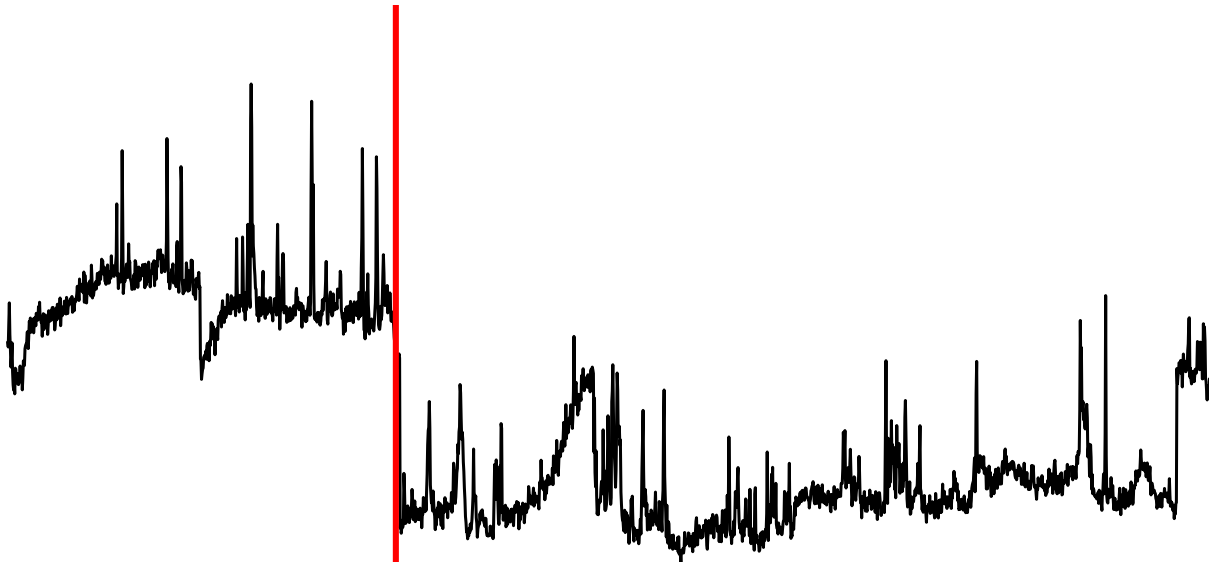

Change point detection in financial time series in connection to purchase behaviours

Master's thesis in Engineering Mathematics and Computational Science

HANNA SKYTT

DEPARTMENT OF MATHEMATICAL SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2022
www.chalmers.se

Master's thesis 2022

Change point detection in financial time series in connection to purchase behaviours

HANNA SKYTT

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden 2022

Change point detection in financial time series in connection to purchase behaviours
HANNA SKYTT

Supervisor: Moritz Schauer, Department of Mathematical Sciences
Examiner: Umberto Picchini, Department of Mathematical Sciences

Master's Thesis 2022
Department of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Time series with marked change point. Source: Time Series Data Library

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2022

Change point detection in financial time series in connection to purchase behaviours
HANNA SKYTT
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

Understanding purchase behaviours of individuals is of interest when the goal is to inspire people to make more environmentally friendly choices. A company with these aspirations is Svalna AB. They have created an app that uses a carbon calculator to give an insight into greenhouse gas emissions based on financial transactions. The aim of this thesis has been to investigate purchase behaviours by comparing the underlying distributions before and after a change point has occurred. This thesis has focused on change point detection in time series using the Metropolis-Hastings algorithm. The model, which has been implemented from scratch, has been tested on well-behaved simulated time series and can accurately find a change point. It has then been used to investigate some specific cases in financial time series provided by Svalna. The results from testing on the simulated time series show a promising start and it is concluded that the overall method is a possibility to investigate the underlying distributions of financial time series.

Keywords: time series, change point detection, bayes, metropolis-hastings, purchase behaviours.

Acknowledgements

I would like to thank my supervisor Moritz Schauer for his support and discussions, and Umberto Picchini for examining my thesis. A big thank you to David Andersson and Ross Linscott at Svalna for giving me this opportunity. I also want to thank my parents and Johan for always being there for feedback and encouragement.

Hanna Skytt, Gothenburg, 2022

Contents

List of Figures

List of Tables

1
Introduction

Change point detection has been a subject of investigation since the 50s [1], with applications within many different subjects, such as dynamical systems [2] or climate changes [3].
The startup company Svalna AB in Gothenburg has developed a technology that uses financial data to estimate greenhouse gas emissions associated with consumption and spending. Their aim is to inspire individuals and companies to make more environmentally friendly choices in regards to their spending. The company states their mission as "By inspiring citizens to change, helping businesses transform, and supporting research about climate-friendly lifestyles, we contribute to creating a sustainable world for ourselves and future generations."[4].
Svalna has a free app that is accessible for individuals. When creating a profile in the app, there are some questions to answer in regards to emissions, with categories such as transport, living situation and food. There is also the possibility to connect to the user's bank account, and with that, give permission to Svalna to use the account data to make emission calculations. Additionally, Svalna is conducting research in cooperation with universities, such as Chalmers University of Technology [5]. Individuals can give consent to their data being used in that research. It is the financial data from the individuals who have given their consent that has been used in this project.

1.1 Aim

The overall aim of this project is to investigate purchase behaviours of individuals before and after an event, a change point in a time series. To specify, investigating the similarities and differences between parameter values of statistical distributions before and after a change point, which includes being able to detect the change point. A possible future continuation of the project is to use the potential similarities to be able to make suggestions in the app for users to make more sustainable choices regarding their spending.

1.2 Limitations and scope

The limitations of this thesis are based on the model implementation (see Section 3.1) along with the time frame of the thesis. The limitations as to how well the questions can be answered lies both in how well the implementation works for simulated

data and in the provided financial data. It also depends on the quality and information content of the data, if there is the possibility to make statistically meaningful conclusions. One of the challenges is that the data could contain outliers that give no meaningful information about a change in behaviour. An example could be that an individual suddenly has a huge payment for a flight, but that might not indicate that they will take a trip every following month.

2
Background

2.1 Definitions and notations

This section describes the different definitions and notations that are used in this thesis. It presents what time series are and gives some real life examples, along with the relevant parts of Bayesian inference.

2.1.1 Time series

The definition of a *time series*, as described in [6], is a set of observations, x_τ, with each observation being recorded at time τ. If the set of observation times is of a discrete nature, for instance where the observations are taken at fixed time intervals, it is a discrete-time time series. If the observations are taken continuously, it is a continuous-time time series. Examples of discrete-time time series can be seen in figure 2.1. Time series can be found in many different areas, the examples show one time series connected to the annual sales of a product, and the other one is connected to temperature changes.

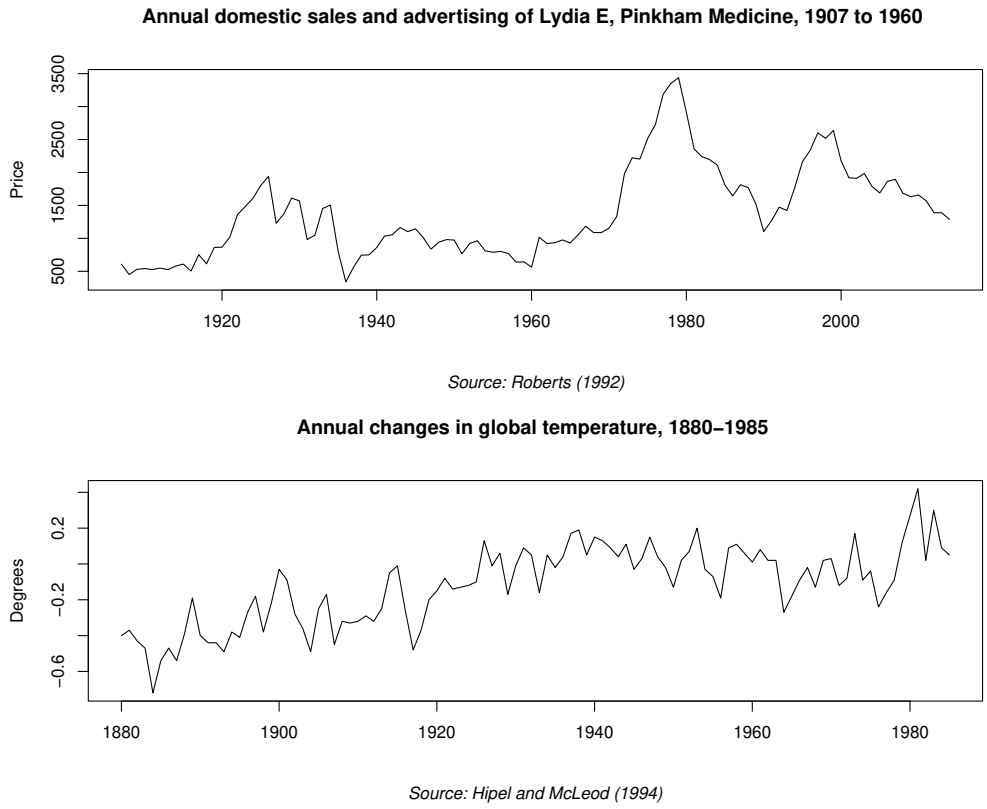

Figure 2.1: Two different examples of time series, both taken from the Time Series Data Library, `tsdl`. [7]

Time series data can be described by time series models that have marginal distributions describing different aspects of the data, such as the mean and variance. There can also be a distribution describing the times of occurrences as opposed to the actual values. A *change point* in a time series is when an abrupt change happens to the underlying distribution. A set of n change points, $S_t = [t_1, \ldots, t_n]$, divides a time series with m points into different segments, $S = [S_1 = \{x_1, \ldots, x_{t_1-1}\}, \ldots, S_{n+1} = \{x_{t_n}, \ldots, x_m\}]$. *Change point detection* is widely used within different areas to find these change points. This change could be a change in the mean or standard deviation of the data, as well as the trend. An example of a time series with distinct change points can be seen in figure 2.2, where the change points are marked by red lines. It shows a measure of radioactivity in the ground at 2 hourly intervals, with a lower radioactivity when the ground is covered by snow. The marginal distribution has a clear change in the mean. For given time series data, we wish to construct time series models that can describe the data. One way of doing this is by using the principles of Bayesian inference.

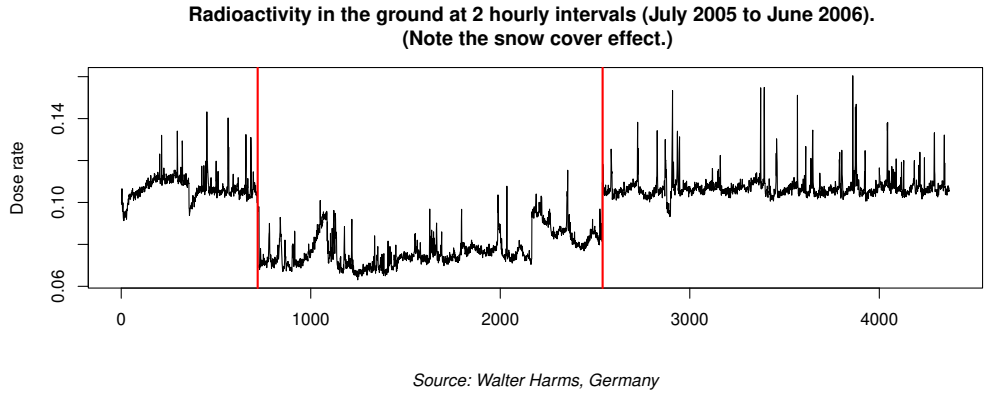

Figure 2.2: Example of a time series with notable change points marked by red lines, taken from the Time Series Data Library, `tsdl`.

2.1.2 Bayesian inference

The following section introduces the relevant theory in Bayesian inference using the notations found in [8, p. 6-7].
Bayesian inference can be used to make conclusions about a parameter θ, conditional on the observed value y. It can be useful to determine if some parameter values are more likely to describe the data than other parameter values, while taking prior information into account. The differentiating thing between Bayesian inference and frequentist inference is that Bayesian inference uses probabilistic assumptions on the parameters. For a time series data y and some parameters θ, the *sampling distribution* $p(y|\theta)$ can be constructed. Using this along with the *density of the prior distribution*, $p(\theta)$, the *joint probability density function* can be written as

$$p(\theta, y) = p(\theta)p(y|\theta). \tag{2.1}$$

This expression is used in Bayes' rule that gives the *posterior density*,

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}, \tag{2.2}$$

with $p(y) = \sum_{\theta} p(\theta)p(y|\theta)$ for the case where θ is a discrete random variable, and $p(y) = \int p(\theta)p(y|\theta)d\theta$ for the case where θ is a continuous random variable. Since it is independent of θ for a fixed value of y, the density $p(y)$ can be seen as a constant, giving rise to the *unnormalized posterior density*,

$$p(\theta|y) \propto p(\theta)p(y|\theta). \tag{2.3}$$

In order to find the posterior density and being able to perform statistical analysis of θ, which is often the goal, the model $p(\theta, y)$ is developed and appropriate computations are done to find and present $p(\theta|y)$. There are some distributions that are conjugate distributions, meaning that the prior distribution and the posterior distribution belong to the same class of parameterized distributions. For many distributions there are no easy conjugate distribution to perform calculations on. Therefore, it can be more efficient to perform sampling methods of the posterior distribution. One method, the Metropolis-Hastings algorithm, is described in more detail in section 2.2.

2.2 Metropolis-Hastings algorithm

This section presents the Metropolis-Hastings algorithm with definitions and notations from [8, p. 275-280]. The Metropolis-Hastings algorithm is useful for sampling from the Bayesian posterior distribution, $p(\theta|y)$. It is a Markov chain Monte Carlo (MCMC) method, that draws values of θ from an approximate distribution, and then correcting to draw the next sample from a better approximation of the target posterior distribution. The samples are drawn sequentially and each draw's distribution is only dependent on the previous draw's value, hence, they form a Markov chain.

The Metropolis-Hastings algorithm starts by drawing the starting point θ^0 from a *starting distribution* with density $p_0(\theta)$, where $p_0(\theta) > 0$. Then it performs the following steps for each time step i, for a number of iterations n.

1. Firstly, the algorithm samples a proposal θ^* from a *proposal density* at time i, $J_i(\theta^*|\theta^{i-1})$. The proposal density should be constructed in a way that it is easy to sample θ^* for any θ. The density should also provide a good trade-off between exploration and exploitation of the space.
2. Secondly, it calculates a *ratio of ratios*,

$$r = \frac{p(\theta^*|y)/J_i(\theta^*|\theta^{i-1})}{p(\theta^{i-1}|y)/J_i(\theta^{i-1}|\theta^*)}.$$

3. Lastly, it accepts the proposal θ^* with probability $\min(r, 1)$ and rejects with probability $1 - \min(r, 1)$. This is equivalent to sampling from a uniform distribution, $U(0, 1)$, and accepting if the sampled value is smaller than $\min(r, 1)$. If it is accepted, $\theta^i = \theta^*$, and if it is rejected, $\theta^i = \theta^{i-1}$. It will always be accepted if the posterior density increases, meaning that $r \geq 1$, and the proposal moves closer to the target.

This type of iterative sampling is useful for when direct sampling from the posterior distribution is not possible. If working with a re-scaled target density, the constants will cancel out one another, see equation 2.3. The ratio of ratios is calculated and used in the acceptance stage in a way that it will always accept θ^* if it will increase the posterior density, but only sometimes accept if it will decrease the density. Even if the value is rejected, the algorithm still counts it as an iteration. The algorithm is summarized in Algorithm 1. The model will use this algorithm as a basis of sampling from the posterior distribution, as well as some parameter values describing different aspects of a time series. The model is described in detail in Section 3.1.

Algorithm 1 Metropolis-Hastings Algorithm

```
function MH(y)
    θ^0 ← p_0(θ)
    for i = 1, 2, ..., n do
        θ* ← J_i(θ*|θ^{i-1})
        r ← [p(θ*|y)/J_i(θ*|θ^{i-1})] / [p(θ^{i-1}|y)/J_i(θ^{i-1}|θ*)]
        a ← U(0, 1)
        if a < r then
            θ^i ← θ*
        else
            θ^i ← θ^{i-1}
        end if
    end for
    return (θ^0, ..., θ^n)
end function
```

3
Method

This chapter will first present the model for change point detection, along with the assumptions that are made, the explicit expressions and functions that are used, and how the model is implemented. The different tests for the simulated time series and the financial time series from Svalna are described in detail in Section 3.2.

3.1 Model

This section describes how the model used to find the change point and relevant parameters is constructed. The input and output is summarized in Table 3.1.

Table 3.1: Descriptions of the input and output of the model.

Input		Output	
Symbol	*Description*	*Symbol*	*Description*
τ	Observation times	t	Change point
X	Observation values	μ	Parameter values describing observation times
		(α, β)	Parameter values describing observation values

3.1.1 Assumptions on the data

This model focuses on finding a change point in a time series, along with parameters describing the observation values before and after the change point. They are used to compare how the behaviours of the observations change. This section describes the assumptions that are made about the data and gives the explicit expressions that are used later in the implementation.

The only known values connected to a time series are the observation times, $\tau \in \mathbb{R}_{\geq 0}$, and the observation values, $X \in \mathbb{R}_{\geq 0}$. The beginning of the time span of the time series is assumed to be 0, and the end is denoted by T.

The first thing that is assumed is that there is a change point, $t \in [0, T]$, in the time series. (This assumption will be relaxed to include the possibility of assuming no change point later in the thesis, and is described in more detail in Section 3.3.) The change point t is assumed to have a higher probability of being closer to the middle of the time series, than to the ends. Therefore, it is assumed to have a truncated normal distribution, $t \sim N_{[a,b]}(\mu, \sigma)$. Different truncated normal distributions are shown in Figure 3.1, and the values $\mu = \frac{T}{2}$ and $\sigma = \frac{T}{3}$ are chosen to cover the given time span $[0, T]$.

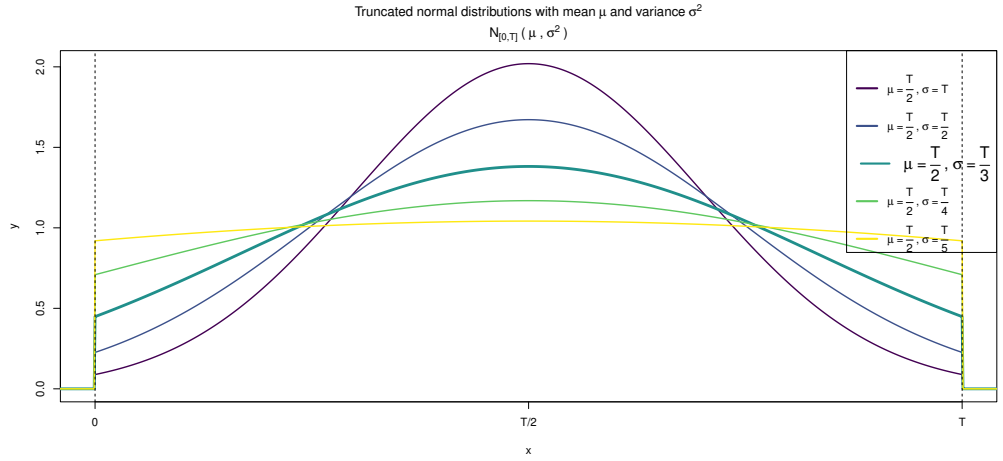

Figure 3.1: Plots of the truncated normal distribution for different parameter values, here $T = 1$. The third plot (dark green) shows the plot for the chosen parameter values.

The frequencies of transactions/purchases are interesting to compare if there is a difference before and after a change point. Therefore, for the number of observations before and after the change point, I_1 and I_2 respectively, the followings assumption are made:

$$\begin{aligned} I_1 &\quad \text{Pois}(\mu_1 \cdot t), \\ I_2 &\quad \text{Pois}(\mu_2 \cdot (T - t)), \end{aligned} \tag{3.1}$$

with some prior assumptions on the parameter $\mu = (\mu_1, \mu_2)$,

$$\log(\mu_1) \quad \text{N}(\mu_0, \sigma_0^2), \tag{3.2}$$
$$\log(\mu_2) \quad \text{N}(\mu_0, \sigma_0^2). \tag{3.3}$$

The observation values, X, the actual transactions or purchases, are assumed to come from the following distribution,

$$X_i \quad \begin{cases} \text{Gamma}(\alpha_1, \beta_1), & \tau_i < t, \\ \text{Gamma}(\alpha_2, \beta_2), & \tau_i \geq t, \end{cases} \tag{3.4}$$

with prior distribution on the Gamma parameter $\beta = (\beta_1, \beta_2)$,

$$\log(\beta_1) \quad \text{N}(\mu_0, \sigma_0^2), \tag{3.5}$$
$$\log(\beta_2) \quad \text{N}(\mu_0, \sigma_0^2). \tag{3.6}$$

The Gamma parameter $\alpha = (\alpha_1, \alpha_2)$ will not be estimated by the Metropolis-Hastings algorithm, instead it uses a plug-in estimator in each iteration. From the mean and variance of the Gamma distribution seen in Table A.1, the following equation can be derived,

$$\left(\frac{\alpha}{\mathrm{E}(\theta)}\right)^2 = \frac{\alpha}{\mathrm{var}(\theta)}, \tag{3.7}$$

$$\alpha = \frac{(\mathrm{E}(\theta))^2}{\mathrm{var}(\theta)}. \tag{3.8}$$

The approximation of α for each iteration is chosen to be the following:

$$\alpha_1 = \frac{(\mathrm{mean}(X_{\tau<t}))^2}{\mathrm{var}(X_{\tau<t})}, \tag{3.9}$$

$$\alpha_2 = \frac{(\mathrm{mean}(X_{\tau\geq t}))^2}{\mathrm{var}(X_{\tau\geq t})}. \tag{3.10}$$

If the variance is 0, or if there is only one data point, the α values are set to mean(X) if it is the first iteration, and for the following iterations it is set to the previous value of α.

3.1.2 Functions

Prior density function

Given the assumptions of the parameter values t, μ, and β from the previous section, Section 3.1.1, and the density functions from Tables A.1 and A.2, the prior function can be written as

$$p(t, \mu_1, \mu_2, \beta_1, \beta_2) = p_t(t) \cdot p_{\mu_1}(\mu_1, \mu_0, \sigma_0) \cdot p_{\mu_2}(\mu_2, \mu_0, \sigma_0) \cdot p_{\beta_1}(\beta_1, \mu_0, \sigma_0) \cdot p_{\beta_2}(\beta_2, \mu_0, \sigma_0), \tag{3.11}$$

with the following priors connected to each parameter:

$$\begin{aligned}
p_t(t) &= \frac{1}{\sqrt{2\pi} \cdot \frac{T}{3} \cdot Z} \cdot \exp\left[-\frac{1}{2 \cdot \left(\frac{T}{3}\right)^2}\left(t - \frac{T}{2}\right)^2\right], \\
p_{\mu_1}(\mu_1, \mu_0, \sigma_0) &= \frac{1}{\sqrt{2\pi}\sigma_0} \cdot \exp\left[-\frac{1}{2\sigma_0^2}(\log(\mu_2) - \mu_0)^2\right], \\
p_{\mu_2}(\mu_2, \mu_0, \sigma_0) &= \frac{1}{\sqrt{2\pi}\sigma_0} \cdot \exp\left[-\frac{1}{2\sigma_0^2}(\log(\mu_2) - \mu_0)^2\right], \\
p_{\beta_1}(\beta_1, \mu_0, \sigma_0) &= \frac{1}{\sqrt{2\pi}\sigma_0} \cdot \exp\left[-\frac{1}{2\sigma_0^2}(\log(\beta_2) - \mu_0)^2\right], \\
p_{\beta_2}(\beta_2, \mu_0, \sigma_0) &= \frac{1}{\sqrt{2\pi}\sigma_0} \cdot \exp\left[-\frac{1}{2\sigma_0^2}(\log(\beta_2) - \mu_0)^2\right].
\end{aligned} \tag{3.12}$$

The value Z is used as a normalizing factor in the density function for the truncated normal distribution with limits A and B, where $Z = \Phi\left(\frac{B-\mu}{\sigma}\right) - \Phi\left(\frac{A-\mu}{\sigma}\right)$. $\Phi(x)$ is the cumulative distribution function for the standard normal distribution.

Likelihood function

From the assumptions of the observation values X and times τ described in Section 3.1.1, and the density functions from Tables A.1 and A.2, the likelihood function can be written as

$$p(X, \tau | t, \mu, \alpha, \beta) = \prod_{i,\tau_i < t} \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} X_i^{\alpha_1 - 1} e^{-\beta_1 X_i} \cdot \prod_{i,\tau_i \geq t} \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)} X_i^{\alpha_2 - 1} e^{-\beta_2 X_i} \cdot \frac{1}{I_1!}(\mu_1 t)^{I_1} e^{-(\mu_1 t)} \cdot \frac{1}{I_2!}(\mu_2 (T - t))^{I_2} e^{-(\mu_2 (T-t))}. \quad (3.13)$$

Posterior density function

As seen in the theory of Bayesian inference, see Section 2.1.2, the posterior density function is proportional to the product of the prior and the likelihood,

$$p(t, \mu, \beta | X, \tau, \alpha) \quad p(t, \mu, \beta) \cdot p(X, \tau | t, \mu, \alpha, \beta). \quad (3.14)$$

Proposal distribution

The proposal distribution is chosen to be where the candidate θ is sampled from a normal distribution with the previous value θ^{t-1} being the mean,

$$J_t(\theta | \theta^{t-1}) \quad N(\theta^{t-1}, \ ^2). \quad (3.15)$$

3.1.3 Parameter values

There are two parameters in the model that are connected to the prior assumptions on the data, μ_0 and σ_0, and one parameter in the proposal distribution, .

μ_0 and σ_0 are parameters to the normal distribution, and the values for $\log(\beta)$ and $\log(\mu)$ are assumed to be samples from the same normal distribution, $N(\mu, \sigma_0)$. Figure 3.2 shows some plots of the lognormal distribution using di erent parameter values. The chosen values are $\mu_0 = 3$ and $\sigma_0 = 0.6$.

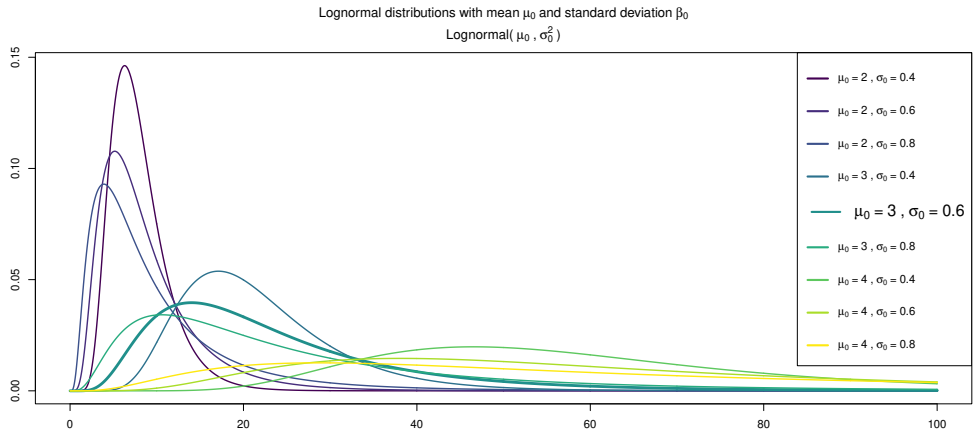

Figure 3.2: Plots of the Lognormal distribution for different parameter values. The fifth plot shows the plot for the chosen parameter values.

The parameter is used in the proposal distribution as a measure of the step size used in generating the new candidate, θ , for each iteration. The step size is an important part in whether or not the candidate is accepted. Too small of a step and it is more likely that more candidates will be accepted, which are not substantially different from the current value. Too large of a step might result in the opposite, too few candidates are likely to be accepted. A good measurement to aim at is an acceptance rate of 23% [9]. Table 3.2 shows a comparison of the mean acceptance rate using different values of . From the table, the chosen value is = 0.07.

Table 3.2: The mean acceptance rate from 10 runs, for time series using the parameters from parameters #13 (found in Table B.1), comparing different values of the step size parameter .

	0.01	0.05	0.07	0.1
Mean acceptance rate	0.70176	0.36739	0.21895	0.14437

3.1.4 Implementation

For the implementation of the model described in the previous sections, the chosen programming language is R [10]. R is useful for statistical computing and graphics [11], as well as being effective at data handling, and it can perform calculations on matrices.

In the implementation of the model, the function to find a change point in a time series is based on the assumptions made in Section 3.1.1, the functions described in Section 3.1.2 with parameters in Section 3.1.3.

The function `Sampling` seen in Algorithm 2, gives the outline of the function used to calculate the random walk based on the Metropolis-Hastings algorithm described in Section 2.2. The main difference is that all the explicit functions are implemented as the logarithm of the functions. This is due to the fact that some of the functions

give too high values to be able to compute, but the logarithm of the same value can be used in computations. The input is a time series consisting of observation times τ, observation values X, and the number of iterations n. The output is a matrix, θ, of the values $(t, \mu_1, \mu_2, \alpha_1, \alpha_2, \beta_1, \beta_2)$ found in each iteration.

Algorithm 2 Metropolis-Hastings sampling implementation (with a plug-in method)

function Sampling(τ, X, n)
  t_1 runif(min(τ), max(τ))
  log($\mu_{1,1}$) rnorm(μ_0, σ_0)
  log($\mu_{2,1}$) rnorm(μ_0, σ_0)
  log($\beta_{1,1}$) rnorm(μ_0, σ_0)
  log($\beta_{2,1}$) rnorm(μ_0, σ_0)
  if var($X_{\tau<t_1}$) is NA or var($X_{\tau<t_1}$) = 0 then
    $\theta_{\alpha_{1,1}} \leftarrow$ mean($X_{\tau<t_1}$)
  else
    $\theta_{\alpha_{1,1}} \leftarrow$ mean($X_{\tau<t_1}$)^2/var($X_{\tau<t_1}$)
  end if
  if var($X_{\tau\geq t_1}$) is NA or var($X_{\tau\geq t_1}$) = 0 then
    $\theta_{\alpha_{2,1}} \leftarrow$ mean($X_{\tau\geq t_1}$)
  else
    $\theta_{\alpha_{2,1}} \leftarrow$ mean($X_{\tau\geq t_1}$)^2/var($X_{\tau\geq t_1}$)
  end if
  $\theta_1 \leftarrow (t_1, \mu_{1,1}, \mu_{2,1}, \alpha_{1,1}, \alpha_{2,1}, \beta_{1,1}, \beta_{2,1})$
  for $i = 1 : n$ do
    $\theta^{\square} \leftarrow$ log-proposal(θ_i)
    if var($X_{\tau<\theta_t}$) is NA or var($X_{\tau<\theta_t}$) = 0 then
      $\theta_{\alpha_1} \leftarrow$ mean($X_{\tau<t}$)
    else
      $\theta_{\alpha_1} \leftarrow$ mean($X_{\tau<t}$)^2/var($X_{\tau<t}$)
    end if
    if var($X_{\tau\geq\theta_t}$) is NA or var($X_{\tau\geq\theta_t}$) = 0 then
      $\theta_{\alpha_2} \leftarrow$ mean($X_{\tau\geq\theta_t}$)
    else
      $\theta_{\alpha_2} \leftarrow$ mean($X_{\tau\geq\theta_t}$)^2/var($X_{\tau\geq\theta_t}$)
    end if
    $r_1 \leftarrow$ log-posterior($\theta^{\square}$, τ, X)
    $r_2 \leftarrow$ log-posterior(θ_i, τ, X)
    $r_3 \leftarrow$ log-proposal-density(θ_i, $\theta^{\square}$)
    $r_4 \leftarrow$ log-proposal-density($\theta^{\square}$, θ_i)
    $r \leftarrow r_1 - r_2 + r_3 - r_4$
    if runif(1) < exp(r) then
      $\theta_{i+1} \leftarrow \theta^{\square}$
    else
      $\theta_{i+1} \leftarrow \theta_i$
    end if
  end for
  return θ
end function

The function Find Parameters seen in Algorithm 3, gives the outline of what calculations are done on the found θ from the Sampling function. It takes θ as input, as well as the number of iterations n, and a warm-up value w. A warm-up value [8, p. 282] is used to discard the first iterations, since the simulation may need some time to reach the posterior distribution. By the same reference, the warm-up value is chosen to be half of the number of iterations, $w = \frac{n}{2}$. The found value of each parameter is set to the mean of the relevant values.

Algorithm 3 Find estimated parameters

```
function Find parameters(θ, n, w)
    θ̂ ← mean(θ[w : n])
    return θ̂
end function
```

The result from running this model on a time series will be a vector of the estimated values, $\hat{\theta}$. Along with the values, a figure is generated for each time series to show where the estimated change point is. It shows the time series, the histogram from the relevant values of t from the Markov chain, along with the estimated change point. The figure also includes trace plots for the parameters that are estimated using the Metropolis-Hastings algorithm: t, μ and β.

3.2 Time series

3.2.1 Simulating time series

In order to test the model before applying it on the financial data, some simulated data was constructed. This section describes how the simulated data was constructed, with descriptions of the parameters, variables and distributions used. Table 3.3 describes the different parameters used in simulating a time series, as well as the variables that are simulated.

Table 3.3: Descriptions of the parameters and variables used for simulating a time series.

Parameters		Variables	
Symbol	*Description*	*Symbol*	*Description*
T	Total time	t	Change point
(α, β)	Parameters for the Gamma distribution	τ	Observation times
μ	Expected frequency of the observation times	X	Observation values

For a chosen total time T, a change point is simulated to lie within the middle 80% of the total time, $t \sim U(0.1 \cdot T, 0.9 \cdot T)$. The number of observations before and after the changepoint, I_1 and I_2 respectively, are simulated by

$$I_1 \sim \text{Pois}(\mu_1 \cdot t), \tag{3.16}$$

$$I_2 \sim \text{Pois}(\mu_2 \cdot (T - t)), \tag{3.17}$$

where μ_1 and μ_2 are given parameter values of the expected number of observations per time unit. The Poisson distribution was chosen to simulate the number of observations since it is often used in data with counts of occurrences that are independent [8, p. 43].
The observation times, τ, are simulated by

$$\tau_i \begin{cases} \sim \mathrm{U}(0, t), & i = 1, \ldots, I_1, \\ \sim \mathrm{U}(t, T), & i = I_1 + 1, \ldots, I_1 + I_2, \end{cases} \tag{3.18}$$

and then sorted so that the following holds: $\tau_1 < \cdots < \tau_{I_1+I_2}$. Each observation time is simulated to be independent of the other times, with no fixed time interval. The observation values connected to each observation time, X, are simulated by

$$X_i \begin{cases} \sim \mathrm{Gamma}(\alpha_1, \beta_1), & \tau_i < t, \\ \sim \mathrm{Gamma}(\alpha_2, \beta_2), & \tau_i \geq t, \end{cases} \tag{3.19}$$

with given parameter values, (α_1, β_1) and (α_2, β_2). The gamma distribution was chosen to simulate the observation values due to the possibility of having a high probability with lower values and a low probability with higher values.
The function used to simulate a time series can be summarized as

$$(t, \tau, X) = \text{Simulating time series}(T, t, \alpha, \beta, \mu), \tag{3.20}$$

and the step by step can be seen in Algorithm 4. An example of a simulated time series can be seen in Figure 3.3.

Algorithm 4 Simulating time series

```
function Simulating time series(T, α, β, μ)
    t ← U(0.1 · T, 0.9 · T)
    I_1 ← Poisson(μ_1 · t)
    I_2 ← Poisson(μ_2 · (T − t))
    for i = 1 : I_1 do
        τ_i ← U(0, t)
    end for
    for i = (I_1 + 1) : (I_1 + I_2) do
        τ_i ← U(t, T)
    end for
    for i = 1 : (I_1 + I_2) do
        if τ_i < t then
            X_i ← Gamma(α_1, β_1)
        else
            X_i ← Gamma(α_2, β_2)
        end if
    end for
    return (t, τ, X)
end function
```

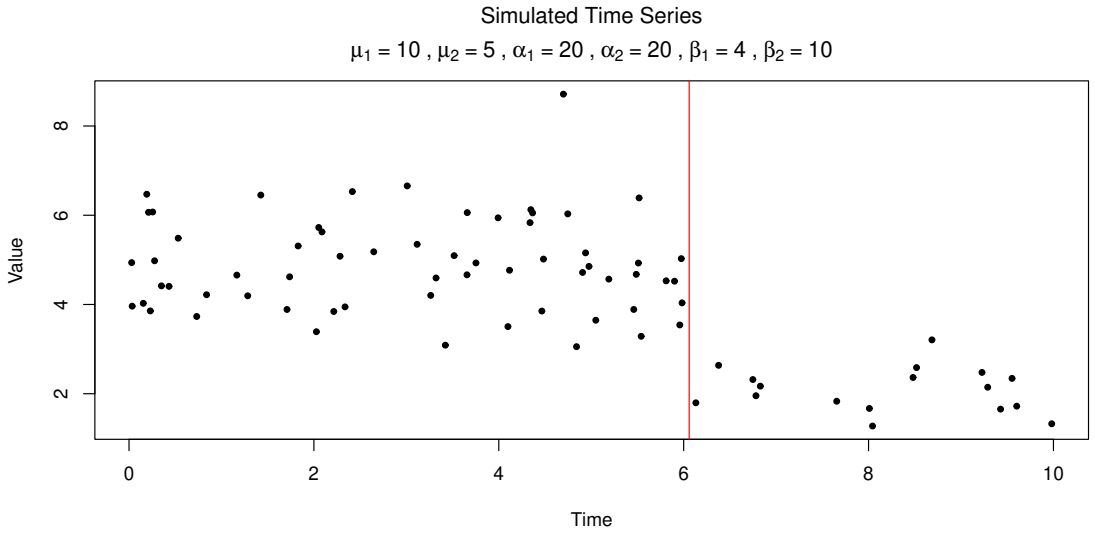

Figure 3.3: Simulated time series where the red line marks the change point.

Test for the simulated time series

To test if the algorithm performs well on the simulated data, the following test was set up.
For 29 different combinations of parameter values P (29 × 6 matrix), found in Table B.1, a time series was simulated using the function Simulating time series, specified with $T = 10$. The parameter values were chosen to produce a wide variety of time series with change points. Some time series have a very clear and distinct change point and others are more difficult to detect through visual inspection only. For one time series, the Sampling function, Algorithm 2, was applied, followed by the Find parameters function, Algorithm 3, in order to estimate values for t, μ_1, μ_2, α_1, α_2, β_1 and β_2. Then the squared error, $SE = (\theta - \hat{\theta})^2$, was calculated for each estimated value $\hat{\theta}$ from the true value θ of the parameters. To get a more comprehensive result, all steps are done for a number of runs, q. Lastly, the root mean squared error, $RMSE = \overline{\text{mean}(SE)}$, of all runs are calculated for each parameter. The test is summarized in Algorithm 5, with the input being the number of iterations ($n = 30000$) for a number of runs ($q = 10$) performed on all parameter combinations (P). The output is a matrix, RMSE, with the root mean squared error for each parameter set. To be able to compare the parameters μ, α and β more easily, they are also normalized by the standard deviations of each corresponding column in table B.1.

Algorithm 5 Test for simulated time series

```
function Testing simulated time series(n, q, P)
    SE ← matrix(length(P), q, 7)
    for i = 1 : length(P) do
        θ ← P_i
        for j = 1 : q do
            (τ, X) ← Simulating time series(T = 10, θ)
            θ□ ← Sampling(τ, X, n)
            θ̂ ← Find changepoint(θ)
            SE_{i,j} ← (θ − θ̂)^2
        end for
    end for
    return RMSE ← sqrt(mean(SE))
end function
```

3.2.2 Financial time series from Svalna

This section gives a description of the data set from Svalna, with some relevant descriptions of the categories and individuals connected to the data.
The data set consists of financial transactions from different individuals. Each transaction has beforehand been given a category according to Svalna's categorizations. Every transaction also has some information connected to the individual who made the transaction, such as their salary and city of residence. Table 3.4 shows some specifics of the data set.

Table 3.4: Specifics of the data set.

Number of transactions	244565
Number of variables connected to each transaction	343
Number of unique individuals	2595
Number of categories	65

Each individual has a unique time span, with a given start date and end date. A transaction for each category is given weekly between the start and end date, and if there has been no transaction for that week, the transaction value is set to zero. These values corresponding to one individual and one category can be seen as a time series.
Since there are 168675 unique time series (#unique individuals · #categories), it was not feasible to investigate all time series. It was more interesting to find some individuals that have something in common and investigate similarities and differences.

Test for the financial time series from Svalna

Based on the aim of this project to investigate differences in purchase behaviours before and after an event, two scenarios were chosen to be the subject of investigation. The first is when the individual has bought a new car or changed car, and the

second scenario is when the individual has moved once within their time span. Both these changes can be found by comparing some of the variables connected to each transaction, such as the city of residence or the number of cars the individual own. It also gives the time of the proposed change point, t, which is the date at which this change is made. This narrowed it down to 85 individuals who has changed car once, and 162 individuals who has moved once.

All values in the time series that equal zero were removed from the time series, since it is the same as an absence of a transaction. Some time series connected to these individuals then had no values or few values. Based on the model used, if there are too few values, nothing statistically interesting can be found. Therefore a limit of at least 50 values in the time series was used to disregard the time series with few values. To make the computations more general, the observation values were normalized and the observation times were changed from specific dates to a time line where $\tau = 1$ is a year from the first transaction. More specifically, each transaction time was ordered by number of weeks from the first transaction and divided by 52.143, which is the average number of weeks per year. Since it was interesting to compare what has happened both before and after the change point, only the individuals that have a change point that lies within the middle 80% of the full time span was investigated, in order to make sure that there were at least some values on both sides of the change point.

For the time series that fulfill these requirements, the same steps as for the simulated time series were done. Using the Sampling function, Algorithm 2, and the Find parameters function, Algorithm 3, to investigate if the change point could be found in that particular time series by calculating the squared error compared to t. If the change point could be found accurately, it would then be worth investigating the differences in the parameters μ, α and β.

Some of the most interesting questions to try to answer for this data were: are there similar categories that accurately show the change point for different individuals, and if so, are the changes in parameters also similar?

3.3 Reversible-jump Metropolis-Hastings

After some initial testing, it was concluded that there was a need for the model to have the possibility of switching between the assumption that there exists a change point or not. The main problem that occurs with switching assumptions is that there is a difference in the number of parameters. When there is no change point, the relevant parameters are only (μ, α, β) instead of $(\mu_1, \mu_2, \alpha_1, \alpha_2, \beta_1, \beta_2, t)$. This requires the model to make transdimensional moves during the iterations. This was done using reversible-jump Metropolis-Hastings [12]. (Since the α variables are approximated in the model and does not use random walk, they are excluded from the following formulas for the sake of clarity.)

For every 10 iterations in the random walk of the model, see Algorithm 2, the candidate θ' is given with the opposite assumption. Then a acceptance ratio, r, is calculated based on the assumptions, and if the candidate is accepted, the next 10 iterations will run with the new assumption. The assumptions can be represented by an extra parameter $k \in \{0, 1\}$, where $k = 0$ corresponds to the assumption that

there is no change point present, and $k = 1$ corresponds to that there is. This gives two different scenarios. (Note that in order to assure linearity, the parameters μ and β work in the log scale. For easier notation; $m = \log(\mu)$, and $b = \log(\beta)$.)

Case 1: $k = 0$, $k' = 1$

Going from the assumption of not having a change point to having a change point in the time series requires the following function for the transformation:

$$g_{0\to 1}(m, u_1, b, u_2, u_3) \to (m_1, m_2, b_1, b_2, t), \tag{3.21}$$

$$g_{0\to 1}(m, u_1, b, u_2, u_3) = (m - u_1, m + u_1, b - u_2, b + u_2, u_3). \tag{3.22}$$

The assumption of having no change point resides in a smaller dimension than the opposite assumption, therefore it is augmented by auxiliary variables, u_1, u_2 and u_3, such that the dimensions agree. The auxiliary variables are sampled from the following distributions:

$$u_1 \quad \mathrm{N}(0, 1), \tag{3.23}$$

$$u_2 \quad \mathrm{N}(0, 1), \tag{3.24}$$

$$u_3 \quad \mathrm{U}(0, T). \tag{3.25}$$

The model uses the acceptance ratio r to determine if the candidate is accepted. The ratio is calculated by

$$r = \frac{p(m_1, m_2, b_1, b_2, t|k')p(k')}{q(u_1, u_2, u_3)p(m, u_1, b, u_2, u_3|k)p(k)}|J_{g_{0\to 1}}|, \tag{3.26}$$

where $p(m_1, m_2, b_1, b_2, t|k')$ is the probability of the model having those parameter values, and $p(k)$ is the probability of the model having the chosen assumption. The probability of this is set as $p(k = 0) = p(k = 1) = \frac{1}{2}$. The auxiliary variables give rise to

$$q(u_1, u_2, u_3) = \varphi(u_1) \cdot \varphi(u_2) \cdot \frac{1}{T}. \tag{3.27}$$

The Jacobian determinant is used since there is a change in variables,

$$J_{g_{0\to 1}} = \begin{vmatrix} 1 & -1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{vmatrix} = 4. \tag{3.28}$$

Case 2: $k = 1$, $k' = 0$

When going from the assumption that there is a change point to that there is no change point, the following function represents the transition:

$$g_{1\to 0}(m_1, m_2, b_1, b_2, t) \to (m, u_1, b, u_2, u_3), \tag{3.29}$$

$$g_{1\to 0}(m_1, m_2, b_1, b_2, t) = \left(\frac{m_1 + m_2}{2}, \frac{m_1 - m_2}{2}, \frac{b_1 + b_2}{2}, \frac{b_1 - b_2}{2}, t\right). \tag{3.30}$$

The acceptance ratio in this scenario is calculated by the following formula:

$$r = \frac{q(u_1, u_2, u_3)p(m, u_1, b, u_2, u_3|k\ )p(k\ )}{p(m_1, m_2, b_1, b_2, t|k)p(k)}|J_{g_{1\to 0}}|. \tag{3.31}$$

The Jacobians of both functions are the inverse to one another, $J_{g_{0\to 1}} = J^{-1}_{g_{1\to 0}}$, and therefore $|J_{g_{1\to 0}}| = \frac{1}{4}$.
These two cases were integrated into the function Sampling (Algorithm 2) and is the model that was used to provide the results.

4

Results and discussions

4.1 Simulated time series

4.1.1 Results

The results for the test run on the simulated time series are presented here. Table 4.1 shows the root mean squared errors (RMSE) for the 10 runs using the different parameters. The closer to 0 a value is, the more accurate the prediction is. The error is only calculated for the runs that have the assumption that there is a change point in the final iteration, and the column #runs shows the number of runs that are used in calculating the error. Table 4.2 shows the percentage of the iterations, after the warm up iterations, that work with the assumption of having a change point to provide a measure of certainty.

Table 4.1: RMSE for testing simulated time series.

	t	μ_1	μ_2	α_1	α_2	β_1	β_2	#runs
Parameters 1	5.48	1.93	2.05	5.32	4.33	4.67	3.68	2
Parameters 2	1.92	0.49	1.30	1.73	1.51	1.76	0.69	10
Parameters 3	0.17	0.81	0.54	2.03	0.37	2.06	0.91	10
Parameters 4	0.85	0.64	0.27	1.57	0.03	1.44	0.08	1
Parameters 5								0
Parameters 6	0.88	0.64	0.67	1.39	1.12	1.37	0.38	9
Parameters 7	2.71	0.78	1.96	1.34	0.46	1.31	1.57	10
Parameters 8								0
Parameters 9	2.13	0.53	0.51	1.80	0.84	0.70	2.17	10
Parameters 10	3.11	0.52	2.09	1.36	0.67	0.47	0.55	10
Parameters 11	1.54	0.77	1.01	1.05	2.38	0.41	2.91	10
Parameters 12	7.11	1.49	4.89	0.30	16.18	0.09	5.07	2
Parameters 13	2.43	1.82	0.95	10.33	0.73	3.88	1.95	10
Parameters 14	0.25	0.58	0.83	0.61	0.42	0.24	0.37	10
Parameters 15	1.66	0.70	0.78	0.45	0.76	1.31	0.94	10
Parameters 16	0.23	0.47	0.50	0.33	1.15	0.84	1.14	10
Parameters 17	2.24	0.81	0.94	0.91	1.37	2.34	0.56	9
Parameters 18	7.29	2.30	2.07	4.83	0.14	10.11	0.56	1
Parameters 19	0.74	0.59	1.00	0.40	2.90	1.06	4.82	10
Parameters 20	5.37	2.22	3.53	3.59	1.99	2.90	2.58	4
Parameters 21	2.46	0.78	1.85	0.35	0.63	0.44	0.33	10
Parameters 22	0.99	0.57	0.56	1.25	0.39	0.94	1.16	10
Parameters 23								0
Parameters 24	0.32	0.67	0.46	1.90	1.89	1.92	0.71	10
Parameters 25	0.32	0.58	0.66	0.79	0.34	0.85	0.92	7
Parameters 26	3.67	3.19	0.57	3.44	1.01	3.03	0.97	1
Parameters 27	1.80	0.59	1.33	3.15	0.77	1.14	1.80	10
Parameters 28	2.04	0.61	0.95	1.75	0.89	0.62	1.10	10
Parameters 29	0.44	0.89	0.73	0.69	0.47	1.96	0.37	9

Table 4.2: Certainty of the assumption of having a change point, in percentages.

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10
Parameters 1	4	1	25	0	3	0	100	0	0	100
Parameters 2	100	100	100	100	100	100	100	100	100	100
Parameters 3	100	100	100	100	100	100	100	100	100	100
Parameters 4	0	0	0	34	0	30	0	12	0	0
Parameters 5	0	0	29	0	30	5	0	78	2	0
Parameters 6	100	100	100	100	100	44	100	100	100	100
Parameters 7	100	100	100	100	100	100	78	100	100	100
Parameters 8	0	0	3	0	0	0	3	14	4	7
Parameters 9	100	100	100	100	100	100	100	100	100	100
Parameters 10	100	100	100	100	100	100	100	100	100	100
Parameters 11	100	100	100	100	100	100	100	100	100	100
Parameters 12	5	1	1	100	0	1	0	0	0	100
Parameters 13	100	100	100	100	100	100	100	100	100	100
Parameters 14	100	100	100	100	100	91	100	100	100	87
Parameters 15	100	100	100	100	100	100	100	78	100	94
Parameters 16	100	100	100	100	100	100	100	100	100	100
Parameters 17	100	100	5	100	100	100	100	100	100	100
Parameters 18	0	2	4	1	67	1	2	0	1	0
Parameters 19	97	100	100	100	100	94	100	97	95	100
Parameters 20	0	0	22	100	71	0	0	47	0	100
Parameters 21	100	100	100	100	100	100	100	100	100	100
Parameters 22	100	100	100	100	100	100	100	100	55	100
Parameters 23	0	0	0	0	1	8	0	0	0	1
Parameters 24	100	100	100	100	100	100	100	100	100	100
Parameters 25	100	100	4	100	100	0	0	100	100	100
Parameters 26	0	2	0	0	0	3	10	0	4	100
Parameters 27	100	100	100	100	100	100	100	100	100	100
Parameters 28	100	100	100	100	100	100	100	100	100	100
Parameters 29	61	100	94	100	67	100	2	94	100	100

4.1.2 Does the model work for simulated time series?

The results of the test for the simulated time series show that they can be divided into five different sections regarding the accuracy of finding the change point:

1. High #runs and low *t*-rmse value.
2. High #runs and high *t*-rmse value.
3. Low #runs and low *t*-rmse value.
4. Low #runs and high *t*-rmse value.
5. No #runs.

The first group with a high number of the runs finding change points (7 to 10) and a low *t*-rmse value (< 1) are the time series where the model accurately finds the change point for all relevant runs. It consists of time series generated from the parameters #3, 6, 14, 16, 19, 22, 24, 25 and 29. From Table 4.2, all relevant runs have a high certainty of the assumption of having a change point. By visually inspecting the time series and comparing the parameter values the common factor is that they all have a difference in the expected mean of the observation values before and after the change point. Figure 4.1 shows examples of the time series along with the corresponding histogram of the *t* parameter from a run. The corresponding

trace plots for the variables t, $_1$, $_2$, $_1$ and $_2$ can be found in Appendix C, gures C.1 and C.2. They show both the convergence of the model and that the found parameter values are close to the correct values.

Figure 4.1: Change points found for time series using parameters #3 and #14 respectively.

The second group also have a high #runs, but higher rmse values (1). The time series are generated from parameters #2, 7, 9, 10, 11, 13, 15, 17, 21, 27 and 28. They all have high certainties of having the assumption that there is a change point in most runs, as can be seen in Table 4.2. The common thread for these combinations of parameters is that they also generate time series with di erent expected means before and after the change point. By visually inspecting all images, the other common thing is that the model can accurately predict the change point in most runs, with a few exceptions. Between 1 to 3 runs for these parameters fail to nd a change point. Two examples of how it looks like when it fails can be seen in Figure 4.2. The corresponding trace plots can be seen in Figures C.3 and C.4.

Figure 4.2: Failing to nd change points for time series using parameters #2 and #17 respectively.

The third group has a low #runs and a lowt-rmse value. This group only consist of a time series generated from parameters #4, that has a di erence in the expected variance before and after the change point. For the one run that ends with the assumption of having a change point, the certainty is low (34%), but the found change point is close to the actual change point, see Figure 4.3. The low certainty can also be seen in the trace plot shown in Figure C.5.

Figure 4.3: Found change point for a time series using parameters #4.

The fourth group has a low #runs and hight-rmse values. It consists of time series generated from the parameters #1, 12, 18, 20 and 26. The time series generated from parameters #1 are the only ones that do not have a change point. The rest of the time series have only a change in the expected variance or the expected frequency of the observation values before and after a change point, or a combination of the two. They all have the same expected mean of the observation values throughout the time series. The certainties of the assumption of having a change point are mostly high, but they fail to accurately nd the change point, see Figure 4.4 for examples with corresponding trace plots found in Figures C.6 and C.7. However,

most of the runs have the assumption that there is no change point.

Figure 4.4: Failing to find change points for time series using parameters #12 and #26 respectively.

The final group is the time series where none of the runs end with the assumption of having a change point and it consists of time series generated from parameters #5, 8 and 23. Most of the runs have low certainties of the assumption of having a change point. Same as for the previous group, the time series have only a change in the expected variance or the expected frequency of the observation values before and after a change point, or a combination of the two.

Based on these results, it is clear that the model works in finding the change points for the time series that has a difference in the expected mean, but has more trouble finding them when there is a change in the expected variance. It also does not seem to work for the time series that varies in the expected frequency.

Table 4.1, as well as the trace plots found in Appendix C, show that the parameter values that are estimated using the reversible-jump Metropolis-Hastings algorithm (λ_1; λ_2; μ_1; μ_2) are also accurately found at the same rate as the change point.

4.2 Financial time series from Svalna

4.2.1 Results

Tables 4.3 and 4.4 show the squared error from the found change point and the corresponding event for each individual with time series that have enough observations. The categories where none of the individuals have enough observations are excluded. If the final iteration in the model has the assumption that there is no change point present, the error is shown as Inf.

Table 4.3: SE from the proposed change point for each category time series, for the individuals who bought/changed car once. Where the model has the assumption that there is no found change point -Inf is used instead of an error.

	2	3	5	6	7	9	12	13	14	16	21	22	23	27	30	32	33	34
Indv. 7				-Inf			0.00	-Inf			3.10			-Inf				-Inf
Indv. 8																		
Indv. 12														-Inf				
Indv. 13														-Inf				
Indv. 14											-Inf							
Indv. 15					4.79		0.93	-Inf						-Inf				-Inf
Indv. 18														-Inf				
Indv. 20									-Inf		-Inf		-Inf	-Inf			0.27	-Inf
Indv. 27	-Inf													-Inf				-Inf
Indv. 33																		
Indv. 34														-Inf				
Indv. 35														-Inf				
Indv. 38														2.34				-Inf
Indv. 43														-Inf				-Inf
Indv. 45								-Inf						-Inf				
Indv. 46														-Inf				-Inf
Indv. 51	-Inf				-Inf	0.53		0.16	3.81				-Inf	-Inf			-Inf	-Inf
Indv. 61	2.15				0.15							0.07		-Inf				-Inf
Indv. 62	-Inf	-Inf			-Inf	-Inf		1.40			-Inf	-Inf	1.69	-Inf				-Inf
Indv. 66														-Inf				
Indv. 67	0.73	1.78											0.06	0.57				
Indv. 69	16.33		-Inf	3.48	-Inf	7.44	23.60	-Inf	10.66		-Inf	-Inf	-Inf	-Inf	-Inf	1.43		8.99
Indv. 77														-Inf			0.12	-Inf
Indv. 80														-Inf				-Inf
Indv. 82						4.60	-Inf			-Inf	4.50		-Inf	0.42			7.32	-Inf
Indv. 84														-Inf				

38	39	40	41	42	43	44	47	48	49	50	51	53	57	58	59	60	62	63
-Inf	0.17		32.20				-Inf		3.59		0.10		-Inf		5.49			-Inf
															0.00			
							0.41				0.01				-Inf			
		1.05	-Inf		-Inf	0.27	-Inf		0.16	-Inf	-Inf				-Inf		-Inf	0.93
	0.02		-Inf		6.02	-Inf	4.76	0.14			-Inf	-Inf		-Inf	1.75			0.95
							-Inf				-Inf				-Inf			
												-Inf				-Inf		
															-Inf			
							-Inf								2.93			
															0.15			
															-Inf			
							0.08				0.99							-Inf
						-Inf	-Inf		0.02		-Inf				0.01			
							-Inf		-Inf	0.31	-Inf				-Inf		-Inf	0.12
-Inf						2.58	2.79		-Inf	1.04	-Inf				-Inf			0.01
							-Inf								0.84			
						0.01					-Inf				0.04			
			-Inf	0.55	23.56	0.43	4.25		3.38	-Inf	16.06	-Inf	-Inf	-Inf	6.38		22.42	-Inf
							-Inf		-Inf						2.21			-Inf
															-Inf			
						-Inf	-Inf	-Inf	0.05		1.68				5.65			5.15
							-Inf								-Inf			

Table 4.4: SE from the proposed change point for each category time series, for the individuals who changed home once. Where the model has the assumption that there is no found change point -Inf is used instead of an error.

	2	3	5	6	7	9	12	13	14	16	17	19	21	22	23	27	30	32	33
Indv. 3																-Inf			
Indv. 5																-Inf			
Indv. 13																			
Indv. 14				-Inf			0.06	-Inf					3.10			-Inf			
Indv. 16																-Inf			
Indv. 19																0.57			
Indv. 21	-Inf					-Inf	-Inf								-Inf	-Inf			
Indv. 22																-Inf			
Indv. 24	11.33		-Inf		-Inf		0.74	-Inf	0.03				-Inf			18.22	32.99		
Indv. 29																-Inf			
Indv. 33																1.54			
Indv. 35																-Inf			
Indv. 43							3.05	-Inf	-Inf	-Inf		6.95	1.47			-Inf			5.52
Indv. 48					-Inf		-Inf			-Inf	17.53		1.72		13.81	0.19			
Indv. 49						1.29		-Inf								-Inf			
Indv. 51										-Inf			1.90			0.02			0.25
Indv. 53	-Inf	-Inf											-Inf			-Inf			
Indv. 55																-Inf			
Indv. 56																0.14			
Indv. 60						-Inf		-Inf							-Inf	-Inf			
Indv. 63					0.00											4.23			
Indv. 65																		0.51	
Indv. 66													0.54			0.75			
Indv. 67																-Inf			
Indv. 68							-Inf	-Inf								0.15			-Inf
Indv. 70								0.07								0.25			
Indv. 72																0.00			
Indv. 77																0.44			
Indv. 79																			
Indv. 80																-Inf			
Indv. 82																			
Indv. 83																-Inf			
Indv. 84								-Inf								-Inf			
Indv. 85				-Inf			4.98	-Inf					-Inf			0.82			
Indv. 86							-Inf									2.50			0.99
Indv. 87																-Inf			
Indv. 88																-Inf			
Indv. 96																-Inf			
Indv. 98																-Inf			
Indv. 101	-Inf															-Inf			
Indv. 102																0.73			
Indv. 106																0.75			
Indv. 112	0.00				0.87											-Inf			
Indv. 114					-Inf	2.20	1.59	-Inf	-Inf				1.87			10.68			-Inf
Indv. 115								-Inf					-Inf			4.58			
Indv. 117																-Inf			
Indv. 123																0.38			
Indv. 126	-Inf	-Inf			-Inf	-Inf		1.58					-Inf	-Inf	0.39	-Inf			
Indv. 127																			
Indv. 133																0.68			0.01
Indv. 142							0.28									-Inf			
Indv. 145							-Inf									-Inf			
Indv. 148																1.73	-Inf		
Indv. 150																-Inf			0.12
Indv. 157						-Inf										0.69			
Indv. 161	-Inf				-Inf				-Inf							0.54			

Continuation of Table 4.4.

34	38	39	41	42	43	44	47	48	49	50	51	53	55	57	58	59	62	63
				-Inf			0.86				0.48					0.00		
	1.40			-Inf			-Inf				-Inf							
							-Inf											
-Inf	-Inf	0.17	-Inf				-Inf		3.54		1.47			-Inf		3.20		-Inf
																-Inf		
							0.41				0.18					-Inf		
0.00				0.03			3.79		0.34		-Inf					0.69		2.22
0.36				0.32			-Inf									0.19		
70.11			-Inf	61.98	19.83		26.68		19.05	-Inf	11.86	10.91				38.62		0.08
							1.08											
																-Inf		
-Inf					53.28		-Inf	0.41	1.76	0.57	-Inf					7.79	-Inf	-Inf
-Inf						15.06	0.90		-Inf		16.76	-Inf				0.64		1.01
-Inf							-Inf				-Inf					-Inf		
-Inf					-Inf	0.16	-Inf	-Inf	0.27		-Inf					1.87		0.11
25.46				-Inf	-Inf	-Inf	2.66		-Inf		-Inf					-Inf		
							-Inf									0.43		
																-Inf		
-Inf							0.55				-Inf					-Inf		
4.36											-Inf					5.42		
0.34							0.13				0.68					1.10		0.46
							-Inf									-Inf		
-Inf							0.20				0.04					3.66		
							0.37			-Inf						-Inf		-Inf
0.06							0.08		0.03							-Inf		
				-Inf			0.00									-Inf		
																-Inf		
-Inf																0.16		
																-Inf		
							-Inf				-Inf							
																-Inf		
-Inf			9.61	-Inf			0.88	2.16	-Inf	-Inf	-Inf		1.12			1.47		1.00
-Inf			-Inf	-Inf		1.32	-Inf				0.60					0.15		
-Inf							0.08				0.99							-Inf
-Inf							0.00									-Inf		
2.59						0.03	0.00		4.53		-Inf					0.80		
							-Inf									-Inf		
							-Inf									0.11		
							0.48				-Inf					-Inf		
							-Inf				0.65					0.00		
-Inf							-Inf									0.95		
-Inf			13.88		2.01	-Inf	0.05	0.04	3.90	-Inf	6.26					0.77	5.30	-Inf
			-Inf				1.59			-Inf	-Inf	-Inf				-Inf		9.96
																-Inf		
-Inf	-Inf					2.58	-Inf		-Inf	1.05	-Inf					-Inf		0.00
-Inf																1.04		
0.08				-Inf											0.67	2.65		
																-Inf		
				-Inf			0.08				1.38					1.11		
			-Inf	0.21			0.04				0.94					3.26		
-Inf							-Inf		-Inf							2.20		-Inf
0.45				-Inf			-Inf									0.01		0.87
0.13							-Inf				0.65					0.52		

4.2.2 Insights into the data

For the two cases (one change in car, one change in home), the results of running the model on the di erent categories for these individuals were not enough to be able to make any conclusions. See Figure 4.5 for one example where the found change point lines up with when the individual changed car, and one example where the change point is found at a di erent time. There are some instances where the model nds the change point in a category at the same point in time as when an individual has bought a new car or moved, but has not found it for other individuals in the same category. Therefore it is not possible to make any certain conclusions about the parameter values of the categories. One possible reason that the model

cannot accurately nd the change point could be that the model uses the wrong assumptions on the parameters. It could also simply be that for these speci c cases, for these chosen time series, there is no change in spending habits, meaning that there is no change point to be found.

Figure 4.5: Found change points in time series from two di erent individuals. The red line in the time series marks the time where the individuals bought/changed car.

4.3 Discussion

4.3.1 Choice of model and assumptions

A non-time dependant model was chosen for the possibility to make general statements and comparisons on the underlying distributions. The reason behind doing the implementation from scratch was to develop a deeper understanding of how this method would be able to answer the questions stated. The assumptions on the data regarding the observation values, was chosen to be from a Gamma distribution to capture the possibility of expensive outliers. The same reasoning is applied to the assumptions on the and parameters. We assume most of them lie within the same span, but still including some outliers. The lognormal distribution that is chosen is very similar in shape to the Gamma distribution, but the implementation of the model becomes more simple if it operates within the log scale.

4.3.2 Possible continuation

The results from running the model on simulated time series show a promising start, along with the need to improve the implementation by using more complex methods, one being automatic tuning of the parameters. Another route is to implement this with already nished functions, found in programming languages such as Stan [13]. Even if the model could be improved to be able to accurately nd the change point for time series where the variance or the frequency is varied, it still might not be possible to nd it in these particular real time series that has been investigated for this thesis. More speci c cases in the real data, that can include more individuals and more time series, could provide more insights.

4.3.3 Ethical aspects

There are some ethical questions that need to be considered when investigating nancial data connected to individuals, and their purchase behaviours. Some principles of data ethics that can be discussed are Ownership, Transparency, Privacy, Intention, and Outcomes[14].
Every individual has ownership over their own data, and therefore it is important that there is some manner of informed consent before collecting and using the data in research. The real data that have been used in this thesis comes from individuals that have given consent in Svalna's app for this purpose. In the meaning of transparency, Svalna provides a link to their ongoing projects that the individual will give consent to, as well as give an explanation as to what data will be included. Svalna works in cooperation with Tink [15] and BankID [16] to provide secure data transfers.
The data provided for this thesis does not divulge any personal information such as name, address, date of birth, phone number, etc. The data connected to an individual instead uses an unique identi cation number consisting of random numbers and letters. Whether or not the information that is given can still be used to connect the data with the correct person has not been investigated, but might still be a question to have in mind. To ensure privacy, the only gure that uses real data, Figure 4.5, uses normalized values so that the actual values are unknown. Nothing that can be connected to an individual has been presented in this thesis.
Regarding the intentions of this thesis, the aim is to investigate purchase behaviours. In the future this information might be used to be able to give recommendations to individuals on how to choose more environmentally friendly options. Even though a possible next step is outside the scope of this thesis, it can give rise to interesting discussions. When and how should this information be presented to someone? What deductions should be made from an individual's purchases? Can this information be used in some nefarious way?

5

Conclusion

We have succeeded in implementing a model that is able to nd the change point in time series where there is a di erence in the mean of the observation values before and after the change point. The model could be improved in order to more accurately nd change points where there is a di erence in the variance of the observation values or the frequency of observations. For the nancial data provided by Svalna, the model is able to nd some change points, or dismiss the assumption of the time series having a change point. In the two speci c cases that were looked at, the model cannot be used to make any certain conclusions of the spending habits of individuals. The overall method, using Bayesian inference and reversible-jump Metropolis-Hastings sampling in order to nd the underlying statistical distributions, is not dismissed. Instead, further improving the model, or implementing it using already existing functions, could be a way forward. Investigating even more cases and including more time series could also provide more insights into spending habits.

Bibliography

[1] E. S. Page. CONTINUOUS INSPECTION SCHEMES . In: Biometrika 41.1-2 (1954), pp. 100 115. issn: 0006-3444. doi : 10.1093/biomet/41.1-2.100 .

[2] Tze Leung Lai. Sequential Changepoint Detection in Quality Control and Dynamical Systems . In: Journal of the Royal Statistical Society: Series B (Methodological) 57.4 (Nov. 1995), pp. 613 644. issn: 00359246. doi : 10.1111/j.2517-6161.1995.tb02052.x .

[3] Colin Gallagher, Robert Lund, and Michael Robbins. Changepoint Detection in Climate Time Series with Long-Term Trends . In: Journal of Climate 26.14 (July 2013), pp. 4994 5006. issn: 0894-8755. doi : 10.1175/JCLI-D-12-00704.1.

[4] Svalna Date accessed: 13/09/2022. url : https://svalna.se/web/en .

[5] Research. Date accessed: 06/12/2022. url : https://svalna.se/web/en/research .

[6] Peter J. Brockwell and Richard A. Davis. Introduction to Time Series and Forecasting Cham: Springer International Publishing, 2016, p. 1. isbn: 978-3-319-29852-8. doi : 10.1007/978-3-319-29854-2 .

[7] Rob Hyndman and Yangzhuoran Yang. tsdl: Time Series Data Library. Date accessed: 03/08/2022. url : https://pkg.yangzhuoranyang.com/tsdl/index.html .

[8] Andrew Gelman, John B. Carlin, Hal S. Stern, et al. Bayesian Data Analysis Third Edition. Chapman and Hall/CRC, Nov. 2013. isbn: 9780429113079. doi : 10.1201/b16018.

[9] Justin Ellis. A Practical Guide to MCMC Part 1: MCMC Basics. Date accessed: 10/08/2022. url : https://jellis18.github.io/post/2018-01-02-mcmc-part1/ .

[10] The R Foundation. The R Project for Statistical Computing Date accessed: 25/07/2022. url : https://www.r-project.org/ .

[11] The R Foundation. What is R? Date accessed: 25/07/2022. url : https://www.r-project.org/about.html .

[12] PETER J. GREEN. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination . In: Biometrika 82.4 (1995), pp. 711 732. issn: 0006-3444. doi : 10.1093/biomet/82.4.711 .

[13] Stan Development Team. Stan Modeling Language Users Guide and Reference Manual. Date accessed: 05/10/2022. 2022. url : https://mc-stan.org .

[14] Catherine Cote. 5 Principles of Data Ethics for Business Date accessed: 10/10/2022. Mar. 2021. url : https://online.hbs.edu/blog/post/data-ethics .

[15] Tink. Date accessed: 10/10/2022. url : https://tink.com/ .
[16] BankID. Date accessed: 10/10/2022. url : https://www.bankid.com/ .

A

Relevant distributions

Here, the relevant distributions are presented for reference. Each distribution has a description of the notation, parameters, density function, mean and variance. Table A.1 shows the continuous distributions, and Table A.2 shows the discrete distributions.

Table A.1: Continuous distributions as seen in [8, Table A.1].

Distribution	Notation	Parameters	Density function	Mean and variance
Uniform	$U(\ ;\ )$ $p(\ ) = U(\ j\ ;\ )$	boundaries $\ ;\ $ with $\ >\ $	$p(\ ) = \frac{1}{\ }; \quad 2 [\ ;\ ]$	$E(\ ) = \frac{+}{2}$ $\mathrm{var}(\ ) = \frac{(\ \ )^2}{12}$
Normal	$N(\ ;\ ^2)$ $p(\ ) = N(\ j\ ;\ ^2)$	location scale $\ >\ 0$	$p(\ ) = \frac{1}{\sqrt{2}}\exp\ \frac{1}{2\ ^2}(\ \ \ )^2$	$E(\ ) =$ $\mathrm{var}(\ ) =\ ^2$
Gamma	$\mathrm{Gamma}(\ ;\ )$ $p(\ ) = \mathrm{Gamma}(\ j\ ;\ )$	shape $\ >\ 0$, rate $\ >\ 0$	$p(\ ) = \frac{\ }{(\ )}\ ^{1}e\ \ , \ >\ 0$	$E(\ ) = \ -$ $\mathrm{var}(\ ) = \frac{\ }{\ ^2}$

Table A.2: Discrete distributions as seen in [8, Table A.2].

Distribution	Notation	Parameters	Density function	Mean and variance
Poisson	$\mathrm{Poisson}(\ )$ $p(\ ) = \mathrm{Poisson}(\ j\ )$	rate $\ >\ 0$	$p(\ ) = \frac{1}{!}\ \ e\ \ , \quad = 0; 1; 2; : : :$	$E(\ ) =$ $\mathrm{var}(\ ) =$

B

Combinations of parameter values for simulated time series

Table B.1: Parameter values used for simulating time series.

	μ_1	μ_2	α_1	α_2	β_1	β_2
Parameters 1	5	5	8	8	4	4
Parameters 2	10	10	20	20	10	4
Parameters 3	10	10	20	8	10	10
Parameters 4	10	10	20	8	10	4
Parameters 5	10	5	20	20	10	10
Parameters 6	10	5	20	20	10	4
Parameters 7	10	5	20	8	10	10
Parameters 8	10	5	20	8	10	4
Parameters 9	10	10	20	8	4	10
Parameters 10	10	10	20	8	4	4
Parameters 11	10	5	20	20	4	10
Parameters 12	10	5	20	20	4	4
Parameters 13	10	5	20	8	4	10
Parameters 14	10	5	20	8	4	4
Parameters 15	10	10	8	8	10	4
Parameters 16	10	5	8	20	10	10
Parameters 17	10	5	8	20	10	4
Parameters 18	10	5	8	8	10	10
Parameters 19	10	5	8	8	10	4
Parameters 20	10	5	8	20	4	10
Parameters 21	10	5	8	20	4	4
Parameters 22	10	5	8	8	4	10
Parameters 23	10	5	8	8	4	4
Parameters 24	5	5	20	20	10	4
Parameters 25	5	5	20	8	10	10
Parameters 26	5	5	20	8	10	4
Parameters 27	5	5	20	8	4	10
Parameters 28	5	5	20	8	4	4
Parameters 29	5	5	8	8	10	4

C

Trace plots

Figure C.1: Trace plots for a run on a time series using parameters #3. The red lines for each trace plot corresponds to the correct value.

Figure C.2: Trace plots for a run on a time series using parameters #14. The red lines for each trace plot corresponds to the correct value.

Figure C.3: Trace plots for a run on a time series using parameters #2. The red lines for each trace plot corresponds to the correct value.

Figure C.4: Trace plots for a run on a time series using parameters #17. The red lines for each trace plot corresponds to the correct value.

Figure C.5: Trace plots for a run on a time series using parameters #4. The red lines for each trace plot corresponds to the correct value.

Figure C.6: Trace plots for a run on a time series using parameters #12. The red lines for each trace plot corresponds to the correct value.

Figure C.7: Trace plots for a run on a time series using parameters #26. The red lines for each trace plot corresponds to the correct value.

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY