Ch3. Estimation by direct maximization of the likelihood

HyeRim Park

July 28, 2023

Table of contents

Table of contents

Introduction

- The likelihood of an HMM is given by

$$L_T = \Pr\left(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}\right) \quad = \boldsymbol{\delta}\mathbf{P}\left(x_1\right)\boldsymbol{\Gamma}\mathbf{P}\left(x_2\right)\boldsymbol{\Gamma}\mathbf{P}\left(x_3\right)\cdots\boldsymbol{\Gamma}\mathbf{P}\left(x_T\right)\mathbf{1},$$

  where $\boldsymbol{\delta}$ is the initial distribution and $\mathbf{P}(x)$ the $m \times m$ diagonal matrix with ith diagonal element the state-dependent probability $p_i(x)$

- Then, in the likelihood formula:

$$\begin{gathered}\boldsymbol{\alpha}_1 = \boldsymbol{\delta}\mathbf{P}\left(x_1\right);\\ \boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}\left(x_t\right) \qquad \text{for } t = 2, 3, \ldots, T;\\ L_T = \boldsymbol{\alpha}_T\mathbf{1}.\end{gathered}$$

Introduction

- Parameter estimation can be performed by numerical maximization of the likelihood with respect to the parameters.

- But there are several problems that need to be addressed when parameters are estimated.
  1. Numerical underflow
  2. Constraints on the parameters
  3. Multiple local maxima in the likelihood function

Table of contents

Numerical underflow

- In the case of discrete state-dependent distributions, the elements of $\boldsymbol{\alpha}_t$ become smaller as t increases and are eventually rounded to zero.
- Then, likelihood approaches 0 exponentially fast with probability 1 (or possibly ∞ in the continuous case)
- We confine our attention to underflow.

Numerical underflow

- To solve the problem, we compute the logarithm of L_T by using a strategy of scaling the vector of forward probabilities $\boldsymbol{\alpha}_t$.

- Effectively we scale the vector $\boldsymbol{\alpha}_t$ at each time t so that its elements add to 1.

- Define the vector

$$\boldsymbol{\phi}_t = \boldsymbol{\alpha}_t / w_t$$

where $w_t = \sum_i \alpha_t(i) = \boldsymbol{\alpha}_t \mathbf{1}$

Scaling the likelihood computation

- Using the definitions of $\boldsymbol{\phi}_t$ and w_t:

$$\begin{aligned} w_0 &= \boldsymbol{\alpha}_0 \mathbf{1} = \boldsymbol{\delta} \mathbf{1} = 1 \\ \boldsymbol{\phi}_0 &= \boldsymbol{\delta} \\ \boldsymbol{\alpha}_t &= \boldsymbol{\alpha}_{t-1} \boldsymbol{\Gamma} \mathbf{P}\left(x_t\right) \\ w_t \boldsymbol{\phi}_t &= w_{t-1} \boldsymbol{\phi}_{t-1} \boldsymbol{\Gamma} \mathbf{P}\left(x_t\right) \\ L_T &= \boldsymbol{\alpha}_T \mathbf{1} = w_T \left(\boldsymbol{\phi}_T \mathbf{1}\right) = w_T. \end{aligned} \tag{1}$$

Scaling the likelihood computation

- Hence $L_T = w_T = \prod_{t=1}^{T} (w_t/w_{t-1})$. From (1) it follows that

$$w_t = \boldsymbol{\alpha_t}\mathbf{1} = w_{t-1}\phi_{t-1}\mathbf{\Gamma P}(x_t)\mathbf{1}$$

$$w_t = w_{t-1}\left(\phi_{t-1}\mathbf{\Gamma P}\left(x_t\right)\mathbf{1}\right),$$

  and so we conclude that

$$\log L_T = \sum_{t=1}^{T} \log\left(w_t/w_{t-1}\right) = \sum_{t=1}^{T} \log\left(\phi_{t-1}\mathbf{\Gamma}\boldsymbol{P}\left(x_t\right)\mathbf{1}\right).$$

Scaling the likelihood computation

- The computation of the log-likelihood is summarized below in the form of an algorithm.

  $\boldsymbol{\delta}$: initial distribution $\boldsymbol{\Gamma}$ and $\mathbf{P}(x_t)$: $m \times m$ matrix
  $\mathbf{v}$ and $\boldsymbol{\phi}_t$: vectors of length m
  l is the scalar in which the log-likelihood is accumulated.

$$
\begin{aligned}
& w_1 \leftarrow \boldsymbol{\delta}\mathbf{P}(x_1)\mathbf{1};\ \boldsymbol{\phi}_1 \leftarrow \boldsymbol{\delta}\mathbf{P}(x_1)/w_1;\ l \leftarrow \log w_1 \\
& \text{for } t = 2, 3, \ldots, T \\
& \qquad \mathbf{v} \leftarrow \boldsymbol{\phi}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(x_t) \\
& \qquad u \leftarrow \mathbf{v}\mathbf{1} \\
& \qquad l \leftarrow l + \log u \\
& \qquad \boldsymbol{\phi}_t \leftarrow \mathbf{v}/u \\
& \text{return } l
\end{aligned}
$$

- The log-likelihood $\log L_T$ is given by the final value of l.

Table of contents

Reparametrization to avoid constraints

- In Poisson–HMM, the elements of $\boldsymbol{\Gamma}$ and those of $\boldsymbol{\lambda}$, the vector of state-dependent means in a are subject to non-negativity and the row sums of $\boldsymbol{\Gamma}$ equal 1.
- Estimates of parameters should satisfy such constraints.
- Thus, when maximizing the likelihood we need to solve a constrained optimization problem.

Reparametrization to avoid constraints

- Special-purpose software can be used to maximize a function of several variables which are subject to constraints.
- However, depending on the implementation and the nature of the data, constrained optimization can be slow.
- For example, the constrained optimizer is slow if the optimum lies on the boundary of the parameter space.
- We shall focus on the use of the unconstrained optimizer **nlm**.

Reparametrization to avoid constraints

- Constraints depends on which state-dependent distribution, and Constraints that applay to the parameters of the Markov chain

- **Example: Poisson-HMM**
  - The means λ_i of the state-dependent distributions must be non-negative.
  - The rows of the transition probability matrix $\mathbf{\Gamma}$ must add to 1, and all the parameters γ_{ij} must be non-negative.

- The constraints can be imposed by making transformations.

Reparametrization to avoid constraints

- Define $\eta_i = \log \lambda_i$, for $i = 1, \ldots, m$. Then $\eta_i \in \mathbb{R}$
- After maximized the likelihood with the unconstrained parameters, the constrained parameter estimates can be obtained by transforming back: $\widehat{\lambda}_i = \exp \widehat{\eta}_i$.
- Note that $\mathbf{\Gamma}$ has m^2 entries but only $m(m-1)$ free parameters, as there are m row-sum constraints

$$\sum_{j=1}^{m} \gamma_{ij} = 1 \quad (i = 1, \ldots, m)$$

Reparametrization to avoid constraints

- Show transformation between the m^2 constrained probabilities γ_{ij} and $m(m-1)$ unconstrained real numbers τ_{ij}, $i \neq j$

- If $m = 3$, define the matrix

$$\mathbf{T} = \begin{pmatrix} - & \tau_{12} & \tau_{13} \\ \tau_{21} & - & \tau_{23} \\ \tau_{31} & \tau_{32} & - \end{pmatrix}$$

, a matrix with $m(m-1)$ entries $\tau_{ij} \in \mathbb{R}$. Now let $g : \mathbb{R} \to \mathbb{R}^+$ be a strictly increasing function, for example, $g(x) = e^x$

Reparametrization to avoid constraints

- Define

$$\nu_{ij} = \begin{cases} g(\tau_{ij}) & \text{for } i \neq j \\ 1 & \text{for } i = j. \end{cases}$$

- Set $\gamma_{ij} = \nu_{ij} / \sum_{k=1}^{m} \nu_{ik}$ (for $i, j = 1, 2, \ldots, m$) and $\mathbf{\Gamma} = (\gamma_{ij})$ and the transfornation in the opposite direction is

$$\tau_{ij} = \log(\gamma_{ij}/\gamma_{ii}), \quad \text{for } i \neq j$$

- We shall refer to the parameters η_i and τ_{ij} as **working parameters**, and to the parameters λ_i and γ_{ij} as **natural parameters**.

Reparametrization to avoid constraints

- Using the transformations of $\mathbf{\Gamma}$ and $\boldsymbol{\lambda}$, the calculation of the likelihood-maximizing parameters in two steps.
  1. Maximize L_T with the working parameters $\mathbf{T} = \{\tau_{ij}\}$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_m)$.
  2. Transform the estimates of the working parameters to estimates of the natural parameters:

$$\widehat{\mathbf{T}} \to \widehat{\mathbf{\Gamma}}, \quad \widehat{\boldsymbol{\eta}} \to \widehat{\boldsymbol{\lambda}}$$

Reparametrization to avoid constraints

- Considering inintial distribution $\boldsymbol{\delta}$ is not supplied.
- If $\boldsymbol{\delta}$ is stationary distribution, compute by solving

$$\boldsymbol{\delta}(\mathbf{I}_m - \boldsymbol{\Gamma} + \mathbf{J}_m) = \mathbf{1}'$$

- Otherwise, transform $\boldsymbol{\delta}$ with the constraints $\delta_i \geq 0$ and $\sum_i \delta_i = 1$ as unconstrained parameters.

Table of contents

Multiple maxima in the likelihood

- The likelihood of an HMM frequently has several local maxima.
- The goal is to find the global maximum, but there is no simple method of determining in general whether a numerical maximization algorithm has reached the global maximum.
- Depending on the starting values, it can easily happen that the algorithm identifies a local, but not the global, maximum.
- A sensible strategy is therefore to use a range of starting values for the maximization, and to see whether the same maximum is identified in each case.

Starting values for the iterations

- It is often easy to find plausible starting values for some of the parameters of an HMM.
- For instance, in a Poisson-HMM with two states to fit the sampe mean for the values of the two state-dependent means.
- To assign a common starting value to all the off-diagonal transition probabilities γ_{ij}.

Table of contents

Example: earthquakes

- Fitting stationary Poisson–hidden Markov models to the earthquakes series by means of the unconstrained optimizer **nlm**.

  1. nature parameter $\rightarrow$ working parameter
  2. iteration of **nlm**(mllk,wp) algorithm;
     mllk(working param $\rightarrow$ nature param, log-likelihood)
  3. estimate of working param $\rightarrow$ estimate of nature param

Example: earthquakes

- The three-state model is

$$\mathbf{\Gamma} = \begin{pmatrix} 0.955 & 0.024 & 0.021 \\ 0.050 & 0.899 & 0.051 \\ 0.000 & 0.197 & 0.803 \end{pmatrix}$$

with $\boldsymbol{\delta} = (0.4436, 0.4045, 0.1519)$, $\boldsymbol{\lambda} = (13.146, 19.721, 29.714)$, and log-likelihood given by $l = -329.4603$.

- The four-state is

$$\mathbf{\Gamma} = \begin{pmatrix} 0.805 & 0.102 & 0.093 & 0.000 \\ 0.000 & 0.976 & 0.000 & 0.024 \\ 0.050 & 0.000 & 0.902 & 0.048 \\ 0.000 & 0.000 & 0.188 & 0.812 \end{pmatrix}$$

with $\boldsymbol{\delta} = (0.0936, 0.3983, 0.3643, 0.1439)$, $\boldsymbol{\lambda} = (11.283, 13.853, 19.695, 29.700)$, and log-likelihood given by $l = -327.8316$.

Example: earthquakes

- The means and variances of the marginal distributions of the four models compare as follows with those of the observations.

	mean	variance
observations:	19.364	51.573
'one-state HMM':	19.364	19.364
two-state HMM:	19.086	44.523
three-state HMM:	18.322	50.709
four-state HMM:	18.021	49.837

Example: earthquakes

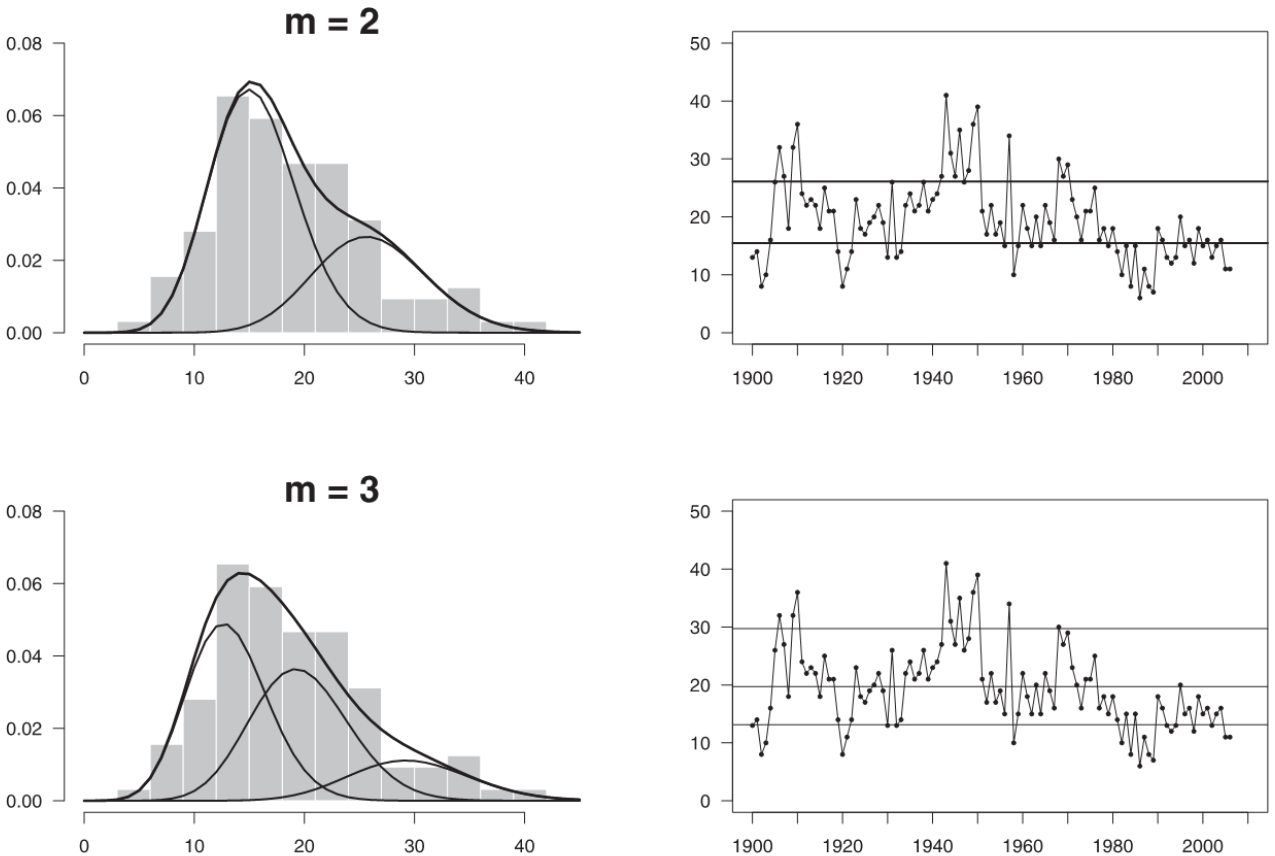

Figure 1: Earthquakes series. Left: marginal distributions of Poisson–HMMs with two and three states, and their components, compared with a histogram of the observations. Right: the state-dependent means (horizontal lines) compared to the observations.

Example: earthquakes

- The three-state model is

$$\mathbf{\Gamma} = \begin{pmatrix} 0.955 & 0.024 & 0.021 \\ 0.050 & 0.899 & 0.051 \\ \mathbf{0.000} & 0.197 & 0.803 \end{pmatrix}$$

- The four-state is

$$\mathbf{\Gamma} = \begin{pmatrix} 0.805 & 0.102 & 0.093 & \mathbf{0.000} \\ \mathbf{0.000} & 0.976 & \mathbf{0.000} & 0.024 \\ 0.050 & \mathbf{0.000} & 0.902 & 0.048 \\ \mathbf{0.000} & \mathbf{0.000} & 0.188 & 0.812 \end{pmatrix}$$

- When one fits models with three or more states is that the estimates of one or more of the transition probabilities turn out to be very close to zero.

Example: earthquakes

- In a stationary Markov chain, the expected number of transitions from state i to state j in a series of T observations is $(T-1)\delta_i\gamma_{ij}$.
- For $\delta_3 = 0.152$ and $T = 107$ in our three-state model, this expectation will be less than 1 if $\gamma_{31} < 0.062$.
- Therefore, it is likely that if γ_{31} is fairly small there will be no transitions from state 3 to state 1, and so when seeking to estimate γ_{31} in an HMM the estimate is likely to be effectively zero.
- As m increases, the probabilities δ_i and γ_{ij} get smaller on average.