Hidden Markov Models: lecture 1

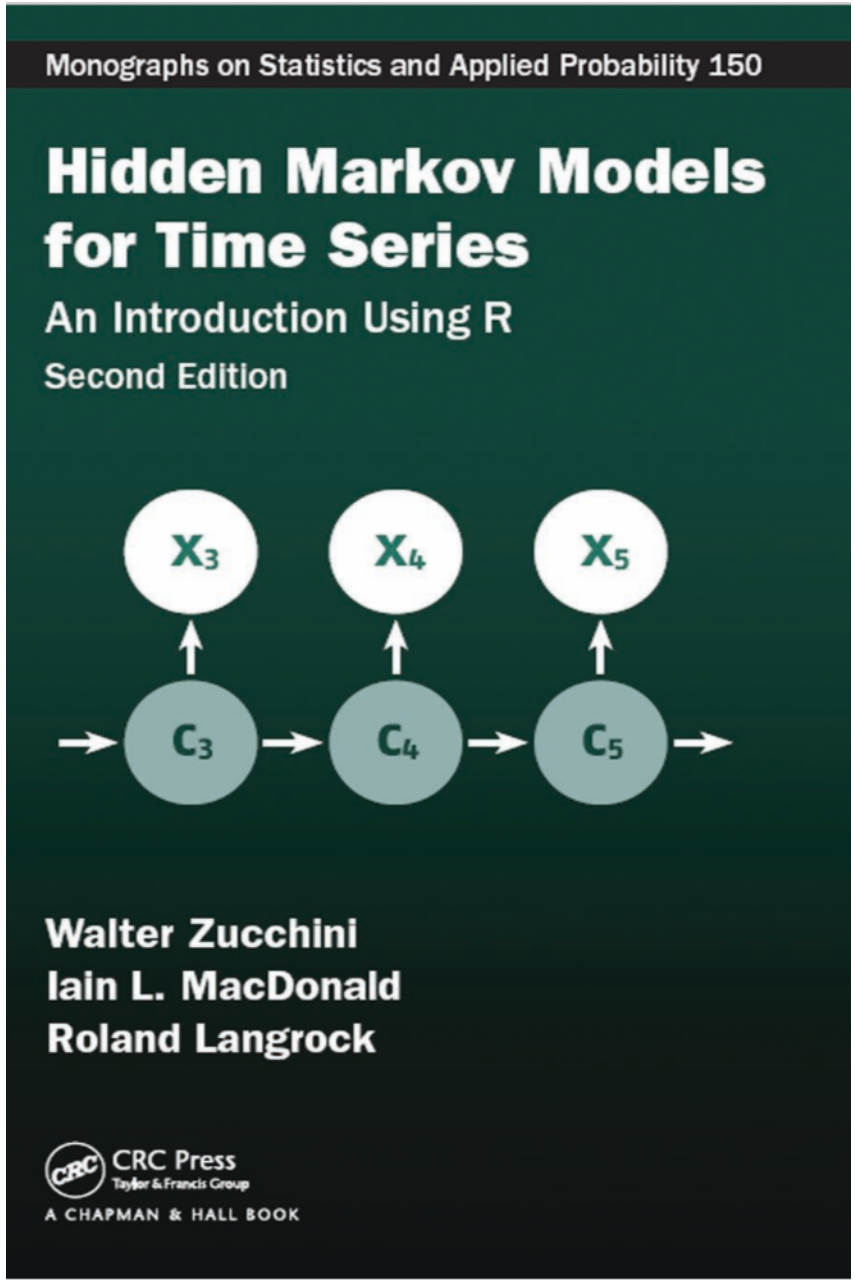
Introduction and definition

Xavier Didelot

Structure

1. Introduction and definition
2. Likelihood computation
3. Local decoding
4. Global decoding
5. Parameter estimation
6. Model selection and checking
7. Bayesian analysis
8. Application to speech recognition
9. Application to genetic data
10. Links and extensions

Companion textbook

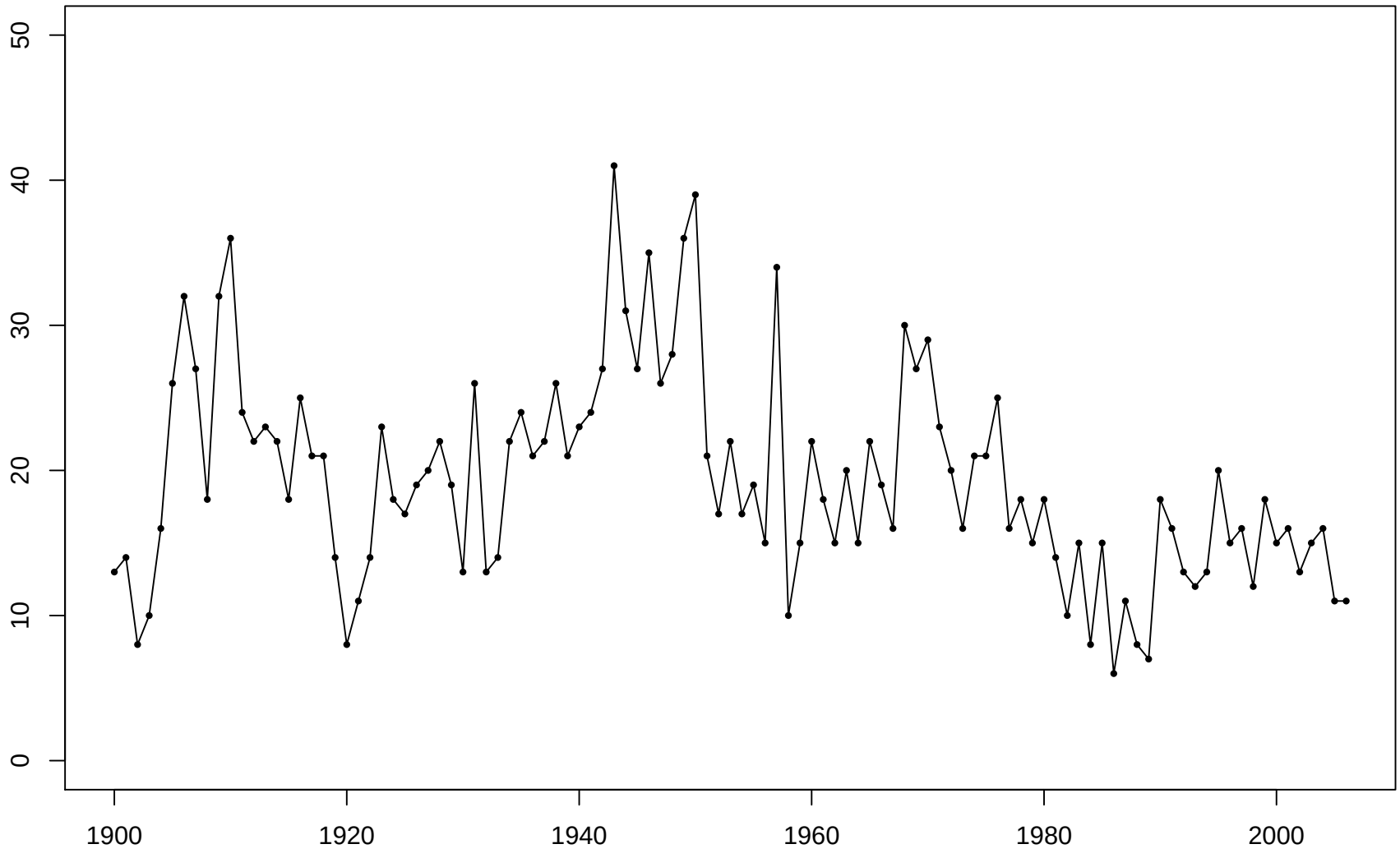


Example

Number of major earthquakes (magnitude 7 or greater) in the world, from 1900 until 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

Plot



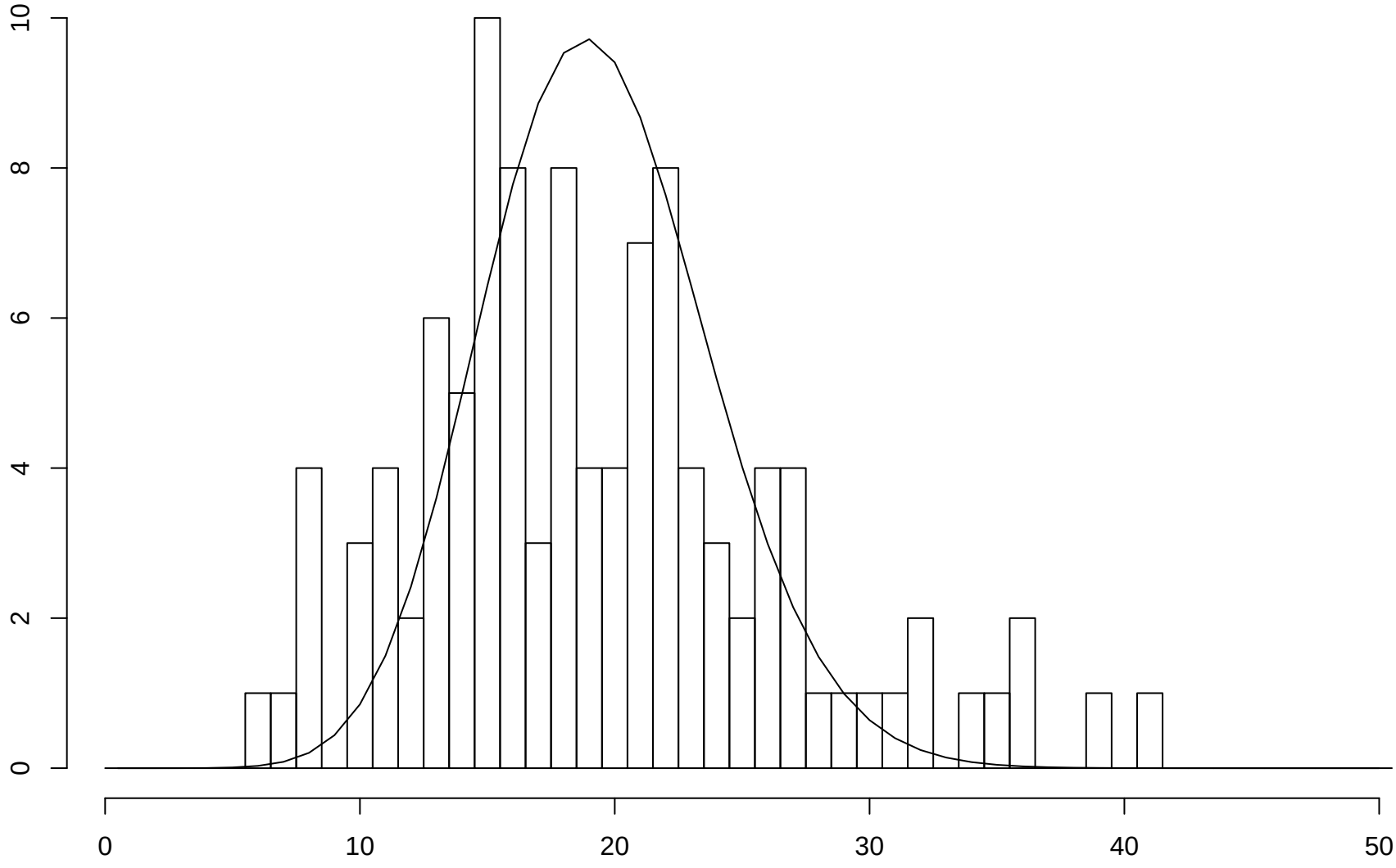
Observations

- This data is made of counts (positive integers)
- The application of standard time series techniques such as autoregressive moving-average (ARMA) models would be inappropriate, because they are based on the normal distribution
- A natural model would be a Poisson distribution with mean λ:

$$p(X_t = x_t) = e^{-\lambda}\lambda^{x_t}/x_t!$$

- The variance of a Poisson distribution is equal to its mean λ. But here we have a mean of 19.364486 and a variance of 51.5734438. So the Poisson model does not fit due to overdispersion of the data.

Data vs Poisson model



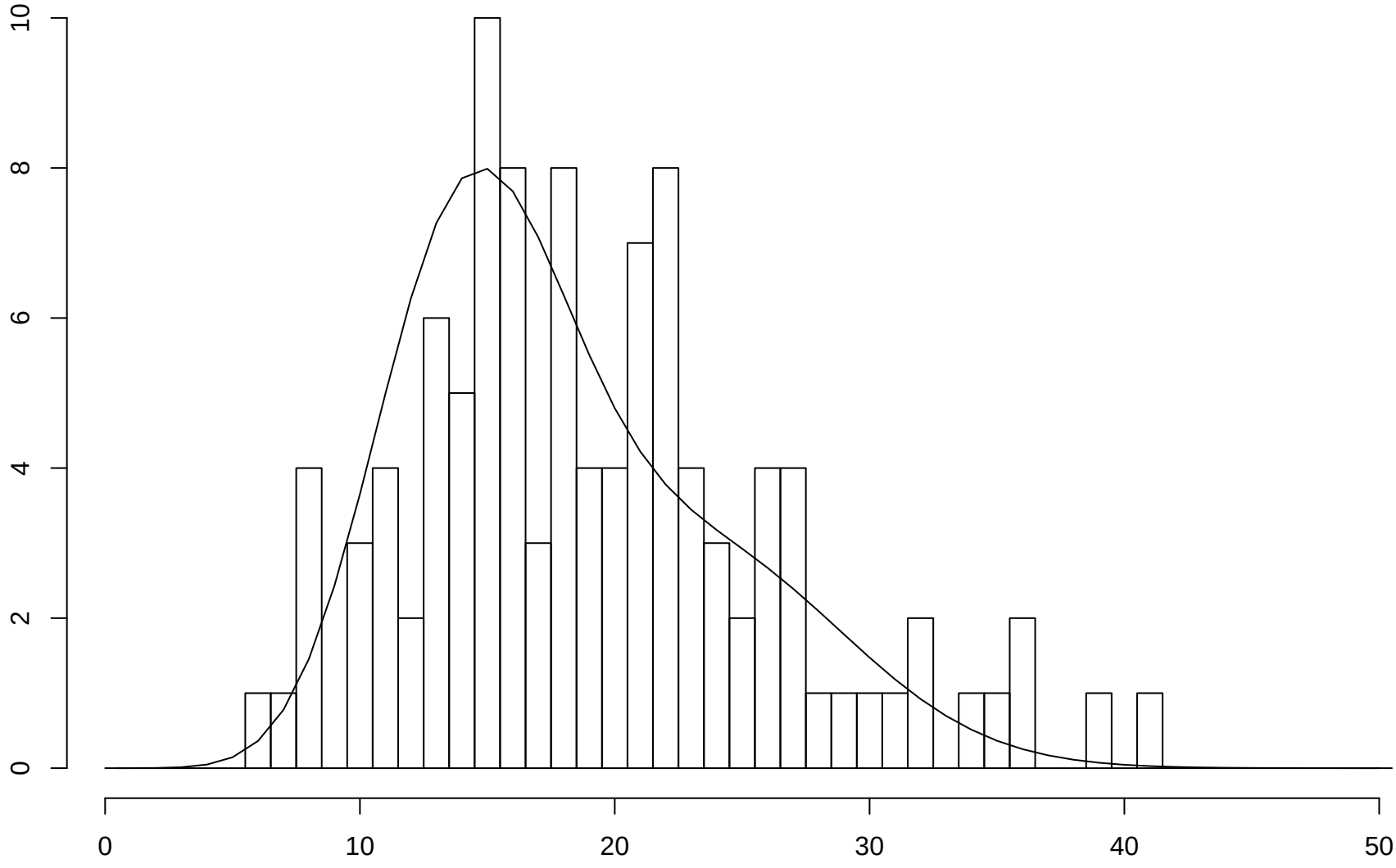
Independent mixture model

- We could consider a model with over-dispersion, for example a Negative Binomial distribution
- However, the distribution seems to have multiple modes which would not be captured
- Instead, let us consider an independent mixture model of (at least) two Poisson distributions
- For example, consider that 70% of the observations are from Poisson($\lambda_1 = 15$) and 30% are from Poisson($\lambda_2 = 25$):

$$p(X_t = x_t) = 0.7e^{-\lambda_1}\lambda_1^{x_t}/x_t! + 0.3e^{-\lambda_2}\lambda_2^{x_t}/x_t!$$

- In other words, the t^{th} observation X_t is generated by sampling C_t from Bernouilli(0.3) and then sampling X_t from Poisson(λ_{1+c_t})
- The fit is improved, and could be improved further by considering mixtures of more components

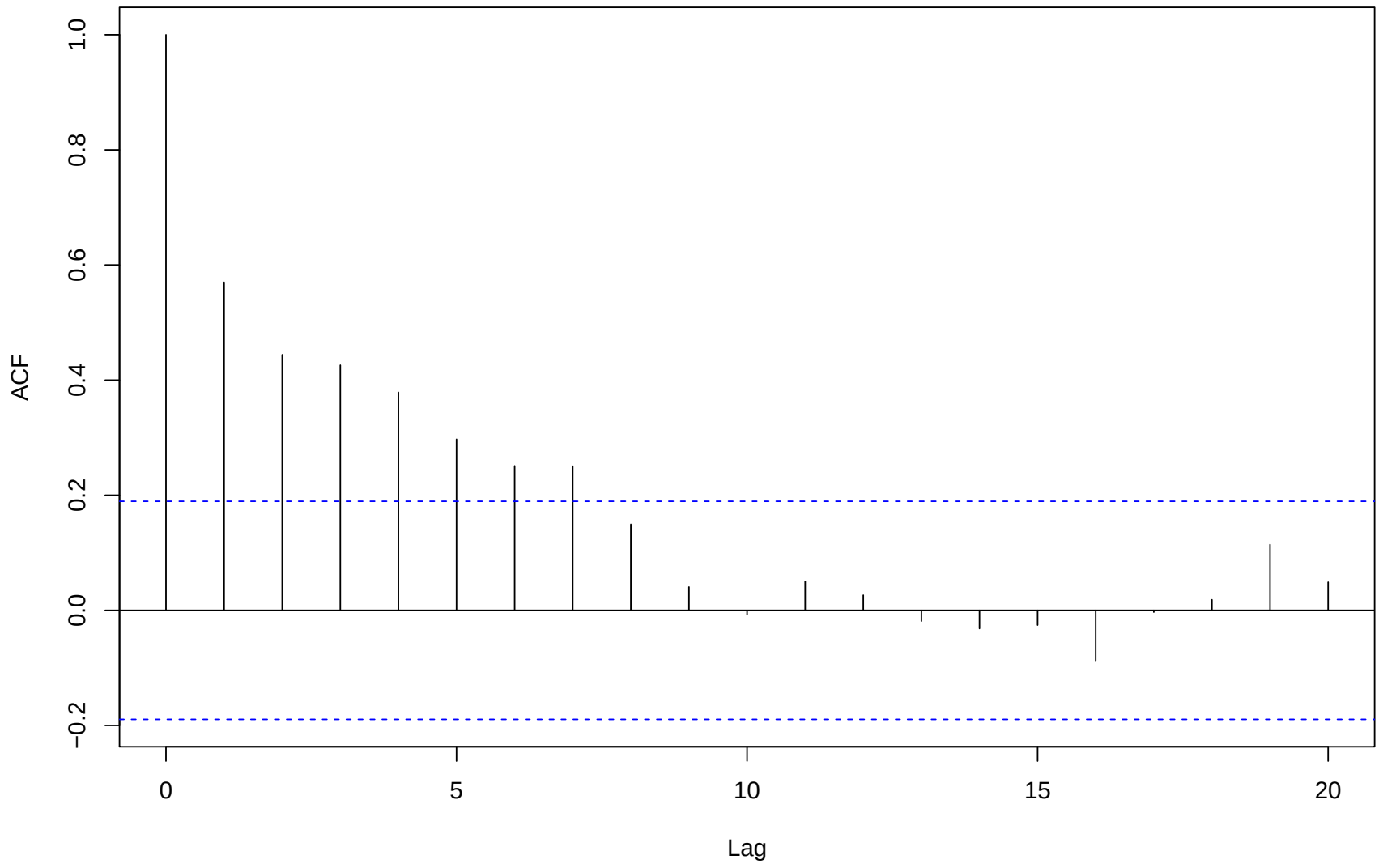
Data vs independent mixture model



Serial dependence

- So far we have been treating the data as if the observations X_t were independently and identically distributed
- However, it is clear by visual inspection that there is significant serial dependence in the data
- This can be made even clearer and tested by plotting the autocorrelation function (ACF) which is the Pearson's correlation at different lag values (lag=length of interval between observations)
- There is significant autocorrelation, with the first 8 ACF values being significantly greater than zero

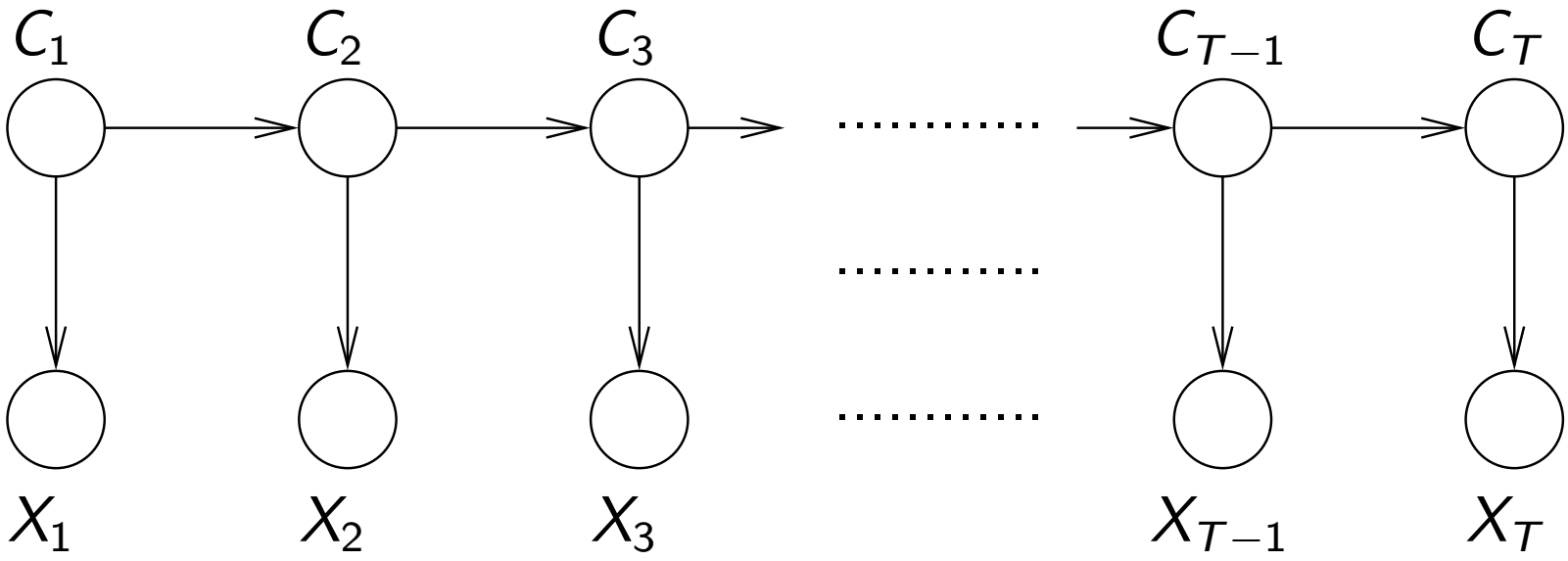
ACF



A first hidden Markov model

- The independent mixture model we described above does not capture serial dependence
- In this independent mixture model, the C_t variables representing which Poisson to sample X_t from were independently identically sampled from Bernouilli(0.3)
- Let us now instead consider that C_t is a Markov chain (homogenous, with order 1)
- The C_t variables are unobserved, but the X_t variables are observed
- The resulting process is therefore called a hidden Markov model
- It can be thought of as a natural extension to the independent mixture model

Dependency graph of a hidden Markov model



$$p(X_1, ..., X_T, C_1, ..., C_T) = p(C_1) \prod_{k=2}^{T} p(C_k|C_{k-1}) \prod_{k=1}^{T} p(X_k|C_k)$$

Refresher on Markov chains

- The Markov property states that:

$$p(C_t|C_{t-1}, C_{t-2}, ..., C_1) = p(C_t|C_{t-1})$$

- A Markov chain is homogeneous, if:

$$p(C_t = j|C_{t-1} = i) = p(C_2 = j|C_1 = i) \text{ for all } t, i, j$$

- In a homogeneous Markov chain, the transition probabilities do not depend on time
- We will always assume that the Markov chains are homogeneous unless otherwise stated

Refresher on Markov chains

- A Markov chain is defined by the transition probabilities:

$$\gamma_{ij} = p(C_t = j | C_{t-1} = i)$$

- The matrix of transition probabilities is denoted $\mathbf{\Gamma}$
- If the Markov chain has m states, $\mathbf{\Gamma}$ is of size $m \times m$ with rows adding up to 1, so that it contains $m \times (m-1)$ free parameters. Such a matrix is sometimes called a stochastic matrix or a Markov matrix:

$$\mathbf{\Gamma 1}' = \mathbf{1}'$$

- We can calculate the probability of a sequence of length L:

$$p(c_1, ..., c_L) = p(c_1) \prod_{t=2}^{T} \gamma_{c_{t-1} c_t}$$

- The first term $p(c_1)$ is given by the initial distribution of the Markov chain

Refresher on Markov chains

- Let $\mathbf{\Gamma}(k)$ denote the matrix of k-step transition probabilities:

$$\gamma_{ij}(k) = p(C_{t+k} = j | C_t = i)$$

- By definition we have:

$$\mathbf{\Gamma}(1) = \mathbf{\Gamma}$$

- The Chapman-Kolmogorov equations state that:

$$\mathbf{\Gamma}(t+u) = \mathbf{\Gamma}(t)\mathbf{\Gamma}(u)$$

- In other words:

$$\gamma_{ij}(t+u) = \sum_{k=1}^{m} \gamma_{ik}(t)\gamma_{kj}(u)$$

- Proof left as exercise
- Consequently:

$$\mathbf{\Gamma}(k) = \mathbf{\Gamma}^k$$

Stationary distribution

- The unconditional distribution of the Markov chain at t is:

$$\boldsymbol{u}(t) = (p(C_t = 1), p(C_t = 2), ..., p(C_t = m))$$

- The initial distribution is therefore $\boldsymbol{u}(1)$
- A Markov chain with transition probability matrix $\boldsymbol{\Gamma}$ has stationary distribution $\boldsymbol{\delta}$ if:

$$\boldsymbol{\delta\Gamma} = \boldsymbol{\delta} \text{ and } \boldsymbol{\delta 1}' = 1$$

- This can be solved as a system of equations, or by finding the eigenvector with eigenvalue equal to 1.
- If the initial distribution is the stationary distribution, the chain has the same unconditional distribution at all points, eg $\boldsymbol{u}(2) = \boldsymbol{u}(1)\boldsymbol{\Gamma} = \boldsymbol{\delta\Gamma} = \boldsymbol{\delta}$. This is a stationary Markov chain.

HMM definition

- A Hidden Markov Model (HMM) is a Markov chain in which the sequence of states is not observed but hidden
- Instead of observing the sequence of states, we observe a stochastic function of them called emissions or observations
- Let $X_1, ..., X_T$ denote the (observed) sequence of T emissions and $C_1, ..., C_T$ denote the (hidden) sequence of states
- A HMM is defined by two quantities:
  - The transition matrix $\mathbf{\Gamma}$ of elements γ_{ij} where i and j are states:

$$\gamma_{ij} = p(C_t = j | C_{t-1} = i)$$

  - The emission probabilities $p_i(x)$ where i is a state and x is an emission:

$$p_i(x) = p(X_t = x | C_t = i)$$

- Note that C_t is discrete, but X_t can be discrete, continuous, multivariate, etc.

A first hidden Markov model for the earthquake dataset

- HMM with two states
- Transition matrix:

$$\mathbf{\Gamma} = \begin{pmatrix} 0.94 & 0.06 \\ 0.14 & 0.86 \end{pmatrix}$$

- This gives the stationary distribution:

$$\boldsymbol{\delta} = (0.7, 0.3)$$

- Emission probabilities:

$$p_i(x) = e^{-\lambda_i}\lambda_i^x/x! \text{ with } \lambda_1 = 15 \text{ and } \lambda_2 = 25$$

- Since the emissions are Poisson distributed, this model is called a Poisson-HMM

Application to data

- Exercise
- Let us consider the two-state Poisson-HMM in the previous slide in stationary mode
- Show that

$$\mathrm{E}(X_t) = \delta_1\lambda_1 + \delta_2\lambda_2 = 18$$

- Show that

$$\mathrm{Var}(X_t) = \delta_1(\lambda_1^2 + \lambda_1) + \delta_2(\lambda_2^2 + \lambda_2) - (\mathrm{E}(X_t))^2 = 39$$

- Thus this model is overdispersed compared to the Poisson model, with mean and variance similar to those observed (19.364486 and 51.5734438, respectively)

Link between HMM and mixture model

- This HMM model has the same marginal distributions as our previous independent mixture model
- But the independent mixture model considered that the X_t are independent, whereas in the HMM the autocorrelation of X_t is explicitly modelled
- The HMM model would reduce to the mixture model if we defined:

$$\mathbf{\Gamma} = \begin{pmatrix} 0.7 & 0.3 \\ 0.7 & 0.3 \end{pmatrix}$$

- The HMM is therefore an extension of the mixture model

A bit of history... (1/2)

- Andrey Markov (Russian Empire) studied Markov chains in the early 20^{th} century
- Chapman (British) and Kolmogorov (USSR) independently worked on Markov processes in the 1930s
- Fundamental HMM results were first described by Ruslan Stratonovich (USSR) in Russian publications in the late 1950s and translated to English in 1960
- Thorough analysis of HMM by Leonard Baum (USA) in the second half of the 1960s

A bit of history... (2/2)

- In the 1970s, HMMs became popular for application to speech recognition

  - A speech signal is recorded and divided each small pieces (frames) of ~10 milliseconds
  - Each frame is classified into 256 categories
  - Aim is to recognise the sequence of words being spoken
- Also attracted interest for military applications (eg target tracking)
- In the second half of the 1980s, HMMs began to be applied in biostatistics
- Became hugely popular for DNA analysis in the 1990s and remains ubiquitous since
- Many extensions still under active research (eg SMC)

Conclusions

- A hidden Markov model is made of:
  - A Markov chain of unobserved states
  - Observed emissions that depend on the states
- We introduced HMM as an extension to independent mixture models, but there are many other ways to think about them
- Hidden Markov models arise in many fields of application: computational finance, speech and handwriting recognition, time series analysis, genetic sequence analysis, gene prediction, protein folding,. . .
- Many interesting questions can be asked, but in the next lecture we will focus on how to calculate the likelihood in a HMM model, ie $p(X_1, ..., X_T)$