) \in A | \mathbf{x}(t) = x)$$

onde A é um átomo de Ω . O kernel de transição tem, necessariamente, as seguintes propriedades:

- $K_t(\mathbf{x}, \cdot)$ é uma medida de probabilidade em $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ para todo $t \in \mathbb{N}$ e $x \in \mathbb{R}^d$;
- $K_t(\cdot, A) \in \mathcal{B}(\mathbb{R}^d)$ para todo $t \in \mathbb{N}$ e $A \in \mathcal{B}(\mathbb{R}^d)$.

Considere-se um espaço mensurável $(X, \mathfrak{F})$, uma sequência crescente de campos- σ

$\mathfrak{F}_0 \subset \mathfrak{F}_1 \subset \dots \subset \mathfrak{F}$, uma medida de probabilidade P_0 definida em $\mathfrak{F}_0$ e uma família de Kernels de transição K_n que formam uma medida de probabilidade em $(X, \mathfrak{F}_n)$. Estes Kernels de transição permitem que se defina indutivamente uma família de medidas de probabilidade em $(X, \mathfrak{F}_n)$ segundo a recorrência

$$P_n(A) \triangleq \int_X K_n(x, A) P_{n-1}(dx).$$

Se os kernels $K_n(x, \cdot)$ são consistentes (ou seja, se $B \in \mathcal{F}_{n+1}$ e $B \cap A = \emptyset$ então $K_{n+1}(x, B) = 0$), o teorema de Tulcea¹ garante que haja uma única medida de probabilidade P tal que $P|_{\mathcal{F}_n} = P_n$ para todo $n \in \mathbb{N}$.

Definindo

$$q_t(A) \triangleq P(\mathbf{x}(t) \in A)$$

segue que q_t satisfaz à seguinte recorrência:

$$q_{t+1} = K_t q_t \quad t > 0$$

onde $K_t q_t$ é a medida definida como

$$K_t q_t(A) \triangleq \int_{\mathbb{R}^d} K_t(\xi, A) q_t(d\xi)$$

de onde segue que

$$q_t = K_{t-1} \dots K_1 K_0 q_0, \quad t > 0.$$

Seja $\mathcal{P}(\mathbb{R}^d)$ o espaço de medidas de probabilidade definido na seção A.1.1. Seja $p \in \mathcal{P}(\mathbb{R}^d)$ uma medida de probabilidade e $\phi(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d$ uma função real mensurável no sentido de Borel tal que $p(\phi) > 0$. O produto projetivo $\phi * p$ é a função de conjunto $\phi * p : \mathcal{B}(\mathbb{R}^d) \rightarrow \mathbb{R}$ definida como

$$\phi * p(A) \triangleq \frac{\int_A \phi(x) p(dx)}{p(\phi)}, \quad \text{para } A \in \mathcal{B}(\mathbb{R}^d)$$

onde

$$p(\phi) = \int_{\mathbb{R}^d} \phi(x) p(dx). \quad (\text{A.3})$$

Uma propriedade importante deste produto é o fato de que sua derivada de Radon-Nikodym com respeito a p vale

$$\frac{d(\phi * p)}{dp} = \frac{\phi}{p(\phi)}.$$

¹(26, p. 299)

A.1.6 Ergodicidade

Defina-se um espaço de probabilidade $(\Omega, \mathfrak{F}, P)$. Seja $\mathbf{x}(k), \mathbf{x}(k+1), \dots, \mathbf{x}(k+n)$ uma sequência de variáveis aleatórias, parte de um processo estocástico $(\mathbf{x}(t))_{t \in \mathbb{N}}$. O processo $(\mathbf{x}(t))_{t \in \mathbb{N}}$ é dito estacionário (no sentido estrito) se, e somente se, para todo $k = 0, 1, 2, \dots$ valer a igualdade

$$P\{\omega : (\mathbf{x}_0(\omega), \dots, \mathbf{x}_n(\omega)) \in B\} = P\{\omega : (\mathbf{x}_k(\omega), \dots, \mathbf{x}_{k+n}(\omega)) \in B\}$$

para todo $B \in \mathfrak{F}$.

Seja $T : (\Omega, \mathfrak{F}) \rightarrow \Omega, \mathfrak{F}$ uma transformação em $(\Omega, \mathfrak{F}, \mu)$. Diz-se que T preserva a medida P (ou, simplesmente, preserva a medida) se, e somente se, $P(T^{-1}A) = P(A)$ para todo $A \in \mathfrak{F}$.

Em particular, considere-se $T = Z$, o operador deslocamento

$$Z(a_k, a_{k+1}, a_{k+2}, \dots) = (a_{k+1}, a_{k+2}, a_{k+3}, \dots).$$

Tem-se que a transformação $Z\mathbf{x}$ preserva P se, e somente se, $P_{\mathbf{x}}$ é estacionário (ver (21, pag. 347)).

O fato de a transformação Z preservar a medida P tem propriedades importantes enunciadas pelo Teorema Ergódico. Considere-se uma função $f \in L^1(\Omega, \mathfrak{F}, P)$, onde $L^1(\Omega, \mathfrak{F}, P)$ representa o espaço de Hilbert formado pelas funções complexas em $(\Omega, \mathfrak{F}, P)$ absolutamente integráveis no sentido de Lebesgue. O teorema garante que existe uma função $\hat{f} \in L^1$ tal que

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(Z^k \omega) = \int_{\Omega} f dP.$$

Isto significa, na prática, que no caso de sequências ergódicas (no sentido estrito), as estatísticas sobre um processo estocástico obtidas por uma amostra de conjunto de várias realizações (horizontal) serão idênticas às estatísticas obtidas mediante a observação de vários períodos (vertical) de uma única realização.

A.1.7 Filtragens

Dado um espaço de probabilidade $(\Omega, \mathfrak{F}, P)$, chama-se filtragem uma sequência crescente de campos- σ

$$\mathfrak{F}_k = \sigma(\mathbf{z}(0), \mathbf{z}(1), \dots, \mathbf{z}(k))$$

tais que, com respeito a um conjunto de índices I totalmente ordenado (que aqui se assume $I = \mathbb{N}$), valha a relação

$$t_1 \leq t_2 \Rightarrow \mathfrak{F}_{t_1} \subset \mathfrak{F}_{t_2}.$$

Em especial, pode-se definir com relação a um processo estocástico $(\mathbf{y}(t))_{t \in \mathbb{N}}$ a filtragem

$$(\mathcal{Y}_t)_{t \in I},$$

a qual é chamada de filtragem gerada por $(\mathbf{y}(t))_{t \in \mathbb{N}}$. Os elementos

$$\mathcal{Y}_t \triangleq \sigma\{\mathbf{y}(t), t \in I\},$$

representam a o campo- σ gerada pelas observações de $\mathbf{y}(t)$.

Seja, para uma medida de probabilidade P ,

$$\mathcal{N} \triangleq \{A : P(A) = 0\}, A \in \mathfrak{F}$$

a coleção de conjuntos nulos de P . Uma filtragem $(\mathfrak{F})_{t \in I}$ é chamada completa se, para cada t ,

$$\mathcal{N} \subset \mathfrak{F}_t.$$

Um processo estocástico $(\mathbf{z}(t))_{t \in I}$ é dito adaptado com relação a uma filtragem $(\mathfrak{F}_t)_{t \in I}$ se, e somente se, $\mathbf{z}(t)$ é mensurável com respeito a $\mathfrak{F}_t$ para todo t .

A filtragem gerada por um processo estocástico é chamada sua filtragem natural. Todo processo é mensurável com relação a sua filtragem natural.

A.2 Estimadores de máxima verossimilhança

A estimação de máxima verossimilhança é uma técnica estatística utilizada na identificação de sistemas que evoluem segundo formas funcionais conhecidas. Trata-se, com raras exceções, de uma estimação paramétrica, cujo objetivo é estimar constantes desconhecidas que determinam o comportamento do sistema.

Na seção A.3.1, desenvolver-se-á um processo mais geral de estimação, que se aplica tanto ao caso paramétrico quanto ao não-paramétrico - permite-se que a própria função de distribuição de probabilidade seja uma variável aleatória. Aqui, permanece-se no caso (prático) em que se supõe que estas funções pertencem a uma determinada família, diferindo entre si apenas em razão de um número finito de parâmetros desconhecidos.

Existe, como se verificará em breve, uma grande correspondência entre a estimação de parâmetros e a estimação de processos não observados. Pode-se tratar a estimação paramétrica como um caso particular do método geral de filtragem, bastando considerar os parâmetros como processos constantes. No entanto, o método de máxima verossimilhança tem algumas particularidades que justificam um tratamento separado. O principal motivo é o fato de este método ter sido desenvolvido por um raciocínio que não levava em conta nenhum critério objetivo de otimalidade (ver (28)), ao contrário do que invariavelmente ocorre em filtragem estocástica.

Nesta seção, seja $(\Omega, \mathfrak{F}, P)$ um espaço de probabilidade, $\mathbf{x} : (\Omega, \mathfrak{F}) \rightarrow (\mathbb{R}^p, \mathbf{B}(\mathbb{R}^p))$ uma variável aleatória (observável). Defina-se $P_{\mathbf{x}}(A|\eta)$, $A \in \mathcal{B}(\Omega)$, a medida de probabilidade induzida por $\mathbf{x}$, condicional a uma coleção de parâmetros η definida em um conjunto H . Seja, além disto, $p_{\mathbf{x}}(\xi|\eta) = dP_{\mathbf{x}}(\xi|\eta)/d\xi$ a função densidade correspondente à derivada de Radon-Nikodym de P_x em relação à medida de Lebesgue $d\xi$, $\xi \in \mathbf{B}(\mathbb{R}^p)$.

A.2.1 Função verossimilhança

Chama-se função verossimilhança um mapeamento $L : H \mapsto \mathbb{R}$ tal que

$$L(\eta|\mathbf{x} = \xi) = p_{\mathbf{x}}(\xi|\eta). \quad (\text{A.4})$$

Note-se que, na fórmula de L , η é considerada uma variável aleatória dependente da observação (fixa) de que $\mathbf{x} = \xi$ (vide (29)). Em geral, omite-se a dependência (evidente) de L com relação a $\mathbf{x}$, utilizando-se a representação

$$L(\eta) \triangleq L(\eta|\mathbf{x} = \xi).$$

Pode-se associar esta formulação a uma inversão das probabilidades condicionais utilizando a fórmula de Bayes, cuja formulação com a notação tradicional seria

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)},$$

embora Fisher (criador da versão moderna do método de verossimilhança) tenha rejeitado esta idéia por fazer alusão às distribuições incondicionais $P(A)$ e $P(B)$ que são desconhecidas na grande maioria dos casos. Nisto, contudo, consiste a essência da constatação de que a única grandeza relevante são as razões de verossimilhança,

$$\frac{L(\eta_2)}{L(\eta_1)},$$

e não os valores individuais. Ou seja, as funções de verossimilhança só fazem sentido quando utilizadas comparativamente.

É útil, além disto, definir a função log-verossimilhança,

$$\uparrow(\eta) \triangleq \log L(\eta),$$

uma vez que muitas das funções verossimilhança fazem parte da família exponencial.

A.2.2 Método de máxima verossimilhança e suas propriedades

O método de máxima verossimilhança consiste em utilizar como estimador para o vetor $\eta \in H$ o parâmetro $\hat{\eta}$ tal que

$$\hat{\eta} = \arg \max_{\bar{\eta}} L(\bar{\eta}),$$

que equivale a

$$\hat{\eta} = \arg \max_{\bar{\eta}} \uparrow(\bar{\eta}),$$

uma vez que o logaritmo é uma transformação monotônica.

Quando se trata de um conjunto de T observações de uma mesma distribuição $p_{\mathbf{x}}$ (ou equivalentemente: observações de um processo estocástico independente e identicamente distribuído $(\mathbf{x}(t))_{t=1}^T$) é evidente que a expressão (A.4) torna-se

$$L_T(\eta) = p_{\mathbf{x}}(\mathbf{x}(1), \dots, \mathbf{x}(T)|\eta),$$

definindo-se $L_T(\eta)$ como a verossimilhança associada à amostra de tamanho T . Devido à suposição de independência, segue que

$$L_T(\eta) = \prod_{t=1}^T p_{\mathbf{x}}(\mathbf{x}(t)|\eta).$$

ou, para a forma logarítmica,

$$\uparrow_T(\eta) = \sum_{t=1}^T \log p_{\mathbf{x}}(\mathbf{x}(t)|\eta).$$

Conforme mencionou-se anteriormente, estes estimadores têm características assintóticas bastante desejáveis. O estimador de máxima verossimilhança é:

- Assintoticamente não viesado (vide prova de consistência e de convergência em (30));

- Assintoticamente normal (ou seja, o erro de estimação assintótico segue uma distribuição normal cuja variância decresce com o número de amostras. (Vide (29));
- Eficiente - na maioria dos casos. A característica de eficiência de um estimador reside no fato de este atingir a fronteira inferior de variância de Cramér-Rao (dada a ausência de viés e a normalidade assintótica, a variância assintótica é a medida natural de eficiência de estimadores) Vide (29);
- Consistente até mesmo em modelos mal especificados (ver p.ex. (31) para os chamados métodos de quasi-máxima verossimilhança) e em condições com observações faltantes (ver (32))

Tendo realizado esta motivação, é necessário, em seguida, especificar as condições sob as quais o método de máxima verossimilhança é consistente, assintoticamente normal e eficiente.

Antes de iniciar a seção que trata das condições de convergência deste método, é importante detalhar melhor uma afirmação feita anteriormente. Mencionou-se que o método de máxima verossimilhança é muito frequentemente empregado sem uma definição explícita de critério (função objetivo) associado ao processo de estimação.

Demonstrou-se, porém, em (28), que a maximização realizada pela forma descrita corresponde ao critério

$$\arg \max E \left[\log f(\mathbf{x}|\hat{\eta}) \right] = \arg \max E \int f(\xi|\eta) \log f(\xi|\hat{\eta}) d\xi.$$

Por sua vez, esta maximização equivale à obtenção do ponto crítico em

$$\arg \max E \left[\log \frac{f(\mathbf{x}|\hat{\eta})}{f(\xi|\eta)} \right] = \arg \max E \int f(\xi|\eta) \log \left[\frac{f(\xi|\hat{\eta})}{f(\xi|\eta)} \right] d\xi.$$

A integral à direita corresponde ao valor esperado da divergência de Kullback-Leibler (vide, por exemplo, (33)) entre a distribuição real (com η) e a distribuição obtida com os parâmetros estimados.

Portanto, pode-se estabelecer que o método de máxima verossimilhança está balizado em um conceito claro da teoria da informação. Mais relações entre a teoria de estimação por máxima verossimilhança e a teoria da informação serão encontradas nas próximas seções.

A.2.3 Condições de consistência, normalidade assintótica e eficiência

Esta seção segue a análise de (29). Um estimador $\hat{\eta}$ é chamado estimador de extremo se existe uma função escalar $J_n(\eta)$ tal que

$$\hat{\eta} = \arg \max_{\bar{\eta}} J_n(\bar{\eta}), \text{ para } \bar{\eta} \in H.$$

Note-se que a função depende das observações $\mathbf{x}(1), \dots, \mathbf{x}(n)$, o que é representado apenas pelo índice n para simplificar a notação. Um estimador de extremo é chamado de M-estimador se a função objetivo J_n satisfizer a

$$J_n(\bar{\eta}) = \frac{1}{n} \sum_{t=1}^n m(\mathbf{x}(t); \bar{\eta}),$$

onde m é uma função real. Ou seja, um M-estimador é um estimador de extremo cuja função objetivo é uma média de conjunto.

O estimador de máxima verossimilhança é, portanto, um M-estimador com

$$m(\mathbf{x}(t); \bar{\eta}) = \log p(\mathbf{x}(t); \bar{\eta}).$$

A.2.3.1 Consistência

Proposição 1. *Seja $(\mathbf{x}(t))_{t \in \mathbb{N}}$ um processo ergódico e estacionário. Suponha-se que*

1. η_0 , o verdadeiro parâmetro, seja um elemento interior de um subespaço convexo $H \subset \mathbb{R}^p$,
2. $\log p(\mathbf{x}(t), \eta)$ é uma função côncava para todo $\eta \in H$,
3. $\log p(\mathbf{x}(t), \eta)$ é uma função mensurável de $\mathbf{x}$ para todo $\eta \in H$,
4. (identificabilidade) $\text{Prob}[p(\mathbf{x}(t), \eta) \neq p(\mathbf{x}(t), \eta)] > 0$
5. $E[|\log p(\mathbf{x}(t), \eta)|]$ existe e é finito para todo $\eta \in H$ (ou seja, $E[|\log p(\mathbf{x}(t), \eta)|] < \infty$).

Se (1-5) se verificam, segue que, para $t \rightarrow \infty$ a estimativa $\hat{\eta}$ existe com probabilidade 1, e $\hat{\eta} \rightarrow \eta$ em probabilidade.²

²Para uma demonstração, vide (29)

As condições (1-5) correspondem, portanto, a um conjunto de condições suficientes para a consistência do estimador de máxima verossimilhança. Note-se que existem resultados que assumem condições menos fortes: é possível relaxar as condições para exigir apenas que o processo $(\mathbf{x}(t))_{t \in \mathbb{N}}$ seja separável; pode-se exigir que $\log p(\cdot)$ seja semicontínua (uma restrição consideravelmente mais branda do que côncava, que já implica continuidade). Além disto, é possível demonstrar a convergência no caso em que $E \log p(\cdot) \rightarrow \infty$, desde que $\log p(\cdot)$ seja ajustável.

Não obstante, utilizar-se á o conjunto de condições (1-5), pois este será suficiente para o caso em questão. A facilidade de verificação de suas hipóteses constitui o principal motivo da escolha deste particular conjunto. Ressalte-se, porém, que não é este o conjunto de postulados menos restritivo possível que demonstra com suficiência que o estimador é consistente.

A.2.3.2 Normalidade assintótica

Dado que se satisfaçam as condições de (1), resulta a convergência em probabilidade do estimador $\hat{\eta}$ para o valor correto, η . Resta, no entanto, determinar de que modo se pode esperar que os erros de estimação $\hat{\eta} - \eta$ estarão distribuídos, quando $t \rightarrow \infty$.

Em condições adequadas, deduz-se que a distribuição assintótica é uma normal. Isto é desejável por ser esta a distribuição de máxima entropia, ou seja, de maior imprevisibilidade. O fato de o erro ser imprevisível é um indicador de que o estimador "extraiu" o máximo possível de informação sobre a quantidade a ser estimada.

Proposição 2. *A normalidade assintótica é garantida quando, além das condições em (1), valerem:*

1. *A função $p(\mathbf{x}(t), \eta)$ é contínua, e tem primeira e segunda derivadas contínuas;*
2. *$E [\nabla_{\eta} \log p(\mathbf{x}(t), \eta)] = 0$ e*

$$-E [H(\log p(\mathbf{x}(t), \eta))] = E [[\nabla_{\eta} \log p(\mathbf{x}(t), \eta)] \cdot [\nabla_{\eta} \log p(\mathbf{x}(t), \eta)]'] = 0,$$

onde $H(\cdot)$ representa a matriz Hessiana;

3. *$E \sup_{\eta \in B} \|H(\log p(\mathbf{x}(t), \eta))\| < \infty$, para uma vizinhança B de η ;*
4. *$E [H(\log p(\mathbf{x}(t), \eta))]$ não singular.*

Caso se verifiquem as condições (1-4), demonstra-se (vide (29)) que o estimador de máxima verossimilhança é assintoticamente normal.

Resta, por fim, determinar qual será a variância deste erro normalmente distribuído. É necessário, antes de apresentar os resultados pertinentes a este tema, introduzir o conceito de fronteira inferior de Cramér-Rao, que determina a menor variância possível de ser atingida por um estimador.

A.2.3.3 Fronteira inferior de Cramér-Rao

Esta seção baseia-se em (33)[p.393-399]. Seja $\hat{\eta} \in H$ um estimador de η , e seja $\hat{\eta} - \eta$ o erro associado à estimação.

Defina-se a matriz de informação de Fisher como

$$\mathcal{I}(\hat{\eta}) \triangleq E_{\hat{\eta}} [H_{\hat{\eta}}(\log p(\mathbf{x}(t)\hat{\eta}))]$$

onde $H_{\eta}(\cdot)$ simboliza a matriz hessiana de p , tomada em função de η .

O teorema de Cramér-Rao garante que

$$E(\hat{\eta} - \eta)^2 = \text{var}(\hat{\eta}) \geq \hat{\eta}(\hat{\eta})^{-1}$$

para estimadores $\hat{\eta}$ não viesados (ou seja, quando o erro esperado é nulo). Isto valerá para o caso em questão, uma vez verificadas as condições que tornam o estimador de máxima verossimilhança assintoticamente consistente.

A.2.3.4 Eficiência

Define-se como estimador eficiente aquele que atinge a fronteira inferior de Cramér-Rao. Dado que chegou-se a uma distribuição assintoticamente normal, é bastante evidente que o este é o critério natural de comparação entre estimadores.

Segundo (29), as condições apresentadas em (2) já implicam a eficiência do estimador. Pode-se arguir que, como tal, elas são demasiado fortes para verificar a normalidade assintótica. No entanto, esta característica será também bastante útil neste contexto, uma vez que se buscam apenas condições suficientes que possam ser verificadas com simplicidade.

Portanto, dado que valham as condições de (1) e (2), segue que

$$\hat{\eta} \sim N(0, \mathcal{I}(\hat{\eta})),$$

para $t \rightarrow \infty$, sendo $\hat{\eta}$ o valor crítico da função verossimilhança.

A.3 Filtro bayesiano não-linear

A.3.1 Metodologia geral

Sejam³ $(\mathbf{x}(t))_{t \in \mathbb{N}}$, $(\mathbf{w}(t))_{t \in \mathbb{N}}$ e $(\mathbf{y}(t))_{t \in \mathbb{N}}$ processos estocásticos. Chame-se $\mathbf{x}$ sinal; $\mathbf{w}$, ruído; $\mathbf{y}$, processo observado, com $\mathbf{x}(t) : \Omega \rightarrow \mathbb{R}^m$, $\mathbf{y}(t) : \Omega \rightarrow \mathbb{R}^d$ e $\mathbf{w}(t) : \Omega \rightarrow \mathbb{R}^m$. Suponha-se que o processo observado $(\mathbf{y}(t))_{t \in \mathbb{N}}$ seja um processo estocástico com valores em $\mathbb{R}^m$ da forma

$$\mathbf{y}(t) \triangleq h(t, \mathbf{x}(t)) + \mathbf{w}(t)$$

supondo, sem perda de generalidade, $\mathbf{y}(0) = 0$. A função $h : \mathbb{N} \times \mathbb{R}^d \rightarrow \mathbb{R}^m$ é uma função mensurável no sentido de Borel e os vetores $\mathbf{w}(t)$ são independentes entre si e distribuídos conforme uma função densidade absolutamente contínua com relação à medida de Lebesgue λ . Denote-se por $g(t, \cdot)$ a densidade (estritamente positiva e limitada) de $\mathbf{w}(t)$ com respeito a λ .

O problema de filtragem consiste em computar a densidade condicional do sinal $\mathbf{x}(t)$ dado o campo- σ gerado pelas observações do processo conhecido. Esta medida de probabilidade é uma variável aleatória em $\mathcal{P}(\mathbb{R}^m)$, o espaço de medidas definido em (A.1.1). Ou seja, é uma variável aleatória $\pi_t(A)$ da forma

$$\pi_t(A) \triangleq P(\mathbf{x}(t) \in A | \sigma(\mathbf{y}_{0:t})), \text{ para todo } A \in \mathcal{B}(\mathbb{R}^d)$$

onde, para $s, t \in \mathbb{N}$ com $s \leq t$,

$$\mathbf{y}_{s:t} \triangleq (\mathbf{y}(s), \dots, \mathbf{y}(t))$$

é uma sequência de vetores aleatórios. Adicionalmente, define-se, em associação a um valor arbitrário $\mathbf{y}_{0:t}^* \triangleq (\mathbf{y}^*(s), \dots, \mathbf{y}^*(t))' \in (\mathbb{R}^m)^{t+1}$, a medida (não aleatória)

$$\pi_t^{\mathbf{y}_{0:t}^*}(A) \triangleq P(\mathbf{x}(t) \in A | \mathbf{y}_{0:t} = \mathbf{y}_{0:t}^*).$$

Definam-se também as medidas p_t e $p_t^{\mathbf{y}_{0:t-1}^*}$, as medidas de probabilidade previstas de $\mathbf{x}(t)$,

³Esta seção segue (26, p.258-259,264)

definidas como

$$p_t(A) \triangleq P(\mathbf{x}(t) \in A | \sigma(\mathbf{y}_{0:t-1})).$$

e

$$p_t^{\mathbf{y}_{0:t-1}^*}(A) \triangleq P(\mathbf{x}(t) \in A | \mathbf{y}_{0:t-1} = \mathbf{y}_{0:t-1}^*),$$

a medida aleatória e a determinística, respectivamente.

A distribuição p é chamada distribuição a priori, enquanto π é chamada distribuição (Bayesiana) a posteriori.

É necessário encontrar, inicialmente, a distribuição de $\mathbf{y}_{s:t}$. Sendo λ a medida de Lebesgue em $((\mathbb{R}^m)^{t-s+1}, \mathcal{B}((\mathbb{R}^m)^{t-s+1}))$, a medida $P_{\mathbf{y}_{s:t}}$ será absolutamente contínua com relação a λ para todo $0 < s \leq t < \infty$, e sua derivada de Radon-Nikodym valerá

$$\frac{dP_{\mathbf{y}_{s:t}}}{d\lambda}(y_{s:t}) = \Upsilon(y_{s:t}) \triangleq \int_{(\mathbb{R}^d)^{t-s+1}} \prod_{i=s}^t g_i(y(i) - h(i, x(i))) P_{\mathbf{x}_{s:t}}(dx_{s:t})$$

onde $P_{\mathbf{x}_{s:t}} \in \mathcal{P}((\mathbb{R}^d)^{t-s+1})$ é a distribuição de probabilidade da sequencia de vetores aleatórios $\mathbf{x}_{s:t}$. Isto decorre do fato de que

$$P(\mathbf{y}_{s:t} \in C_{s:t} | \mathbf{x}_{s:t} = \mathbf{x}_{s:t}^*) = \prod_{i=s}^t \int_{C_i} g_i(\mathbf{y}^*(i) - h(i, \mathbf{x}(i))) d\mathbf{y}^*(i).$$

A relação expressa em (A.3.1) demonstra o motivo de a função

$$g_t^{\mathbf{y}^*(t)} \triangleq g_t(\mathbf{y}^*(t) - h(t, \mathbf{x}(t)))$$

ser conhecida, em inglês, como *likelihood function*, ou função probabilidade.

Obtem-se ((26, p.261-262)) que, para um determinado caminho $(\mathbf{y}(0), \mathbf{y}(1), \dots, \mathbf{y}(t), \dots)$ a sequencia de medidas de probabilidade (determinísticas) $(\pi_t^{\mathbf{y}^*})_{t \geq 0}$ satisfaz à recursão

$$\pi_t^{\mathbf{y}_{0:t}^*} = g_t^{\mathbf{y}^*(t)} * K_{t-1} \pi_{t-1}^{\mathbf{y}_{0:t-1}^*}$$

e

$$p_t^{\mathbf{y}_{0:t-1}^*} = K_{t-1} \pi_{t-1}^{\mathbf{y}_{0:t-1}^*}$$

onde $*$ representa o produto projetivo apresentado em (A.1.5).

Combinando-se ambas as relações, segue que se pode obter um mecanismo geral de estimação das distribuições a priori e a posteriori da forma

$$\pi_{t-1}^{\mathbf{y}_{0:t-1}^*} \mapsto p_t^{\mathbf{y}_{0:t-1}^*} = K_{t-1} \pi_{t-1}^{\mathbf{y}_{0:t-1}^*} \mapsto \pi_t^{\mathbf{y}_{0:t}^*} = g_t^{\mathbf{y}^*(t)} * p_t^{\mathbf{y}_{0:t-1}^*}.$$

Embora este resultado pareça bastante simples e elegante, sua implementação encontra dificuldades substanciais, excetuando-se casos excepcionais, cujo tratamento é possível em forma fechada (filtro linear gaussiano). O principal obstáculo reside no cálculo do produto projetivo, que é uma transformação não-linear que envolve integração em um espaço possivelmente multidimensional. Na próxima seção, apresentar-se-á a forma explícita do filtro Bayesiano, o que tornará mais claro este argumento.

A.3.2 Forma explícita do filtro bayesiano não-linear

Suponha-se⁴, como na seção precedente, que os processos estocásticos $(\mathbf{x}(t))_{t \in \mathbb{N}}$, $(\mathbf{w}(t))_{t \in \mathbb{N}}$ e $(\mathbf{y}(t))_{t \in \mathbb{N}}$ estejam definidos em um espaço de probabilidades $(\Omega, \mathfrak{F}, P)$. Seja $(\mathbf{y}(t))_{t \in \mathbb{N}}$ o processo observável determinado pela fórmula (A.3.1). Considere-se, ainda, que o processo $(\mathbf{x}(t))_{t \in \mathbb{N}}$, determinado anteriormente em função de seus Kernels de transição, possa ser expresso como funcional explícito da forma

$$\mathbf{x}(t+1) = f(t, \mathbf{x}(t)) + \mathbf{v}(t)$$

onde $(\mathbf{v}(t))_{t \in \mathbb{N}}$ é um processo estocástico definido em $(\Omega, \mathfrak{F}, P)$, que se assume independente e identicamente distribuído, e independente de $\mathbf{w}(t)$ para todo $t \in \mathbb{N}$.

O objetivo do filtro bayesiano é encontrar a distribuição

$$\pi_t^{\mathbf{y}_{0:t}^*}.$$

Pode-se obtê-la partindo da distribuição a priori

$$p_t^{\mathbf{y}_{0:t-1}^*}(A) = \int_A P(\mathbf{x}(t) \in A | \mathbf{x}(t-1) = \xi) \pi_{t-1}^{\mathbf{y}_{0:t-1}^*}(d\xi)$$

e, lembrando-se de que

$$\pi_t^{\mathbf{y}_{0:t}^*} = g_t^{\mathbf{y}^*(t)} * p_t^{\mathbf{y}_{0:t-1}^*}$$

segue que

$$\pi_t^{\mathbf{y}_{0:t}^*}(A) = \frac{\int_A g_t^{\mathbf{y}^*(t)} p_t^{\mathbf{y}_{0:t-1}^*}(\xi) d\xi}{\int_{\mathbb{R}} g_t^{\mathbf{y}^*(t)} p_t^{\mathbf{y}_{0:t-1}^*}(\xi) d\xi}.$$

Tomando-se o conjunto A como o átomo que contém o ponto $\xi = \bar{x}$, pode-se utilizar a propriedade (A.1.5) para expressar a igualdade (A.3.2) em termos de sua função densidade.

Ou seja,

$$\frac{d(\phi * p)}{d\lambda} = \frac{\phi dp}{d\lambda} \frac{1}{p(\phi)}$$

⁴Esta seção segue (22, p.463-466)

onde λ é a medida de Lebesgue. Convencionando-se $q(\cdot) \triangleq dP(\cdot)/d\lambda$, e aplicando a propriedade (A.3.2) a (A.3.2), obtêm-se as expressões mais usualmente encontrada na literatura para o filtro bayesiano:

$$q(\mathbf{x}(t)|\sigma(\mathbf{y}_{0:t})) = \frac{q(\mathbf{y}(t)|\mathbf{x}(t))q(\mathbf{x}(t)|\sigma(\mathbf{y}_{0:t-1}))}{\int q(\mathbf{y}(t)|\mathbf{x}(t) = \xi)q(\mathbf{x}(t) = \xi|\sigma(\mathbf{y}_{0:t-1})) d\xi}$$

que determina a distribuição a posteriori, e

$$q(\mathbf{x}(t)|\sigma(\mathbf{y}_{0:t-1})) = \int q(\mathbf{x}(t)|\mathbf{x}(t-1) = \xi)q(\mathbf{x}(t-1) = \xi|\sigma(\mathbf{y}_{0:t-1})) d\xi.$$

para a distribuição a priori.

Uma solução analítica para estas expressões está disponível em muito poucos casos. Em particular, para o caso linear-gaussiano, esta recursão se reduz ao familiar filtro de Kalman. No caso geral, o problema requer uma solução aproximada para que haja tratabilidade numérica.

Referências

- 1 BAR-SHALOM, Y.; LI, X. *Estimation and Tracking: Principles, Techniques and Software*. Norwood: Artech House, 1993.
- 2 COSTA, O.; FRAGOSO, M. D.; MARQUES, R. P. *Discrete-time Markov Jump Linear Systems*. London: Springer Verlag, 2005.
- 3 HAMILTON, J. D. *Time series analysis*. Princeton: Princeton University Press, 1994.
- 4 ELLIOTT, R. J.; AGGOUN, L.; MOORE, J. B. *Hidden Markov models: estimation and control*. New York: Springer, 1995.
- 5 WILLIAMS, J. L. *Gaussian mixture reduction for tracking multiple maneuvering targets in clutter*. Ohio: Air Force Institute of Technology, 2003.
- 6 TAFAZOLI, S.; SUN, X. Hybrid system state tracking and fault detection using particle filters. *IEEE Transactions on Control Systems Technology*, v. 14, n. 6, p. 1078–1087, November 2006.
- 7 HWANG, I. et al. A survey of fault detection, isolation, and reconfiguration methods. *IEEE Transactions on Control Systems Technology*, v. 18, n. 3, p. 636–653, May 2010.
- 8 HAMILTON, J. D. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, v. 57, n. 2, p. 357–384, Março 1989.
- 9 VAL, J. B. R. do; BASAR, T. Receding horizon control of jump linear systems and a macroeconomic policy problem. *Journal of Economic Dynamics and Control*, v. 23, n. 8, p. 1099 – 1131, 1999. ISSN 0165-1889.
- 10 TUGNAIT, J. Adaptive estimation and identification for discrete systems with Markov jump parameters. *Automatic Control, IEEE Transactions on*, v. 27, n. 5, p. 1054–1065, Oct 1982. ISSN 0018-9286.
- 11 HAMILTON, J. D. Analysis of time series subject to changes in regime. *Journal of Econometrics*, v. 45, p. 39–70, 1990.
- 12 BACCARELLI, E.; CUSANI, R. Recursive kalman-type optimal estimation and detection of hidden Markov chains. *Signal Processing*, v. 51, n. 1, p. 55–64, 1996.
- 13 ELLIOTT, R.; KRISHNAMURTHY, V. New finite-dimensional filters for parameter estimation of discrete-time linear gaussian models. , *IEEE Transactions on Automatic Control*, v. 44, n. 5, p. 938–951, Maio 1999. ISSN 0018-9286.
- 14 BLOM, H.; BAR-SHALOM, Y. The interacting multiple model algorithm for systems with markovian switching coefficients. *IEEE Transactions on Automatic Control*, v. 33, n. 8, p. 780–783, Aug 1988. ISSN 0018-9286.

- 15 COSTA, O. Linear minimum mean square error estimation for discrete-time markovian jump linear systems. *IEEE Transactions on Automatic Control*, v. 39, n. 8, p. 1685–1689, Aug 1994. ISSN 0018-9286.
- 16 GERMANI, A.; MANES, C.; PALUMBO, P. State estimation of stochastic systems with switching measurements: A polynomial approach. *International Journal of Robust and Nonlinear Control*, v. 19, p. 1632–1655, 2009.
- 17 DOUCET, A.; GORDON, N. J.; KRISHNAMURTHY, V. Particle filters for state estimation of jump Markov linear systems. *IEEE Transactions on Signal Processing*, v. 49, p. 613–624, 1999.
- 18 DOUCET, A.; RISTIC, B. Recursive state estimation for multiple switching models with unknown transition probabilities. *IEEE Transactions on Aerospace and Electronic Systems*, v. 38, n. 3, p. 1098–1104, July 2002.
- 19 JILKOV, V. P.; LI X., R. Online bayesian estimation of transition probabilities for markovian jump systems. *IEEE Transactions on Signal Processing*, v. 52, n. 6, p. 1620–1630, June 2004.
- 20 ORGUNER, U.; DEMIREKLER, M. Maximum likelihood estimation of transition probabilities of jump markov linear systems. *IEEE Transactions on Signal Processing*, v. 56, n. 10, p. 5093 – 5108, October 2008.
- 21 ASH, R. B.; DOLÉANS-DADE, C. *Probability and measure theory*. San Diego, CA, USA: Academic Press, 2000.
- 22 SIMON, D. *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*. Hoboken, NJ, USA: Wiley-Interscience, 2006. ISBN 0471708585.
- 23 HERSHEY, J. R.; OLSEN, P. A. Approximating the kullback leibler divergence between gaussian mixture models. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. [S.l.: s.n.], 2007.
- 24 COSTA, O.; VAL, J. do; GEROMEL, J. Continuous-time state-feedback h2-control of markovian jump linear systems via convex analysis. *Automatica*, v. 35, p. 259–268, February 1999.
- 25 BAIN, A. Stochastic calculus. Disponível em: <<http://www.chiark.greenend.org.uk/~alanb/stoc-calc.pdf>>.
- 26 BAIN, A.; CRISAN, D. *Fundamentals of stochastic filtering*. New York, NY, USA: Springer, 2008.
- 27 KLEBANER, F. C. *Introduction to stochastic calculus with applications*. London, UK: Imperial College Press, 2005.
- 28 AKAIKE, H. *Information theory and an extension of the maximum likelihood princile*. New York: Springer Verlag, 1992. 610–624 p.
- 29 HAYASHI, F. *Econometrics*. Princeton, NJ, USA: Princeton University Press, 2000.

- 30 WALD, A. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 20, n. 4, p. 595–601, December 1949.
- 31 GOURIEROUX, C.; MONFORT, A.; TROGNON, A. Pseudo maximum likelihood methods: Theory. *Econometrica*, The Econometric Society, v. 52, n. 3, p. 681–700, 1984.
- 32 DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, Blackwell Publishing for the Royal Statistical Society, v. 39, n. 1, p. 1–38, 1977.
- 33 COVER, T. M.; THOMAS, J. A. *Elements of information theory*. New York, NY, USA: Wiley-Interscience, 1991. ISBN 0-471-06259-6.