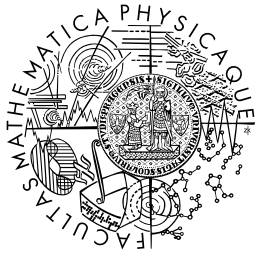


FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Jan Kislinger

Dynamic Fare Model

Department of Probability and Mathematical Statistics

Supervisor of the master thesis:	Doc. RNDr. Petr Lachout, CSc.
Study programme:	Mathematics
Study branch:	Probability, Mathematical Statistics and Econometrics

Prague 2017

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In Prague on January 4, 2017 Jan Kislinger

Název práce: Dynamický model ceny jízdného

Autor: Bc. Jan Kislinger

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Doc. RNDr. Petr Lachout, CSc., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Problém hledání dynamického modelu pro ceny jízdného se skládá ze dvou úloh – odhadování poptávky po vlakových jízdenkách a vícestupňová optimalizace ceny jízdného. V této práci představujeme model nehomogenního markovského řetězce, který používáme pro vývoj prodeje jízdenek. Z důvodu velikosti stavového prostoru je nutné řešit optimalizační úlohu pomocí simulované optimalizace. Řešení jednostupňového a dvoustupňového problému je implementováno v jazyce R. Před samotným praktickým problémem shrnujeme teorii nehomogenních markovských řetězců, kde se podrobněji zaměřujeme na procesy se separovatelnou nehomogenitou. Dále navrhujeme metody odhadování intenzity markovského procesu založené na teorii maximální věrohodnosti. Také popisujeme a srovnáváme dva algoritmy simulované optimalizace.

Klíčová slova: nehomogenní markovský proces, odhadování intenzity, simulovaná optimalizace, dynamické jízdné

Title: Dynamic Fare Model

Author: Bc. Jan Kislinger

Department: Department of Probability and Mathematical Statistics

Supervisor of the master thesis: Doc. RNDr. Petr Lachout, CSc., Department of Probability and Mathematical Statistics

Abstract: The problem of creating dynamic fare model consists of two tasks – estimating demand for train tickets and multistage optimization of price of fare. We introduce a model of inhomogeneous Markov process for the process of selling the tickets in this thesis. Because of the complexity of the state space the optimization problem needs to be solved using simulation methods. The solution was implemented in R language for single-stage and two-stage problems. Before this application we summarize the theory of inhomogeneous Markov process with special attention to process with separable inhomogeneity. Then we propose methods for estimating the intensity using maximum likelihood theory. We also describe and compare two algorithms for simulated optimization.

Keywords: inhomogeneous Markov process, estimating intensity, simulated optimization, dynamic fare

I would like to thank doc.RNDr. Lachout Petr, CSc., my supervisor of this thesis, for leading me in the right direction. I also thank to Prof. RNDr. Antoch Jaromír, CSc., who helped me understand and use simulation methods in this thesis.

Contents

Introduction

In this thesis we build a model for pricing train fares using continuous-time inhomogeneous Markov process. Such model can be useful for any kind of service that happens in predetermined time. The train transportation has its own specific competition on the demand side. That is, two passengers can compete each other even if they are not buying the same service. By the same service we mean transportation between the same stations in this particular case.

Williams [2013] solved similar problem with airplane ticket prices. He mainly studied how the demand and its price elasticity changes in time and how dynamic models can significantly increase the return form sold tickets.

We suppose that the demand for the train tickets depends on their price. This indicates that the problem is of endogenous character because the decision influences the stochastic part of the problem. Sometimes, we say that the randomness of the problem is decision-dependent. Some properties of exogenous problems can be found in Dupačová [2006].

We propose to model the occupancy of the train (number of tickets sold for each pair of boarding and exiting stations) as states of Markov process. Problem is that the number of states (all possible combinations of number of sold tickets) is too high to calculate transition probabilities analytically. Instead of that we estimate the expected return from sold tickets from simulated demands. This procedure is computationally demanding itself because we need to generate inhomogeneous Markov process in order to simulate the demand. In addition, we want to optimize the problem with respect to vector of prices which has $\binom{K}{2}$ elements, where K is the number of stations. Even for small K the problem becomes high-dimensional and one needs to adjust the number of simulated data accordingly to be able to make any statistical inference on so many independent variables.

The thesis is divided into two parts. Theory contains general results that are both own and taken from other literature. In Practical Applications we apply the results from the first part to the example with the train ticket prices.

In first chapter we introduce the theory of inhomogeneous Markov processes. Ordering of this chapter follows textbook of stochastic processes Prášková and Lachout [2012], where all the theory is derived for the homogeneous Markov process only. Most of the proofs of theorems in this chapter are also adapted from this textbook. Some theorems from the textbook are left out mostly because they are not relevant to the topic or does not hold for the inhomogeneous process. On the other hand, some definitions and theorems that was not included in the textbook had to be added to this thesis to complete the theory of inhomogeneous Markov processes. An example of such theory might be Chapter Separable Inhomogeneity whose content makes no sense in homogeneous case.

In Chapter Estimating Transition Intensity we introduce several models for the intensity of Markov process. The models are ordered by their complexity from the homogeneous case to general inhomogeneous case and regression model in which the intensity depends on exogenous variables. Individual models are illustrated on simple examples for more clarity. The intensity is estimated using maximum likelihood estimates. The theory of maximum likelihood is not introduce in this thesis. Instead, a literature where the theory is clearly summarized is recommended to the reader. All models described in this chapter are author's own contribution.

We describe two methods for simulation optimization in the third chapter. The first method, a response surface method, is partly based on own study and partly based on the results from Kroese et al. [2011]. The second method, cross entropy, is taken from that book. We also introduce a method for multistage stochastic optimization that is taken from Pflug and Pichler [2014].

In the second part of this thesis we try to implement the theory into a statistics and optimization problem - dynamic fare optimization. The implementation is done in R language and the methods are packed into a library[1]. The results show that Markov process is not efficient way to solve such problem.

[1] Available at https://github.com/jankislinger/dynFare.

Part I

Theory

Chapter 1

Markov Process

Prior to any theory of Markov process, we need to at least introduce some general terms of stochastic processes that are necessary to undestand the context in following chapters and in the practical applications.

In following sections we consider probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

1.1 Stochastic Process

There are several books about stochastic processes and most of the theory can be found in books about probability in general. In this section, we mention some of the sources while referring to the proofs of propositions that does not have to be proven in this thesis.

Let us begin with the definition of a stochastic process.

Definition 1.1. *Let* $\boldsymbol{X} = \{X_t, t \geq 0\}$ *be a family of random variables defined on probability space* $(\Omega, \mathcal{A}, \mathbb{P})$. *Then we call* $\boldsymbol{X}$ *a* stochastic process.

Kolmogorov consistency theorem states about necessary condition for existence of stochastic process.

Theorem 1.1 (Daniell-Kolmogorov)**.** *Let* $\{\mathbb{P}_{t_1,\ldots,t_n} : 0 \leq t_1 \leq \ldots \leq t_n < \infty, n \in \mathbb{N}\}$ *be a system of finite-dimensional distributions for which holds*

$$\begin{aligned}\mathbb{P}_{t_1,\ldots t_{k-1},t_{k+1},\ldots,t_n}(B_1 \times \ldots \times B_{k-1} \times B_{k+1} \times \ldots \times B_n) \\ = \mathbb{P}_{t_1,\ldots,t_n}(B_1 \times \ldots \times B_{k-1} \times \mathbb{R} \times B_{k+1} \times \ldots \times B_n)\end{aligned}$$

for all $n \in \mathbb{N}$, $1 \geq k \geq n$, $0 \leq t_1 \leq \ldots \leq t_n < \infty$ *a* $B_i \in \mathcal{B}$. *Then there exists a stochastic process, whose finite-dimensional distributions are given by this system.*

Proof. For example in Štěpán [1987], Theorem I.10.3. □

Following definitions will become useful when studying Markov processes. We state them just to introduce the terminology. We will not discuss their properties and relations too much in detail. The reader can find details in any introductory book of stochastic processes.

Definition 1.2. *Let* $\boldsymbol{X} = \{X_t, t \geq 0\}$ *and* $\boldsymbol{Y} = \{Y_t, t \geq 0\}$ *are stochastic processes defined on the same probability space* $(\Omega, \mathcal{A}, \mathbb{P})$. *If* $\mathbb{P}[X_t = Y_t] = 1$ *for all* $t \geq 0$ *then we say that* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are* stochastically equivalent.

Definition 1.3. *A stochastic process* $\boldsymbol{X} = \{X_t,\, t \geq 0\}$ *is called* stochastically continuous at point $t \geq 0$ *if for any* $\epsilon > 0$

$$\lim_{s \to t} \mathbb{P}(|X_s - X_t| > \epsilon) = 0.$$

If $\boldsymbol{X}$ *is stochastically continuous at every point* $t \geq 0$ *then we say that it is* stochastically continuous.

Definition 1.4. *Let* $\boldsymbol{X} = \{X_t,\, t \geq 0\}$ *be a stochastic process defined on probability space* $(\Omega,\, \mathcal{A},\, \mathbb{P})$, $D \subset [0, \infty)$ *be a countable dense set and* $\Lambda \subset \Omega$ *be a zero-probability event. If for any close set* $C \subset \mathbb{R}$ *and any open interval* $J \subset [0, \infty)$ *holds*

$$\{\omega : X_t(\omega) \in C, t \in J \cap D\} \backslash \{\omega : X_t(\omega) \in C, t \in J\} \subset \Lambda$$

then we say that $\boldsymbol{X}$ *is* separable.

Definition 1.5. *We will call* $\boldsymbol{X} = \{X_t,\, t \geq 0\}$ *a* measurable stochastic process *if mapping* $(\omega, t) \to X_t(\omega)$ *is measurable with respect to product* σ*-algebra* $\mathcal{A} \otimes \mathcal{B}([0, \infty))$.

We state two theorems providing relations between the two of the defined terms that are necessary for this thesis. Both the propositions are taken from Doob [1990] and are left without any proof. The reader can find one in said literature.

Proposition 1.2. *Let* $\boldsymbol{X} = \{X_t,\, t \geq 0\}$ *be a stochastically continuous process. Then there exists separable and measurable stochastic process that is equivalent to process* $\boldsymbol{X}$.

By previous proposition we can consider only process that is separable and measurable when we are given distribution of a stochastically continuous process.

Lemma 1.3. *Let* $\boldsymbol{X} = \{X_t,\, t \geq 0\}$ *be a stochastically continuous and separable process and* $\{D_n = t_0^{(n)} < \ldots < t_n^{(n), n \in \mathbb{N}}\}$ *be a sequence of subdivisions of an interval* $[s, s + h]$ *for which* $\|D_n\| \to 0$ *as* $n \to \infty$. *Then*

$$\lim_{n \to \infty} \mathbb{P}(X_{t_0^{(n)}} = i, \ldots, X_{t_n^{(n)}} = i) = \mathbb{P}(X_t = i, s \leq t \leq s + h)$$

Finally, let us introduce the idea of time of random event. This concept is a basis for Markov process with continuous time and its transformation into discrete time.

Definition 1.6. *Let* $\boldsymbol{X} = \{X_t,\, t \geq 0\}$ *be a stochastic process and* $\tau : \Omega \to [0, \infty]$ *be a measurable function. If* $[\tau \leq t] \in \mathcal{F}_t \equiv \sigma(X_s, s \leq t)$ *then we call* τ *a* Markov time *of process* $\boldsymbol{X}$.

Family of σ*-algebras* $\mathcal{F} = \{\mathcal{F}_t, t \geq 0\}$ *is called a* natural filtration *of process* $\boldsymbol{X}$.

Note that definition of Markov time can be generalized to any *filtration*. However, it is not necessary for the purpose of this thesis. Markov time is a random variable of time of certain event about which we can say whether it occurred until time t based on information about process $\boldsymbol{X}$ up to time t. Note that complementary set $[\tau \leq t]^C = [\tau > t]$ is also included in the σ-algebra $\mathcal{F}_t$. One of the simplest example of Markov time is the first entry into a set or the first exit from a set.

1.2 Markov Process

Definition 1.7. *Let* $\boldsymbol{X} = \{X_t, t \geq 0\}$ *be a stochastic process with values in discrete set* $\mathcal{S}$ *for which*

$$\mathbb{P}[X_t = j | X_{s_1} = i_1, \ldots, X_{s_n} = i_n, X_s = i] = \mathbb{P}[X_t = j | X_s = i] \tag{1.1}$$

for every $i, i_1, \ldots, i_n, j \in \mathcal{S}$, $n \in \mathbb{N}$ *and* $0 \leq s_1 < \ldots < s_n < s < t$ *such that* $\mathbb{P}[X_{s_1} = i_1, \ldots, X_{s_n} = i_n, X_s = i] > 0$. *Then we call* $\boldsymbol{X}$ *a* Markov process *with (discrete) state space* $\mathcal{S}$.

Without loss of generality we will consider only state space $\mathcal{S} = \{0, 1, \ldots, N\}$ or $\mathcal{S} = \mathbb{N}_0$.

Let us denote probabilities from (1.1) by $p_{ij}(s,t)$ and call these *transition probabilities*. Similarly, we denote *absolute probabilities* by $p_j(t) = \mathbb{P}[X_t = j]$ and *initial probabilities* by $p_j = p_j(0) = \mathbb{P}[X_0 = j]$. Further, we denote *transition matrix* by $\boldsymbol{P}(s,t) = (p_{ij}(s,t))_{i,j\in\mathcal{S}}$, *absolute probability vector* $\boldsymbol{p}(t) = (p_j(t))_{j\in\mathcal{S}}$ and *initial probability vector* by $\boldsymbol{p} = \boldsymbol{p}(0)$.

It is clear that for any $t \geq s \geq 0$ vector $\boldsymbol{p}(t)$ is a *stochastic vector* and matrix $\boldsymbol{P}(s,t)$ is a *stochastic matrix*, i.e.

$$p_j(t) \geq 0, \qquad j \in \mathcal{S}; \qquad \sum_{j\in\mathcal{S}} p_j(t) = 1, \tag{1.2}$$

and

$$p_{ij}(s,t) \geq 0, \qquad i,j \in \mathcal{S}; \qquad \sum_{j\in\mathcal{S}} p_{ij}(s,t) = 1, \qquad i \in \mathcal{S}. \tag{1.3}$$

If transition probabilities are independent of the beginning time and depend only on time difference, i.e. $\boldsymbol{P}(s, s+t)$ does not depend on s for any $t \geq 0$, we call the Markov process *homogeneous* and we can denote $\boldsymbol{P}(s, s+t) = \boldsymbol{P}(t)$. Otherwise, we talk about *inhomogeneous* Markov process. We consider mostly inhomoheneous process in this thesis.

Proposition 1.4 (Chapman-Kolmogorov equality)**.** *Let* $\boldsymbol{X} = \{X_t, t \geq 0\}$ *be a Markov process with system of transition matrices* $\{\boldsymbol{P}(s,t), t \geq s \geq 0\}$. *Then for any* $t \geq r \geq s \geq 0$ *and* $i, j \in \mathcal{S}$ *holds*

$$p_{ij}(s,t) = \sum_{k\in\mathcal{S}} p_{ik}(s,r) p_{kj}(r,t). \tag{1.4}$$

Proof. Let us define $\mathcal{S}' = \{k \in \mathcal{S} : \mathbb{P}[X_s = i, X_r = k] > 0\}$. Using Markov

property and Bayes formula we can write

$$\begin{aligned}
p_{ij}(s,t) = \mathbb{P}[X_t = j | X_s = i] &= \sum_{k\in\mathcal{S}} \mathbb{P}[X_t = j, X_r = k | X_s = i] \\
&= \sum_{k\in\mathcal{S}'} \mathbb{P}[X_t = j, X_r = k | X_s = i] \\
&= \sum_{k\in\mathcal{S}'} \mathbb{P}[X_t = j | X_s = i, X_r = k]\, \mathbb{P}[X_r = k | X_s = i] \\
&= \sum_{k\in\mathcal{S}'} \mathbb{P}[X_t = j | X_r = k]\, \mathbb{P}[X_r = k | X_s = i] \\
&= \sum_{k\in\mathcal{S}} \mathbb{P}[X_t = j | X_r = k]\, \mathbb{P}[X_r = k | X_s = i] \\
&= \sum_{k\in\mathcal{S}} p_{ik}(s,r) p_{kj}(r,t).
\end{aligned}$$

□

Using matrix notation the Chapman-Kolmogorov equality can be rewritten as

$$\boldsymbol{P}(s,t) = \boldsymbol{P}(s,r)\boldsymbol{P}(r,t)$$

for any $t \geq r \geq s \geq 0$.

The converse implication of previous theorem which ensures the existence of Markov process for given vector of initial probabilities and system of transition matrices also holds.

Proposition 1.5. *Let $\mathcal{P} = \{\boldsymbol{P}(s,\, t),\, t \geq s \geq 0\}$ be system of stochastic matrices fulfilling (1.4) and $\boldsymbol{p}$ be a stochastic vector. Then there exists Markov process with initial probability $\boldsymbol{p}$ and system of transition matrices $\mathcal{P}$.*

Proof. We will prove required by using Kolmogorov consistency theorem. To fulfill assumptions of that theorem we need to find consistent system of finite-dimensional distributions generated by $\mathcal{P}$. Since we are looking for a process with countable space of states it suffices to show that holds

$$\begin{aligned}
&\mathbb{P}_{(X_{t_1},\ldots,X_{t_{k-1}},X_{t_{k+1}},\ldots,X_{t_1})}(\{i_1\} \times \ldots \times \{i_{k-1}\} \times \{i_{k+1}\} \times \ldots \times \{i_n\}) \\
&\qquad = \mathbb{P}_{(X_{t_1},\ldots,X_{t_1})}(\{i_1\} \times \ldots \times \{i_{k-1}\} \times \mathcal{S} \times \{i_{k+1}\} \times \ldots \times \{i_n\})
\end{aligned}$$

for any $0 \leq t_1 < \ldots < t_n$ and $i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_n \in \mathcal{S}$.

Let us denote left hand side and right hand side of the previous equation by Ξ_L and Ξ_R respectively. Then we can write

$$\begin{aligned}
\Xi_L &= \mathbb{P}[X_{t_1} = i_1, \ldots, X_{t_{k-1}} = i_{k-1}, X_{t_{k+1}} = i_{k+1}, \ldots, X_{t_n} = i_n] \\
&= \mathbb{P}[X_{t_1} = i_1]\, \mathbb{P}[X_{t_2} = i_2 | X_{t_1} = i_1] \ldots \mathbb{P}[X_{t_{k-1}} = i_{k-1} | X_{t_1} = i_1, \ldots, X_{t_{k-2}} = i_{k-2}] \\
&\qquad \ldots \mathbb{P}[X_{t_{k-1}} = i_{k-1} | X_{t_1} = i_1, \ldots, X_{t_{k-2}} = i_{k-2}] \\
&\qquad \ldots \mathbb{P}[X_{t_{k+1}} = i_{k+1} | X_{t_1} = i_1, \ldots, X_{t_{k-1}} = i_{k-1}] \\
&\qquad \ldots \mathbb{P}[X_{t_n} = i_n | X_{t_1} = i_1, \ldots, X_{t_{k-1}} = i_{k-1}, X_{t_{k+1}} = i_{k+1}, \ldots, X_{t_{n-1}} = i_{n-1}] \\
&= \mathbb{P}[X_{t_1} = i_1] p_{i_1 i_2}(t_1, t_2) \ldots p_{i_{k-2} i_{k-1}}(t_{k-2}, t_{k-1}) p_{i_{k-1} i_{k+1}}(t_{k-1}, t_{k+1}) \\
&\qquad \ldots p_{i_{n-1} i_n}(t_{n-1}, t_n)
\end{aligned}$$

Assuming that (1.4) holds we have

$$p_{i_{k-1}i_{k+1}}(t_{k-1}, t_{k+1}) = \sum_{i_k \in \mathcal{S}} p_{i_{k-1}i_k}(t_{k-1}, t_k) p_{i_k i_{k+1}}(t_k, t_{k+1}).$$

Hence

$$\Xi_L = \sum_{k \in \mathcal{S}} \mathbb{P}[X_{t_1} = i_1] \prod_{\nu=2}^{n} p_{i_{\nu-1}i_\nu}(t_{\nu-1}, t_\nu).$$

The right hand side of the equation might be written as

$$\begin{aligned}
\Xi_R &= \sum_{k \in \mathcal{S}} \mathbb{P}[X_{t_1} = i_1, \ldots, X_{t_n} = i_n] \\
&= \sum_{k \in \mathcal{S}} \mathbb{P}[X_{t_1} = i_1] \, \mathbb{P}[X_{t_2} = i_2 | X_{t_1} = i_1] \ldots \mathbb{P}[X_{t_n} = i_n | X_{t_1} = i_1, \ldots, X_{t_{n-1}} = i_{n-1}] \\
&= \sum_{k \in \mathcal{S}} \mathbb{P}[X_{t_1} = i_1] \prod_{\nu=2}^{n} p_{i_{\nu-1}i_\nu}(t_{\nu-1}, t_\nu).
\end{aligned}$$

We met required by obtaining $\Xi_L = \Xi_R$ and the proof is done. □

Following proposition describes how to construct finite-dimensional distributions of Markov process.

Proposition 1.6. *Let $\boldsymbol{X} = \{X_t,\, t \geq 0\}$ be a stochastic process with state space $\mathcal{S}$. Let $\boldsymbol{p}$ be a stochastic vector fulfilling (1.2) and $\mathcal{P} = \{\boldsymbol{P}(s,t), t \geq s \geq 0\}$ be a system of stochastic matrices fulfilling (1.3) and (1.4) for all $t \geq r \geq s \geq 0$. Then X is a Markov process with initial probability $\boldsymbol{p}$ and system of transition matrices $\mathcal{P}$ if and only if*

$$\mathbb{P}[X_0 = i_0, X_{t_1} = i_1, \ldots, X_{t_n} = i_n] = p_{i_0} p_{i_0 i_1}(0, t_1) \ldots p_{i_{n-1} i_n}(t_{n-1}, t_n) \tag{1.5}$$

for any $n \in \mathbb{N}_0$, $i_0, \ldots, i_n \in \mathcal{S}$ and $0 < t_1 < \ldots t_n$.

Proof. Let X be a Markov process with initial probability $\boldsymbol{p}$ and system of transition matrices $\mathcal{P}$. If $\mathbb{P}[X_0 = i_0] = 0$ then both sides of (1.5) equal trivially.

If there exists $0 < k \leq n$ such that both following statements hold

$$\begin{aligned}
\mathbb{P}[X_0 = i_0, X_{t_1} = i_1, \ldots, X_{t_k} = i_k] &= 0 \\
\mathbb{P}[X_0 = i_0, X_{t_1} = i_1, \ldots, X_{t_{k-1}} = i_{k-1}] &> 0,
\end{aligned}$$

then, by applying Bayes formula, we obtain

$$p_{i_{k-1}i_k}(t_{k-1}, t_k) = \frac{\mathbb{P}[X_0 = i_0, X_{t_1} = i_1, \ldots, X_{t_k} = i_k]}{\mathbb{P}[X_0 = i_0, X_{t_1} = i_1, \ldots, X_{t_{k-1}} = i_{k-1}]} = 0$$

and equation (1.5) holds trivially as well.

If such k does not exist then the left hand side of (1.5) is greater than 0 and all the following conditioned probabilities are well defined. With $t_0 = 0$ we can write

$$\begin{aligned}
\mathbb{P}[X_0 = i_0, X_{t_1} = i_1, \ldots, X_{t_n} = i_n] &= \prod_{m=0}^{n} \mathbb{P}[X_{t_m} = i_m | X_{t_\nu} = i_\nu, \nu < m] \\
&= \prod_{m=0}^{n} \mathbb{P}[X_{t_m} = i_m | X_{t_{m-1}} = i_{m-1}] = p_{i_0} p_{i_0 i_1}(0, t_1) \ldots p_{i_{n-1} i_n}(t_{n-1}, t_n)
\end{aligned}$$

and equation (1.5) holds.

Let equation (1.5) holds. With choice $n = 0$ we obtain $\mathbb{P}[X_0 = i_0] = p_{i_0}$ for any $i_0 \in \mathcal{S}$. Hence the initial probability is equal to $\boldsymbol{p}$.

Take $0 \leq s < t$ and suppose that $\mathbb{P}[X_s = i] > 0$. Then we can write

$$\begin{aligned}
\mathbb{P}[X_t = j | X_s = i] &= \frac{\mathbb{P}[X_s = i, X_t = j]}{\mathbb{P}[X_s = i]} \\
&= \frac{\sum_{k \in \mathcal{S}} \mathbb{P}[X_0 = k, X_s = i, X_t = j]}{\sum_{k \in \mathcal{S}} \mathbb{P}[X_0 = k, X_s = i]} \\
&= \frac{\sum_{k \in \mathcal{S}} p_k p_{ki}(0, s) p_{ij}(s, t)}{\sum_{k \in \mathcal{S}} p_k p_{ij}(0, s)} \\
&= p_{ij}(s, t)
\end{aligned}$$

Hence $\{\boldsymbol{P}(s,t), t \geq s \geq 0\}$ is system of transition matrices of process $\boldsymbol{X}$.

To show the Markov property of the process $\boldsymbol{X}$, we take any $i_1, \ldots, i_n \in \mathcal{S}$ and $0 \leq s_1 < \ldots < s_n < s < t$ for which $\mathbb{P}[X_{s_1} = i_1, \ldots, X_{s_{n-1}} = i_{n-1}, X_s = i] > 0$. Then we can write

$$\begin{aligned}
&\mathbb{P}[X_t = j | X_{s_1} = i_1, \ldots, X_{s_{n-1}} = i_{n-1}, X_s = i] \\
&\qquad = \frac{\mathbb{P}[X_{s_1} = i_1, \ldots, X_{s_n} = i_n, X_s = i, X_t = j]}{\mathbb{P}[X_{s_1} = i_1, \ldots, X_{s_{n-1}} = i_{n-1}, X_s = i]} \\
&\qquad = \frac{\sum_{k \in \mathcal{S}} \mathbb{P}[X_0 = k, X_{s_1} = i_1, \ldots, X_{s_n} = i_n, X_s = i]}{\sum_{k \in \mathcal{S}} \mathbb{P}[X_0 = k, X_{s_1} = i_1, \ldots, X_{s_{n-1}} = i_{n-1}, X_s = i, X_t = j]} \\
&\qquad = \frac{\sum_{k \in \mathcal{S}} p_k p_{ki_1}(0, s_1) \ldots p_{i_{n-1} i_n}(s_{n-1}, s_n) p_{i_n i}(s_n, s) p_{ij}(s, t)}{\sum_{k \in \mathcal{S}} p_k p_{ki_1}(0, s_1) \ldots p_{i_{n-2} i_{n-2}}(s_{n-2}, s_{n-2}) p_{i_n i}(s_n, s)} \\
&\qquad = p_{ij}(s, t) = \mathbb{P}[X_t = j | X_s = i].
\end{aligned}$$

Hence we have proven the Markov property of process $\boldsymbol{X}$. □

To be able to state and prove deeper theorems we need to make several assumptions about the Markov process.

(A1) Transition probabilities are right continuous at the axis of the first quadrant, i.e.

$$\lim_{t \to s^+} p_{ij}(s, t) = \delta_{ij}$$

for all $i, j \in \mathcal{S}$ and $s \geq 0$.

(A2) Transition probabilities are differentiable, i.e. for all $i \in \mathcal{S}$ and $t \geq 0$ there exists a limit

$$q_i(t) = \lim_{h \to 0^+} \frac{1 - p_{ii}(t, t+h)}{h} \leq \infty$$

and for all $i, j \in \mathcal{S}$, $i \neq j$ and $t \geq 0$ there exists a limit

$$q_{ij}(t) = \lim_{h \to 0^+} \frac{p_{ij}(t, t+h)}{h} < \infty.$$

(A3) The order of limit and summation is interchangeable, i.e.

$$q_i(t) = \sum_{j \in \mathcal{S}} q_{ij}(t)$$

for all $i \in \mathcal{S}$ and $t \geq 0$.

(A4) For all $j \in \mathcal{S}$ and $t \geq 0$ the limit $q_{ij}(t)$ converges uniformly with respect to $i \in \mathcal{S}$.

Obviously, there exist relations between individual assumptions. For example, Chung [1967] proved that if the Markov process is homogeneous then (A1) implies (A2). Also, if $\mathcal{S}$ is finite then both (A3) and (A4) hold.

It is sometimes convenient to use equation (that is based on (A2))

$$p_{ij}(t, t+h) = q_{ij}(t)h + o(h)$$

for any $i, j \in \mathcal{S}$, $t \geq 0$, and $h > 0$.

Definition 1.8. *Function $q_{ij}(t)$ defined at* (A2) *is called* transition rate *from state i to state j at time t. Function $q_i(t)$ defined at (A2) is called* total transition rate *at time t. Matrix $\boldsymbol{Q}(t) = (q_{ij}(t))_{i,j \in \mathcal{S}}$, where $q_{ii}(t) = -q_i(t)$ is called* transition rate matrix *at time t.*

Proposition 1.7 (Kolmogorov Differential Equations)**.** *Suppose that (A1)-(A3) hold. Then*

$$\frac{\partial \boldsymbol{P}(s,t)}{\partial s} = -\boldsymbol{Q}(s)\boldsymbol{P}(s,t). \tag{1.6}$$

If also (A4) holds then

$$\frac{\partial \boldsymbol{P}(s,t)}{\partial t} = \boldsymbol{P}(s,t)\boldsymbol{Q}(t). \tag{1.7}$$

Proof. For example in Mandl [1985], page 36, Proposition 1 and Proposition 2. □

Proposition 1.8. *Let $\boldsymbol{X} = \{X_t,\, t \geq 0\}$ be a Markov process, for which (A1) holds. Then $\boldsymbol{X}$ is stochastically continuous.*

Proof. We know that

$$\mathbb{P}[|X_t - X_{t+h}| > \epsilon] \leq \mathbb{P}[X_t \neq X_{t+h}] = 1 - \mathbb{P}[X_t = X_{t+h}].$$

Then, for right-sided limit, it is sufficient to show that

$$\lim_{h \to 0^+} (X_t = X_{t+h}) = 1.$$

Holds

$$\mathbb{P}[X_t = X_{t+h}] = \sum_{j \in \mathcal{S}} \mathbb{P}[X_t = j, X_{t+h} = j] = \sum_{j \in \mathcal{S}} p_j(t) p_{jj}(t, t+h).$$

Because

$$|p_j(t)p_{jj}(t,t+h)| \leq p_j(t) \qquad \& \qquad \sum_{j \in \mathcal{S}} p_j(t) = 1,$$

the summands are dominated by summable sequence and the order of the limit and summation can be changed

$$\lim_{h\to 0^+}\sum_{j\in\mathcal{S}} p_j(t)p_{jj}(t,t+h) = \sum_{j\in\mathcal{S}} p_j(t)\lim_{h\to 0^+} p_{jj}(t,t+h) = \sum_{j\in\mathcal{S}} p_j(t) = 1.$$

We can show similar for the left-sided limit. With both one-sided limits together we obtain $\lim_{h\to 0}\mathbb{P}(|X_t - X_{t+h}| > \epsilon) = 0$. □

By Proposition 1.2 we know that for there exists a version of any Markov process that is separable and measurable. In the following, we will consider only processes that are separable and measurable.

Proposition 1.9. *Let $\boldsymbol{X} = \{X_t,\, t \geq 0\}$ be a Markov process with total intensities $q_i(t) < \infty$ on interval $[s, s+h]$. Then for any $s \geq 0$ and $h > 0$ holds*

$$\mathbb{P}\left[X_t = i, s \leq t \leq s+h | X_s = i\right] = \exp\left(-\int_s^{s+h} q_i(t)dt\right). \tag{1.8}$$

Proof. Because Markov process is stochastically continuous and separable, by Proposition 1.3 with choice $D_n = \{s + \frac{k}{n}h, k = 0, \ldots, n\}$ holds

$$\mathbb{P}\left[X_t = i, s \leq t \leq s+h | X_s = i\right] = \lim_{n\to\infty}\mathbb{P}\left[X_{s+\frac{k}{n}h} = i, k = 0, \ldots, n\right]$$

for all $i \in \mathcal{S}$ and $s \geq 0$. That can be rewritten, with use of the Markov property, as

$$\begin{aligned}\lim_{n\to\infty}\prod_{k=1}^{n}\mathbb{P}\left[X_{s+\frac{k}{n}h} = i | X_{s+\frac{k-1}{n}h} = i\right] &= \lim_{n\to\infty}\prod_{k=1}^{n} p_{ii}\left(s+\frac{k}{n}h, s+\frac{k-1}{n}h\right)\\ &= \lim_{n\to\infty}\prod_{k=1}^{n}\left(1 - q_i\left(s+\frac{k}{n}h\right) + o\left(\frac{h}{n}\right)\right) = \exp\left(-\int_s^{s+h} q_i(t)dt\right).\end{aligned}$$

□

If the Markov process is homogeneous, the transition rate $q_i = q_i(t)$ is constant in time and the probability 1.2 is equal to e^{-hq_i}.

Proposition 1.10. *Let Markov process $\boldsymbol{X} = \{X_t,\, t \geq 0\}$ enters the state i at time s. Then the dwell time is a random variable with cumulative distribution function*

$$F_{s,i}(x) = 1 - \exp\left(-\int_s^{s+x} q_i(t)dt\right)$$

and probability density function

$$f_{s,i}(x) = \exp\left(-\int_s^{s+x} q_i(t)dt\right) q_i(s+x).$$

Proof. Define

$$\tau_{s,i} = \min\left\{t > s : X_t \neq i\right\} \tag{1.9}$$

with additional convention $\min\{\} = \infty$. Because we assume only right continuous Markov process minimum always exists (either finite or infinite). Then the cumulative distribution function is

$$\begin{aligned}\mathbb{P}\left[\tau_{s,i} \leq s + x | X_s = i\right] &= 1 - \mathbb{P}\left[\tau_{s,i} > s + x | X_s = i\right] \\ &= 1 - \mathbb{P}\left[X_t = i, s \leq t \leq s + x | X_s = i\right] = 1 - \exp\left(-\int_s^{s+x} q_i(t)dt\right).\end{aligned}$$

Probability density function is then the derivative of cumulative distribution function with respect to x. □

If $q_i = 0$ the dwell time is almost surely ∞. If $\int_s^{s+x} q_i(t)dt = \infty$ for all $x > 0$, then the distribution function has single jump of size 1 at point 0, i.e. dwell time is almost surely 0. If $\int_s^{\infty} q_i(t)dt < \infty$ then $\mathbb{P}[\tau_{s,i} = \infty] > 0$. Relations between transition rate function and cumulative distribution function of dwell time are illustrated in Figure 1.1.

Example 1.1 (Lomax distributed dwell time)**.** Suppose that a Markov process has space of states $\mathcal{S} = \{0, 1\}$ and transition rate matrix

$$\boldsymbol{Q}(t) = \begin{pmatrix} -\dfrac{\lambda_0}{t+\varepsilon} & \dfrac{\lambda_0}{t+\varepsilon} \\ \dfrac{\lambda_1}{t+\varepsilon} & -\dfrac{\lambda_1}{t+\varepsilon} \end{pmatrix},$$

where $\lambda_0, \lambda_1, \varepsilon > 0$ are given constants. Suppose that process has transitioned into state $i \in \{0, 1\}$ at time $s \geq 0$. Thus the dwell time has cumulative distribution function

$$\begin{aligned}1 - \exp\left(-\int_s^{s+x} \frac{\lambda_i}{t+\varepsilon} dt\right) &= 1 - \exp\left(\lambda_i \log(s+\varepsilon) - \lambda_i \log(s+x+\varepsilon)\right) \\ &= 1 - \exp\left(\lambda_i \log\left(\frac{s+\varepsilon}{s+x+\varepsilon}\right)\right) = 1 - \left(1 + \frac{x}{s+\varepsilon}\right)^{-\lambda_i}.\end{aligned}$$

That is the Lomax distribution with scale $s + \varepsilon$ and shape λ_i. △

Proposition 1.11. *Suppose that Markov process $\boldsymbol{X}$ is at state i up to time t and at time t leaves. Then it transitions into state j with probability*

$$\mathbb{P}[X_t = j | \lim_{s \to t^+} X_s = i, X_t \neq i] = \frac{q_{ij}(t)}{q_i(t)}.$$

Proof. Let $\{s_k, k \in \mathbb{N}_0\}$ be any increasing sequence with $s_k \to t$ as $k \to \infty$. Clearly the elements

$$A_k = [X_s = i, s \in [s_0, s_k], X_t \neq i]$$

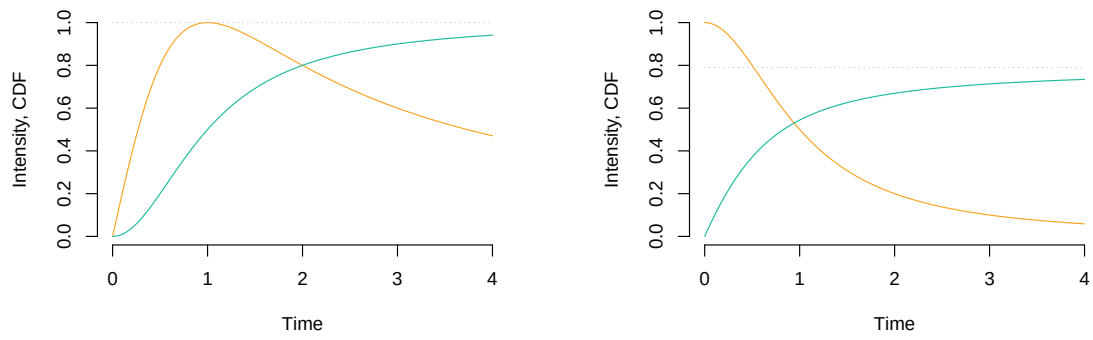
are monotone as well with $A_1 \supset A_2 \supset ....$ Then we can write

$$\mathbb{P}[X_t = j | \lim_{s \to t^+} X_s = i, X_t \neq i] = \lim_{k \to \infty} \mathbb{P}[X_t = j | A_k].$$

Because $\boldsymbol{X}$ is Markov process it only depends on the state at the last known time and we have

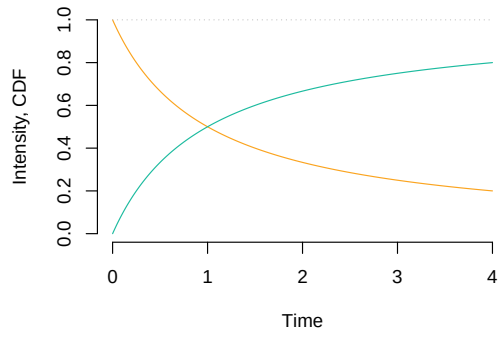
$$\mathbb{P}[X_t = j | A_k] = \mathbb{P}[X_t = j | X_{s_k} = i, X_t \neq i] = \frac{p_{ij}(s_k, t)}{1 - p_{ii}(s_k, t)}.$$

After applying the limits we arrive to required. □

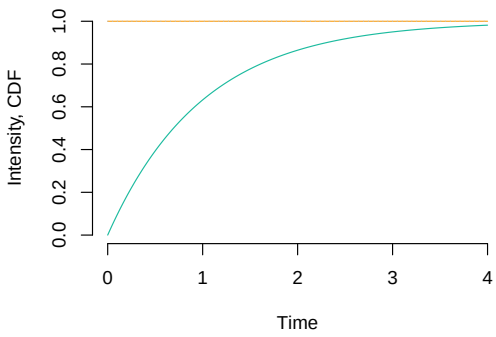


(a) $q(t) = \frac{2t}{t^2+1}$

(b) $q(t) = \frac{1}{t^2+1}$



(c) $q(t) = \frac{1}{t+1}$



(d) $q(t) = 1$

Figure 1.1: Comparison of several transition rates (orange lines), their respective cumulative distribution functions of dwell time (blue lines), and limits of CDF (gray dashed lines. In the last plot we can see CDF of exponential distribution Exp(1).

Proposition 1.12. *Let random variable $\tau_{s,i}$ be the first time process leaves state i after time $s \geq 0$ and $X_s = i$, i.e. the random variable defined in* (1.9)*. The process first transitions into state j with probability*

$$\mathbb{P}[X_{\tau_{s,i}} = j | X_s = i] = \mathbb{E}\left[\frac{q_{ij}(\tau_{s,i})}{q_i(\tau_{s,i})}\right]. \tag{1.10}$$

Proof. Denote F_τ the cumulative distribution function of random variable $\tau_{s,i}$ and recall the support of $\tau_{s,i}$ is interval (s, ∞). Then, by applying continuous Bayes' formula, we get

$$\begin{aligned}
&\mathbb{P}[X_{\tau_{s,i}} = j | X_s = i] = \int_s^\infty \mathbb{P}[X_{\tau_{s,i}} = j | X_s = i, \tau_{s,i} = t] dF_\tau(t) \\
&= \int_s^\infty \mathbb{P}[X_t = j | X_t \neq j, X_r = i, s \leq r < t] dF_\tau(t) = \int_s^\infty \frac{q_{ij}(t)}{q_i(t)} dF_\tau(t) = \mathbb{E}\left[\frac{q_{ij}(\tau_{s,i})}{q_i(\tau_{s,i})}\right].
\end{aligned}$$

□

Note that the transition probability (1.10) from previous proposition can be rewritten using the density of dwell time as

$$\mathbb{E}\left[\frac{q_{ij}(\tau_{s,i})}{q_i(\tau_{s,i})}\right] = \int_0^\infty \exp\left(-\int_s^{s+x} q_i(t)dt\right) q_{ij}(s+x)dx.$$

Definition 1.9. *Let $\boldsymbol{X} = \{X_t,\, t \geq 0\}$ be a Markov process with transition times $T_1, T_2, \ldots$. Then we call a sequence of random variables*

$$\{Y_k = X_{T_k}, k \in \mathbb{N}_0\}$$

an embedded chain.

Prášková and Lachout [2012] proved that embedded chain of homogeneous Markov process is homogeneous Markov chain. It is clear that embedded chain of inhomogeneous Markov process is again inhomogeneous. We can further show that the chain does not have Markov property. Take, for example, intensity matrix

$$\boldsymbol{Q}(t) = \begin{pmatrix} -\lambda & 0 & \lambda & 0 & 0 \\ 0 & -1/\lambda & 1/\lambda & 0 & 0 \\ 0 & 0 & -\lambda & \lambda\mathbb{I}[t \leq 1] & \lambda\mathbb{I}[t > 1] \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

and initial probability vector $\boldsymbol{p} = (1/2, 1/2, 0, 0, 0)$ with state space $\mathcal{S} = \{1, ..., 5\}$. Then it holds for the embedded chain $\{Y_t, t \in \mathbb{N}_0\}$

$$\begin{aligned}
\mathbb{P}[Y_3 = 4 | Y_2 = 3, Y_1 = 1] &\to 1 \qquad \text{as} \qquad \lambda \to \infty \\
\mathbb{P}[Y_3 = 4 | Y_2 = 3, Y_1 = 2] &\to 0 \qquad \text{as} \qquad \lambda \to \infty
\end{aligned}$$

Hence this chain does not have Markov property. However, we are still able to derive its distribution.

Corollary 1.13. *Let Markov process $\boldsymbol{X} = \{X_t,\, t \geq 0\}$ enter the state i at time s. Then the joint distribution of the dwell time and the new state has the density*

$$f_{s,i}(x, j) = \exp\left(-\int_s^{s+x} q_i(t)dt\right) q_{i,j}(s + x)$$

with respect to product of Lebesgue measure on $(0, \infty)$ and counting measure on $\mathcal{S} \setminus \{i\}$.

Proof. The joint density is a product of marginal density of dwell time (Proposition 1.10) and conditional transition probability (Proposition 1.11) □

1.3 Separable Inhomogeneity

In this section, we will study family of inhomogeneous Markov processes for which transition rates vary over time in the same manner for each transition between pairs $i, j \in \mathcal{S}$, $i \neq j$.

Definition 1.10. *Let $\boldsymbol{X} = \{X_t,\, t \geq 0\}$ be inhomogeneous Markov process with transition matrix $\boldsymbol{Q}(t)$. If there exists a function $\lambda : [0, \infty) \to (0, \infty)$ and a (constant-in-time) matrix $\boldsymbol{\Gamma}$ such that*

$$\boldsymbol{Q}(t) = \lambda(t)\boldsymbol{\Gamma}, \qquad \forall t \geq 0$$

then we say that $\boldsymbol{X}$ is Markov process with separable inhomogeneity.

We will call λ a separated inhomogeneity *and $\boldsymbol{\Gamma}$ a* constant rate matrix.

Definition 1.11. *Suppose that $\boldsymbol{X} = \{X_t,\, t \geq 0\}$ is an inhomogeneous Markov process and $\boldsymbol{Y} = \{Y_t,\, t \geq 0\}$ is a homogeneous Markov process. We say that $\boldsymbol{Y}$ is* homogeneous transformation *of $\boldsymbol{X}$ if there exists an increasing and differentiable function $\tau : [0, \infty) \to [0, \infty)$, $\tau(0) = 0$ such that $X_{\tau(t)} = Y_t$ a.s.*

Following theorem justifies the equivalency of previous definitions. The proof of the theorem provides a way how to construct the time-transformation function τ and separated intensity function λ.

Proposition 1.14. *Markov process $\boldsymbol{X} = \{X_t,\, t \geq 0\}$ has separable inhomogeneity if and only if there exists a homogeneous transformation of $\boldsymbol{X}$.*

Proof. Let $\boldsymbol{X}$ has separable inhomogeneity with notation from Definition 1.10. Define $\boldsymbol{Y}$ as $Y_t = X_{\tau(t)}$, where $\tau(t) = \Lambda^{-1}(t)$. Because the function τ is increasing, the Markov property holds for the process $\boldsymbol{Y}$ as well. The transition probabilities of $\boldsymbol{Y}$ are given by

$$\begin{aligned} p_{ij}^{(Y)}(t, t + h) =& \mathbb{P}[Y_{t+h} = j | Y_t = i] = \mathbb{P}[X_{\tau(t+h)} = j | X_{\tau(t)} = i] \\ =& \gamma_{ij}\lambda(\tau(t)) \left(\Lambda^{-1}(t + h) - \Lambda^{-1}(t)\right) + o(h). \end{aligned}$$

The derivative of inverse function is given by

$$\frac{\partial \Lambda^{-1}(t)}{\partial t} = \frac{1}{\lambda(\Lambda^{-1}(t))},$$

which implies that the intensity of $\boldsymbol{Y}$ is $q_{ij}^{(Y)}(t) = \gamma_{ij}$ and does not depend on time t.

Suppose there exists a process $\boldsymbol{Y}$ that is a homogeneous transformation of $\boldsymbol{X}$ with time-transformation function τ as in Definition 1.11 and the intensity matrix of $\boldsymbol{Y}$ is $\boldsymbol{\Gamma}$. The transition probabilities of $\boldsymbol{X}$ are given by

$$\begin{aligned} p_{ij}^{(X)}(t, t+h) =& \mathbb{P}[X_{t+h} = j | X_t = i] = \mathbb{P}[Y_{\tau^{-1}(t+h)} = j | Y_{\tau^{-1}(t)} = i] \\ =& \gamma_{ij} \left(\tau^{-1}(t+h) - \tau^{-1}(t) \right) + o(h). \end{aligned}$$

Therefore $\boldsymbol{X}$ has separated inhomogeneity with constant rate matrix $\boldsymbol{\Gamma}$ and function

$$\lambda(t) = \frac{\partial \tau^{-1}(t)}{\partial t}.$$

□

In previous chapter we showed that embedded chain of inhomogeneous Markov process does not have to be Markov chain. On the other hand, if the process has separable inhomogeneity its embedded chain is homogeneous Markov Chain. That is because such process has homogeneous transformation by Proposition 1.14 and both the process and its homogeneous transformation have almost surely the same embedded chain. This property is convenient, for example, when simulating such process. One can simulate first the embedded chain and then the dwell times.

1.4 Inhomogeneous Poisson Process

Poisson process is widely used as a counting process for number of events that occur independently over a time interval. In practice, we sometimes face problems that violate the assumptions of homogeneity of the process. In this section we introduce process that generalizes the definition of Poisson process. Let us begin with definition of inhomogeneous Poisson process.

Definition 1.12. *Suppose that $N = \{N_t, t \geq 0\}$ is a counting process with values in $\mathbb{N}_0$ and $\lambda : [0, \infty) \to [0, \infty)$ be a real function. We call N an* inhomogeneous Poisson process *if it follows*

i) $N_0 = 0$ *a.s.;*

ii) Increments of N are independent;

iii) $\mathbb{P}[N_{t+h} = i + 1 | N_t = i] = \lambda(t)h + o(h)$*;*

iv) $\mathbb{P}[N_{t+h} > i + 1 | N_t = i] = o(h)$.

Clearly, if λ is constant, non-zero function then N is homogeneous Poisson process. Inhomogeneous Poisson process is a special case of Markov process with separable inhomogeneity with transition rate matrix

$$\boldsymbol{Q}(t) = \begin{pmatrix} -\lambda(t) & \lambda(t) & 0 & \cdots \\ 0 & -\lambda(t) & \lambda(t) & \\ \vdots & & \ddots & \ddots \end{pmatrix} = \lambda(t) \begin{pmatrix} -1 & 1 & 0 & \cdots \\ 0 & -1 & 1 & \\ \vdots & & \ddots & \ddots \end{pmatrix}.$$

The distribution of Poisson process and its increments is again Poisson, i.e.

$$N_t \sim Poi(\Lambda(t)),$$
$$N_s - N_t \sim Poi(\Lambda(s) - \Lambda(t)),$$

where $\Lambda(t) = \int_0^t \lambda(t)dt$. The reader can find a proof for example in Lewis and Shedler [1979]. Another way to show this is to use the fact that the inhomogeneous Poisson process is also an inhomogeneous Markov process with separable inhomogeneity. Then there exists by a homogeneous transformation with unit intensity and transformation function $\tau(t) = \Lambda^{-1}(t)$. This can be used to simulate an inhomogeneous Poisson process on time interval $[0, T]$.

Algorithm 1.1 (Simulating Poisson process)**.**

1. *Simulate number of transitions* $N \sim Poi(\Lambda(T))$.
2. *Simulate independent random variables* $R_i \sim Exp(1)$ *for* $i = 1, ..., N + 1$.
3. *Calculate ordered transition times of the homogeneous transformation on interval* $[0, 1]$
$$S_k = \frac{\sum_{i=1}^{k} R_i}{\sum_{i=1}^{N+1} R_i}, \qquad k = 1, ..., N.$$
4. *Transform to transition times of the inhomogeneous Poisson process*
$$T_k = \Lambda^{-1}(S_k \times \Lambda(T)), \qquad k = 1, ..., N,$$
5. *Return the transition times* $T_1, ..., T_N$.

Denote by $T_1, T_2, ...$ the increasing sequence of (random) jump times of inhomogeneous Poisson process. Their distributions are given by Proposition 1.10. Let $Z_1, Z_2, ...$ be a sequence of independent random variables with distribution measures μ_{T_i}. Then we can define random process $\boldsymbol{M} = \{M - t, t \geq 0\}$ by

$$M_t = \sum_{i=1}^{\infty} Z_i \times \mathbb{I}[t \geq T_i].$$

Let us call this process an *inhomogeneous compound Poisson process.* In following Proposition we show the property of such process if the distributions of Z_i are Bernoulli.

Proposition 1.15. *Let* $N = \{N_t, t \geq 0\}$ *be an inhomogeneous Poisson process with intensity* $\lambda(t)$ *and jump times* $T_1, T_2, ...$ *Let* $\pi : [0, \infty) \to (0, 1)$ *be a right-continuous function and* $Z_i \sim Alt(\pi(T_i))$. *Then the family of random variables*

$$M_t = \sum_{i=1}^{\infty} Z_i \times \mathbb{I}[t \geq T_i].$$

is an inhomogeneous Poisson process with intensity $\lambda(t)\pi(t)$.

Proof. Since M is non-negative process and $M_t \leq N_t$ a.s. for any $t \geq 0$, then also $M_0 = 0$ a.s. The increments of N are independent by definition and also the values of Z_i are mutually independent. Therefore the increments of M are also independent.

Because the function π is right continuous, one can write $\pi(t+h) = \pi(t)+o(h)$. The probability of jump in a small interval is bounded by

$$\begin{aligned}\mathbb{P}[M_{t+h} = j+1 | M_t = j] &\geq \mathbb{P}[N_{t+h} = i+1 | N_t = i] \times (\pi(t) + o(h)) \\ &= (\lambda(t)h + o(h)) \times (\pi(t) + o(h)) = \lambda(t)\pi(t)h + o(h)\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}[M_{t+h} = j+1 | M_t = j] &\leq \lambda(t)\pi(t)h + o(h) + \mathbb{P}[N_{t+h} > i+1 | N_t = i] \\ &= \lambda(t)\pi(t)h + o(h),\end{aligned}$$

Which implies that the probability is equal to both the boundaries. Also

$$\mathbb{P}[M_{t+h} > j+1 | M_t = j] \leq \mathbb{P}[N_{t+h} > i+1 | N_t = i]$$

implies point *iv* in Definition 1.12. □

Chapter 2

Estimating Transition Intensity

In this chapter, we build several parametric models for intensity functions of inhomogeneous Markov process. For each of the models we propose a method for estimating the parameter using maximum likelihood. The theory of maximum likelihood estimates (MLE) is not described in this thesis. The reader can find general results for example in Anděl [2011], Chapter 7.6. We try to keep the notation close to the notation at said book.

We will estimate the parameter using embedded Markov chain and dwell times of multiple observations of Markov process on the same time interval. Let $\mathbf{X}^1, \ldots, \mathbf{X}^n$ be a random sample from distribution of inhomogeneous Markov process with state space S, intensity $\boldsymbol{Q}(t; \boldsymbol{\theta}_X)$, and initial probability $\boldsymbol{p}(\boldsymbol{\theta}_X)$, where $\boldsymbol{\theta}_X \in \mathbb{R}^H$ is the unknown parameter. Denote by K^ν the number of jumps of process $\mathbf{X}^\nu$, by $T_1^\nu, ..., T_{K^\nu}^\nu$ the times of jumps and by $Y_0^\nu, ..., Y_{K^\nu}^\nu$ the embedded Markov chain.

The likelihood function is of form

$$\begin{aligned} L_n(\boldsymbol{\theta}; \mathbf{X}^1, \ldots, \mathbf{X}^n) \\ = \prod_{\nu=1}^{n} \Bigg[p_{Y_0}(\boldsymbol{\theta}) \prod_{k=1}^{K^\nu} \left[f_{T_{k-1}^\nu, Y_{k-1}^\nu}(T_k^\nu - T_{k-1}^\nu; \boldsymbol{\theta}) \, \frac{q_{ij}(T_k^\nu; \boldsymbol{\theta})}{q_i(T_k^\nu; \boldsymbol{\theta})} \right] \\ \times \left(1 - F_{T_{K^\nu}^\nu, Y_{K^\nu}^\nu}(T - T_{K^\nu}^\nu; \boldsymbol{\theta})\right) \Bigg], \end{aligned} \tag{2.1}$$

where

$$F_{s,i}(x; \boldsymbol{\theta}) = 1 - \exp\left(- \int_s^{s+x} q_i(t; \boldsymbol{\theta}) dt \right)$$

is the cumulative distribution function of the dwell time given that the process transitions at time s into state i and

$$f_{s,i}(x; \boldsymbol{\theta}) = \exp\left(- \int_s^{s+x} q_i(t; \boldsymbol{\theta}) dt \right) q_i(s + x; \boldsymbol{\theta})$$

is the respective probability density function.

Let use denote the likelihood function by $L_n(\boldsymbol{\theta})$ and exclude the observed data $\mathbf{X}^1, \ldots, \mathbf{X}^n$ from the notation. Further we use standard notation $\ell_n(\boldsymbol{\theta}) = \log(L_n(\boldsymbol{\theta}))$ for log-likelihood, $\boldsymbol{U}_n(\boldsymbol{\theta}) = \partial \ell_n(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ for the score function, $\boldsymbol{I}_n(\boldsymbol{\theta}) = -1/n \times \partial \boldsymbol{U}_n(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ for the observed information matrix, and $\boldsymbol{I}(\boldsymbol{\theta}) = \mathbb{E}\left[\boldsymbol{I}_n(\boldsymbol{\theta})\right]$ for the Fisher information matrix.

The true vector parameter $\boldsymbol{\theta}_X$ is estimated by parameter $\hat{\boldsymbol{\theta}}_n$ that maximizes the likelihood function. In most cases, that is equivalent to solving the system of equations $\boldsymbol{U}_n(\boldsymbol{\theta}) = \mathbf{0}$.

In general case, one cannot calculate $\hat{\boldsymbol{\theta}}_n$ directly by solving the system of equations $\boldsymbol{U}_n(\boldsymbol{\theta}) = \mathbf{0}$. The MLE must be calculated by some numerical method. Here, we explain Newton-Raphson method. The method is iterative and updates the vector parameter at each iteration. Denote by $\boldsymbol{\theta}^{(r)}$ the r-th iteration of the algorithm. The $(r+1)$-th iteration is given by

$$\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^{(r)} + \left[n\boldsymbol{I}_n(\boldsymbol{\theta}^{(r)})\right]^{-1} \boldsymbol{U}_n(\boldsymbol{\theta}^{(r)}).$$

The algorithm is iterated until the value of $\boldsymbol{U}_n(\boldsymbol{\theta}^{(r)})$ is sufficiently close to $\mathbf{0}$. If the starting value $\boldsymbol{\theta}^{(0)}$ is well chosen, the algorithm converges to the solution, i.e. $\boldsymbol{\theta}^{(r)} \to \hat{\boldsymbol{\theta}}_n$ as $r \to \infty$.

2.1 Homogeneous Markov Process

The simplest Markov process is the homogeneous case. The intensity is constant in time and depends only on the unknow parameter $\boldsymbol{\theta}$, i.e. for each $i, j \in S$ we can write

$$\begin{aligned} q_{ij}(t;\boldsymbol{\theta}) &= q_{ij}(\boldsymbol{\theta}), \\ q_i(t;\boldsymbol{\theta}) &= q_i(\boldsymbol{\theta}). \end{aligned}$$

Also the cumulative distribution function and density can be simplified as

$$\begin{aligned} F_{t,i}(x;\boldsymbol{\theta}) &= 1 - \mathrm{e}^{-xq_i(\boldsymbol{\theta})}, \\ f_{t,i}(x;\boldsymbol{\theta}) &= q_i(\boldsymbol{\theta})\,\mathrm{e}^{-xq_i(\boldsymbol{\theta})}\,. \end{aligned}$$

Finally, the log-likelihood function is

$$\begin{aligned} &\ell_n(\boldsymbol{\theta}; X^1, \ldots, X^n) \\ &\quad = \sum_{\nu=1}^{n} \Bigg[\log(p_{Y_0^\nu}(\boldsymbol{\theta})) + \sum_{k=1}^{K^\nu} \left[\log(q_{Y_{k-1}^\nu Y_k^\nu}(\boldsymbol{\theta})) - q_{Y_{k-1}^\nu}(\boldsymbol{\theta})(T_k^\nu - T_{k-1}^\nu) \right] - \\ &\qquad\qquad - q_{Y_{K^\nu}^\nu}(\boldsymbol{\theta})(T - T_{K^\nu}^\nu) \Bigg]. \end{aligned}$$

This setup is still to general to derive any result that could be applicable to every parametrization of the intensity. Instead of that, we derive the properties of MLE for one special case. This process and its properties are taken from Prášková and Lachout [2012], Chapter 3.5.

Example 2.1 (Linear Growth Process)**.** The process is defined on state space $S = \mathbb{N}$ with inception at point $i = 1$ and intensities

$$q_{ij}(\theta) = \begin{cases} -i\theta & \text{if } j = i \\ i\theta & \text{if } j = i+1 \\ 0 & \text{else,} \end{cases}$$

where $\theta > 0$ is an unknown parameter (one-dimensional).

The distribution of the process at fixed time is given by (see the referred book for proof)

$$p_{1j}(0,t) = \mathrm{e}^{-jt\theta}(\mathrm{e}^{t\theta} - 1)^{j-1}.$$

That implies that the expected value is

$$\mathbb{E}\left[X_t\right] = \sum_{j=1}^{\infty} j\,\mathrm{e}^{-jt\theta}(\mathrm{e}^{t\theta} - 1)^{j-1} = \mathrm{e}^{t\theta}\,. \tag{2.2}$$

Because the process is straightforward and for each state $i \in S$ there exists exactly one state $j \in S$ for which the process can transition from i to j, we know that $Y_k = k + 1$. With that, we can write

$$\begin{aligned} q_{Y_{k-1},Y_k}(\theta) &= q_{k,k+1}(\theta) = k\theta \\ q_{Y_{k-1}}(\theta) &= q_k(\theta) = k\theta \\ p_{Y_0} &= p_1 = 1 \end{aligned}$$

Suppose we observe a random sample of size n over time interval $[0, T]$. The log-likelihood function is

$$\ell_n(\theta) = \sum_{\nu=1}^{n}\left[\sum_{k=1}^{K^\nu}\left[\log(k\theta) - k\theta(T_k^\nu - T_{k-1}^\nu)\right] - (K^\nu + 1)\theta(T - T_{K^\nu}^\nu)\right],$$

the score function is

$$U_n(\theta) = \sum_{\nu=1}^{n}\left[K^\nu\frac{1}{\theta} - \sum_{k=1}^{K^\nu} k(T_k^\nu - T_{k-1}^\nu) + (K^\nu + 1)(T - T_{K^\nu}^\nu)\right],$$

and the observed information matrix is

$$I_n(\theta) = \frac{1}{n}\sum_{\nu=1}^{n} K^\nu\frac{1}{\theta^2} = \overline{K}\frac{1}{\theta^2}$$

where $\overline{K} = \frac{1}{n}\sum_{\nu=1}^{n} K^\nu$ denotes the mean number of jumps.

Solving the equation $U_n(\theta) = 0$ one get the MLE

$$\hat{\theta}_n = \frac{\sum_{\nu=1}^{n} K^\nu}{\sum_{\nu=1}^{n}\left[\sum_{k=1}^{K^\nu} k(T_k^\nu - T_{k-1}^\nu) + (K^\nu + 1)(T - T_{K^\nu}^\nu)\right]}$$

Given the fact that $K^\nu = X_T^\nu - 1$ and (2.2) one can calculate the Fisher information matrix

$$I(\theta) = \mathbb{E}\left[I_n(\theta)\right] = \frac{\mathrm{e}^{T\theta} - 1}{\theta^2}.$$

The asymptotic distribution of the estimate is

$$\sqrt{n}(\hat{\theta}_n - \theta_X) \xrightarrow{d} N\left(0, \frac{\theta_X^2}{\mathrm{e}^{T\theta_X} - 1}\right) \text{ as } n \to \infty,$$

where θ_X denotes the true parameter.

△

At the end of this section, let us state a proposition about estimation of full model. The result is straightforward and intuitive and the proposition is left without a proof.

Proposition 2.1. *Let $\boldsymbol{X}^1, ..., \boldsymbol{X}^n$ be a random sample from homogeneous Markov process with transition rate $\boldsymbol{Q} = (q_{ij})_{i,j\in S}$ and initial probability vector $\boldsymbol{p} = (p_i)_{i\in S}$. Denote by N_{ij} total number of transitions from state i to state j, by M_i number of processes that initiates at state i, and by T_i total time spent in state i. If $T_i > 0$, the MLE estimator of transition rates and initial probabilities are*

$$\widehat{q_{ij}} = \frac{N_{ij}}{T_i}, \qquad \widehat{p_i} = \frac{M_i}{n}.$$

If $T_i = 0$, the result for initial probabilities holds and any estimate $\widehat{q_{ij}} \geq 0$ maximizes the (log-)likelihood.

Note that MLE estimators from Proposition 2.1 does not fulfill regularity conditions and not all results valid from MLE theory hold.

2.2 Separated Inhomogeneity

From this chapter on, we assume that the distribution of initial probability $\boldsymbol{p}$ is known and does not depend on the unknown parameter $\boldsymbol{\theta}$. In most of the examples we show in this thesis, the initial distribution is concentrated in single state space (either $p_0 = 1$ or $p_1 = 1$).

Assume that $\boldsymbol{X}$ is Markov process process with separable inhomogeneity, i.e. its transition rate matrix is

$$\boldsymbol{Q}(t) = \lambda(t)\boldsymbol{\Gamma}, \qquad \forall t \geq 0.$$

Let us further assume that the constant rate matrix $\boldsymbol{\Gamma} = \{\gamma_{ij}, i, j \in S\}$ is known and separated inhomogeneity λ follows a log-linear model

$$\lambda(t; \boldsymbol{\beta}) = \exp\left(\sum_{h=0}^{H} \beta_h \varphi_h(t)\right), \qquad \forall t \geq 0, \tag{2.3}$$

where $\varphi_h : [0, \infty) \to \mathbb{R}$ is given continuous function for $h = 0, ..., H$ and $\boldsymbol{\beta} = (\beta_0, ..., \beta_H)$ is vector of unknown parameters. Usually, we take $\varphi_0 = 1$ and call β_0 an intercept. The derivative of separated inhomogeneity with respect to parameter vector is

$$\frac{\partial \lambda(t; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \lambda(t; \boldsymbol{\beta})(\varphi_0(t), ..., \varphi_H(t))^\top \equiv \lambda(t; \boldsymbol{\beta})\boldsymbol{\varphi}(t).$$

We can adjust the likelihood function (2.1) for this particular case and calculate log-likelihood function

$$\ell_n(\boldsymbol{\beta}) = \sum_{\nu=1}^{n} \Bigg[\log(p_{Y_0^\nu}) + \sum_{k=1}^{K^\nu} \Bigg[\log(\gamma_{Y_{k-1}^\nu Y_k^\nu}) + \boldsymbol{\beta}^\top \boldsymbol{\varphi}(T_k^\nu) - \\ - \gamma_{Y_{k-1}^\nu} \int_{T_{k-1}^\nu}^{T_k^\nu} \lambda(t; \boldsymbol{\beta}) dt \Bigg] - \gamma_{Y_{K^\nu}^\nu} \int_{T_{K^\nu}^\nu}^{T} \lambda(t; \boldsymbol{\beta}) dt \Bigg],$$

the score statistic

$$\boldsymbol{U}_n(\boldsymbol{\beta}) = \sum_{\nu=1}^{n}\left[\sum_{k=1}^{K^\nu}\left[\boldsymbol{\varphi}(T_k^\nu) - \gamma_{Y_{k-1}^\nu}\int_{T_{k-1}^\nu}^{T_k^\nu}\lambda(t;\boldsymbol{\beta})\boldsymbol{\varphi}(t)dt\right] - \right.$$
$$\left. - \gamma_{Y_{K^\nu}^\nu}\int_{T_{K^\nu}^\nu}^{T}\lambda(t;\boldsymbol{\beta})\boldsymbol{\varphi}(t)dt\right],$$

and the observed information matrix

$$\boldsymbol{I}_n(\boldsymbol{\beta}) \quad = \quad \frac{1}{n}\sum_{\nu=1}^{n}\left[\sum_{k=1}^{K^\nu}\gamma_{Y_{k-1}^\nu}\int_{T_{k-1}^\nu}^{T_k^\nu}\lambda(t;\boldsymbol{\beta})\boldsymbol{\Omega}(t)dt \;+\; \gamma_{Y_{K^\nu}^\nu}\int_{T_{K^\nu}^\nu}^{T}\lambda(t;\boldsymbol{\beta})\boldsymbol{\Omega}(t)dt\right],$$

where $\boldsymbol{\Omega}(t) = \boldsymbol{\varphi}(t)(\boldsymbol{\varphi}(t))^\top$ for all $t \geq 0$.

It is not possible to solve the system of equations $\boldsymbol{U}_n(\boldsymbol{\beta}) = \mathbf{0}$ directly and one needs to use the Newton-Raphson iterative method to find the MLE $\hat{\boldsymbol{\beta}}_n$.

Example 2.2 (Testing homogeneity of Poisson Process)**.** While testing homogeneity of Poisson process, we are interested in test of null hypothesis H_0 against alternative hypothesis H_1, where

$$H_0 : \forall s,t \in [0,T] : \lambda(s) = \lambda(t),$$
$$H_1 : \exists s,t \in [0,T] : \lambda(s) \neq \lambda(t).$$

This hypothesis is hard to test. Instead of that, we will provide a weaker test in model (2.3)

$$H_0 : \beta_1 = ... = \beta_H = 0$$
$$H_1 : \exists h \in \{1,...,H\} : \beta_h \neq 0.$$

Denote the vector parameter under null hypothesis by $\boldsymbol{\beta}_0 = (\beta_{0,0}, 0, ..., 0)^\top$, where $\beta_{0,0} \in \mathbb{R}$ is nuisance parameter. The intensity under null hypothesis is $\lambda(t, \boldsymbol{\beta}_0) = \mathrm{e}^{\beta_{0,0}}$. The MLE of $\beta_{0,0}$ under null hypothesis is $\tilde{\beta}_{0,n} = \log(\frac{1}{n}\sum_{\nu=1}^{n} X_T^\nu)$.

We already know (from Chapter 1.4) that inhomogeneous Poisson process is inhomogeneous Markov process with unit constant rate from state i to $i+1$, where $i \in \mathbb{N}_0$, i.e. $\gamma_i = \gamma_{i,i+1} = 1$ for each $i \in S$.

The score statistic is simplified into

$$\boldsymbol{U}_n(\boldsymbol{\beta}) = \sum_{\nu=1}^{n}\sum_{k=1}^{K^\nu}\boldsymbol{\varphi}(T_k^\nu) - n\int_0^T \lambda(t;\boldsymbol{\beta})\boldsymbol{\varphi}(t)dt.$$

The observed information matrix does not depend on the data and, therefore, is equal to Fisher information matrix, i.e.

$$\boldsymbol{I}_n(\boldsymbol{\beta}) = \int_0^T \lambda(t;\boldsymbol{\beta})\boldsymbol{\Omega}(t)dt = \boldsymbol{I}(\boldsymbol{\beta}).$$

Without loss of generality assume $T = 1$. Suppose the model is given by $\varphi_h(t) = t^h$. Then the matrix $\boldsymbol{\Omega}(t)$ at position i,j is equal to t^{i+j-2} and the

respective position of Fisher information matrix $\boldsymbol{I}(\boldsymbol{\beta}_0)$ is (if the null hypothesis holds)

$$\int_0^1 \mathrm{e}^{\beta_{0,0}}\, t^{i+j-2} dt = \frac{\mathrm{e}^{\beta_{0,0}}}{i+j-1}.$$

Since $\tilde{\boldsymbol{\beta}}_n = (\tilde{\beta}_{0,n}, 0, ..., 0)^\top$ is consistent estimate of $\boldsymbol{\beta}_0$ under null hypothesis, also the matrix

$$\boldsymbol{I}(\tilde{\boldsymbol{\beta}}_n) = \left\{ \frac{\sum_{\nu=1}^n X_T^\nu}{n(i+j-1)}; i,j = 1, ..., H+1 \right\}$$

is consistent estimate of $\boldsymbol{I}(\boldsymbol{\beta}_0)$. The Rao score statistic is

$$R_n = \frac{1}{\sum_{\nu=1}^n X_T^\nu} \boldsymbol{U}_{-1,n}(\tilde{\boldsymbol{\beta}}_n)^\top \left[[\mathcal{H}_{H+1}]^{-1} \right]_{-1,-1} \boldsymbol{U}_{-1,n}(\tilde{\boldsymbol{\beta}}_n),$$

where $\boldsymbol{U}_{-1,n}(\tilde{\boldsymbol{\beta}}_n)$ is the score function at point $\tilde{\boldsymbol{\beta}}_n$ where the first element is excluded and $\mathcal{H}_k$ is the $k \times k$ Hilbert matrix. By notation $A_{-i,-j}$ we indicate the matrix A without i-th row and j-th column excluded. The statistic has (under null hypothesis) χ^2 distribution with H degrees of freedom so one can reject the null hypothesis if $R_n \geq \chi^2_H(1-\alpha)$.

Note that we are able to calculate the Rao score statistic without calculating the MLE of $\boldsymbol{\beta}$ (without assumption that the null hypothesis holds).

△

2.3 Constant Rate Matrix

In this chapter we follow notation from Chapter 1.3. Suppose that the separated inhomogeneity function λ is known and we are interested in estimating the constant rate matrix $\boldsymbol{\Gamma}$. We suppose that the matrix depends on unknown vector parameter $\boldsymbol{\theta}$. We denote the matrix by $\boldsymbol{\Gamma}(\boldsymbol{\theta})$. By Proposition 1.14, we are able to find a homogeneous transformation of the process $\boldsymbol{X}$. Then, we can use this homogeneous transformation and results from Chapter 2.2 to estimate the parameter $\boldsymbol{\theta}$.

If also the separated inhomogeneity were unknown, we cannot find the homogeneous transformation $X_{\tau(t)}$ and we need to estimate both the constant rate matrix and the separated inhomogeneity function. Suppose that the vector parameter is divided into $(\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$, where $\boldsymbol{\beta} \in \mathbb{R}^{H_1}$, $\boldsymbol{\theta} \in \mathbb{R}^{H_2}$, and $H_1 + H_2 = H$. Further suppose that function $\lambda = \lambda(\boldsymbol{\beta})$ follows log-linear model similar to Chapter 2.2 and matrix $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}(\boldsymbol{\theta})$ has similar parametrization as in Chapter 2.1.

Note that we do not need to keep the intercept in the log-linear term for separated inhomogeneity function in most of reliable models. If for each $\boldsymbol{\theta} \in \Theta$ and $c > 0$ there exists $\boldsymbol{\theta}' \in \Theta$ such that $\boldsymbol{\Gamma}(\boldsymbol{\theta}') = c\boldsymbol{\Gamma}(\boldsymbol{\theta})$, then it is easier to set $c = \mathrm{e}^{\beta_0} = 1$, i.e. the intercept $\beta_0 = 0$ needs not be included in the model.

The log-likelihood function

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{\nu=1}^n \Bigg[\log(p_{Y_0^\nu}) + \sum_{k=1}^{K^\nu} \Bigg[\log(\gamma_{Y_{k-1}^\nu Y_k^\nu}(\boldsymbol{\theta})) + \boldsymbol{\beta}^\top \boldsymbol{\varphi}(T_k^\nu) - \\ - \gamma_{Y_{k-1}^\nu}(\boldsymbol{\theta}) \int_{T_{k-1}^\nu}^{T_k^\nu} \lambda(t; \boldsymbol{\beta}) dt \Bigg] - \gamma_{Y_{K^\nu}^\nu}(\boldsymbol{\theta}) \int_{T_{K^\nu}^\nu}^{T} \lambda(t; \boldsymbol{\beta}) dt \Bigg]$$

is again hard to work with without any assumption about the constant intensity rate $\boldsymbol{\Gamma}(\boldsymbol{\theta})$. Instead of deriving other statistics in general, we show one example of such model.

Example 2.3 (Independent queuing systems)**.** Suppose that there exist H_2 independent queuing systems for which the arrivals follow inhomogeneous processes. Further suppose that the queuing systems are mutually independent and the intensities of individual systems differs only in scale. Denote by $\lambda_h(t;\boldsymbol{\beta})$ the intensity of h-th system at time t for $h = 1, ..., H_2$. Then the ratio $\lambda_h(t;\boldsymbol{\beta})/\lambda_h(s;\boldsymbol{\beta})$ does not depend on the system h for any $s, t \geq 0$ such that $\lambda_h(s;\boldsymbol{\beta}) > 0$. One reliable model for intensities could be

$$\lambda_h(t;\boldsymbol{\beta},\boldsymbol{\theta}) = \theta_h \exp\left(\sum_{h=1}^{H} \beta_h \varphi_h(t)\right), \qquad \forall t \geq 0, h = 1, ..., H_2.$$

Denote by $(i,j) \sim h$ the fact that that the transition from i to j is equivalent to one arrival in system h. Now, we can write the constant intensities as

$$\gamma_{i,j}(\boldsymbol{\theta}) = \sum_{h=1}^{H_2} \theta_h \mathbb{I}[(i,j) \sim h]$$

and their derivations are $\partial\gamma_{i,j}(\boldsymbol{\theta})/\partial\theta_h = \mathbb{I}[(i,j) \sim h]$. For each state $i \in S$ and system $h = 1, ..., H_2$ there exist exactly one state $j \in S$ such that $(i,j) \sim h$. Therefore, the total constant intensities are $\gamma_i(\boldsymbol{\theta}) = \sum_{h=1}^{H_2} \theta_h \equiv s(\boldsymbol{\theta})$ for all $i \in S$.

The first and the second derivatives of log-likelihood function with respect to vector parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are

$$\begin{aligned}
\frac{\partial \ell_n(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} &= \sum_{\nu=1}^{n}\sum_{k=1}^{K^\nu} \boldsymbol{\varphi}(T_k^\nu) - n s(\boldsymbol{\theta}) \int_{T_{K^\nu}^\nu}^{T} \lambda(t;\boldsymbol{\beta})\boldsymbol{\varphi}(t)dt, \\
\frac{\partial \ell_n(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial \theta_h} &= \frac{1}{\theta_h}\sum_{\nu=1}^{n}\sum_{k=1}^{K^\nu} \mathbb{I}[(Y_{k-1}^\nu, Y_k^\nu) \sim h] - n\int_0^T \lambda(t;\boldsymbol{\beta})dt, \\
\frac{\partial^2 \ell_n(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top} &= -n s(\boldsymbol{\theta}) \int_0^T \lambda(t;\boldsymbol{\beta})\boldsymbol{\Omega}(t)dt. \\
\frac{\partial^2 \ell_n(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial \boldsymbol{\beta}\partial \theta_h} &= -n \int_0^T \lambda(t;\boldsymbol{\beta})\boldsymbol{\varphi}(t)dt, \\
\frac{\partial^2 \ell_n(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial \theta_{h_1}\partial \theta_{h_2}} &= \frac{-1}{\theta_{h_1}^2}\sum_{\nu=1}^{n}\sum_{k=1}^{K^\nu} \mathbb{I}[(Y_{k-1}^\nu, Y_k^\nu) \sim h_1] \times \mathbb{I}[h_1 = h_2],
\end{aligned}$$

Denote by $N_{n,h} = \sum_{\nu=1}^{n}\sum_{k=1}^{K^\nu} \mathbb{I}[(Y_{k-1}^\nu, Y_k^\nu) \sim h]$ the total number of arrivals in all observations of system h. Then we have the score function

$$\boldsymbol{U}_n(\boldsymbol{\beta},\boldsymbol{\theta}) = \begin{pmatrix} \sum_{\nu=1}^{n}\sum_{k=1}^{K^\nu} \boldsymbol{\varphi}(T_k^\nu) - n s(\boldsymbol{\theta}) \int_{T_{K^\nu}^\nu}^{T} \lambda(t;\boldsymbol{\beta})\boldsymbol{\varphi}(t)dt \\ \dfrac{N_{n,1}}{\theta_1} \displaystyle\sum_{\nu=1}^{n} -n\int_0^T \lambda(t;\boldsymbol{\beta})dt \\ \vdots \\ \dfrac{N_{n,H_2}}{\theta_{H_2}} \displaystyle\sum_{\nu=1}^{n} -n\int_0^T \lambda(t;\boldsymbol{\beta})dt \end{pmatrix}$$

and the observed information matrix

$$\boldsymbol{I}_n(\boldsymbol{\beta},\boldsymbol{\theta}) = \begin{pmatrix} s(\boldsymbol{\theta})\int_0^T \lambda(t;\boldsymbol{\beta})\boldsymbol{\Omega}(t)dt & \boldsymbol{D}_{\boldsymbol{\theta},\boldsymbol{\beta}}^\top \\ \boldsymbol{D}_{\boldsymbol{\theta},\boldsymbol{\beta}} & \boldsymbol{D}_{\boldsymbol{\theta},\boldsymbol{\theta}} \end{pmatrix},$$

where

$$\boldsymbol{D}_{\boldsymbol{\theta},\boldsymbol{\beta}} = \begin{pmatrix} \int_0^T \lambda(t;\boldsymbol{\beta})\boldsymbol{\varphi}(t)^\top dt \\ \vdots \\ \int_0^T \lambda(t;\boldsymbol{\beta})\boldsymbol{\varphi}(t)^\top dt \end{pmatrix}, \qquad \boldsymbol{D}_{\boldsymbol{\theta},\boldsymbol{\theta}} = \begin{pmatrix} \frac{N_{n,1}}{n\theta_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{N_{n,H_2}}{n\theta_1^2} \end{pmatrix}$$

are matrices of dimensions $H_2 \times H_1$ and $H_2 \times H_2$, respectively.

The MLE of $(\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$ can be found using the Newton-Rhapson algorithm. △

2.4 General Markov Process

In the most general case, we suppose that each transition intensity follows a log-linear model. However, we do not assume any relations between individual intensities as we did in previous chapters.

Let $\mathcal{I} = \{1, ..., H\}$ be an index set with subsets $\mathcal{I}_{ij} \subset \mathcal{I}$ for $i, j \in S$, $i \neq j$. Let $\varphi_h : [0, \infty) \to \mathbb{R}$ be real functions for each $h \in \mathcal{I}$. Assume that the intensity of transition from state $i \in S$ to state $j \in S \setminus \{i\}$ at time $t \geq 0$ are

$$q_{ij}(t;\boldsymbol{\beta}) = \begin{cases} \exp\left(\sum_{h\in\mathcal{I}_{ij}} \beta_h \varphi_h(t)\right) & |\mathcal{I}_{ij}| > 0 \\ 0 & \text{else,} \end{cases}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_H)^\top \in \mathbb{R}^H$ is unknown parameter. We do not assume that the subsets $\{\mathcal{I}_{ij}; i, j \in S, i \neq j\}$ are mutually disjoint. Let us use the vector notation for the functions $\boldsymbol{\varphi}(t) = (\varphi_1(t), ..., \varphi_H(t))^\top$.

The log-likelihood function can be written as

$$\ell_n(\boldsymbol{\beta}) = \sum_{\nu=1}^n \left[\log(p_{Y_0}) + \sum_{k=1}^{K^\nu} \left[\log(q_{Y_{k-1}^\nu Y_k^\nu}(T_k^\nu;\boldsymbol{\beta})) - \int_{T_{k-1}^\nu}^{T_k^\nu} q_{Y_{k-1}^\nu}(t;\boldsymbol{\beta})dt \right] - \int_{T_{K^\nu}^\nu}^{T} q_{Y_{K^\nu}^\nu}(t;\boldsymbol{\beta})dt \right].$$

Let us begin with the derivatives of individual terms:

$$\begin{aligned} \frac{\partial \log(q_{ij}(t;\boldsymbol{\beta}))}{\partial \beta_h} &= \varphi_h(t)\mathbb{I}[h \in \mathcal{I}_{ij}] \\ \frac{\partial^2 \log(q_{ij}(t;\boldsymbol{\beta}))}{\partial \beta_{h_1}\partial \beta_{h_2}} &= 0 \\ \frac{\partial q_i(t;\boldsymbol{\beta})}{\partial \beta_h} &= \sum_{j\in S\setminus\{i\}} \varphi_h(t) q_{ij}(t;\boldsymbol{\beta})\mathbb{I}[h \in \mathcal{I}_{ij}] \\ \frac{\partial^2 q_i(t;\boldsymbol{\beta})}{\partial \beta_{h_1}\partial \beta_{h_2}} &= \sum_{j\in S\setminus\{i\}} \varphi_{h_1}(t)\varphi_{h_2}(t) q_{ij}(t;\boldsymbol{\beta})\mathbb{I}[h_1, h_2 \in \mathcal{I}_{ij}] \end{aligned}$$

The h-th element of the score function is given by

$$\boldsymbol{U}_{n,h}(\boldsymbol{\beta}) = \sum_{\nu=1}^{n} \Bigg[\sum_{k=1}^{K^\nu} \varphi_h(T_k^\nu) \mathbb{I}[h \in \mathcal{I}_{Y_{k-1}^\nu Y_k^\nu}] - \sum_{k=1}^{K^\nu} \sum_{j \in S \setminus \{Y_{k-1}^\nu\}} \mathbb{I}[h \in \mathcal{I}_{Y_{k-1}^\nu j}] \int_{T_{k-1}^\nu}^{T_k^\nu} \varphi_h(t) q_{Y_{k-1}^\nu j}(t; \boldsymbol{\beta}) dt - \sum_{j \in S \setminus \{Y_{K^\nu}^\nu\}} \mathbb{I}[h \in \mathcal{I}_{Y_{K^\nu}^\nu j}] \int_{T_{K^\nu}^\nu}^{T} \varphi_h(t) q_{Y_{K^\nu}^\nu j}(t; \boldsymbol{\beta}) dt \Bigg]. \quad (2.4)$$

The element of the observed information matrix in row h_1 and column h_2 is given by

$$\boldsymbol{I}_{n,h_1,h_2}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{\nu=1}^{n} \Bigg[\sum_{k=1}^{K^\nu} \sum_{j \in S \setminus \{Y_{k-1}^\nu\}} \mathbb{I}[h \in \mathcal{I}_{Y_{k-1}^\nu j}] \int_{T_{k-1}^\nu}^{T_k^\nu} \varphi_{h_1}(t) \varphi_{h_2}(t) q_{Y_{k-1}^\nu j}(t; \boldsymbol{\beta}) dt + \sum_{j \in S \setminus \{Y_{K^\nu}^\nu\}} \mathbb{I}[h \in \mathcal{I}_{Y_{K^\nu}^\nu j}] \int_{T_{K^\nu}^\nu}^{T} \varphi_{h_1}(t) \varphi_{h_2}(t) q_{Y_{K^\nu}^\nu j}(t; \boldsymbol{\beta}) dt \Bigg]. \quad (2.5)$$

The MLE of vector parameter $\boldsymbol{\beta}$ may be estimated using Newton-Raphson method.

2.5 Regression Model

In the last part of this chapter, we introduce a model that allows dependence of the Markov Process on other variables. This model is similar to Cox proportional hazard model (see Cox [1972]). Suppose we observe a piecewise constant and right-continuous function (covariate) $\boldsymbol{z} : [0, \infty) \to \mathbb{R}^M$ along with each process.

Further let function $\boldsymbol{\varphi} : [0, \infty) \times \mathbb{R}^M \to \mathbb{R}^H$ depend both on time and the covariate. Denote the vector function by $\boldsymbol{\varphi}(t, \boldsymbol{z}(t))$. We keep the assumptions from Chapter 2.4 and modify only the definition of intensities

$$q_{ij}(t, \boldsymbol{z}(t); \boldsymbol{\beta}) = \begin{cases} \exp\left(\sum_{h \in \mathcal{I}_{ij}} \beta_h \varphi_h(t, \boldsymbol{z}(t))\right) & |\mathcal{I}_{ij}| > 0 \\ 0 & \text{else.} \end{cases}$$

All the results from Chapter 2.4 hold. One only needs to add an argument $\boldsymbol{z}^\nu(t)$ to functions q_i, q_{ij}, and φ_h in formulas for the log-likelihood function, the score function, and the observed information matrix. The function $\boldsymbol{z}^\nu(t)$ denotes the observed covariate of the process $\boldsymbol{X}^\nu$.

Chapter 3

Stochastic Optimization

In this chapter we discuss methods for finding an optima of functions that we are unable to evaluate exactly, but we are able to sample from distribution that is related to such objective function. For example, the objective function might be some characteristics of a random variable. Suppose that pair of random variables $(\mathbf{X}, Y) \in \mathbb{R}^{d+1}$ follows model

$$\begin{aligned} \mathbb{E}\left[Y \middle| \mathrm{X} = \mathbf{x}\right] &= \mu(\mathbf{x}), \\ \mathrm{var}[Y \middle| \mathrm{X} = \mathbf{x}] &= \sigma^2(\mathbf{x}). \end{aligned}$$

where $\mu : \mathbb{R}^d \to \mathbb{R}$ and $\sigma : \mathbb{R}^d \to [0, \infty)$ are unknown functions. Further suppose that we are able to sample from the conditional distribution $Y \,|\, \mathrm{X}$.

Denote by $f(y|\mathbf{x})$, $F(y|\mathbf{x})$, and $F^{-1}(\alpha|\mathbf{x})$ the conditional density, cumulative distribution function, and quantile function of $Y \,|\, \mathrm{X} = \mathbf{x}$.

For the optimization problem

$$\begin{aligned} \max_{\mathbf{x}} \quad & g(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{3.1}$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is given set, we will consider following objective functions

i) $g(\mathbf{x}) = \mu(\mathbf{x})$,

ii) $g(\mathbf{x}) = \mu(\mathbf{x}) - \alpha\sigma^2(\mathbf{x}), \qquad \alpha > 0,$

iii) $g(\mathbf{x}) = F^{-1}(\alpha|\mathbf{x}), \qquad (0 < \alpha < 1)$, and

iv) $g(\mathbf{x}) = \mathbb{E}\left[Y | Y \leq F^{-1}(\alpha|\mathbf{x}), \mathrm{X} = \mathbf{x}\right], \qquad (0 < \alpha < 1).$

The optimization problem with objective function in *ii* is called Markowitz problem in portfolio theory. The last two objective functions stand for Value at Risk (VaR) and Conditional Value at Risk (CVaR). These optimization problems are well known and the reader can find more about these in any literature about stochastic optimization. We will focus on situations when the functions μ and σ are not known but we are able to generate pairs (X, Y).

Denote by $\mathbf{x}^*$ the optimal point and by $g^* = g(\mathbf{x}^*)$ the optimum of the problem. Note that $\mathbf{x}^*$ need not be unique for some problems. In such cases, we

will use it to denote whole set of optimal points. The maximization and minimization problems are mutually reversible and, therefore, we will only consider maximization problems in this thesis.

We have already stated, that the distribution of the pair of random variables $(\mathbf{X}, Y)$ is unknown. Also, the function g, optimal point $\mathbf{x}^*$, and optimum g^* are unknown (in sense that we cannot explicitly state it). In this chapter we will propose and study simulation methods for estimating solution to the problem. For that, we simulate random sample $(\mathrm{X}_1, Y_1), ..., (\mathrm{X}_n, Y_n)$ from a population (X, Y). This data will be used for estimation of the objective function.

The data needs to be generated in way that $Y_1 | \mathrm{X}_1, ..., Y_n | \mathrm{X}_n$ are mutually independent. Let us not make any assumptions about the distribution of X. Variables $\mathrm{X}_1, ..., \mathrm{X}_n$ need not to be independent or identically distributed.

In following chapters, we briefly introduce two optimization methods for given problem. Both of them are iterative and use the simulated data described above. The first method, a response surface method, seeks for the optimum locally and updates the decision variable based on data generated close to the actual value of the decision variable. On the other hand, the second method, a cross entropy method for optimization, seeks for the optimum globally, uses all information available and selects the values that perform the best.

3.1 Response Surface Method

The idea of the response surface method is to estimate the objective function g and maximize the estimate. Let us denote $\widehat{g}_n$ the estimate of the function g based on random sample of size n. Let us consider a bounded set $\mathcal{X}_0 \subseteq \mathcal{X}$ that contains the optimal point, i.e. $\mathbf{x}^* \in \mathcal{X}_0$. The set $\mathcal{X}_0$ is our best guess of where the optimal point could be and we are able to estimate the objective function on it. The reason for reduction of the set will be described later. Than we can derive a maximization problem from the initial problem (3.1). The problem can be written in form

$$\begin{aligned} \max_{\mathbf{x}} \quad & \widehat{g}_n(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{X}_0. \end{aligned} \tag{3.2}$$

Let us also denote by $\widehat{\mathbf{x}}_n^*$ the optimal point and by $\widehat{g}_n^* \equiv \widehat{g}_n(\widehat{\mathbf{x}}_n^*)$ the optimum of problem (3.2). We will call $\widehat{\mathbf{x}}_n^*$ and $\widehat{g}_n^*$ *estimated optimal point* and *estimated optimum* of problem (3.1). We are also interested in $g(\widehat{\mathbf{x}}_n^*)$, which is the value of the objective function at the estimated optimal point. This value is, however, unknown and we try to estimate it.

Obviously, the problem has two main parts that should be solved. First, we need to estimate the objective function $\widehat{g}_n(\mathbf{x})$. Second, this objective function needs to be maximized. Here we start with the second part. Later, we find out that more than the objective function g needs to be estimated for the purpose of optimization.

We briefly introduce two iterative methods for numerical optimization. That is we take an initial point $\mathbf{x}^{(0)}$. Then we iteratively update $x_k \mapsto x_{k+1}$ until the sequence $x_0, x_1, ...$ reaches selected criterion. Both of the methods are described in Kroese et al. [2011], Appendix C. Then, we generalize these two numerical methods to be applicable in simulated optimization. In the same book, in Chapters

11 and 12, different methods for simulated optimization are described.

Let us suppose that the derivatives of g exist and are known. Denote by $\nabla_g(\mathbf{x})$ the gradient of g at $\mathbf{x}$. The *gradient descent* method updates the sequence in form

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \nabla_g(\mathbf{x}^{(k)}),$$

where $(\alpha_k)_{k=1}^{\infty}$ is sequence of (typically small) step sizes. Intuitively, it is appropriate to take sequence for which

$$\alpha_k \to 0 \text{ as } k \to \infty, \qquad \sum_{k=0}^{\infty} \alpha_k = \infty.$$

Let us further assume that also the second derivatives of g exist and are known and the Hessian, $\nabla_g^2(\mathbf{x})$, is invertible on $\mathcal{X}_0$. The *Newton's method* updates the sequence in form

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - (\nabla_g^2(\mathbf{x}^{(k)}))^{-1} \nabla_g(\mathbf{x}^{(k)}).$$

These methods are applicable only if the derivatives were known. However, alike the function g, its derivatives are also unknown. These needs to be estimated as well. Denote by $\widehat{\nabla_g}(\mathbf{x})$ and $\widehat{\nabla_g^2}(\mathbf{x})$ the estimated gradient and estimated Hessian of g at point $\mathbf{x}$. Instead of estimating single function, one needs to estimate $d+1$ functions (function g and its derivatives) for gradient method and $(d+1)(d+2)/2$ functions for Newton's method. This indicates that the gradient method becomes more suitable for high-dimensional problems.

With this setup, there are several questions to be answered. In this chapter we try to find answers for these questions.

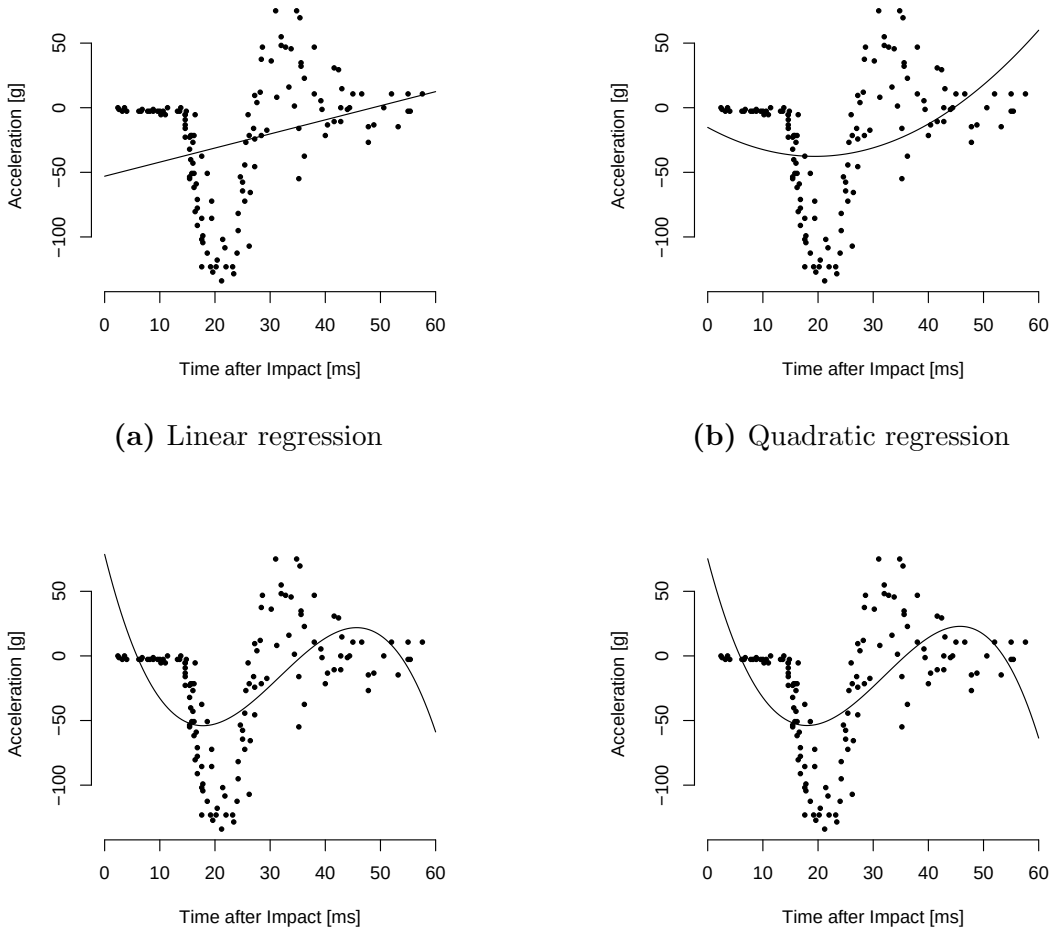
- How to estimate the objective function g and its derivatives? This, obviously, depends on the type of the function. Different techniques will be used, for example, for mean and for quantile.
- How large should the simulated sample be in order to get close to the actual optimum? For example, in sense of

  $$\mathbb{P}[g(\widehat{\mathbf{x}}_n^*) \leq g^* - \varepsilon] < \delta$$

  for suitable parameters $\varepsilon, \delta > 0$.
- How to construct subset $\mathcal{X}_0$? How to sample variables $\mathrm{X}_1, ..., \mathrm{X}_n$? These two questions are strongly related and need to be answered together.

Clearly, there are no general answers to those questions that could be applied to every single problem. The reader should be aware that the proposed solutions are reliable to the problems stated in this thesis. On the other hand, it should give an intuition of how to find answer to different problems.

One of the simplest possibilities to estimate the objective function is polynomial regression. The advantages of this approach are that fitting a polynomial is relatively computationally undemanding and polynomial function $\widehat{g}$ is easy to maximize numerically (or even analytically). On the other hand, there are many dependencies for which the polynomial regression fails. See, for example, Figure 3.1.



(a) Linear regression

(b) Quadratic regression

(c) Kubic regression

(d) Quartic regression

Figure 3.1: Polynomial fits to the motorcycle data. The bias of estimates is large even in case of high-order polynomials. The plots are adapted from Fan and Gijbels [2003].

Another solution is to estimate the objective function locally for each value $\mathbf{x} \in \mathcal{X}$. For that, we assign weights to individual observations based on their distance from $\mathbf{x}$. The further the point is, the lower the weight is. Also, it is possible to assign zero weights to observations from certain distance and exclude these points from estimation. Let us assign the weights to individual observations with a kernel function.

Definition 3.1. *Let $K : \mathbb{R} \to \mathbb{R}$ be a real function satisfying*

$$\int_{\mathbb{R}} K(x)dx = 1, \qquad \int_{\mathbb{R}} xK(x)dx = 0, \qquad 0 < \int_{\mathbb{R}} x^2K(x)dx < \infty$$

Then we call K a kernel function.

Cleveland [1979] proposed also other properties that a weight function should have. These are non-negativity, symmetry around zero, monotonicity for all non-negative values, and zero values outside the unit circle. The first three properties

are fulfilled for the majority of common kernel functions. The fourth property is reasonable especially for computational purposes, but there are several kernel functions that violate this property. Later, we will see that symmetry of kernel function is unnecessary for our purpose.

Clearly, any density function of a non-degenerate random variable with zero mean and finite variance is a kernel function. Three common examples are the Gaussian kernel

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),$$

the Epanechnikov kernel

$$K(x) = \begin{cases} \frac{3}{4}(1-x^2) & \text{if } |x| \leq 1 \\ 0 & \text{else,} \end{cases}$$

and the Tricube kernel

$$K(x) = \begin{cases} \frac{70}{81}(1-|x|^3)^3 & \text{if } |x| \leq 1 \\ 0 & \text{else.} \end{cases}$$

The Tricube kernel was originally proposed by Cleveland [1979] and is implemented in R software, package stats, function loess.

Assume that the estimated function is $\mu(\mathbf{x}) = \mathbb{E}[Y | \mathrm{X} = \mathbf{x}]$. Then the *local polynomial estimator* is in form of

$$\begin{aligned} \widehat{\mu}(\mathbf{x}) =& \widehat{\varphi}_{\mathbf{x}}(0) \\ \widehat{\varphi}_{\mathbf{x}} =& \underset{\varphi}{\operatorname{argmin}} \quad \sum_{i=1}^{n} K\left(\frac{\|\mathrm{X}_i - \mathbf{x}\|}{h_\alpha(\mathbf{x})}\right) (Y_i - \varphi(\mathrm{X}_i - \mathbf{x}))^2 \\ & \text{s.t.} \qquad \varphi \text{ is a polynomial of degree } p. \end{aligned} \tag{3.3}$$

Function $\|\cdot\|$ is the Euclidean norm of a vector. Function $h_\alpha(\mathbf{x})$ is distance between $\mathbf{x}$ and r-th closest variable within $\mathrm{X}_1, ..., \mathrm{X}_n$, where $r = \lceil \alpha n \rceil$. By selecting the kernel function K we control the weights of individual simulations. The span α and the degree p of polynomial φ are also of our choice.

We can use the estimated polynomial $\widehat{\varphi}_{\mathbf{x}}$ to estimate also derivatives of mean function by setting

$$\begin{aligned} \widehat{\nabla_\mu}(\mathbf{x}) &= \nabla_{\widehat{\varphi}_{\mathbf{x}}}(0), \\ \widehat{\nabla^2_\mu}(\mathbf{x}) &= \nabla^2_{\widehat{\varphi}_{\mathbf{x}}}(0). \end{aligned}$$

For the gradient descent method, we need the local polynomial at least linear ($p \geq 1$). For the Newton's method, we need at least quadratic polynomials ($p \geq 2$).

Let us introduce a method that combines the gradient descent method and the Newton's method. We estimate the second-order derivative of the objective function only as a diagonal matrix. There are two main reasons for not estimating whole Hessian. First, the number of parameters that needs to be estimated is too high and the method is computationally demanding for high-dimensional problems. Second, the Newton's method needs the Hessian (or estimated Hessian)

to be invertible. If the matrix is diagonal with non-zero diagonal elements, the inverse always exists. We can define the iterative method as

$$x_i^{(k+1)} = \begin{cases} x_i^{(k)} - \dfrac{\widehat{\nabla}_{g,i}(\mathbf{x}_i^{(k)})}{\widehat{\nabla}^2_{g,i,i}(\mathbf{x}_i^{(k)})}, & \widehat{\nabla}^2_{g,i,i}(\mathbf{x}_i^{(k)}) < -\frac{1}{\alpha_k} \\ x_i^{(k)} + \alpha_k \widehat{\nabla}_{g,i}(\mathbf{x}_i^{(k)}), & \text{else,} \end{cases} \tag{3.4}$$

where $\widehat{\nabla}_{g,i}$ and $\widehat{\nabla}^2_{g,i,i}$ denote the estimated derivatives of the objective function with respect to x_i. The condition on the first line of the definition of the next iteration has two reasons. First, we want to use the second derivatives only if the function is locally estimated as concave down. Second, we want to keep the algorithm stable even around estimated inflexion points where the estimated second derivative is close to zero.

Algorithm 3.1 (Response Surface)**.**

1. *Set initial value* $\mathbf{x}^{(0)}$ *and put* $k := 0$*.*
2. *Generate data (multiple independent observations) from distribution* (X, Y)*, where the predictors* X *have the distribution centered at* $\mathbf{x}^k$*. Join the data with previously generated data (if any).*
3. *Fit the local polynomial regression at point* $\mathbf{x}^{(k)}$*. If the estimated first-order derivatives of* μ *are sufficiently close to zero then continue with step 5.*
4. *Update* $\mathbf{x}^{(k+1)}$ *by (3.4) based on estimated derivatives, increase* $k := k + 1$ *and repeat steps 2–4.*
5. *Return value* $\mathbf{x}^{(k)}$*.*

Note that the algorithm does not return the estimated optimum of given problem. The reason for that is that local polynomial regression provides, in general, biased estimate of the estimated function. One can generate multiple observations from distribution $Y \,|\, \mathrm{X} = \widehat{\mathbf{x}}^*$.

Let us illustrate the method on simple example of finding an optimum of exponential of an paraboloid.

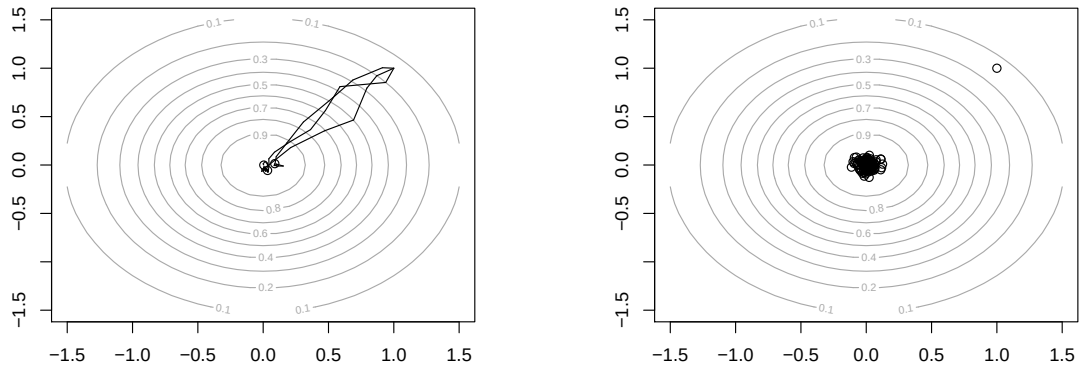
Example 3.1. Let $Y = \mathrm{e}^{-\|\mathbf{x}\|^2} + \varepsilon$, where $\mathbf{x} \in \mathbb{R}^2$ and $\varepsilon \sim N(0, 1)$. Suppose that the optimization problem is

$$\begin{aligned} \max_{\mathbf{x}} \quad & \mu(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathbb{R}^2. \end{aligned}$$

Clearly, the optimum is $\mu^* = 1$ at point $\mathbf{x} = (0, 0)^\top$. However, we want to estimate the optimum using the Algorithm 3.1.

We set the initial guess $\mathbf{x}^{(0)} = (1, 1)^\top$ and at k-th iteration we generate 10 000 simulations with distribution of predictors as $\mathrm{X} \sim N_2(\mathbf{x}^{(k)}, 0.01 \times \mathbb{I}_2)$, where $\mathbb{I}_2$ is 2×2 unit matrix. The sequence of steps lengths was defined as $\alpha_k = \frac{0.6}{\sqrt[3]{k+1}}$. At k-th iteration we used $\frac{0.75}{\sqrt{k+1}} \times 100\%$ nearest observations to fit the polynomial regression with Epanechnikov kernel as weights.

The average distance of the estimated optimum from the actual optimum was 0.074 (confidence interval: 0.046–0.101). The mean number of iterations was



(a) Paths of three runs. **(b)** End points of 100 runs.

Figure 3.2: Visualized results of response surface method from Example 3.1.

23.3. Some paths of the algorithm and all the estimated optimums are visualized in Figure 3.2.

△

Until now, we supposed only mean function as the objective function. The method can be generalized for various other functions. One can easily replace the local polynomial regression with any other local model. For example, by using the local quantile regression we can maximize the Value at Risk of random variable $Y \,|\, \mathrm{X} = \mathbf{x}$. For more local models that could be applied in the response surface method see Fan and Gijbels [2003].

3.2 Cross Entropy Method

Another way to find the optimum of given function with a noise is to use the cross entropy method. This method searches for the optimum by narrowing the set over which the objective function is optimized. In general, this method does not estimate the maximum of the conditional mean as the response surface method does. Also, only a little is known about the convergence of the algorithm.

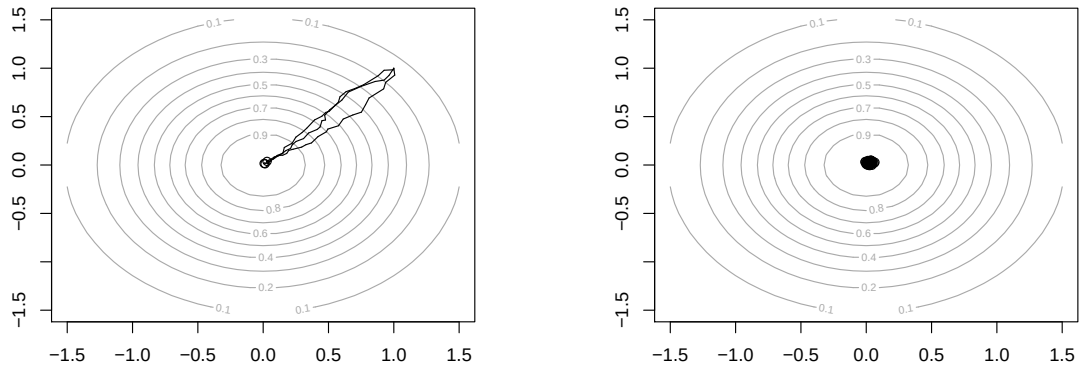
The algorithm has two main steps. In the first step, the independent values are generated from some distribution and the values that performs the best are selected for the second step. In the second step, the sampling distribution is updated based on the best performances from the first step. It is suitable to use the sampling distributions fixed up to the choice of the distribution parameter and update only the parameter.

Following algorithm is adapted from Kroese et al. [2011].

Algorithm 3.2 (Cross Entropy for Noisy Optimization)**.**

1. *Set initial parameter* $\boldsymbol{\psi}^{(0)}$ *and counter* $k := 1$.

2. *Generate data (multiple independent observations) of size* n *from distribution* (X, Y)*, where the predictors have the distribution* $\mathrm{X} \sim f(\,\cdot\,; \boldsymbol{\psi}^{(k-1)})$*. Let* γ_k *be the* $(1 - \varrho)$*-quantile of* $Y_1, ..., Y_n$.



(a) Paths of three runs. **(b)** End points of 100 runs.

Figure 3.3: Visualized results of cross entropy method from Example 3.2.

3. *Estimate the parameter* $\boldsymbol{\psi}^{(k)}$ *from the best performing observations using maximum likelihood*

$$\boldsymbol{\psi}^{(k)} = \max_{\psi} \sum_{\mathrm{X}_i \geq \gamma_k} \log(f(\mathrm{X}_i; \boldsymbol{\psi})).$$

4. *If the distribution* $f(\,\cdot\,; \boldsymbol{\psi}^{(k)})$ *is almost degenerated then continue with step 6, otherwise increase* $k := k + 1$ *and repeat steps 2–4.*

5. *Return value* $\mathbf{x}$ *such that the distribution* $f(\,\cdot\,; \boldsymbol{\psi}^{(k)})$ *is almost degenerated at point* $\mathbf{x}$*.*

In the previous algorithm, the stopping criterion is vaguely described as *almost degenerated distribution* $f(\,\cdot\,; \boldsymbol{\psi})$. We use the 1-norm of the variance matrix of the distribution. If the norm is sufficiently close to zero we say that the distribution is almost degenerated.

To compare the algorithm with the response surface method, we use the cross entropy to solve the same problem as we did in previous chapter.

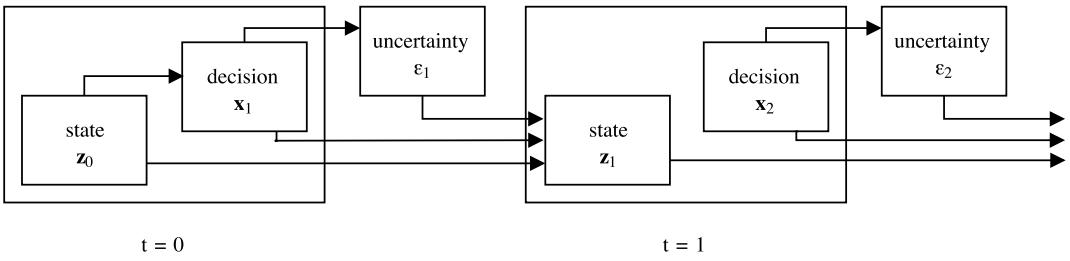
Example 3.2 (Continuation of Example 3.1)**.** We choose to use bivariate normal distribution as the sampling distribution. We set the initial parameters for the distribution as $\mu = (1, 1)^\top$ and $\sigma^2 = 5 \times \mathbb{I}_2$, where $\mathbb{I}_2$ is 2×2 unit matrix. The best performances in each iteration are selected using median, i.e. $\varrho = 0.5$.

The average distance of the estimated optimum from the actual optimum was 0.036 (confidence interval: 0.033–0.038). The mean number of iterations was 28.2. Some paths of the algorithm and all the estimated optima are visualized in Figure 3.3.

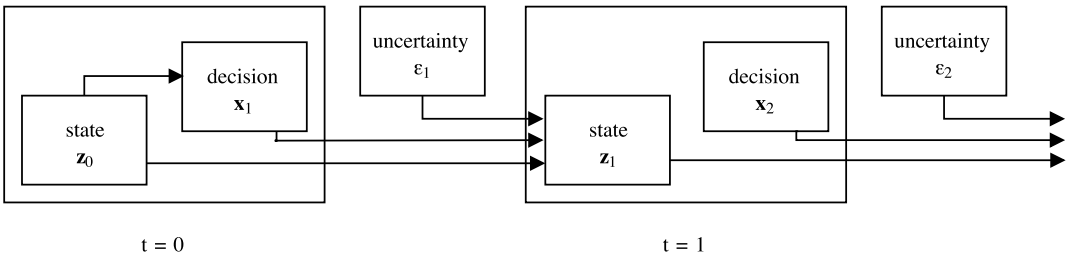
The cross entropy algorithm provided better results in only a few more iterations. Additionally, this method is computationally less demanding, because we do not need to estimate regression model at each iteration and we do not need to keep all the generated data.

△

Another possibility is to combine both described methods. One could use the cross entropy method to find a solution that will be used as the initial point for the response surface method. By doing this, there are already generated data before the first iteration of the response surface algorithm. Also, the initial point might be closer to the actual solution than the initial guess would be.



(a) Endogenous uncertainty.



(b) Exogenous uncertainty.

Figure 3.4: Comparison between the dynamics of the state-space decision model with endogenous and exogenous uncertainty. Adapted from Pflug and Pichler [2014].

3.3 Multistage Optimization

Until now, we considered only single-stage problems with one (d-dimensional) decision variable. In this chapter, we introduce a method for solving problems with several points in time when the decision has to be done. However, we still consider problems that need to be solved using one of the simulation method. We also need to add one variable to the model to track the history of the realized uncertainty. We considered general objective function g for the simple problems. For the multistage problems we consider only mean function $g = \mu$ for the objective function.

For the M-stage stochastic problem we consider following assumptions. Let $\mathbf{x}_1, ..., \mathbf{x}_M \in \mathcal{X} \subset \mathbb{R}^d$ be the decision variables. Suppose that the constraints for the decision variable $\mathbf{x}_m$ depends on the very previous decision $\mathbf{x}_{m-1}$ for $m = 2, ..., M$. Denote the largest set that fulfills such constraints by $\mathcal{X}(\mathbf{x}_{m-1})$. Let $\varepsilon_1, ...\varepsilon_M$ be the sequence of random variables that we call the uncertainty of the problem. Suppose the uncertainty depends on the last decision. Then we talk about the *endogenous uncertainty*. Let $\mathbf{z}_0, ..., \mathbf{z}_M \in \mathcal{S}$ be the random variables identifying the state of the decision model. Suppose the number of states is at most countable. The states depend on the very previous state, the last decision and the realization of the uncertainty. One can write $\mathbf{z}_m = \mathbf{z}_m(\mathbf{z}_{m-1}, \mathbf{x}_m, \varepsilon_m)$ for $m = 1, ..., M$. Since the state has a random variable as an argument it is random itself as well and it is natural to talk about its distribution. The state $\mathbf{z}_0$ is known (deterministic) prior to the first decision. The dynamics of the model is illustrated in Figure 3.4 (a).

Let the objective random variables be denoted by $Y_m = h(\mathbf{z}_m, \mathbf{z}_{m-1}, \mathbf{x}_m)$ for $m = 1, ..., M$, where $h : \mathbb{R}^{2c+d} \to \mathbb{R}$ is a measurable function. The variable Y_m is

usually called a *transition reward*. Denote the expected reward in the m-th time interval by $g_m(\mathbf{x}_m, \mathbf{z}_{m-1}) = \mathbb{E}\left[Y_m | \mathbf{x}_m, \mathbf{z}_{m-1}\right]$. Let the multistage problem be

$$\begin{aligned} G_0 = \max_{\mathbf{x}_1,...,\mathbf{x}_M} \quad & \sum_{m=1}^{M} g_m(\mathbf{x}_m, \mathbf{z}_{m-1}) \\ \text{s.t.} \quad & \mathbf{x}_1 \in \mathcal{X}, \\ & \mathbf{x}_m \in \mathcal{X}(\mathbf{x}_{m-1}), \qquad m = 2, ..., M. \end{aligned} \tag{3.5}$$

Similarly, we can write recursive problems for $m = 1, ..., M-1$

$$\begin{aligned} G_m(\mathbf{x}_m, \mathbf{z}_m) = \max_{\mathbf{x}_{m+1},...,\mathbf{x}_M} \quad & \sum_{k=m+1}^{M} g_k(\mathbf{x}_k, \mathbf{z}_{k-1}) \\ \text{s.t.} \quad & \mathbf{x}_k \in \mathcal{X}(\mathbf{x}_{k-1}), \qquad k = m, ..., M. \end{aligned} \tag{3.6}$$

Note the state $\mathbf{z}_m$ is unknown (depends on unobserved the uncertainty ε_m) at time m and, therefore, is random. Since the problem (3.6) depends on $\mathbf{z}_m$ it is random as well. Pflug and Pichler [2014] showed that the problem (3.5) is equivalent to

$$\begin{aligned} \max_{\mathbf{x}_1,...,\mathbf{x}_M} \quad & g_1(\mathbf{x}_1, \mathbf{z}_0) + \mathbb{E}\left[G_1(\mathbf{x}_1, \mathbf{z}_1)\right] \\ \text{s.t.} \quad & \mathbf{x}_0 \in \mathcal{X}. \end{aligned} \tag{3.7}$$

Suppose that we are able to evaluate (generate random variable from such distribution) the function $G_1(\mathbf{x}_1, \mathbf{z}_1)$. Then we can use some of the method described in Chapters 3.1 and 3.2 to find the optimum of (3.7).

Evaluation of $Q_1(\mathbf{x}_0, \mathbf{z}_1)$ consists of two parts: generating the uncertainty ε_1 and solving the stochastic problem with $M-1$ stages and known initial state $\mathbf{z}_1(\varepsilon_1)$. If $M = 2$ then the problem is easy to solve. For larger problems we need to iterate through the procedure until the problem at the end is single-stage.

Part II

Practical Applications

Chapter 4

Stochastic Demand

In this chapter, we introduce a model for stochastic demand for a service with limited inventory and clearly bounded time when the service may be sold. We also propose a method for estimating the parameters of the demand. The model is illustrated on example of train tickets.

Suppose that the demand follows Poisson distribution with intensity $\Lambda(p)$, where p is the price. Denote the period when the service is being sold by $[0, T]$. Assume there exists positive intensity function $\lambda(p, t)$ that identifies the change of the demand during time. It has to hold that $\Lambda(p) = \int_0^T \lambda(p, t)dt$. In this example we suppose the demand intensity has the elasticity constant in price and linear in time. Constant elasticity functions are widely used in microeconomics. Williams [2013] investigated the data from airline ticket prices and concluded that passengers are more likely to accept high prices as the departure time encloses. Function with mentioned attributes may be

$$\lambda(p, t; \boldsymbol{\beta}) = \exp(\beta_1 + \beta_2 \log(p) + \beta_3 t + \beta_4 t \log(p)), \tag{4.1}$$

for which the price elasticity is

$$\frac{\partial \lambda(p, t; \boldsymbol{\beta})}{\partial p} \frac{p}{\lambda(p, t)} = \beta_2 + \beta_4 t.$$

Note that the intensity lacks any periodicity (daily, weekly, etc.). This model is, however, able to capture the trend in demand intensity which is the most important part. The imperfection of the model is balanced by computational speed. This will become more clear in Chapter 5 where we need the inverse function of the cumulative intensity to simulate the demand.

This model for demand may be used for the tickets for each route (the pair of boarding and exiting stations). It is possible to assume that some of the parameters may be the same for each route or some group of routes, e.g. one can assume that the elasticity is the same for routes with similar route length. We do not make any assumptions about the relations between parameters of individual routes. Further, we assume that the demands for individual routes are mutually independent.

This market for train tickets is specific in following ways:

i) The tickets can be sold only up to given time (departure of the train). This time is different for each boarding station.

ii) The inventory level (number of seats available) is set before the tickets are purchased and cannot be changed dynamically.

iii) There exists a rivalry between customers even if they are purchasing different service. Such case may happen when two passengers have no station (nor boarding or exiting) in common, but their journeys have an intersection. These passengers cannot buy the ticket for the same seat in the train. This implies that the number of seats available is random for each par of boarding and exiting station. The reservation system may be managed in one of the following ways.

a) The passenger may choose any of the seats that is available for whole his route. If there is no such seat, the ticket is unavailable for the route.

b) The system selects the seat that would provide the best seating availability for upcoming passengers and sends the seat number to the passenger at the time of the purchase.

c) The system sends the seat number to the passenger right before the train departures (e.g. at the time the purchasing period is over).

The model we introduce in this thesis is applicable to system c). System b) requires additional combinatorial optimization for seat selection. System a) may result in significant revenue losses due to ineffectiveness in allocation of seats to passengers.

iv) The tickets may be in some cases returned to the seller (depends on the conditions of individual carrier). The customer has to pay a ratio of original price as a cancellation fee, say $r \in [0, 1]$. In our example, the ticket cannot be returned, i.e. $r = 1$.

v) The service is provided in classes that differ in price and number of supplementary services. The demand for each class should be modeled individually. The price elasticity is usually lower (higher in absolute value) for higher classes. We suppose only one class in the train.

vi) There are several passenger types with different demands. In some countries, the carrier is obliged[1] to make a discount for some of the passenger types (e.g. students, seniors, handicapped person, etc.). We suppose only one passenger type.

Suppose the train goes through K stations (including the first and the last one) and M seats. Denote these stations by $\{1, ..., K\}$. Since the arrivals of passengers for each route follow inhomogeneous Poisson process and the upper bound for the number of passengers depends on the demand for tickets for different routes it is suitable to model number of passengers for all routes with an inhomogeneous Markov process. The states identify number of tickets sold for each route, i.e. the state space $\mathcal{S}$ is $\binom{K}{2}$-dimensional with values in $\mathbb{N}_0$. Even though it is possible to index the dimensions of the states with only one number (e.g. s_i would be the number of passengers for i-th route for state $\boldsymbol{s} \in \mathcal{S}$, where $i \in \{1, ..., \binom{K}{2}\}$) we

[1]For example based on Assessment Notice 01/2016 of Ministry of Finance of the Czech Republic.

index the dimensions with two numbers denoting the boarding and exiting station (e.g. $s_{k,l}$ is the number of passengers traveling on the route (k,l) from station k to station l, where $k,l \in \{1,...,K\}$ and $l > k$). The initial state denoting empty train is $\mathbf{0} = (0,...,0)$. Similarly, we denote the prices as a $\binom{K}{2}$-dimensional vector $\boldsymbol{p}$ with indexes $\{(k,l) : 1 \leq k < l \leq K\}$.

There is a capacity condition that every state of the state space must fulfill, i.e.

$$\mathcal{S} = \left\{ \boldsymbol{s} \in \mathbb{N}_0^{\binom{K}{2}} : \sum_{k=1}^{h} \sum_{l=h+1}^{K} s_{k,l} \leq M, \forall h = 1,...,K-1 \right\}$$

Denote by $\omega(\boldsymbol{s},k,l)$ the state which the system transitions into after arrival of a passenger for route (k,l) when the system was in state $\boldsymbol{s} \in \mathcal{S}$. It is clear that

$$\omega(\boldsymbol{s},k,l) = \boldsymbol{s} + \omega(0,k,l).$$

The state $\omega(\boldsymbol{s},k,l)$ is an element of state space $\mathcal{S}$ if and only if there is an empty seat for route (k,l) is state $\boldsymbol{s} \in \mathcal{S}$. The transition intensity from state $\boldsymbol{s} \in \mathcal{S}$ to state $\omega(\boldsymbol{s},k,l) \in \mathcal{S}$ is equal to $\lambda_{k,l}$, i.e. the demand intensity for route (k,l). The total transition intensity from state $\boldsymbol{s} \in \mathcal{S}$ at time $t \in [0,T]$ with prices $\boldsymbol{p}$ is given by

$$q_{\boldsymbol{s}}(\boldsymbol{p},t;\boldsymbol{\beta}) = \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} \lambda_{k,l}(p_{k,l},t,;\boldsymbol{\beta}_{k,l}) \times \mathbb{I}[\omega(\boldsymbol{s},k,l) \in \mathcal{S}],$$

where $\boldsymbol{\beta}_{k,l} = (\beta_{1,k,l},...,\beta_{4,k,l})^\top$ denotes the parameters that matter for route (k,l) and

$$\boldsymbol{\beta} = \{\beta_{h,k,l} : 1 \leq h \leq 4, 1 \leq k < l \leq K\}$$

denotes the aggregated parameter of the demand.

Suppose we have the data from selling system from n trains with equal demands (e.g. trains at the same time and day in week in multiple consequence weeks). Denote by N^ν the number of passengers in ν-th train. Use following notation for the observed data about i-th passenger in ν-th train:

T_i^ν	purchase time,
p_i^ν	price of the ticket,
$(\mathrm{k}_i^\nu, l_i^\nu)$	route (boarding station, exiting station).

Suppose that the price was constant over J time intervals that were the same for each train. Denote by $[\tau_{j-1},\tau_j]$ the interval over which the price $\boldsymbol{p}^{(j)}$ was effective for $j = 1,...,J$. Denote by $\hat{\tau}_{k,l}^\nu$ the time the tickets for route (k,l) on train ν became unavailable due to the capacity limitations or due to the departure of the train.

Let us denote a set of passengers that travel route (k,l) for each train by

$$\mathcal{I}_{k,l}^\nu = \{i \in \{1,...,N^\nu\} : k_i^\nu = k, l_i^\nu = l\}.$$

The elements of the score function (2.4) that matter for route (k,l) can be written

as

$$\boldsymbol{U}_{n,(k,l)}(\boldsymbol{\beta}) = \sum_{\nu=1}^{n} \left[\sum_{i \in \mathcal{I}_{k,l}^{\nu}} \begin{pmatrix} 1 \\ \log(p_i^{\nu}) \\ T_i^{\nu} \\ T_i^{\nu} \log(p_i^{\nu}) \end{pmatrix} - \sum_{j=1}^{J} \int_{\tau_{j-1}}^{\tau_j} \boldsymbol{V}_{k,l}^{(j)}(t)\ \lambda_{k,l}(p_{k,l}^{(j)}, t; \boldsymbol{\beta}_{k,l})\ \mathbb{I}[t < \hat{\tau}_{k,l}^{\nu}] dt \right],$$

where

$$\boldsymbol{V}_{(k,l)}^{(j)}(t) = \begin{pmatrix} 1 \\ \log(p_{k,l}^{(j)}) \\ t \\ t \log(p_{k,l}^{(j)}) \end{pmatrix}.$$

The observed information matrix (2.5) is block diagonal. The block that matters for route (k, l) is

$$\boldsymbol{I}_{n,(k,l)}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{\nu=1}^{n} \left[\sum_{j=1}^{J} \int_{\tau_{j-1}}^{\tau_j} \left(\boldsymbol{V}_{k,l}^{(j)}(t) \right) \left(\boldsymbol{V}_{k,l}^{(j)}(t) \right)^{\top} \lambda_{k,l}(p_{k,l}^{(j)}, t; \boldsymbol{\beta}_{k,l})\ \mathbb{I}[t < \hat{\tau}_{k,l}^{\nu}] dt \right].$$

The estimate $\hat{\boldsymbol{\beta}}_n$ needs to be found using Newton-Raphson method described in Chapter 2.

Note that even though the variance matrix of the estimate $\hat{\boldsymbol{\beta}}_n$ is block diagonal the elements of the estimate that matter for different routes are still dependent. The seat availability determined by passengers on all routes influences the variance of the estimate.

The method was used for an analysis of simulated data of processes of selling the tickets for 500 trains. The demand was both simulated and estimated using model (4.1). The parameters for demand intensity for each route individually are in Table 4.1. The tickets began to sell 60 days before the departure, say at time 0. For computational purposes we assume the time period of 60 days as $[0, 1]$. The train departed (and the tickets ended sells) from stations $1, ..., 5$ at times $\frac{59.4}{60}$, $\frac{59.44}{60}$, $\frac{59.48}{60}$, $\frac{59.52}{60}$, $\frac{59.56}{60}$, and $\frac{59.6}{60}$. Prices that we used for the simulations are in Table 4.1. The times of price changes were at the end of the 30th, 46th, 54th, and 58th day.

Distribution of times of purchases is showed in Figure 4.1. There are significant drops in demand right after the price has increased. Also at the end of the selling period the number of tickets sold decreases because the train departed from its begging station and some of the tickets cannot be sold anymore.

The number of simulated trains is clearly way higher that the railways company may be using for modeling the demand of single train. The assumption that the demand is the same for almost ten years is most certainly wrong. On the other hand, the demand for specific train might be similar to other trains departing in similar times or the same type of day in a week (workday or weekend) or the train in opposite direction. If the company dispatches for example 16 trains a day in each direction the total number of trains in a week is 224. That is why the number of simulated trains may be reasonable.

route	β_1	β_2	β_3	β_4	$\boldsymbol{p}^{(1)}$	$\boldsymbol{p}^{(2)}$	$\boldsymbol{p}^{(3)}$	$\boldsymbol{p}^{(4)}$	$\boldsymbol{p}^{(5)}$
(1,2)	17.92	-4.19	3.77	1.17	86	99	114	131	151
(1,3)	17.17	-4.04	4.20	1.69	157	181	208	239	275
(1,4)	23.36	-4.20	3.52	0.77	223	256	295	339	390
(1,5)	18.54	-3.52	4.42	1.11	286	329	378	435	500
(1,6)	18.28	-3.62	4.74	1.33	347	399	458	527	606
(2,3)	14.54	-3.38	4.02	1.28	86	99	114	131	151
(2,4)	20.58	-4.27	3.94	1.41	157	181	208	239	275
(2,5)	22.10	-4.34	3.55	1.20	223	256	295	339	390
(2,6)	26.05	-5.02	3.77	1.22	286	329	378	435	500
(3,4)	18.29	-3.69	4.32	0.55	86	99	114	131	151
(3,5)	17.35	-3.43	4.18	0.95	157	181	208	239	275
(3,6)	21.56	-4.19	3.58	1.22	223	256	295	339	390
(4,5)	21.19	-4.54	3.19	1.07	86	99	114	131	151
(4,6)	16.76	-3.43	4.60	0.89	157	181	208	239	275
(5,6)	18.24	-4.54	4.07	1.37	86	99	114	131	151

Table 4.1: Actual parameters of the demand and prices for individual time periods used in simulation of demand for train tickets.

However, high correlation of regressors implies high correlation and variance of all estimates. Theoretical Fisher information matrix (with assumption of unlimited train capacity) indicates that the estimated values might not have meaningful values. In many cases the standard deviation is greater than the expected value (actual value of the parameter) so that the sign of the estimate has different value in many cases. However, the Newton-Raphson method still converges and we are able to calculate the estimates. Their values are meaningless because of their standard errors. For example, the variance matrix for estimates that matters for route between stations 1 and 2 is

$$\boldsymbol{I}_{(1,2)} = \begin{pmatrix} 124.41 & -26.15 & -108.82 & 23.27 \\ -26.15 & 5.50 & 22.78 & -4.88 \\ -108.82 & 22.78 & 99.31 & -21.06 \\ 23.27 & -4.88 & -21.06 & 4.47 \end{pmatrix}.$$

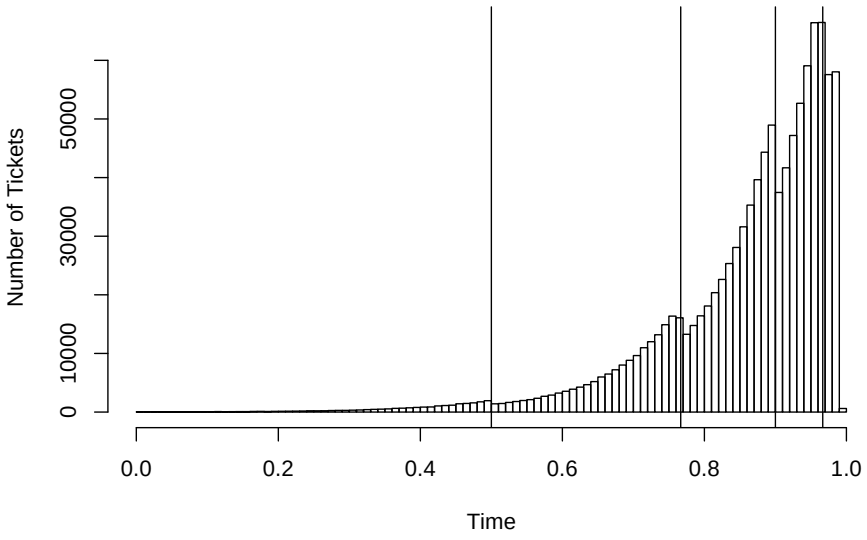


Figure 4.1: Histogram of times of purchase over all simulated trains. Vertical lines indicate the times of price changes.

Chapter 5

Optimal Price of Fare

We use the demand intensity from previous chapter to find an optimal price of fare. First, we solve the single-stage problem to find the optimal price for each route that would apply for the whole selling period. Then we add the possibility to change the initial decision on the price during the selling period. The decision can be done only at certain moments, *decision times*. The decision times are chosen before the selling period and cannot be change later.

The objective function to which the price is optimized is expected return, that is the sum of prices of all sold tickets. We use the same state space $\mathcal{S}$ as we did in Chapter 4. The objective variable for each stage of the multistage problem is

$$Y_m = h(\mathbf{z}_m, \mathbf{z}_{m-1}, \boldsymbol{p}_m) = \boldsymbol{p}_m^\top(\mathbf{z}_m - \mathbf{z}_{m-1}),$$

where $\boldsymbol{p}_m$ is the decision variable for m-th stage, i.e. the prices for all stations for m-th period. The randomness of Y_m is caused via $\mathbf{z}_m$ (number of tickets sold at the end of m-th period) which depends on the uncertainty factor ε_m (i.e. the randomness in demand).

The set of general feasible solutions $\mathcal{X}$ that must hold for prices between each decisions is given by following constraints:

i) If one route is part of another route the longer route must not have cheaper tickets, i.e.

$$p_{k,l} \leq p_{k',l'}, \qquad k' \leq k < l \leq l'. \tag{5.1}$$

ii) A route must not have more expensive ticket than sum of ticket prices of routes that combine to the original route, i.e.

$$p_{k,l} \leq p_{k,h} + p_{h,l}, \qquad k < h < l. \tag{5.2}$$

iii) All tickets must have positive price. Due to condition i it is sufficient to have only constraints on the routes between adjacent stations, i.e.

$$p_{k,k+1} > 0, \qquad k = 1, ..., K-1.$$

Since $\lambda(p, t; \boldsymbol{\beta}) \to \infty$ as $p \to 0$ for reasonable values of $\boldsymbol{\beta}$ ($\beta_2 + \beta_4 t < 0$) and the demand is simulated as Poisson process with such intensity, it is reasonable to forbid prices close to zero. Because of that we set constraints to

$$\boldsymbol{p} \geq \boldsymbol{p}_0, \tag{5.3}$$

where $\boldsymbol{p}_0 > 0$ is $\binom{K}{2}$-dimensional constant (in our case $\theta = (50, ..., 50)^\top$). Constraints (5.1) and (5.2) are replaced by linear penalty function

$$\Phi_1(\boldsymbol{p}) = \sum_{k' \le k < l \le l'} [p_{k',l'} - p_{k,l}]_+ + \sum_{k<h<l} [p_{k,h} + p_{h,l} - p_{k,l}]_+$$

where $[\,\cdot\,]_+ = \max\{\,\cdot\,, 0\}$ denotes the positive part and $\mathbf{1}$ is vector of ones with respective length.

The specific conditions for m-th decision variables is given by

$$\mathcal{X}(\boldsymbol{p}_{m-1}) = \{\boldsymbol{p}_m \in \mathcal{X} : \boldsymbol{p}_m \ge \boldsymbol{p}_{m-1}\}, \qquad m = 2, ..., M \tag{5.4}$$

so the tickets are at least that expensive as they were after previous decision. The penalty function for constraints (5.3) and (5.4) is

$$\Phi_2(\boldsymbol{p}_m) = \mathbf{1}^\top [\boldsymbol{p}_{m-1} - \boldsymbol{p}_m]_+.$$

We can write the maximization problem using penalty functions as

$$\begin{aligned} \max_{\boldsymbol{p}_1, ..., \boldsymbol{p}_M} \quad & \mathbb{E}\left[\boldsymbol{p}_m^\top (\mathbf{z}_m - \mathbf{z}_{m-1})\right] - \theta \left[\sum_{m=1}^{M} \Phi_1(\boldsymbol{p}_m) + \sum_{m=2}^{M} \Phi_2(\boldsymbol{p}_m)\right] \\ \text{s.t.} \quad & \boldsymbol{p}_m \in \mathbb{R}^{\binom{K}{2}}, \qquad m = 1, ..., M \end{aligned}$$

where $\theta > 0$ is a *penalty coefficient*. In our case, we set $\theta = 1000$.

5.1 Single-stage model

Let us begin with the single-stage problem, that is $M = 1$ and the selected prices hold during whole selling period. The problem is solved in two phases. In the first phase, we use the cross entropy algorithm to find a solution that is close to the actual optimum. The algorithm is useful to begin with if we have no good guess where the solution is because it searches through the set of feasible solutions globally and narrows this set. On the other hand, it does not have to converge to the actual optimum. We used normal distribution as the sampling distribution. The initial mean is as in Table 5.1 (a) and the initial variance matrix is diagonal with diagonal elements $\sigma^2_{(k,l),(k,l)} = 500$. The algorithm was iterated 20-times and during each iteration 2000 new data point were generated. New sampling distribution was estimated using 25% best performing observations. The result of the algorithm is in Table 5.1 (b).

In the second phase, we used the response surface algorithm on the data generated in the first phase. In addition, we generated 500 new data during each iteration of the algorithm. The prices were generated from multivariate (mutually independent) normal distribution with mean equal to current approximation of the optimum price and standard deviation 10. The derivative of the objective function was estimated locally using 30% of the observations. The new approximation of the price were calculated by

$$\boldsymbol{p}^{(r+1)} = \boldsymbol{p}^{(r)} + \frac{\widehat{\nabla}(\boldsymbol{p}^{(r)})}{20}.$$

	2	3	4	5	6
1	250.0	420.5	569.9	707.1	835.9
2		250.0	420.5	569.9	707.1
3			250.0	420.5	569.9
4				250.0	420.5
5					250.0

(a) Initial prices.

	2	3	4	5	6
1	317.9	419.7	478.0	624.2	794.3
2		187.8	208.4	496.1	671.3
3			176.6	418.2	591.6
4				320.1	501.2
5					283.9

(b) Cross Entropy.

	2	3	4	5	6
1	231.8	344.3	363.2	454.0	476.3
2		317.3	312.2	444.0	442.7
3			293.8	398.5	423.4
4				124.1	316.3
5					225.3

(c) Response Surface.

	2	3	4	5	6
1	224.6	336.9	351.7	447.2	478.6
2		308.4	307.5	443.6	439.7
3			275.9	398.9	413.1
4				127.2	306.2
5					221.6

(d) Dynamic pricing.

Table 5.1: Comparison of prices between the optimization phases. Rows stand for boarding stations and columns stand for exiting stations.

Results for this method are in Table 5.1 (c). Note that the solution does not fulfill the constraints (5.1). This solution may be justified by projection into the set of feasible solutions.

We compared the results from both phases to verify that the second phase actually improved the solution. The expected return with price given from the first phase is 557,481 (556,505–558,457). With the prices obtained after the second phase the expected return is 732,923 (731,893–733,953). Both these results were estimated using 1,000 observations.

To approximate the solution of the optimization problem we had to simulate the demand 140,000-times in both phases combined. This took more than two hours on regular PC.

5.2 Dynamic model

Suppose that we are given a sequence of decision times $0 = t_0 < t_1 < ... < t_M = 1$ (t_M is actually not a decision time since it is at the time of the departure of the train). We want to find an optimal price $\boldsymbol{p}_m$ for time interval $[t_{m-1}, t_m)$ conditionally on state $\mathbf{z}_{m-1}$ for $m = 1, ..., M$. We want to use the method described in Chapter 3.3. However, there is a problem with computational complexity.

We did 300 iterations with 1,000 simulated demands in the static (single-stage) model. Each simulation of the demand consists of $\binom{K}{2}$ simulations of inhomogeneous Poisson process with hundreds of arrival times. To do the same in the dynamic model we would need to repeat this procedure for each time interval and each state that we arrived into in one of the simulations of previous interval. To have the same precision as in the static model we would need to simulate the demand $300,000^M$-times. That is far beyond the limits of computing capability even for small M. We will discuss two methods that would simplify the problem but they would produce only suboptimal solutions. On the other hand, the simulated optimization only provides an approximation of the optimal

solution so the simplifications should not deteriorate the estimated solution too much.

The easiest way to solve this is to solve a single-stage problem at each decision time. The prices are optimizes as if they were set up until the train departure. This prices hold up to following decision when they are changed again as if the were set up until the train departure. The prices for the first time interval are the same as in Table 5.1.

The other way is to solve only two-stage problem in each decision time (except the last decision time t_{M-1}). The m-th decision is made as if the prices hold until the very following decision time and at that time they were changed until the departure of the train. The initial point for the prices in the first time interval in the dynamic model are taken from the result of the single-stage problem. In each iteration the seat occupancy is generated along with return over the first interval. Then the optimal price till the departure of the train is found. Expected return in this price for the rest of the selling period is estimated. Finally, the returns from both time intervals are added up to generate the outcome for given price.

Because the two-stage problem is much more complex problem for simulated optimization we reduced the number of simulations. First, we did 3 iterations of Cross Entropy algorithm with 100 iterations each. The response surface method with 20 iterations by 20 simulations followed. This means we made 490,000 simulations but the precision decreased because of less number of simulations

The estimated optimum for the dynamic model is in Table 5.1 (d). Note that the optimum prices are in general less than the optimal prices of the single-stage problem. This is caused by the constraint (5.4).

5.3 Reformulation with exogenous uncertainty

The methods for simulated optimization are generally time demanding especially for multistage optimization. The reason is that we need to simulate the output every time we want to change the decision variable. If we were able to simulate the randomness (*scenario*) only once and then recalculate the output every time we want to change the decision variable.

Until now we simulated the process of selling tickets for given prices. Another way is to simulate all potential passengers and their maximum acceptable price and eventual time of purchase. This allows us to rewrite the problem into quadratic programming with following inputs from simulated scenario:

N	number of potential passengers,
$\pi_1, ..., \pi_N$	maximum acceptable prices,
$\tau_1, ..., \tau_N$	eventual times of purchase,
$(\mathrm{k}_1, l_1), ..., (k_N, l_N)$	boarding and exiting stations.

While simulating such scenario we need to keep the intensity as in (4.1). We can achieve this by setting low[1] price p_0 and simulate the process of demand as if it was the desired price, i.e. simulate Poisson process with intensities $\lambda(p_0, t; \boldsymbol{\beta})$

[1]The price p_0 needs to be low enough to be sure that is is less than the optimum. On the other hand, too low price implies large number of potential passengers, which increases computational complexity for both the simulations and consecutive optimization.

for every pair of stations without any capacity constraint. Than we simulate the maximum acceptable price for each potential passenger from Pareto distribution with shape $-\beta_2 - \beta_4 t$ and scale p_0. Its probability distribution function is

$$\mathbb{P}[\pi_i \leq x] = 1 - \left(\frac{x}{p_0}\right)^{\beta_2+\beta_4 t_i}, \qquad x \geq p_0.$$

The process intensity and required distribution of total return is ensured by Proposition 1.15.

Denote by $\boldsymbol{R} \in \mathbb{R}^{N\times(K-1)}$ and $\boldsymbol{S} \in \mathbb{R}^{N\times\binom{K}{2}}$ matrices indicating the occupancies and routes for each potentional passenger, i.e.

$$\boldsymbol{R}_{i,k} = \begin{cases} 1, & k_i \leq k \wedge k+1 \leq l_i \\ 0, & \text{else,} \end{cases}$$

$$\boldsymbol{S}_{i,(k,l)} = \begin{cases} 1, & k_i = k \wedge l_i = l \\ 0, & \text{else.} \end{cases}$$

The quadratic programming problem is in form of

$$\begin{aligned}
\max_{\boldsymbol{p},\boldsymbol{y},\boldsymbol{z}} \quad & (\boldsymbol{p}^\top, \boldsymbol{y}^\top, \boldsymbol{z}^\top)\boldsymbol{D}(\boldsymbol{p}^\top, \boldsymbol{y}^\top, \boldsymbol{z}^\top)^\top \\
\text{s.t.} \quad & \boldsymbol{p} \in \mathbb{R}^{\binom{K}{2}} \\
& \boldsymbol{y} \in \{0,1\}^N \\
& \boldsymbol{z} \in \{0,1\}^{N\times(K-1)} \\
& \boldsymbol{R}^\top \boldsymbol{y} \leq (M, ..., M)^\top \\
& y_i = 0 \vee \boldsymbol{S}_{i*}\boldsymbol{p} \leq \pi_i, \qquad i = 1, ..., N \\
& y_i \leq y_j \vee \boldsymbol{S}_{i*}\boldsymbol{p} > \pi_i \vee \sum_{k=1}^{K-1}(1 - z_{(i,k)}) \geq 1, \qquad i < j \\
& z_{(i,k)} = 1 \vee (\boldsymbol{R}_{i,k} = 1 \wedge \sum_{j=1}^{i-1} \boldsymbol{R}_{j,k} y_j \leq M), \\
& \qquad\qquad (i,k) = (1,1), ..., (N, K-1)
\end{aligned}$$

where $\boldsymbol{S}_{i*}$ denotes i-th row of matrix $\boldsymbol{S}$ and

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{S}^\top & \boldsymbol{0} \\ \boldsymbol{S} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}$$

is block matrix with block sizes equivalent to sizes of vectors $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$. All the constraints above are easily transferable into linear constraints. An issue is that the matrix $\boldsymbol{D}$ is indefinite and standard solvers cannot solve this problem.

Conclusion

We introduced the theory of inhomogeneous Markov process and specifically of inhomogeneous Poisson process which is a special case. The theory generalizes the properties of homogeneous Markov process that is generally well known and described in many literature. We paid special attention to processes with separable inhomogeneity and proved that it is necessary and sufficient condition for transformability to homogeneous Markov process. The separable inhomogeneity also implies that the process has embedded chain that has Markov property and is homogeneous.

The intensity of Markov process can be estimated using maximum likelihood theory. We explained in detail how to calculate the estimates for different types of the process. First, we derived the method for homogeneous Markov process. The we did the same for separated inhomogeneity and for constant rate matrix – both separately and together. Finally we described a model for general inhomogeneous Markov process that could be also dependent on any exogenous variable. Every section was accompanied by en example. This theory can be also used for testing, e.g. testing of homogeneity of Poisson process, as it was shown in an example.

The two algorithms – response surface and cross entropy – was introduced as methods for solving optimization problems for which the objective function is unknown and observable only with an error. Both methods require simulations of objective functions. In a simple example we saw that cross entropy algorithm converges faster to actual optimum. Both these methods were intended to be used in multistage optimization as well. However, there is a lot of literature about multistage optimization that it was not necessary to describe it along with simulated optimization in this thesis.

Both statistics and optimization from Chapters 2 and 3 were implemented in a complex example of train ticket fare optimization. The occupancy of the train is interpreted as inhomogeneous Markov process and return from sold tickets is maximized by controlling the intensity via fare price. The only input into the model are historical data of selling the tickets. This means that the intensity was about to be estimated. However, we found out that the variability of the estimates of the intensity parameters is too high to be any useful for any reasonable size of the data.

The optimization of fare was done using actual intensity because it was impossible to estimate it from generated data. The Markov process representing sold tickets has too many states to be able to calculate transition probabilities from intensity matrix. Therefore the optimization requires simulating the demand and estimating the objective function from the simulated returns. That implies that both the number of iterations of the algorithm and the number of simulations in each iteration must be limited. This is even bigger issue when calculating multi-

stage optimization because the number of simulations grows exponentially with number of stages. Therefore we managed to approximate the optimum for only two-stage problem. The number of iterations was still insufficient.

One simplification is to reformulate the problem with exogenous uncertainty. We proved that it is possible to simulate the randomness even before the price is known and then calculate the return using this scenario (or multiple scenarios). This allows us to reformulate the problem into quadratic programming but with indefinite matrix in objective function.

Bibliography

Jiří Anděl. *Základy matematické statistiky*. Third edition. Matfyzpress, Praha, 2011.

Kai Lai Chung. *Markov Chains With Stationary Transition Probabilities*. Springer Verlag, New York, 1967.

William S. Cleveland. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

William S. Cleveland and Susan J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

Mark H. A. Davis. *Markov Models & Optimization*. Monographs on Statistics & Applied Probability. Taylor & Francis, 1993.

Joseph Leo Doob. *Stochastic processes*. Wiley publications in statistics. Wiley, 1990.

Jitka Dupačová. Optimization under exogenous and endogenous uncertainty. 2006.

Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applicatons*. Monographs on Statistics & Applied Probability. Chapman & Hall/CRC, 2003.

Vikas Goel and Ignacio E. Grossmann. A class of stochastic programs with decision dependent uncertainty. *Mathematical Programming*, 108:355–394, 2006.

Dirk P. Kroese, Thomas Taimre, and Zdravko I. Botev. *Handbook of Monte Carlo methods*. Wiley series in probability and statistics. Hoboken, N.J. Wiley, 2011.

P. A. W Lewis and G. S. Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413, 1979.

Petr Mandl. *Pravděpodobnostní dynamické modely*. First edition. Academia, Praha, 1985.

Georg Ch. Pflug and Alois Pichler. *Multistage Stochastic Optimization*. First edition. Springer, New York, 2014.

Zuzana Prášková and Petr Lachout. *Základy náhodných procesů I.* Second edition. Matfyzpress, Praha, 2012.

Josef Štěpán. *Teorie pravděpodobnosti.* Academia, Praha, 1987.

Kevin R. Williams. Dynamic airline pricing and seat availability. 2013.

List of Figures

List of Tables