Modelos Oultos de Markov

Introdução

Parte I. Fundamentos

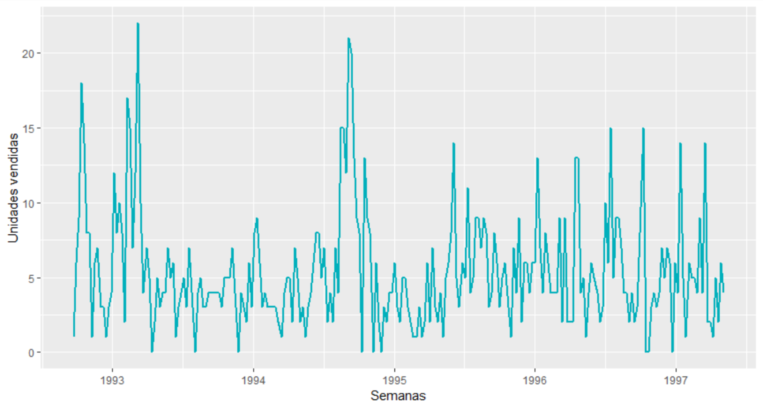
Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Agosto, 2020

Os Modelos Ocultos de Markov (HMMs) são uma classe de modelos em que a distribuição que gera uma observação depende do estado de uma Cadeia de Markov subjacente e não observada. Eles se mostram promissores como modelos genéricos de propósito geral para séries temporais univariadas e multivariadas, especialmente para séries de valor discreto, incluindo séries categóricas e séries de contagens (Zucchini MacDonald, 1998).

Considere, por exemplo, uma série de vendas semanais de um produto específico de sabão em um supermercado. As unidades semanais inteiras de um sabonete de cóodigo 3700031165 foram fornecidos pelo Kilts Center for Marketing da Escola de Pós-Graduação em Administração da Universidade de Chicago. O produto é Zest White Water 15 oz. Uma onça, abreviada oz, é uma unidade de medida de massa, uma onça equivale a 28.349523125 gramas. Os dados são mostrados na figura a seguir.



Série de vendas semanais de um produto de sabão específico. Dados fornecidos pelo Grupo de Marketing da Universidade de Chicago.

Nesse caso a aplicação de modelos de séries temporais padrão, como os modelos ARMA, é restrita pois eles são baseados na distribuição normal. Em vez disso, o modelo básico para contagens ilimitadas é a distribuição Poisson. No entanto, o modelo Poisson padrão não é apropriado neste caso uma vez que, como será demonstrado mais tarde, a série apresenta considerável superdispersão em relação à distribuição Poisson e forte dependência serial positiva. Além disso, parecem existir alguns períodos com baixa taxa de vendas semanais e outros períodos com uma taxa relativamente alta de vendas semanais.

A classe de modelos de séries temporais Ocultas de Markov, que modelam a distribuição de probabilidade S_t na dependência do estado não observado, ou seja, oculto C_t de uma Cadeia de Markov com m estados e que pode acomodar tanto a superdispersão quanto a dependência serial, parece ser uma ferramenta útil para modelar esta série e tentando entender sua estrutura. O ajuste de um Modelo Oculto de Markov Poisson à série de vendas semanais de sabão constituirá parte integrante desta nota, ou seja, a maioria dos aspectos dos HMMs introduzidos aqui será demonstrada por meio desta série.

Os HMMs têm sido utilizados há mais de duas décadas em aplicações de processamento de sinais, especialmente no contexto do reconhecimento automático de voz, mas o interesse na teoria e nas aplicações de HMMs está se expandindo rapidamente para outros campos, por exemplo:

- todos os tipos de reconhecimento: rostos, fala, gesto, caligrafia e/ou assinatura,
- bioinformática: análise de sequências biológicas, ambiente: direção do vento, chuvas, terremotos,
- finanças: série de retornos diários.

A bibliografia aqui apresentada lista vários artigos e monografias que lidam com a aplicação dos HMMs nesses campos e podem ser de interesse para leitura adicional: Durbin et al. (1998), Elliott, Aggoun e Moore (1995), Koski (2001), Rabiner (1989) e Ephraim e Merhav (2002).

Entre as características atrativas dos HMMs estão sua versatilidade, sua facilidade matemática e o fato de que a probabilidade é relativamente direta (Zucchini e MacDonald, 2001).

Em detalhe, os HMMs são caracterizados pelas seguintes propriedades:

- todos os momentos disponíveis: média, variância, autocorrelações,
- probabilidade de fácil cálculo: cálculo linear no tempo,
- distribuições marginais disponíveis: observações faltantes sem problemas,
- distribuições condicionais disponíveis: identificação de outliers; previsão de k-passos adiante, distribuição conjunta de várias previsões.

Além disso, os HMMs são interpretáveis em muitos casos e podem facilmente acomodar covariáveis adicionais. Além disso, eles são moderadamente parcimoniosos, ou seja, um modelo simples de dois estados geralmente fornece um ajuste razoável.

Os principais objetivos ao lidar com Modelos Ocultos de Markov são os seguintes:

- revelam a estrutura dos dados, ou seja, tendência, variação sazonal e dependência serial,
- prever valores futuros, incluindo intervalos de previsão,
- identificar valores incomuns,
- relacionar as observações a outras séries, ou seja, covariáveis.

Esta nota basicamente pretende dar uma introdução simples ao Modelo Oculto de Markov (HMM). É simples no sentido de que é restrito a séries temporais estacionárias, ou seja, sem tendência ou variação sazonal. As observações podem ser de valor discreto ou contínuo, mas neste curso vamos supor que elas sejam univariadas e iremos ignorar qualquer informação que possa estar disponível nas covariáveis. Apenas no final desta nota, daremos uma breve visão geral das possíveis extensões do Modelo Oculto de Markov.

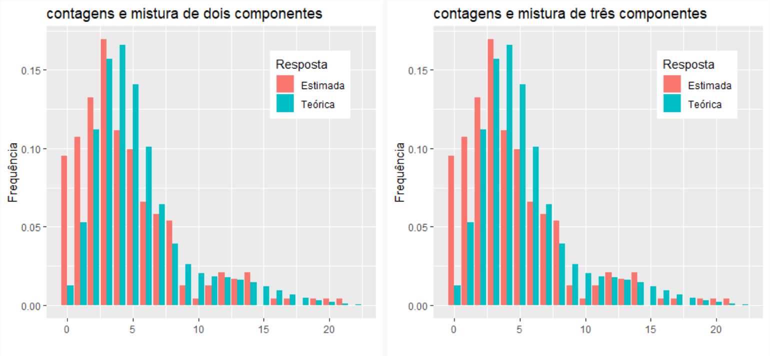
A ênfase estará na aplicação dos modelos, em particular na especificação de modelos, na estimação de parâmetros, na seleção de modelos, na verificação de diagnósticos e na previsão.

Como suporte computacional utilizamos a linguagem de programação e ambiente de desenvolvimento integrado para cálculos estatísticos e gráficos R, última versão 3.5.2, Eggshell Igloo de 20 de dezembro de 2018.

Temos por objetivo aqui apresentar uma breve introdução a dois conceitos fundamentais que são necessários para compreender a estrutura básica dos Modelos Ocultos de Markov (HMM).

Reconsiderar a série de vendas semanais de sabonetes introduzida na Introdução e vamos assumir que não há correlação serial na série, ou seja, as vendas semanais são contagens independentes. O modelo básico para contagens independentes não vinculadas é a distribuição de Poisson com função de probabilidade $P(S) = \lambda^s e^{-\lambda}/s!$ e a propriedade restritiva que a variância iguala a esperança $\mathrm{E}(S) = \mathrm{Var}(S)$.

Para a série de vendas de sabão, tem-se $\overline{S} = 5.442149$ e $S^2 = 15.4012$, indicando uma forte dispersão em relação à distribuição de Poisson e, portanto, inapropriação desta última. Isto é apresentado na figura a seguir que apresenta um histograma das observações e a distribuição de Poisson teórica. Além disso, a distribuição das vendas semanais de sabonetes é bimodal, enquanto a distribuição de Poisson tem apenas uma moda.



Série de vendas semanais de um produto de sabão específico e a distribuição Poisson estimada.

Um método de superar as deficiências da distribuição Poisson em tal caso é usar um modelo de mistura independente. As distribuições de misturas são muito úteis no tratamento de observações ou dados superdispersos com distribuições bimodais ou multimodais mais gerais. Os modelos de mistura assumem que a superdispersão e/ou multimodalidade das observações de uma variável podem ser devidas à heterogeneidade não observada na população, ou seja, a população consiste em grupos diferentes com distribuições diferentes para essa variável. Imagine, por exemplo, a distribuição do número de pacotes de cigarros comprados por um número de clientes em um supermercado. Os clientes podem ser separados em vários grupos: não-fumantes, pessoas que fumam ocasionalmente, etc. Isso leva à superdispersão em relação à Poisson e talvez até a uma distribuição multimodal de suas compras.

No caso de duas distribuições componentes, a distribuição de mistura é caracterizada pelas duas variáveis aleatórias X_1 e X_2 e suas funções de probabilidade ou funções de densidade, respectivamente:

- Variável aleatória X_1 X_2
- função de probabilidade p_1 p_2
- função de densidade f_1 f_2

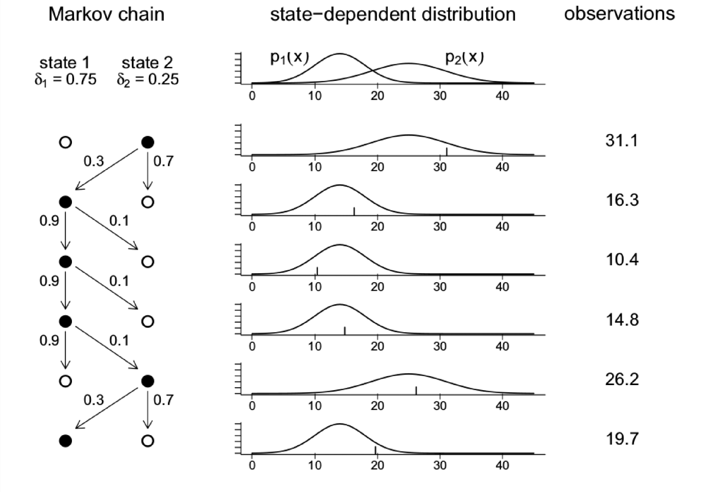
Além disso, para o processo de parâmetros, é necessária uma variável aleatória discreta C para realizar a mistura:

$$C = \begin{cases} 1, & \text{com probabilidade } \delta_1 \\ 2, & \text{com probabilidade } \delta_2 = 1 - \delta_1 \end{cases} . \tag{1}$$

Pode-se imaginar C como o resultado de jogar uma moeda: Se C assume o valor 1, obtenha uma observação de X_1. De acordo com isso, se C assumir o valor 2, obtenha uma observação de X_2. A estrutura desse processo para o caso de duas distribuições de componentes contínuos é mostrada na figura seguinte.

Em situações práticas, não sabemos para que lado a moeda pousou. Apenas as observações geradas por X_1 ou X_2 podem ser observadas e, na maioria dos casos, elas não podem ser atribuídas a variáveis aleatória distintas.

Dada a probabilidade cada componente e as respectivas distribuições de probabilidade, a função de probabilidade ou de densidade da mistura pode ser encontrada.



Estrutura do processo de uma distribuição de mistura de dois componentes.
Figura retirada do livro Zucchini MacDonald (2009).

A extensão para o caso de m componentes é simples. Sejam $\delta_1, \delta 2, \cdots, \delta_m$ os pesos atribuídos aos diferentes componentes, $p_1, p_{2,m}$ ou $f_1, f_2, \cdots, f_m$ denotando suas funções de probabilidade ou funções de densidade, respectivamente. Então, a distribuição da mistura dada pela variável aleatória X pode ser facilmente calculada por uma combinação linear das distribuições componentes:

$$\begin{array}{rcll} p(x) & = & \delta_1 p_1(x) + \delta_2 p_2(x) & \qquad \text{(caso discreto)}, \\ f(x) & = & \delta_1 f_1(x) + \delta_2 f_2(x) & \qquad \text{(caso contínuo)}\cdot \end{array}$$

Mais geral, o k-ésimo momento $\mathrm{E}(X^k)$ de uma mistura é simplesmente uma combinação linear dos respectivos momentos de seus componentes:

$$\mathrm{E}(X^k) = \sum_{i=1}^{m} \delta_i \mathrm{E}(X_i^k), \qquad k = 1, 2, \cdots \cdot$$

Agora, observe que isto não é válido para os momentos centrais, por exemplo, a variância de uma mistura:

$$\mathrm{Var}(X) \neq \sum_{i=1}^{m} \delta_i \mathrm{Var}(X_i)\cdot$$

Ver o Exemplos I.2 de estimação de misturas de densidades.

Uma Cadeia de Markov é por vezes um modelo probabilístico adequado para determinada série temporal em que a observação em um determinado momento é uma categoria à qual um indivíduo corresponde. A mais simples Cadeia de Markov é aquela na qual existe um número finito de estados ou categorias, um número finito de pontos no tempo equidistantes em que são feitas as observações, a cadeia é de primeira ordem e as probabilidades de transição são as mesmas para cada intervalo de tempo. Vamos considerar agora como obter estimadores da matriz de probabilidades transição.

A típica Cadeia de Markov a tempo discreto limita a descrição do histórico de cada sujeito a pontos de tempo igualmente espaçados. Em outras palavras, em vez de modelar a possibilidade de progressão a cada instante no tempo, ou seja, dia, mês ou ano. O intervalo entre esses pontos de tempo é conhecido como o comprimento do ciclo.

Presume-se que a matriz de transição vai ser estimada a partir de dados de coorte longitudinais, com intervalos de observação comuns a todos os sujeitos. A atenção é restrita a obtenção da estimativa de máxima verossimilhança da matriz de transição no caso em que os intervalos de observação são constantes e coincidem com a duração do ciclo.

Vamos considerar $\{c_0, c_1, \cdots, c_n\}$ uma amostra de uma Cadeia de Markov com probabilidades de transição $p_{x,y}$ e distribuição inicial π_0. Observe que $\{c_0, \cdots, c_n\}$ deve ser uma sequencia de $n+1$ estados. Então, a probabilidade de que $c_0, \cdots, c_n$ seja essa sequência é justamente

$$\pi(x_0)p_{c_0,c_1}\cdots p_{c_{n-1},c_n}.$$

Para $x, y = 1, 2, \cdots, d$, seja $n_{x,y}$ o número de transições, assim a matriz $(n_{x,y})$ vai ser chamada de matriz de contagens de transições da sequência. Dado que

$$\pi(x_0)p_{c_0,c_1}\cdots p_{c_{n-1},c_n} = \pi(x_0)\prod_{x,y} p_{x,y}^{n_{x,y}},$$

a contagem das transições junto com o estado inicial formam uma estatística suficiente. Sabemos que a probabilidade de obtiver uma sequência em particular, que comece com com c_0 e tenha matriz de transição $(p_{x,y})$ foi dada acima e, com o objetivo de encontrarmos a distribuição da estatística é necessário e suficiente somente contar o número de tais sequências. Esse número de conhece como Fórmula de Whittle (Whittle, 1955).

Exemplo

Seja, por exemplo, a sequência de 12 valores observados $\{0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 0, 1\}$. Esta sequência tem $u = 0$ e $v = 1$ e matriz de contagens de transição

$$(n_{x,y}) = \begin{pmatrix} 1 & 4 \\ 3 & 3 \end{pmatrix}.$$

Podemos utilizar a seguinte função **R** para encontrarmos a matriz de contagens de transição

```
> x = c(0,1,1,0,1,0,1,1,1,0,0,1)
> library(markovchain); library(matlab); library(matlib)
> Matriz = createSequenceMatrix(x, sanitize=FALSE)
> Matriz
    0  1
0   1  4
1   3  3
```

Vemos, da expressão na Fórmula de Whittle que $n^*_{x,y} = \begin{pmatrix} \frac{4}{5} & -\frac{4}{5} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$ e $C_{0,1} = 4/5$.

Exemplo (continuação)

Substituindo, temos que

$$N_{0,1}^{(12)}(n_{x,y}) = \frac{5! \cdot 6!}{1! \cdot 3! \cdot 3! \cdot 4!} \cdot \frac{4}{5} = 80 \cdot$$

Logo, 80 é o número de sequências $\{0, c_1, \cdots, c_{10}, 1\}$ tendo contagens de transição $(n_{x,y})$, dada acima. Desenvolvemos uma função **R** para encontrarmos o número de sequências $\{c_0, c_1, \cdots, c_n\}$ tendo contagens de transição $(n_{x,y})$ e satisfazendo $c_0 = u$ e $c_n = v$:

```
> Whittle = function(M, u, v)
    n = length(rowSums(M))
    Prod1 = 1;
    for(i in 1:n) Prod1 = Prod1*gamma(rowSums(M)[[i]]+1)
    Prod2 = 1;
    for(i in 1:dim(M)[1])
    for(k in 1:dim(M)[2]) Prod2 = Prod2*gamma(M[i,k]+1)
    u = which(row.names(M) == u)
    v = which(row.names(M) == v)
    C = cofactor(eye(n)-M/rowSums(M), v, u)
    return((Prod1/Prod2)*C)

> Whittle(Matriz, 0, 1)
 80
```

Suponhamos que nos seja dado a realização de uma Cadeia de Markov e que se deseja estimar a matriz de probabilidades de transição. Uma abordagem é encontrar as contagens de transição e estimar as probabilidades de transição de uma forma óbvia, pelo método dos momentos.

Exemplo

Este é uma situação hipotética. Consideremos uma Cadeia de Markov com três estados da qual é observada a sequência:

233211111221313233212222323233222221313233221221323213223 2
313233222321323233123222323233122212323213212323313233 2121

Por simples contagem, segue-se que, a matriz do número de transições entre os estados é

$$(n_{x,y}) = \begin{pmatrix} 4 & 8 & 10 \\ 13 & 17 & 22 \\ 6 & 26 & 9 \end{pmatrix},$$

onde $n_{x,y}$ denota o número de transições observadas desde o estado x ao estado y.

Uma vez que o número de transições do estado 2 para o estado 3 é 22 e o número total de transições do estado 2 é $13 + 17 + 22$, uma estimativa empírica de $\widehat{p}_{2,3}$ é 22/52.

Exemplo (continuação)

Uma estimativa empírica para a matriz de transição seria então

$$\widehat{\Gamma} = \begin{pmatrix} \frac{4}{22} & \frac{8}{22} & \frac{10}{22} \\ \frac{13}{52} & \frac{17}{52} & \frac{22}{52} \\ \frac{6}{41} & \frac{26}{41} & \frac{9}{41} \end{pmatrix} \cdot$$

Este é, de fato, a estimativa de máxima verossimilhança condicional de Γ, condicionada à primeira observação. Suponhamos, então, que nós queremos estimar os $d^2 - d$ parâmetros de uma Cadeia de Markov $\{C_n\}$ com d estados a partir uma realização $c_0, c_1, \cdots, c_T$. A função de verossimilhança condicional à primeira observação é

$$\ell = \sum_{x=1}^{d} \left(\sum_{y=1}^{d} n_{x,y} \log p_{x,y} \right) = \sum_{x=1}^{d} \ell_x,$$

a qual podemos maximizar maximizando cada somando separadamente.

Exemplo

Fazendo uso do pacote de funções **markovchain**, construímos a matriz de contagens de transição utilizando o comando **createSequenceMatrix**, como mostrado a seguir:

```
> x = c(2,3,3,2,1,1,1,1,1,2,2,1,3,1,3,2,3,3,2,1,2,2,2,2,3,2,3,2,3,3,2,2,2,2,2,1,3,1,3,2,3,3,
      2,2,1,2,2,1,3,2,3,2,1,3,2,2,3,2,3,1,3,2,3,3,2,2,2,3,2,1,3,2,3,2,3,3,1,2,3,2,2,2,3,2,
      3,2,3,3,1,2,2,2,1,2,3,2,3,2,1,3,2,1,2,3,2,3,3,1,3,2,3,3,2,1,2,1)
> Matriz = createSequenceMatrix(x, sanitize=FALSE)
> Matriz
    1   2   3
1   4   8  10
2  13  17  22
3   6  26   9
```

Às vezes o número de Whittle é impraticável, como nesta situação. Utilizando a função anterior obtemos que:

```
> Whittle(Matriz, u = 2, v = 1)
"[1]"   8.462769e+44
```

Exemplo (continuação)

A questão agora é transformar as frequências observadas em probabilidades, para isso utilizamos o comando **markovchainFit** do qual temos por resposta uma lista com diversas informações. A primeira resposta é a matriz de probabilidades de transição estimada, a qual pode ser obtida também digitando:

```
> markovchainFit(x)[[1]]
MLE Fit
  A 3 - dimensional discrete Markov Chain defined by the following states:   1,   2,   3
  The transition matrix (by rows) is defined as follows:
          1         2         3
1 0.1818182 0.3636364 0.4545455
2 0.2500000 0.3269231 0.4230769
3 0.1463415 0.6341463 0.2195122
```

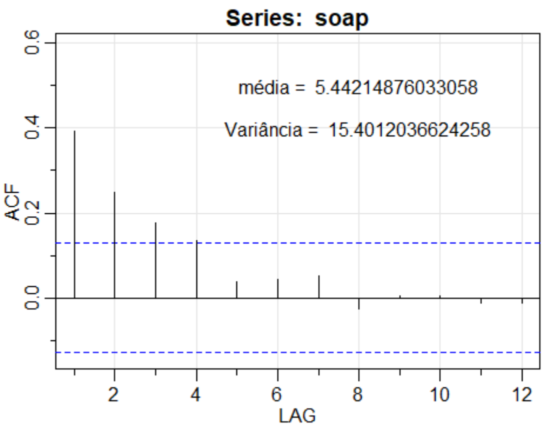
Teremos ocasião de comparar a função de autocorrelação (ACF) de um Modelo Oculto de Markov com a de uma Cadeia Markov. Portanto, discutiremos esta última agora. Assumimos, naturalmente, que os estados da Cadeia de Markov são quantitativos e não meramente categóricos. A função de autocorrelação (ACF) de uma Cadeia de Markov $\{C_n\}$ em $\{1, 2, 3, \cdots, m\}$ assumida estacionária e irredutível, pode ser obtida da seguinte forma

$$\rho(k) \; = \; \mathrm{Corr}\big(C_t, C_{t+k}\big)\cdot$$

Exemplo

Aplicando uma mistura de distribuições discretas à série de vendas de sabão, como mostrado, pode-se pensar em alguns estados subjacentes não observados que determinam de qual componente da mistura extrair uma observação. Em um modelo de mistura independente, esses estados ou as observações, respectivamente, devem ser independentes. No entanto, esse não é o caso da série de vendas de sabão, como pode ser visto na figura abaixo, que mostra a função de autocorrelação dessa série.

Exemplo (continuação)



Função de autocorrelação da série de vendas de sabão..

Exemplo (continuação)

A figura acima foi obtida utilizando os comandos:

```
> library(astsa)
> acf1(soap, max.lag = 12)
0.39 0.25 0.18 0.14 0.04 0.04 0.05 -0.03 0.01 0.01 -0.01 -0.01
> text(8,0.5,paste("média = ",mean(soap)))
> text(8,0.4,paste("Variância = ",var(soap)))
```

A figura acima revela que as observações de vendas de sabão estão significativamente correlacionadas entre um a três, indicando dependência serial na série. Assim, uma mistura independente não é um modelo apropriado aqui, pois não considera todas as informações contidas nos dados e deve-se pensar em modelos alternativos. Uma maneira de modelar séries de dados de contagem com correlação serial é aplicar um Modelo Oculdo de Markov Poisson, que é um caso especial de uma mistura dependente.

Nos casos em que as observações sobre um processo com espaço de estados finito parecem não satisfazer a propriedade de Markov, uma possibilidade que se sugere é usar uma Cadeia de Markov de ordem superior, ou seja, um modelo $\{C_n\}$ que satisfaça a seguinte generalização da propriedade de Markov para alguns $l \geq 2$.

Vamos considerar sequências $\{C_n\}$ com espaço de estados S. No modelo proposto, assumimos que a distribuição de probabilidade da sequência no tempo n depende da distribuição de probabilidade da sequência no tempo $n-1, \cdots, n-m$.

Definição

Seja $\{C_n\}$ uma sequência de variáveis aleatórias categóricas dependentes. Diz-se que a sequência satisfaz a propriedade de Markov de ordem m se

$$P(C_{n+1} = c_{n+1} \,|\, C_0 = c_0, C_1 = c_1, \cdots, C_n = c_n) = \\ = P(C_{n+1} = c_{n+1} \,|\, C_{n-m} = cx - m, C_{n-m+1} = c_{n-m+1}, \cdots, C_n = c_n)\cdot$$