university of groningen

faculty of science
and engineering

mathematics and applied
mathematics

# Optimization of the Bayesian Poisson Changepoint Model

Bachelor's Project Mathematics

July 2017

Student: D.M. Heeg

First supervisor: dr. M.A. Grzegorczyk

Second assessor: dr. W.P. Krijnen

### Abstract

The Bayesian Poisson changepoint model, which is used to analyse non-homogeneous data sets, does not deal with over-dispersion and is therefore suboptimal to analyse real world taxi pick-up counts. In this thesis several adjustments to the model are suggested and tested and they aim for an optimization of the Bayesian Poisson changepoint model. The Poisson-Gamma model will be replaced by the Negative-Binomial-Beta model and information exchange on both global and sequential level will be implemented. Eventually the improved model is used to perform simulations on the taxi data set and an interpretation of the results is given.

# Preface

In this bachelor thesis an attempt will be made to optimize the Poisson changepoint model (CPS) in its Bayesian framework as used in a recent paper by Grzegorczyk and Kamalabad [1]. Grzegorczyk and Kamalabad performed a comparative evaluation study on popular non-homogeneous Poisson models for count data, among those model was the Poisson changepoint model. In their conclusion they stated that the Bayesian CPS might have been suboptimal during their simulations on certain data sets and they made a request for further research on the optimization of this model. This thesis will answer their request for further research and is intended for undergraduate students with a basic knowledge of Bayesian Statistics.

# Contents

# 1 Introduction

The field of statistical inference is divided into two main philosophical approaches. The first one is the *Bayesian* approach and the second is the *frequentist*, or sometimes referred to as *classical*, approach.

The divergence between the frequentist and Bayesian approach finds its origin in the two different interpretations of probabilities. Bayesians interpret probabilities as subjective statements about how likely it is that an event will occur. Frequentists on the other hand, interpret probabilities as long-run relative frequency of an event when an experiment is repeated many times. For this reason, frequentists find bayesians approaches objectionable, and sometimes even unacceptable. They themselves, do not make probabilistic statements about parameters.

From the 20th century Bayesian methods increased in popularity. Statisticians such as Finetti, Lindley and Wallace developed a complete method of Statistical inference, based on Bayes theorem [3]. In this method they stated that, since we are uncertain about the true value of the parameters, we might as well consider them to be random variables. Following this approach, Bayes' theorem is used in situations where $y$ represents the observed data and $x$ depicts the unknown parameter that is to be estimated. Hence, the rules of probability are directly used to learn about the parameter. Those probabilistic statements about parameters must be seen as *degrees of belief.* Everyone can make their personal beliefs about parameters. The degree of belief measures how likely someone considers a value of the parameter to be, before having observed the data. After obtaining and observing the data, those beliefs about the parameters are updated.

Heated debates still take place between classical statisticians and Bayesian statisticians. These debates are rather philosophical and there will probably never be a definitive answer to the question which approach is the best one. However, lots of studies have been dedicated to comparison of the two approaches, among which a recent study by Grzegorczyk and Kamalabad (2017).

Grzegorczyk and Kamalabad performed a comparative evaluation study on popular non-homogeneous Poisson models for count data. In their study they implemented the models in both Bayesian and frequentist framework and made a pairwise comparison between the four Bayesian and the four frequentist models. This pairwise comparison was made to see to which extent the results relying on the two frameworks differed. [1]

The study was performed on various Poisson synthetic data sets and on real-world taxi pick-up counts, extracted from the recently published New York City Taxi database. From the pairwise comparison it was concluded that the Poisson changepoint model was the only non-homogeneous model, out of three, that was superior to its frequentist counterpart. However, in the conclusion of their papaer, Grzezgorczyk and Kamalabad formulated an important note: potential over-dispersion was not taken into account within the presented study. Over-disperion means that the observed variance of the data is higher than the variance of the corresponding theoretical model (in this case the Poisson variance). When the data was actually sampled from the Poisson distributions, over-dispersion could not have arisen. However, in the case of the taxi pick-up counts, over-dispersion might very well occur. Therefore Grzegorczyk and Kamalabad concluded that the Bayesian Poisson changepoint model might have been suboptimal for the real-world data set and made a suggestion for further research to study possible improvement of the model. Based on his suggestion for further research, the main goal of this thesis is to improve the Poisson changepoint model in its Bayesian framework as used in the paper of Grzegorczyk and Kamalabad.

Before starting off with the ideas for improvement, first a short introduction in Bayesian statistics is given followed by an extension of all the mathematical models that will be used during this thesis. Then in chapter 4 the original Bayesian Poisson changepoint model is defined and explained and the improvements that might be made are discussed as well. Those improvements will form the next directions of this thesis. Chapter 5 will provide all the derivations that were necessary. In chapter 6 the first improvement to the model will be tested and chapter 7 the search for more improvements will continue. Chapter 8 will explain the concept of information exchange on global and sequential level and eventually simulations will be performed and the results are sketched in chapter 9. Then the model is applied to the real-world taxi pick-up counts and the results are analysed. Eventually the thesis will and with a conclusion where the most important findings during this thesis are summarized and suggestions for further research will be made in the discussion.

# 2 Short introduction in Bayesian statistics

Bayesian statistics is a mathematical procedure that applies probabilities to statistical problems. Probabilities are interpreted as subjective degrees of belief and the goal is to state and analyse beliefs. The most important and basic concepts of Bayesian statistics which will be used through this thesis will be outlined in this section.

In general the concept of Bayesian inference is quite simple. A certain event can occur and a mathematical model can be used as mathematical formulation of the observed events. A belief is constructed in the realization of such an event. The belief is updated after observing the data and the updated belief is implemented into the model.

The models that will be used to represent the observed events are *probability density functions* (pdf). The pdfs represent the likeliness of the data given a certain parameter, i.e. $P(X|\theta)$, which is referred to as the *likelihood*. Given a parameter $\theta$, a belief is constructed and the *prior* is the strength of this belief in the parameter, also denoted as $P(\theta)$.

The *marginal likelihood* (mll) can be derived with help of the likelihood and the prior. The mll represents the likeliness of the observed data:

$$\begin{aligned} P(X) &= \int_\theta P(X, \theta)\, \mathrm{d}\theta \\ &= \int_\theta P(X|\theta)P(\theta)\, \mathrm{d}\theta. \end{aligned}$$

The mll represents the likeliness of the data in general in contrast to the likelihood function, which shows the likeliness of the data given a fixed parameter.

For example, suppose a coin is flipped $n$ times and the outcomes are denoted $X = (x_1, x_2, ..., x_n)$. Where $x_i$ is equal to 0 (heads) or 1 (tails). The parameter $\theta$ represents the fairness of the coin. If $\theta = 0.5$ the coin is completely fair. The prior, $P(\theta)$, represents the belief in the fairness of the coin. If $P(\theta)$ is equal to 1 for $\theta = 0.5$ and 0 for all other $\theta$, we have absolute belief that the coin is fair. Usually, in Bayesian statistics, $P(\theta)$ is nonzero for multiple values of $\theta$. Instead of focusing on one optimal value of the parameter given the data, the distribution of the data is estimated using every value of the parameter and its belief.

In general, the Bayesian inference procedure is as follows:

1. A prior distribution $P(\theta)$ is chosen, expressing the beliefs about the parameter $\theta$.
2. A pdf $P(X|\theta)$ that reflects the likeliness of the data is chosen.
3. After observing the data $X$, the posterior distribution $P(\theta|X)$ is determined and update our beliefs.

The *posterior belief* $P(\theta|X)$ can be used to update the prior and is proportional to the likelihood times the prior:

$$P(\theta|X) \propto P(X|\theta) \cdot P(\theta).$$

If the prior and the posterior belief are from the same distribution family the prior is called *conjugate.* The posterior belief gives a distribution of the parameter given the data. The posterior can be used to update the prior, by using the parameters of the posterior as parameters for the prior. The updated prior gives a better fitting density for a new data set. This density is called *posterior predictive distribution.* The likeliness of a new data set $X_{new}$, based on an old data set $X_{old}$ and can be obtained in the following way:

$$
\begin{aligned}
P(X_{new}|X_{old}) &= \int_0^\infty P(X_{new}|\theta, X_{old})P(\theta|X_{old})\,\mathrm{d}\theta \\
&= \int_0^\infty P(X_{new}|\theta)P(\theta|X_{old})\,\mathrm{d}\theta.
\end{aligned}
$$

This distribution will be used to check whether adjustments to the model lead to the optimization aimed for. A higher value of the posterior predictive distribution when filling in the data is evidence for a better fit of the model. So changes to the model will be considered improvements only if the value of the posterior predictive is higher than before the adjustments to the model.

The mll can also be used to evaluate changes in the model. For every change made, a higher value of the mll indicates a better fit to the data.

Now that the basic notions in Bayesian inference have been explained, the specific models used in this thesis will be elaborated on in the next section.

# 3 Mathematical models

Mathematical models are used to make a mathematical formulation of the observed events. In this thesis, three mathematical models are used and for all three of them a brief explanation will be given.

## 3.1 Models for homogeneous data sets

The main goal of this thesis is to analyse the real-world taxi pick-up counts. Hence, the model should be able to deal with count data, i.e. integer-valued samples. If such a data set does not depend on time it is called *homogeneous* and one of the most popular statistical standard tools to deal with such a data set is the Poisson distribution with parameter $\lambda$. For one single observation $x$ the Poisson distribution is

$$
P(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}.
$$

The parameter $\lambda$ is a positive integer and a natural prior for $\lambda$ is the Gamma distribution with parameters $a$ and $b$:

$$P(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}.$$

This model is called the *Poisson-Gamma model.*

Another useful Bayesian model that could be used to analyse homogeneous data sets is the *Negative-Binomial-Beta model* (Neg-Bin-Beta). For this model the pdf of a Negative-Binomial is used to describe the data. There are several interpretations of a Negative-Binomial distribution, so to avoid any confusion this distribution will be defined and its expectation and variance are mentioned as well.

The Neg-Bin distribution counts the number $x$ of failures before reaching the $r$th success. Parameter $\theta$ is the probability of having a success.

$$P(x|r,\theta) = \frac{(r+x-1)!}{x!(r-1)!}\theta^r(1-\theta)^x$$

For the expectation we have:

$$\begin{aligned}
\mathrm{E}(x) &= \sum_{x=0}^{n} x \cdot \frac{(r+x-1)!}{x!(r-1)!}p^r(1-p)^x \\
&= \sum_{x=1}^{n} \frac{(r+x-1)!}{(x-1)!(r-1)!}p^r(1-p)^x \\
&= \sum_{x=1}^{n} \frac{r(1-p)}{p}\frac{(r+x-1)!}{(x-1)!(r-1)!}p^{r+1}(1-p)^{x-1} \\
&= \frac{r(1-p)}{p}\sum_{z=0}^{n} \frac{(r+1+z-1)!}{z!r!}p^{r+1}(1-p)^z \\
&= \frac{r(1-p)}{p}.
\end{aligned}$$

In a similar way we can obtain:

$$\mathrm{Var}(x) = \frac{r(1-p)}{p^2}.$$

The parameter $\theta$ is assumed to be a Beta distributed prior in this model and the Beta distribution has parameters $a$ and $b$ and is defined as:

$$P(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}.$$

The Poisson-Gamma model and the Neg-Bin-Beta model are the two standard Bayesian models that will be used in this thesis.

### 3.2 Models for non-homogeneous data

The models discussed so far can deal with homogeneous data sets. However, in the case of the real-world taxi data, the rate at which events occur might change somewhere over the time range. The frequency of occurring events may increase or decrease at certain moments in the time range. These moments will be referred to as *changepoints*. Since the number of events that occur differ over time, we are dealing with non-homogeneous count data. A real life interpretation of such a shift in rate could be rush hour where obviously higher counts can be expected.

If the existence of such changepoints is already suspected, intuitively it does not make sense to analyse such a data set with the same belief in all parameters for the whole time range. In those cases the model can be improved by embedding it in a Changepoint model architecture. In the next chapter, the concept of a Changepoint model is explained and an elaboration on the application of the model on the Poisson-Gamma model is given.

## 4 The Bayesian Poisson Changepoint model

Before starting off with the characteristics and details of this model, first the general concept of the Changepoint model will be discussed. Afterwards a detailed explanation of the application of the model will follow.

### 4.1 The concept of the changepoint model

Consider the Poisson-Gamma model from the previous chapter. This model can be extended in such a way that it can also be used to analyse non-homogeneous data sets. This can be done by embedding it in a Changepoint model architecture. This architecture will make sure that the optimal number of changepoints and their related locations in the time series are determined such that the model becomes the best possible fit to the given data set. The Poisson-Gamma model has turned into a Bayesian *Poisson Changepoint model* (CPS).

The *Metropolis-Hastings Markov Chain Monte Carlo sampling scheme* is used to determine the changepoints as well as their exact locations. Eventually, the Bayesian CPS algorithm will show several vectors which give the changepoints that make sure the model is the best fit to the given data. Based on these vectors, for each time point a probability of having a changepoint there can be calculated. So, important to realise is that not all possible changepoints are implemented in the model, only those that make the model a good fit for the data set.

### 4.2 A detailed outline of the Bayesian CPS

There are various ways to implement a Bayesian changepoint model and Grzegorczyk and Kamalabad used the classical one from Green (1995). [2] Since we are looking for an optimization of their Bayesian CPS, this implementation of the Bayesian changepoint model to the Poisson-Gamma model will be used.

Grzegorczyk and Kamalabad have written an algorithm where a data set is given as input to the Bayesian CPS and the changepoints that make the model the best fit to this data set are given as output. In this thesis this model and its algorithm form the starting point. From here adjustments will be made, which optimize the model. The codes can be reproduced if one is interested in them.

As stated before the classical form of the Bayesian changepoint model from Green (1995) is used. The model will be applied to various Poisson synthetic data sets and on real-world taxi pick-up counts. The data set given as input for the model will be divided in $K$ components and is identified with $K-1$ changepoints. Hence, after each changepoint a new component begins. $K$ is a truncated Poisson distribution, e.g. $K$ must be a positive integer. Conditional on $K$, the changepoints are assumed to have the even-numbered order statistics of $L := 2(K-1)$ points uniformly distributed on the data set. This implies that two changepoints can not be located at two consecutive time points.

At each iteration the MCMC randomly selects either the based on changepoint birth, death or re-allocation move and performs the chosen move (Green (1995)). Each move has a probability of $\frac{1}{3}$ to be selected. The three move types can be briefly described as follows:

1. *The changepoint birth move*: This move randomly places one single new changepoint at one of its possible locations. The data becomes divided in $K+1$ components instead of $K$.
2. *The changepoint re-allocation move*: This move randomly picks one of the existing changepoints and relocate this changepoint at one of the

possible locations between the surrounding changepoints. Suppose we have a current state with three changepoints, $c_1, c_2, c_3$ and in the next iteration of the MCMC the re-allocation move was selected. Suppose $c_2$ was randomly selected to be re-allocated. The new location of $c_2$ will be one of the possible locations of a changepoint between the changepoints $c_1$ and $c_3$. The number of components stays the same.

3. *The changepoint death move*: This move randomly selects one of the existing changepoints and deletes it. $K$ becomes $K - 1$.

Each of these three moves are proposals and will not be performed automatically. For each move the *Metropolis-Hastings acceptance probability* for the candidate state is used to see whether to accept or reject the move. If the move is accepted the move is performed and accept the new state. If the move is rejected the state is left unchanged. The Metropolis-Hastings acceptance probability is defined

$$A = \frac{P((x_1, ..., x_n)|\text{New changepoints})}{P((x_1, ..., x_n)|\text{Old changepoints})} \cdot \frac{P(\text{New changepoints})}{P(\text{Old changepoints})} \cdot Q.$$

$Q$ is the Hastings ratio, which can be computed straightforwardly for each of the three move types. It usually is the ratio of the probability of going from the new changepoints to the old ones and vice versa from the old changepoints to the new ones.

In Grzegorczyk and Kamalabad the Bayesian CPS is explained in more rigorously. However, this thesis is focused on the adjustments of this model rather than the mathematics behind it.

## 4.3 Room for improvement of the Bayesian CPS

As Grzegorczyk and Kamalabad stated in their paper, the Bayesian CPS is suboptimal for the type of data set they analysed, namely the taxi pick-up counts. The main goal in this thesis is to improve the Bayesian CPS for this kind of data set. A short outline of the ideas for improvements are given in this section. The purpose of this section is to familiarise the reader with the directions this thesis will follow.

The Poisson Changepoint model is based on a standard Bayesian model with conjugate prior, namely the Poisson-Gamma model. One of the key features of the Poisson distribution is that the mean equals the variance. However, this is unrealistic when applying the model on real-world taxi pick-up counts. In these type of data sets, the data exhibits over-dispersion, e.g. the variance is larger than the mean. Therefore, the Poisson distribution can not be called optimal to deal with these kind of data sets. Grzegorczyk and Kamalabad proposed to use a Negative Binomial distribution instead. It has a variance larger than the mean and hence, deals with over-dispersion. Therefore the first attempt to improve the Bayesian CPS is by replacing the Poisson distribution with a Negative Binomial one and see if this leads to a better fitting model.

The next main idea for improvement is by *information exchange* among components. Information exchange could be implemented by constructing a prior belief that is used in an iteration of the MCMC scheme, based on data observed so far. The information exchange can be implemented on two different levels: global and sequential. Global information exchange means that at the end of each MCMC iteration, the parameters for the prior will be updated and used for every component of the data set. In sequential information exchange the prior for each component is updated based on the observed data from the previous component. Information exchange on global and sequential level will form the second direction this thesis will follow.

# 5 Derivations

The concepts of mll and posterior predictive distribution have already been discussed briefly. They will be used to check if adjustments to the model can be seen as improvements. This section will provide all the necessary derivations of equations that will be used in this thesis.

## 5.1 Poisson-Gamma model

The pdf for the Poisson distribution has already been given. However, to analyse data sets with multiple data points the *joint probability density function* is needed. If a serie of independent identically distributed (i.i.d.) observations is sampled from $P(x|\lambda)$, the joint pdf of $x_1, x_2, \ldots, x_n$ is the product of the individual pdfs:

$$\begin{aligned} P(x_1,\ldots,x_n|\lambda) &= \prod_{i=1}^{n} P(x_i|\lambda) \\ &= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \\ &= \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}. \end{aligned}$$

The marginal likelihood of this model will use the joint pdf and the Gamma prior:

$$\begin{aligned} P(x_1,\ldots,x_n) &= \int_{\lambda} P(x_1,\ldots,x_n,\lambda)\,\mathrm{d}\lambda \\ &= \int_{\lambda} P(x_1,\ldots,x_n|\lambda)P(\lambda)\,\mathrm{d}\lambda \\ &= \int_0^{\infty} \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \cdot \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}\,\mathrm{d}\lambda \\ &= \frac{b^a}{\Gamma(a)} \frac{1}{\prod_{i=1}^{n} x_i!} \int_0^{\infty} \lambda^{\sum_{i=1}^{n} x_i+a-1} e^{-(b+n)\lambda}\,\mathrm{d}\lambda \\ &= \frac{b^a}{\Gamma(a)} \frac{1}{\prod_{i=1}^{n} x_i!} \frac{\Gamma(\sum_{i=1}^{n} x_i + a)}{(b+n)^{\sum_{i=1}^{n} x_i+a}} \int_0^{\infty} \frac{(b+n)^{\sum_{i=1}^{n} x_i+a}}{\Gamma(\sum_{i=1}^{n} x_i + a)} \lambda^{\sum_{i=1}^{n} x_i+a-1} e^{-(b+n)\lambda}\,\mathrm{d}\lambda \\ &= \frac{b^a}{\Gamma(a)} \frac{1}{\prod_{i=1}^{n} x_i!} \frac{\Gamma(\sum_{i=1}^{n} x_i + a)}{(b+n)^{\sum_{i=1}^{n} x_i+a}}. \end{aligned}$$

Note that the integral in the second step is hard to solve analytically. However, no additional calculus was used to solve it. Instead the term necessary to obtain a Gamma density inside the integral were added. The inverse of these terms is added in front of the integral for obvious reasons. The integral taken of a density is always equal to 1. In this case a Gamma density with parameters $\sum_{i=1}^{n} x_i + a$ and $b+n$ is obtained and therefore the integral becomes 1.

To apply Bayesian inference, the parameters of the posterior can be used as updated prior parameters. The posterior distribution for the Poisson-Gamma model look like this:

$$\begin{aligned} P(\lambda|x_1,\ldots,x_n) &\propto \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \cdot \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} \\ &\propto \lambda^{\sum_{i=1}^{n} x_i+a-1} e^{-\lambda(b+n)}. \end{aligned}$$

The posterior belief is Gamma distributed with parameters $\sum_{i=1}^n x_i + a$ and $b+n$. The posterior predictive displays the likeliness of having the new data set $\tilde{x}_1, \ldots, \tilde{x}_n$, given the old data set $x_1, ..., x_n$.

$$\begin{aligned}
P(\tilde{x}_1, \ldots, \tilde{x}_n | x_1, \ldots, x_n) &= \int_0^\infty P(\tilde{x}_1, \ldots, \tilde{x}_n | \lambda, x_1, \ldots, x_n) P(\lambda | x_1, \ldots, x_n) \, \mathrm{d}\lambda \\
&= \int_0^\infty P(\tilde{x}_1, \ldots, \tilde{x}_n | \lambda) P(\lambda | x_1, \ldots, x_n) \, \mathrm{d}\lambda \\
&= \int_0^\infty \frac{\lambda^{\sum_{i=1}^n \tilde{x}_i} e^{-n\lambda}}{\prod_{i=1}^n \tilde{x}_i!} \cdot \frac{(b+n)^{\sum_{i=1}^n x_i + a}}{\Gamma(\sum_{i=1}^n x_i + a)} \lambda^{\sum_{i=1}^n x_i + a - 1} e^{-(b+n)\lambda} \, \mathrm{d}\lambda \\
&= \frac{1}{\prod_{i=1}^n \tilde{x}_i!} \frac{(b+n)^{\sum_{i=1}^n x_i + a}}{\Gamma(\sum_{i=1}^n x_i + a)} \int_0^\infty \lambda^{\sum_{i=1}^n \tilde{x}_i + \sum_{i=1}^n x_i + a - 1} e^{-(b+2n)\lambda} \, \mathrm{d}\lambda \\
&= \frac{1}{\prod_{i=1}^n \tilde{x}_i!} \frac{(b+n)^{\sum_{i=1}^n x_i + a}}{\Gamma(\sum_{i=1}^n x_i + a)} \frac{\Gamma(a + \sum_{i=1}^n x_i + \sum_{i=1}^n \tilde{x}_i)}{(b+2n)^{a + \sum_{i=1}^n \tilde{x}_i + \sum_{i=1}^n x_i}} \cdot \\
&\int_0^\infty \frac{(b+2n)^{a + \sum_{i=1}^n \tilde{x}_i + \sum_{i=1}^n x_i}}{\Gamma(a + \sum_{i=1}^n x_i + \sum_{i=1}^n \tilde{x}_i)} \lambda^{\sum_{i=1}^n \tilde{x}_i + \sum_{i=1}^n x_i + a - 1} e^{-(b+2n)\lambda} \, \mathrm{d}\lambda \\
&= \frac{1}{\prod_{i=1}^n \tilde{x}_i!} \frac{(b+n)^{\sum_{i=1}^n x_i + a}}{\Gamma(\sum_{i=1}^n x_i + a)} \frac{\Gamma(a + \sum_{i=1}^n x_i + \sum_{i=1}^n \tilde{x}_i)}{(b+2n)^{a + \sum_{i=1}^n \tilde{x}_i + \sum_{i=1}^n x_i}}.
\end{aligned}$$

## 5.2 Negative-Binomial-Beta model

The joint pdf of $x_1, x_2, \ldots, x_n$ where the $x_i$ are i.i.d. observations from $P(x|\theta)$ is:

$$\begin{aligned}
P(x_1, \ldots, x_n | r, \theta) &= \prod_{i=1}^n P(x_i | r, \theta) \\
&= \prod_{i=1}^n \frac{(r + x_i - 1)!}{x_i!(r-1)!} \theta^r (1-\theta)^{x_i} \\
&= \frac{\prod_{i=1}^n (r + x_i - 1)!}{\prod_{i=1}^n x_i! ((r-1)!)^n} \theta^{nr} (1-\theta)^{\sum_{i=1}^n x_i}.
\end{aligned}$$

The mll is:

$$\int_0^1 \frac{\prod_{i=1}^n (r+x_i-1)!}{\prod_{i=1}^n x_i!((r-1)!)^n} \theta^{nr} (1-\theta)^{\sum_{i=1}^n x_i} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} \, \mathrm{d}\theta$$
$$= \frac{\prod_{i=1}^n (r+x_i-1)!}{\prod_{i=1}^n x_i!((r-1)!)^n} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta^{nr+a-1} (1-\theta)^{b+\sum_{i=1}^n x_i - 1} \, \mathrm{d}\theta$$
$$= \frac{\prod_{i=1}^n (r+x_i-1)!}{\prod_{i=1}^n x_i!((r-1)!)^n} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(nr+a)\Gamma(b+\sum_{i=1}^n x_i)}{\Gamma(nr+a+b+\sum_{i=1}^n x_i)} \int_0^1 \frac{\Gamma(nr+a+b+\sum_{i=1}^n x_i)}{\Gamma(nr+a)\Gamma(b+\sum_{i=1}^n x_i)} \cdot$$
$$\theta^{nr+a-1} (1-\theta)^{b+\sum_{i=1}^n x_i - 1} \, \mathrm{d}\theta$$
$$= \frac{\prod_{i=1}^n (r+x_i-1)!}{\prod_{i=1}^n x_i!((r-1)!)^n} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(nr+a)\Gamma(b+\sum_{i=1}^n x_i)}{\Gamma(nr+a+b+\sum_{i=1}^n x_i)}.$$

Note that the integral is taken of a Beta density function with parameters $nr + a$ and $b + \sum_{i=1}^n x_i$.

The posterior belief is:

$$\begin{aligned} P(\theta|x_1,\ldots,x_n) &\propto P(x_1,\ldots,x_n|r,\theta) \cdot P(\theta|r) \\ &\propto \frac{\prod_{i=1}^n (r+x_i-1)!}{\prod_{i=1}^n x_i!((r-1)!)^n} \theta^{nr} (1-\theta)^{\sum_{i=1}^n x_i} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} \\ &\propto \theta^{nr+a-1} (1-\theta)^{b+\sum_{i=1}^n x_i - 1}. \end{aligned}$$

This posterior distribution is Beta distributed with parameters $nr + a$ and $b + \sum_{i=1}^n x_i$. The posterior predictive distribution is obtained as follows:

$$
\begin{aligned}
&P(\tilde{x}_1, \ldots, \tilde{x}_n | x_1, \ldots, x_n) \\
&= \int_0^1 P(\tilde{x}_1, \ldots, \tilde{x}_n | r, \theta, x_1, ..., x_n) P(\theta | r, x_1, ..., x_n) \, \mathrm{d}\theta \\
&= \int_0^1 P(\tilde{x}_1, \ldots, \tilde{x}_n | r, \theta) P(\theta | r, x_1, ..., x_n) \, \mathrm{d}\theta \\
&= \int_0^1 \frac{\prod_{i=1}^n (r + \tilde{x}_i - 1)!}{\prod_{i=1}^n \tilde{x}_i!((r-1)!)^n} \theta^{nr} (1-\theta)^{\sum_{i=1}^n \tilde{x}_i} \cdot \frac{\Gamma(nr + a + b + \sum_{i=1}^n x_i)}{\Gamma(nr+a)\Gamma(b + \sum_{i=1}^n x_i)} \theta^{nr+a-1} (1-\theta)^{b + \sum_{i=1}^n x_i - 1} \, \mathrm{d}\theta \\
&= \frac{\prod_{i=1}^n (r + \tilde{x}_i - 1)!}{\prod_{i=1}^n \tilde{x}_i!((r-1)!)^n} \cdot \frac{\Gamma(nr + a + b + \sum_{i=1}^n x_i)}{\Gamma(nr+a)\Gamma(b + \sum_{i=1}^n x_i)} \int_0^1 \theta^{2nr+a-1} (1-\theta)^{b + \sum_{i=1}^n \tilde{x}_i + \sum_{i=1}^n x_i - 1} \, \mathrm{d}\theta \\
&= \frac{\prod_{i=1}^n (r + \tilde{x}_i - 1)!}{\prod_{i=1}^n \tilde{x}_i!((r-1)!)^n} \cdot \frac{\Gamma(nr + a + b + \sum_{i=1}^n x_i)}{\Gamma(nr+a)\Gamma(b + \sum_{i=1}^n x_i)} \frac{\Gamma(2nr+a)\Gamma(b + \sum_{i=1}^n \tilde{x}_i + \sum_{i=1}^n x_i)}{\Gamma(2nr + a + b + \sum_{i=1}^n \tilde{x}_i + \sum_{i=1}^n x_i} \cdot \\
&\int_0^1 \frac{\Gamma(2nr + a + b + \sum_{i=1}^n \tilde{x}_i + \sum_{i=1}^n x_i)}{\Gamma(2nr+a)\Gamma(b + \sum_{i=1}^n \tilde{x}_i + \sum_{i=1}^n x_i)} \theta^{2nr+a-1} (1-\theta)^{b + \sum_{i=1}^n \tilde{x}_i + \sum_{i=1}^n x_i - 1} \, \mathrm{d}\theta \\
&= \frac{\prod_{i=1}^n (r + \tilde{x}_i - 1)!}{\prod_{i=1}^n \tilde{x}_i!((r-1)!)^n} \cdot \frac{\Gamma(nr + a + b + \sum_{i=1}^n x_i)}{\Gamma(nr+a)\Gamma(b + \sum_{i=1}^n x_i)} \frac{\Gamma(2nr+a)\Gamma(b + \sum_{i=1}^n \tilde{x}_i + \sum_{i=1}^n x_i)}{\Gamma(2nr + a + b + \sum_{i=1}^n \tilde{x}_i + \sum_{i=1}^n x_i}.
\end{aligned}
$$

## 6 Poisson-Gamma versus Neg-Bin-Beta model

The first suggestion for improvement of the Bayesian CPS is replacing the Poisson-Gamma model with the Neg-Bin-Beta model. In this chapter it will be explored if the Neg-Bin-Beta model can be used as a substitute for the Poisson-Gamma model. This will be investigated by calculating the mll and the posterior predictive distribution for both models using the same data sets. Replacing the Poisson-Gamma model will be an improvement since the Neg-Bin-Beta model also deals with over-dispersion, whereas he Poisson-Gamma model does not. Hence, we would obtain a model that handles Poisson generated data as good as the original model, but can deal with data sets with over-dispersion as well.

The upcoming plots, histograms, tables et cetera are all modeled with Matlab. However, not all plots and histograms will be included in this thesis. The results that are necessary to understand the conclusions are shown, but those are not the only results that form the basis for the conclusion. A few results are shown that present the general trend of the results.

## 6.1 Data

In this chapter the data sets that are used to compare both models are synthetic count data sets generated from both the Poisson and Negative Binomial distribution. For the comparison in this chapter homogeneous data sets are used. Hence, the data sets are generated with one value for either $\lambda$ (Poisson) or $\theta$ (Neg-Bin). This is sufficient: if the Negative-Binomial is as good as the Poisson for homogeneous data sets, it will be as good as the Poisson for a non-homogeneous data set, since we can divide these data sets into components with each its own parameter. In other words, a non-homogeneous data set can be divided in multiple homogeneous components.

## 6.2 The marginal likelihood

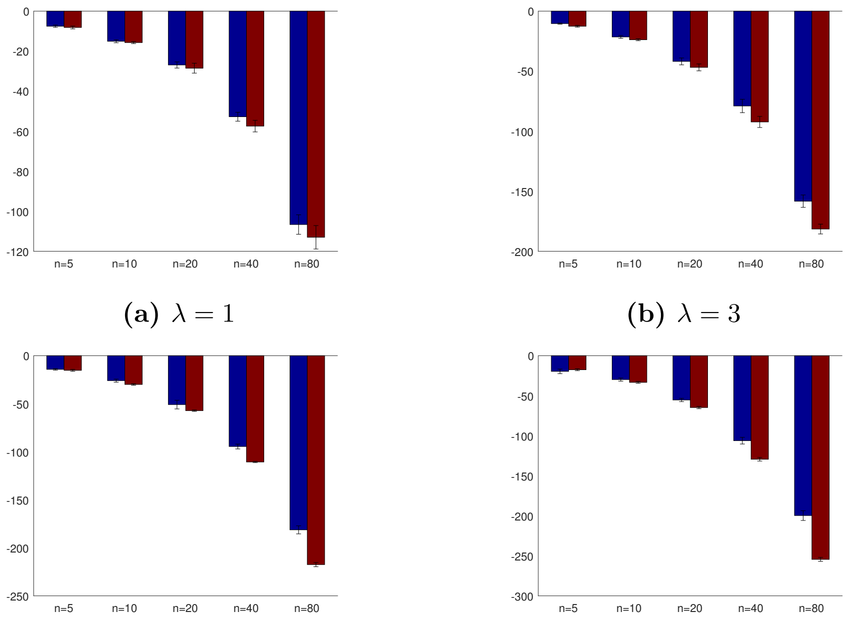
The values of the mll and posterior predictive are very small numbers, therefore their log values are taken. Note that when taking the logarithm, negative values appear. In figure 1, histograms are shown where the Poisson log marginal likelihood is compared to the Negative Binomial log marginal likelihood on Poisson generated data with different sample sizes and different $\lambda$.

It is easy to see that the Poisson distribution is superior to the Negative binomial distribution for all values of $\lambda$ and all sample sizes. This is a surprising result, since the Negative Binomial deals with counting data as well. To improve the Bayesian CPS, ideal would be that the Negative Binomial is as least as good as the Poisson. Therefore the next step is to find different ways to improve the Negative Binomial as an approximation for the Poisson.

The Negative-Binomial distribution has two different parameters, $r$ and $\theta$, where $\theta$ is Gamma distributed and has parameters $a$ and $b$. Hence, $a, b$ and $r$ are the three parameters that the Neg-Bin-Beta model depends on. So far they have all been fixed at 1. For these values the Neg-Bin turned out not to be an optimal approximation for the Poisson. However, for other values of these parameters it may be.

## 6.3 Parameter $r$ of the Negative-Binomial distribution

The first improvement of the Neg-Bin-Beta as an approximation of the Poi-Gam, might lay in the parameter $r$. In figure 2, the log predictive distribution are given for different values of $r$. The values of the posterior predictive increase quite fast when $r$ goes to approximately 10 and after that the changes in $r$ do not seem to have a large impact on the values of the posterior predictive.



**(a)** $\lambda = 1$

**(b)** $\lambda = 3$

**(c)** $\lambda = 5$

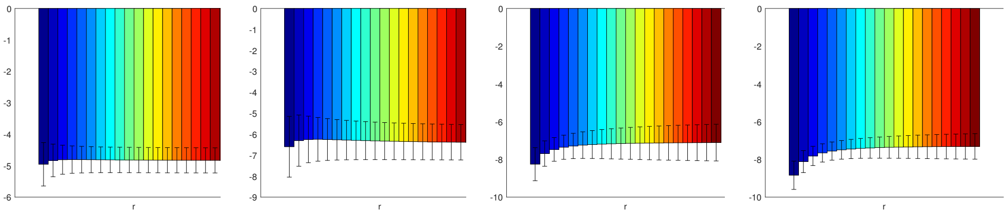
**(d)** $\lambda = 8$

**Figure 1:** Comparison marginal likelihood. Histograms of the average log marginal likelihood based on Poisson generated data with the given $\lambda$, for different sample sizes $n$. The blue (first) bar represents the log mll of the Poi-Gam and the red (second) bar represents the log mll of the Neg-Bin-Beta. The error bars represent standard deviations.

Based on these results the following hypothesis is formulated: the Neg-Bin-Beta model approximates the Poi-Gam model better for higher values of $r$. This hypothesis is tested by plotting the likelihood function of the Neg-Bin-Beta model for different values of $r$ and the likelihood function of the Poi-Gam model in one figure. The results are shown in figure 3. For $r = 1$ the Neg-Bin is clearly not a good approximation of the Poisson distribution. However, for values of $r = 5$ or higher, the negative binomial approaches the Poisson distribution a lot better. In this plot a sample size of 40 was used to test both models. However, eventually the changepoint model will divide the input data set is components and those components are likely to have smaller sample sizes. Therefore this hypothesis is tested even further for smaller sample sizes. Again the log mll is used to test both models. In figure 4 those results are displayed.

Figure 4 shows the log mll of the Poi-Gam in the blue (first) bars and for the Neg-Bin-Beta with different values of $r$ in the red (second) bars. One data set was generated with a sample size of 3 and for different values of $r$ the mll was calculated. Therefore there is no change in the mll of the Poi-Gam model, since this function does not depend on $r$. From the histograms it can be seen that there is a clear improvement of the Neg-Bin-Beta model.

A similar result can be seen in figure 5 where the log posterior predictions



**(a)** $\lambda = 1$ **(b)** $\lambda = 3$ **(c)** $\lambda = 5$ **(d)** $\lambda = 8$

**Figure 2:** Log posterior predictions for different $r$ values. Histograms of the average Negative Binomial log predictive probability based on Poisson generated data with the given $\lambda$. The value of $r$ increases from 1 to 20.

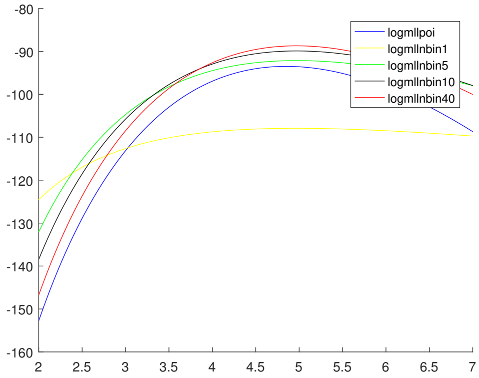


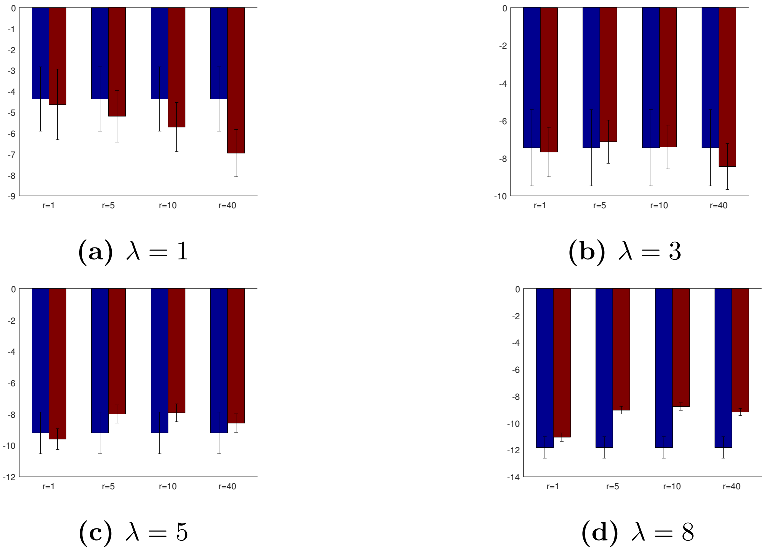
**Figure 3:** Plot of the average likelihood for different $r$ values. On the $x$-axis values for $\lambda$ used to generate the Poisson data and on the $y$-axis the likelihood. A sample size of $n = 40$ was used.

have been displayed in the histograms. Again one data set with sample size 3 was analysed for different values of $r$. For $r = 1$ the Poi-Gam model gives higher values of the posterior predictions and the Negative-Binomial becomes superior to the Poisson at $r = 5$ or higher. Note that very large differences are not needed. Since it is not expected to see that the Neg-Bin would be superior to the Poisson distribution on Poisson generated data sets. However, since we are not dealing with the pure pdf but Bayesian models, this can happen. But if the Poi-Gamma might give slightly higher values of posterior predictions than the Neg-Bin-Beta, those would be good results. The goal is to aim for parameters that make the Neg-Bin-Beta good approximations of the Poi-Gam, not necessarily better.

Based on these results the conclusion can be drawn that the Negative Binomial is as good as the Poisson for Poisson generated data sets when the parameter $r$ is at least 5. However, what is the optimal value of $r$? There are two different options. Parameter $r$ can be either fixed at a certain value or $r$ is made a free parameter. Some experiments were done to analyse the behavior of this parameter. The plots showed that there were no signs of convergence of the value of $r$. The new values of $r$ were uniformly distributed and the plots showed that $r$ would make a random walk. Hence, the value of $r$ does not have a significant influence on the marginal likelihood. Therefore, to avoid making the model unnecessarily complicated, the parameter $r$ will be fixed for a certain value higher than five. From figures 3, 4, 5 one can conclude that for $r = 10$ the Negative Binomial is as good as the Poisson for Poisson generated data, so from now on $r$ would be fixed at 10.

At this point it is shown that for $r = 10$ the Neg-Bin-Beta model is as good as the Poisson-Gamma model on Poisson generated data, in these kind of data sets over-dispersion does not occur. Left to show is how this change would actually be an improvement.

Instead of analysing a data set generated from a Poisson distribution, data sets from the Negative-Binomial distribution are gathered and again both models were tested. In figure 6 three rows of histograms are displayed. The first row shows the log mll values of the two models where different sample sizes were used. In the second row the log mll is shown where different values of $r$ are used and in the last row the log posterior predictives were calculated, again with different values of $r$. It is easy to see that the Neg-Bin-Beta model is a better fit to these kind of data sets than the Poisson-Gamma model. This is not a surprising results since it is general knowledge that the Poisson distribution does not deal with over-dispersion and the Negative-Binomial does.



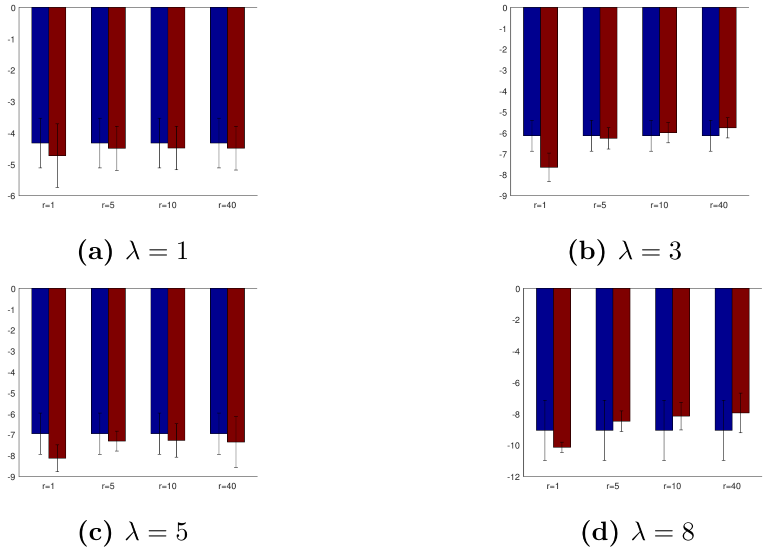
**(a)** $\lambda = 1$

**(b)** $\lambda = 3$

**(c)** $\lambda = 5$

**(d)** $\lambda = 8$

**Figure 4:** Histograms of the average log marginal likelihood based on Poisson generated data with the given $\lambda$. The blue (first) bars represent the log marginal likelihood of the Poisson distribution. The red (second) bars represent the log marginal likelihood of the Negative binomial for a certain value of $r$. The sample sizes of the data sets are all 3.



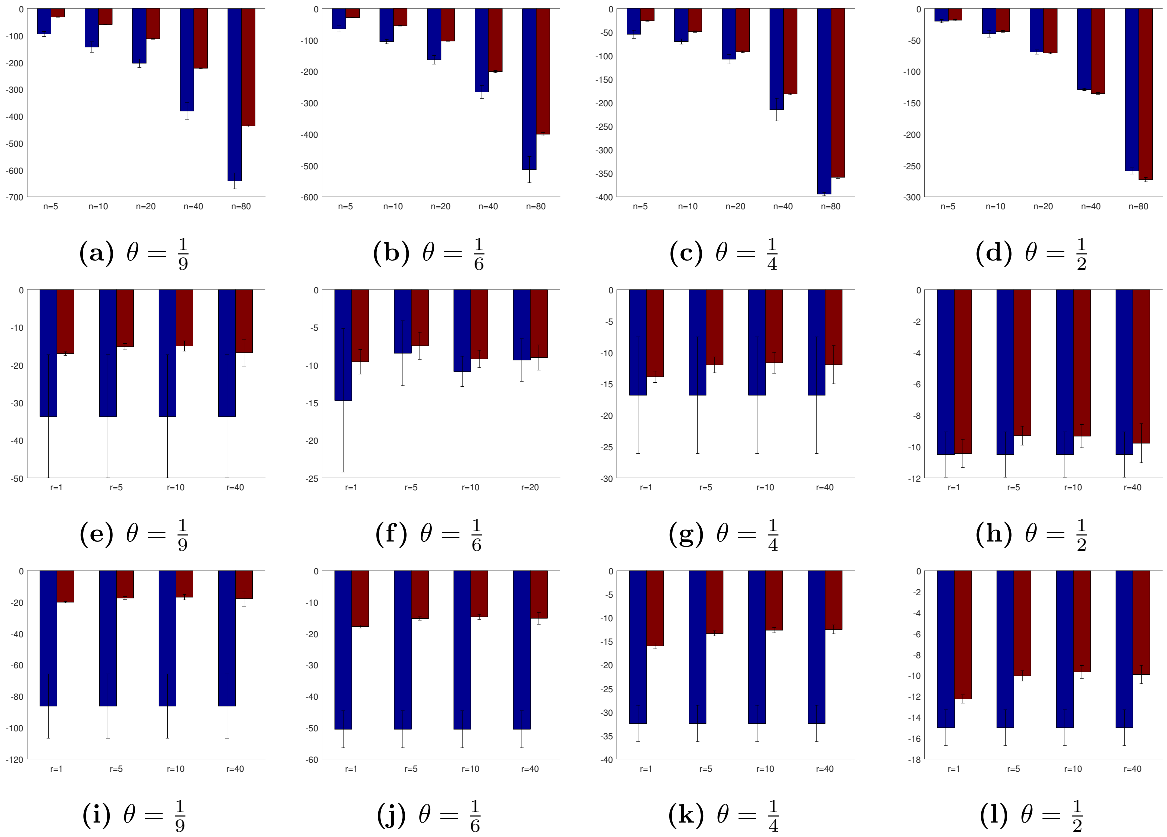
**(a)** $\lambda = 1$

**(b)** $\lambda = 3$

**(c)** $\lambda = 5$

**(d)** $\lambda = 8$

**Figure 5:** Histograms of the average log predictive distribution based on Poisson generated data with the given $\lambda$. The blue (first) bars represent the log predictive distribution of the Poisson distribution. The red (second) bars represent the log predictive distribution of the Negative binomial for a certain value of $r$. The sample sizes of the data sets are all 3.



**(a)** $\theta = \frac{1}{9}$ **(b)** $\theta = \frac{1}{6}$ **(c)** $\theta = \frac{1}{4}$ **(d)** $\theta = \frac{1}{2}$

**(e)** $\theta = \frac{1}{9}$ **(f)** $\theta = \frac{1}{6}$ **(g)** $\theta = \frac{1}{4}$ **(h)** $\theta = \frac{1}{2}$

**(i)** $\theta = \frac{1}{9}$ **(j)** $\theta = \frac{1}{6}$ **(k)** $\theta = \frac{1}{4}$ **(l)** $\theta = \frac{1}{2}$

**Figure 6:** Histograms of the average log marginal likelihood probability based on Negative Binomial generated data with the given $\theta$. The blue (first) bars represent the values for the Poisson model. The red (second) bars represent the values for the Negative binomial model. The error bars represent the standard deviations. The whole upper row represent the log marginal likelihoods for different increasing sample sizes and $r = 1$. The second row shows the log mll, where one data set generated for the given $\theta$ was analysed for different values of $r$. The sample sizes were 40. The third row shows the same as the second row, but instead the log mll, the log posterior predictions are displayed.

Therefore the adjustment of replacing the Poisson-Gamma by the Neg-Bin-Beta model with $r = 10$ is an actual improvement. The Bayesian Poisson Changepoint model changes in a Bayesian Negative-Binomial Changepoint model and is a good fit to data sets in which over-dispersion might occur.

Analysing the data sets that were generated from the Neg-Bin distribution resulted in one other surprising result. The parameter $r^*$ from the Negative-Binomial distribution that was used to generate the data turned out to have a significant influence on the values of the mll and the posterior predictive. In figure 6 data sets were generated with the given values of $\theta$, sample size 40 and $r^* = 10$. However, if higher values of $r^*$ were used to generate the data, higher values of the mll and posterior predictive were reached.

With this new knowledge one last test was performed to decide if $r$ could stay fixed. Different data sets were gathered with all a different $r^*$ value and the mll was calculated multiple times, each time with a different $r$ values. Where $r^*$ is the parameter that was used to generate the data and $r$ the parameter that belongs to the Negative-Binomial model and therefore the mll and posterior predictive formulas. Hence, it was tested if different values of $r^*$ had different values for $r$ that lead to the optimal mll or posterior predictive. It turned out that despite the fact higher $r^*$ resulted in higher mll values of the Neg-Bin-Beta model, the parameter $r$ in the model did not have any furhter influence on these values. Hence, $r$ can stay fixed.

In conclusion, the first step in the optimization of the Bayesian Poisson Changepoint model is replacing the Poisson-Gamma model with the Neg-Bin-Beta model where $r$ is fixed at 10.

# 7 Parameters $a$ and $b$ of the Beta prior

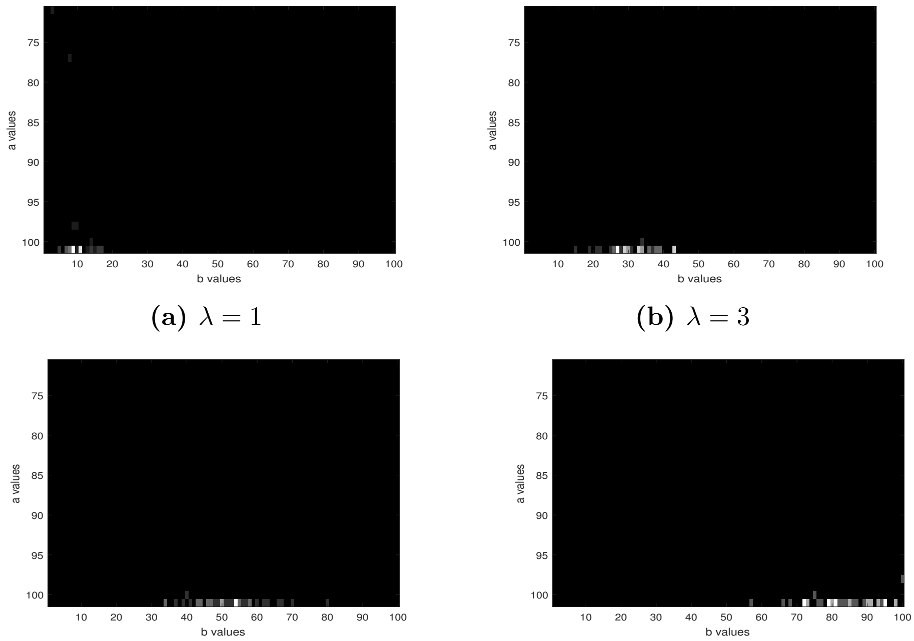
The Neg-Bin-Beta model has three parameters it depends on. So far it has been established that $r$ should be fixed at a value of 10. The other parameters $a$ and $b$ are still left to examine. These parameters were until now fixed at a value of one. However, making one or perhaps both parameters free might result in a better fitting model to the various data sets.

Before making the parameters free, different values of $a$ and $b$ were tested to see if they have any kind of influence on the marginal likelihood of the Neg-Bin-Beta model. Figure 7 displays four grey scaled color maps. On the x-axis the values for parameter $b$ are given and the $y$-axis shows the values of parameter $a$. Experiments were performed in which the number of times a combination of $a$ and $b$ values, gave the maximum average log mll based on Poisson generated data was calculated. The dark colors stand for a low number of times. The lighter the color, the more often that combination gave the maximum log mll. For $a$ the displayed values start at $a = 70$, due to the reason that there were no optimal combination with parameter $a$ lower than 70.

Figure 7 clearly shows that the optimal values for $a$ and $b$ change for each $\lambda$. An increase of the $\lambda$ results in an increase of the $b$ value. This intuitively makes sense. When $\lambda$ gets higher a data set with higher counts is obtained. The higher the counts, the longer the waiting time for the $r$th success and hence, the lower the probability of having a success. The probability of having a success is denoted by $\theta$ and in the current model $\theta$ was given a Beta distribution with parameters $a$ and $b$. The expectation of parameter $\theta$ is therefore:

$$E(\theta) = \frac{a}{a+b}.$$

In figure 7 the optimal $a$ values stayed the same, namely all 100. However, $b$ increased when $\lambda$ increased. The expectation shows that when $a$ keeps the same value but $b$ increases the expected value of $\theta$ decreases. Hence, a higher $b$ value results in a lower probability of having a success. Figure 7 makes therefore intuitively sense. The exact behaviour of $a$ and $b$ are however not clear. Figures 8 and 9 show various color maps which were used to find a pattern in the behavior of parameters $a$ and $b$.



**(a)** $\lambda = 1$

**(b)** $\lambda = 3$

**(c)** $\lambda = 5$

**(d)** $\lambda = 8$

**Figure 7:** Grey scaled colormaps. The colormaps represent the number of times a certain combination of values $a$ and $b$ gave the maximum log mll. A light color stands for a high number of times that combination of $a$ and $b$ gave the maximum log mll for Poisson generated data with the given $\lambda$.

The pattern became very clear. The values of $a$ and $b$ that gave the maximum mll were the values that made parameter $\theta$ close to the maximum likelihood estimator of the Negative Binomial distribution. This maximum likelihood estimator is obtained by solving:

$$\frac{\mathrm{d}}{\mathrm{d}\,\theta} log(\frac{\prod_{i=1}^{n}(r+x_i-1)!}{\prod_{i=1}^{n} x_i!((r-1)!)^n}\theta^{nr}(1-\theta)^{\sum_{i=1}^{n} x_i}) = 0.$$

Hence,

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\,\theta} log(\prod_{i=1}^{n}(r+x_i-1)!) - (log(\prod_{i=1}^{n} x_i!) + nlog((r-1)!)) + nrlog(\theta) + \sum_{i=1}^{n} x_i\, log(1-\theta) = 0 \\
\frac{nr}{\theta} - \frac{\sum_{i=1}^{n} x_i}{1-\theta} = 0 \\
\frac{nr}{\theta} = \frac{\sum_{i=1}^{n} x_i}{1-\theta} \\
nr(1-\theta) = \sum_{i=1}^{n} x_i \cdot \theta \\
nr = \theta(\sum_{i=1}^{n} x_i + nr) \\
\theta = \frac{nr}{\sum_{i=1}^{n} x_i + nr} \\
\theta = \frac{r}{\frac{\sum_{i=1}^{n} x_i}{n} + r} \\
\theta \approx \frac{r}{\lambda + r}
\end{aligned}
$$

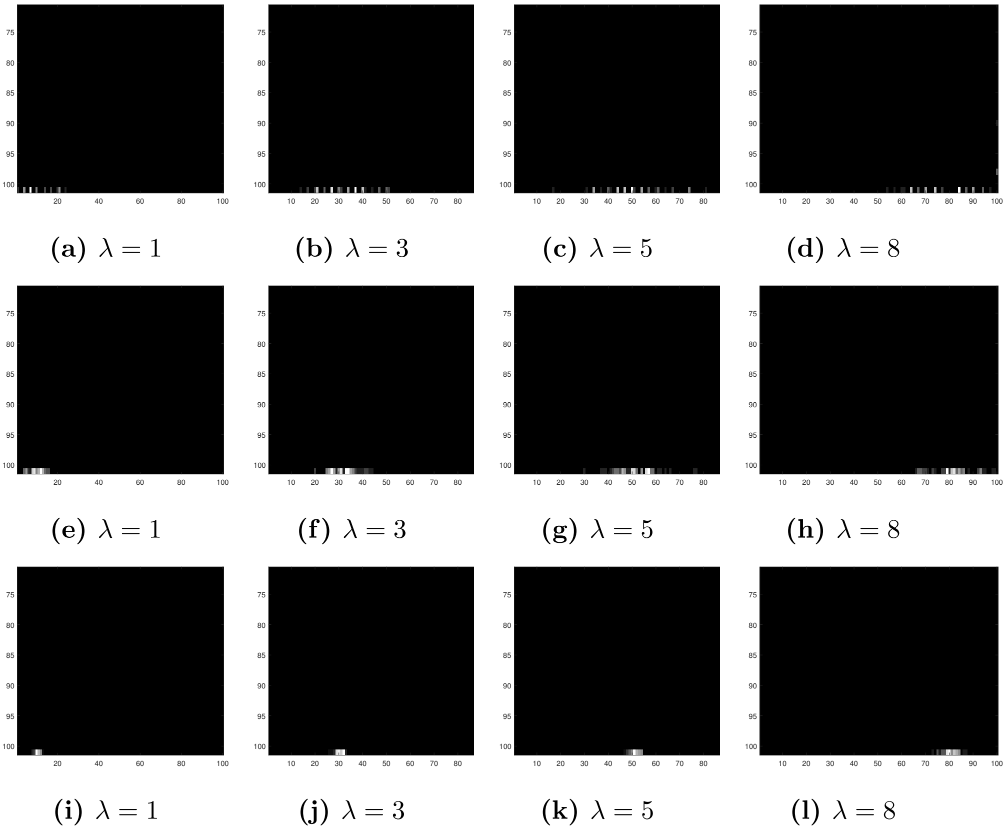
Since $\theta$ is Beta distributed the values $a$ and $b$ that maximize the model therefore satisfy:

$$\frac{r}{\lambda + r} \approx \frac{a}{a+b}.$$

From the Negative-Binomail Changepoint model $r = 10$ and the data sets used in figure 7 are generated from the Poisson distribution, hence $\sum_{i=1}^{n} x_i \approx n\lambda$. It is easy to check that the values of $a$ and $b$ in figures 7, 8, 9 satisfy this equation.

From figures 7, 8 and 9 it can also be concluded that the optimal value for $a$ is always the maximum value of $a$ for which the equation can still be satisfied. Hence, the optimal $a$ is the highest value of $a$ possible if the corresponding $b$ value is available as well. This will be illustrated with an example. Take a data set that is generated with $\lambda = 8$. Parameters $a$ and $b$ must satisfy

$$
\begin{aligned}
\frac{10}{10+8} = \frac{a}{a+b} \\
18a = 10a + 10b \\
8a = 10b \\
b = \frac{4}{5}a.
\end{aligned}
$$

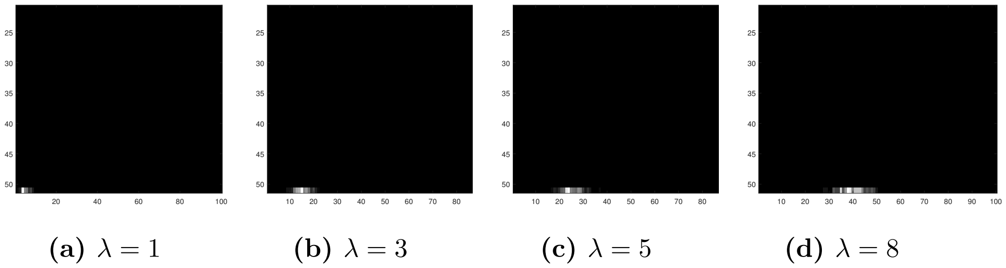


**(a)** $\lambda = 1$ **(b)** $\lambda = 3$ **(c)** $\lambda = 5$ **(d)** $\lambda = 8$

**(e)** $\lambda = 1$ **(f)** $\lambda = 3$ **(g)** $\lambda = 5$ **(h)** $\lambda = 8$

**(i)** $\lambda = 1$ **(j)** $\lambda = 3$ **(k)** $\lambda = 5$ **(l)** $\lambda = 8$

**Figure 8:** Grey scaled colormaps. The colormaps represent the number of times a certain combination of values $a$ and $b$ gvae the maximmum log mll. A light color stands for a high number of times. Grids a: 70:100, b; 1:100. First row: n=3. Second row: n=10. Third row: n=100



**(a)** $\lambda = 1$ **(b)** $\lambda = 3$ **(c)** $\lambda = 5$ **(d)** $\lambda = 8$

**Figure 9:** Grey scaled colormaps. The colormaps represent the number of times a certain combination of values $a$ and $b$ gvae the maximmum log mll. A light color stands for a high number of times. Grids a: 1:50, b: 1:100, n=10.

Suppose the maximum likelihood is calculated for different values of $a$ and $b$ where for both parameters a grid from 1 to 100 is used. The highest $a$ value possible is in this case 100. This will be the optimal value of $a$ but only if the corresponding $b$ value is available. The corresponding $b$ value is $\frac{4}{5} \cdot 100 = 80$. This value of $b$ lays in the grid and hence the optimal values are $a = 100$ and $b = 80$.

However, if a grid for $b$ from 1 to 70 is used. Then the corresponding $b$ value for $a = 100$ is not on the grid. In this case the highest $b$ value is taken, so $b = 70$ together with its corresponding $a$ value, $a = \frac{5}{4} \cdot 70 = 87.5$. So either the maximum $a$ with the corresponding $b$ is taken, or vice versa, the maximum $b$ is taken with the corresponding $a$ value. Depending on the grid.

In figure 8 the highest value of $a$ on the grid is 100 and the corresponding $b$ value is 80 for $\lambda = 8$. As the sample sizes get higher and hence the mll depends more on the data, it is easy to see that this combination of $a$ and $b$ is the optimal one. In figure 9 the maximum $a$ value is 50 and the corresponding $b$ value for $\lambda = 8$ is 40, which can be seen in figure 9(d). Hence, the discussed behavior of $a$ and $b$ are in line with the different color maps.

Based on these observations parameters $a$ and $b$ are suspected to take high values such that the expectation of $\theta$ will be close to the maximum likelihood estimator of $\theta$. This will result in a high value of of the mll however, a new problem arises.

The variance of the Beta distribution is

$$Var(\theta) = \frac{ab}{(a+b)^2 \cdot (a+b+1)}.$$

The higher the parameters are, the lower the variance of the Beta distribution gets. A low variance means that there is a strong belief in the value that $\theta$ takes. The prior will be very concentrated on the maximum likelihood estimator, hence:

$$P(\theta \approx \frac{r}{r+x}) \approx 1.$$

Which automatically yields that:

$$\int_\theta P(x_1, \ldots, x_n|\theta) P(\theta) \, \mathrm{d}\theta \approx P(x_1, \ldots, x_n|\theta\_ML) \cdot 1.$$

Since there were different optimal values of $a$ and $b$ for different kind of data sets, these two parameters should be made free. However, in order to make them free the problem of getting a low variance should be dealt with. To solve this problem there are three potential solution strategies:

1. *Fix $a$*: If $a$ is fixed the corresponding optimal $b$ value, as mentioned before, will be chosen which will give the optimal mll and prevents the prior to extremely peak.

2. *Impose a restriction*: A restriction such as $a + b = constant$ where the constant could get values as $10, 100, 200$, will prevent that the variance will tend to a very small value.

3. *Penalize the sizes of $a$ and $b$*: Instead of choosing a uniform distribution for the hyperpriors, where each $a$ and $b$ value is equally likely, the hyperpriors will be chosen such that $P(a1) < P(a2)$ if $a1 < a2$.

All three strategies would solve the problem. In further research of this thesis strategy (2) will be used when parameters $a$ and $b$ are made free. The behavior of the two parameters is studied by running a program which will make a certain number of iterations. The values of the parameters will start at $a = constant/2$ and $b = constant/2$ and during each iteration it imposes to either state:

$$\begin{array}{lcl} a_{new} = a_{old} + 1 & \quad \text{or} \quad & a_{new} = a_{old} - 1 \\ b_{new} = b_{old} - 1 & & b_{new} = b_{old} + 1 \end{array}$$

each with a probability of a half. The acceptance probability is the ratio of the mll with the old parameters and the mll with the new parameters. After running the program, the values of $a$ and $b$ can be analysed and a clear convergence of both parameters is found. Both parameters converged to the values that were seen in the beginning of the chapter. Namely, the ones making sure that the expectation of the prior is approximately equal to the maximum likelihood estimator of the Negative-Binomial distribution.

In conclusion, with respect to the original model it would be an improvement to make parameters $a$ and $b$ free. However, the variance must not get a very small value. Therefore some kind of restriction must be implemented if the two parameters are set free. Since $a$ and $b$ have different optimal values for each $\lambda$ generated data set, it would be optimal to learn these parameters from the data. This could be done through information exchange.

# 8 Information Exchange

Let $X = (x_1, x_2, ..., x_n)$ be a data set and suppose this data set was given three changepoints at the time points $a$, $b$ and $c$ where $1 < a < b < c < n$. Which divide the data set into 4 segments.

Without information exchange the likeliness of the data in each segment will be calculated with an uninformative prior, namely $\theta \sim Beta(a, b)$ where $a = b = 1$ for each segment. However, information exchange will make sure that the informative priors will be used. This can be done on two different levels: global and sequential.

In this section the two levels will be explained and the different strategies to implement these kind of information exchange into the model will be discussed as well.

## 8.1 Global Information Exchange

When information Exchange on global level is applied, one prior is used for all segments of the data. This can be implemented by making parameters $a$ and $b$ of the prior free. At the end of each MCMC iteration new values for the parameters are proposed, either rejected or accepted and used in the next iteration of the MCMC scheme. In chapter 7 it was concluded that a restriction on $a$ and $b$ is necessary if they are made free. During the simulation studies in chapter 10, this restriction on $a$ and $b$ will be tested for three different values of the constant, namely $10, 100, 200$.

The first iteration will use $a_1 = \frac{constant}{2}$ and $b_1 = \frac{constant}{2}$. At the end of each iteration we either consider:

$$\begin{aligned} a_{i+1} &= a_i + 1 \\ b_{i+1} &= b_i - 1 \end{aligned} \qquad \text{or} \qquad \begin{aligned} a_{i+1} &= a_i - 1 \\ b_{i+1} &= b_i + 1, \end{aligned}$$

where $i$ represents the $i$th iteration of the MCMC scheme.

## 8.2 Sequential Information Exchange

During global information exchange the prior is updated after each MCMC iteration and used for all the data components. However, for information exchange on sequential level the prior must be updated per data segment. The prior for component $i$ is the updated version of the prior of component $i - 1$ after having observed the data of this component. The parameters for the first component will be the parametes for the uninformative prior, hence $a_1 = b_1 = 1$. After that, updated priors will be used.

Let $a_i$ and $b_i$ be the parameters for the prior for component $i$ of the data. The most basic way to update the prior based on observed data is by using the posterior. In chapter 6 the posterior for the Neg-Bin-Beta model was calculated and the parameters were $nr + a$ and $\sum_{i=1}^{n} x_i + b$. Therefore, one way to update the priors is the following way:

$$\begin{array}{ll} a_1 = 1 & b_1 = 1 \\ \mathrm{a}_{i+1} = a_1 + n_i \cdot r & b_{i+1} = b_1 + \sum_{k=1}^{n_i} x_k^{(i)} \end{array}$$

However, this strategy let the parameters $a_{i+1}$ and $b_{i+1}$ depend on the size of data component $i$. To avoid this the parameters should be normalized. This can be done in the following way:

$$\begin{array}{ll} a_1 = 1 & b_1 = 1 \\ \mathrm{a}_{i+1} = a_1 + r & b_{i+1} = b_1 + \frac{\sum_{k=1}^{n_i} x_k^{(i)}}{n_i}. \end{array}$$

This strategy to update the priors might work however, there might be one disadvantage of this approach. There exists a probability that the data that is observed in component $i$ is not of any relevance for component $i+1$. In those cases the uninformative prior must be used instead of the informative prior. This can be done by introducing a coupling parameter $c$. This parameter will make sure that if the observed data is not informative for the new data component, we go back to the uninformative prior $a = 1$ and $b = 1$ for all components $i = 1, ..., k$. Hence,

$$\begin{array}{ll} a_1 = 1 & b_1 = 1 \\ \mathrm{a}_{i+1} = a_1 + r \cdot c & b_{i+1} = b_1 + \frac{\sum_{k=1}^{n_i} x_k^{(i)}}{n_i} \cdot c. \end{array}$$

A low value for $c$ will make sure that the prior parameters take the value of the uninformative prior. The higher $c$ gets, the stronger the belief in the parameters and the lower the variance is. The start value of $c$ will be 1 and at the end of each iteration a new value for $c$ will be proposed. Either $c_{new} = c_{old} + \mu$ or $c_{new} = c_{old} - \mu$ is proposed and rejected or accepted. Here $\mu$ will be uniformly distributed on the interval $[0, 0.1]$.

# 9 Simulation Studies

In this chapter various simulations will be discussed. These simulations are performed to see how the different levels of information exchange influences the model and if it benefits from it. In other words, the simulations that are performed serve as a check if the information exchange on global level, sequential level or both levels are an improvement on the current Negative-Binomial Changepoint model or not.

The simulations work the following way. First two data sets will be generated. Both $X^{(1)}$ and $X^{(2)}$ are generated from the same distribution with equal parameters. Then the model that must be tested will be used to find the changepoints for $X^{(1)}$, those changepoints will be placed at the exact same time points in $X^{(2)}$. Then the mll will be calculated of the second data set. The marginal likelihood of the whole data set equals the product of the marginal likelihood of all components separately and if the log values are taken the mll is equal to:

$mll(X^{(2)}) = \sum_{i=1}^{k} mll(X_i^{(2)})$.

Hence, by observing one data set the mll of the second data set is obtained. Since the goal is to make improvements on the model, the aim is to retrieve higher values of the mll. First both levels of information exchange will be compared to the Negative-Binomial changepoint model. After that a general comparison will be made between the global, sequential and regular Neg-Bin changepoint model.

During the simulations the maximal number of components is set to $K_{MAX} = 10$ for all types of models. For all the models with the Neg-Bin-Beta model $r$ is fixed at 10. Grzegorczyk and Kamalabad have performed a pre-study to determine the required number of MCMC iterations. This pre-study indicated that the following MCMC setting is sufficient for the simulations that are performed in this chapter: The burn-in phase is set to 25.000 MCMC iterations, before $R = 250$ equidistant samples are taken from the subsequent 25.000 MCMC iterations. Hence, 250 steps are taken in the MCMC simulation with each 100 iteration steps.

## 9.1 Synthetic data

To perform the simulations various data sets must be generated. Let $\mathbf{s}_m$ denote a row vector of length $m$, whose elements are all equal to $s \in \mathbb{N}$, $\mathbf{s}_m = (s, ..., s)$. Here $s$ stands for the specific Poisson parameter which was used to generate the data with. Hence, a data set where each element $x_{\lambda=k}$ is generated with a Poisson parameter $\lambda = k$, can be compactly defined as:

$$X = (\underbrace{x_{\lambda=2}, ..., x_{\lambda=2}}_{\text{m-times}}, \underbrace{x_{\lambda=4}, ..., x_{\lambda=4}}_{\text{k-times}}, \underbrace{x_{\lambda=6}, ..., x_{\lambda=6}}_{\text{l-times}}) =: (\mathbf{2}_m, \mathbf{4}_k, \mathbf{6}_l).$$

With regard to the real-world Taxi data, described in section 11.1, where each data matrix $\mathbf{D}$ is built with $T = 96$ columns (time points) and $n \in 1, 2, 4, 8, 16$ rows (independent samples per time point). The simulations on synthetic gathered data sets use data matrices that are also built with a varying number of rows $n \in 1, 2, 4, 8, 16$ and $T = 96$ columns. For each model the procedure of making these data sets is repeated 5 times. In total this sums op to an amount of $5 \cdot 5 = 25$ data matrices per model.

## 9.2 Global information exchange

Before testing this type of information exchange on data sets as described in the previous section, first a check is performed to see if this strategy of information exchange actually works. This is done by creating two different data sets. $X^{(1)} = (10, 10, 10, 10, 1, 1, 1, 1)$ and $X^{(2)} = (10, 10, 10, 10, 10, 10, 10, 10)$. In both data sets a changpeoint was set in the middle, hence after the fourth element and divides the data set into two components. In data set 1 the first component of the data set is very informative for the second component of the data set. For data set 2 this is not the case. To check if the global information exchange forms an improvement on the model as suspected, the mll of both data sets is calculated with both the original Neg-Bin changepoint model and the Neg-Bin changepoint model with global information exchange. The mll was calculated multiple times where each time it took other values for parameters $a$ and $b$ in the global information exchange. There were three different values of the constant that is used in the restriction on the two parameters. The results are given in figure 10.

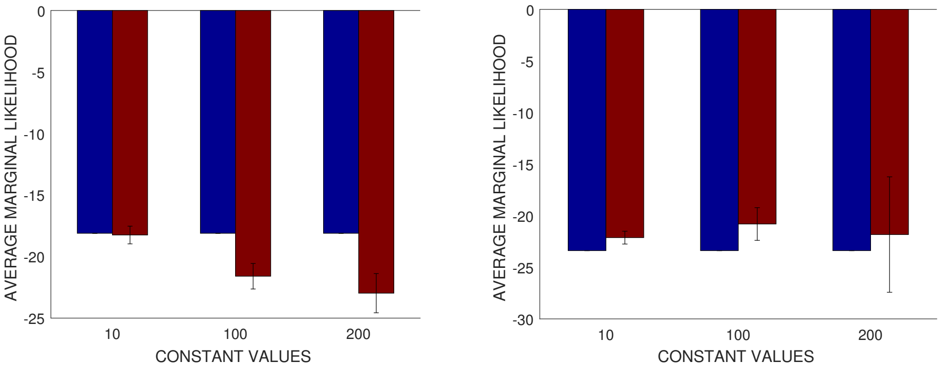


**Figure 10:** Results for various synthetic data sets. The histograms show the average mll (averaged across 25 data sets; i.e. 5 randomly sampled data instantiations for each sample size $n = 1, 2, 4, 8, 16$) for different kind of values for the constant, $a + b = constant$. The first (blue) bars represent the mll for the origanl Neg-Bin-Beta changepoint model. The second (red) bars represent the mll for the Neg-Bin-Beta with global information exchange. The error bars represent the standard deviations. *Left* $X^{(1)}$, *right* $X^{(2)}$ .

The first (*left*) histogram in figure 10 shows the results for data set $X^{(1)}$. The global information exchange turns out not to be an improvement on the model. In fact, it gives worse results for the mll. This is in line with the expectations since the two components of this data set are not informative to each other.

The second (*right*) histogram shows the results for data set $X^{(2)}$. Here the two data components are very informative to each other and as can be concluded from the histogram, the global information exchange forms an improvement to the Neg-Bin changepoint model. Hence, it can be concluded that the strategy of global information exchange works. The next step is to apply it on synthetic data sets described in section 9.1 and perform simulations as discussed in the beginning of this chapter, namely: Two data sets will be taken. With the model that must be tested the changepoints for the first data set are learned and placed in the second data set. After placing the changepoints the mll of this new data set will be calculated. Figure 11 shows the results.

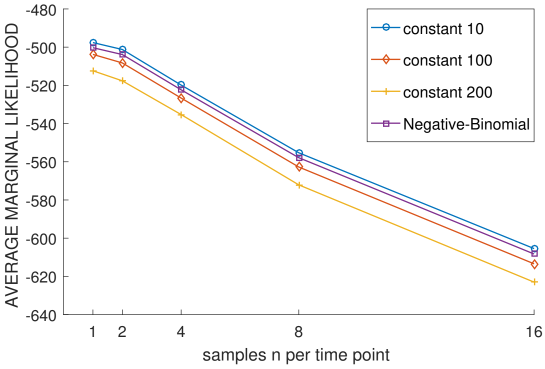


**Figure 11:** Results for various synthetic data sets. The average mll (averaged across 5 data sets) have been plotted against the number of samples $n$ per time point $t$. The five symbols on each line correspond to the values obtained for the sample sizes $n \in 1, 2, 4, 8, 16$. $X^{(i)} = (\mathbf{1}_m, \mathbf{2}_m, \mathbf{4}_m, \mathbf{6}_m, \mathbf{8}_m)$ for $i = 1, 2$.

In figure 11 the Neg-Bin changepoint model is tested and the global Neg-Bin changepoint model with three different kind of values for the constant. They were only tested on a data set in which the number of counts slowly increases, since figure 10 showed that those are the kind of data sets the global information exchange might be an improvement for the model. When the constant is set at 10 it clearly is an improvement. For the other values it is not.

So, the global information exchange can be used to improve the original Neg-Bin changepoint model, however if data sets are analysed that are not smooth, it might lead to worse results. Hence, applying this model requires carefulness.

### 9.3 Sequential information exchange

Again, this level of information exchange requires a test before the simulations can be performed. $X^{(1)} = (10, 10, 10, 10, 1, 1, 1, 1)$ and $X^{(2)} = (10, 10, 10, 10, 10, 10, 10, 10)$ are considered again and the behaviour of the coupling parameters is analysed. The idea behind the coupling parameters is that if the data components are very uninformative, the coupling parameter lets the model go back to the Neg-Bin changepoint model without information exchange, by making the prior parameters equal to $a = b = 1$ again. However, if the data components are informative, the coupling parameters would let the parameters of the prior increase. The results are shown in figure 12, where the behavior of the coupling parameter is plotted.

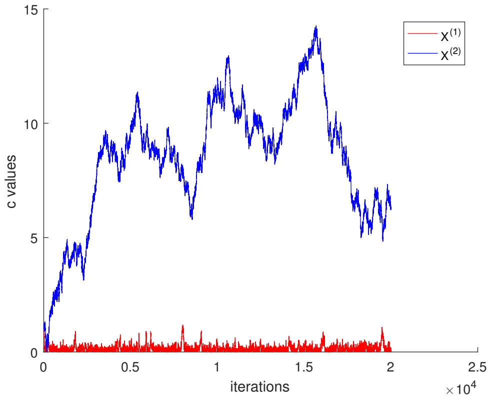


**Figure 12:** The values of the coupling parameters. The values of the coupling parameter have been plotted agianst the number of the corresponding iteration. The blue line shows the behaviour of $c$ in the smooth data set. The red line shows the behaviour of $c$ in the not so smooth data set. The starting value of $c$ was 1. For the given data set 20000 iterations were done. During each iteration a new value for $c$ was imposed and either rejected or accepted.

For data set $X^{(1)}$ where the first data component was uninformative for the second one, the coupling parameter takes values around zero. Hence, for each segment the uninformative parameters $a = b = 1$ are used. In this case the model with information exchange is actually the same as the regular Neg-Bin changepoint model. In the case of data set $X^{(2)}$ the coupling parameter has higher values and therefore the prior parameters for the second component are learned from the first component and the belief in those values is quite high, since high parameters imply a low variance. Therefore, the strategy of sequential information exchange works and can be applied to the synthetic data as discussed in section 9.1.

In the first (*left*) plot in figure 13 a data set with smoothly increasing number of counts has been analysed. Based on the results in figure 12 high values of the coupling parameters are suspected and therefore information exchange between the segments. The plot shows that the Neg-Bin changepoint model with sequential information exchange forms a better fit to the data than the regular Neg-Bin changepoint model. In the second (*right*) plot data sets were used where the number of counts is not smoothly increasing. In this case the coupling parameter makes sure that the uninformative priors are used. This can be seen in the plot since the models perform almost equally well.

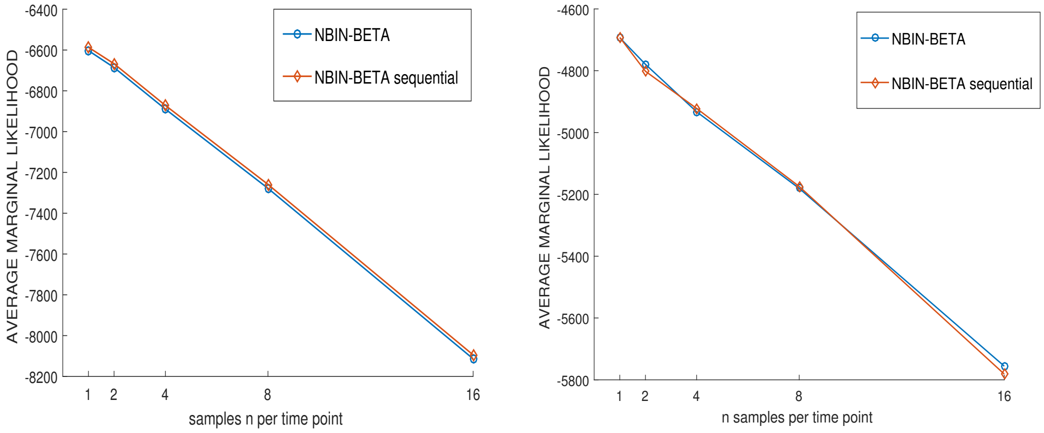


**Figure 13:** Results for various synthetic data sets. The average mll (averaged across 5 data sets) have been plotted against the number of samples $n$ per time point $t$. The five symbols on each line correspond to the values obtained for the sample sizes $n \in 1, 2, 4, 8, 16$. *Left* $X^{(i)} = (\mathbf{1}_m, \mathbf{2}_m, \mathbf{4}_m, \mathbf{6}_m, \mathbf{8}_m)$, *right* $X^{(i)} = (\mathbf{1}_m, \mathbf{5}_m, \mathbf{1}_m, \mathbf{5}_m, \mathbf{1}_m)$ .

### 9.4 Comparison of sequential and global information exchange

From the previous two sections it can be concluded that both levels of information exchange are improvements on the Neg-Bin changepoint model. However, the information exchange on sequential level is saver to use, since it will not perform any information exchange if the data component are uninformative to each other. The global information exchange performs information exchange, even when the data component are uninformative. Hence, this kind of information exchange might be a disadvatage to the model. On sequential level it only forms an improvement to the model if possible and otherwise it is not used. Therefore the information exchange on sequential level is prefered over the global level and this strategy will be used to analyse the real-world taxi pick-up counts.

## 10 Application of the Neg-Bin changepoint model with sequential information exchange

Several changes were suggested for the original Poisson changepoint model and based on the results in chapter 9 the model has been optimized the best when it is transformed in a Negative-Binomial changepoint model with sequential information exchange. Therefore this type of model will be used in this chapter to see how the model can be used and how it analyses the real world taxi data. First the data is defined and after that the data set is analysed with the model, all the probabilities of having a changepoint will be found and a real life interpretation of these changepoints will be given.

## 10.1 The New York city Taxi (NYCT) data from 2013

The University of Illinois (Donovan and Work 2015) stored and published a data set covering information of about 700 million taxi trips in New York City (USA) from the calendar years 2010-2013. In the NYCT database, for each trip various details are provided. For the simulations in this thesis the pick-up dates and daytimes of about 170 million taxi rides in the most recent year 2013 are of particular interest. Each pick-up is considered as a *taxi call*, so that it can be analysed how the number of taxi calls varies over the daytime. The data is summarised in the same way as Grzegorczyk and Kamalabad (2017) did. For more details on this it is recommended to study their paper. Discretising the daytimes into $T = 96$ equidistant time intervales, each covering 15 minutes of the 24 hours day and binning the pick-up times of each individual day into the $t = 96$ time intervals, gives a 355-by-96 matrix $\mathbf{D}$, whose elements $d_{i,t}$ are the number of taxi calls on the $i$th day in time interval $t$, $t \in T$. Since the seven weekdays might show different patterns, the data set matrix $\mathbf{D}$ is subdivided into seven $n_w$-by-T sub-matrices $\mathbf{D}_w(w = 1, ..., 7)$, where $w$ indicates the weekday, and $n_w \in (46, 50, 51, 52)$. The week starts at sunday, i.e. $w = 1$ means it is sunday. $n_w$ represents the number of weekdays in the year 2013. If $n_1 = 46$ it means that there were 46 sundays in the year 2013. For each weekday a random number of $n$ rows is selected from $\mathbf{D}_w$. Hence, data sets such as $\mathbf{D}_{w,n}$ where $w$ is the weekday and $n$ the number of rows randomly selected from this weekday are used.

## 10.2 Results for the taxi data

For each weekday the number of rows were set to 8, hence $\mathbf{D}_{w,8}$ for $w = 1, 2, 3, 4, 5, 6, 7$ are analysed. Each of these data sets have been analysed by the Neg-Bin changepoint model with sequential information exchange. Every time a data set has been analysed by the model an output of 250 vectors is given. Each of these vectors show the locations of the changepoints. Hence, the probability of having a changepoint at location $t$ can be calculated. This is done for each weekday and the results are given in figure 14.

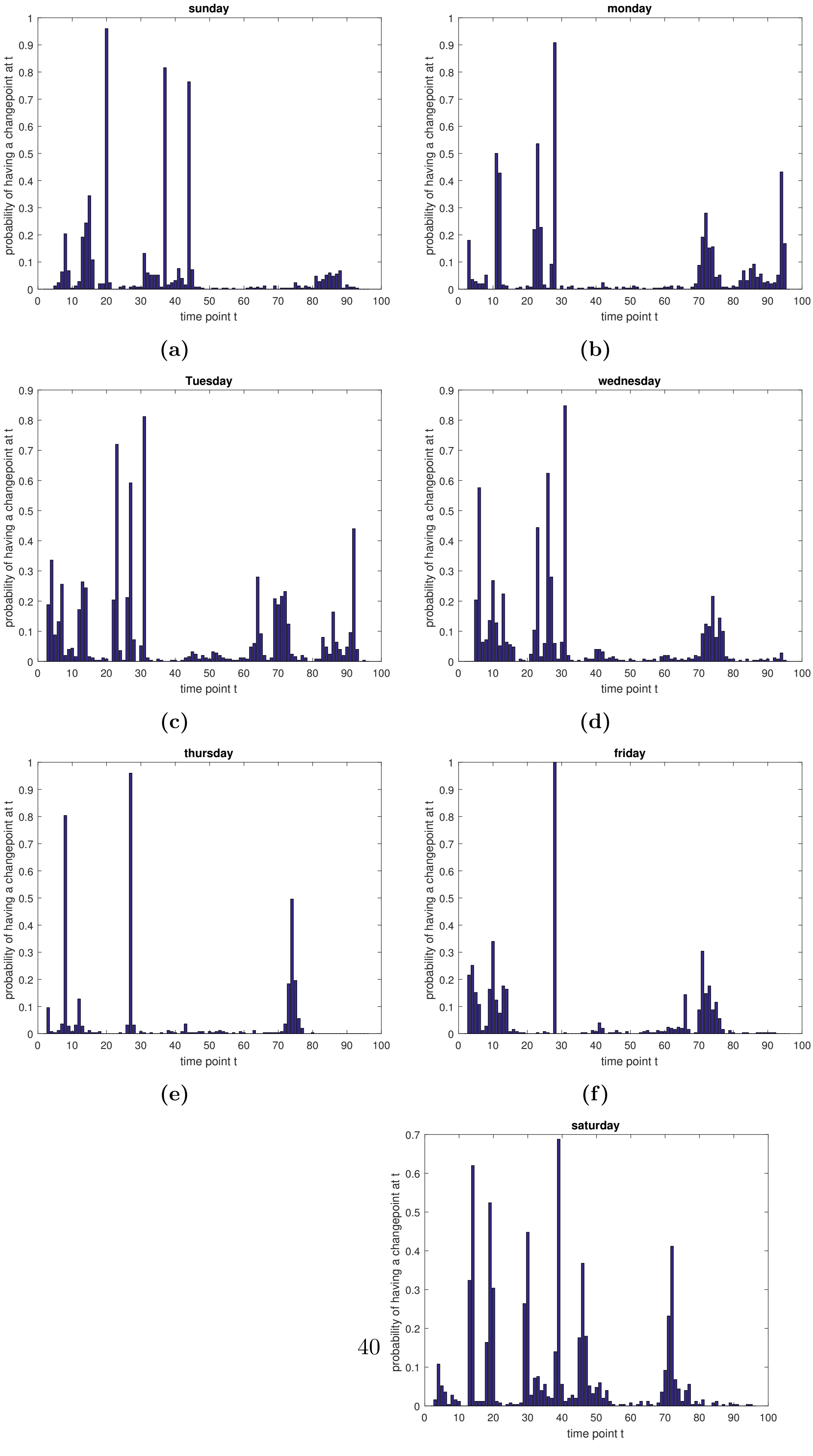
First of all it should be clear that if there are multiple peaks at neighbouring time points, it is most likely that around these points there is one changepoint. This changepoint is slightly shifted each time, therefore the sum of these separate probabilities can be taken and be placed at one corresponding time point.

For the workdays $w = 2, 3, 4, 5, 6$ general trends can be noted. There are four changepoints that are set for each of these days. Located at (or around) 2, 12, 29 and 75. These changepoints correspond with times 00:30,03:00, 07:15 and 18:45. A logical explanation for the locations of these changepoints can be argued. After 00:30 everyone can be at home so the taxi calls would be lower. Then somewhere in the morning the call will slightly increase since the workdays is starting and probably around 7:30 most people in New York already arrived at their work and will not leave until the end of the day. From the changepoints is might be concluded that after 18:30 everyone is going home. However, not all at the same time. The calls have been divided over the next few hours. Therefore a clear changepoint at the end of the evening is not visible. From figure 14 (b) and (c) a changepoint around 92 is detected, 23:00. This could be interpreted as a time where most of the people got home and hence, a decreasing number of taxi calls is made. Figure 14(a) and (g) correspond with the weekend days where changepoints occur between 7:30 and 18:30 as well. In the weekend most people are free of work and hence, more shifts in the frequency of taxi calls during the day are not a surprising result. The clear changepoint that is located around the time point 30 is in the weekend days a bit later. Around 35 or 40 and another changepoint is around 45 or 50. Hence, in the weekend for most people the day starts an hour later which results in probably a busy period between 9:00 and 11:30. The time points interval from 1 to 15 is for each day different. There are no clear indications of changepoints, but multiple time points here have some small probabilities. This might be the result of slow increasing numbers. Therefore, the shift in frequency of the taxi calls is not clear and therefore placed at different times in this interval. However, it makes sense that from some certain moment in the morning people wake up and the number taxi calls start to increase. It is clear where they stop on workdays, namely at 07:15. From here the number of calls clearly decreased a lot.

# 11 Conclusion

In summary, based on the finding of Grzegorczyk and Kamalabad that the Poisson changepoint model was not optimal to analyse a real world taxi data set, this thesis was dedicated to the optimization of this model. And the original Poisson changepoint model that was used as starting point can be improved in multiple ways. An enumeration of all the different improvements will follow.

1. *The Negative-Binomial distribution*: The biggest shortcoming of the original Poisson changepoint model is the fact that is does not deal with over-dispersion. This problem could be solved by replacing the



(a)

(b)

(c)

(d)

(e)

(f)

(g)

**Figure 14:** Vertical bar plots. Probabilities of changepoints for each time point.

Poisson distribution with the Neg-Bin distribution. The Neg-Bin distribution turned out to be as good as the Poisson distribution for the same type of data sets and it performed even better when data sets with over-dispersion had to be analysed. Hence, changing the Poisson changepoint model into a Negative-Binomial changepoint model is an improvement. However, one surprising result was found. In order to let the Negative-Binomial be as good as the Poisson distribution, the parameter $r$ of the Neg-Bin distribution had to be fixed at a value of at least 10. Unfortunately, a mathematical explanation of why this has to be at least 10 is not found yet. But, the test results were clear and it was safely concluded that the model worked when this parameter was fixed. Hence, the first improvement to the model was made.

2. *Free parameters a and b of the Beta prior*: The current Negative-Binomial changepoint model depends on three parameters, namely: $r, a$, and $b$. Parameter $r$ was already fixed at 10. The other two parameters were until now fixed at a value of 1. Some experiments with different values of $a$ and $b$ showed quite fast that they should be made free. For different kind of data sets, different values of $a$ and $b$ turned out to be optimal. However, experiments showed that if both parameters would be made free they automatically will take high values. Such high values that the variance would be too small and hence the prior would be peaked around certain values. In order to prevent this, a restriction on the parameters should be made if they are free. A restriction such as $a + b = constant$ would be a solution to this problem. Thus, making parameters $a$ and $b$ free and give them a restriction such as $a + b = constant$ would be a second improvement to the model.

3. *Global information exchange*: In the current Negative-Binomial changepoint model the goal is the find the optimal number of changepoints and their related locations. In order to do this all components of the data will be analysed with a Beta distribution with parameters $a$ and $b$. However, from the experiments on the parameters $a$ and $b$ it was concluded that the values of those parameters could be learned from the data. The concept of global information exchange is that at the end of each iteration of the MCMC scheme, new values of parameters $a$ and $b$ were imposed and either accepted if they were improvements on the models fit, or rejected if not. Those new values of the prior parameters are used for all the data components and hence, it is called global information exchange. The simulations in chapter 9 showed that when the constant was set at 10, the global information exchange formed an improvement to the Neg-Bin changepoint model.

4. *Sequential information exchange*: In global information exchange during each iteration new values for parameters $a$ and $b$ are imposed and used for all the data components. However, this kind of information exchange

can also be implemented on a sequential level. When a data component is observed the prior parameters will be updated based on the observed data. The updated prior will then be used for the next data component. Hence, the information exchange is carried out on sequential level. To update the priors based on the data of the previous component, the posterior is used. But the prior parameters should be updated a bit further, namely with a coupling parameter. The coupling parameter will make sure that the information exchange will only happen if the data components are actually informative to each other. If not, then the coupling parameter will make sure that the basic uninformative prior is used, the one from the original model where $a$ and $b$ were fixed at 1. This resulted in the following prior parameters $a_i$ and $b_i$ for component $i$: $a_{i+1} = a_1 + r \cdot c$ and $b_{i+1} = b_1 + \frac{\sum_{k=1}^{n_i} x_k^{(i)}}{n_i} \cdot c$. From the simulation studies in chapter 9 it was concluded that this type of information exchange was an improvement the model. Moreover, it is safe to implement it in the Neg-Bin changepoint model, since it will not harm the model in situations where the model does not benefit from information exchange, since the prior parameters will be near 1 in that case.

Based on these four ways to improve the model it was concluded that the Negative-Binomial changepoint model with sequential information exchange is at this point the optimization of the Poisson changepoint model. This is due to the fact that the global information exchange could actually make the model a worse fit to the data than the Negative-Binomial changepoint model would be. Since the information exchange is always executed, even if the data components are not informative to each other. The sequential information exchange uses the coupling parameter to control this and is therefore the best way to improve the Poisson changepoint model.

So the most optimal model has been used to analyse the real world taxi data. For each weekday, each time point during the day was given a probability of having a changepoint there. For some time points there were changepoint detected with probabilities of almost one. Some conclusions could be drawn from the taxi calls. For example, during the workdays a clear shift in the frequency of taxi calls was denoted at a time of 07:15. Another shift was visible around 18:15. This clearly indicates that during the working hours of most people the taxi calls probably had a lower frequency than before 07:15 and after 18:15. In the weekend there were more shifts in frequency of taxi calls during the day, which makes sense since more people are free during the weekends and therefore more movement in the city during the day.

So, the optimization of the Poisson changepoint model has lead to the Negative-Binomial changepoint model with sequential information exchange, which can be applied to real world data sets in which over-dispersion may occur.

# 12 Discussion

The optimization of the Poisson changepoint model provided in this thesis lays still open for more improvements. Unfortunately this thesis has a deadline and due to the time pressure not all the scheduled experiments have been executed. Therefore, in this section some ideas for further research are proposed.

First of all one may search for the mathematical explanation behind the need to fix parameter $r$ in the Neg-Bin distribution. There is a lot of literature that mention the Neg-Bin as a good replacement for the Poisson distribution however, nowhere is mentioned what should be done with parameter $r$.

Secondly, different strategies for both global and sequential information exchange might be investigated. For the sequential information exchange a quite simple coupling parameter is used at this point. The coupling parameter is given a certain value and used for all the components. Perhaps it might be possible to make a coupling parameter for each segment separately. In the simulations studies data sets such as $X^{(i)} = (\mathbf{1}_m, \mathbf{5}_m, \mathbf{1}_m, \mathbf{5}_m, \mathbf{1}_m)$ have been analysed. However, as a whole the data components should not be seen as informative to each other. However, the first component is informative to the second one and the third one to the fourth. Hence, maybe some sort of strategy can be formulated which makes sure that the coupling parameter is determined per segment not per data set as a whole.

For global information exchange the parameters were made free and were given some kind of restriction. At each iteration parameter $a$ either increases with one and $b$ decreased with 1 or the other way around. However, perhaps the values of these parameters could already been centered around certain values, which will make sure that those values are found faster and used more often during the MCMC iterations. Or maybe the posterior can be used to update the priors. Maybe for all segments the updated priors are determined and the average of all these priors can be taken and globally used.

In short, there are many more ways to implement sequential and global information exchange. It might be that one strategy of doing so is better than the other. Therefore the strategies of implementing the information exchange should be explored further before it can be concluded that this model is an optimal version.

**Acknowledgements**

I would like to thank Marco Grzegorczyk for all the help during the writing of this thesis. Our weekly meetings have been beneficial to my understanding of the material and were useful to decide which directions this thesis had to take. Furthermore, I would like to thank Wim Krijnen for being my second

supervisor and taking the time to read and evaluate this work.

# 13 Bibliography

[1] Grzegorczyk, M. (2017). Comparative evaluation of various frequentist and bayesian non-homogeneous poisson counting models. Computational Statistics, 32(1), 1-33.

[2] Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711-732.

[3] Bolstad, W. (2017). Introduction to bayesian statistics (Third edition. ed.). Hoboken, N.J.: John Wiley 1-13.

[4] Savage, L. (1962). The foundations of statistical inference (Methuen's monographs on applied probability and statistics). London: Methuen.