Modelos Ocultos de Markov

Parte IV Previsão e decodificação.

Parte V. Seleção e validação de modelos.

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Setembro, 2020

Parte IV Previsão e decodificação.

Como mencionado, uma característica atraente dos HMMs é que as distribuições condicionais e as distribuições de previsão estão disponíveis. Isso torna mais fácil, por exemplo, verificar se há outliers ou fazer previsões por intervalos. Aqui, primeiro mostramos como calcular as distribuições condicionais de um HMM (Seção IV.1). Então, com base na fórmula para a distribuição condicional, derivamos a distribuição das previsões de um HMM (Seção IV.2). Finalmente, na última seção (Seção IV.3), demonstramos como se podem obter informações sobre os estados ocultos da Cadeia de Markov subjacente, dado o valor de HMM. Esta inferência é chamada decodificação.

Parte V. Seleção e validação de modelos.

Na Seção IV.3, mencionamos a tarefa de selecionar um HMM apropriado para uma série de observações e também a tarefa de verificar o ajuste do modelo selecionado. Neste capítulo, apresentaremos, na Seção V.1, uma introdução à seleção de modelos nos HMMs e ao respectivo diagnóstico de modelos usando pseudo-resíduos, isso na Seção V.2.

Utilizando a verossimilhança de um HMM tal como previsto na Seção II.3 e a definição das probabilidades para a frente e para trás (ver Seção III.1), é possível obter uma fórmula para a distribuição de S_t condicionada às outras observações do HMM.

Primeiro, é preciso introduzir uma nova notação:

$$\begin{aligned} S^{(-u)} &= \{S_1, \cdots, S_{u-1}, S_{u+1}, \cdots, S_T\}, \\ s^{(-u)} &= \{s_1, \cdots, s_{u-1}, s_{u+1}, \cdots, s_T\}. \end{aligned}$$

Isto significa que S^u e s^u denotam as sequências das variáveis aleatórias S_t e as observações, respectivamente, com S_u e s_u excluídos. Então, para $u = 1, 2, , T$ a distribuição condicional de S_u, dadas todas as outras observações é dada por

$$P\left(S_u = s \,|\, S^{(-u)} = s^{(-u)}\right) = \frac{\alpha_{u-1} \Gamma P(s) \beta_u^\top}{\alpha_{u-1} \Gamma \beta_u^\top}.$$

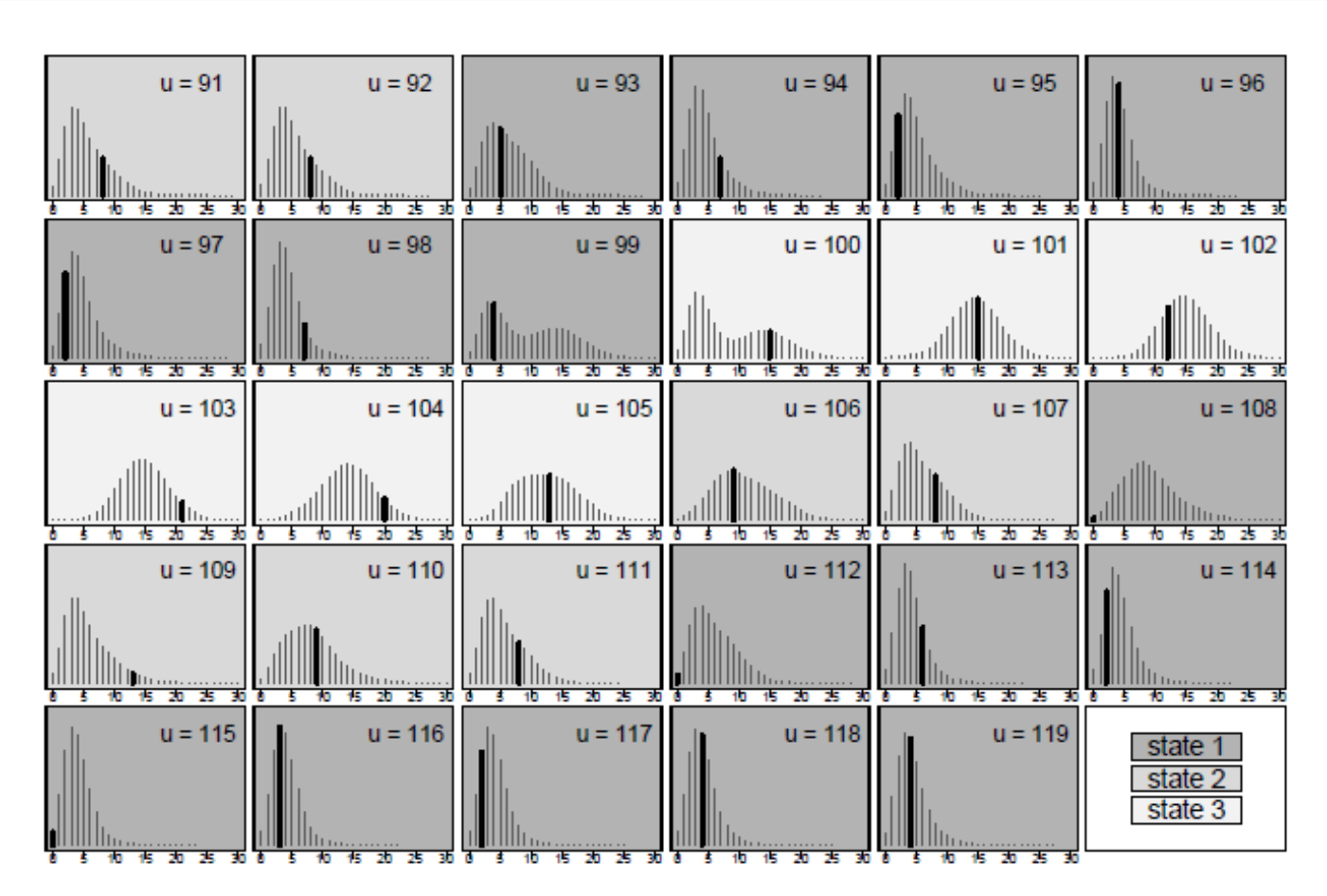

Figura IV.1: Distribuição condicional de S_u para $u \in \{91, \cdots, 119\}$.
Figura retirada do livro Zucchini & MacDonald (2009).

A idéia da previsão é calcular a probabilidade de certas observações ocorrerem no futuro. Ou seja, alguém está interessado na derivação da função de probabilidade condicional de S_{t+h}, dado $S^{(T)} = s^{(T)}$, em que h é chamado horizonte de previsão.

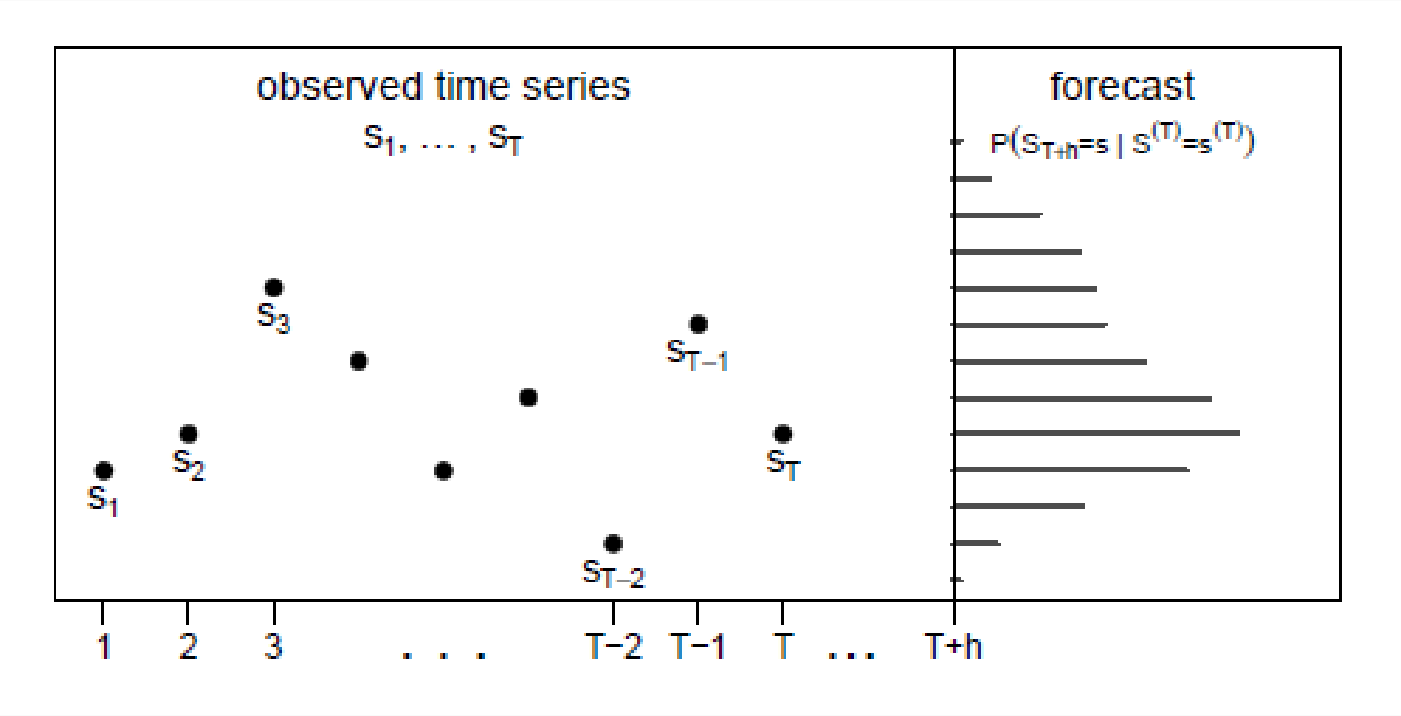

Figura IV.2: Previsão da distribuição de um HMM.
Figura retirada do livro Zucchini & MacDonald (2009).

Como a distribuição da previsão $P\big(\{S_{T+h} = s \,|\, S^{(T)} = s^{(T)}\}\big)$ de um HMM é um caso especial de uma distribuição condicional, ela pode ser calculada de maneira semelhante à distribuição condicional $P(S_u = s \,|\, S^{(-u)} = s^{(-u)})$ derivada na seção anterior:

$$\begin{aligned} P\big(\{S_{T+h} = s \,|\, S^{(T)} = s^{(T)}\}\big) &= \frac{P\big(\{S_{T+h} = s, S^{(T)} = s^{(T)}\}\big)}{P\big(\{S^{(T)} = s^{(T)}\}\big)} \\ &= \frac{\delta B_1 \cdots B_T \Gamma^h P(s) \mathbf{1}^\top}{\delta B_1 \cdots B_T \mathbf{1}^\top} \\ &= \frac{\alpha_T \Gamma^h P(s) \mathbf{1}^\top}{\alpha_T \mathbf{1}^\top}. \end{aligned}$$

Pode-se mostrar que, à medida que o horizonte de previsão h aumenta, a distribuição da previsão converge para a distribuição estacionária do HMM, ou seja,

$$\lim_{h \to \infty} P\big(\{S_{t+h} = s \,|\, S^{(T)} = s^{(T)}\}\big) = \delta P(s) \mathbf{1}^\top.$$

Normalmente, a distribuição das previsões se aproxima da distribuição estacionária relativamente rápido. Por exemplo, considere as distribuições das previsões do HMM Poisson de três estados da série de vendas de sabão para $h = 1, , 6$, isto é, $t = 243, , 248$, mostrados na Figura IV, também retirada do livro Zucchini & MacDonald (2009).

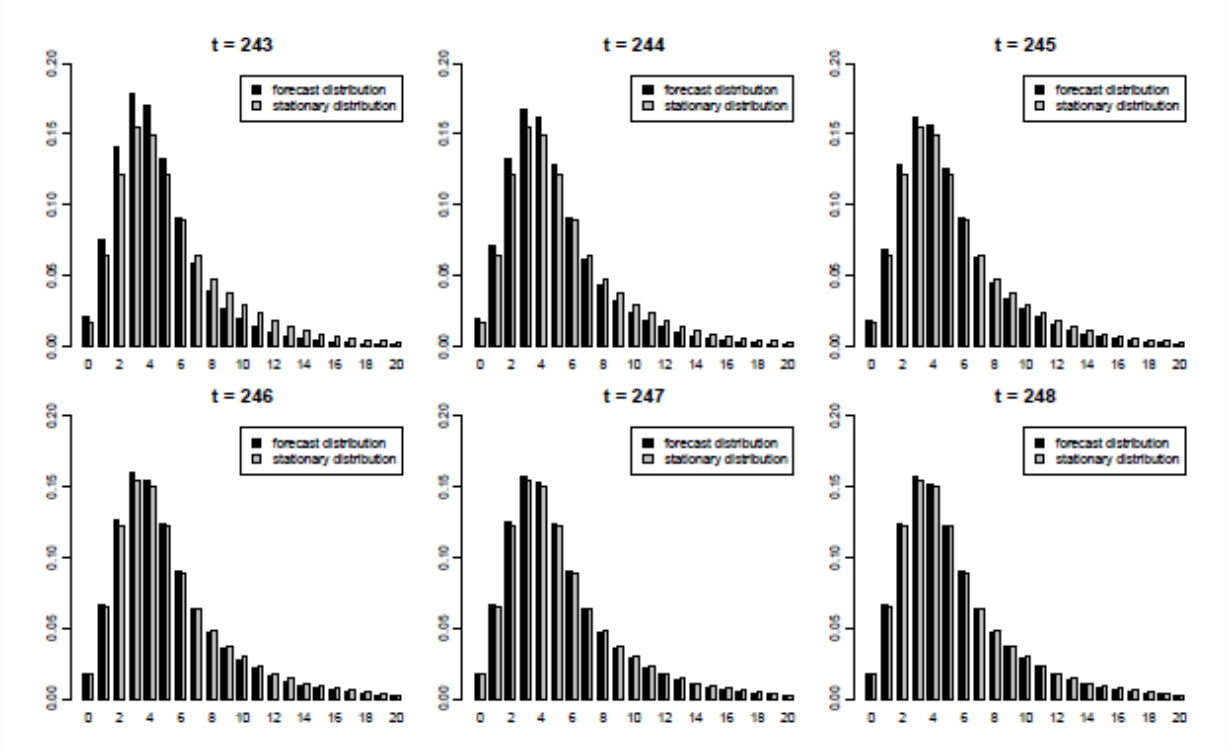

Figura IV.3: Distribuições das previsões, série de vendas de sabão.

Uma questão muito natural surge no contexto de muitas aplicações, por exemplo, todos os tipos de reconhecimento, quais são, considerando o HMM estimado, os estados mais prováveis da Cadeia de Markov subjacente?

Por exemplo, no campo do reconhecimento de fala onde os HMMs foram aplicados pela primeira vez, os estados subjacentes da Cadeia de Markov representam as sílabas das palavras que devem ser reconhecidas, enquanto as observações são palavras faladas e barulhos. Portanto, a matriz de probabilidades de transição pode ser gerada analisando palavras, por exemplo, de um dicionário e contando a aparência de certas sequências de sílabas. Em seguida, as observações ruidosas devem ser analisadas e uma sequência adequada de estados da Cadeia de Markov, ou seja, uma sequência de sílabas e, portanto, de palavras, devem ser encontradas.

Esse tipo de investigação de um HMM é chamado de decodificação. Distingue-se entre decodificação local que determina o estado mais provável no tempo t, independente dos outros tempos, e decodificação global que deriva a sequência mais provável de estados. Esses dois serão descritos nas seções a seguir.

Reconsidere as probabilidades de avanço (forward) e retrocesso (backward), α_t e β_t, conforme definidas na Seção III.1. Para a derivação do estado mais provável da Cadeia de Markov no momento t, a seguinte declaração sobre essas probabilidades é muito útil:

$$\alpha_t(i)\beta_t(i) \;=\; P\big(\big(S^{(T)} = s^{(T)}, C_T = i\big)\cdot$$

Este resultado pode ser interpretado da seguinte maneira. $\alpha_t(i)\beta_t(i)$ é a probabilidade conjunta das observações $S^{(T)} = s^{(T)}$ e a Cadeia de Markov C_t estar no estado i no momento t. Portanto, a distribuição condicional de C_t, dadas as observações, $P\big(\{C_t = i \,|\, S^{(T)} = s^{(T)}\}\big)$, pode ser obtida como

$$P\big(\{C_t = i \,|\, S^{(T)} = s^{(T)}\}\big) \;=\; \frac{P\big(\{C_t = i, S^{(T)} = s^{(T)}\}\big)}{P\big(\{S^{(T)} = s^{(T)}\}\big)} \;=\; \frac{\alpha_t(i)\beta_t(i)}{L_T}.$$

Aqui, o L_T pode ser calculado usando o algoritmo de dimensionamento apresentado na Seção III.3. No entanto, para impedir o subfluxo numérico na avaliação do termo $\alpha_t(i)\beta_t(i)$, é necessário definir um novo algoritmo de escala para o cálculo do $\beta_t(i)$. Isso pode, por exemplo, ser alcançado usando os mesmos fatores de escala obtidos para o $\alpha_t(i)$ no algoritmo da Seção III.3.

Assim, para cada vez que $t \in \{1, \cdots, T\}$ pode-se determinar toda a distribuição do estado C_t, dado o HMM observado. Por exemplo, no caso de dois estados, um tem

$$\begin{aligned} P(\{C_t = 1 \,|\, S^{(T)} = s^{(T)}\}) &= \frac{\alpha_t(1)\beta_t(1)}{L_T}, \\ P(\{C_t = 2 \,|\, S^{(T)} = s^{(T)}\}) &= \frac{\alpha_t(2)\beta_t(2)}{L_T}. \end{aligned}$$

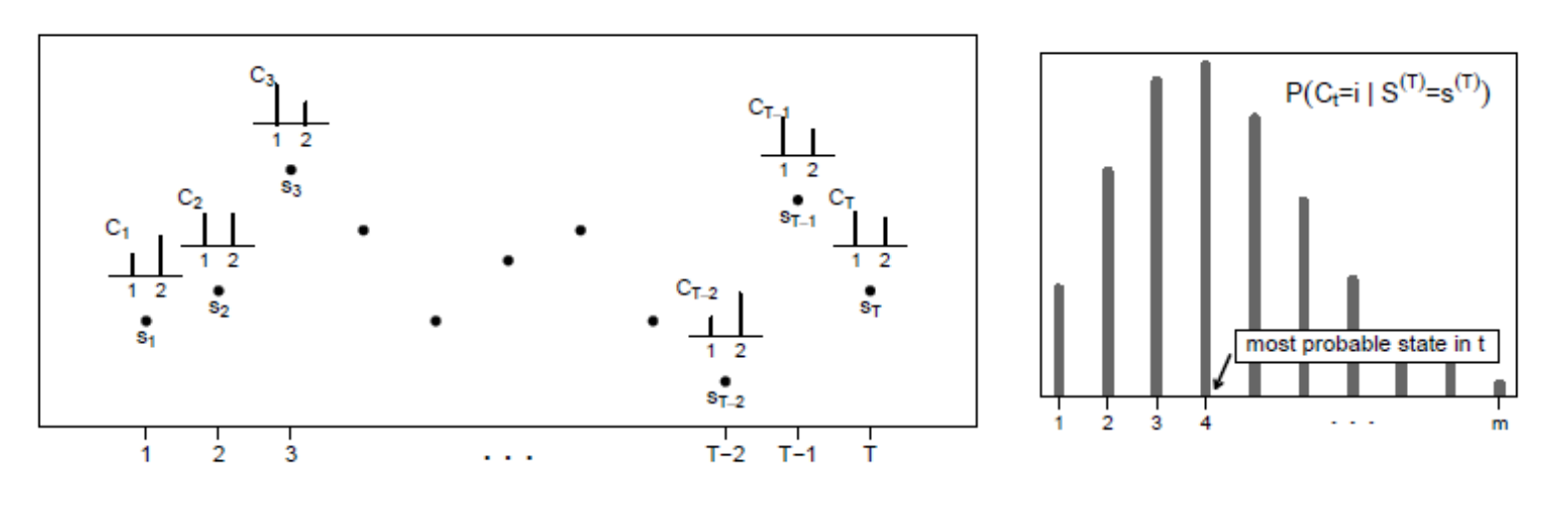

Figura IV.: Decodificação local para um HMM.
Figura retirada do livro Zucchini & MacDonald (2009).

Em muitas aplicações, especialmente em todos os campos de reconhecimento, não estamos interessados pelo estado mais provável para um tempo separado t, conforme fornecido pela decodificação local, mas pela sequência mais provável de estados ocultos. Isso significa que, em vez de maximizar o valor da probabilidade condicional de um estado, $P\big(\{C_t = i \,|\, S^{(T)} = s^{(T)}\}\big)$, tenta-se determinar a sequência de estados $(i_1^*, \cdots, i_T^*)$ que maximiza a probabilidade condicional

$$(i_1^*, \cdots, i_T^*) = \underset{(i_1^*, \cdots, i_T^*) \in \{1, \cdots, m\}}{\operatorname{argmax}} P\big(\{C_1 = i_1, \cdots, C_T = i_T \,|\, S^{(T)} = s^{(T)}\}\big)$$

ou de forma equivalente e mais conveniente, a probabilidade conjunta

$$\begin{aligned}(i_1^*, \cdots, i_T^*) &= \underset{(i_1^*, \cdots, i_T^*) \in \{1, \cdots, m\}}{\operatorname{argmax}} P\big(\{C_1 = i_1, \cdots, C_T = i_T, S^{(T)} = s^{(T)}\}\big) \\ &= \underset{(i_1^*, \cdots, i_T^*) \in \{1, \cdots, m\}}{\operatorname{argmax}} (\delta_{i_1} \gamma_{i_1, t_2}, \cdots, \gamma_{i_{T-1}, i_T})\big(p_{i_1}(s_1) \cdots, p_{i_T}(s_T)\big).\end{aligned}$$

Um método de programação dinâmica eficiente que pode ser aplicado no contexto dos HMMs é o algoritmo de Viterbi. Para aplicar esse algoritmo, precisamos definir uma variável auxiliar $\nu_t(i)$:

$$\nu_t(i) = \max_{i_1,\cdots,i_{t-1}} P\big(\{C_1 = i_1, \cdots, C_{t-1} = i_{t-1}, C_t = i, S^{(t)} = s^{(t)}\}\big),$$

$t \in \{2, \cdots, T\}$, onde $\nu_t(1)$ é dado por $\nu_t(1) = P(C_1 = i, S_1 = s_1)$.

Isto significa que $\nu_t(i)$ é a verossimilhança da distribuição conjunta das observações $S^{(t)} = s^{(t)}$, a sequência de estados $C_1 = i_1, \cdots, C_{t-1} = i_{t-1}$ e o estado $C_t = i$, maximizado em todas as sequências de estados possíveis $i_1, \cdots, i_{t-1}$. Pode-se demonstrar que as probabilidades $\nu_t(i)$ satisfazem a seguinte recursão para $t \in \{2, \cdots, T-1\}$:

$$\nu_{t+1}(j) = \left[\max_i \left(\nu_t(i)\gamma_{i,j}\right)\right] \left(p_j(s_{t+1})\right).$$

Esta recursão fornece um meio eficiente de calcular a matriz $T \times m$ dos valores $\nu_t(j)$, $t = 1, \cdots, T$; $j = 1, \cdots, m$. A sequência necessária de estados $i_1^*, \cdots, i_T^*$ pode então ser determinada recursivamente a partir de

$$i_T^* = \underset{i \in \{1, \cdots, m\}}{\operatorname{argmax}} \nu_T(i)$$

e

$$i_t^* = \underset{i \in \{1, \cdots, m\}}{\operatorname{argmax}} \nu_t(i) \gamma_{i, i_{t+1}^*}, \qquad t = T - 1, \cdots, 1.$$

Finalmente, gostaríamos de nos referir à Figura IV.1 novamente, onde as cores de fundo das distribuições condicionais indicam os estados subjacentes da Cadeia de Markov. Esses estados foram calculados usando o algoritmo Viterbi, ou seja, fazem parte da sequência de estados mais provável.

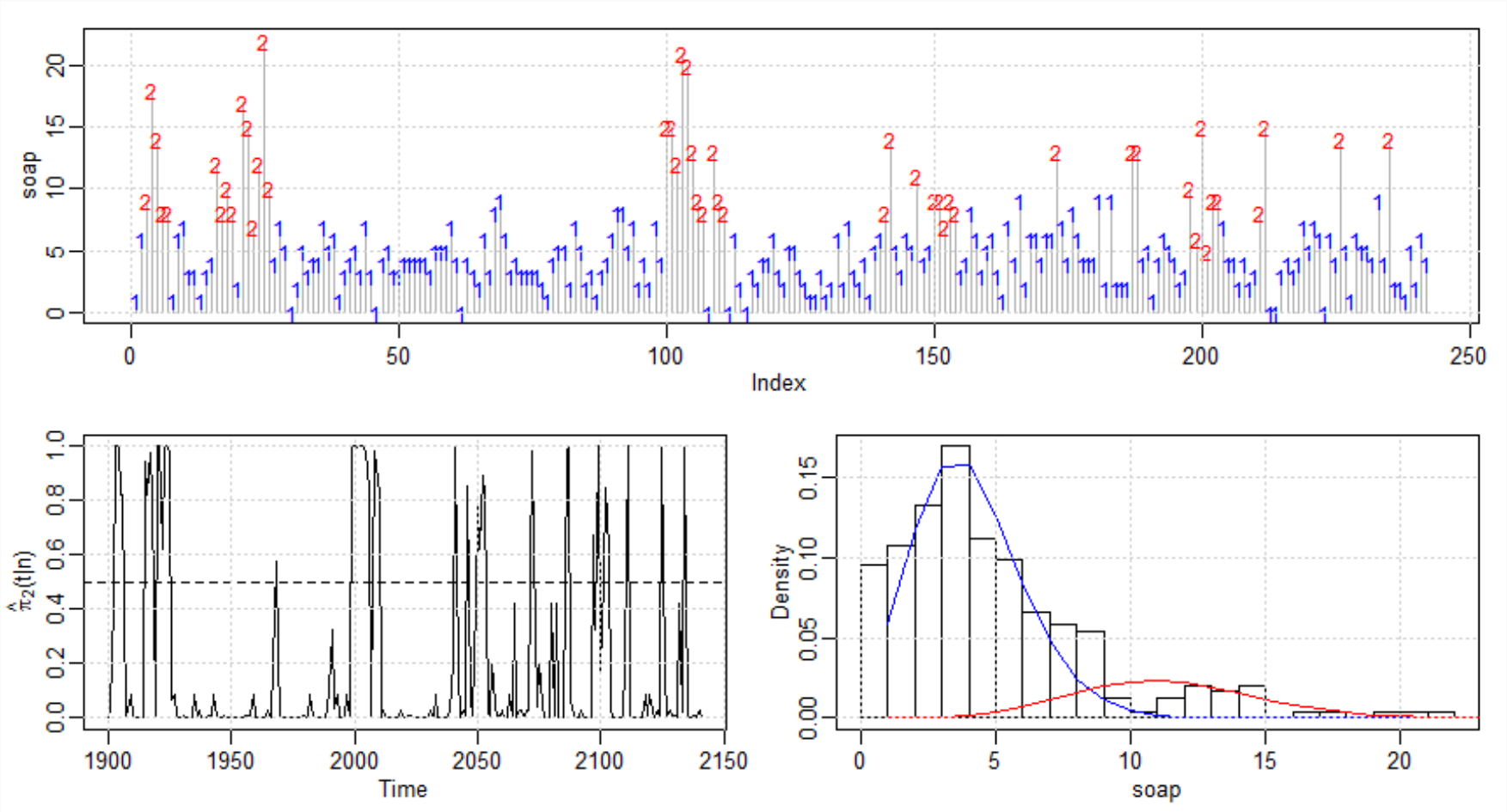

Figura IV.5: Série de vendas semanais de um produto de sabão específico e estados estimados. Probabilidades de suavização. Histograma dos dados com as duas densidades estimadas sobrepostas (linhas sólidas).

Um problema que surge naturalmente ao trabalhar com HMMs é o de selecionar um modelo apropriado, isto é, de escolher o número apropriado de estados m. Como os resultados assintóticos acerca da ordem m de um HMM ainda não são claros veja, por exemplo, Rydén (1995), é necessário especificar um critério para a comparação de modelos.

À escolha da discrepância de Kullback-Leibler leva ao chamado critério de informação de Akaike (AIC), que pode ser calculado da seguinte forma:

$$AIC = -2\log(L) + 2p,$$

onde $\log(L)$ é a log-verossimilhança do modelo ajustado e p representa o número de parâmetros do modelo. O primeiro termo é uma medida de ajuste, diminui com o aumento do número de estados m. O segundo termo é um termo de penalidade, aumenta com o aumento de m. O AIC é a escolha canônica para comparação de modelos. Para uma introdução mais detalhada ao AIC e sua derivação, ver Zucchini (2000).

A idéia da segunda abordagem da seleção de modelos, a abordagem bayesiana, é selecionar a família estimada com maior probabilidade de ser verdadeira.

Essa abordagem resulta no Critério de Informação Bayesiano (BIC), que tem um termo de penalidade ligeiramente modificado em comparação com o AIC:

$$BIC = -2\log(L) + p\log(T),$$

onde $\log(L)$ e p são como para o AIC e T é o número de observações.

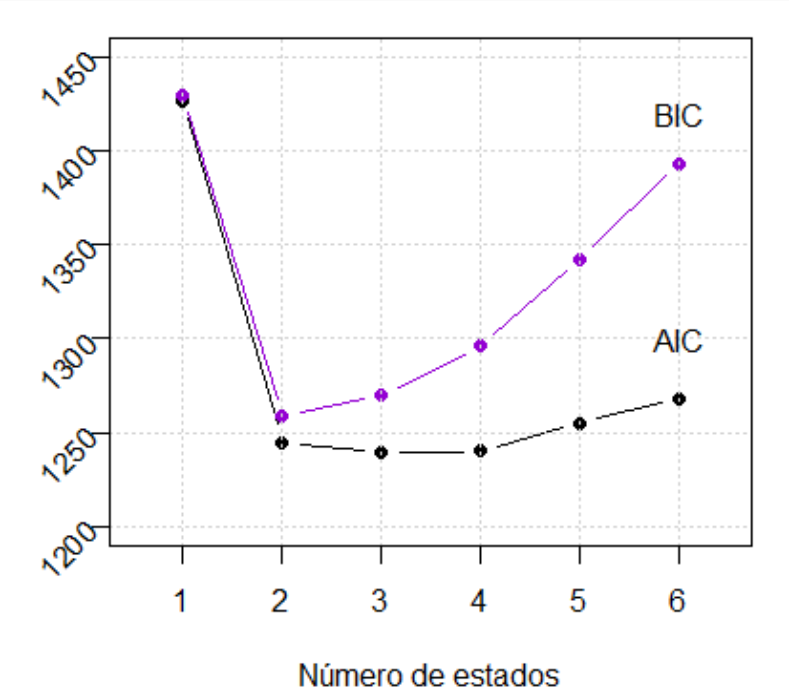

Figura V.1: Critérios de seleção de modelos AIC e BIC, série de vendas de sabão.

Observemos que: modelo HMM Poisson $m = 3$ estados, **AIC = 1239** e **BIC = 1270**, enquanto que para o modelo HMM Poisson $m = 4$ estados, **AIC = 1240** e **BIC = 1296**.

De acordo com a AIC, o modelo com $m = 3$ é o mais apropriado, enquanto a BIC seleciona o modelo de dois estados. Assim, o modelo selecionado depende da abordagem de seleção de modelos que se deseja seguir.

Suponha que se considere a AIC como critério de seleção, ou seja, escolhemos o modelo de três estados. As respectivas estimativas de parâmetros são fornecidas por:

$$\widehat{\Gamma} = \begin{pmatrix} 0.86 & 0.12 & 0.02 \\ 0.44 & 0.54 & 0.02 \\ 0.00 & 0.30 & 0.70 \end{pmatrix}, \qquad \begin{array}{rcl} \widehat{\delta} & = & (0.72 \quad 0.22 \quad 0.06) \\ \widehat{\lambda} & = & (3.74 \quad 8.44 \quad 14.93) \end{array}.$$

As distribuições do estado dependentes componentes desse modelo, juntamente com as distribuições marginais, são ilustradas na Figura V.2.

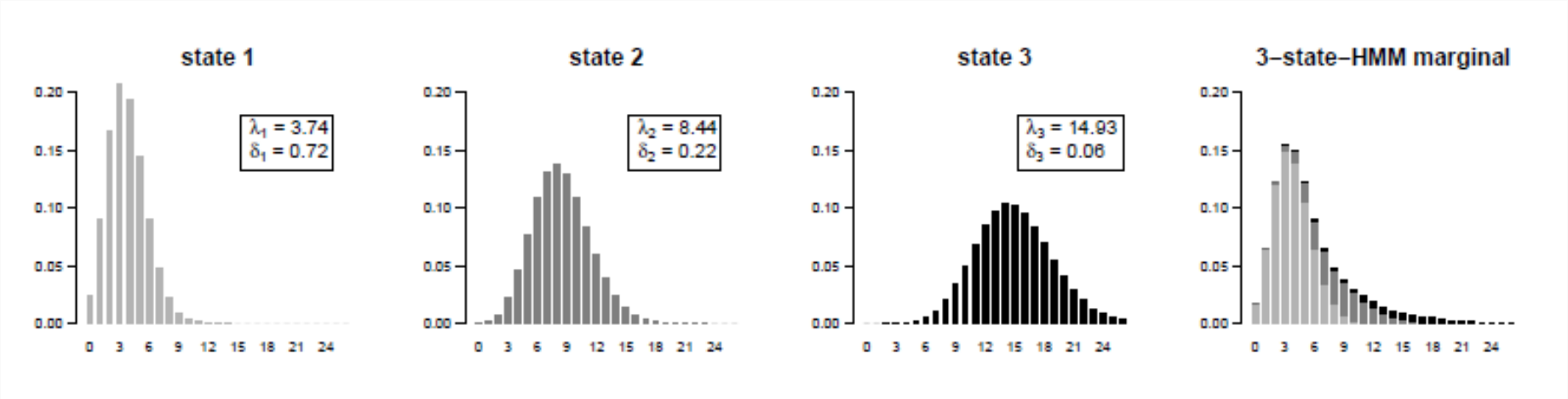

Figura V.2: Distribuições do estado dependentes componentes do modelo HMM Poisson $m = 3$ estados nos três primeiros gráficos, de esquerda para direita. Último mostrando as distribuições marginais. Dados da série de vendas de sabão.

Na seção anterior, consideramos dois critérios para a seleção de modelos nos HMMs. Esses critérios fornecem uma regra de decisão para selecionar o melhor de vários modelos ajustados, no entanto, eles não garantem que o modelo selecionado seja realmente apropriado. Portanto, ainda é preciso avaliar o ajuste do modelo escolhido.

Em geral, os resíduos são uma ferramenta popular para avaliar o ajuste de um modelo. No caso ideal, eles são distribuídos de forma independente e idêntica. Além disso, é de grande vantagem se, no caso de um modelo válido, eles forem distribuídos $U(0;1)$ ou $N(0;1)$, independentemente do modelo ajustado.

Por exemplo, para procurar outliers ou verificar estruturas e dependência específicas, pode-se desenhar um gráfico do índice dos resíduos ou plotá-los na variável dependente ou covariáveis. Caso certos padrões ocorram, o modelo deve ser revisado e melhorado, por exemplo, adicionando termos quadráticos ou cúbicos ou mesmo novas covariáveis. Outra possibilidade é testar as premissas distribucionais usando histogramas ou gráficos qq dos resíduos. Algumas dessas parcelas residuais são ilustradas na Figura V.4.

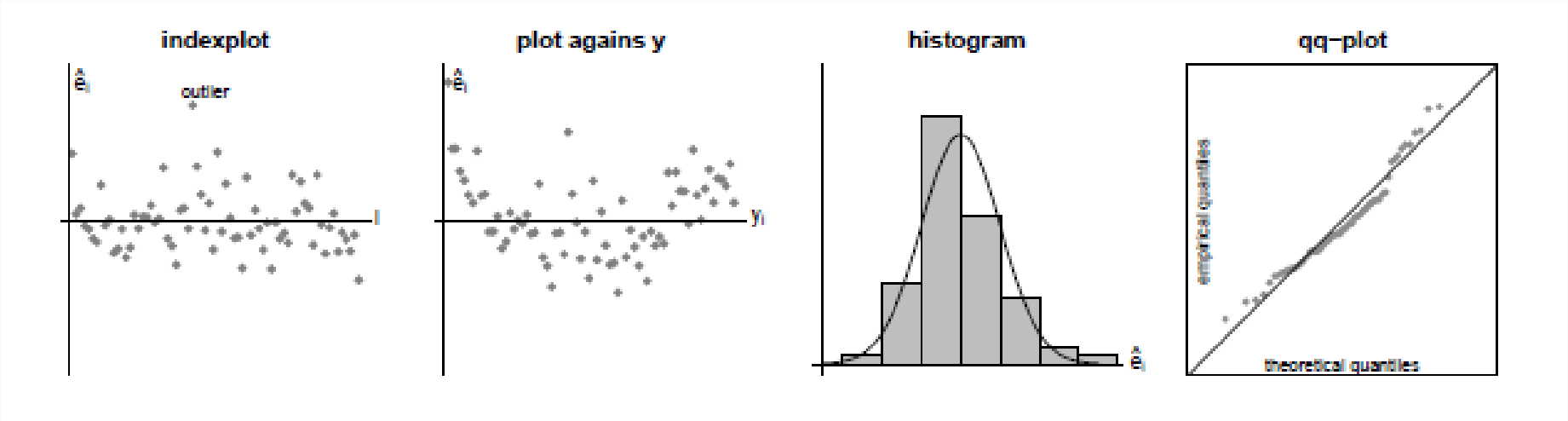

Figura V.4: Gráficos de resíduos.

O conceito de pseudo-resíduos, que se baseia no conceito de p-valores, satisfaz as propriedades desejadas dos resíduos delineadas no início desta seção pelo menos aproximadamente, pelo que é um meio adequado para avaliar o ajuste de um modelo. Os pseudo-resíduos podem ser definidos para quase todos os modelos e, caso o modelo seja válido, são distribuídos de forma independente e idêntica de forma uniforme ou normal.

Para a construção de pseudo-resíduos consideramos o seguinte teorema: Que X seja uma variável aleatória com função de distribuição F. Então, $U = F(X)$ é uniformemente distribuída no intervalo $(0; 1)$, ou seja:

$$U = F(X) \sim U(0, 1)\cdot$$

Com base neste teorema, uma primeira versão do pseudo-resíduo de uma observação x_i de uma variável aleatória contínua X_i pode ser definida como a probabilidade de obter uma observação inferior a x_i no modelo ajustado:

$$u_i = P\left(X_i \leq x_i\right) = F(x_i)\cdot$$

Dada a validade do modelo ajustado este tipo de pseudo-resíduo é distribuído $U(0; 1)$, com resíduos para observações extremas próximas de 0 ou 1 (Zucchini e MacDonald, 1999). A construção do chamado pseudo-resíduo uniforme ilustra-se na Figura V.5.

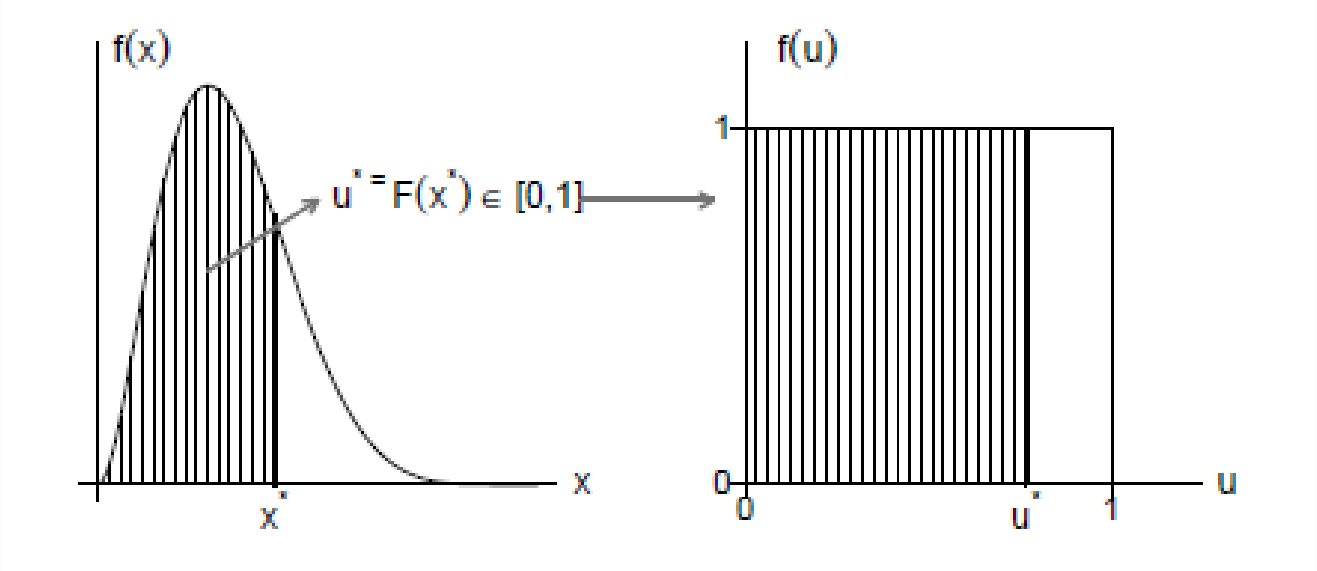

Figura V.5: Construção de pseudo-resíduos uniformemente distribuídos.

Assim, utilizando o conceito de pseudo-resíduos uniformes, observações de diferentes distribuições podem ser comparadas. Suponha ter observações $x_1, \cdots, x_n$ e um modelo $X_i \sim F_i$, isto é, cada x_i tem a sua própria função de distribuição e portanto os valores x_i não podem ser comparados diretamente. No entanto, o conceito de pseudo-resíduos pode ser utilizado para gerar pseudo-resíduos u_i que, no caso do modelo estar correto, são independentes e identicamente distribuídos $U(0;1)$ e, portanto, podem ser comparados.

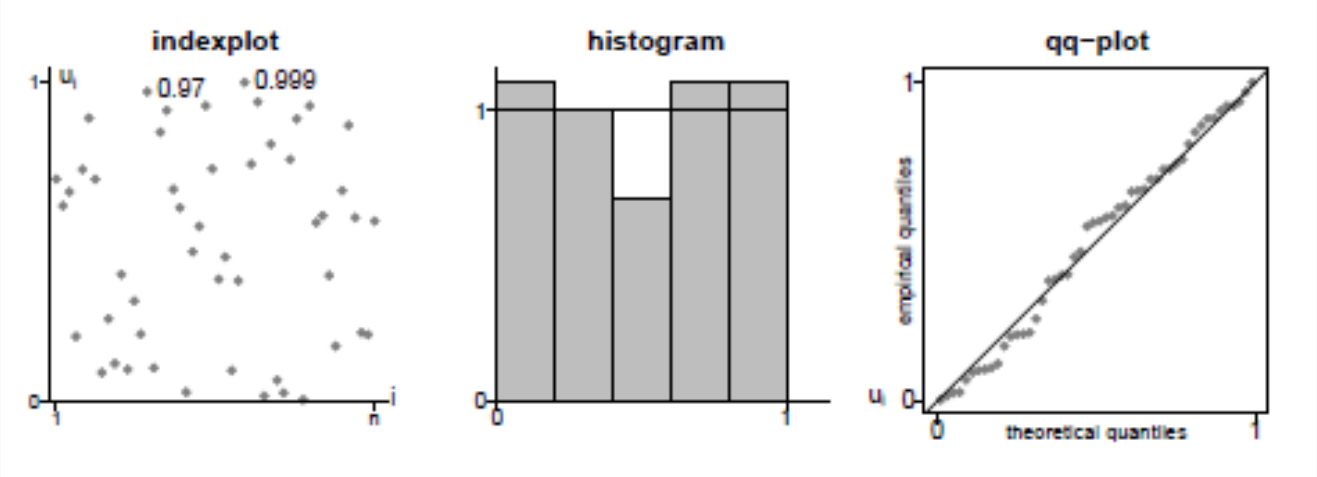

Figura V.6: Gráficos de diagnóstico para pseudo-resíduos uniformes.

Se o histograma e o gráfico qq-plot dos pseudo-resíduos uniformes ui não parecerem como deveriam, pode-se deduzir que os resíduos não são distribuídos uniformemente e, portanto, o modelo não é válido. Certamente, o conceito de pseudo-resíduos uniformes é muito útil, no entanto, pode levar a problemas na identificação externa. Por exemplo, considere o gráfico de índice fornecido na Figura V.6 e observe os valores próximos a 0 ou 1. É difícil ver se um valor é muito improvável ou não. Como um valor de 0.999 é difícil de distinguir de um valor de 0.97, o gráfico de índice é quase inútil para uma análise visual rápida.