Master's Thesis

Modelling the Spread of Innovations by a Markov Process in a Bayesian Framework

Niklas Wulkow

Fachbereich für Mathematik und Informatik
Freie Universität Berlin

Supervisor - Prof. Dr. Christof Schütte
Secondary supervisor - Privatdozent Dr. Marcus Weber

Submitted: September 2017
Revised version, February 2018

Eidesstattliche Erklärung:

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Ausführungen, die anderen veröffentlichten oder nicht veröffentlichten Schriften wörtlich oder sinngemäß entnommen wurden, habe ich kenntlich gemacht. Die Arbeit war in gleicher oder ähnlicher Form noch nicht Bestandteil einer Studien- oder Prüfungsleistung.

Berlin, den 06.09.2017, Niklas Wulkow

Acknowledgements:
I would like to thank the members and supervisors of the project I was a part of during the work on this thesis: Luzie Helfmann and Johannes Zonker for their excellent work and our good chemistry, Natasa Djurdjevac Conrad for valuable advice in all aspects and Marcus Weber and Christof Schütte for always knowing exactly what is going on.
Additionally, this project would not have been possible without the archaeological insights from Ana Grabundzija, Brigitta Schütt, Wolfram Schier, Daniel Fürstenau and especially Martin Park.
A special thanks goes out to Ilja Klebanov for some insightful discussions about the field of Bayesian modelling.

Contents

Introduction

This thesis presents a mathematical method to model the spread of an innovation over space and time. According to a common definition by E.M. Rogers, an innovation is an 'idea, practice or object that is perceived as new by an individual or other unit of adoption' [41]. Examples include the wheel or paper (objects), animal domestication (practice) or the thought that the Earth is round instead of flat (idea).
It has rarely been the case that one innovation was made independently at multiple places at the same time. Rather, it originated from one area and spread from there via e.g. trade routes, migration or wars. Reconstructing how exactly an innovation spread is often a very difficult task. Archaeologists have to dig for instances of the innovation and date an instance back to the time when it was created. This information then allows to reconstruct at which time the innovation was known and in use in which area of the world. But, depending on which innovation one discusses, the reliability of those findings should not be overstated for several reasons. One reason is the phenomenon of decaying material as buried indications of the innovation often vanish over the millennia. Another reason is that instances that are found in an area need not have been created there. An example could be an instance of innovative weaponry that was designed in one country, used in a war in a different place in the world and never taken back to its origin. Moreover of course, archaeological excavations cannot be made everywhere.
Therefore, when it comes to the reconstruction of the spreading of a particular innovation there is room for improvement.
An intuitive approach could be a model similar to a common modelling of disease spreading where the concept of 'being infected by the disease' would be replaced by the knowledge or the use of the innovation. As a standard framework [22], people are either susceptible to, infected by or recovered from the disease. The development of the portion of each of these groups over time is then modelled via partial differential equations. We could thereby model how long it takes until a certain portion of the world population has taken up the innovation. Unfortunately, in this framework we would not model how the innovation spreads over space.
The relatively little existing work on the mathematical modelling of innovation spreading includes similar approaches to the disease spreading framework. As a popular tool has emerged the wave-of-advance model [23] which is based on a PDE-system, too, and also takes the spatial component of the spreading into account. Ackland et. al. [40] used the wave-of-advance model for the spreading of farming habits (opposed to hunting). It requires, however, the setting of culture-, biology- and technology-dependent parameters that can clearly vary dependent on the innovation. A still similar but more straightforward approach is the determination of

spreading velocity of the practice of farming measured in distance per time and a subsequent model of the portion of an area that is accustomed to that innovation [38, 39]. The approaches in these sources, however, are restricted to one specific innovation in farming.

In this thesis, we will derive an abstract data-driven probabilistic model that is not restricted to a single specific innovation. For this model, we will divide a part of the Earth into several regions and construct a time-continuous Markov process between them. This process will simulate how the innovation spreads from region to region. Later on we will use data to estimate the underlying parameters of this Markov process and try to reconstruct how a particular innovation spread across the land.

Chapter 1 of this thesis covers some theory about the two major cornerstones of the model. One of them is the concept of time-continuous Markov processes on a discrete state space. The behaviour of such a Markov process can be described by the generator matrix which contains the inverses of the average waiting times until the process transitions from one state to the other. We will eventually try to estimate this object from data and therefore introduce basic theory about parameter estimation. As the result of parameter estimation can be highly non-unique and easily affected by slight modifications of the conditions around it, we will use a Bayesian framework whose basics will be discussed in the first chapter, too.

In Chapter 2 the model will be explained in detail. The regions that we divided the map into will be the only players in that model with the property whether the innovation is known inside of them or not. According to so-called spread rates, they will then 'infect' other regions with the innovation. We will make two important assumptions on that spreading process that will force us to construct a Markov process not on the intuitive state space of regions but on a more complex one. It will then be shown how to estimate those spread rates from archaeological data.

In Chapter 3 we will ultimately apply this model to two different datasets about the spreading of the woolly sheep across Eastern Europe and Western Asia. For that aim we are given archaeological data on several hundred findings of sheep indications. With the use of these data we will try to determine along which route people became acquainted with the practice of keeping woolly sheep and see which type of information the model can and which it cannot deliver.

1 Fundamentals

1.1 Markov Processes

As the spread of innovations is a highly random process this first part of the first chapter is meant to provide knowledge and intuition about fundamentals of Markov processes that will be vital for the machinery used later on. Before we get there, we will start with the very basics of probability theory.

1.1.1 Basic Probability Theory

Let us imagine a people of ants that is on the way to its anthill. One after another they come to a fork of the way with both paths leading to the anthill. Every ant now has to choose between the path to the left or to the right. Without further knowledge of the factors that influence the ants in taking the decision, such as wind or neglecting herd behaviour where one follows the other, i.e. assuming that every ant makes its decision independently and under the same conditions as the other ants, what do we expect how many of the total population of the ant people take the left respectively the right path? That number divided by the number of ants is denoted as the **probability** for an ant to go left respectively right.
Another, much more common example is the following: Consider a dice with six sides. If we perform an experiment and take a high number of rolls, what do we expect will be the ratio between seeing a 1 on top and the number of rolls? Certainly, one would say that this value is 1/6. Hence we have already defined a simple 'model' for the outcome of the roll of a dice. However, if for example the silver dots that represent the number on each side of the dice bias its center of gravity and hence alter the probabilities of the outcome the model would not exactly represent the reality.
Here, one should pay extra attention to the word 'expected'. The ratio of occurrences of an event divided by the total number of tries as observed after performing the experiment is something different. This is usually called 'relative frequency'. Probability refers to our subjective expectation of the ratio.
Besides giving an idea about what probability means, these two examples shall already illustrate that when trying to map natural processes into a mathematical framework, one will usually not be able to comprehend all influences of a model and therefore the model will only live by some assumptions.

We will now formalize the concept of probability.

Definition 1.1 (Probability Space)**.** *A triple* $(\Omega, \mathcal{A}, P)$ *is called a* ***probability space***

if the following conditions hold:
$\mathcal{A} \subseteq 2^{\Omega} = \{A \subseteq \Omega\}$ *is a sigma-algebra over* Ω*, i.e.:*

- $\emptyset, \Omega \in \mathcal{A}$,
- $A \in \mathcal{A} \Rightarrow \Omega \setminus A \in \mathcal{A}$,
- $A_1, A_2, \cdots \in \mathcal{A} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$

and P *is a probability measure, i.e.* $P : \Omega \to [0, 1]$ *with:*

- $P(\Omega) = 1$,
- $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=0}^{\infty} P(A_i)$ *for pairwise disjoint* $A_i \in \mathcal{A}$.

The set Ω denotes all possible events that can occur as the outcome of a probabilistic experiment. $\mathcal{A}$ is a set of sets of those elementary events. As in the dice example it is a sensible question to ask what the probability is for rolling either a 4,5 or 6, $\mathcal{A}$ contains those sets of combined elementary events from Ω that the probability measure P is defined on.

Often one has to deal with a very big event set Ω but is only interested in knowing whether the event is in a certain class of events. An example for that is the lottery where there are millions of possible combinations of numbers but a participant is mostly interested in the money won. Multiple different number combinations yield the same win, i.e. they are mapped to the same value in a different space that contains the potential wins. That leads us to the next definition:

Definition 1.2 (Random Variable)**.** *Given a measurable space* $(E, \mathcal{E})$ *(i.e. a set* E *with a sigma-algebra* $\mathcal{E}$ *on* E*), a* ***random variable*** *on the probability space* $(\Omega, \mathcal{A}, P)$ *is a map* $X : \Omega \to E$ *with*

$$X^{-1}(A) = \{\omega \in \Omega | X(\omega) \in A\} \in \mathcal{A} \; \forall A \in \mathcal{E}. \tag{1.1}$$

For every $A \in \mathcal{E}$ the probability that $X(\omega) \in A$ is then given by $P_X(A) := P(X^{-1}(A))$. Hence a new probability space $(E, \mathcal{E}, P_X)$ is constructed.

The average value that a random variable X produces given the probabilities in P_X is often a quantity of interest. The law of large numbers [4] states that for a sequence of samples from E this value converges to the following:

Definition 1.3 (Expected value)**.** *For a random variable* X *on* $(\Omega, \mathcal{A}, P)$ *the* ***expected value*** *is defined as*

$$\mathbb{E}(X) := \int_{\Omega} X(\omega)P(\omega)d\omega. \tag{1.2}$$

It holds that

$$P(A) = \mathbb{E}(\mathbb{1}_A) = \int_A P(\omega)d\omega, \tag{1.3}$$

where $\mathbb{1}_A$ is the indicator function on A. This puts the way we introduced probability in the beginning of the chapter into a formula.

In general, information about the outcome of an experiment affects the probabilities for its outcome. For example, when rolling a fair six-sided dice, the probability for the dice to show a 2 is 1/6. If we roll the dice so that we do not see which number is on top and someone tells us that it is an even number the chance increases to 1/3 because the result has to come from the set $\{2, 4, 6\}$. We write: $P(A|B) = \frac{1}{3}$ where A is the event $(\omega = 2)$ and $B = (\omega \in \{2, 4, 6\})$.
$P(A|B)$ is called the **conditional probability** of A and B. We will introduce the definition of the conditional probability but before that motivate it with the following illustration:
Recall that if we execute an experiment n times, the probability for the outcome ω to be equal to a specific $\tilde{\omega}$ is the expected ratio of the number of occurrences of $\tilde{\omega}$ divided by the number of executions. Likewise, the probability that $\omega = \tilde{\omega}$ given that $\omega \in B$ (assuming that also $\tilde{\omega} \in B$) is the expected number of occurrences of $\tilde{\omega}$ divided by the expected number of occurrences of B since only those executions have to be taken into account where $\omega \in B$.
For any $\tilde{\omega} \notin B$ the events $(\omega = \tilde{\omega})$ and $(\omega \in B)$ exclude each other, so the probability is 0. In summary, the probability of $(\omega = \tilde{\omega})$ knowing that $\omega \in B$ is given by

$$P(\omega|B) = \begin{cases} \frac{P(\omega)}{P(B)} & , \omega \in B \\ 0 & , \omega \notin B. \end{cases}$$

This yields:

$$P(A|B) = \sum_{\omega \in A \cap B} P(\omega|B) = \sum_{\omega \in A \cap B} \frac{P(\omega \cap B)}{P(B)} = \sum_{\omega \in A \cap B} \frac{P(\omega)}{P(B)} = \frac{P(A \cap B)}{P(B)}.$$

We hence define:

Definition 1.4 (Conditional Probability)**.** *The* ***conditional probability*** *of* $\omega \in A$

given $\omega \in B$ with $P(B) > 0$ is given by:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ for } P(B) > 0. \tag{1.4}$$

*We analogously define the **conditional expected value** of a random variable X by*

$$\mathbb{E}(A|B) = \frac{\mathbb{E}(\mathbb{1}_B X)}{B}. \tag{1.5}$$

Basic probability theory ([5],p.14) yields a helpful result about the conditional probability of an event:

Theorem 1.1 (Law of total probability)**.** *Let A, B be subsets of Ω and let $B_1, B_2, \ldots$ be a countable partition of B where $B_i \in \mathcal{A}$ for all i. Then*

$$P(A|B) = \sum_{B_i} P(A|B_i)P(B_i). \tag{1.6}$$

Note that for $B = \Omega$ this equation reads:

$$P(A|\Omega) = P(A) = \sum_{B_i} P(A|B_i)P(B_i). \tag{1.7}$$

This also implies for the conditional expected value:

Corollary 1.1.

$$\mathbb{E}(A|B) = \sum_{B_i} \mathbb{E}(A|B_i)P(B_i). \tag{1.8}$$

Corollary 1.2.

$$P(A \cap B|C) = P(A|B \cap C)P(B|C) \tag{1.9}$$

Proof.

$$\begin{aligned} P(A \cap B|C) &= \frac{P(A \cap B \cap C)}{P(C)} = \frac{P(A|B \cap C)P(B \cap C)}{P(C)} \\ &= P(A|B \cap C)P(B|C). \end{aligned}$$

□

The scenario $P(B) = 0$, at least in the case that Ω is discrete, does not pose a problem since the question 'What is the probability of A given B if B is impossible?' does not make sense. For a continuous Ω, $P(\omega) = 0$ for elementary events $\omega \in \Omega$. This is a problem if we are given values from a non-discrete Ω, such as the height of a person when we want to compute the probability for the weight of that person to

be in a certain interval given their exact height.
More formally, for a non-discrete set Ω it holds for $\omega, \omega_0 \in \Omega$ and $A \subset \Omega$:

$$P(\omega = \omega_0) = 0 \text{ but in general } P(\omega \in A) > 0. \tag{1.10}$$

For this case the conditional probability can be defined using the conditional expectation:

Definition 1.5 (Conditional expectation)**.** *Given a random variable X on the probability space $(\Omega, \mathcal{A}, P)$ and a sub sigma-algebra $\mathcal{F} \subset \mathcal{A}$ (i.e. a subset of $\mathcal{A}$ that is a sigma-algebra itself), let $Y : \Omega \to \mathbb{R}$ be a random variable defined as $Y = \mathbb{E}[X|\mathcal{F}]$. Y is a **conditional expectation** of X if it fulfills:*

- *Y is $\mathcal{F}$-measurable, i.e. $Y^{-1}(A) \in \mathcal{F}\ \forall A \in \mathbb{R}$,*
- *$\forall A \in \mathcal{F}$ it holds $\mathbb{E}[X\mathbb{1}_A] = \mathbb{E}[Y\mathbb{1}_A]$.*

We can now infer a conditional probability again by:

$$P(A) = \mathbb{E}[\mathbb{1}_A] \Rightarrow P(A|\mathcal{F}) = \mathbb{E}[\mathbb{1}_A|\mathcal{F}]. \tag{1.11}$$

Note that the conditional expected value and the conditional expectation are not the same. The first is a fixed value, the latter is a random variable.
The meaning of $\mathbb{E}[X|\mathcal{F}](\omega)$ is the following:
If after drawing ω from Ω we are given the information whether $\omega \in A$ for every $A \in \mathcal{F}$ what is the expected value of X?
Note that $P(A|\mathcal{F})(\omega)$ is a random variable depending on ω. The difference between this conditional probability and the one that depends on certain events B is that we do not restrict the probability space to B any longer. Instead, we assume ω is drawn and then consider the information in which sets from $\mathcal{F}$ it lies. For $\mathcal{F} = \{\emptyset, \Omega, B, \Omega \setminus B\}$ we ask the question 'Does ω lie in B or does it not?' and only then adapt the probabilities instead of weighting the option that $\omega \in B$ with its probability when this can be 0. If now $P(B) = 0$ then we know that ω will lie in $\Omega \setminus B$ so in this case $P(A|\mathcal{F})(\omega) = P(A)\ \forall \omega \in \Omega$.

Lemma 1.1. *([6], p.189ff.) If $\mathbb{E}[Y] < \infty$, the conditional expectation under $\mathcal{F}$ exists and is almost surely unique.*

Example 1. Let $\Omega = [0, 2\pi]$, P the uniform distribution on $[0, 2\pi]$ (i.e. for an interval $B = [a, b], P(\omega \in B) = \frac{b-a}{2\pi}$), $X(\omega) = sin(\omega)$ and $\mathcal{F}$ be given as $\mathcal{F} = \sigma([0, \pi), [\pi, 2\pi])$ where $\sigma(A_1, \ldots, A_n)$ is the sigma-algebra constructed by the sets $A_1, \ldots, A_n$, i.e. all

unions of those sets and their complements.
Then the first condition on a conditional expectation yields that $\mathbb{E}[X|\mathcal{F}]$ is constant on both intervals because otherwise its preimage would not be a proper subset of either $[0, \pi)$ or $[\pi, 2\pi]$ and therefore not a subset of $\mathcal{F}$. The second condition gives that on both intervals the value of $\mathbb{E}[X|\mathcal{F}]$ is equal to the average value of X on the respective interval. Using the uniform distribution, these are given by $\frac{2}{\pi}$ and $-\frac{2}{\pi}$ (Figure 1.1).

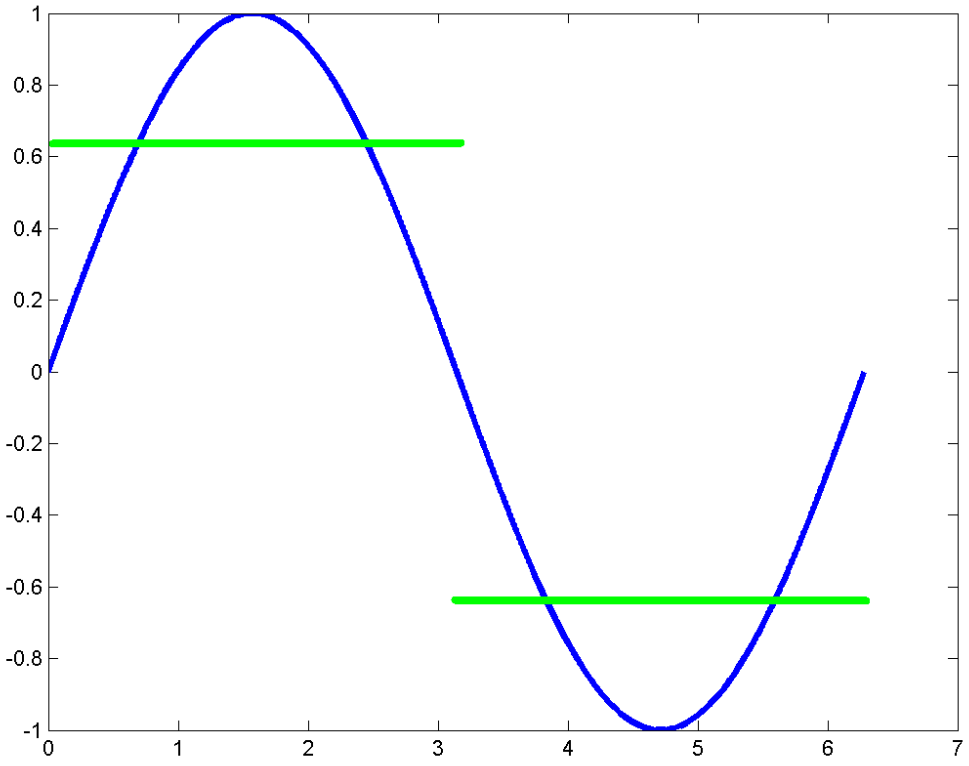


Figure 1.1: Conditional Expectation (green) of the random variable $X(\omega) = sin(\omega)$ (blue).

Definition 1.6 (Independence)**.** *Two events A and B are called **independent** if it holds:*

$$P(A \cap B) = P(A)P(B). \tag{1.12}$$

This also gives $P(A|B) = P(A)$ and $P(B|A) = P(B)$. Two random variables $X : \Omega \to \mathcal{E}_X$ and $Y : \Omega \to \mathcal{E}_Y$ are called independent if for all pairs of events $A \in \mathcal{E}_X, B \in \mathcal{E}_Y$ $X^{-1}(A)$ and $Y^{-1}(B)$ are independent.

So, if A and B are independent, additional information in the form of B resp. A does not change the probability for A resp. B.

1.1.2 Stochastic Processes

We will now expand the concept of random variables and introduce a notion for a sequence $(X_t)_{t\in I}$ of random variables that are in general not independent. Often the index set I can be interpreted as time and every value of the sequence is heavily influenced by the previous ones.
An example is the stock market: The behaviour of stock courses obeys many random factors but the value of a share at time t will usually be around the value it had at a time shortly before t.

Definition 1.7 (Stochastic Process)**.** *Given an index set I and a state space E, a **stochastic process** is a family of random variables*

$$(X_i)_{t\in I},\ X_t : \Omega \to E\ \forall t \in I. \tag{1.13}$$

Consecutive rolls of a dice yield a stochastic process, too, although not a very interesting one, if we assume all rolls to be independent from each other.
A trajectory, i.e. a realisation of the form $(X_{t_1}, X_{t_2}, \dots)$ for a fixed ω of this process, could read $(3, 4, 1, 6, 6, 6, 3, \dots)$ whereas for the stock market it would rather read $(25.5, 25.7, 26.7, 26.6, 26.9, \dots)$.

If a stochastic process has the characteristic that a value is influenced by many of the previous values, the process will both computationally and analytically be difficult to analyse as there will be lots of different cases to distinguish. Does maybe the trajectory $(1, 4, 3)$ until the third value in the index set give different probabilities for the upcoming value than the trajectory $(2, 5, 3)$? For that reason, natural processes are often modelled in a simplified way where only the last value of a trajectory influences the next value. Therefore, we will now define a special group of stochastic processes that are called **Markov processes** and start with the case of a discrete index set. Since P will denote a matrix that will soon be introduced, we will use $\mathbb{P}$ for the underlying probability measure.

Definition 1.8 (Markov chain)**.** *A stochastic process $(X_t)_{t\in I}$ with discrete time space I is called **Markov chain** if it fulfills*

$$\mathbb{P}[X_{k+1} = x_{k+1} | X_k = x_k, \dots, X_0 = x_0] = \mathbb{P}[X_{k+1} = x_{k+1} | X_k = x_k] \tag{1.14}$$

$\forall x_k \in E, k \in I.$

The condition (1.14) is called the **Markov Property**. It means that:

The probabilities for the future of a process depend solely on its present and not

on the past.

We now want to give an analogous definition of a Markov process for a continuous time and state space. The left hand side of (1.14) is an expression that contains the values of the process at all times until a final time. As noted in (1.10), in general a trajectory of the form $(X_s)_{0\leq s\leq t}$ will then have probability 0 which causes problems using the conditional probability on events. So now is the time to apply the conditional probability to sigma-algebras. We need a sigma-algebra that comprises the entire history of a stochastic process. It is given by

$$\mathcal{F}_t = \sigma\{X_s^{-1}(A), s \leq t, A \in \mathcal{E}\}. \tag{1.15}$$

It holds:

- $\mathcal{F}_t \subset \mathcal{A}$,
- $\mathcal{F}_s \subset \mathcal{F}_t$ for $s \leq t$.

We can now define Markov processes on a continuous time and state space:

Definition 1.9 (Markov process)**.** *A stochastic process (X_t) is called **Markov process** if it fulfills:*

$$\mathbb{P}[X_t \in A|\mathcal{F}_s] = \mathbb{P}[X_t \in A|\sigma(X_s)] \text{ with } \sigma(X_s) = \sigma\{X_s^{-1}(A), A \in \mathcal{E}\}. \tag{1.16}$$

We will write:

$$\mathbb{P}[X_t \in A|\mathcal{F}_s] = \mathbb{P}[X_t \in A|X_s]. \tag{1.17}$$

1.1.3 Markov chains

In the analysis of Markov processes there are typical questions that usually are of interest, such as:

- When the process is in state x where is it likely to go next?
- When observing the process at some point in time, what is the probability for the value of the process to be exactly x (or in a set A)?

From now on, we will only consider Markov processes on discrete state spaces because later on in this thesis we will construct a Markov process on a discrete state space. For more information on Markov processes with a continuous state space the reader is referred to [1, 3].
For time-discrete Markov processes, i.e. Markov chains, there is a simple object that helps answering those questions:

Let a Markov chain (and from now on all Markov processes in this thesis) be **time–homogeneous**, i.e. the transition probabilities

$$\mathbb{P}[X_{k+1} = j | X_k = i]$$

from time step k to $k+1$ do not depend on k. Then we can write those probabilities into an $n \times n$-matrix P where n is the number of states in the state space E. Then P is defined by

$$P_{ij} = \mathbb{P}[X_1 = j | X_0 = i]. \tag{1.18}$$

Obviously it must hold:

$$P_{ij} \geq 0, \ \sum_{j \in E} P_{ij} = 1. \tag{1.19}$$

Such a matrix is called a **stochastic matrix**. In order to fully describe the behaviour of the process such a matrix, i.e. the transition probabilities, is already almost enough. We are lacking only one more piece of information. That is: Where does the process start respectively what are the probabilities for every state i that the process starts in i, i.e. that $X_0 = i$?
We will capture those probabilities in the **initial distribution** ν_0:

$$\nu_0(i) = \mathbb{P}[X_0 = i].$$

Note that the values in ν_0 and the values in P do not have anything to do with each other. The information about ν_0 has to come from somewhere else.
In general, by **distribution** we denote a vector ν that has the property

$$\sum_{i=1}^{n} \nu(i) = 1.$$

Every stochastic matrix generates a Markov chain:

Theorem 1.2. *([7], p.69) Given a stochastic matrix P together with an initial distribution ν_0 there is a Markov chain with*

$$\mathbb{P}[X_{k+1} | X_k] = P_{ij} =: p_{ij}, \ \mathbb{P}[X_0 = i] = \nu_0(i). \tag{1.20}$$

We then call this matrix the **transition matrix** of the Markov chain.

Given the initial distribution, one could ask for the probabilities for the process to be in each state after one or multiple time steps. In other words, what is the distribution after some time?
Those distributions can easily be computed using the transition matrix:

Let us define the linear operator $\mathcal{P}$ by:

$$(\mathcal{P}\nu)(j) = \sum_{i \in E} P(i,j)\nu(i). \tag{1.21}$$

$\mathcal{P} : l^1 \to l^1$ where $l^1 = \{\nu : E \to \mathbb{R} | \sum_{i \in E} |\nu(i)| < \infty\}$. With $\mathcal{P}$ being restricted to l^1 it is guaranteed that the right hand side in (1.21) is bounded so that $\mathcal{P}$ is well-defined. This is always the case for finite state spaces anyway but not necessarily for infinite ones.

The operator has the following effect on an initial distribution ν_0:

$$(\mathcal{P}\nu_0)(j) = \sum_{i \in E} P(i,j)\nu_0(i) = \sum_{i \in E} \mathbb{P}[X_1 = j | X_0 = i]\mathbb{P}[X_0 = i] = \mathbb{P}[X_1 = j]. \tag{1.22}$$

In matrix-vector notation this can be written as

$$\mathcal{P}\nu_0 = P^T \nu_0. \tag{1.23}$$

So the application of $\mathcal{P}$ to the initial distribution gives the distribution after one time step. Using the time-homogeneity of the process we can also derive for $\nu_1 := \mathcal{P}\nu_0$

$$(\mathcal{P}\nu_1)(j) = \mathbb{P}[X_2 = j].$$

Iterating this k times gives

$$(\mathcal{P}^k \nu_0)(j) = \mathbb{P}[X_k = j]. \tag{1.24}$$

In summary, $\mathcal{P}$ propagates distributions over time using the transition probabilities of the Markov process that are noted in the transition matrix.

Example 2. Let $E = \{1, 2, 3\}$ and the transition probabilities and initial distribution be given by

$$P = \begin{pmatrix} 1/2 & 1/3 & 1/6 \\ 0 & 1/2 & 1/2 \\ 1/4 & 1/2 & 1/4 \end{pmatrix}, \nu_0 = \begin{pmatrix} 1/3 \\ 1/6 \\ 1/2 \end{pmatrix}.$$

Then

$$P\nu_0 = P^T \nu_0 = \begin{pmatrix} 21/72 \\ 32/72 \\ 19/72 \end{pmatrix}.$$

Under some conditions, $\mathcal{P}\nu$ converges to a fixed point μ of $\mathcal{P}$, that is

$$\mathcal{P}\mu = \mu. \tag{1.25}$$

The application of $\mathcal{P}$ to μ does not change the probabilities for every state in any later time step. For that reason, this distribution is called **stationary distribution**. Such a μ always exists, since it is the right eigenvector corresponding to the eigenvalue 1 of $\mathcal{P}$, i.e. the right eigenvector of P^T. As $P\mathbb{1} = \mathbb{1}$ due to (1.19), P has the eigenvalue 1. That guarantees that also P^T has that eigenvalue.

1.1.4 Time–continuous Markov processes

We will now loosen the restriction that transitions between states can only happen at multiples of a fixed time step. Instead, we now allow them to happen at any time, that is, the process stays in a state for a random time and then transitions to a different state. We thereby add another uncertainty to the behaviour of the process: Not only do we not know where the process will go next but neither when the next transition will occur. As we have seen, the first question can be answered to some degree with the transition probabilities. For the second question we would like to know a distribution about the time that the process stays in one state before it goes elsewhere, too.

Let (X_t) be a homogeneous time-continuous Markov process on a discrete state space that starts in $x \in E$, i.e. $X_0 = x$.
Please note: For the sake of visual clarity, for the rest of this chapter we will use the notation $X(t)$ for the value of the process at time t synonymously with X_t.
We denote by T_x the time at which the process makes a transition away from x and call it the **holding time** of x. Formally:

$$T_x := \min\{t : X(t) \neq x | X(0) = x\} \tag{1.26}$$

In order to figure out the distribution of T_x we make the following observation:

Let $s, t \geq 0$, then

$$\begin{aligned}
&\mathbb{P}[T_x > s+t | T_x > s] \\
&= \mathbb{P}[X(r) = x, r \in [0, s+t] | X(r) = x, r \in [0, s]] \\
&= \mathbb{P}[X(r) = x, r \in [s, s+t] | X(r) = x, r \in [0, s]] \\
&\overset{\text{Markov property}}{=} \mathbb{P}[X(r) = x, r \in [s, s+t] | X(s) = x] \\
&\overset{\text{Time homogeneity}}{=} \mathbb{P}[X(r) = x, r \in [0, t] | X(0) = x] \\
&= \mathbb{P}[T_x > t].
\end{aligned} \tag{1.27}$$

This equality means that the time s that has already passed does not affect the probability that the holding time lasts at least for another additional amount of time t. A distribution with this property is called **memoryless**.
By the definition of conditional probability it follows that:

$$\begin{aligned}
\frac{\mathbb{P}[T_x > s+t]}{\mathbb{P}[T_x > s]} = \mathbb{P}[T_x > s+t | T_x > s] &\overset{(1.27)}{=} \mathbb{P}[T_x > t] \\
\Rightarrow \mathbb{P}[T_x > s+t] &= \mathbb{P}[T_x > s]\mathbb{P}[T_x > t].
\end{aligned} \tag{1.28}$$

If we write $\mathbb{P}[T_x > t]$ as a function $G(t)$, then it becomes evident that

$$G(s+t) = G(s)G(t).$$

G then also fulfills

$$G(t) = G(1)^t \; \forall t \geq 0.$$

It follows that G has to be an exponential function, i.e.

$$G(t) = e^{-\lambda t} \tag{1.29}$$

where $\lambda = -ln(G(1))$.

If $\mathbb{P}[T_x > t] = e^{-\lambda t}$, then $\mathbb{P}[T_x \leq t] = 1 - e^{-\lambda t}$.
Also, $\mathbb{P}[T_x \leq t]$ is the cumulated distribution over the function $\mathbb{P}[T_x = t]$, that is:

$$\mathbb{P}[T_x \leq t] = 1 - e^{-\lambda t} = \int_0^t \mathbb{P}[T_x = s] ds.$$

This yields:

$$\mathbb{P}[T_x = t] = \frac{d}{dt}\mathbb{P}[T_x \leq t] = \lambda e^{-\lambda t}. \tag{1.30}$$

We also demand that the holding time is non-negative. We hence define:

$$\mathbb{P}[T_x < 0] = 0.$$

The distribution of $\mathbb{P}[T_x = t]$ then reads:

$$\mathbb{P}[T_x = t] = \begin{cases} \lambda e^{-\lambda t} & , t \geq 0 \\ 0 & , t < 0. \end{cases} \tag{1.31}$$

for a $\lambda \geq 0$.

This distribution is called the **exponential distribution**.

Note that it holds:

$$\mathbb{E}[T_x] = \frac{1}{\lambda} \tag{1.32}$$

which means, the higher the value of λ the earlier is the holding time expected to be over.

Observe that for non-negative values of t the exponential distribution is actually a monotonically decreasing function (Figure 1.2). This might be counter-intuitive because it indicates that the shorter the holding time, the more probable it is. However, here is where the meaning of memorylessness has to be remembered: Regardless of how much time has passed, the probability that the holding time will be over in time t from that point on is the same. That is to say, the holding time ends inside the time interval $(0, t]$ with probability α. If it does not end inside this interval (with probability $1 - \alpha$) then it ends in the interval $(t, 2t]$ with probability α. Thus the probability for the holding time to end in the interval $(t, 2t]$ is $(1 - \alpha)\alpha$. In general, the probability for the holding time to end in the interval $(kt, (k + 1)t]$ is given by $(1 - \alpha)^k \alpha$ which is monotonically decreasing in k.

Now we know about the distribution of the holding time of a state x but this does

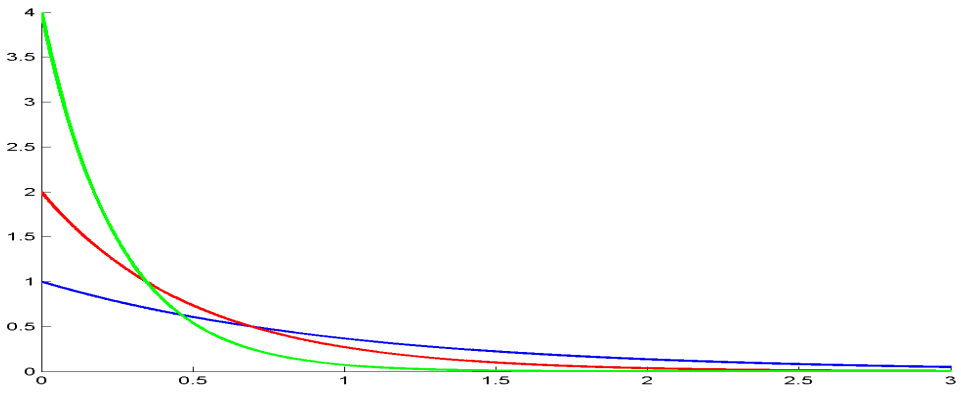


Figure 1.2: Exponential distribution with $\lambda = 1$ (blue), $\lambda = 2$ (red) and $\lambda = 4$ (green).

not yield information about the probabilities of where to go from x when the holding

time is over. We define these values by:

$$p_{xy} := \mathbb{P}[X(T_x) = y | X(0) = x] \; \forall y \neq x.$$

Moreover, we define

$$\lambda(x, y) := p_{xy}\lambda(x),$$

where $\lambda(x)$ is the parameter of the distribution of the holding time of state x and therefore the inverse of the expected holding time as a consequence of (1.32).
Now observe that by the series representation of the exponential function it holds:

$$\mathbb{P}[T_x \leq t] = 1 - e^{-\lambda(x)t} = \lambda(x)t + o(t). \tag{1.33}$$

This helps us to compute

$$\begin{aligned} \mathbb{P}[X(t) = y | X(0) = x] &= \mathbb{P}[T_x < t, X(T_x) = y | X(0) = x] + o(t) \\ &= \mathbb{P}[T_x < t]\mathbb{P}[X(T_x) = y | X(0) = x] + o(t) \\ &= \lambda(x)tp_{xy} + o(t). \end{aligned} \tag{1.34}$$

This results in

$$\mathbb{P}[X(t) = y | X(0) = x] = \lambda(x, y)t + o(t). \tag{1.35}$$

The $o(t)$ captures the probability of two or more transition in the time interval $[0, t]$. The second step uses the independence of T_x and $X(T_x)$ which holds because of the Markov property: If T_x affected the distribution of $X(T_x)$, then that would mean that X would depend on events from the past (namely the waiting time that has passed) which would contradict the Markovianity of X_t.
So, the higher $\lambda(x, y)$ the higher the probability to transition from x to y until a given time t. For small values of t this relation is almost linear. This is why $\lambda(x, y)$ is called the **transition rate** to go from x to y.
Noting that

$$\begin{aligned} &\sum_{y \neq x} \lambda(x, y) = \sum_{y \neq x} \lambda(x) p_{xy} = \lambda(x) \\ &\text{because } \sum_{y \neq x} p_{xy} = 1, \end{aligned}$$

we see that in the case that we only know the rates $\lambda(x, y)$ for all states x and y then not only parameters for the holding time can be computed directly from the $\lambda(x, y)$ but also the transition probabilities by

$$p_{xy} = \frac{\lambda(x, y)}{\lambda(x)}. \tag{1.36}$$

Let us get back to the distribution of the process at any time t. Since in (1.35) we assume a small value for t, we cannot yet predict where the process will be after a longer time. In the time-discrete case we could do that by applying the operator $\mathcal{P}^t$. We will now derive an equivalent object for the time-continuous case using (1.35). Firstly, we define

$$P_{ij}(t) = \mathbb{P}[X(t) = j | X(0) = i]. \tag{1.37}$$

We will now take the derivative of this expression with respect to t, reformulate it and see that it yields a differential equation of which the solution is easy to compute. This will provide us with an operator that gives the values of (1.37) [20].

$$\begin{aligned}
P'_{ij}(t) &= \lim_{h\to 0} \frac{P_{ij}(t+h) - P_{ij}(t)}{h} \\
&= \lim_{h\to 0} \frac{1}{h}(\mathbb{P}[X(t+h) = j | X(0) = i] - \mathbb{P}[X(t) = j | X(0) = i]) \\
&= \lim_{h\to 0}(\sum_{y\in E} \mathbb{P}[X(t+h) = j | X(t) = y, X(0) = i]\mathbb{P}[X(t) = y | X(0) = i] - \\
&\mathbb{P}[X(t) = j | X(0) = i]).
\end{aligned}$$

Now we use some of the previously done work:

$$\begin{aligned}
&\sum_{y\in E} \mathbb{P}[X(t+h) = j | X(t) = y, X(0) = i]\mathbb{P}[X(t) = y | X(0) = i] \\
&= \mathbb{P}[X(t+h) = j | X(t) = j, X(0) = i]\mathbb{P}[X(t) = j | X(0) = i] + \\
&\sum_{y\neq j} \mathbb{P}[X(t+h) = j | X(t) = y, X(0) = i]\mathbb{P}[X(t) = y | X(0) = i] \\
&\overset{(1.33)}{=} (1 - \lambda(j)h)P_{ij}(t) + \sum_{y\neq j} \mathbb{P}[X(t+h) = j | X(t) = y]\mathbb{P}[X(t) = y | X(0) = i] + o(h) \\
&\overset{(1.34)}{=} (1 - \lambda(j)h)P_{ij}(t) + \sum_{y\neq j} \lambda(y,j)hP_{iy}(t) + o(h).
\end{aligned}$$

In the first step all we did was write the j-term of the sum separately. In the second step we used that the probability to be in a state j at time $t+h$ after being there at time t is the complementary probability of the event of staying in j for at least a time of h up to $o(h)$. In the last step we used the Markov property to reduce the condition '$X(t) = y, X(0) = i$' to '$X(t) = y$' and used (1.34).

We apply that in order to reformulate $P'_{ij}(t)$ to:

$$P'_{ij}(t) = \lim_{h\to 0} \frac{1}{h}((1 - \lambda(j)h - 1)P_{ij}(t) + \sum_{y\neq j} \lambda(y,j)hP_{iy}(t) + o(h))$$

which is

$$P'_{ij}(t) = -\lambda(j)P_{ij}(t) + \sum_{y \neq j} \lambda(y,j)P_{iy}(t). \tag{1.38}$$

Let us now define a matrix in which we store the rates:

Definition 1.10 (Generator)**.** *Let X_t be a time-continuous Markov process on a discrete state space. Let transition rates be given by $\lambda(i,j)$. Then the matrix*

$$R_{ij} := \begin{cases} \lambda(i,j) & ,i \neq j \\ -\lambda(i) & ,i = j \end{cases}$$

*is called the **generator** of the Markov process.*

Using the generator (1.38) can be written as

$$P^{'}(t) = P(t)R. \tag{1.39}$$

This is the promised differential equation which has the solution

$$P(t) = P(0)e^{tR} = e^{tR}, \tag{1.40}$$

$P(0)$ is the identity matrix because $P_{ij}(0) = \mathbb{P}[X(0) = j | X(0) = i] = \delta_{ij}$ where δ_{ij} is the Kronecker delta.

Analogously to the time-discrete case, for a distribution ν_0 we can determine the distribution ν_t at time t by

$$\nu_t = P(t)^T \nu_0. \tag{1.41}$$

As a last note here, we must distinguish here between the concepts of distribution and **density**. As mentioned in (1.10), for continuous state spaces $\nu_0(x) = 0$ for elementary events x. We therefore introduce the concept of a density by calling a function $\nu : E \to \mathbb{R}$ a density of the random variable Y if it fulfills

$$\nu(x) \geq 0 \; \forall x \in E \text{ and } \int_E \nu(x)dx = 1. \tag{1.42}$$

Then ν should be specified so that it contains the information

$$\nu(A) := \int_A \nu(x)dx = \mathbb{P}[Y \in A]. \tag{1.43}$$

We write

$$Y \sim \nu. \tag{1.44}$$

1.1.5 Transition Path Theory

This section will explain tools to investigate several more properties of both time-discrete and time-continuous Markov processes. We have derived the probabilities to be in a certain state at a given time. But one could be interested in even more detailed information such as the average time that this state is reached for the first time or how probable it is that a particular different state is reached before it. Again think of the stock market: There those questions can arise in the form: How long will it expectedly take for a stock to reach a certain value? Respectively, is it likely for the stock to be below a value before it rises to a specific height?

We start with answering the first question and define for every subset of the state space the first time that it is reached by the process:

Definition 1.11 (First hitting time)**.** *Let (X_t) be a Markov process with state space E. Then for every $A \subset E$ the **first hitting time** is defined as the random variable*

$$\tau^A := \inf\{t : X_t \in A\}. \tag{1.45}$$

For every single-element set $\{i\}$ we write τ^i instead of $\tau^{\{i\}}$.
The expected value for τ^A is defined as:

Definition 1.12 (Mean first hitting time)**.** *Given that the process starts in state i, the **mean first hitting time** of a set A is defined as:*

$$m_i^A := \mathbb{E}[\tau^A | X_0 = i]. \tag{1.46}$$

Again we write m_i^j for $m_i^{\{j\}}$.
These can be computed by solving a linear system of equations. For the time-discrete case it holds:

Lemma 1.2. *Let (X_t) be a time-discrete Markov process on a discrete state space E. Then the mean first hitting times can be computed by:*

$$m_i^A = \begin{cases} 0 & , i \in A \\ 1 + \sum_{j \in E} m_j^A p_{ij} & , i \notin A. \end{cases} \tag{1.47}$$

Proof. If $i \in A$, then $\tau^A = 0$ so $m_i^A = 0$. For $i \notin A$ we can reformulate m_i^A using the law of total expectation (1.8):

$$\begin{aligned} m_i^A &= \mathbb{E}[\tau^A | X(0) = i] = \sum_{j \in E} \mathbb{E}[\tau^A | X(1) = j] \mathbb{P}[X(1) = j | X(0) = i] \\ &= \sum_{j \in E} (1 + \mathbb{E}[\tau^A | X(0) = j]) p_{ij} = 1 + \sum_{j \in E} m_i^A p_{ij}. \end{aligned}$$

□

In this proof we split m_i^A into two parts: As by assumption $i \notin A$, the process has to take at least one more step to reach A. This gives the 1 in the third step as we assume one time step to have length 1. After this step we have to reevaluate the situation: With probability p_{ij} we will be in state j and the average time to reach A from then on is given by m_j^A.

Theorem 1.3. *Let P be the transition matrix of the Markov chain and Q be a matrix that is constructed by deleting the i-th row and i-th column from P for all states $i \in A$. Then the vector $m^A := (m_i^A)_{i \notin A}$ is given by:*

$$(Id - Q)^{-1} \mathbf{1} = m^A, \tag{1.48}$$

where Id is the identity matrix and $\mathbf{1}$ the vector of ones.

Proof. The equation for m_i^A for $i \notin A$ in (1.47), i.e. for m_i^A, can be written as:

$$m_i^A = 1 + \sum_{j \notin A} m_j^A Q_{ij}.$$

We can write this in matrix-vector-notation as

$$m^A = \mathbf{1} + Q m^A.$$

This is equivalent to

$$(Id - Q)^{-1} \mathbf{1} = m^A.$$

□

Lemma 1.3. *(1.47) is solvable if and only if for every state $j \notin A$ there is a sequence of states $j = j_0, j_1, j_2, \ldots, j_{l-1}, j_l = k$ with $j_1, \ldots, j_{l-1} \notin A$, $k \in A$ with $p_{j_p, j_{p+1}} > 0 \; \forall \; p \in \{0, \ldots, l-1\}$.*

Proof. Firstly, let us note the following:

$$\begin{aligned} &(1.47) \text{ is solvable} \\ \Leftrightarrow\ &Id - Q \text{ is invertible} \\ \Leftrightarrow\ &Id - Q \text{ does not have eigenvalue } 0 \\ \Leftrightarrow\ &Q \text{ does not have eigenvalue } 1. \end{aligned} \tag{1.49}$$

Let j be such that there is no such sequence of states and let us partition E into sets $B := \{i \in E | \text{There is such a sequence}\}$ and $C := \{i \in E | \text{There is no such sequence}\}$. Then $j \in C$ and B and C are separated, i.e. the probability to reach C from B is 0. Thus, in all rows that correspond to entries in C, the entry that is missing in Q compared to P was 0. This means that all row sums of Q corresponding to C are 1 and hence Q has the eigenvalue 1 with the eigenvector v defined by

$$v_i = \begin{cases} 1 & , i \in C \\ 0 & , i \in B. \end{cases}$$

If for every $j \in E$ there is such a sequence then $C = \emptyset$. Let us define $D = \{i \in B | p_{iA} > 0\} \neq \emptyset$ where $p_{iA} := \sum_{l \in A} p_{il}$. For an eigenvector v of Q corresponding to the eigenvalue 1 let us further define $K = \{k \in E | v_k = \max_i v_i\} \neq \emptyset$. For $k \in D$ the sum of the entries of the k-th row of Q is less than 1. Hence K has to be disjoint from D because otherwise for $k \in K \cap D$ the sum of entries of the k-th row of Q would be less than 1 and in consequence the scalar product of the k-th row of Q and v would be less than v_k as

$$\sum_{j=1}^{n} Q_{kj} v_j \leq v_k \sum_{j=1}^{n} Q_{kj} < v_k.$$

On top of that, it has to hold that $Q_{kj} = 0$ for all $j \notin K$ for the same reason. Especially this means that $Q_{kj} = 0$ for all $j \in D$. But this has as a consequence that there is no sequence from any state in K to A. □

In the time-continuous case, that first step away from state i does not necessarily occur after time 1 but after the expiration of the exponentially distributed holding time. This has to be taken into account in order to modify the proof of the time-continuous version of the theorem.

Lemma 1.4. *Let (X_t) be a time-continuous Markov process on a discrete state space E. Then the mean first hitting times can be computed by*

$$m_i^A = \begin{cases} 0 & , i \in A \\ \frac{1}{\lambda(i)} + \sum_{j \in E} m_j^A p_{ij} & , i \notin A, \end{cases} \tag{1.50}$$

which is equivalent to

$$m_i^A = \begin{cases} 0 & , i \in A \\ \frac{1}{\lambda(i)}(1 + \sum_{j \in E, j \neq i} m_j^A \lambda(i,j)) & , i \notin A. \end{cases} \tag{1.51}$$

Proof. For $i \notin A$,

$$\begin{aligned} m_i^A &= \mathbb{E}[\tau^A | X(0) = i] = \sum_{j \in E} \mathbb{E}[\tau^A | X(T_i) = j] \mathbb{P}[X(T_i) = j | X(0) = i] \\ &= \mathbb{E}[T_i] + \sum_{j \in E} \mathbb{E}[\tau^A | X(0) = j] p_{ij} = \frac{1}{\lambda(i)} + \sum_{j \in E} m_j^A p_{ij} \\ &= \frac{1}{\lambda(i)} + \sum_{j \in E, j \neq i} m_j^A \frac{\lambda(i,j)}{\lambda(i)} = \frac{1}{\lambda(i)}(1 + \sum_{j \in E, j \neq i} m_j^A \lambda(i,j)). \end{aligned}$$

□

Let us now state Theorem 1.3 in matrix-vector-notation, too.

Theorem 1.4. *Let R be the generator of the Markov process and let V be a matrix that is constructed by deleting the i-th row and i-th column from R for all states $i \in A$ and setting all entries on the diagonal to 0. Let further denote $\Gamma := (\frac{1}{\lambda(i)})_{i \notin A}$. Then the vector $m^A := (m_i^A)_{i \notin A}$ is given by:*

$$(Id - diag(\Gamma)V)^{-1}\Gamma = m^A. \tag{1.52}$$

$diag(\Gamma)V$ is the matrix of transition probabilities for all states in A as given in (1.36). The only real difference to (1.48) is the replacement of $\mathbf{1}$ by the expected holding times in Γ.
(1.52) is solvable under analogous conditions as (1.48) where the proof follows the same argumentation.

Lemma 1.5. *(1.51) is solvable if and only if for every state $j \notin A$ there is a sequence of states $j = j_0, j_1, j_2, \ldots, j_{l-1}, j_l = k$ with $j_1, \ldots, j_{l-1} \notin A$, $k \in A$ with $\lambda(j_p, j_{p+1}) > 0$ $\forall\, p \in \{0, \ldots, l-1\}$.*

Theorem 1.3 might look like a special case of Theorem 1.4 because if for a given transition matrix we choose $\lambda(i,j) = p_{ij}$ then (1.47) is identical to (1.51). Although the solutions for the mean first hitting times are indeed equal in that case, the setting is different: Theorem 1.4 assumes that the holding time of every state is

exponentially distributed, whereas in Theorem 1.2 the holding time is always fixed to a time step of length 1. So the two Theorems discuss Markov processes of very different characteristics. We should even generalize this statement: A time-discrete Markov process is not a special case of a continuous-time Markov process where we just fix the holding time. A time-continuous Markov process always has exponentially distributed holding times.

Let us now answer the second of the two questions asked at the beginning of this section: What is the probability to reach a set A before another set B?

Definition 1.13 (Discrete forward committor)**.** *For fixed sets A and B, the **discrete forward committor** q_i^+ is the probability to reach B before A when starting in i, that is:*

$$q_i^+ := \mathbb{P}[\tau^B < \tau^A | X(0) = i]. \tag{1.53}$$

Lemma 1.6. *The q_i^+ can be computed by:*

$$\begin{cases} \sum_{j \in E} p_{ij} q_i^+ = 0 & , i \in E \setminus (A \cup B) \\ q_i^+ = 0 & , i \in A \\ q_i^+ = 1 & , i \in B. \end{cases} \tag{1.54}$$

Proof. ([9], p.64) For the proof we first define the probability that the process ever reaches a set:

Definition 1.14 (Hitting probability)**.** *The **hitting probability** of a set A given that the process starts in i is defined as*

$$h_i^A := \mathbb{P}[\tau^A < \infty | X(0) = i]. \tag{1.55}$$

They can easily be computed by:

$$h_i^A = \begin{cases} \sum_{j \in E} p_{ij} h_j^A & , i \notin A \\ 1 & , i \in A. \end{cases} \tag{1.56}$$

This again follows directly from the law of total probability:

$$\begin{aligned} &\mathbb{P}[\tau^A < \infty | X(0) = i] = \mathbb{P}[\tau^A < \infty | X(1) \in E, X(0) = i] \\ &= \sum_{j \in E} \mathbb{P}[\tau^A < \infty | X(1) = j, X(0) = i] p_{ij} = \sum_{j \in E} \mathbb{P}[\tau^A < \infty | X(1) = j] p_{ij} \\ &= \sum_{j \in E} \mathbb{P}[\tau^A < \infty | X(0) = j] p_{ij} = \sum_{j \in E} p_{ij} h_j^A. \end{aligned}$$

The law of total probability is used in the second step. The third step follows from the Markov property.
With $X(1)$ we consider the value of X after one time step, so this proof is only valid for Markov chains. For time-continuous Markov processes, we have to replace the 1 by T_i.

Now to the computation of the committor equations:
Let us pretend that all states in A are absorbing states, i.e. the process can never transition away from any of them. This would mean that the transition rates from those states to any other are 0. We hence have to define a new process $\tilde{X}_t$ with the same transition rates stored in R but set $R_{ij} = 0$ for all $i \in A$ and get subsequent new transition probabilities $\tilde{p}_{ij}$.
Then the statement 'X_t reaches B before A' is equivalent to the statement '$\tilde{X}_t$ ever reaches B'. In other words, the discrete forward committors of X_t are equal to the hitting probabilities $\tilde{h}_j^B$ of $\tilde{X}_t$. It follows:

$$q_i^+ = \begin{cases} \sum_{j \in E} \tilde{p}_{ij} \tilde{h}_j^B & , i \notin B \\ 1 & , i \in B. \end{cases}$$

This is equivalent to

$$\begin{cases} \sum_{j \in E} p_{ij} q_i^+ = 0 & , i \in E \setminus (A \cup B) \\ q_i^+ = 0 & , i \in A \\ q_i^+ = 1 & , i \in B. \end{cases}$$

□

The discrete forward committors are given by the same rule for both time-discrete and time-continuous Markov processes. If we assume the time-continuous setting, then the p_{ij} are again given by the transition probabilities for when the holding time is over. In (1.54) and (1.56) they can be replaced by R_{ij} as for every i R_{ij} is the value of p_{ij} scaled by $\lambda(i)$ and scaling does not change the solutions.

1.2 Parameter estimation and Bayesian modelling

In the first section of this chapter we have introduced the framework of Markov processes and learned how to make predictions about the behaviour of a process given some parameters, namely the transition probabilities or transition rates. We will now reverse this: Given a simulation of a Markov process, what would we guess what those parameters are?

This question, however, addresses only a very special application of a more general framework that is known as **parameter estimation**. In general, one is interested in the following: Given data that is generated according to a probability distribution that is only known up to a parameter θ, how can we estimate θ so that the data fits to that distribution?
In the example of a Markov process, the data could be the portion of transitions from one state to any other and θ would be the underlying transition matrix.

1.2.1 Basics and notation

Let $(\mathcal{Y},\sigma(\mathcal{Y}),\mathbb{P}_{\theta^\star})$ be the probability space of the data, i.e. data are taken from $\mathcal{Y}$ according to the distribution $\mathbb{P}_{\theta^\star}$. Given **D** taken from $\mathcal{Y}$, the goal is to find $\mathbb{P}_{\theta^\star} \in \mathtt{P} := \{\mathbb{P}_\theta, \theta \in \Theta\}$ where $\Theta \subset \mathbb{R}^p$ is the set of all possible parameter values. So, essentially, we are interested in finding $\theta^\star$.
We denote the **estimator** that gets data **D** as input and maps it to the best fitting parameter by

$$\begin{aligned} \varphi &: \mathcal{Y} \to \Theta \\ \varphi(\mathbf{D}) &=: \tilde{\theta}. \end{aligned} \tag{1.57}$$

Please note that in general the $\tilde{\theta} \neq \theta^\star$. But usually the hope is that the data are representative enough for $\mathbb{P}_\theta$ that $\tilde{\theta} \approx \theta^\star$.
We now have to specify the way φ chooses the optimal θ. Thus, we have to introduce a measure for the distance between the data and the expected result if we draw samples from $\mathcal{Y}$ if the true parameter was θ. This distance is called a **loss function**

$$\rho : \Theta \times \mathcal{Y} \to \mathbb{R} \tag{1.58}$$

$$\rho(\theta, \mathbf{D}) = dist(g(\theta), \mathbf{D}). \tag{1.59}$$

$dist$ is some a priori specified distance and $g : \Theta \to \mathcal{Y}$ is a function that maps a parameter value into the observation space so that the data samples **D** and θ can be compared.
Then usually φ estimates $\theta^\star$ by

$$\varphi(Y) = \arg\min_{\theta \in \Theta} \rho(\theta, \mathbf{D}). \tag{1.60}$$

It is worth mentioning that ρ has to be chosen by us. Different functions for it can yield very different estimations. Before we discuss how to pick ρ, let us get familiar with the notation in the following example:

Example 3. We consider the experiment of flipping a potentially manipulated coin n times. We store the result (either 'heads' or 'tails') of every coin flip and calculate the ratio of 'heads' in

$$\mathbf{D} = \frac{\#heads}{n}.$$

Let us now try to infer the probability θ for the event that one coin toss yields the result 'heads'.
A very intuitive approach is to estimate $\theta^\star$ by $\mathbf{D}$ because by the Law of Large Numbers $\mathbf{D}$ will converge to $\theta^\star$. Formalizing this approach using the previously introduced notation, we set

$$g(\theta) = \theta \text{ and } dist(\mathbf{D}, x) = |\mathbf{D} - x|.$$

Then $\rho(\mathbf{D}, \theta) = |\mathbf{D} - \theta|$. This gives that $\varphi(\mathbf{D}) = \mathbf{D}$.

1.2.2 Regression

We now discuss the common case that the data consist of two pieces: **output data** $Y = (y_1, \ldots, y_n)$ and **input data** $X = (x_1, \ldots, x_n)$. We assume there is a rule between the input and the output data given by the function f that maps values from the input space $\mathcal{X}$ to the observation space $\mathcal{Y}$, i.e. it holds:

$$y = f(x) \text{ for } y \in \mathcal{Y}, x \in \mathcal{X}. \tag{1.61}$$

Again, we assume f to be known up to the parameter θ where the true parameter value for θ is $\theta^\star$, i.e. (1.61) should be written as

$$y = f(x, \theta^\star). \tag{1.62}$$

Thus, we try to find $\tilde{\theta}$ such that:

$$y_i = f(x_i, \tilde{\theta}) \; \forall \; i = 1, \ldots, n. \tag{1.63}$$

Often the y_i are subject to small randomized errors, i.e. the relation between $\mathcal{X}$ and $\mathcal{Y}$ is not exactly given as in (1.62) but rather by

$$y = f(x, \theta^\star) + \varepsilon \tag{1.64}$$

where ε is a random term with $\mathbb{E}[\varepsilon] = 0$.
Denoting $\varepsilon_i(\tilde{\theta}) = y_i - f(x_i, \tilde{\theta})$, we therefore want that

$$\|\varepsilon_i(\tilde{\theta})\| \approx 0 \; \forall \; i = 1, \ldots, n.$$

In other words, we want to minimize the distance between Y and $f(X, \tilde{\theta})$ and hope that the data are representative so that the relation given by $f(\cdot, \tilde{\theta})$ is close to the true relation $f(\cdot, \theta^\star)$.
It is hence worthwhile to introduce some sensible distance measures.

Least squares estimation:
When least squares estimation is applied $\tilde{\theta}$ is determined in such a way that the squared errors between the y_i and $f(x_i)$ are minimal:

$$\rho_{LSE}(Y, \theta) := \sum_{i=1}^{n} (y_i - f(x_i, \theta))^2. \tag{1.65}$$

So $\tilde{\theta}_{LSE} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} (y_i - f(x_i, \theta))^2$.

Least absolute deviation estimation:
Here not the squared errors but the absolute errors between the output data and the mappings of the input data are minimized:

$$\rho_{LAD}(Y, \theta) := \sum_{i=1}^{n} |y_i - f(x_i, \theta)| \tag{1.66}$$

and $\tilde{\theta}_{LAD} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} |y_i - f(x_i, \theta)|$.

Example 4. We use the example of the relation of body weight and height of a person to illustrate the effect of both estimates.
Let the input data X denote the height in cm and the output data Y the weight in kg, i.e. given the height of a person we want to estimate its weight. We are given the following values:

$$X = (170, 183, 179, 195, 168),$$
$$Y = (72, 79, 81, 92, 95).$$

We assume here a linear relation $f(x, \theta) = \theta_1 + x\theta_2$. So θ is a 2-dimensional parameter.

Then $\tilde{\theta}_{LSE} \approx (47.9, 0.20)$ and

$$f(X, \tilde{\theta}_{LSE}) \approx (82.0, 84.6, 83.8, 87.0, 81.6)$$

which gives $\rho_{LSE}(Y, \tilde{\theta}_{LSE}) \approx 18.5$.

On the other hand:
$\tilde{\theta}_{LAD} \approx (1, 0.45)$ with

$$f(X, \tilde{\theta}_{LAD}) \approx (76.9, 82.7, 80.9, 88.0, 76.0)$$

and $\rho_{LAD}(Y, \tilde{\theta}_{LAD}) \approx 31.6$.

In $\rho_{LSE}(Y, \cdot)$ large distances between f and Y are punished quadratically. This is why the outlier $(x_5, y_5) = (168, 95)$ is taken into account more and $f(x_5, \tilde{\theta}_{LSE})$ is closer to 95 than $f(x_5, \tilde{\theta}_{LAD})$. However, for all the other values, $f(X, \tilde{\theta}_{LAD})$ yields a better fit.
It should be noted that the values $\rho_{LSE}(Y, \tilde{\theta}_{LSE})$ and $\rho_{LAD}(Y, \tilde{\theta}_{LAD})$ should not be compared because they are evaluations of two different error measures. By definition of $\tilde{\theta}_{LSE}$ and $\tilde{\theta}_{LAD}$ it will hold:

$$\begin{aligned}
&\rho_{LSE}(Y, \tilde{\theta}_{LSE}) \leq \rho_{LSE}(Y, \tilde{\theta}_{LAD}) \text{ and} \\
&\rho_{LAD}(Y, \tilde{\theta}_{LAD}) \leq \rho_{LAD}(Y, \tilde{\theta}_{LSE}).
\end{aligned}$$

Let us introduce one more distance measure:

Maximum likelihood estimation:
This approach demands knowledge about the distribution ν of the errors ε_i. It then answers the question 'Under which parameter is the output data the most probable?'.
Formally, $\nu(\varepsilon)$ is the density that for a randomly drawn (x, y) it holds that $y - f(x, \theta) = \varepsilon$.
The maximum-likelihood loss function is given by

$$\rho_{ML}(Y, \theta) := -\prod_{i=1}^{n} \nu(y_i - f(x_i, \theta)). \tag{1.67}$$

Hence the maximum-likelihood estimator is

$$\begin{aligned}\tilde{\theta}_{ML} &= \arg\min_{\theta\in\Theta} -\prod_{i=1}^{n} \nu(y_i - f(x_i,\theta)) \\ &= \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} \nu(y_i - f(x_i,\theta)).\end{aligned} \tag{1.68}$$

assuming that the y_i are independent of each other.

Lemma 1.7. *In the case of **Gaussian** errors with expected value equal to zero, i.e. $\nu(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}\frac{x^2}{\sigma^2})$ with known σ, the maximum-likelihood-estimate is equal to the least-squares-estimate.*

Proof. The maximizer of a non-negative function is always equal to the maximizer of the logarithm of that function. For $\rho(Y,\theta) = \prod_{i=1}^{n} \nu(y_i - f(x_i,\theta))$ it holds that $log(\rho(Y,\theta)) = \sum_{i=1}^{n} log(\nu(y_i - f(x_i,\theta))$. With ν given as a Gaussian function this means:

$$log(\rho(Y,\theta)) = \sum_{i=1}^{n} log(\frac{1}{\sqrt{2\pi}\sigma}) - \frac{1}{2\sigma^2}(y_i - f(x_i,\theta))^2. \tag{1.69}$$

This is equivalent to minimizing only $\sum_{i=1}^{n}(y_i - f(x_i,\theta))^2$. □

As noted in (1.60), in all these approaches the minimum of the corresponding loss function is of interest. Finding this minimum can be of varying difficulty. We will now discuss a straightforward example.

Example 5. [13] Let the data $(X,Y) = (x_i,y_i)_{i=1,\dots,n}$ be values that are created by taking n points on a polynomial curve of up to degree 4 and slightly shifting them up or down (Figure 1.3). We now want to figure out the underlying polynomial, i.e. what is the polynomial that the data points lie closest to? (Figure 1.3)
Then $f(x,\theta) = \theta_0 + x\theta_1 + x^2\theta_2 + x^3\theta_3 + x^4\theta_4$. Here, we use the least-squares-estimation and write

$$f(X,\theta) = A\theta$$

where

$$A = \begin{pmatrix} 1 & \dots & x_1^4 \\ \vdots & \ddots & \vdots \\ 1 & \dots & x_n^4 \end{pmatrix}.$$

Then $\rho_{LSE}(Y,\theta) = \sum_{i=1}^{n}(y_i - f(x_i,\theta))^2 = \|Y - A\theta\|_2^2 = (Y - A\theta)^T(Y - A\theta)$.

Minimizing this expression is equivalent to solving

$$\nabla_\theta \rho_{LSE}(Y, \theta) = 0$$

which is

$$(A^T A)\theta - A^T Y = 0.$$

So

$$\theta = (A^T A)^{-1} A^T Y.$$

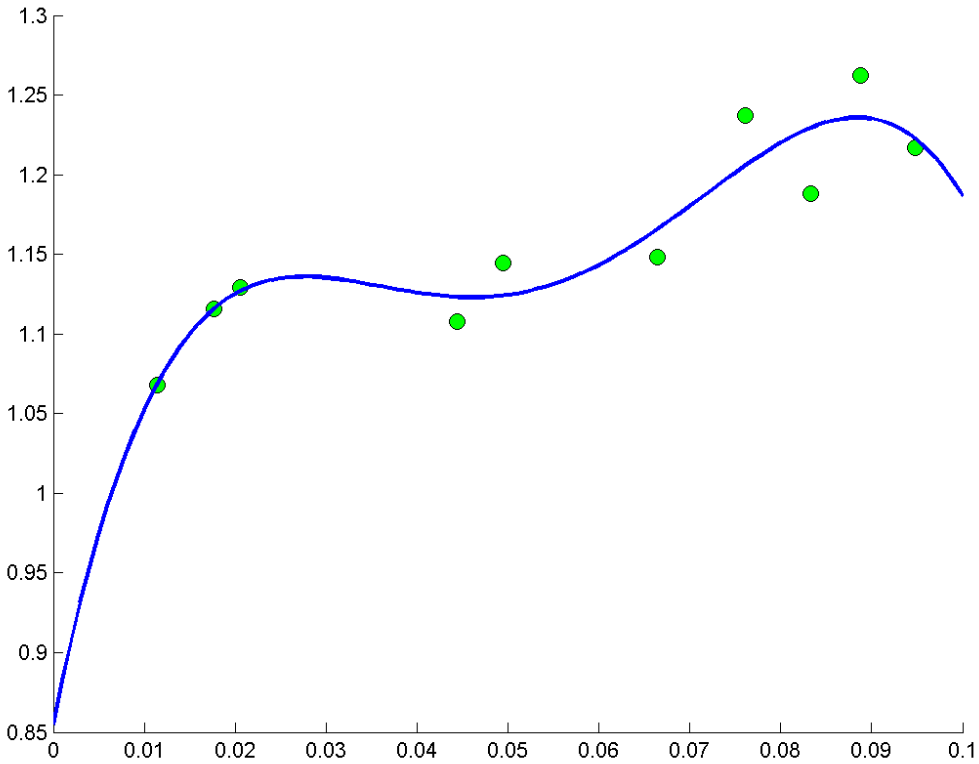


Figure 1.3: Least squares minimizing polynomial (blue) of degree 4 for 10 points (green).

Apparently, the matrix $A^T A$ has to be inverted. For some matrices this can be a very ill-conditioned task, i.e. small changes in A yield a very different result for the inverse. As A consists of the input data X, this gives reason to worry that perturbed data, be it input or output data, can severely change the result of the parameter estimation. An even simpler example illustrates this.

Example 6. [12] Consider the problem

$$A\theta = Y \tag{1.70}$$

where A is a 5×5-matrix given by $A_{ij} = \frac{1}{i+j-1}$, i.e.

$$A = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} \end{pmatrix}.$$

Assuming that not even errors ε_i are given,

$$\theta = A^{-1}Y.$$

For $Y = Y_1 = (2.2813, 1.4490, 1.0922, 0.8840, 0.7452)^T$ we get

$$\theta_1 = A^{-1}Y_1 = (0.9982, 0.9992, 0.9997, 1.0003, 1.0010)^T.$$

If now Y_1 is slightly perturbed to $Y_2 = (2.283, 1.450, 1.093, 0.885, 0.746)^T$, then

$$\theta_2 = A^{-1}Y_2 = (2.105, -20.28, 94.29, -141.4, 71.19)^T.$$

The estimation is then still executed in the correct way. But often times due to prior knowledge one knows that θ has to lie in a certain region, e.g. around $(1, 1, 1, 1, 1)^T$. Then the latter result of the parameter estimation is unreasonable and the former would still make for a very good fit because

$$A\theta_1 = Y_1 \approx Y_2.$$

In the next section we will get familiar with a framework that takes account for such prior knowledge and instead of finding the single best fitting parameter computes a distribution of the probabilities for every parameter to be the true underlying parameter $\theta^\star$.

1.2.3 Bayesian modelling

This approach is based on the following observation:

Theorem 1.5 (Bayes' Theorem). *Let A and B be two events with $\mathbb{P}[B] > 0$. Then it holds:*

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}. \tag{1.71}$$

Proof.

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \frac{\frac{\mathbb{P}[A \cap B]\mathbb{P}[A]}{\mathbb{P}[A]}}{\mathbb{P}[B]} = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}$$

□

We now translate this to the parameter estimation problem and denote by A the event that $\theta^\star = \theta$ and by B the event that the data **D** is given by D. For the case of continuous Θ or $\mathcal{Y}$ we cannot work with the expression $\mathbb{P}[A|B]$. Therefore, let us define the following densities in the notation of (1.44):

$$\begin{aligned}\theta^\star|\mathbf{D} = \mathsf{D} &\sim p(\cdot|\mathbf{D}),\\ \mathbf{D}|\theta^\star = \theta &\sim l(\cdot|\theta),\\ \theta^\star &\sim \pi,\\ \mathbf{D} &\sim d.\end{aligned}$$

We then get:

$$p(\theta|\mathbf{D}) = \frac{l(\mathbf{D}|\theta)\pi(\theta)}{d(\mathbf{D})}. \tag{1.72}$$

$l(\mathbf{D}|\theta)$ is the density of the event that the data **D** will occur if the true parameter is θ. We call this function the **likelihood** of the data.
But $\pi(\theta) \sim \mathbb{P}[\theta^\star = \theta]$ also features in the expression. This comes from our estimate about what θ is. We have to remember the intuition about probability that we introduced earlier.It suggests that probability is always subject to one's point of view because otherwise we would have to realise that

$$\mathbb{P}[\theta^\star = \theta] = \begin{cases} 1 & ,\theta = \theta^\star \\ 0 & ,\text{else.}\end{cases}$$

Here we have to use that prior estimate of ours in the form of a probability distribution of θ that is called the **prior** denoted by π. It denotes the probability that we assign to every θ that it is equal to $\theta^\star$ without any further information given.
$d(\cdot)$ poses a similar problem but we can simply avoid that by using the fact that $\int\limits_\Theta \mathbb{P}[\theta|\mathbf{D}]d\theta = 1$: We ignore d in the computation of $p(\theta|\mathbf{D})$ and later on normalise $p(\theta|\mathbf{D})$.
This enables us to derive a probability distribution over all $\theta \in \Theta$:

$$p(\theta|\mathbf{D}) = \frac{l(\mathbf{D}|\theta)\pi(\theta)}{\int\limits_\Theta l(\mathbf{D}|\theta)\pi(\theta)d\theta}. \tag{1.73}$$

This distribution is called the **posterior** distribution.

Example 7. ([10], p.185) Let us consider a so-called Bernoulli experiment of flipping a coin ten times. We are interested in the probability $\theta^\star$ for the coin to show heads and use the prior estimate that $\mathbb{P}[\theta^\star = 0.4] = 0.25$, $\mathbb{P}[\theta^\star = 0.5] = 0.5$ and $\mathbb{P}[\theta^\star =$

$0.6] = 0.25$. For some reason, we are sure that any other value for $\theta^\star$ is impossible. Let us assume that the coin shows heads all ten times, i.e. $\mathbf{D} = (\text{'heads'}, \ldots, \text{'heads'})$. Since the space of possible values for θ and $\mathbf{D}$ is finite we do not have to use densities here but can formulate the probabilities for the elementary events $\theta^\star = 0.4$, $\theta^\star = 0.5$ and $\theta^\star = 0.6$.
For every value of θ, the probability to see heads after every flip is given by

$$\mathbb{P}[\mathbf{D}|\theta] = \theta^{10}.$$

We can now compute the posterior to be:

$$\mathbb{P}[\theta^\star = 0.4|\mathbf{D}] \approx 0.01, \mathbb{P}[\theta^\star = 0.5|\mathbf{D}] \approx 0.24, \mathbb{P}[\theta^\star = 0.6|\mathbf{D}] \approx 0.75$$

The probabilities are shifted further towards the highest value 0.6. This makes sense because the lower the probability for heads the more unlikely it is to see heads ten out of ten times.

We will now see where the Bayesian framework can be helpful compared to the classical parameter estimation where only the optimal solution is of interest as presented in the previous section.

Example 8. [13] Let us go back to Example 6. We observed there that for two sets of output data Y_1 and Y_2 with very little difference between them the optimal solutions for the parameters θ_1 and θ_2 can be significantly different although θ_1 still fits to Y_2 very well.
In order to apply the Bayesian approach we have to specify a prior distribution and a likelihood function: Assume for some reason it is sensible to expect the optimal parameters to be around $\theta_0 = (1, 1, 1, 1, 1)^T$ and hence define the prior

$$\pi(\theta) = \frac{1}{Z_\pi} \exp(-\frac{1}{2}(\theta - \theta_0)^T(\theta - \theta_0))$$

where Z_π is a normalisation constant so that $\int_\Theta \pi(\theta)d\theta = 1$.
We define the likelihood l in a similar way by

$$l(\theta|Y) = \frac{1}{Z_l} \exp(-\frac{1}{2}(A\theta - Y)^T(A\theta - Y)).$$

Then the posterior reads

$$p(\theta|Y) = \frac{1}{Z_p} \exp(-\frac{1}{2}((\theta - \theta_0)^T(\theta - \theta_0) + (A\theta - Y)^T(A\theta - Y)) \tag{1.74}$$

where again Z_l and Z_p are normalisation constants.
This gives the following expected value of θ according to the posterior. With again $Y_2 = (2.283, 1.450, 1.093, 0.885, 0.746)^T$:

$$\mathbb{E}[\theta|Y_2] \approx (0.78, 0.38, 1.03, -0.38, 2.60)^T.$$

It holds

$$A\mathbb{E}[\theta|Y_2] \approx (1.7383, 1.315, 0.8691, 0.7134, 0.6079)^T.$$

So we have found an estimate for $\theta^\star$ that takes into account the information that $\theta^\star$ should be close to $(1,1,1,1,1)^T$ and still $A\mathbb{E}[\theta|Y_2]$ is not too far away from Y_2.

A downside of this approach is the cost of the computation of the posterior. Technically, for every $\theta \in \Theta$ the likelihood and the prior value at this point have to be calculated. Especially in high-dimensional spaces this poses a problem [14].
Here $\mathbb{E}[\theta|Y_2]$ was computed by approximating

$$\mathbb{E}[\theta|Y_2] = \int_\Theta \theta' p(\theta'|Y_2) d\theta'$$

by a Monte Carlo method [17].

If we do not have any prior information, the Bayesian approach can still be applied and becomes surprisingly simple: We can use the uniform distribution over Θ and derive

$$p(\theta|\mathbf{D}) = \frac{l(\mathbf{D}|\theta)\frac{1}{\mu(\Theta)}}{\int\limits_\Theta l(\mathbf{D}|\theta)\frac{1}{\mu(\Theta)}d\theta} = \frac{l(\mathbf{D}|\theta)}{\int\limits_\Theta l(\mathbf{D}|\theta)d\theta} \tag{1.75}$$

where $\mu(\Theta) = \int\limits_\Theta 1 d\theta$.
So here the posterior distribution is equal to the normalised likelihood.

1.2.4 The Metropolis–Hastings algorithm

In the last example from the previous section we could actually compute the posterior analytically in (1.74). If we are interested in the probability of θ according to the posterior p, we simply have to evaluate $p(\theta)$. But this requires knowledge about the normalisation constant Z_p, i.e. about $\int\limits_\Theta l(\mathbf{D}|\theta)\pi(\theta)d\theta$, which can yield a very high computational effort especially in high dimensional spaces. It is then often preferable to approximate the posterior distribution by 'sampling' from it. That means that we randomly choose elements from Θ, the 'samples', that should be distributed approximately according to the distribution. From these samples we can than get an impression about the distribution. In this section an effective way to

do that is presented, that is the **Metropolis–Hastings algorithm** [15, 16]. The

Algorithm 1: Metropolis-Hastings algorithm

input : A distribution p, number of iterations *numIter*
output: Samples that are distributed according to p

Pick a **proposal distribution** $q : \Omega \times \Omega \to [0, 1]$
Pick a starting point x_0
for *i = 1:numIter* **do**
  Proposal step: Pick x by $q(\cdot|x_{i-1})$
  Acceptance step: Compute $\alpha(x_{i-1}, x) = \min\{1, \frac{q(x_{i-1}|x)p(x)}{q(x|x_{i-1})p(x_{i-1})}\}$
  Draw u from $\mathcal{U}([0, 1])$ (uniform distribution on $[0, 1]$)
  if $\alpha(x_{i-1}, x_i) > u$ **then**
    $x_i = x$
  else
    $x_i = x_{i-1}$

algorithm produces samples x_i from a set Ω that are distributed according to p.
It consists of two main steps: The proposal and the acceptance step.
For every x_{i-1} the **proposal step** gives a sample x from an arbitrarily chosen distribution $q(\cdot|x_{i-1})$.
The **acceptance step** then decides whether this sample is included in the output. This is done via the **acceptance probability**

$$\alpha(x, y) = \min\{1, \frac{q(x|y)p(y)}{q(y|x)p(x)}\} \tag{1.76}$$

that guarantees that all the samples in the output are (approximately) distributed by p.
Naturally the question arises: Why does the algorithm work, i.e. why are the samples in the output distributed by p?

To answer this question we have to leave the Bayesian framework for a moment and go back to Markov chains. This algorithm was actually designed to produce Markov chains with a given stationary distribution (1.25). We make use of it here by constructing a Markov chain whose stationary distribution is the desired posterior distribution p. This means if we sample from that Markov chain after the algorithm has finished, the probability for a drawn x_i to be in $A \subset \Omega$ will be equal (or very close) to $p(A)$. As we are now in the Markov process setting again we will write E instead of Ω.
It can be seen in Algorithm 1 that the transition probability from $x_{i-1} = x$ to $x_i = y$

is given by

$$L(x,y) = \begin{cases} q(y|x)\alpha(x,y) & x \neq y \\ 0 & x = y \end{cases} \tag{1.77}$$

Hence the probability to remain at x is

$$r(x) = 1 - \sum_{y \in E} L(x,y). \tag{1.78}$$

This allows us to write the transition probabilities $P(x,y)$ for this Markov chain as

$$P(x,y) = q(y|x)\alpha(x,y) + r(x)\delta_{xy}. \tag{1.79}$$

Noting that L fulfills the **detailed balance** property

$$p(x)L(x,y) = p(y)L(y,x) \tag{1.80}$$

(proof below) we can see that p is in fact the stationary distribution of the Markov chain:

$$\begin{aligned} \sum_{x \in E} P(x,y)p(x) &= \sum_{x \in E} L(x,y)p(x) + r(x)\delta_{xy}p(x) \\ &= \sum_{x \in E} L(y,x)p(y) + r(x)\delta_{xy}p(x) \\ &= (1 - r(y))p(y) + \sum_{x \in E} r(x)\delta_{xy}p(x) \\ &= (1 - r(y))p(y) + r(y)p(y) \\ &= p(y). \end{aligned}$$

This is exactly the condition for p to be the stationary distribution of the Markov chain with transition matrix P as noted in (1.25).

(1.80) holds because of the following observation:

$$\begin{aligned} p(x)L(x,y) = p(x)q(y|x)\alpha(x,y) &= p(x)q(y|x)\min\{1, \frac{q(x|y)p(y)}{q(y|x)p(x)}\} \\ &= \min\{p(x)q(y|x), p(y)q(x|y)\}. \end{aligned}$$

Analogously:

$$\begin{aligned} p(y)L(y,x) &= \min\{p(y)q(x|y), p(x)q(y|x)\} \\ &\Rightarrow p(x)L(x,y) = p(y)L(y,x). \end{aligned}$$

In (1.76) $p(y)$ is divided by $p(x)$. This means that p does not need to be normalized as the normalisation constants of p in the denominator of (1.73) would cancel out here. For that reason, when trying to sample from a posterior that was constructed by the Bayesian approach under the data $\mathbf{D}$, we can simply compute $\mathbb{P}[\mathbf{D}|x_i]\pi(x_i)$ for every step of the algorithm and use that for p.

Example 9. Let us try to approximate the unknown posterior p on $E = \{1, 2, 3\}$. We apply the Metropolis-Hastings-algorithm with the proposal distribution $q(\cdot|x) = \mathcal{U}(\{1, 2, 3\})$ for all $x \in E$ and get a Markov chain of length 10000 on E. Say, out of those 10000 values 5034 are ones, 3281 are twos and 1686 are threes. This suggests that

$$\begin{aligned} p(1) &\approx \frac{5034}{10000} \approx \frac{1}{2}, \\ p(2) &\approx \frac{3281}{10000} \approx \frac{1}{3} \\ p(3) &\approx \frac{1686}{10000} \approx \frac{1}{6}. \end{aligned}$$

An excerpt of the produced Markov chain is given in Figure 1.4.

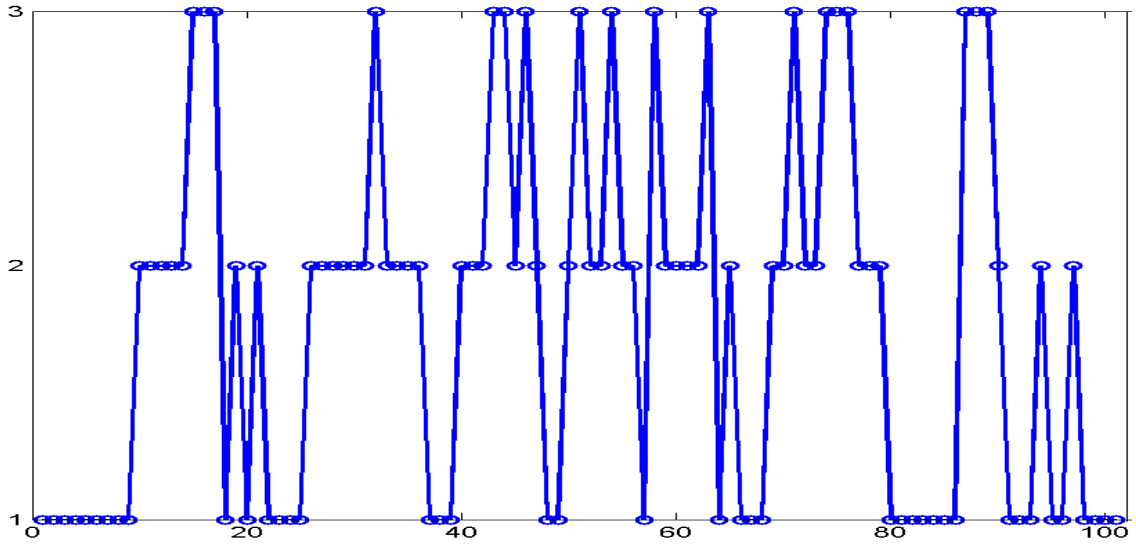


Figure 1.4: First 100 values of the Markov chain from the Metropolis-Hastings algorithm on $E = \{1, 2, 3\}$. The chain tends to stay in a state 1 because the acceptance probability (1.76) is low for transitions away from 1. This is due to the relatively high value of $p(1)$.

2 Modelling the spread of an innovation

In the following main part of this thesis we will develop a workflow of the modelling of the spread of an innovation over time. A cooperation between the Zuse Institut Berlin (ZIB) and TOPOI (The Formation and Transformation of Space and Knowledge in Ancient Civilizations) that was established in 2016 aims to reconstruct the way in which the woolly sheep spread across Eastern Europe and Western Asia. The woolly sheep is itself not really an innovation that was made by people but rather the result of a genetic mutation that came in handy for the people of ancient times.

When we say 'it spread', one could ask us to specify whether we mean the sheep spread on their own or whether they were lead by people, more precisely: farmers. On top of that, when we say that not only the woolly sheep but any innovation spreads, do we mean it spreads physically or is it rather the idea that spreads across the land?

When simulating the spread of the woolly sheep, we will assume they are lead by farmers and in fact spread physically but the model we will derive works without that information and only lives in an abstract setting where we do not need to specify in which sense the spreading takes place in a very detailed way.

Some meaningful inventions in human history

Over the course of this thesis there will be several excursuses over the history and spreading of innovations in human history. We start with a list of some of the most important ones.

Fire: Stone Age [28] - Presumably the first use of fire was made by maintaining bush fires. The oldest proof of man-made fire goes back to 790.000 BC in Israel.

Rope: 28.000 BC [27] - First very thin documents of rope can be dated back to multiple thousand years ago in Europe. But only around 4000 BC the Egyptians started to invent tools to produce ropes.

Copper smelting: 5000 BC [26] - Smelting of copper requires a temperature of 1100°C. This was first done in Serbia.

Wheel: 3500 BC [25] - Although humans had found out that objects can be moved more easily if they are round in the Stone Age, the first wheel out of potter was manufactured in 3500 BC in Mesopotamia and the first wooden wheel in 3200 BC in Slovenia.

Alphabet: 1800 BC [33] - The alphabet as a convention of elementary ingredients of

written communication was invented by Semitic workers in Egypt who adapted the Egyptian version that used a mixture of logographic, syllabic and alphabetic hieroglyphs.
Paper: 150 AD [29] - Paper was first developed in Western China before it spread over Asia, Arabia and Europe within the next 1500 years.
Crane: 500 BC [24] - The invention of the crane came in helpful in Greece in the building of temples.
Sulfur match: 900 AD [30] - Shortly before their first use of gun powder in 1000 AD, the Chinese invented an easy method to ignite fire.
Eyeglasses: 1300 AD [31] - After it had been noticed much earlier that objects appear larger when observed through water than through air it took until late in the Middle Age that eyeglasses were invented in Italy.
Mercator projection: 1569 AD [32] - The Mercator projection which allows to display a map of the world in 2D was compiled by the Flemish geographer Gerardus Mercator.

The cooperation between ZIB and TOPOI includes work on two different approaches. The results of both approaches depend on parameters that are estimated from data which is provided by TOPOI.

2.1 The agent-based approach

One approach [21] that is only touched on here, is **agent-based**. It contains the implementation of a model that simulates multiple thousand farmers (the 'agents') walking the landscape and taking decisions about where to go next individually. Agents can carry woolly sheep with them and thereby spread it across the land. When an agent that does not have a woolly sheep with him meets an agent that does then the former agent can adopt the innovation. More precisely, the properties of those agents are:

- Each agent represents a group of people that herd animals
- Each agent has a position state in $\mathbb{R}^2$
- Each agent has an information state in $\{0, 1\}$ where 1 means that the agent has woolly sheep and 0 means it does not
- Agents can pass the innovation of herding woolly sheep on to other agents if they are close to each other in distance

The movement of the agents is governed by two influences:

- Every point on the land has an attraction value and agents are driven towards points with high attraction value

- Agents are interested in joining an existing group of agents that is close to them in distance

Figure 2.1 illustrates how the agents distribute themselves over the land. It can be seen how the agents avoid several patches of the land and how they gather in certain areas. The red dots in Figure 2.1 denote agents that carry woolly sheep with them, the blue dots denote the ones that do not.

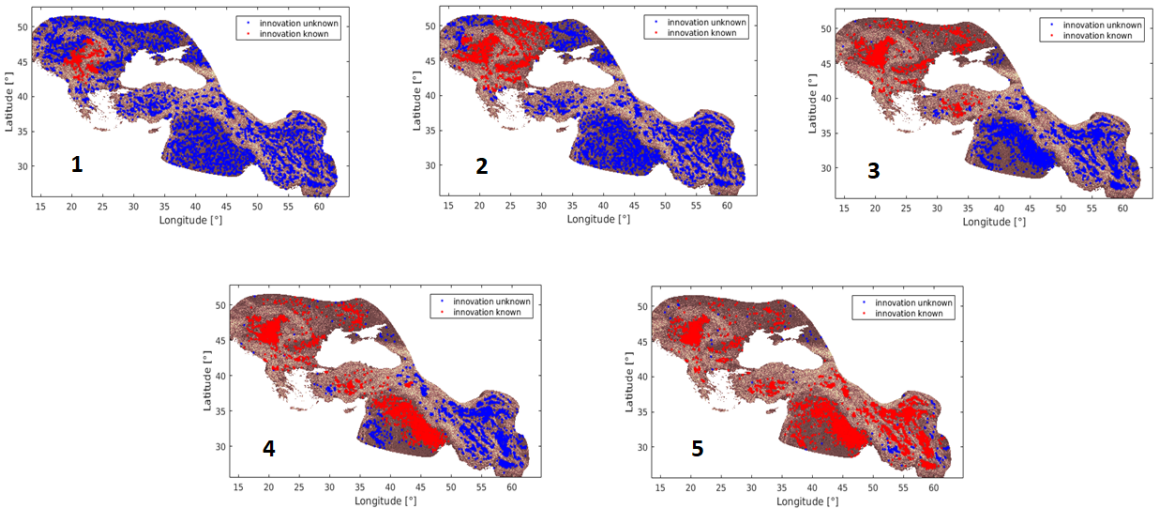


Figure 2.1: Simulated spreading of the woolly sheep with the agent-based approach over time. The blue dots represent agents without the woolly sheep. The red dots represent agents with the woolly sheep. The agents seek attractive areas. After being distributed randomly across the land in the beginning they gather in specific areas. The woolly sheep is spread over almost all agents and thus almost the entire land in the end.

Every step of an agent depends only on the current position and information states of itself and agents nearby so the agents follow a Markov process. The exact decision making process of the agents depends on parameters that are estimated by comparing the times when the agents reach certain points on the map in the model with the TOPOI data.
The second approach that is featured in this thesis is in some ways similar to the agent-based approach as it will also construct a Markov process on a state space of positions. It is, however, much more abstract and has a far smaller position space. For reasons that will become clear later on, we call this approach **network-based**.

2.2 The network-based approach

We divide the map into n connected **regions** (Figure 2.2). Every one of those regions will only have one property: Does it **have the innovation** or not?
In this model we do not specify any further what 'having the innovation' or in other

words 'being **infected** by the innovation' or also that 'the innovation is known in a region' means in detail. To get back to the agent-based approach, it could mean that all the agents that are in a certain region have a woolly sheep with them but we assume a level of abstractness in the model that allows us to leave this question open for now. It will become important, however, much later on.

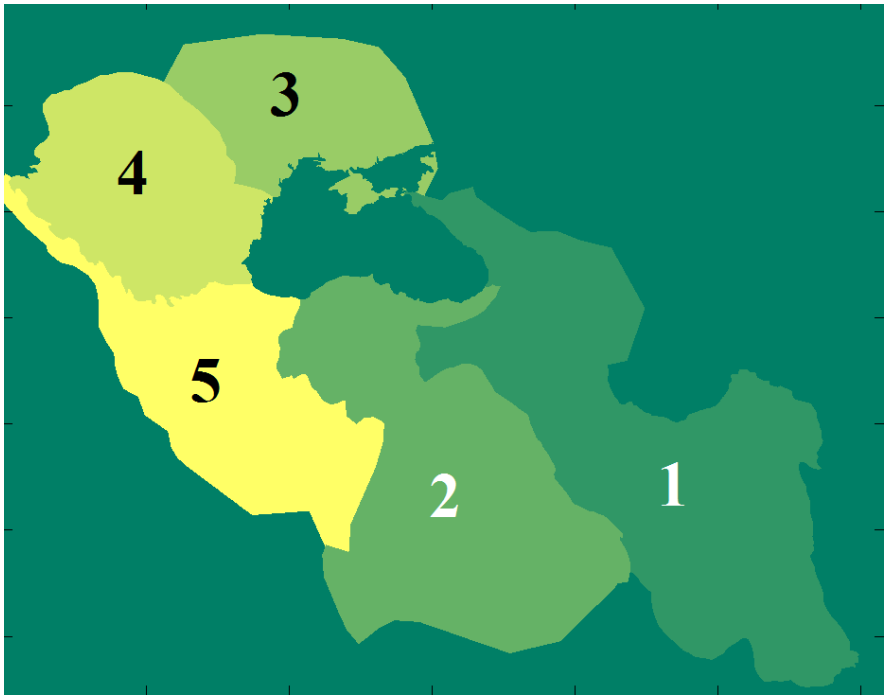


Figure 2.2: Division of the land into five regions. The darkest green patch does not belong to that area of interest.

We assume interaction between regions: By interaction we mean that an infected region infects a non-infected region after a random time, i.e. the innovation is passed on from region to region.
The essence of this approach is now a time-continuous Markov process on a discrete state space of regions where every state represents a region. The process is said to be in a state if the corresponding region has the innovation.
As explained earlier, every time-continuous Markov process has underlying transition rates that we capture in the $n \times n$-matrix R. In this setting, these transition rates then govern the time that passes until the innovation is transferred from one region to another.
Without further assumptions, such a Markov process could look like this:

Figure 2.3: Potential realisation of a Markov process on the state space of regions. If a region is coloured in red and white it means that the process is currently in that region.

This contradicts two important and intuitive assumptions that have to be made about the spread of an innovation and that need to be incorporated into the model:

1. The innovation can be present in multiple regions at the same time.
2. If the innovation is known in a region once it will never leave it.

Hence the following series of images gives a more sensible realisation:

Figure 2.4: More sensible simulation of the spreading of an innovation. If the process is in a region once, i.e. that region is infected once, it stays there forever. If a region is coloured in red and white it means that the innovation is known in that region.

Describing such a spreading process by a stochastic process is a difficult task because if the process was in two states at the same time it would not be well defined any more.
One option to tackle this problem is the following:

2.2.1 Idea 1: A non–Markov process

Naturally, when considering a Markov process one wants to have an intuition about what it means that 'the process is in a state'. In order to construct a process between the regions, let us try the following:
Let us comprehend the value of the process at time t to be the position of a little fairy that transports the innovation over the land. Noone but her can bring the innovation to a region but she obeys the calls of infected regions to go and travel to a non-infected region to bring the innovation there. Once she gets to a region that region is infected and the fairy waits there until she is called away to another region. So the value of the process X_t represents

$$X_t = \text{Region the fairy is in at time t}$$

where at all times $\tau^i = \inf\{t : X_t = i\}$ the fairy brings the knowledge about the innovation to a region and at all other times is set in a region waiting for a call to go

somewhere else.
Again we assume transition rates between the regions that determine how long it will take until an infected region orders the fairy to visit a particular non-infected region.
Then all regions that the little fairy has visited should be marked as infected. When simulating where she goes next it therefore has to be taken into account where she was before because:

- She is never sent to already infected regions. Hence when the fairy visits a region all rates towards this region should be set to 0
- The more regions are infected the faster non-infected regions will be infected because more regions can order the fairy to go somewhere

Thus the complete history of the process is important. As a consequence, X_t is not a Markov process. This can make the process very difficult to predict and analyse and gives us a reason to try to model the spreading process in a different way so that we can treat it as a Markov process.

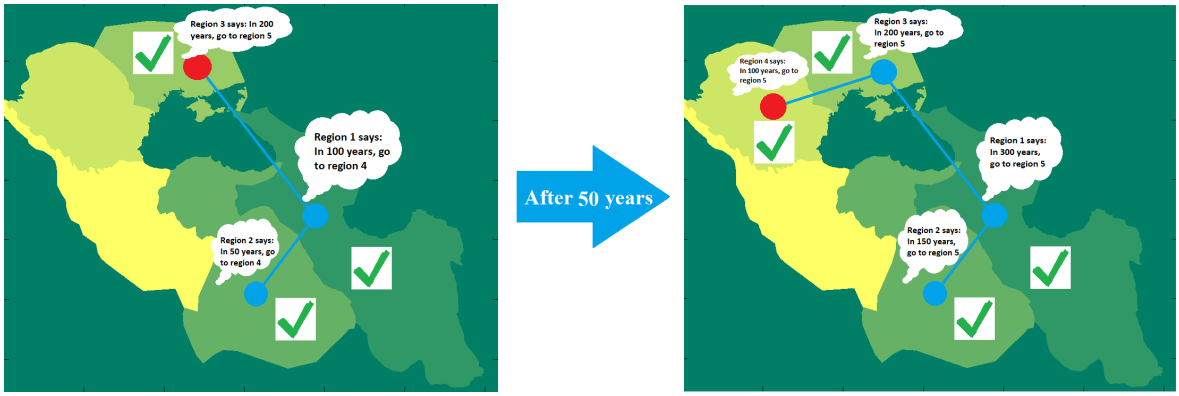


Figure 2.5: Illustration of a possible way the little fairy spreads the innovation across the land. The current position of the fairy (the region it is in) is represented by the red dot. The blue dots represents the regions the fairy was in before. Infected regions (denoted by the check mark) interact with non-infected regions. This is represented by them commanding the fairy to go to a non-infected region.

The spread of paper [29]

For thousands of years, people used to write on very different materials dependent on the part of the world they lived in. From 3000 BC on, papyrus had been the most commonly used material to record text on in ancient Egypt and around the Mediterranean, whereas in China it was primarily captured on bamboo and even on silk. In between, in the Middle East, it was usual to write on parchments made out of calfskin. Those materials often were either too heavy (bamboo), too expensive (silk and calfskin) or too

effortful to produce (papyrus) for large parts of the population to communicate through writing.
The first version of paper, which is made of vegetable fibres, originates from China, more precisely, from the Chinese province of Dunhuang in 150 AD. It spread over Western China within the next 250 years and emerged in Japan until 610 before spreading to India in the 7th century. Since China tried to keep the knowledge of papermaking a secret it took until 751 that it reached the Middle East: During the Battle of Talas in today's Kyrgyzstan between the Chinese Tang dynasty and the Arab Abbasid Caliphate along with the Tibetan Empire, Chinese soldiers were captured and passed the secret of papermaking on to the Arabs. This lead to paper being the predominant writing material in the Middle East by the end of the 10^{th} century.
Europe had its first paper mill, a machine to produce paper out of its ingredients, in 1056 in today's Spain. After the Islamic conquest of the Iberian Peninsula in the 8^{th} century, large parts of the area were still more Islamic in terms of culture. This is why paper reached the most Western part in Europe before becoming popular in France in the 12^{th} and most of the rest of Central Europe in the 14^{th} century. England only built their first paper mill in 1490 and Sweden in 1612. For many years paper had been proclaimed as unholy to write on by the church which decelerated the spreading process.
Paper production arrived in Mexico by 1575 and in Philadelphia by 1690.

2.2.2 Idea 2: Changing the state space

Another option to model the spread of innovations in this region-based framework is by a Markov process on an altered state space:
In every moment, let us keep track of which regions are infected and which are not. Formalizing that, we use the state space

$$S = \{0,1\}^n \setminus \{0\}^n$$

and call every element of S a **configuration**. A configuration contains the information which regions are infected in the following way:

$$\phi(j) = \begin{cases} 1, & \text{region } j \text{ is infected} \\ 0, & \text{else} \end{cases} \tag{2.1}$$

We exclude the zero-configuration $(0,\dots,0)^T$ from the state space because the innovation can only be transported if it already exists and since we only want to model the spreading and not the inventive process, we neglect the case that the innovation has still to be invented.
S then consists of $s := 2^n - 1$ configurations that are denoted by $\phi_1, \dots, \phi_s$.
If $\phi(j) = 1$ we say that j is **checked** in ϕ.

We will denote the process on the state space of configurations by

$$Y_t : [0, \infty) \to S.$$

Again we assume time-independent transition rates between the regions that govern the average time until an infected region infects a non-infected one that we store in the transition rate matrix R.
With a given R on the state space of regions, how do we compute the transition rates between the configurations, i.e. the rate matrix on S? For the answer we need two definitions:

Definition 2.1 (Covered)**.** *For two configurations ϕ_1,ϕ_2 we say that ϕ_1 is **covered** by ϕ_2 if it holds*

$$\phi_1(i) = 1 \Rightarrow \phi_2(i) = 1.$$

Definition 2.2 (Simply covered)**.** *For two configurations ϕ_1,ϕ_2 we say that ϕ_1 is **simply covered** by ϕ_2 if ϕ_1 is covered by ϕ_2 and there is only one i s.t. $\phi_1(i) = 0$ and $\phi_2(i) = 1$.*

Example 10. ϕ_1 is covered by ϕ_3 and simply covered by ϕ_2. ϕ_2 is simply covered by ϕ_3:

$$\phi_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \phi_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \phi_3 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}. \tag{2.2}$$

Please note: We will denote the i-th entry of a configuration ϕ by $\phi(i)$ and by ϕ_i the i-th configuration in (a subset of) S.
The configuration that is checked only at j will be denoted by ϕ^j.

Lemma 2.1. *Let R be the matrix of transition rates for a state space of cardinality n. Then for the corresponding spreading process on the state space of configurations $S = \{0,1\}^n \setminus \{0\}^n$ the generator $\bar{R}$ is given by:*

$$\bar{R}_{ij} = \begin{cases} \sum\limits_{l \text{ s.t. } \phi_i(l)=1} R_{lk}, & \text{if } \phi_i \text{ is simply covered by } \phi_j \text{ and } \phi_j - \phi_i = \phi^k \\ -\sum\limits_{k=1, k\neq i}^{s} \bar{R}_{ik}, & \text{if } i = j \\ 0, & \text{else} \end{cases} \tag{2.3}$$

Proof. Direct transitions can only be made between configurations that differ in exactly one entry since the exponentially distributed waiting times yield that it is infinitely improbable that two transitions from region to region are made exactly at

the same time. Therefore we only have to discuss the case that two configurations ϕ_i and ϕ_j differ in exactly one entry and w.l.o.g. ϕ_i is simply covered by ϕ_j:
Remember that $R_{lk} = \lambda(l,k)$ represents the inverse of the expected waiting time until a Markov process transitions from state l to k. In the case of multiple infected states according to ϕ_i in the process of configurations multiple states can make a transition to a non-infected state k independently from each other. The state k will be infected as soon as the first transition from any of those states occurs. This yields for the waiting time $\bar{T}_i$ of ϕ_i:

$$\bar{T}_i = \min_{l\ s.t. \phi_i(l)=1} T_l \tag{2.4}$$

Denoting by $\{l_1, \dots, l_m\} = \{l : \phi_i(l) = 1\}$, it holds [18]:

$$\begin{aligned}
\mathbb{P}[\bar{T}_i > x] &= \mathbb{P}[T_{l_1} > x, \dots, T_{l_m} > x] \\
&= \mathbb{P}[T_{l_1} > x] \cdot \ldots \cdot \mathbb{P}[T_{l_m} > x] \\
&= \prod_{u=1}^{m} e^{-\lambda_{l_u} x} = e^{-(\sum_{u=1}^{m} \lambda_{l_u})x} \\
&= e^{-(\sum_{l:\phi_i(l)=1} \lambda_l)x}.
\end{aligned} \tag{2.5}$$

This shows that the waiting time of ϕ_i is exponentially distributed with parameter $\sum_{l:\phi_i(l)=1} \lambda_l = \sum_{l:\phi_i(l)=1} R_l =: \bar{R}_i$. The rate $\bar{R}_{ij}$ is then given by

$$\bar{R}_{ij} = \bar{R}_i \bar{p}_{ij}$$

with

$$\begin{aligned}
\bar{p}_{ij} = \mathbb{P}[Y(\bar{T}_i) = \phi_j | Y(0) = \phi_i] &= \sum_{u=1}^{m} \mathbb{P}[T_{l_u} = \min_{l\ s.t. \phi_i(l)=1} T_l] p_{l_u k} \\
&= \sum_{u=1}^{m} \mathbb{P}[T_{l_u} = \min_{l\ s.t. \phi_i(l)=1} T_l] \frac{R_{l_u k}}{R_{l_u}}.
\end{aligned} \tag{2.6}$$

For the final step, it will be helpful that

$$\mathbb{P}[T_{l_u} = \min_{l\ s.t. \phi_i(l)=1} T_l] = \frac{R_{l_u}}{R_{l_1} + \cdots + R_{l_m}} \tag{2.7}$$

because for m independent exponentially distributed random variables $Z_1, \dots, Z_m$

with parameters $\lambda_1, \ldots, \lambda_m$ it holds [19]:

$$\begin{aligned}
&\mathbb{P}[Z_i = \min_{j=1,\ldots,m} Z_j] \\
&= \mathbb{P}[Z_i \leq Z_j \ \forall j = 1, \ldots, m] \\
&= \mathbb{P}[Z_i < Z_j \ \forall j = 1, \ldots, m] \\
&= \int_0^\infty \mathbb{P}[t < Z_j \ \forall j = 1, \ldots, m] \mathbb{P}[Z_i = t] dt \\
&\overset{\text{Indep. + (1.30)}}{=} \int_0^\infty \prod_{j \neq i} \mathbb{P}[t < Z_j] \lambda_i e^{-\lambda_i t} dt \\
&\overset{(1.29)}{=} \lambda_i \int_0^\infty e^{-(\lambda_1 + \cdots + \lambda_m)} dt \\
&= \lambda_i [\frac{-e^{-(\lambda_1 + \cdots + \lambda_n)t}}{\lambda_1 + \cdots + \lambda_n}]_0^\infty \\
&= \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_m} \\
&= \frac{R_i}{R_1 + \cdots + R_m}.
\end{aligned}$$

Then

$$\bar{p}_{ij} = \sum_{u=1}^{m} \frac{R_{l_u}}{R_{l_1} + \cdots + R_{l_m}} \frac{R_{l_u k}}{R_{l_u}} = \sum_{u=1}^{m} \frac{R_{l_{u_k}}}{R_{l_1} + \cdots + R_{l_m}}$$

and consequently,

$$\begin{aligned}
\bar{R}_{ij} = \bar{R}_i \bar{p}_{ij} &= (R_{l_1} + \cdots + R_{l_m}) \sum_{u=1}^{m} \frac{R_{l_{u_k}}}{R_{l_1} + \ldots R_{l_m}} \\
&= \sum_{u=1}^{m} R_{l_u k}.
\end{aligned}$$

□

2.2.3 Neighbouring regions

Let us augment the model by one further assumption: The innovation can only be transferred between **neighbouring** regions which means that regions have to share a sufficiently big part of their boundaries in order to be able to infect each other with the innovation.

We hence have to set all transition rates between non-neighbouring regions to 0 as the average waiting time until a transition is made between two of them is infinite.

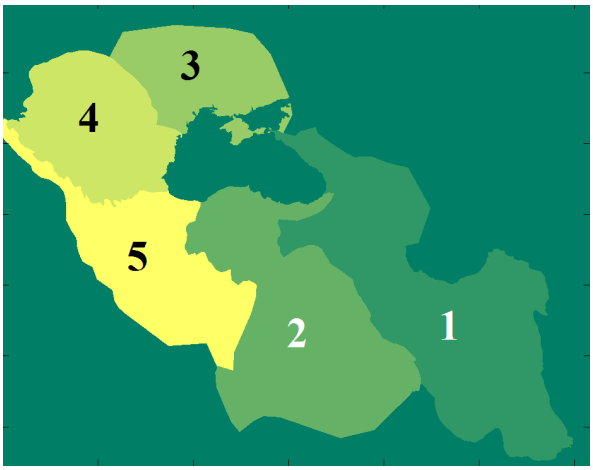


Figure 2.6: Neighbouring regions are 1 and 2, 2 and 5, 3 and 4, 4 and 5.

With the regions given as in Figure 2.6 it can be seen that the neighbouring regions are:

1 and 2

2 and 5

3 and 4

4 and 5.

This yields that only some regions are directly connected to each other and the connections are of different intensities (different rates). We have thereby created a **network** of regions.

We store the neighbourhood information in the set $\mathcal{N} \subset \mathbb{R}^2$ and demand for every transition rate matrix R that

$$R_{ij} = 0 \; \forall \; (i,j) \notin \mathcal{N}. \tag{2.8}$$

2.3 Computing path probabilities

As mentioned in the very beginning of this chapter, the goal is to be able to find out the **path** that an innovation took. By a path $(i_1, \ldots, i_n)$, we denote the order $(I_1, \ldots, I_n)$ in which the regions get infected.
Given the rate matrix R we will now calculate the probability for a path $(i_1, i_2, \ldots, i_n)$.

Lemma 2.2. *Given $R \in \mathbb{R}^{n \times n}$ and assuming that the first infected region was i_1 with*

probability $\mathbb{P}[I_1 = i_1]$ *the probability for the path* $(i_1, \ldots, i_n)$ *is*

$$\mathbb{P}[(i_1, \ldots, i_n)] = \mathbb{P}[I_1 = i_1] \prod_{k=2}^{n} \frac{\sum_{l=1}^{k-1} \lambda(i_l, i_k)}{\sum_{u=1}^{k-1} \sum_{v=1}^{n-k+1} \lambda(i_u, i_{n-v+1})}.$$

Proof. Let ϕ be the configuration with $\phi(i) = 1$ exactly for $i \in \{i_1, \ldots, i_{k-1}\}$ and ϕ' be the configuration with $\phi'(i) = 1$ exactly for $i \in \{j, i_1, \ldots, i_{k-1}\}$. Then the probability that region j is reached next with $\{i_1, \ldots, i_{k-1}\}$ already infected is

$$\begin{aligned} \mathbb{P}[i_k = j | i_1, \ldots, i_{k-1}] &= \mathbb{P}[Y_{T_\phi} = \phi' | Y_0 = \phi] \\ &= \frac{\bar{R}_{\phi\phi'}}{\bar{R}_\phi} = \frac{\sum_{l=1}^{k-1} R_{i_l,j}}{\sum_{\tilde{\phi} \in S} R_{\phi\tilde{\phi}}} = \frac{\sum_{l=1}^{k-1} R_{i_l,j}}{\sum_{u=1}^{k-1} \sum_{v=1}^{n-k+1} R_{i_u,i_{n-v+1}}} \\ &= \frac{\sum_{l=1}^{k-1} \lambda(i_l, j)}{\sum_{u=1}^{k-1} \sum_{v=1}^{n-k+1} \lambda(i_u, i_{n-v+1})}. \end{aligned} \tag{2.9}$$

Remember, Y_{T_ϕ} in the second term is the value of Y_t at T_ϕ which is the time the process leaves ϕ. In the denominator of the fifth expression, $\sum_{u=1}^{k-1} \sum_{v=1}^{n-k+1} R_{i_u,i_{n-v+1}}$ is the sum of rates from every infected to every non-infected region.
Then

$$\begin{aligned} \mathbb{P}[(i_1, \ldots, i_n)] &= \mathbb{P}[I_1 = i_1] \mathbb{P}[i_2|(i_1)] \cdot \mathbb{P}[i_3|(i_1, i_2)] \cdot \ldots \cdot \mathbb{P}[i_n|(i_1, \ldots, i_{n-1})] \\ &= \mathbb{P}[I_1 = i_1] \prod_{k=2}^{n} \mathbb{P}[i_k|(i_1, \ldots i_{k-1})] \\ &= \mathbb{P}[I_1 = i_1] \prod_{k=2}^{n} \frac{\sum_{l=1}^{k-1} \lambda(i_l, i_k)}{\sum_{u=1}^{k-1} \sum_{v=1}^{n-k+1} \lambda(i_u, i_{n-v+1})}. \end{aligned}$$

□

We will now make the concept of paths more precise and for every infected region specify where this infection came from.

Definition 2.3 (Originating path)**.** $((i_1, \ldots, i_n), (j_1, \ldots, j_{n-1})) \in \mathbb{R}^n \times \mathbb{R}^{n-1}$ *is an* ***originating path*** *where* $(i_1, \ldots, i_n)$ *is a path and region* i_{k+1} *was infected by region* j_k*.*

Lemma 2.3. *Given $R \in \mathbb{R}^{n\times n}$ and assuming that the first infected region was i_1 with probability $\mathbb{P}[I_1 = i_1]$ the probability for the originating path $((i_1, \dots, i_n), (j_1, \dots, j_{n-1}))$ is*

$$\mathbb{P}[((i_1, \dots, i_n), (j_1, \dots, j_{n-1}))] = \mathbb{P}[I_1 = i_1] \prod_{k=2}^{n} \frac{\lambda(j_{k-1}, i_k)}{\sum_{u=1}^{k-1} \sum_{v=1}^{n-k+1} \lambda(i_u, i_{n-v+1})}.$$

Proof.

$$\mathbb{P}[i_k = i, j_{k-1} = j | (i_1, \dots, i_{k-1})] \overset{(1.9)}{=} \mathbb{P}[j_{k-1} = j | (i_1, \dots, i_{k-1}, i_k)] \mathbb{P}[i_k = i | (i_1, \dots, i_{k-1})]$$

$$\overset{(2.9)}{=} \frac{\lambda(j, i)}{\sum_{l=1}^{k-1} \lambda(i_l, i)} \frac{\sum_{l=1}^{k-1} \lambda(i_l, j)}{\sum_{u=1}^{k-1} \sum_{v=1}^{n-k+1} \lambda(i_u, i_{n-v+1})} = \frac{\lambda(j, i)}{\sum_{u=1}^{k-1} \sum_{v=1}^{n-k+1} \lambda(i_u, i_{n-v+1})}.$$

Hence

$$\mathbb{P}[((i_1, \dots, i_n), (j_1, \dots, j_{n-1}))] = \mathbb{P}[I_1 = i_1] \prod_{k=2}^{n} \frac{\lambda(j_{k-1}, i_k)}{\sum_{u=1}^{k-1} \sum_{v=1}^{n-k+1} \lambda(i_u, i_{n-v+1})}.$$

□

We will now go one step further and assume that we do not know the start region. Given an estimate about what the **end region**, i.e. the region that was infected last, was, we can infer probabilities for every region to be the start region. More generally, given an estimate about every region to be the k-th infected region I_k we can calculate the probability for every region to be the start region, i.e. given

$$\mathbb{P}[I_k = i]$$

we can obtain

$$\mathbb{P}[I_1 = j].$$

Note that

$$\begin{aligned} \mathbb{P}[I_k = i] &= \sum_{l=1}^{n} \mathbb{P}[I_1 = l] \mathbb{P}[I_k = i | I_1 = l] \\ &= \sum_{l=1}^{n} \mathbb{P}[I_1 = l] \sum_{\text{Paths } p \text{ with } i_1 = l, i_k = i}^{n} \mathbb{P}[p | I_1 = l]. \end{aligned} \tag{2.10}$$

Therefore, we derive the following linear system

$$\mathbb{P}[I_1 = 1] \sum_{\text{Paths } p \text{ with } i_1=1, i_k=1}^{n} \mathbb{P}[p|I_1 = 1] + \cdots + \mathbb{P}[I_1 = n] \sum_{\text{Paths } p \text{ with } i_1=n, i_k=1}^{n} \mathbb{P}[p|I_1 = n] = \mathbb{P}[I_k = 1]$$

$$\vdots$$

$$\mathbb{P}[I_1 = 1] \sum_{\text{Paths } p \text{ with } i_1=1, i_k=n}^{n} \mathbb{P}[p|I_1 = 1] + \cdots + \mathbb{P}[I_1 = n] \sum_{\text{Paths } p \text{ with } i_1=n, i_k=n}^{n} \mathbb{P}[p|I_1 = n] = \mathbb{P}[I_k = n] \tag{2.11}$$

which can be written as

$$Ax = r \tag{2.12}$$

where $r_i = \mathbb{P}[I_k = i]$, $A_{ij} = \sum_{\text{Paths } p \text{ with } i_1=j, i_k=i}^{n} \mathbb{P}[p|I_1 = i]$ and $x_i = \mathbb{P}[I_1 = i]$.
Then

$$x = A^{-1}r. \tag{2.13}$$

Of course, this only holds if A is invertible.

2.4 Finding the best fitting rates

We can now simulate the spreading process given the transition rates between the regions but how do we find out which rates best fit to the process that historically took place? In other words, given which rates will a simulation of the spreading process be the most similar to what actually happened? If we know those values we can calculate which spreading path has most likely taken place with the results we have derived in the previous section.
To that end we introduce here the 'real' process $\hat{Y}_t : [0, \infty] \to S$ with

$$(\hat{Y}_t)_i = \begin{cases} 1 & \text{, region } i \text{ was infected until time } t \textbf{ in reality} \\ 0 & \text{, else.} \end{cases} \tag{2.14}$$

Using the term 'infected...in reality' requires a notion of what it should mean if a region was 'infected' by an innovation thousands of years ago on Earth. Later on we will specify such a notion but for now let us assume we have one at hand so that we can objectively measure whether an innovation has spread out within a region enough for us to call this region 'infected'.
We assume here that the spreading process in reality obeys the rules of a Markov process with average waiting times for an infection spread from one region to another that are governed by rates in the rate matrix $R^\star \in \mathbb{R}^{n \times n}$ that is unknown for us.

Please note: We set the starting time of $\hat{Y}_t$ to 0 although depending on the innovation the process can have started at other times, e.g. 5000 BC.
Let us assume we know the **mean first hitting times** $m(R^\star)$ of $\hat{Y}_t$, which for a chosen rate matrix R are defined as $m(R) \in \mathbb{R}^{2^n-1\times n}$ with

$$m(R)_{ij} = \mathbb{E}[\tau^{A_j}|Y_0 = \phi_i] \tag{2.15}$$

where $\tau^A = \inf\{t : Y_t \in A\}$ and $A_j = \{\tilde{\phi} \in S : \tilde{\phi}(j) = 1\}$.

Then for a rate matrix R we use the similarity of $m(R)$ and $m(R^\star)$ as the measurement of how close R comes to $R^\star$. In other words, we measure the quality of the fit of R by the loss function

$$\rho(R, m(R^\star)) = dist(m(R), m(R^\star)) \tag{2.16}$$

and try to find

$$\arg\min_{R\in\mathcal{R}_n} \rho(R, m(R^\star)$$

where $\mathcal{R}_n$ is the space of $n \times n$ rate matrices, i.e.

$$\mathcal{R}_n := \{R \in \mathbb{R}^{n\times n}|R_{ii} = -\sum_{j\neq i} R_{ij}, R_{ij} \geq 0 \text{ for } i \neq j\}. \tag{2.17}$$

Unfortunately, we will not come across any examples where we actually know $m(R^\star)$.
Instead, we look for a replacement. For that we need data that informs us about the way the innovation spread in reality and gives at least the following information:
For every region, what was the time the region was infected by the innovation?
Assume we have those data. Then we can deduce the **first hitting times** $f_D \in \mathbb{R}^{2^n-1\times n}$ from them where

$$(f_D)_{ij} := \max\{0, \hat{\tau}^{A_j} - \hat{\tau}^{\phi_i}\} \tag{2.18}$$

with $\hat{\tau}^A = \inf\{t : \hat{Y}_t \in A\}$. If $\hat{Y}_t \neq \phi_i$ at all times then $\tau^{\phi_i} = \infty$ and $(f_D)_{ij} = 0$.
So $(f_D)_{ij}$ denotes the time that passed between the process $\hat{Y}_t$ assuming configuration ϕ_i and the infection of region j if this value is positive.
As $\hat{Y}_t$ is one realisation of the spreading process with the rate matrix $\bar{R}$ our hope is

that

$$\text{maybe not } m(R^\star) = f_D$$
$$\text{but } m(R^\star) \approx f_D$$

so that

$$\arg\min_{R\in\mathcal{R}_n} \rho(R, f_D) \approx \arg\min_{R\in\mathcal{R}_n} \rho(R, m(R^\star)). \tag{2.19}$$

It can be observed, that the distance function $dist$ does not have to work between the wholes of $m(R)$ and f_D but only between one row of them:
We denote by

$$\phi_{i_1} = \hat{Y}_0$$

the first configuration that the real process had according to the data. Then all rows other than the i_1-row are redundant and can be inferred from it because then

$$\begin{aligned}&(f_D)_{i_1 j} = \tau^{A_j} - \tau_{\phi_{i_1}}, (f_D)_{i_1 k} = \tau^{A_k} - \tau_{\phi_{i_1}}\\ &\Rightarrow (f_D)_{jk} = \max\{0, \tau^{A_k} - \tau^{A_j}\} = \max\{0, (f_D)_{i_1 k} - (f_D)_{i_1 j}\}.\end{aligned}$$

In $m(R)$ we have information about the expected times from the process starting in each configuration until reaching every region. But in reality there can only have been one scenario, i.e. there was only one starting configuration that we have to consider. This will have the value 1 where the innovation was known first according to the data and 0 elsewhere. Hence we only compare the row of $m(R)$ that corresponds to the starting configuration to the i_1-row of f_D.
In general, let us denote the i_1-row of a matrix M by M_{i_1}.

We thus have identified the problem we want to solve as a replacement of

$$\arg\min_{R\in\mathcal{R}_n} dist(m(R), m(R^\star)). \tag{2.20}$$

It reads

$$\arg\min_{R\in\mathcal{R}_n} dist(m(R)_{i_1}, (f_D)_{i_1}). \tag{2.21}$$

In the next chapter we will apply this setting to a spreading process that we have data about. It will be explained there how to infer f_D from the data. In the next section already, we will see how to infer $m(R)$ from R.

2.4.1 Computing the mean first hitting times

In (1.51) we saw how the mean first hitting times m^A of a set $A \subset E$ of a time-continuous Markov process on E with rate matrix R can be computed:

$$m_i^A = \begin{cases} 0 & , i \in A \\ \frac{1}{\lambda(i,1)+\cdots+\lambda(i,n)}(1 + \sum_{j \in E, j \neq i} m_j^A \lambda(i,j)) & , i \notin A. \end{cases} \tag{2.22}$$

From (2.3) we know the generator $\bar{R}$ of the spreading process on the state space of configurations. As we are interested in the expected times that Y_t reaches the set A_j of configurations that are checked at j, we can reformulate (2.22) by setting $A = A_j$ and using the rates between the configurations that are captured in $\bar{R}$ for the $\lambda(i,j)$.

$$m_{\phi_i}^{A_j} = \begin{cases} 0 & , i \in A_j \\ (\bar{R}_{i1} + \cdots + \bar{R}_{in})^{-1}(1 + \sum_{\phi_u \in S, u \neq i} m_{\phi_u}^{A_j} \bar{R}_{ij}) & , i \notin A_j \end{cases} \tag{2.23}$$

where the sum $\bar{R}_{i1} + \cdots + \bar{R}_{in}$ excludes $\bar{R}_{ii}$.

Theorem 2.1. *(2.23) is solvable if and only if for every $\phi \notin A_j$ there is a sequence of regions $r = r_0, r_1, \ldots, r_{l-1}, r_l = j$ with $\lambda(j_p, j_{p+1}) > 0 \; \forall \; p \in \{0, \ldots, l-1\}$ and $\phi(r) = 1$.*

Proof. From Theorem 1.5 we know that the solvability of (2.23) is equivalent to the existence of a sequence of configurations $\underbrace{\phi = \phi_{j_0}, \phi_{j_1}, \ldots, \phi_{j_{l-1}}}_{\notin A_j}, \underbrace{\phi_{j_l}}_{\in A_j}$ with $\bar{R}_{j_p, j_{p+1}} > 0$ for all $p \in \{0, \ldots, l-1\}$ and $\phi \notin A_j$. By the construction of $\bar{R}$ it is clear that for every sequence of configurations $(\phi_{j_p})_{p \in \{0,\ldots,l\}}$ with $\bar{R}_{j_p, j_{p+1}} > 0$, ϕ_{j_p} always has to be simply covered by $\phi_{j_{p+1}}$. Moreover, letting $\phi_{j_{p+1}} - \phi_{j_p} = \phi^k$, $\bar{R}_{j_p, j_{p+1}} > 0$ is equivalent to the existence of an index i with $\phi_{j_p}(i) = 1$ at such that $R_{ik} > 0$. This proves the forward direction.

If the sequence of regions in the Theorem exists there is also such a desired sequence of configurations because for a $\phi \in S$ we can take any checked state r and take the sequence that leads to j in the state space of regions. Then the sequence of configurations is given by

$$\phi_{j_0} = \phi, \phi_{j_1} = \phi + \phi^{r_1}, \phi_{j_2} = \phi_{j_1} + \phi^{r_2}, \ldots, \phi_{j_l}$$

where $\phi_{j_l} \in A_j$. □

This means that only by looking at the rates between the regions we can decide

whether $\bar{R}$ yields mean first hitting times.
Now let us reverse this: If we are given values $m_\phi = (m_\phi^{A_1}, \dots, m_\phi^{A_n})$ for the mean first hitting times is there always a rate matrix R such that $m(R)_i = m_\phi$?
If this is the case then we can solve (2.23) for $\bar{R}$ with given m_ϕ. But as we can see in (2.3) we make severe demands on some entries of $\bar{R}$. Before we investigate when (2.23) is solvable for $\bar{R}$ let us first discuss the when (2.22) is solvable by any rate matrix R.

Lemma 2.4. *For any finite mean first hitting times $m_i = (m_i^1, \dots, m_i^n)$ that are pairwise different, (2.22) is solvable for R.*

Proof. We order the entries of m_i ascendingly to $m_i^{j_1} < \dots < m_i^{j_n}$. $m_i^i = 0$ so j_1 will be i. Then we build a rate matrix R that is 0 on all non-diagonal entries except the following ones:

$$\text{Set } R_{j_1,j_2} = \frac{1}{m_i^{j_2}}$$
$$\text{and generally } R_{j_k,j_{k+1}} = \frac{1}{m_i^{j_{k+1}} - m_i^{j_k}}.$$

Thus, ignoring the time between transitions a simulation of the induced Markov process of R will be $j_1, j_2, \dots, j_n$ where the expected waiting time between every two transition is given by $m_i^{j_{k+1}} - m_i^{j_k}$ and thus the mean first hitting time of state k is

$$m_i^{j_{k+1}} - m_i^{j_k} + m_i^{j_k} - m_i^{j_{k-1}} + \dots - \underbrace{m_i^{j_1}}_{=0} = m_i^{j_{k+1}}.$$

□

We can now make the analogous statement for the spreading process.

Lemma 2.5. *For any finite mean first hitting times, $m_i = (m_\phi^{A_1}, \dots, m_\phi^{A_n})$ (2.23) is solvable for R.*

Proof. The proof is very similar to the previous one: Let $I = \{i_1, \dots, i_m\}$ be the checked states of ϕ and $J = \{i_{m+1}, \dots, i_n\}$ the non-checked ones. Again we order the entries of m_ϕ ascendantly to $\underbrace{m_\phi^{A_{i_1}}, \dots, m_\phi^{A_{i_m}}}_{=0}, \underbrace{m_\phi^{A_{i_{m+1}}}, \dots, m_\phi^{A_{i_n}}}_{>0}$. We then pick any $i \in I$ and proceed as in the proof above by setting all entries of R to 0 except

$$R_{i,i_{m+1}} = \frac{1}{m_\phi^{A_{i_{m+1}}}},$$
$$R_{i_{m+1+k},i_{m+2+k}} = \frac{1}{m_\phi^{A_{i_{m+2+k}}} - m_\phi^{A_{i_{m+1+k}}}}.$$

□

We assumed here that we can assign a positive value to any entry of R. Taking the concept of neighbouring regions in (2.8) into account, this does not have to be the case. In consequence, given the scenario of non-neighbouring regions Lemma 2.5 is not true:
Given three regions $1, 2$ and 3 with neighbourhood $\mathcal{N} = \{(1,2),(2,3)\}$, i.e. 1 and 2 and 2 and 3 are neighbouring. Let mean first hitting times for the configuration $\phi = (1,0,0)^T$ be given by $(0, 2a, a)$. Then A_3 has to be reached before A_2 although this cannot happen because starting in region 1 region 2 has to be infected in order for region 3 to be infected.

Given a neighbourhood such that (2.23) is solvable for R, the solution is not unique as the following counter example shows:
Let $R \in \mathbb{R}^{3\times 3}$ be a rate matrix and $\bar{R}$ its corresponding generator of the spreading process. Let the starting configuration be $\phi_{i_1} = (1,1,0)^T$. Then

$$\begin{aligned} m_\phi^{A_3} &= \frac{1}{R_{13}+R_{23}} \\ \Leftrightarrow R_{13} &= \frac{1 - R_{23} m_\phi^{A_3}}{m_\phi^{A_3}} =: r(R_{23}). \end{aligned} \tag{2.24}$$

We will shortly use

$$r^{-1}(R_{13}) = \frac{1 - R_{13} m_\phi^{A_3}}{m_\phi^{A_3}}. \tag{2.25}$$

This gives a relation between R_{13} and R_{23}. Of course, independently on them both $m_\phi^{A_1}$ and $m_\phi^{A_2}$ are 0. So for $\phi_{i_1} = (1,1,0)^T$ we can always find different pairs of R_{13} and R_{23} that preserve the mean first hitting times. The $(1,2)$- ,$(2,1)$-, $(3,1)$- and $(3,2)$-entries do not affect $m_\phi^{A_3}$ so they can be chosen arbitrarily.

2.4.2 The problem of non-uniqueness

We will now quickly illustrate this problem of non-uniqueness in two examples of mean first hitting times for a 3×3-matrix.

Example 11. Let $R^{(1)}, R^{(2)}$ be 3×3 rate matrices and we want to know how to choose them so that they contain the real rates between the regions $R^{\star(1)}$ and $R^{\star(2)}$. Please note: There can only be one real rate matrix but for the sake of this example, we assume two parallel universes where the rates between the regions are different in both.
We assume we know the exact mean first hitting times $m_\phi = (m_\phi^{A_1}, m_\phi^{A_2}, m_\phi^{A_3})$ of

both $R^{\star(1)}$ and $R^{\star(2)}$ for $\phi = (1,1,0)^T$:

$$m_\phi^{(1)} = (0,0,1),$$
$$m_\phi^{(2)} = (0,0,10).$$

Figure 2.9 shows the relation between $R_{13}^{(1)}$ and $R_{23}^{(1)}$ (in blue) and $R_{13}^{(2)}$ and $R_{23}^{(2)}$ (in green and red).
For $R^{(1)}$ and $R^{(2)}$ to be rate matrices all entries have to be non-negative. This yields a restriction on $R_{13}^{(1)}$ and $R_{23}^{(1)}$ resp. $R_{13}^{(2)}$ and $R_{23}^{(2)}$. It gives that

$$R_{13}^{(1)}, R_{23}^{(1)} \in [0,1]$$
$$\text{and } R_{13}^{(2)}, R_{23}^{(2)} \in [0,0.1].$$

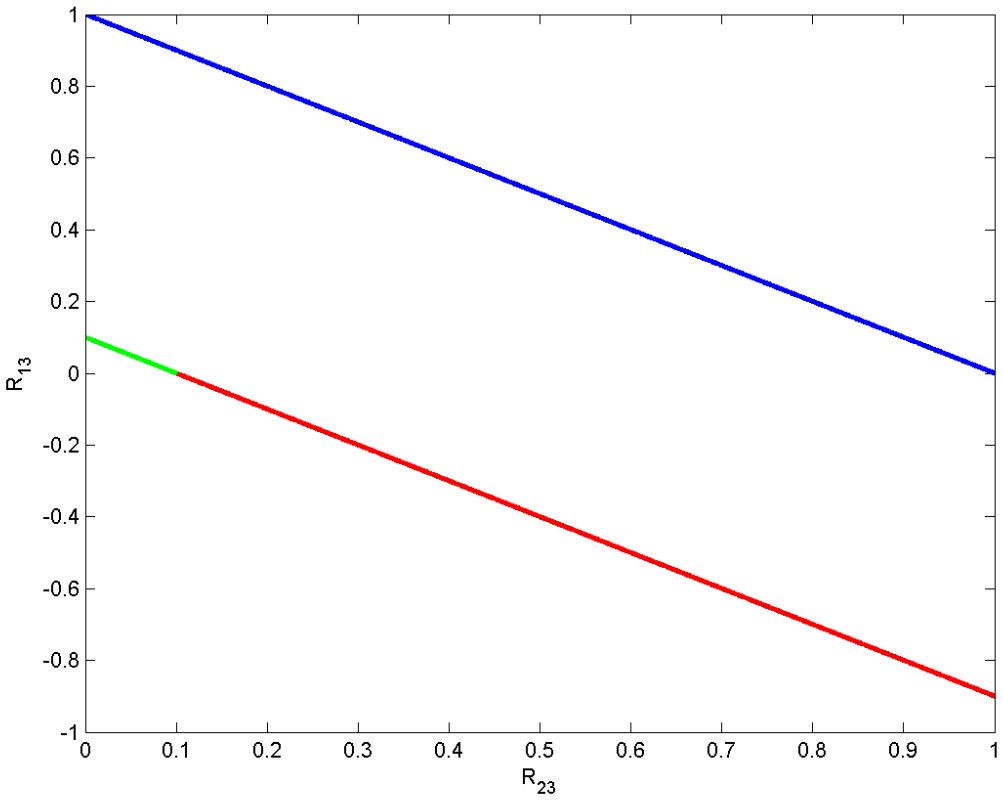


Figure 2.7: Relation between $R_{13}^{(1)}$ and $R_{23}^{(1)}$ with given mean first hitting time $m_{\phi^1}^{A_3} = 1$ (blue) and between $R_{13}^{(2)}$ and $R_{23}^{(2)}$ with given $m_{\phi^1}^{A_3} = 10$ (green and red). As we can see, both $R_{13}^{(1)}$ and $R_{23}^{(1)}$ have to lie in the interval $[0,1]$. On the other hand, $R_{13}^{(2)}$ and $R_{23}^{(2)}$ are restricted to the interval $[0,0.1]$. The lower graph is painted in red in the domain where $R_{13}^{(2)}$ would have to be negative in order to preserve (2.24).

Without any further knowledge, we can only guess the values of $R_{13}^{\star(1)}$ and $R_{23}^{\star(1)}$

resp. $R_{13}^{\star(2)}$ and $R_{23}^{\star(2)}$ from the sets $[0,1]$ resp. $[0,0.1]$ without a preference. In other words, our own probability distributions for $\mathbb{P}[R_{k3}^{(l)} = R_{k3}^{\star(l)}]$, $k,l \in \{1,2\}$, are uniform on the intervals $[0,1]$ resp. $[0,0.1]$ so let us write

$$\pi_{13}^{(1)}(R_{13}) = \begin{cases} 1 & , R_{13} \in [0,1] \\ 0 & , \text{else.} \end{cases}$$

where $\pi_{13}^{(1)}$ is the **density** of $R_{13}^{\star(1)}$, i.e.

$$\int_0^x \pi_{13}^{(1)}(y)dy = \mathbb{P}[R_{13}^{\star(1)} \leq x]$$

and

$$\pi_{13}^{(2)}(R_{13}) = \begin{cases} \frac{1}{0.1} & , R_{13} \in [0,0.1] \\ 0 & , \text{else.} \end{cases}$$

In this example, a mean first hitting time of $m_\phi^{A_3} = 10$ yields a better estimate for the rates because the set of possible values is smaller than for the mean first hitting time of 1.

Let us not forget we could only make that restriction to an interval because we know that the rates have to be non-negative in order to be called rates. Someone who does not know that will not be able to find this restriction. In other words: One needs that **prior knowledge** about rates.
In fact, we might have even more of that prior knowledge. Let us assume that because of archaeological evidence, we strongly suspect that all rates in $R^{\star(1)}$ are between 0 and 0.1 because higher rates would yield a much faster innovation spread than the archaeological evidence suggests. Then it would be very unreasonable that $R_{13}^{(1)}$ or $R_{23}^{(1)}$ were close to 1.
This gives rise to the idea of a prior function that contains the a priori probability of every value for a rate. If we restrict the domain of possible values for $R_{13}^{(1)}$ manually to $[0,0.1]$, then also the domain for $R_{23}^{(1)}$ shrinks to $[0,0.1]$.
We have thereby altered the probability for $R_{13}^{(1)}$ to be equal to $R_{13}^{\star(1)}$ to

$$\pi_{13}^{(1)}(R_{13}) = \begin{cases} \frac{1}{0.1} & , R_{13} \in [0,0.1] \\ 0 & , \text{else.} \end{cases} \tag{2.26}$$

Of course, we can make the restriction much smoother, e.g. in the following way:

$$\pi^{(1)}(R_{13}) \sim \begin{cases} \exp(\frac{-(0.05-R_{13})^2}{0.01}) & , R_{13} \geq 0 \\ 0 & , \text{else.} \end{cases} \tag{2.27}$$

Then $\pi_{13}^{(1)}$ looks like the following graph:

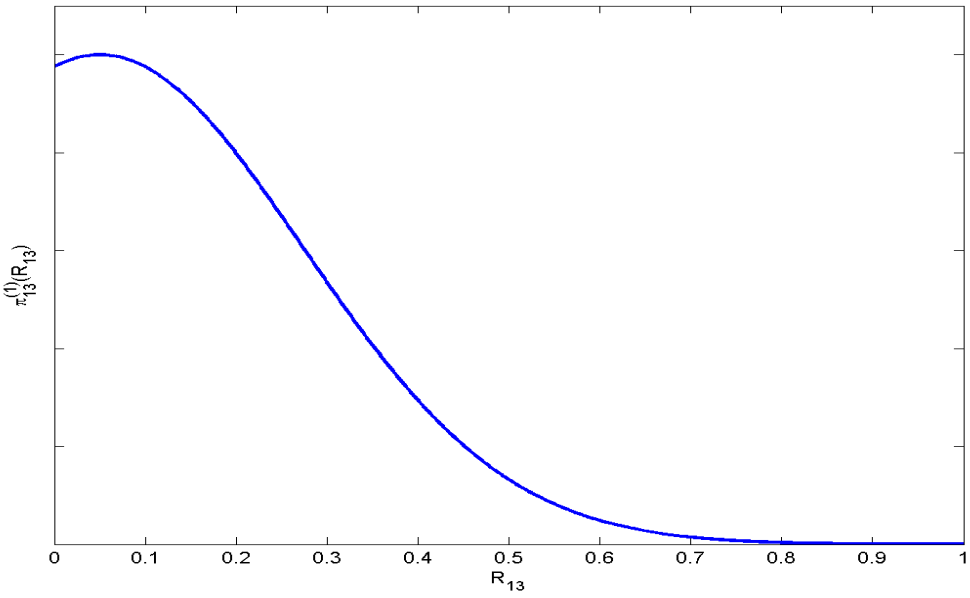


Figure 2.8: Graph of $\pi_{13}^{(1)}$ dependent on the change of R_{13}.

Remembering basic theory about the probability measure of a random variable (Definition 1.2) this yields for the probability of $R_{23}^{(1)} = R_{23}^{\star(2)}$:

$$\pi_{23}^{(1)}(R_{23}) \sim \begin{cases} \exp(\frac{-(0.05-r^{-1}(R_{23}))^2}{0.01}) & , R_{23} \geq 0 \\ 0 & , \text{else.} \end{cases}$$

The graph of $\pi_{23}^{(1)}$ has the form

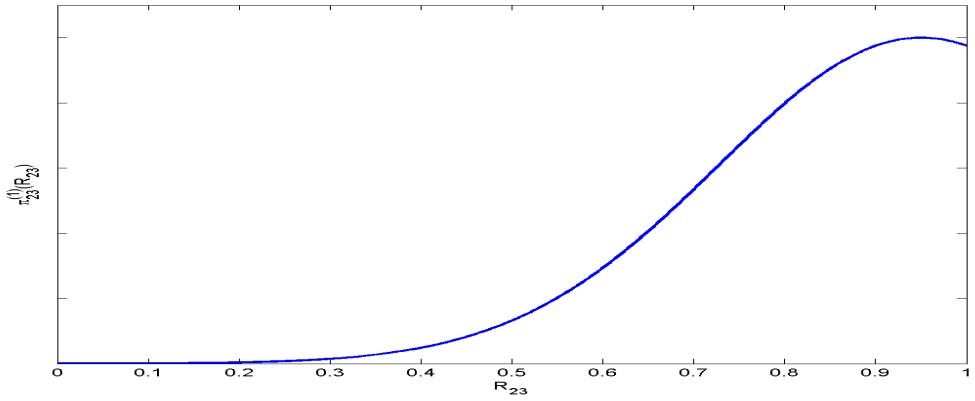


Figure 2.9: Graph of $\pi_{23}^{(1)}$ dependent on the change of R_{23}.

Still, we admit here only values for R_{13} and R_{23} so that $R^{(1)}$ and $R^{(2)}$ solve (2.20) because we assumed here that we know the exact mean first hitting times of $R^\star$. If we are only given f_D instead of $m(R^\star)$, then it is possible that a rate matrix does not solve (2.21) but still solves (2.20) because in general $m(R) \neq f_D$. So we also have to take those rate matrices into account, too, but also not forget that they should not be the top candidates for $R^\star$.
Therefore, it makes sense to quantify the probability that a realisation of the spreading process with underlying spread rates $R^\star$ has the first hitting times f_D. More generally: The probability that data occur given a parameter.
We already know this concept from the Bayesian modelling. There we denoted this probability by the **likelihood** $\mathbb{P}[\theta|\mathbf{D}]$ where θ was the parameter and **D** the observed data. In this setting, we define the likelihood

$$l_{f_D} \text{ as the density of } f_D|R^\star = R. \tag{2.28}$$

Together with the prior π we can infer the **posterior distribution** ν_{f_D} of the rate matrices:

$$\nu_{f_D} \text{ is the density of } R^\star|f_D \sim l_{f_D}(R)\pi(R). \tag{2.29}$$

This puts us into a Bayesian setting. We are now not interested in solving (2.21) exactly but want to find the probability distribution over the space of rate matrices.

Another reason for the Bayesian framework is the fact that available data about the spreading is likely to be incomplete. The reason for that is that an innovation could have been known in a region very early but some of its indication might have vanished due to natural processes in the ground or not been found because of lacking archaeological efforts. This means that instead of f_D the data only give us $\tilde{f}_D$ where not necessarily $f_D = \tilde{f}_D$ or only $f_D \approx \tilde{f}_D$.

As we have seen in Example 6, even small perturbations to the data can cause severe changes to the result of the estimator so $\arg\min_{R\in\mathcal{R}_n} \rho(R, f_D)$ can be very different from $\arg\min_{R\in\mathcal{R}_n} \rho(R, \tilde{f}_D)$. We compensate for that by using Bayesian modelling here.

In summary, again note the three reasons why we estimate the rates in a Bayesian setting:

> We might have a priori knowledge about where the rates should be and quantify that in the **prior**.
>
> A rate matrix can be the wanted $R^\star$ even if its mean first hitting times are not equal to f_D. We take that into account with the **likelihood**.
>
> The data can be incomplete and the result of the classical parameter estimation on that data can be far away from the truth.

We will apply this framework to data in the next chapter and specify sensible prior and likelihood functions. Before that we will discuss the probability for every spreading path in the Bayesian framework.

2.5 Path probabilities in the Bayesian framework

As explained in section 2.3, by a path we denote the order in which every region was infected, i.e. an ordering $i_1, \dots, i_n$ such that $\tau^{A_{i_1}} < \dots < \tau^{A_{i_n}}$. We saw earlier how, given a rate matrix R, we can compute the probability for every path and obtain the **path probability vector** $c(R) \in \mathbb{R}^z$ for all possible paths (i.e. along neighbouring regions) $p_1, \dots, p_z$ according to R, i.e.

$$c_j(R) := \mathbb{P}[\text{path } p_j]. \tag{2.30}$$

At that point we did not know yet that we will have to deal with not only one rate matrix but with the whole space of rate matrices where every rate is given a weight (its value in the posterior). We therefore define the **Bayesian path probability vector** by

$$b_j = \frac{1}{Z_b} \int_{\mathcal{R}_n} \nu_{f_D}(R) c_j(R) dR \tag{2.31}$$

where Z_b is a normalisation constant such that $\sum_{j=1}^{z} b_j = 1$.

b_j is finite because

$$\int_{\mathcal{R}_n} \nu_{f_D}(R) \underbrace{c_j(R)}_{\leq 1} dR \leq \int_{\mathcal{R}_n} \nu_{f_D}(R) dR = 1.$$

The latter statement hold because ν_{f_D} is the normalisation of $l_{f_D} \cdot \pi$ which itself is finite because $\int_{\mathcal{R}_n} \pi(R)dR = 1$ by construction and l_{f_D} is bounded by 1.

3 Application of the network-based approach

We will now apply the network-based approach to some real world data. As explained in the previous chapter, TOPOI provides some data on the spread of the woolly sheep and other ovicaprids (sheep and goats). We will make a division of the land into regions and use these data to

- Determine the spread rates between these regions
- Compute probabilities for every path that the spreading took

What makes the woolly sheep special [37]

The name woolly sheep might seem redundant at first because usually one associates sheep with wool and vice versa. There are, however, multiple types of sheep that do not wear any wool but hair instead. Although nowadays it is predominant at least in farming culture the woolly sheep is only the product of a genetic mutation of hairy sheep. Those hairy sheep, along with goats, were spread over Europe and Western Asia already around 10.000 BC which is presumably long before any sheep wore wool.
The eventual emergence of the woolly sheep meant a benefit for people as the wool has since then been used for textile production. When people realised how much they would profit from that feature of that new version of a sheep, they cultivated the herding of woolly sheep and invented instruments such as spindles for the post production of the wool.

3.1 Numerical preparations for the application of the model

As we have seen in the previous chapter, finding the rate matrix that fits best to $(f_D)_{i_1}$, i.e. solving (2.21)

$$\arg\min_{R\in\mathcal{R}_n} dist(m(R)_{i_1}, (f_D)_{i_1})$$

can have multiple very different solutions of which we might regard some as improbable by the prior function π. On top of that we remembered that **what happened in nature was not necessarily the most probable option.** As a consequence, the true rates of the spreading process can lie outside of the set of solutions of (2.21).

For both reasons we apply a Bayesian framework here that gives us a distribution over the space of rate matrices. The information that the data carry will then influence the result in the likelihood function l. Consequently, we will derive a probability distribution (the posterior) that will be centred around the rate matrices that actually fit best to the data.
We hence compute the posterior distribution ν_{f_D} for which should hold

$$R^\star | f_D \sim \nu_{f_D}.$$

However, it is infeasible to calculate the full posterior neither analytically nor computationally because the space of rate matrices is $n \cdot (n-1)$-dimensional. Thus, we will use the Metropolis-Hastings algorithm (section 1.2.4) to sample a sufficient number of rate matrices distributed according to the posterior and from them infer an approximate distribution of the rate matrices.

For every non-diagonal entry of a rate matrix we will use a partly **Gaussian prior** in

$$\pi_{ij}(R) \propto \begin{cases} \exp((R_{ij}-a)^2/\sigma) & , R_{ij} \geq 0 \\ 0 & , \text{else} \end{cases} \tag{3.1}$$

and for every rate matrix the likelihood

$$l(R) \propto exp(-\|m(R)_{i_1} - (f_D)_{i_1}\|_2). \tag{3.2}$$

We choose this likelihood because we want it to denote the probability that the first hitting times f_D occur given the rate matrix R. $l(R)$ lies in the interval $[0,1]$ but also preserves that it is maximal whenever the least squares distance $\|m(R)_{i_1} - (f_D)_{i_1}\|_2^2$ is minimal due to Lemma 1.7. This means that the likelihood is in this sense consistent with the intuitive and very common choice of the least squares estimation.
As the prior over $\mathcal{R}_n$ we use

$$\pi(R) = \prod_{i \neq j} \pi_{ij}(R).$$

In the Metropolis-Hastings algorithm we take the proposal distribution q with

$$q(R_{ij}, R^{(k)}) = \mathcal{U}[0,d] \tag{3.3}$$

for a $d > 0$ with starting value

$$R_{ij} \sim \mathcal{U}[0,d] \text{ for all } i \neq j. \tag{3.4}$$

We pick every entry independently of the others. After K iterations we obtain a sequence of rate matrices $R_1, \ldots, R_K$.

As the last step, we approximate the Bayesian path probability vector b by

$$\tilde{b}_j = \frac{1}{Z_{\tilde{b}}} \sum_{k=1}^{K} \nu_{f_D}(R_k) c_j(R_k) \tag{3.5}$$

where again $Z_{\tilde{b}}$ is a normalisation constant so that $\sum_{j=1}^{z} \tilde{b}_j = 1$.

3.2 Testing the method

Before we confront the method with real world data let us test it on toy data and make some observations about the meaningfulness of its results.

Example 12. Let us assume only three regions that are all neighbouring to each other and the rates between them given by

$$R = \begin{pmatrix} -0.101 & 0.1 & 0.01 \\ 0.1 & -0.2 & 0.1 \\ 0.1 & 0.1 & -0.2 \end{pmatrix}.$$

Let the start configuration be $\phi^1 = (1, 0, 0)^T$. Then the mean first hitting times of regions 2 and 3 are

$$m_{\phi^1}^{A_2} \approx 10 \text{ and } m_{\phi^1}^{A_3} \approx 19.7.$$

We simulate the spreading process on the corresponding state space of configurations and let us assume the first hitting times of this process are given in

$$(f_D)_1 = (0, 8, 17),$$

so relatively close to the expected first hitting times.
We take $K = 10000$ iterations of the Metropolis-Hastings algorithm with $a = 0.05$ and $\sigma = 1$ in the prior and $d = 0.2$ in the proposal distribution and end up with the following distributions over the rates.

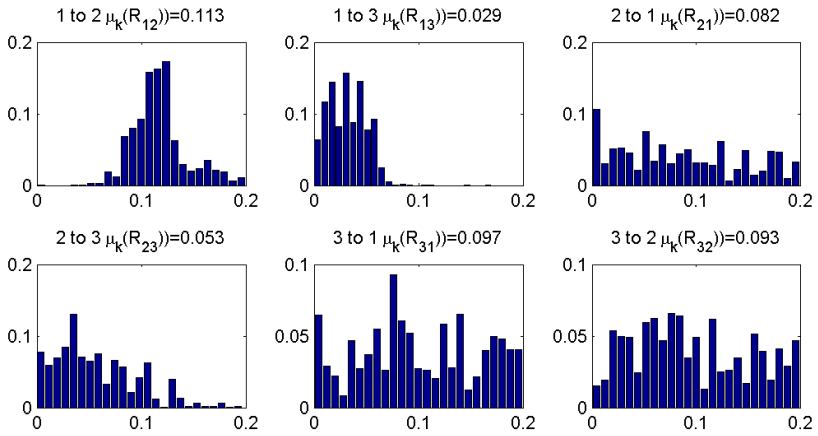


Figure 3.1: Histogram of the rates between neighbouring regions obtained from the network-based approach. The x-axis represents the value of the rate. The y-axis represents the probability of every value. First hitting times of a simulation given by $(f_D)_{i_1} = (0, 8, 17)$. $a = 0.5, \sigma = 1, d = 0.2$.

The distributions of R_{12} and R_{13} have a clear peak whereas the others are more homogeneous. The expected values (let us call them $\mu_K(R_{ij})$) make a relatively good fit with the rates in R. We should note that the rates that point into the start region (R_{21} and R_{31}) do not affect the hitting times of any state so these are meaningless. The fact that both of them are still close to the true value 0.1 as $\mu_K(R_{21}) = 0.082$ resp. $\mu_K(R_{31}) = 0.097$ is due to the choice of d in the proposal distribution because they are chosen uniformly at random from the interval $[0, 0.2]$ and should be close to 0.1 in average.
$\mu_K(R_{23})$ is about half the value of R_{23} and $\mu_K(R_{13})$ is almost three times the value of R_{13} but what the result clearly reproduces is

- The order of magnitude the rates lie in
- The fact that $R_{12} >> R_{13}$.

The Bayesian path probability vector of $R_1, \ldots, R_K$ is

$$\begin{aligned}
&\text{Path} \rightarrow \text{Probability} \\
&(3, 2, 1) \rightarrow 0.17 \\
&(3, 1, 2) \rightarrow 0.16 \\
&(2, 3, 1) \rightarrow 0.14 \\
&(2, 1, 3) \rightarrow 0.20 \\
&(1, 3, 2) \rightarrow 0.07 \\
&(1, 2, 3) \rightarrow 0.27
\end{aligned}$$

when we set $\mathbb{P}[I_1 = 1] = \mathbb{P}[I_1 = 2] = \mathbb{P}[I_1 = 3] = \frac{1}{3}$. If we assume that the start region I_1 is 1 then we get

$$\begin{aligned} &\text{Path} \to \text{Probability} \\ &(1,3,2) \to 0.20 \\ &(1,2,3) \to 0.80. \end{aligned}$$

In our example the spreading process took the path $(1,2,3)$ as we can see in f_D.

We discuss a second toy example that allows a seemingly disappointing observation.

Example 13. Let

$$R = \begin{pmatrix} -0.16 & 0.060 & 0.87 & 0.012 \\ 0.157 & -0.457 & 0.137 & 0.163 \\ 0.193 & 0.098 & -0.367 & 0.076 \\ 0.012 & 0.008 & 0.118 & -0.139 \end{pmatrix}.$$

with

$$m_{\phi^1}^{A_2} = 10.27, m_{\phi^1}^{A_3} = 8.02 \text{ and } m_{\phi^1}^{A_4} = 11.63.$$

Let us assume a simulation of the spreading process starting in ϕ^1 gives exactly $(f_D)_{i_1} = m(R)_{i_1}$. After $K = 100000$ iterations of the Metropolis-Hastings algorithm (10 times more than in the previous example) with $a = 0.1, \sigma = 1$ and $d = 0.2$ we get the following distributions over the rates.

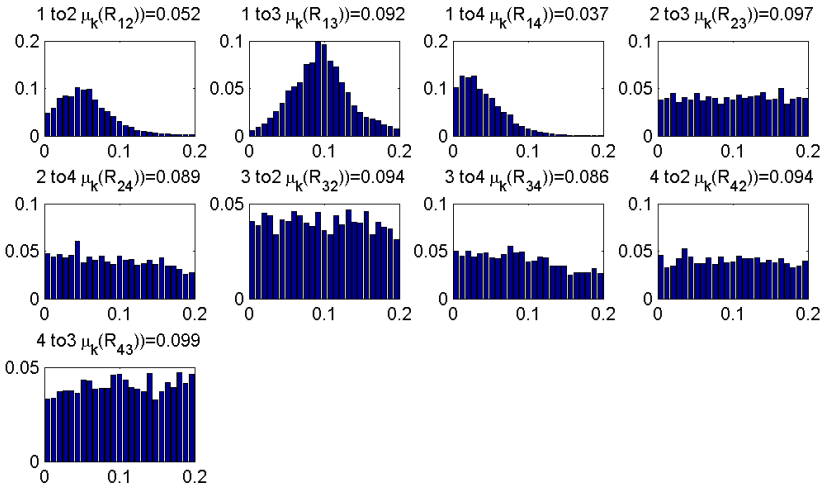


Figure 3.2: Histogram of the rates between neighbouring regions obtained from the network-based approach. The x-axis represents the value of the rate. The y-axis represents the probability of every value. First hitting times of a simulation given by $(f_D)_{i_1} = (0, 10.27, 8.02, 11.63)$. $a = 0.5, \sigma = 1, d = 0.2$.

Again, the values of R_{12}, R_{13} and R_{14} have clear peaks. The others are very homogeneous over the interval $[0, 0.2]$ so they can not be discerned in a thin region. This, however, does not mean that any rate matrix that is created by randomly choosing those entries from $[0, 0.2]$ fits well to f_D. As we have seen in section 2.4.1 different combinations of the entries can yield the same mean first hitting times. What those combinations are can not be seen in the histograms of the rates!

3.3 Processing the data

Let us now use bigger datasets that actually carry real content about the spreading of the woolly sheep. These data will be introduced and it will be explained how to infer f_D from them.

Dataset 1:

The data from TOPOI contain 565 findings of ovicaprids (different types of sheep and goats in general) over Eastern Europe and Western Asia. To be more precise, all findings lie within the **area of interest**

$$[13.5°E, 64.9°E] \times [25.6°N, 52.9°N].$$

Unfortunately, this dataset only contains findings of woolly sheep in a small patch of the area of interest. When we divide this area into regions, most of the regions will never have been infected according to this dataset because there are no findings in these regions. Therefore, the data about the woolly sheep alone is not

sufficient for the application of the network-based approach because for most of the regions, there are no first hitting times that serve as the evaluation criterion for every rate matrix. The dataset however, does contain well distributed data about ovicaprids. Although most sheep and goat types were spread over the area of interest much earlier than the earliest finding in the dataset, we will test our method on this dataset.
Every data point describes one finding of ovicaprid bones. Among others, it consists of the following properties:

X-coordinate
Y-coordinate
Time that the bones are dated back to

Dataset 2:
We use data that are created by the agent-based approach (section 2.1). As explained, it simulates the movements of several thousand farmers potentially carrying a woolly sheep with them in the area of interest over time. We generate a dataset that has the structure of Dataset 1 by taking multiple time stamps of the simulation (i.e. every 10 simulated years note the position and information state of every agent) and converting every agent that has the innovation (i.e. carries a woolly sheep with them) into a data point.

```
for every time stamp t ∈ {t_0, ..., t_p} do
    for every agent A do
        A has properties (X-coordinate,Y-coordinate, has innovation or not)
        if A has the innovation now then
            Create new data point (X-coordinate,Y-coordinate,t)
```

3.3.1 Division into regions

We have to divide the area of interest into regions in a reasonable way. This refers to both the number of regions n and where to draw the borders. TOPOI offers a division into nine regions that is based on both cultural and climate-related criteria (Figure 3.3). A significantly higher number of regions would yield a much higher computational cost (Figure 3.11).

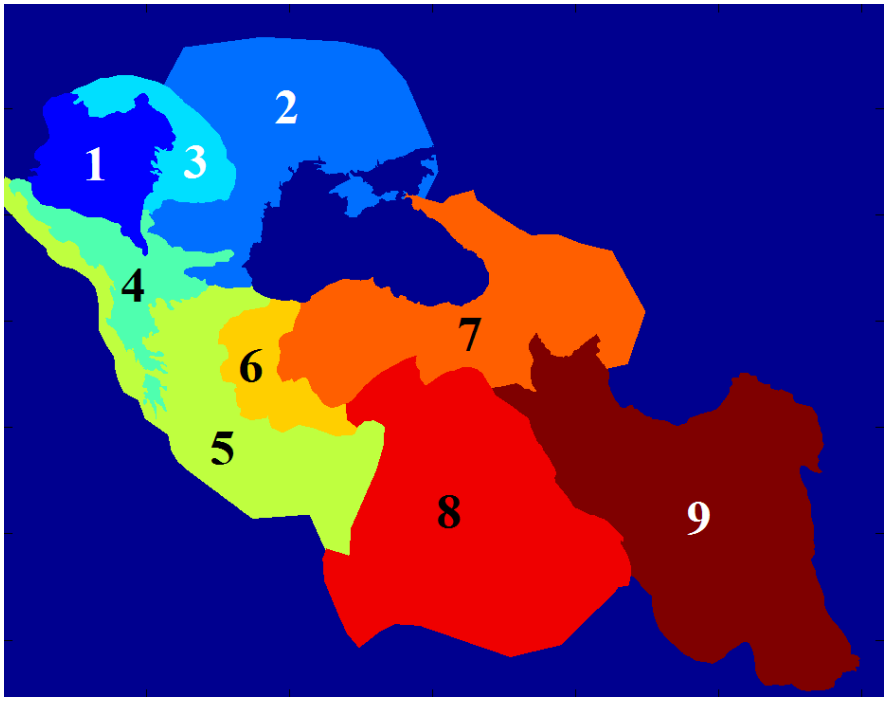


Figure 3.3: Division into nine regions.
Region 1 consists of: The Pannonian Basin
Region 2: The Pontic Steppe and the Lower Danube reach
Region 3: The Carpathian mountain range
Region 4: The Balkan and Peleponnes moutain range
Region 5: The Northern mediterranean coast region
Region 6: The Marmara region, Aegean region and Anatolian South coast
Region 7: Anatolia and Caucasus
Region 8: Syrian desert, Levant and Mesopotamia
Region 9: Azerbaijan, Zagros system and the Iranian highlands and mountains

Figure 3.4: The area of interest is highlighted. Image taken from Google Earth.

As not all regions share a border, only those entries of R should be positive that correspond to neighbouring regions. The set of those entries is

$$\begin{aligned}\mathcal{N} = \{&(1,3),(1,4),(2,3),(2,4),(2,5),(4,5),(5,6),(5,8),\\ &(6,7),(6,8),(7,8),(7,9),(8,9)\}.\end{aligned} \tag{3.6}$$

In Figure 3.3 it looks as if region 2 and 7 are neighbouring but we specify them as non-neighbouring because they only have a slim border and are separated by the Caucasus Mountains.

3.3.2 Inferring the first hitting times from the data

As explained earlier, we need to know the first hitting times f_D, defined as in (2.18). To that end, we first discretise the time space into slices of $\triangle t$ years beginning at the time of the first finding and ending when all regions were already infected (in Dataset 1 and 2 from 6750 BC until 500 BC). We then define a new process

$$\hat{Y}'_t : \mathcal{T} := \{-6750, -6750 + \triangle t, -6750 + 2\triangle t, \ldots, -500\} \to S \tag{3.7}$$

with

$$\begin{aligned}&\hat{Y}'_t = \phi \in S \text{ if there has been at least one finding}\\ &\text{of ovicaprids until } t \text{ in exactly the regions}\\ &\text{that are checked in } \phi.\end{aligned} \tag{3.8}$$

In order to make the process start at time 0, i.e. to make it consistent with the notation of first hitting times introduced in (1.45), we shift the process by 6750 and define

$$\hat{Y}_t = \hat{Y}'_{t+6750}. \tag{3.9}$$

Then again the first hitting time of j is $\tau_j := \tau^{A_j} = \inf\{t : \hat{Y}_t \in A_j\}$, where $A_j = \{\phi \in S : \phi(j) = 1\}$.

Example 14. Let us set $\triangle t = 250$ (years) and assume the following findings of at least one bone:

Time	Region
6750 BC	2
6400 BC	2
6300 BC	1
6050 BC	3

This yields that

$$\hat{Y}_0 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \hat{Y}_{250} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \hat{Y}_{500} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \hat{Y}_{750} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}. \tag{3.10}$$

Now actually using $\triangle t = 250$ years, Dataset 1 produces the following trajectory:

Region	Y_0	Y_{250}	Y_{500}	Y_{750}	Y_{1000}	Y_{1250}	Y_{1500}
1	0	1	1	1	1	1	1
2	0	0	1	1	1	1	1
3	0	0	0	0	1	1	1
4	0	0	0	1	1	1	1
5	0	1	1	1	1	1	1
6	1	1	1	1	1	1	1
7	0	0	0	0	0	1	1
8	1	1	1	1	1	1	1
9	0	1	1	1	1	1	1

Table 3.1: Spread of the ovicaprid over time according to the data.

This gives the following first hitting times:

From region \ To region									
	0	1	3	2	0	0	4	0	0
	0	0	2	1	0	0	3	0	0
	0	0	0	0	0	0	1	0	0
	0	0	1	0	0	0	2	0	0
	0	1	3	2	0	0	4	0	0
	1	2	4	3	1	0	5	0	1
	0	0	0	0	0	0	0	0	0
	1	2	4	3	1	0	5	0	1
	0	1	3	2	0	0	4	0	0

Table 3.2: First hitting times of regions according to Dataset 1 in 250 years.

The start configuration of $\hat{Y}_t$ is $\phi_{\text{start}} = (0,0,0,0,0,1,0,1,0,0)^T$. As explained, in order to evaluate the quality of a given rate matrix we compare its mean first hitting times starting from the configuration ϕ_{start} to $(f_D)_6$, i.e. we are interested in $(m^{A_1}_{\phi_{start}}, \ldots, m^{A_n}_{\phi_{start}})$.

Dataset 2 generated by the agent-based approach gives:

Region	Y_0	Y_{250}	Y_{500}	Y_{750}	Y_{1000}	Y_{1250}	Y_{1500}
1	0	0	0	0	1	1	1
2	0	0	0	1	1	1	1
3	0	0	0	1	1	1	1
4	0	0	0	1	1	1	1
5	0	0	0	1	1	1	1
6	0	0	1	1	1	1	1
7	0	1	1	1	1	1	1
8	1	1	1	1	1	1	1
9	0	1	1	1	1	1	1

Table 3.3: Spread of the wooly sheep over time according to a simulation of the agent-based model.

The corresponding first hitting times are:

From region \ To region								
0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
2	1	1	1	1	0	0	0	1
3	2	2	2	2	1	0	0	0
4	3	3	3	3	2	1	0	1
3	2	2	2	2	1	0	0	0

Table 3.4: First hitting times of regions according to Dataset 2 in 250 years.

We scale the first hitting times by $\frac{1}{250 \text{ years}}$ for the sake of simplicity. Then for Dataset 1

$$(f_D)^{(1)}_{i_1} = (1, 2, 4, 3, 1, 0, 5, 0, 1)$$

and for Dataset 2

$$(f_D)^{(2)}_{i_1} = (4, 3, 3, 3, 3, 2, 1, 0, 1).$$

The history of animal domestication [35], [36]

Animal domestication has been executed by humans since as early as 12.000 BC. The first animals to have been domesticated are believed to be dogs who first served as guards or hunting companions and later on for various reasons such as shepherd dogs. The Romans kept them as lap dogs because they believed that the warmth of the dog would cure stomach ache. People started using selective breeding to obtain dogs that matched their needs. For example, shepherd dogs were preferably white in order to distinguish them from wolves. The domestication of dogs took place in multiple places around the world in parallel as it was not necessarily intended by humans. Dogs saw the

benefit of the nearness to humans in food leftovers whereas humans were automatically warned about dangerous animals or hostile groups of people by the alarmed behaviour of dogs nearby.
One should not confuse domestication with taming. The latter means conditioned modification of the behaviour of only an individual animal whereas the former refers to intended genetic modifications that among other things lead to native tolerance of humans. A tiger that was raised in a zoo might be tamed but is not domesticated as its descendants will usually not be tame. On the other hand, elephants were in a way domesticated in India in 2000 BC and have been made to serve for means of transportation, entertainment, physical work and as war animals. But as they can fall into a rage at all times they can hardly be considered as tamed.
Between dogs and elephants there lie the domestications of many other animals: Sheep and goats followed around 8000-11.000 BC, about the same time pigs and cattle were domesticated.
Cats were moved into service 7500 BC and were helpful as they did not allow mice and rat epidemics to arise. Bigger wild cats, however, until now proved to have a too deep carnivorous instinct to be domesticated.
No earlier than 5000 BC, horse domestication took place in Eastern Europe.
As many benefits as domesticated animals have brought with them, they also carry the downsides of animal diseases such as rhinovirus, tuberculosis or influenza.
Evolutionary biologist Jared Diamond [35] named six criteria for an animal species to be suited for domestication in 1997:
1. Accept wide range of food
2. Grow up to maturity fast
3. Be able to breed in captivity
4. Docility
5. Do not panic easily
6. Agree to a social hierarchy

With those criteria in mind it makes sense why people have been able to domesticate some animals instead of others. For example, in contrast to pet cats wild cats regularly test the hierarchy within their group which contradicts criterion 6. Another example is the rhinoceros which is related to horses but is not known to be very docile and takes up to ten years to reach maturity.

3.4 Numerical Results

With prior and likelihood function as explained in section 3.1, we create results based on both datasets.
In both cases we use $a = 0.3, \sigma = 1$ in the prior and $d = 1$ in the proposal distribution of the Metropolis-Hastings algorithm.

Dataset 1:

Dataset 1 gives us the following distributions over all entries:

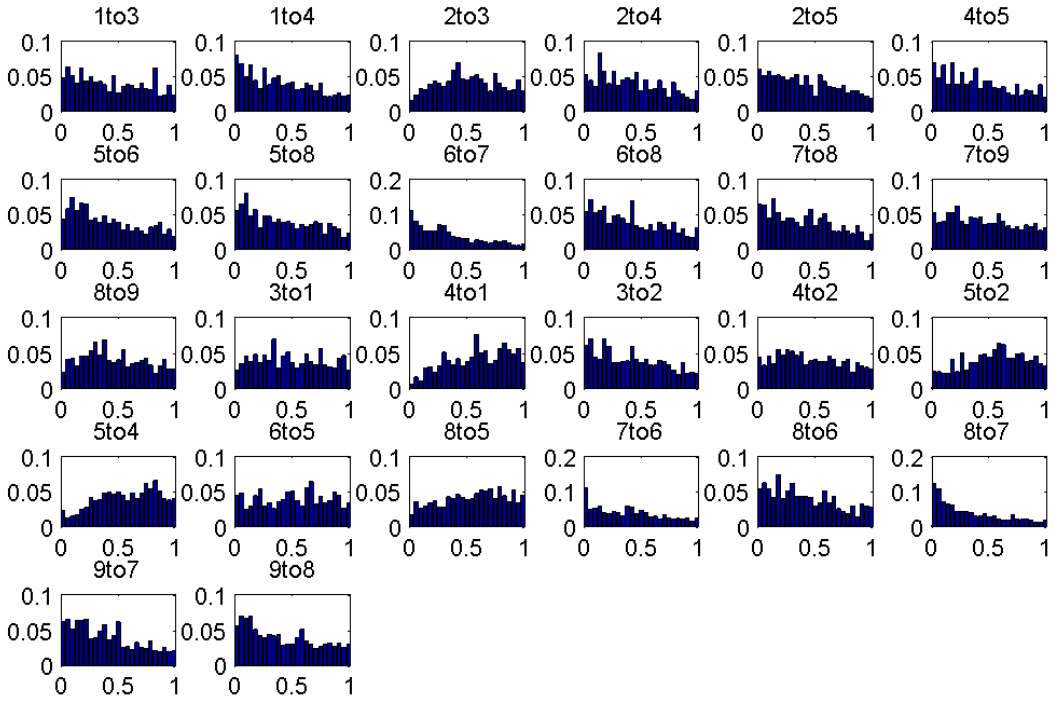


Figure 3.5: Histogram of the rates between neighbouring regions obtained from the network-based approach on Dataset 1. The x-axis represents the value of the rate. The y-axis represents the probability of every value. All rates between non-neighbouring regions are 0. $a = 0.3$, $\sigma = 1$ in the prior. $d = 1$ in the proposal distribution.

Even more evidently than in Example 13, the distributions of the rates do not have clearly visible peaks but are rather homogeneously distributed. Even the order of magnitude is difficult to deduce. As the probabilities for some of the rates decrease towards a value of 1 (e.g. 2 to 5, 9 to 7 or 6 to 8), we have reason to believe that they should be not much higher than 1 and probably lower. Since we scaled the first hitting times from of $\hat{Y}_t$ by a factor of $1/250$ years, the result suggests that an infected region should infect a neighbouring non-infected region within the next 250 years.

Still, we can find the probabilities for every path of the innovation spread. Figure 3.7 shows the approximated Bayesian path probability vector $\tilde{b}$.

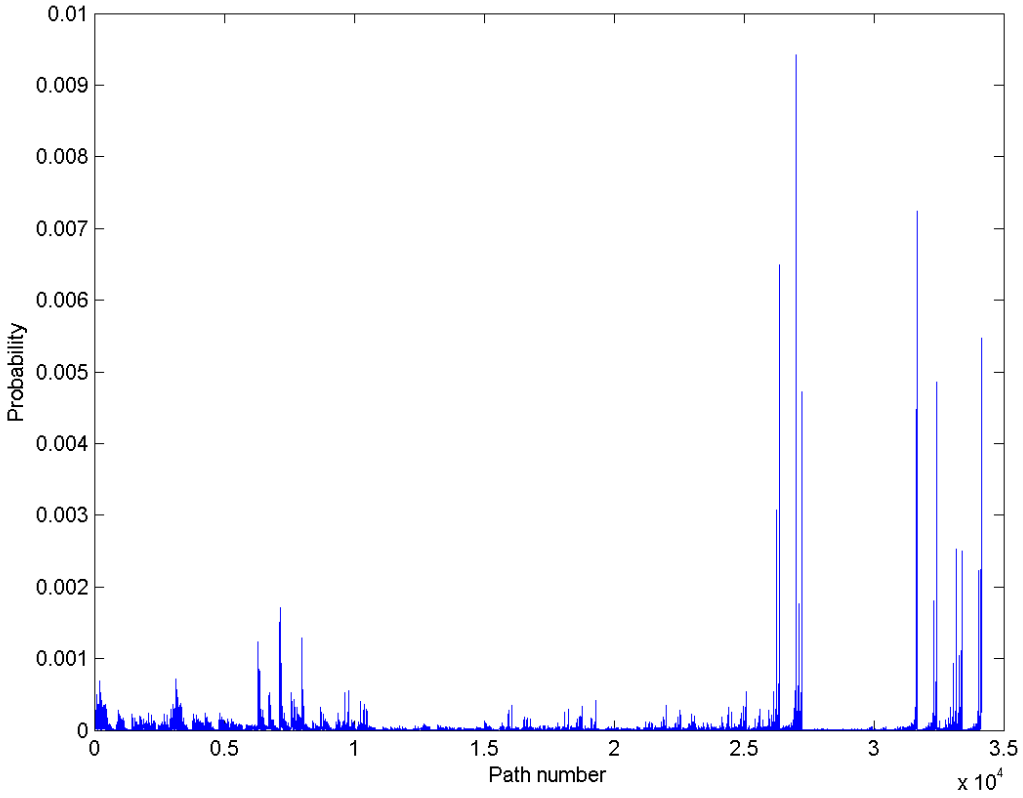


Figure 3.6: Probabilities of all paths based on the rates obtained from Dataset 1. For the probability of the start region, we set $\mathbb{P}[I_1 = i] = \frac{1}{n}$.

The three most probable paths are:

(1) : $3 \to 2 \to 1 \to 4 \to 5 \to 6 \to 8 \to 9 \to 7,$ Probability: 0.0094

(2) : $3 \to 2 \to 1 \to 4 \to 5 \to 6 \to 9 \to 8 \to 7,$ Probability: 0.0086

(3) : $2 \to 3 \to 1 \to 4 \to 5 \to 6 \to 8 \to 9 \to 7,$ Probability: $0.0072.$

Assuming uniform probabilities for the start region the spreading process is likely to have started in the North-West and ended in the South-East. Region 7 is always infected last. This is the case because it is only neighbouring to three regions (Figure 3.3) and the rates that point into it are lower than most of the other rates (Figure 3.5).

Archaeological evidence, however, suggests that the start region is region 8, therefore we now use $\mathbb{P}[I_1 = 8] = 1$.

Then $\tilde{b}$ restricted to the paths that start in region 8 is

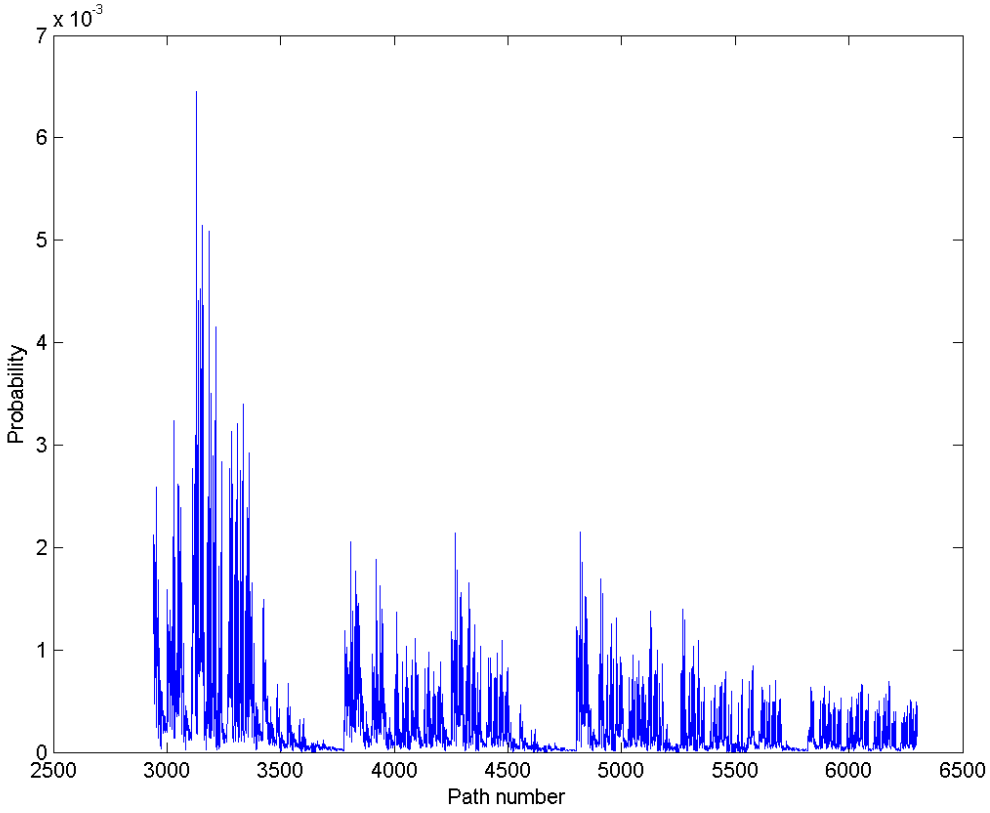


Figure 3.7: Probabilities of all paths that start in region 8 based on the rates obtained from Dataset 1.

The three most probable paths that start in region 8 are:

(1) : $8 \to 9 \to 5 \to 6 \to 4 \to 2 \to 1 \to 3 \to 7,$ Probability: 0.0064

(2) : $8 \to 9 \to 5 \to 6 \to 2 \to 3 \to 1 \to 4 \to 7,$ Probability: 0.0051

(3) : $8 \to 9 \to 5 \to 4 \to 6 \to 2 \to 1 \to 3 \to 7,$ Probability: $0.0051.$

These paths do not differ much as they all see region 9 and afterwards 5 infected first. To be more precise, the first and third paths only differ in two entries and the beginnings are identical. Moreover, their probabilities among all of the other thousands of possible paths are not high enough to yield a reliable prediction about how the spreading of the innovation took place.
Therefore, it makes sense to summarise paths by a criterion, e.g. their first two infected regions. With start region 8 this gives:

Region	1	2	3	4	5	6	7	9
Probability	0	0	0	0	0.41	0.15	0.07	0.37

Table 3.5: Probabilities for every region to be infected second after region 8 according to Dataset 1.

Surprisingly, region 5 is most likely to be infected second instead of region 9. The reason for that is that the probabilities of the paths in which region 5 is the second infected region (denoted by **85–paths**) are much more homogeneous than of the ones that go first from 8 to 9 (89-paths) (or 6 or 7). Due to the neighbourhood specification, there are also many more different 85-paths than 86-, 87- and 89-paths namely 2100 compared to 420 each. That means, if region 9 is infected as the second region the ongoing path is more predetermined than if it is region 5. So, the probabilities for all 85-paths sum up to the highest value but that can not be inferred from the most likely paths.

Ignoring the archaeological evidence about the start region, we now solve (2.11) in order to find a distribution about the start region.
With no information about the end region, we could choose a uniform distribution about the regions. But we set $e_5 = 0$ because as we can see on Figure 3.3 region 5 separates the regions 1,2,3 and 4 from the regions 6,7,8 and 9. This means that all paths that go from one of those two sets of regions to the other one have to run through 5. Hence region 5 cannot be the end region.

$$e = (0.11, 0.11, 0.11, 0.11, 0, 0.11, 0.11, 0.11, 0.11)$$

The matrix A given in (2.12) is not invertible in this case. Thus, we seek a least squares solution of (2.12), i.e. we want to find x such that

$$\begin{aligned} &\|Ax - r\|_2 \\ &\text{s.t. } x \geq 0, \sum_{i=1}^{n} x_i = 1 \end{aligned} \tag{3.11}$$

is minimal.
As a reminder, by construction of the linear system (2.11) $x_i = \mathbb{P}[I_1 = i]$.
The solution is

$$x = (0, 0, 0.37, 0, 0, 0, 0.63, 0, 0).$$

which indicates that either region 7 or 3 must have been start region. But since we know that according to Dataset 2 the start region is region 8 this result seems unreliable and is probably is partly due to the lower rates that point into region 7

and it is unreasonable that the result is so one-sided its meaning is questionable. Even very different choices for e yield the same result.

Dataset 2:

We now use Dataset 2 generated by the agent-based approach. It gives the following distributions over the rates:

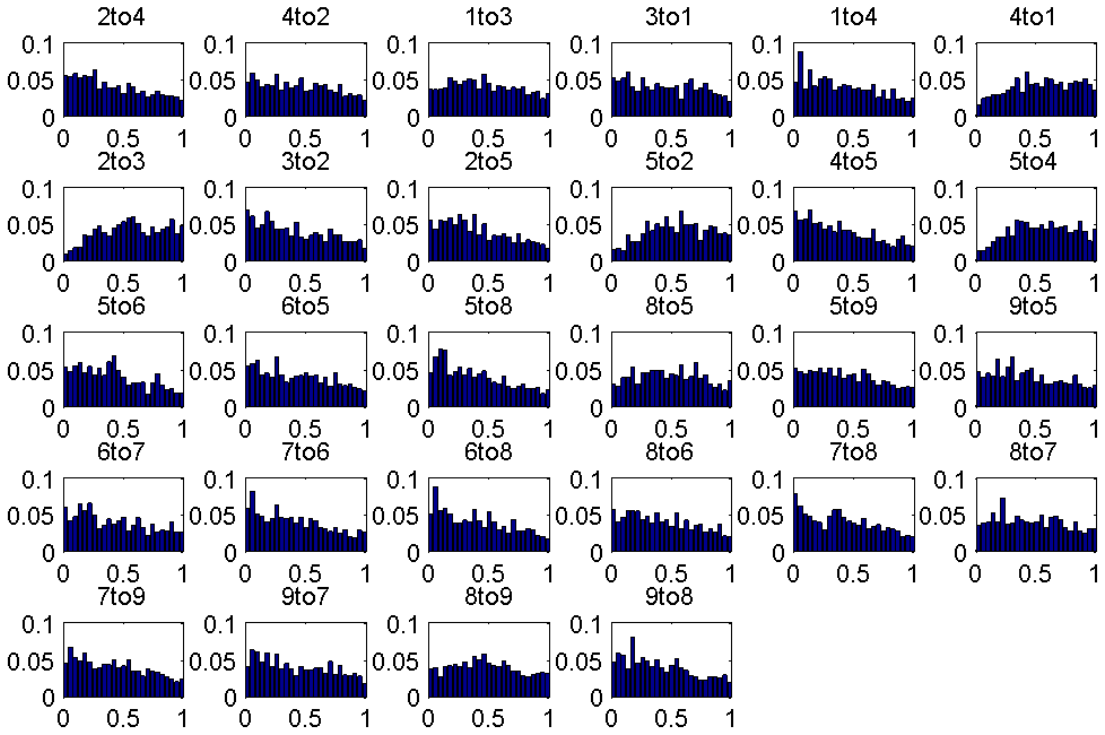


Figure 3.8: Histogram of the rates between neighbouring regions obtained from the network-based approach on Dataset 2. The x-axis represents the value of the rate. The y-axis represents the probability of every value. All rates between non-neighbouring regions are 0. $a = 0.3$, $\sigma = 1$ in the prior. $d = 1$ in the proposal distribution.

The path probability vector with uniformly weighted start region is heterogeneous.

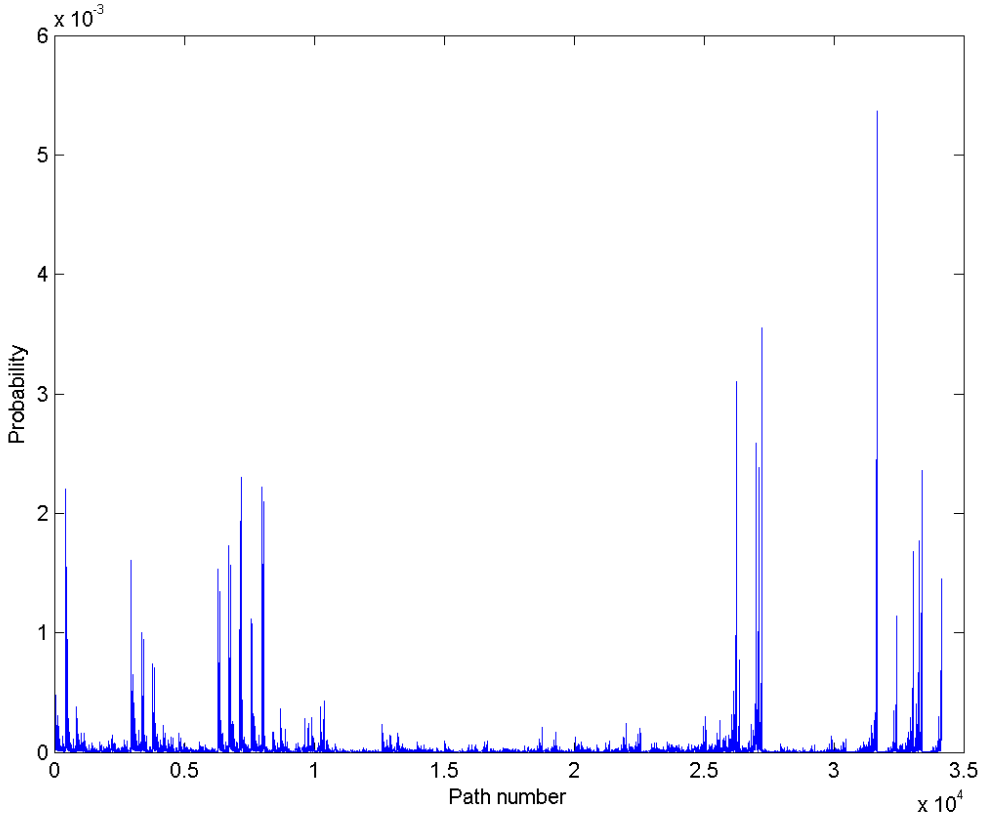


Figure 3.9: Probabilities of all paths based on the rates obtained from Dataset 2. For the probability of the start region we set $\mathbb{P}[I_1 = i] = \frac{1}{n}$.

Assuming that $\mathbb{P}[I_1 = 8] = 1$ we get

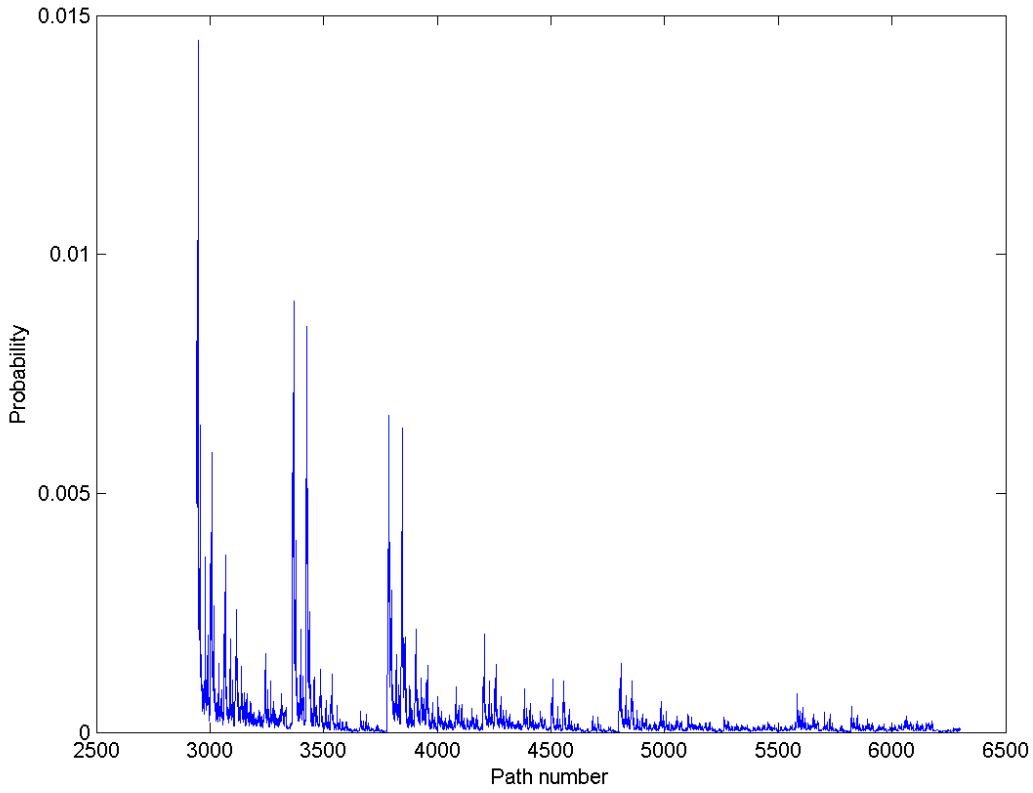


Figure 3.10: Probabilities of all paths that start in region 8 based on the rates obtained from Dataset 2. For the probability of the start region we set $\mathbb{P}[I_1 = 8] = 1$.

The three most probable paths are:

(1) : $8 \to 9 \to 7 \to 6 \to 5 \to 2 \to 3 \to 1 \to 4,$ Probability: 0.0145

(2) : $8 \to 9 \to 7 \to 6 \to 5 \to 2 \to 4 \to 1 \to 3,$ Probability: 0.0103

(3) : $8 \to 7 \to 9 \to 6 \to 5 \to 2 \to 3 \to 1 \to 4,$ Probability: $0.0090.$

The same phenomenon arises that we saw with Dataset 1: The three paths are very similar and their probabilities are fairly low. Again we summarise all paths that start in region 8 by their second infected region and get a similar result as with Dataset 1 because regions 5 and 9 are the most likely to be infected second.

Region	1	2	3	4	5	6	7	9
Probability	0	0	0	0	0.25	0.21	0.20	0.33

Table 3.6: Probabilities for every region to be infected second after region 8 according to Dataset 2.

For computing an estimate about the start region again the matrix A is not invertible.

With again

$$e = (0.11, 0.11, 0.11, 0.11, 0, 0.11, 0.11, 0.11, 0.11)$$

the least squares solution to (2.12) is

$$x = (0, 0, 0.49, 0, 0, 0, 0.51, 0, 0).$$

This again suggests that either region 3 or 7 was the starting region. Analogously to Dataset 1 the start region in Dataset 2 is region 8, so this method of determining the start region seems unreliable.

Computation time:
It has been mentioned that a finer division into regions would bring a much higher computational cost with it. This is mostly due to the calculation of all path probabilities. The amount of time this takes in our MATLAB implementation increases exponentially with additional regions (Figure 3.11). On the Windows machine used (2.6 GHz, 8 GB RAM) the computation of the probabilities of all paths for one rate matrix in a division into nine regions with every region neighbouring to every other one takes approximately 30 minutes.

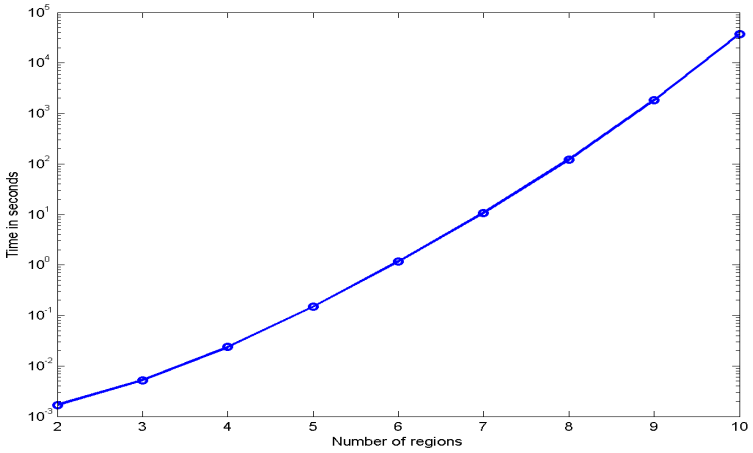


Figure 3.11: Logarithmic plot of the time in seconds for the computation of path probabilities in our MATLAB implementation for different numbers of regions. We assumed that all regions are neighbouring to each other.

Since in general not all regions are neighbouring to each other, in that case the amount of computation time is lower because there are fewer possible paths but its increase is very similar. With a similar portion of neighbouring regions as in our division into nine regions, the computation time develops approximately exponentially,

too (Figure 3.12), but the absolute time needed is lower by a factor of 5 to 10:

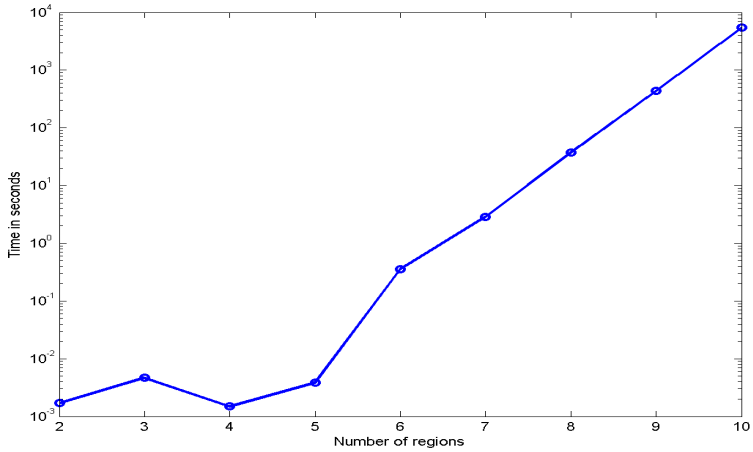


Figure 3.12: Logarithmic plot of the time in seconds for the computation of path probabilities in our MATLAB implementation for different numbers of regions. For the graph on the left we assumed that all regions are neighbouring to each other. For every number of regions the portion of neighbouring regions is similar to the one that we used in our division into nine regions.

4 Outlook

The network-based approach on modelling and analysing the spread of an innovation consists of the following basic steps:

1. Assume that multiple regions can have the innovation at the same time and a region never loses the innovation.
2. Model a stochastic process between the regions that does not violate the assumptions.
3. Construct a Markov process on the state space of configurations.
4. Estimate rates between regions by comparison of mean first hitting times of a rate matrix and empirical first hitting times from data.
5. Calculate probabilities for every order in which the regions got to have the innovation.

This leaves room for several ideas to improve the model and implement concepts that have already been explained in this thesis but are beyond the scope of it.

Originating paths

The concept of originating paths (Def. 2.3) is not difficult to implement. The reason why no results on it are included in this thesis is the computational cost. The division into nine regions and the subsequent neighbourhood (3.6) yields 34152 (non-originating) paths. The number of originating paths, however, is $8!$ the number of it, which is 1.377.000.000. Computing the probability for every one of those paths for multiple rate matrices (as done in (3.5)) is highly infeasible on most computers.

Delay of the spreading

Under the current state of the model a region can either be infected by an innovation or not. It could yield an improvement to incorporate a more continuous measurement of the spreading of the innovation within a region.
This could be done by augmenting the space of configurations from $S = \{0,1\}^n \backslash \{0\}^n$ to

$$S^k := \{0,k\}^n \setminus \{0\}^n$$

where k is the number of different states of the infection, i.e. we measure how well the innovation is known in a region i by a number from $\{0, \dots, k\}$. Let us call that value the **information state** of a region and make two additional definitions.
1: ϕ_1 is **i–covered** by ϕ_2 if $\phi_2(l) \geq \phi_1(l) \; \forall \; l$.
2: ϕ_1 is **simply i–covered** by ϕ_2 if ϕ_1 is i-covered by ϕ_2 and $\|\phi_2 - \phi_1\|_1 = 1$.
Then the spread rates from one region i to another region j depend on the information state ι of region i. Intuitively, they should increase for increasing information state of region i because the further the innovation has spread in one region the more likely should it be that it jumps over to another region and increases the information state of that other region. We denote the rate by $\lambda(i, j, \iota)$. This would give up to k times the number of parameters that have to be fitted. For configurations ϕ_u and ϕ_v, the rate from ϕ_u to ϕ_v is given by

$$\bar{R}_{uv} = \begin{cases} \sum\limits_{l \; s.t. \; \phi_u(l) > 0} \lambda(l, j, \phi_u(l)), & \text{if } \phi_u \text{ is simply i-covered by } \phi_v \text{ and } \|\phi_u - \phi_v\|_1 = \phi^j \\ -\sum\limits_{l=1, l \neq u}^{|S^k|} \bar{R}_{ul}, & \text{if } u = v \\ 0, & \text{else.} \end{cases} \tag{4.1}$$

The way the innovation spreads within a region is another parameter that has to be quantified. One approach to that is to have the information state increase by 1 regardless of other regions after an exponentially distributed waiting time with parameters $\mu_1, \dots, \mu_n$ and call these parameters the **information rates**.
Together with the spread rates an infected region can both influence its own information state (governed by the information rates) and the information state of other regions (governed by the spread rates). The information rates influence the rates between configurations as given in (4.1) by $\lambda(j, j, \phi_u(j)) = \mu_j$, whereas the spread rates are represented by $\lambda(l, j, \phi_u(l))$ for $l \neq j$.
Then denoting by $A_j := \{\phi \in S^k | \phi(j) = k\}$, i.e. all configurations that have region j fully infected, the mean first hitting time $m_\phi^{A_j}$ from a configuration ϕ to this set is analogously to (2.23) given by

$$m_{\phi_i}^{A_j} = \begin{cases} 0 & , i \in A \\ (\bar{R}_{i1} + \dots + \bar{R}_{in})^{-1} (1 + \sum\limits_{\phi_u \in S^k, u \neq i} m_{\phi_u}^{A_j} \bar{R}_{ij}) & , i \notin A \end{cases} \tag{4.2}$$

The size of the information rates in relation to the size of the spreading rates affects the model: For much bigger information rates than spread rates, the information state increases much more independently from the other regions than given relatively high spread rates. That relation is, however, hard to estimate at this point

without scientific archaeological expertise. Moreover, and that yields a profound drawback of the implementation of the delay into the model, for the fitting process data about the information state would be required because the question arises: When do we assign to a region the information state $1, 2 \dots k$, i.e. if there are 100 findings in that region, is an information value of 2 appropriate or rather of close to k?

Clustering Regions

The division into nine regions was done with the help of scientific archaeological and and geological expertise. Although this can be trusted, it would be interesting to create a division that does not rely on people's knowledge but has a more mathematical basis. To that aim, we could make a division into many more regions (e.g. ≈ 50), apply the network-based approach to it and then cluster the regions into nine groups again by the rates between them.
A much higher number of regions, however, would yield an infeasible computational cost without a supercomputer at hand. But even with only those nine regions we could perform a cluster analysis, drawing on one of the many common algorithms for that, and obtain a very compact division, for example into North, West and South. We could then summarise the paths by them, e.g. figure out what the probability is for the innovation to go from region 8 (South) rather North than West.

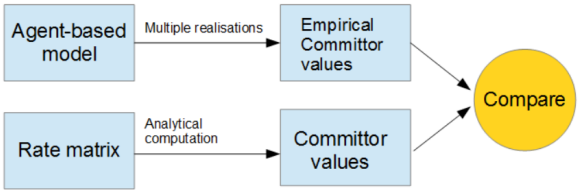
Committors

Building on a compact clustering, e.g. into three groups of regions $North$, $South$ and $West$, we could use committors, explained in section 1.1.5, to compute the probabilities for the innovation to reach e.g. North rather than West. As the discrete forward committor $q_i^+ := \mathbb{P}[\tau^B < \tau^A | X(0) = i]$ describes the probability of a Markov process to hit a set B earlier than another set A given that the process started in state i we would define

$$\begin{aligned} B &= \{\phi \in S | \phi(j) = 1 \text{ for any } j \in North\} \\ A &= \{\phi \in S | \phi(j) = 1 \text{ for any } j \in West\} \\ i &= \phi^8, \end{aligned}$$

if we again assume region 8 as the start region, and then compute the q_i^+.
Even without a clustering, we could use committors by again making use of the agent-based model: Let us run multiple simulations of the agent-based model and keep track with which relative frequency a region i is reached before a different

region j. We could then use these values as a reference for the corresponding discrete forward committors of any given rate matrix instead of only using its mean first hitting times, i.e.



As reality as we know it only happens once, i.e. is only one simulation of what it could be, we would have to rely on the accuracy of the agent-based model and use it to produce multiple simulations.

5 Summary

In this thesis a method described, defended and tested that reconstructs the spreading path of an innovation given rough information about when it was spread in which part of the world. It models the spreading process as a time-continuous Markov process on a state space of configurations where every state contains information about the infection of a certain region of the area of interest. Every time-continuous Markov process is governed by a rate matrix that contains the frequencies of transitions between the states. We measure how well a given rate matrix mimics the frequencies of the real world process by comparing its mean first hitting times, i.e. the expected time for a state - or, here rather, region - to be reached, with the earliest time every region was infected by the innovation according to the data. We are not solely interested in the best fitting rate matrix for multiple reasons:

> The best fitting rate matrix might not be unique.
>
> The data are potentially incomplete in the sense that the spreading process might have taken place differently than displayed by the data.
>
> The path that the innovation took in reality need not have been the most probable one. Thus the best fitting rates could be very different from the actual transition frequencies.

Therefore we use a Bayesian framework that instead of only a single element gives a distribution over the space of all rate matrices. Since that space is far too big to be analysed analytically, we take the Metropolis-Hastings algorithm to sample rate matrices from it according to that distribution. An advantage of this Bayesian approach is its robustness to errors in the data.
As a final result, we can subsequently compute the probabilities for every spreading path based on the result from the Bayesian framework. By path we mean the order in which the regions were infected.
We applied this method to data about the spread of ovicaprids (sheep and goat) for a test purpose and on data about the spread of the woolly sheep in the Neolithic Period. The results do not allow for a clear determination of the rates because within a certain order of magnitude very different combinations of rates between regions make for a good fit with the data. We can, however, give reasonable estimates about the path probabilities. Assuming by archaeological indication that the woolly sheep originates from Mesopotamia, according to the results it spread East to Azerbaijan first and then North to Anatolia. Only then did it spread West towards Europe.

A downside of particularly the Bayesian aspect of the method is the high computational cost. Even though we used the Metropolis-Hastings algorithm in order to estimate the posterior, the computation of the path probabilities for 10000 rate matrices on a space of nine regions takes multiple hours. Moreover, the method requires archaeological findings in every region in order to compare the first hitting times of the real process to the mean first hitting times of a given rate matrix.
In conclusion, our method only needs data about findings and a division of the area of interest into discrete regions as input in order to give an answer as to along which path a certain innovation spread. A benefit of it is that it uses very few assumptions about the innovation whose spread it is applied to. Given data about e.g. copper, the wheel or any other innovation from ancient times we could apply this method to it.

The spread of the smart phone [34]

Smart phones, commonly known as mobile devices with means to make phone calls and access the Internet, have spread across the world faster than almost any other major invention in human history.
The first version of a smart phone was the IBM Simon Personal Computer©in 1993 as it featured a touch screen. 29th June 2007, the day the Apple© iPhone© went on sale in the United States, marks the day the saturation of smart phones increased rapidly. After only 715.000 smart phones of various brands had been sold until then iPhone clocked at over one million after its first quarter in October 2007.
Within three years the smart phone increased its saturation of the U.S American population from 10% to 40% which is only about half the time that it took for the radio and one fifth of the the time it took for the computer.
The iPhone debuted in Europe in November 2007 and not before November 2009 it was released in China with the competing brands following shortly after. But by 2013, over 400 million Chinese were smart phone users. In Africa smart phones have struggled to replace the standard cell phone but by 2015 200 million people had become owners of a smart phone.
Overall, in 2015, 1.86 billion people worldwide were smart phone users. That is an increase by a factor of over 900 compared to October 2007.

Bibliography

[1] M. Sarich, "Projected transfer operators: Discretization of Markov processes in high-dimensional state spaces," Ph.D. dissertation, Freie Universität Berlin, 2012.

[2] J. Podlesny, "Multilevel methods for eigenvalue problems with stochastic matrices," Master's thesis, Freie Universität Berlin, 2014.

[3] A. Nielsen, "Von Femtosekunden zu Minuten - ein verallgemeinerter Operatoransatz in der Molekülsimulation," Master's thesis, Freie Universität Berlin, 2012.

[4] I. D. Dinov, N. Christou, and R. Gould, "Law of Large Numbers: The Theory, Applications and Technology-based Education," *J Stat Educ. 2009 March ; 17(1): 1–19*, 2009.

[5] G. Grimmett and D. Welsh, *Probability: An Introduction*. Oxford Science Publications, ISBN 0-19-853264-4, 1986, p. 14.

[6] R. Durrett, *Probability: Thoery and Examples*. Cambridge University Press, 2010.

[7] S. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. Springer, 1993.

[8] F. Noé, B. Keller, and J.-H. Prinz, "Lecture notes on stochastic processes," DFG Research center Matheon, FU Berlin, 2013.

[9] P. Metzner, "Transition path theory for markov processes," Ph.D. dissertation, Freie Universität Berlin, 2008.

[10] V. Spokoiny and T. Dickhaus, *Basics of Modern Mathematical Statistics*. Springer, 2015.

[11] S. Roeblitz and P. Deuflhard, *A Guide to Numerical Modelling in Systems Biology*. Springer, 2015.

[12] C. Schütte, "Parameter estimation: How difficult is it really?" Berlin, 9th-11th November 2016, lecture held at Workshop "Robust modelling in process optimization".

[13] ——, "Bayesian analysis and deep learning: New approaches," Berlin, 9th-11th November 2016, lecture held at Workshop "Robust modelling in process optimization".

[14] N. Chopin, S. Gadat, B. Guedj, A. Guyader, and E. Vernet, "On some recent advances in high dimensional Bayesian statistics," *Esiam: Proceedings and surveys, Vol. 51, p. 293-319*, 2015.

[15] I. Yildirim, "Bayesian Inference: Metropolis-Hastings Sampling," University of Rochester: Department of Brain and Cognitive Sciences, 2012.

[16] L. Tierney, "Markov chains for exploring posterior distributions," *The Annals of statistics*, 1994.

[17] C. E. Rasmussen and Z. Ghahramani, "Bayesian Monte Carlo," Gatsby Computational Neuroscience Unit, University College London.

[18] J. Li, "Course notes: Stochastic modeling," 2015, Penn State University, http://www.personal.psu.edu/jol2/course/stat416/hints.html.

[19] G. Takahara, "Course notes: Stochastic processes," Queen's University at Kingston: Department of Mathematics and Statistics, 2016.

[20] G. Amir, "Continuous time Markov chains," Bar Ilan University, http://u.math.biu.ac.il/ amirgi/CTMCnotes.pdf.

[21] J. Zonker, L. Helfmann, and N. Conrad, "Mathematical modelling for the spreading of innovations in ancient times," 2017, in preperation.

[22] G. Abramson, "Mathematical modeling of the spread of infectious diseases," 2001, a series of lectures given at PANDA, UNM.

[23] R. Fisher, "The wave of advance of advantegeous genes," *The annals of Eugenics*, 1937.

[24] J. Coulton, "Lifting in early greek architecture," *The Journal of Hellenic Studies*, 1974.

[25] U. S. Dixit, M. Hazarika, and J. P. Davim, *A Brief History of Mechanical Engineering*, 2016.

[26] M. Radivojevic, T. Rehren, E. Pernicka, D. Sljivar, M. Brauns, and D. Boric, "On the origins of extractive metallurgy: New evidence from Europe," *Journal of Archaeological Science*, 2010.

[27] M. F. Small, "String theory: The tradition of spinning raw fibers dates back 28,000 years (at the museum)," *Natural History*, 2002.

[28] N. Goren-Inbar, N. Alperson, M. E. Kislev, O. Simchoni, Y. Melamed, A. Ben-Nun, and E. Werker, "Evidence of hominin control of fire at gesher benot ya'aqov, Israel," *Science*, 2004.

[29] D. Hunter, "Papermaking. The history and technique of an ancient craft," *Dover publications, New York Starbird, Margaret*, 1943.

[30] J. Needham, *Science and civilization in China: Volume 4, Physics and Physical Technology; Part 1, Physics.*, 1962.

[31] T. C. Kriss and V. M. Kriss, "History of the operating microscope: From magnifying glass to microneurosurgery," *Neurosurgery*, 1998.

[32] P. Osborne, *The Mercator Projections: The normal and transverse Mercator projections on the sphere and the ellipsoid with the full derivations of all formulae*, 2013.

[33] B. E. Colless, "The history of the alphabet: An examination of the Goldwasser hypothesis," *Cuadernos del centro de estudios de historia del antiguo oriente*, 2014.

[34] M. DeGusta, "Are smart phones spreading faster than any technology in human history?" https://www.technologyreview.com/s/427787/are-smart-phones-spreading-faster-than-any-technology-in-human-history/, 2012.

[35] J. Lear, "Our furry friends: The history of animal domestication," *Journal of Young Investigators*, 2012.

[36] C. A. Driscoll, D. W. Macdonald, and S. J. O'Brien, "From wild animals to domestic pets, an evolutionary view of domestication," *Proceedings of the National Academy of Sciences*, 2009.

[37] C. Becker, N. Benecke, A. Grabundzija, H.-C. Küchelmann, S. Pollock, W. Schier, C. Schoch, I. Schrakamp, B. Schütt, and M. Schumacher, "The textile revolution. Research into the origin and spread of wool production between the Near East and Central Europe," *eTopoi Journal for ancient studies, Special Volume 6 (2016): Space and Knowledge. Topoi Research Group Articles, ed. by Gerd Graßhoff and Michael Meyer, pp. 102–151.*, 2016.

[38] C. Lemmen, D. Gronnenborn, and K. W. Wirtz, "A simulation of the neolithic transition in Western Eurasia," *Journal of Archaeological Science 38*, 2011.

[39] C. Lemmen, "Mechanisms shaping the transition to farming in Europe and the North American woodland," *Archaeology, ethnology and anthropology of Eurasia*, 2013.

[40] G. J. Ackland, M. Signitzer, K. Stratford, and M. H. Cohen, "Cultural hitchhiking on the wave of advance of beneficial technologies," *Proceedings of the National Academy of Sciences*, 2007.

[41] E. M. Rogers, *Diffusion of Innovations*. A Division of Macmillan Publishing Co., Inc, 1983.