Ch1. Preliminaries: mixtures and Markov chains

HyeRim Park

July 14, 2023

Table of contents

1. Introduction
2. Independent mixture models
3. Markov chains

Table of contents

1. Introduction

2. Independent mixture models

3. Markov chains

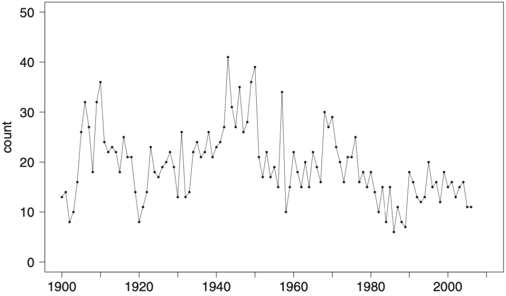
Introduction

- **Hidden Markov models (HMMs)** are models in which the distribution that generates an observation depends on the state of an unobserved Markov process.
- HMMs provide flexible general-purpose models for time series.
- This chapter is about
  1. brief and informal introduction of **HMMs**.
  2. **finite mixture distribution** that is marginal distribution of HMM.
  3. **Markov chains** which provide the underlying 'parameter process' of HMM .

Example: the series of annual counts of earthquakes for 1900-2006

13	14	8	10	16	26	32	27	18	32	36	24	22	23	22	18	25	21	21	14
8	11	14	23	18	17	19	20	22	19	13	26	13	14	22	24	21	22	26	21
23	24	27	41	31	27	35	26	28	36	39	21	17	22	17	19	15	34	10	15
22	18	15	20	15	22	19	16	30	27	29	23	20	16	21	21	25	16	18	15
18	14	10	15	8	15	6	11	8	7	18	16	13	12	13	20	15	16	12	18
15	16	13	15	16	11	11													

Table: Number of major earthquakes in the world, 1900–2006



Example: the series of annual counts of earthquakes for 1900-2006

- For this series, the application of standard models such as ARMA models would be not appropriate.
  $\rightarrow$ such models are based on the normal distribution.

- The usual model for unbounded counts is the Poisson distribution
  $\rightarrow$ but the series displays overdispersion and strong positive serial dependence.

- **HMMs** can accommodate both overdispersion and serial dependence.

Introduction of Hidden Markov model

- Attractive features of HMMs
  - their general mathematical tractability.
  - the likelihood is relatively straightforward to compute.
- Introduce the **basic HMM**: is univariate and based on a homogenous Markov chain and has neither trand nor seasonal variation.
- Ignore information that may be available on covariates

Table of contents

1. Introduction

2. Independent mixture models

3. Markov chains

Definition and properties

- Consider again the series of earthquake counts.
- A standard model for unbounded counts is the Poisson distribution, with its probability function $p(x) = e^{-\lambda}\lambda^x/x!$.
- However, for the earthquakes series the sample variance, $s^2 \approx 52$ is much larger than the sample mean, $\bar{x} \approx 19$, which indicates strong overdispersion.

Definition and properties

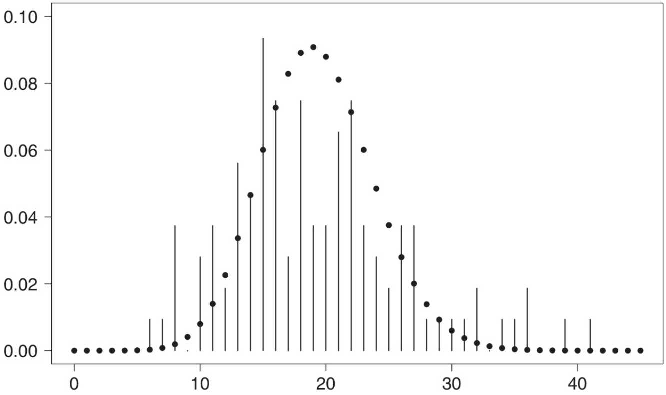


Figure: Major earthquakes, 1900-2006: bar plot of relative frequencies of counts, and fitted Poisson distribution.

Definition and properties

- One method of dealing with overdispersed observations is to use a mixture model.
- **Mixture model** is a probabilistic model for representing the presence of subpopulations within an overall population
- the population may consist of unobserved groups, each having a distinct distribution for the observed variable.
- That provides a principled approach to modeling such complex data.

Example: the series of annual counts of earthquakes for 1900-2006

- Suppose that each count in the earthquakes series is generated by one of two Poisson distributions, with means λ_1 and λ_2.
- The choice of mean is determined by some other random mechanism which we call the **parameter process**.
- Suppose also that λ_1 is selected with probability δ_1 and λ_2 with probability $\delta_2 = 1 - \delta_1$.
- If the parameter process is a series of independent random variables, the term 'independent mixture'.

Definition and properties

- In general, an independent mixture distribution consists of a finite number m, of component distributions and a 'mixing distribution' which selects from these components.

- In the case of two components, the mixture distribution depends on two probability or density functions:

component	1	2
probability or density function	$p_1(x)$	$p_2(x)$

- To specify the component, one needs a discrete random variable C which performs the mixing:

$$C = \begin{cases} 1 & \textit{with } \delta_1 \\ 2 & \textit{with probability } \delta_2 = 1 - \delta_2 \end{cases}$$

Definition and properties

- Let $\delta_1, ..., \delta_m$ denote the probabilities assigned to the different components, and let $p_1, ..., p_m$ denote their probability or density functions.
- Let X denote the random variable which has the mixture distribution.
- In discrete case, the probability function of X is given by

$$\begin{aligned} p(x) &= \sum_{i=1}^{m} Pr(X = x|C = i)Pr(C = i) \\ &= \sum_{i=1}^{m} \delta_i p_i(x) \end{aligned}$$

Definition and properties

- The expectation of the mixture can be given in terms of the expectations of the component distributions.
- Letting Y_i denote the random variable with probability function p_i,

$$E(X) = \sum_{i=1}^{m} Pr(C = i)E(X|C = i) = \sum_{i=1}^{m} \delta_i E(Y_i)$$

- The same result holds for a mixture of continuous distributions.

Parameter estimation

- The estimation of the parameters of a mixture distribution is often performed by maximum likelihood (ML).
- The likelihood of a mixture model with m components is given by,

$$L(\theta_1, ..., \theta_m, \delta_1, ..., \delta_m | x_1, ..., x_m) = \prod_{j=1}^{n} \sum_{i=m}^{m} \delta_i p_i(x_j, \theta_i) \quad (1)$$

- $\theta_1, ..., \theta_m$: the parameter vectors of the component distributions
  $\delta_1, ..., \delta_m$: the mixing parameters, totalling 1
  $x_1, ..., x_m$: n observations

Parameter estimation

- In the case of component distributions each specified by one parameter, $2m - 1$ independent parameters have to be estimated.
- Except perhaps in special cases, analytic maximization of such a likelihood is not possible.

Parameter estimation

- Suppose that $m = 2$ and the two components are Poisson-distributed with mean λ_1 and λ_2.
- Let δ_1 and $\delta_2 = 1 - \delta_1$ be the mixing parameters.
- The mixture distribution p is then given by

$$p(x) = \delta_1 \frac{\lambda_1^x e^{-\lambda_1}}{x!} + \delta_2 \frac{\lambda_2^x e^{-\lambda_2}}{x!}$$

- Since $\delta_2 = 1 - \delta_1$, there are only three parameters to be estimated: $\lambda_1, \lambda_2, \delta_1$
- The likelihood is

$$L(\lambda_1, \lambda_2, \delta_1 | x_1, ..., x_m) = \prod_{i=1}^{n} (\delta_1 \frac{\lambda_1^{x_i} e^{-\lambda_1}}{x_i!} + (1 - \delta_1) \frac{\lambda_2^{x_i} i e^{-\lambda_2}}{x_i!})$$

Parameter estimation

- The analytic maximization of L with respect to $\lambda_1, \lambda_2, \delta_1$ would be awkward, as *L* is the product of *n* factors, each of which is a sum.
- Taking the logarithm and then differentiating does not greatly simplify matters either.
- Therefore parameter estimation is more conveniently carried out by **direct** numerical maximization of the likelihood or **using EM algorithm**.

Unbounded likelihood in mixtures

- The likelihood of mixtures of continuous distributions is unbounded.
- For instance, in the case of a mixture of normal distributions, the likelihood becomes **arbitrarily large** if one sets a component mean equal to one of the observations and the corresponding variance to tend to zero.
- If the likelihood is thus unbounded, the ML estimates simply 'do not exist'.

Unbounded likelihood in mixtures

- Thus, replace each density value in a likelihood by the probability of the interval corresponding to the recorded value.
- In the context of independent mixtures one replaces the expression (1) for the likelihood by the discrete likelihood

$$L = \prod_{j=1}^{n} \sum_{i=m}^{m} \delta_i \int_{a_j}^{b_j} p_i(x, \theta_i) dx \tag{2}$$

where the interval (a_j, b_j) consists of those values which, if observed, would be recorded as x_j .

Examples of fitted mixture models: Poisson distribution

Model	i	δ_i	λ_i	$-logL$	Mean	Variance
$m = 1$	1	1.000	19.364	391.9189	19.364	19.364
$m = 2$	1	0.676	15.777	**360.3690**	19.364	46.182
	2	0.324	26.840			
$m = 3$	1	0.278	12.736	356.8489	19.364	51.170
	2	0.593	19.785			
	3	0.130	31.629			
$m = 4$	1	0.093	10.584	356.7337	19.364	51.638
	2	0.354	15.528			
	3	0.437	20.969			
	4	0.116	32.079			
observation					19.364	51.573

Table: Poisson independent mixture models fitted to the earthquakes series

Examples of fitted mixture models: Poisson distribution

- There is a very clear improvement in likelihood resulting from the addition of a second component.
- There is very little improvement from addition of a fourth apparently insufficient to justify the additional two parameter.
- It is clear that the mixtures fit the observations much better than does a single Poisson distribution, and visually the three- and four-state models seem adequate.

Table of contents

1. Introduction

2. Independent mixture models

3. Markov chains

Definition

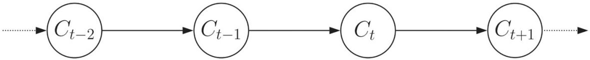
- Our treatment is restricted to those few aspects of discrete time Markov chains that we need.
- A sequence of discrete random variables $\{C_t : t \in \mathbb{N}\}$ is said to be a (discrete-time) **Markov chain** (MC) if, for all $t \in \mathbb{N}$, it satisfies the **Markov property**

$$Pr(C_{t+1}|C_t, ..., C_1) = Pr(C_{t+1}|C_t)$$

- To conditioning only on the most recent value C_t.

Definition

- The random variables $\{C_t\}$ are displayed in the following directed graph in which the future are dependent **only through the present**.



- Define **transition probabilities**:

$$\gamma_{ij}(t) = Pr(C_{s+t} = j | C_s = i)$$

- If these probabilities do not depend on s, the Markov chain is called **homogeneous**.
- We shall assume that the Markov chain under discussion is homogeneous.

Definition

- The matrix $\Gamma(t)$ is defined as the matrix with (i, j) element $\gamma_{ij}(t)$.
- An important property of all finite state-space homogeneous Markov chains is that they satisfy the **Chapman-Kolmogorov equations**: $\Gamma(t + u) = \Gamma(t)\Gamma(u)$.
- For all $t \in \mathbb{N}, \quad \Gamma(t) = \Gamma(1)^t$
- The matrix of t-step transition probabilities is the tth power of $\Gamma(1)$, the matrix of one-step transition probabilities.

Definition

- The matrix $\Gamma(1)$ which will be abbreviated as Γ, is a square matrix of probabilities with row sums equal to 1:

$$\Gamma = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1m} \\ \vdots & \ddots & \vdots \\ \gamma_{m1} & \cdots & a_{mm} \end{pmatrix}$$

  where m denotes the number of states of the Markov chain.

- The row sums are equal to 1 can be written as $\Gamma\mathbf{1}' = \mathbf{1}'$

- We shall refer to Γ as the (one-step) **transition probability matrix** (t.p.m.).

Definition

- The unconditional probabilities $Pr(C_t = j)$ of a Markov chain being in a given state at a given time t are often of interest.

- Denote these by the row vector

$$u(t) = \ (Pr(C_t = 1), \cdots, Pr(C_t = m)), \ t \in \mathbb{N}$$

- We refer to $u(1)$ as the initial distribution of the Markov chain.

- The distribution at time $t+1$ from that at t we postmultiply by the transition probability matrix Γ:

$$u(t+1) = u(t)\Gamma. \tag{3}$$

Stationary distributions

- A Markov chain with transition probability matrixΓ is said to have **stationary distribution** δ if $\delta\Gamma = \delta$ and $\delta\mathbf{1}' = 1$.
- Since $u(t+1) = u(t)\Gamma$, a Markov chain started from its stationary distribution will continue to have that distribution at all subsequent time points.
- An irreducible (homogeneous, discrete-time, finite state-space) Markov chain has a unique, strictly positive, stationary distribution.
- Always assume aperiodicity and irreducibility of stationary Markov chains.

Autocorrelation function

- To compare the autocorrelation function (ACF) of a hidden Markov model with that of its underlying Markov chain $\{C_t\}$, on the states $1, 2, ..., m$.
- Assume that these states are quantitative and not merely categorical, and stationary and irreducible.
- Then, for all non-negative integers k,

$$Cov(C_t, C_{t+k}) = \delta V \Gamma^k v' - (\delta v')^2 \tag{4}$$

Autocorrelation function

- If Γ is diagonalizable, and its eigenvalues (other than 1) are denoted by $\omega_2, \omega_3, \cdots, \omega_m$, then Γ can be written as

$$\Gamma = U \Omega U^{-1}$$

where Ω is $diag(1, \omega_2, \omega_3, \cdots, \omega_m)$ and the columns of U are corresponding right eigenvectors of Γ.

- Then, for all non-negative integers k,

$$\begin{aligned} Cov(C_t, C_{t+k}) &= \delta V U \Omega^k U^{-1} v' - (\delta v')^2 \\ &= a \Omega^k b' - a_1 b_1 \\ &= \sum_{i=2}^{m} a_i b_i \omega_i^k, \end{aligned}$$

where $a = \delta V U$ and $b' = U^{-1} v'$.