Stockholms universitet

Characterizing Processing Times with Mixture Distributions

Simon Berggren

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2018:2
Matematisk statistik
Mars 2018

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Master Thesis **2018:2**
http://www.math.su.se

Characterizing Processing Times with Mixture Distributions

Simon Berggren*

March 2018

Abstract

In this text, we characterize the processing times of a large number of errands. Our goal with this characterization is to provide certain key values, such as a number of quantiles and the expected value. We propose a (finite) mixture distribution, more specifically a weighted mixture of a lognormal distribution and a truncated normal distribution. Goodness of fit of this mixture model is tested using the Kolmogorov- Smirnof statistic. We also develop a methodology, using maximum likelihood techniques and an iterative Expectation Maximization (EM) algorithm, for estimation of the parameters that define the mixture distribution. Occasionally, our processing times are only partially observed, i.e. some of the values are unknown at the time of estimation. To deal with this, an extension of the Stochastic EM (SEM) algorithm is adapted, and combined with the above mentioned estimation methodology for fully observed processing times. In this extended algorithm, repeatedly updated intermediate estimates, based on both the observable and randomly generated data, are weighted to produce the parameter estimates. Computer programs that are needed for parameter estimation, computation of the key values, evaluation of the models, etc. are implemented in C++ and R as a part of the work. The outlined methodology is assessed on simulated data, and finally applied to subsets of our real word data set.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: simon.berggren@gmail.com. Supervisor: Ola Hössjer.

Contents

1 Introduction **4**
- 1.1 Characterization of Processing Times . . . 5
- 1.2 Available Work . . . 5
- 1.3 Key Values . . . 6
- 1.4 Notation and Preliminaries . . . 6
- 1.5 Outline . . . 7

2 Model Description **8**
- 2.1 Overview . . . 8
- 2.2 Model – $\mathcal{M}$ . . . 9
- 2.3 Discussion . . . 10

3 Data **13**
- 3.1 Description of Data . . . 13
- 3.2 Cohorts . . . 14
- 3.3 Truncation of Data . . . 14

4 Theory **15**
- 4.1 Mixture of a *Lognormal* and a *Truncated Normal* Distribution 15
- 4.2 The *Lognormal* and *Truncated Lognormal* Distribution . . . 18
- 4.3 The Truncated Normal Distribution . . . 22
- 4.4 Testing Hypothesis . . . 23
  - 4.4.1 Discussion . . . 25
  - 4.4.2 Kolmogorov-Smirnof Test . . . 25
  - 4.4.3 Kolmogorov-Smirnof Test and Composite Hypotheses . 27
- 4.5 Parameter Estimation in a Mixture Distribution . . . 28
  - 4.5.1 A General Description . . . 28
  - 4.5.2 Parameter Estimation - $\mathcal{M}$ . . . 30
  - 4.5.3 Initialization and Convergence . . . 32
  - 4.5.4 Heuristics . . . 33
- 4.6 Estimation with Partially Observed Data . . . 35
  - 4.6.1 General Description . . . 35
  - 4.6.2 Application to $\mathcal{M}$ . . . 36
  - 4.6.3 Heuristics . . . 36
  - 4.6.4 Initialization and Convergence . . . 37
  - 4.6.5 Remark . . . 37
  - 4.6.6 Discussion . . . 37

5 Computer Software – Implementation **39**

5.1 Discussion 39

6 Experimental Results **40**
6.1 Parameter Estimation in a Mixture Model. 40
6.2 Parameter Estimation with Partially Observed Data 40
6.2.1 Experiment 1 42
6.2.2 Experiment 2 43
6.2.3 Experiment 3 43
6.2.4 Discussion 44
6.3 Real World Data - $\mathcal{D}$ 50
6.3.1 Discussion 51

7 Discussion **60**
7.1 Limitations and Further Work 60
7.1.1 Smaller Cohorts 60
7.1.2 Theoretical Motivation 60
7.1.3 Estimation 61
7.1.4 A Shorter Horizon 61
7.2 Summary 62

Appendices **63**

A Computer Software - List of Functions **63**

1 Introduction

A phenomenon that appears across numerous fields arises when many similar tasks are being processed in a similar fashion. Examples of such phenomena are the processing of applications for membership in a community, or the legal reviews of many similar cases. Developing models that provide the answer to questions about such processes are therefore important. These models may be used to, for example:

- optimize the process,
- provide information and increase our understanding about the process,
- make predictions for future instances to be processed,
- etc.

In this text, we develop a model in order to characterize processing times.

1.1 Characterization of Processing Times

We have data, collected by an agency that processes a large number of errands. The agency would like to gain understanding of the underlying process, to be able to optimize it, and to communicate information about the processing times in terms of a number of key values. These key values are presented in Section 1.3.

The main task of this work is therefore to characterize the distribution of the processing times. A parametric model, flexible enough to capture complex patterns in data, but simple enough to not overfit it, is desired. Ideally, one would consider exclusively closed errands, i.e. data points where the processing time is known, when characterizing the processing times. However, it is reasonable to believe that the distribution of the processing times vary over time, and to wait for all errands to be processed before using a cohort may not be reasonable. Therefore, a strategy for parameter-estimation using "partially observed" data is developed and applied.

In summary, the main tasks of this work involve

1) deduction of a parametric model $\mathcal{M}$, that describes the processing times,

2) verification of $\mathcal{M}$,

3) development of a strategy for estimation of the parameters that define $\mathcal{M}$ on partially observed data,

4) provide runtime-feasible implementations to perform the above listed tasks.

Due to privacy policies, very little information about data, and about the underlying process, will be presented in this text.

1.2 Available Work

In similar settings, processing times have been characterized using the *lognormal* distribution, see [7] and [8]. The *law of the proportionate effect*, or *Gibrat's Law*, also provides a theoretical explanation to why the *lognormal* assumption is reasonable, see e.g. [1] and [5]. However, a *lognormal* distribution does not fit our processing times, and therefore a more complex model has to be developed.

1.3 Key Values

Information about the processing times will be communicated in terms of a number of key values, namely:

- the median ($Q_{0.5}$),
- the 75th percentile ($Q_{0.75}$),
- the 90th percentile ($Q_{0.9}$),
- the 95th percentile ($Q_{0.95}$),
- the mean, and
- the standard deviation.

1.4 Notation and Preliminaries

In this section we introduce notation and terminology to be used in the following text.

- $\mathcal{N}(\mu, \sigma)$ denotes the normal distribution with mean μ and standard deviation σ.
- $\phi(\cdot)$ and $\Phi(\cdot)$ denote the *density* and the *distribution function* of a *standard normal*, i.e. a $\mathcal{N}(0, 1)$-distributed, random variable, respectively.
- $\mathbb{1}$ denotes the *indicator function*, i.e. $\mathbb{1}(bool) = 1$ if *bool* evaluates to true, and $\mathbb{1}(bool) = 0$ otherwise.
- Q_d, where $d \in [0, 1]$, denotes the $d \cdot 100$th percentile. Hence, if Q refers to a random variable X, then

  $$Q_d = \inf\{x : \mathbb{P}(X \leq x) \geq d\}. \tag{1}$$

  Which distribution Q refers to should be clear from the context.
- CDF will refer to the *Cumulative Distribution Function.*
- The shorthand std. dev. will occasionally be used to denote *standard deviation.*
- The shorthand edf will occasionally be used to denote the *empirical distribution function* of a sample.

1.5 Outline

In Section 2, the model that will be used to characterize our processing times is introduced, and the problem at hand is described in greater detail. Section 3 describes the data that inspired this work, and carries the information about our processing times. In Section 4, the mixture distribution that is used to characterize the processing times, the algorithms that are used to estimate parameters, the methods that are used to assess the developed methodology, etc. are described, from a theoretical viewpoint. Section 5 (and Appendix A) provides a brief treatment of the computer software that has been developed during the work. Results from experiments, i.e. from applying the developed methods on simulated data and on the "real world" data described in Section 3, are presented in Section 6. Finally, a short summary and a discussion about limitations and further work is presented in Section 7.

2 Model Description

In this section, the model that is used to characterize the processing times is introduced and discussed. More detailed descriptions of the components of the model are given in Section 4.

2.1 Overview

According to the *law of proportionate effect*, also referred to as *Gibrat's Law*, it is reasonable to assume that processing times, with similar characteristics, follow a *lognormal* distribution, see [13] and [5].[1] Figure 1 presents a histogram over processing times in cohort 1, which is specified in Section 3.2, where this *lognormal* assumption holds. The red curve is the density of the mixture model, explained in Section 4.1, parametrized by

$$\theta = (3.2, 1.3, 1, 99.5, 79),$$

where the two first numbers define a *lognormal* component, the last two define a truncated normal component, and the third weights those components into a mixture distribution. This value of θ corresponds to a pure *lognormal* distribution, since the weight for the normally distributed component, 1 minus the middle component of θ, equals 0. The statistic from a *Kolmogorov Smirnof*-test, a test to assess validity of the hypothesis that data follows the hypothesized, *lognormal* distribution,

$$\sqrt{n}D_n = 0.777,$$

indicates a good fit (see Section 4.4.2). Figure 2 presents the empirical distribution function (the black piecewise constant "curve") for cohort 1, together with the CDF of the mixture distribution, parametrized by θ, i.e. that of a *lognormal* distribution (the red curve).

The assumption about lognormally distributed processing times holds only as an exception for our data, and generally the processing times follow a more complex pattern. The proper distribution has, at least, a bi-modal density, which motivates a mixture distribution.[2] Figure 3 presents a histogram over processing times in cohort 2, which is specified in Section 3.2,

[1] A *Weibull distribution*, or more generally a *Generalized Gamma-distribution* has also been proposed in this context. However, neither of them fit the complex distribution of our processing times very well, and therefore only the *lognormal* distribution is exemplified here.

[2] Occasionally, the density is rather *multi-modal*, than *bi-modal*.

together with the density from the mixture distribution presented in Section 4.1, parametrized by

$$\theta = (3.8, 1.8, 0.6, 397, 116.5).$$

Figure 4 presents the empirical distribution (the black piecewise constant curve) of cohort 2, together with the CDF of the mixture distribution (the red curve) for this parameters. The *Kolmogorov Smirnof*-statistic

$$\sqrt{n} D_n = 1.31,$$

provides no strong evidence against the mixture distribution (see Section 4.4.2).

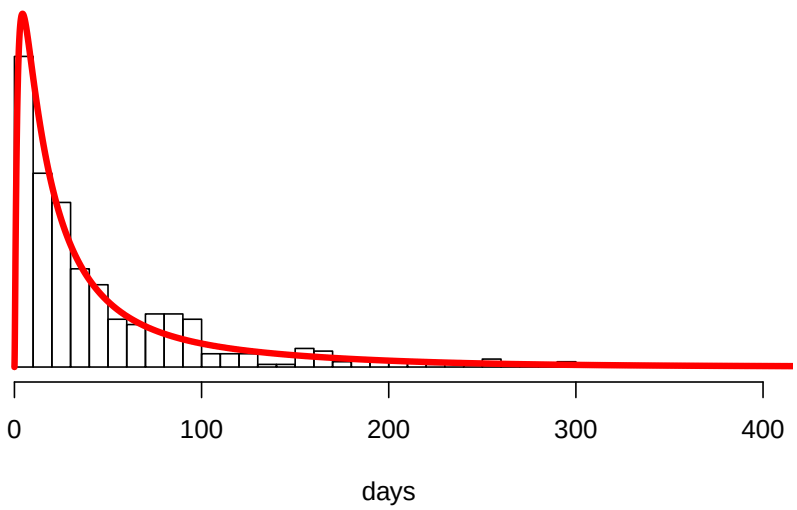

Figure 1: Histogram over processing times in cohort 1 (see Section 3.2), together with the density of the mixture model, parametrized by $\theta = (3.2, 1.3, 1, 99.5, 79)$. This parameter vector θ corresponds to a pure *lognormal* distribution, see Section 2.1.

2.2 Model – $\mathcal{M}$

The model that will be used to characterize the processing times, denote it by $\mathcal{M}$, is

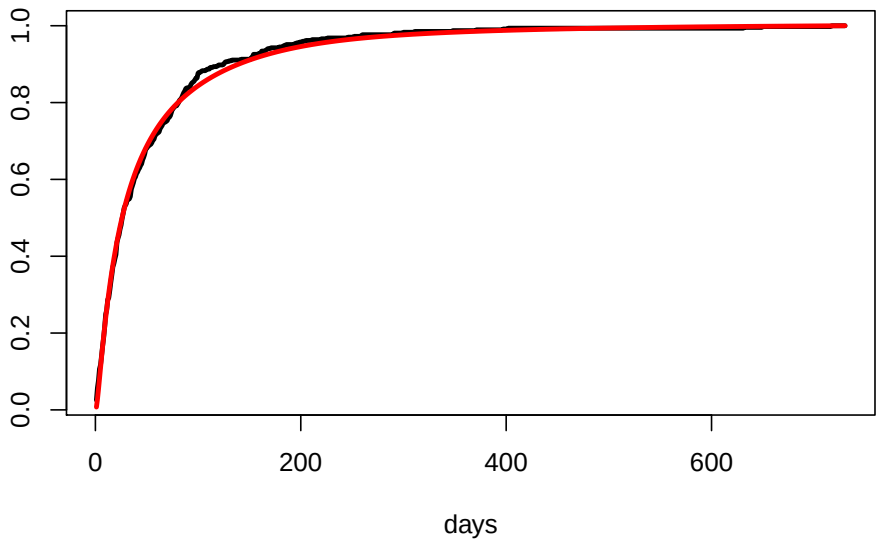

Figure 2: Empirical distribution (black curve) of processing times in cohort 1 (see Section 3.2), and theoretical CDF (red curve) of the mixture distribution, parametrized by $\theta = (3.2, 1.3, 1, 99.5, 79)$. This parameter vector θ corresponds to a pure *lognormal* distribution, see Section 2.1.

$\mathcal{M}$: *a mixture of a lognormal and a truncated normal distribution.*[3]

$\mathcal{M}$ will be explained in greater detail in Section 4, and used to analyze data in Section 6.

2.3 Discussion

The processing times are integer-valued, they measure a number of days, but $\mathcal{M}$ is based on continuous probability distributions. This is motivated by the fact that the underlying phenomenon, i.e. that of processing errands, takes place in continuous time. Furthermore, statistical evidence for $\mathcal{M}$ based on processing times that are "rounded of to integers", i.e. to discrete days, are still adequate. If anything, a good fit on such integer valued processing times strengthens our belief in $\mathcal{M}$.

It is reasonable to believe that deviation from the *lognormal* distribution is

[3]The actual distribution to be considered is truncated from above at a constant U, see Section 3.3.

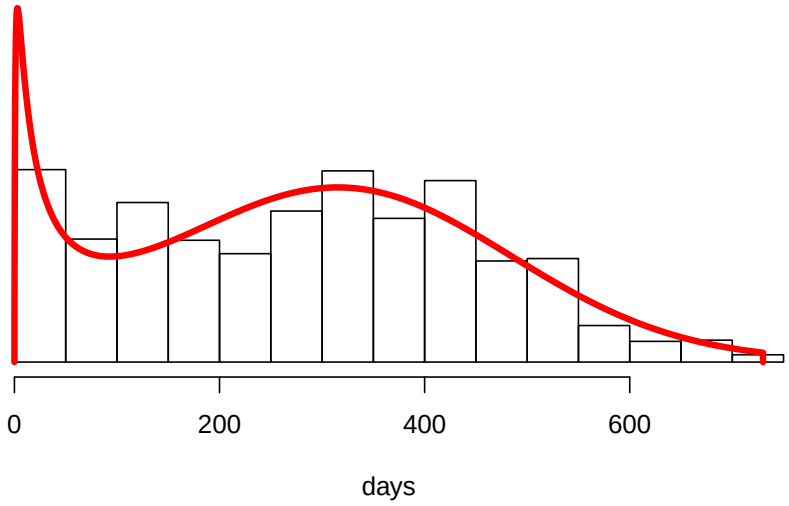

Figure 3: Histogram over processing times in cohort 2 (see Section 3.2), together with the density of the mixture distribution, parametrized by $\theta = (3.8, 1.8, 0.6, 397, 116.5)$, see Section 2.1.

due to some disturbance, and that lack of fit is due to insufficiently well-defined cohorts. However, there are no means to partition our data into such cohorts, and after all the point of having a probabilistic modeling framework is that of simplification and generalization. See Section 7.1.1 for a further discussion about partitioning our data into smaller cohorts.

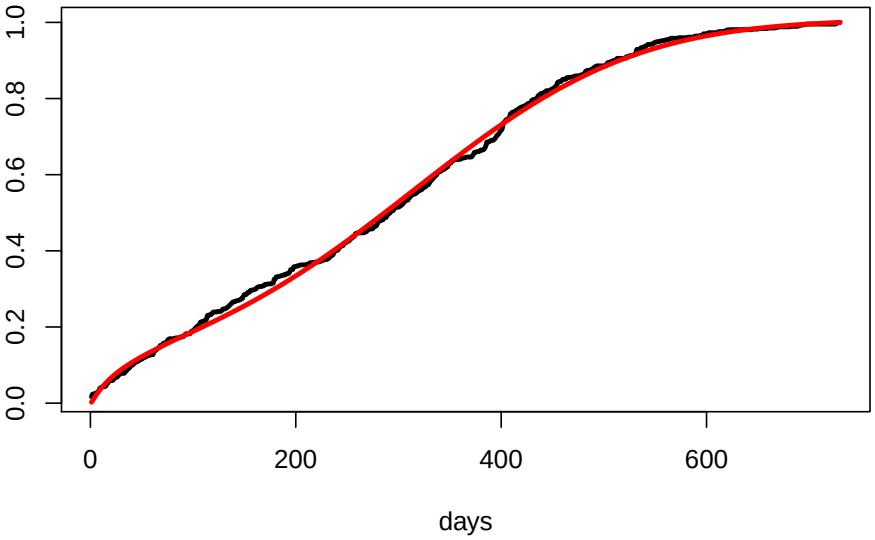

Figure 4: Empirical distribution (black curve) of processing times in cohort 2 (see Section 3.2), and theoretical CDF (red curve) of the mixture distribution, parametrized by $\theta = (3.8, 1.8, 0.6, 397, 116.5)$, see Section 2.1.

3 Data

In this section, we present the data that carries information about the processing times that inspired this work.

3.1 Description of Data

The data is collected by an agency that continuously processes a large number of errands. It contains information on:

- date of opening (YYYY-MM-DD),
- date of closing (YYYY-MM-DD), and
- several categorical variables.

From this we deduce the crucial entity

$$\textit{processing time} := \text{date of closing} - \text{date of opening} + 1, \tag{2}$$

which will be denoted by X. Two of the categorical variables, `COUNTRY` and `CATEGORY`, will be used to partition the data into smaller cohorts. `COUNTRY` describes the country of origin of the object being processed, and `CATEGORY` categorizes the processes into different groups, based on other information about the objects. Due to privacy policies, no further information about the data, or about the underlying process it describes, is provided.

In this text, only a subset of data, denote this by $\mathcal{D}$, where

- the date of opening belongs to the first half of 2015, and
- the processing times satisfy $X \leq 2 \cdot 365$,[4]

will be used.

There are no *missing values* in $\mathcal{D}$.[5]

[4]Occasionally, a tighter upper bound on the processing times are used, i.e. $2 \cdot 365$ is replaced by a smaller number. In those cases, the reader will be informed.

[5]To be precise, erroneous registrations that result in undefined fields are filtered out, and therefore $\mathcal{D}$ is free from missing values.

3.2 Cohorts

In this section, the cohorts that are used in examples and experiments in this text, are presented.

cohort 1:
Data in cohort 1 is aggregated over all levels of `COUNRTY` and no restrictions on the opening date, except for those presented in Section 3.1, are imposed. `CATEGORY` is fixed on one of its levels.

cohort 2:
Data in cohort 2 is aggregated over all levels of `CATEGORY`. `COUNTRY` is fixed on one of its levels, and all errands are opened during the first month of 2015.

cohort 3 − 6:
Data in these cohorts are aggregated over all levels of `CATEGORY`, but contains exclusively one level of `COUNTRY`. Processes are opened during months 2015-01, 2015-02, 2015-03, and 2015-04 in cohorts 3, 4, 5 and 6 respectively. Furthermore, the processing times satisfy

$$X \leq 1.8 \cdot 365 = 657.$$

Cohorts 3 through 6 are used in the experiments presented in Section 6.3.

3.3 Truncation of Data

Due to errors in the registration of closing dates of processed errands, some data points have unreasonably long processing times. Knowledge about the underlying process motivates the truncation of $\mathcal{D}$ that was presented in Section 3.1 ($X \leq 2{\cdot}365$).[6] This truncation also effects the model $\mathcal{M}$, as discussed in Section 4.1.

[6]Note that cohorts 3-6 are truncated at a lower value, namely $1.8 \cdot 365$.

4 Theory

This section provides a description about the model that is used to characterize our processing times. We also describe how this model is assessed, and how the parameters that define it are estimated.

4.1 Mixture of a *Lognormal* and a *Truncated Normal* Distribution

$\mathcal{M}$ specifies that a processing time, X, has density

$$f_X(x) = \beta f_Y(x) + (1-\beta) f_V(x), \tag{3}$$

where $f_Y(x)$ is the density of a *lognormal* random variable Y with parameters $\mu_1, \sigma_1 > 0$, f_V is the density of a normal random variable V with parameters $\mu_2, \sigma_2 > 0$, truncated from below at 0, and $\beta \in [0,1]$ is a real number. More specifically,

$$V = W | W > 0, \tag{4}$$

where $W \in \mathcal{N}(\mu_2, \sigma_2)$. Equivalently, X has distribution function

$$\mathbb{P}(X < x) = \beta \mathbb{P}(Y < x) + (1-\beta)\mathbb{P}(V < x) =$$

$$\beta \Phi(\tfrac{\log x - \mu_1}{\sigma_1}) + (1-\beta) \left(\frac{\Phi(\frac{x-\mu_2}{\sigma_2}) - \Phi(\frac{-\mu_2}{\sigma_2})}{1 - \Phi(-\frac{\mu_2}{\sigma_2})} \right), \tag{5}$$

for $x > 0$. Hence, $\mathcal{M}$ is completely defined by five parameters

$$\theta = (\mu_1, \sigma_1, \beta, \mu_2, \sigma_2). \tag{6}$$

The two first moments of the mixture distribution,

$$\mathbb{E}(X) = \beta \mathbb{E}(Y) + (1-\beta)\mathbb{E}(V), \tag{7}$$

and

$$\mathbb{E}(X^2) = \beta \mathbb{E}(Y^2) + (1-\beta)\mathbb{E}(V^2), \tag{8}$$

follow directly from integrating Equation (3) componentwise. We get the variance

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 =$$

$$\beta(\mathbb{E}(Y^2) - \mathbb{E}(Y)^2) + \beta\mathbb{E}(Y)^2 + (1-\beta)(\mathbb{E}(V^2) - \mathbb{E}(V)^2) + (1-\beta)\mathbb{E}(V)^2$$
$$-\beta^2\mathbb{E}(Y)^2 - (1-\beta)^2\mathbb{E}(V)^2 - 2\beta(1-\beta)\mathbb{E}(Y)\mathbb{E}(V) =$$

$$\beta\text{Var}(Y) + (1-\beta)\text{Var}(V) + \beta(1-\beta)\left(\mathbb{E}(Y) - \mathbb{E}(V)\right)^2, \tag{9}$$

which is the familiar linear combination of the two components' variances, plus a term that is proportional to the difference between their expected values.

Due to practical reasons – at each point in time there is an upper bound on the length of processing times that can be observed – the observed processing times in $\mathcal{D}$ is truncated from above, see Section 3.3.[7] Therefore, when e.g. estimating the parameters θ from $\mathcal{D}$, one should consider

$$X_U := X | X < U, \tag{10}$$

for some upper bound $U > 0$, and the density in Equation (3) becomes

$$f_{X_U}(x) = \frac{f_X(x)}{\mathbb{P}(X < U)} \mathbb{1}(x \leq U) \tag{11}$$

where $\mathbb{1}(\cdot)$ is the *indicator*-function (see Section 1.4).

The two first moments for the truncated variable X_U are

$$\mathbb{E}(X_U) = \frac{\int_0^U x f_X(x)dx}{\mathbb{P}(X<U)} =$$

$$\frac{1}{\mathbb{P}(X<U)}\left(\beta\int_0^U x f_Y(x)dx + (1-\beta)\int_0^U x f_V(x)dx\right) =$$

$$\frac{1}{\mathbb{P}(X<U)}\left[\beta\mathbb{E}(Y_U)\Phi(\tfrac{\log U - \mu_1}{\sigma_1}) + (1-\beta)\mathbb{E}(V_U)\left(\Phi(\tfrac{U-\mu_2}{\sigma_2}) - \Phi(\tfrac{-\mu_2}{\sigma_2})\right)\right], \tag{12}$$

[7] The fact that some processes might abort without being deregistered sometimes motivate an even "tighter" truncation.

and

$$\mathbb{E}(X_U^2) = \frac{\int_0^U x^2 f_X(x)dx}{\mathbb{P}(X<U)} =$$

$$\frac{1}{\mathbb{P}(X<U)} \left(\beta \int_0^U x^2 f_Y(x)dx + (1-\beta) \int_0^U x^2 f_V(x)dx\right) =$$

$$\frac{1}{\mathbb{P}(X<U)} \left[\beta \mathbb{E}(Y_U^2)\Phi(\tfrac{\log U - \mu_1}{\sigma_1}) + (1-\beta)\mathbb{E}(V_U^2)\left(\Phi(\tfrac{U-\mu_2}{\sigma_2}) - \Phi(\tfrac{-\mu_2}{\sigma_2})\right)\right], \tag{13}$$

where

$$Y_U \stackrel{d}{=} Y \Big| Y < U \tag{14}$$

and

$$V_U \stackrel{d}{=} W \Big| 0 < W < U. \tag{15}$$

No explicit formulas for the mean and variance of X_U are derived, but those are computed as functions of the moments of Y_U and V_U, through (12), (13) and

$$\text{Var}(X_U) = \mathbb{E}(X_U^2) - \mathbb{E}(X_U)^2. \tag{16}$$

See Sections 4.2 and 4.3 for details about the *lognormal* and the *truncated normal* distributions respectively, and e.g. [10] to read more about mixture models.

Table 1 presents the mean value, the standard deviation, and three quantiles of X_U, as defined in Equation (11), for parameters

$$\theta = (4.7,\ 0.6,\ 0.8,\ 365.0,\ 80.0),$$

and $U = 547$. Figure 5 present the density of X_U, together with some characteristics, for this parameter vector θ and U.

	mean	std. dev.	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.9}$
X_U	189.5	131.4	82.7	143.9	394.8

Table 1: Characteristics for X_U, defined in Equation (11) in Section 4.1, with $\theta = (4.7,\ 0.6,\ 0.8,\ 365.0,\ 80.0)$ and $U = 547$.

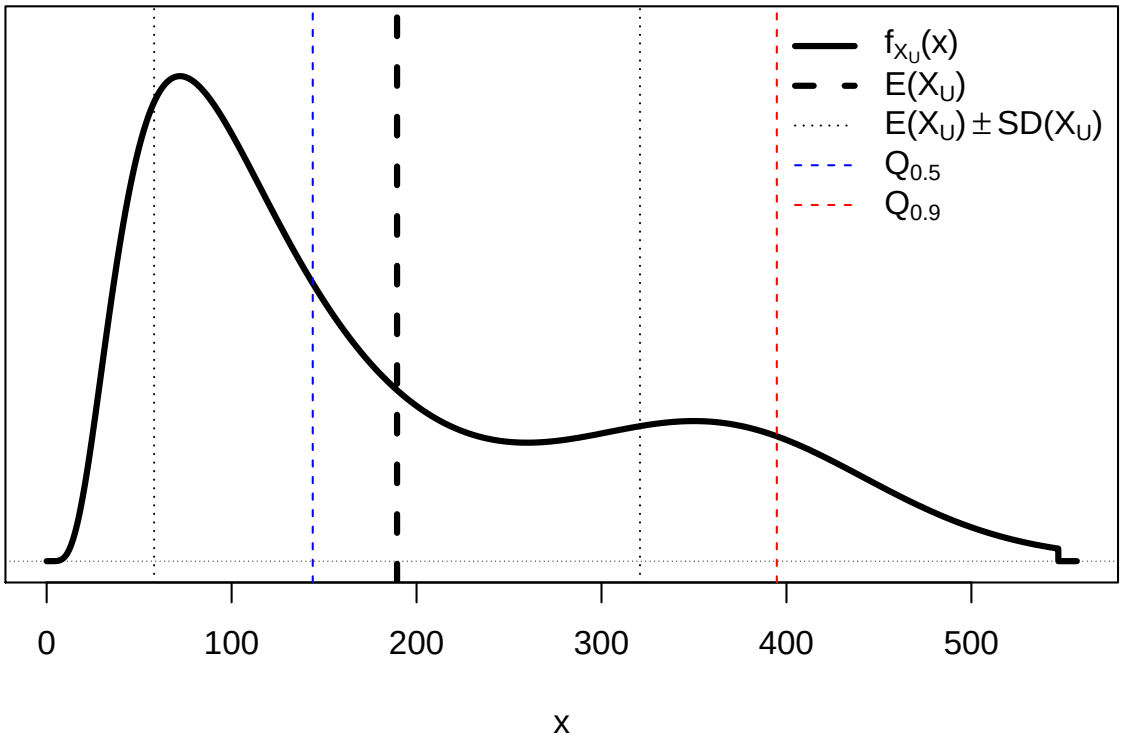

Figure 5: Density of X_U, defined in Equation (11), Section 4.1, for $\theta = (4.7,\ 0.6,\ 0.8,\ 365.0,\ 80.0)$ and $U = 547$. The vertical lines represents some characteristics of X_U, and SD refer to the *standard deviation*.

4.2 The *Lognormal* and *Truncated Lognormal* Distribution

The *lognormal* distribution constitutes one of the components in the mixture distribution $\mathcal{M}$. Y has a *lognormal* distribution, with parameters μ_1 and σ_1, if

$$Y \stackrel{d}{=} e^{\sigma_1 Z + \mu_1}, \tag{17}$$

where Z is a standard normal random variable, or equivalently if

$$\log Y \in \mathcal{N}(\mu_1, \sigma_1). \tag{18}$$

Clearly, $Y \geq 0$. Y has distribution function

$$\mathbb{P}(Y < y) = \mathbb{P}(e^{\sigma_1 Z + \mu_1} < y) = \Phi\left(\frac{\log y - \mu_1}{\sigma_1}\right), \tag{19}$$

and density

$$f_Y(y) = \Phi'\left(\tfrac{\log y - \mu_1}{\sigma_1}\right) = \frac{1}{\sigma_1 y}\phi(\tfrac{\log y - \mu_1}{\sigma_1}) =$$

$$\frac{1}{\sqrt{2\pi}\sigma_1\ y}\exp\left(-\frac{(\log y - \mu_1)^2}{2\sigma_1^2}\right). \tag{20}$$

Using the variable substitution

$$y = e^{v\sigma_1 + \mu} \Leftrightarrow v = (\log y - \mu_1)/\sigma_1, \tag{21}$$

we derive the mean

$$\mathbb{E}(Y) = \int_0^\infty y f_Y(y)dy = \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(v-\sigma_1)^2} dv \cdot e^{\mu + \sigma_1^2/2} = e^{\mu_1 + \sigma_1^2/2}, \tag{22}$$

the variance

$$\textsc{Var}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 =$$

$$\int_0^\infty \frac{y}{\sqrt{2\pi}\sigma_1}\exp(-\tfrac{1}{2\sigma_1^2}(\log y - \mu_1)^2)dy - e^{2\mu_1 + \sigma_1^2} =$$

$$(e^{\sigma_1^2} - 1)e^{2\mu_1 + \sigma_1^2}, \tag{23}$$

and the median e^{μ_1}, since

$$\int_0^{e^{\mu_1}} \frac{1}{\sqrt{2\pi}\sigma_1\ y}\exp(-\frac{1}{2\sigma_1^2}(\log y - \mu_1)^2)dy = \int_{-\infty}^0 \frac{1}{\sqrt{2\pi}} e^{-v^2/2} dv = \frac{1}{2}. \tag{24}$$

Note that the median satisfies

$$e^{\mu_1} \leq e^{\mu_1} e^{\sigma_1^2/2} = \mathbb{E}(Y),$$

with equality in the degenerate case $\sigma_1 = 0$, i.e. the *lognormal* distribution is right skewed.

We derive the moments of

$$Y_U = Y | Y < U \tag{25}$$

for some $U > 0$, that was used in Section 4.1. We have density

$$f_{Y_U}(y) = \frac{f_Y(y)}{\mathbb{P}(Y < U)}\mathbb{1}(y < U) = \frac{f_Y(y)}{\Phi(\frac{\log U - \mu_1}{\sigma_1})}\mathbb{1}(y < U), \tag{26}$$

and mean

$$\mathbb{E}(Y_U) = \frac{1}{\mathbb{P}(Y < U)} \int_0^U y f_Y(y) dy =$$

$$\frac{\mathbb{E}(Y) - \int_U^\infty \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(\log y - \mu_1)^2}{2\sigma_1^2}} dy}{\Phi(\frac{\log U - \mu_1}{\sigma_1})}. \tag{27}$$

Using the variable substitution presented in Equation (21), we get

$$\int_U^\infty \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(\log y - \mu_1)^2}{2\sigma_1^2}} dy = \int_{(\log U - \mu_1)/\sigma_1}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(v - \sigma_1)^2} dv \; e^{\mu_1 + \sigma_1^2/2} =$$

$$e^{\mu_1 + \sigma_1^2/2} \int_{(\log U - \mu_1)/\sigma_1 - \sigma_1}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt =$$

$$e^{\mu_1 + \sigma_1^2/2} \left[1 - \Phi(\frac{\log U - \mu_1 - \sigma_1^2}{\sigma_1}) \right], \tag{28}$$

and with $\mathbb{E}(Y) = e^{\mu_1 + \sigma_1^2/2}$,

$$\mathbb{E}(Y_U) = e^{\mu_1 + \frac{\sigma_1^2}{2}} \frac{\Phi(\frac{\log U - \mu_1 - \sigma_1^2}{\sigma_1})}{\Phi(\frac{\log U - \mu_1}{\sigma_1})}. \tag{29}$$

Similarly we obtain

$$\int_U^\infty y^2 f_V(y) dy = e^{2(\sigma_1^2 + \mu_1)} \int_{(\log U - \mu_1)/\sigma_1}^\infty \frac{e^{-\frac{1}{2}(y - 2\sigma_1)^2}}{\sqrt{2\pi}} dy =$$

$$e^{2\sigma_1^2 + 2\mu_1} \left[1 - \Phi(\frac{\log U - \mu_1 - 2\sigma_1^2}{\sigma_1}) \right], \tag{30}$$

and

$$\mathbb{E}(Y_U^2) = \frac{\int_0^U y^2 f_V(y) dy}{\Phi(\frac{\log U - \mu_1}{\sigma_1})} = \frac{\mathbb{E}(Y^2) - \int_U^\infty y^2 f_V(y) dy}{\Phi(\frac{\log U - \mu_1}{\sigma_1})} =$$

$$e^{2\mu_1 + 2\sigma_1^2} \frac{\Phi(\frac{\log U - \mu_1 - 2\sigma_1^2}{\sigma_1})}{\Phi(\frac{\log U - \mu_1}{\sigma_1})}. \tag{31}$$

More details about the *lognormal* distribution can be found in e.g. [9].

Table 2 presents the mean, the standard deviation, and three quantiles, of a *lognormal* random variable Y with parameters $\mu_1 = 4.8$ and $\sigma_1 = 0.8$, and of it's truncated version Y_U, as defined in Equation (26), for $U = 547$. Figure 6 presents the densities of Y and Y_U, together with some characteristics, for these parameters and U.

	mean	std. dev.	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.9}$
Y	174.4	179.5	68.5	121.5	361.2
Y_U	148.9	110.5	66.7	116.6	310.9

Table 2: Characteristics of a *lognormal* variable Y, and its' truncated version Y_U, as defined in Equation (26), for $\mu_1 = 4.8$, $\sigma_1 = 0.8$ and $U = 547$.

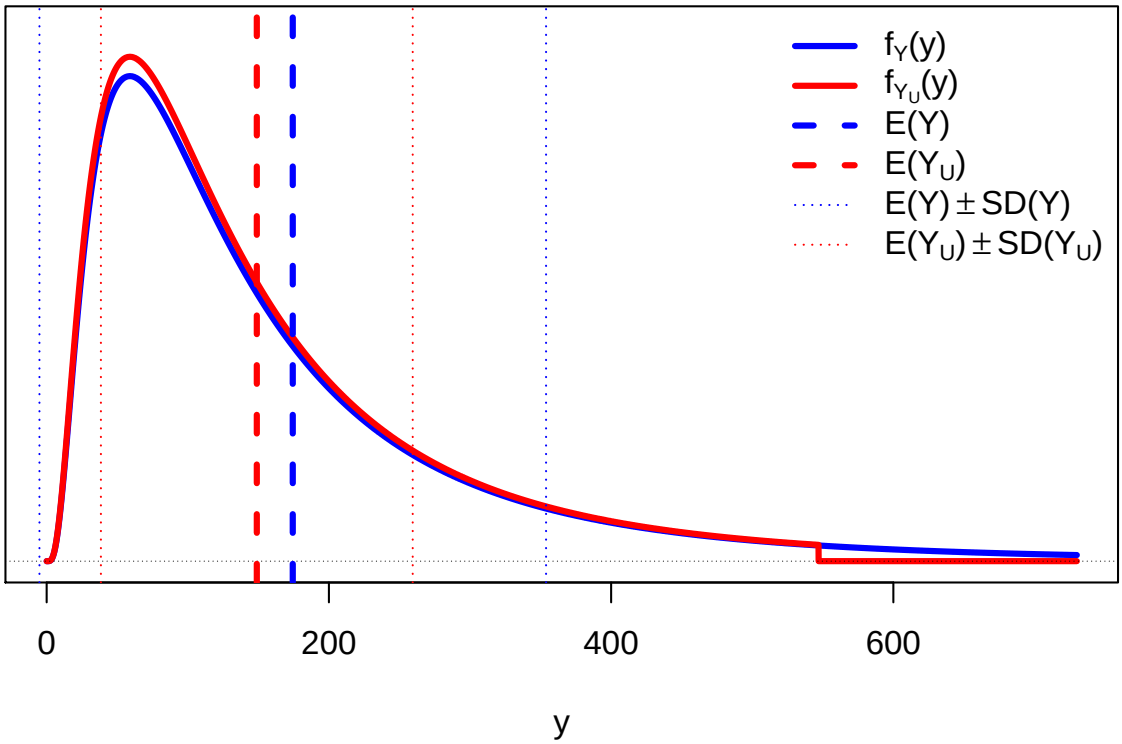

Figure 6: Densities of a *lognormal* variable Y, and its' truncated version Y_U, as defined in Equation (26), for $\mu_1 = 4.8$, $\sigma_1 = 0.8$ and $U = 547$. The vertical lines represents some characteristics of Y and Y_U, and SD refer to the *standard deviation*.

4.3 The Truncated Normal Distribution

The *truncated normal distribution* constitutes one of the components in the mixture distribution $\mathcal{M}$. Let W be a normal random variable with parameters μ_2 and σ_2, and consider

$$V_U = W|0 < W < U, \tag{32}$$

for some real number $U > 0$.

In order to deduce some important properties of the truncated normal distribution, we derive the moment-generating function

$$M_{V_U}(t) = \mathbb{E}(e^{tW}|0 < W < U) =$$

$$\tfrac{1}{\mathbb{P}(0<W<U)} \int_0^U \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\tfrac{1}{2\sigma_2^2}(y^2 - 2y(\mu_2 + t\sigma_2^2) + \mu_2^2)\right) dy =$$

$$\tfrac{1}{\mathbb{P}(0<W<U)} \exp\left(-\tfrac{1}{2\sigma_2^2}(\mu_2^2 - (\mu_2 + t\sigma_2^2)^2)\right) \int_0^U \tfrac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\tfrac{1}{2\sigma_2^2}(y - \mu_2 - t\sigma_2^2)^2\right) dy =$$

$$e^{\mu_2 t + t^2\sigma_2^2/2} \frac{\Phi(\frac{U-\mu_2-\sigma_2^2 t}{\sigma_2}) - \Phi(\frac{-\mu_2-\sigma_2^2 t}{\sigma_2})}{\Phi(\frac{U-\mu_2}{\sigma_2}) - \Phi(-\frac{\mu_2}{\sigma_2})} \tag{33}$$

Differentiation of $M_{V_U}(t)$ with respect to t gives

$$M'_{V_U}(t) = e^{\mu_2 t + t^2\sigma_2^2/2}(\mu_2 + t\sigma_2^2) \frac{\Phi(\frac{U-\mu_1-\sigma_2^2 t}{\sigma_2}) - \Phi(\frac{-\mu_2-\sigma_2^2 t}{\sigma_2})}{\Phi(\frac{U-\mu_2}{\sigma_2}) - \Phi(-\frac{\mu_2}{\sigma_2})}$$

$$-e^{\mu_2 t + t^2\sigma_2^2/2}\sigma_2 \frac{\phi(\frac{U-\mu_1-\sigma_2^2 t}{\sigma_2}) - \phi(\frac{-\mu_2-\sigma_2^2 t}{\sigma_2})}{\Phi(\frac{U-\mu_2}{\sigma_2}) - \Phi(-\frac{\mu_2}{\sigma_2})}. \tag{34}$$

Letting $t \to 0$ we get

$$\mathbb{E}(V_U) = M'_{V_U}(t)\Big|_{t=0} = \mu_2 - \sigma_2 \frac{\phi(\frac{U-\mu_2}{\sigma_2}) - \phi(\frac{\mu_2}{\sigma_2})}{\Phi(\frac{U-\mu_2}{\sigma_2}) - \Phi(-\frac{\mu_2}{\sigma_2})}. \tag{35}$$

Similarly

$$\mathbb{E}(V_U^2) = M''_{V_U}(t)\Big|_{t=0} =$$

$$\mu_2^2 + \sigma_2^2 - 2\mu_2\sigma_2 \left(\frac{\phi(\frac{U-\mu_2}{\sigma_2}) - \phi(\frac{\mu_2}{\sigma_2})}{\Phi(\frac{U-\mu_2}{\sigma_2}) - \Phi(-\frac{\mu_2}{\sigma_2})} \right) + \sigma_2 \frac{\left(\phi(\frac{U-\mu_2-\sigma_2^2 t}{\sigma_2}) - \phi(\frac{-\mu_2-\sigma_2^2 t}{\sigma_2})\right)'\Big|_{t=0}}{\Phi(\frac{U-\mu_2}{\sigma_2}) - \Phi(-\frac{\mu_2}{\sigma_2})} =$$

$$\mu_2^2 + \sigma_2^2 - 2\mu_2\sigma_2 \left(\frac{\phi(\frac{U-\mu_2}{\sigma_2}) - \phi(\frac{\mu_2}{\sigma_2})}{\Phi(\frac{U-\mu_2}{\sigma_2}) - \Phi(-\frac{\mu_2}{\sigma_2})} \right) + \sigma_2^2 \left(\frac{\frac{U-\mu_2}{\sigma_2} \cdot \phi(\frac{U-\mu_2}{\sigma_2}) + \frac{\mu_2}{\sigma_2} \cdot \phi(\frac{\mu_2}{\sigma_2})}{\Phi(\frac{U-\mu_2}{\sigma_2}) - \Phi(-\frac{\mu_2}{\sigma_2})} \right), \tag{36}$$

and

$$\textsc{Var}(V_U) = \mathbb{E}(V_U^2) - \mathbb{E}(V_U)^2$$

$$= \sigma_2^2 \left(1 + \left(\frac{\frac{U-\mu_2}{\sigma_2} \cdot \phi(\frac{U-\mu_2}{\sigma_2}) + \frac{\mu_2}{\sigma_2} \cdot \phi(\frac{\mu_2}{\sigma_2})}{\Phi(\frac{U-\mu_2}{\sigma_2}) - \Phi(-\frac{\mu_2}{\sigma_2})} \right) - \left(\frac{\phi(\frac{U-\mu_2}{\sigma_2}) - \phi(\frac{\mu_2}{\sigma_2})}{\Phi(\frac{U-\mu_2}{\sigma_2}) - \Phi(-\frac{\mu_2}{\sigma_2})} \right)^2 \right). \tag{37}$$

Figure 7 presents W (the blue curve) and V_U (the red curve), together with some characteristics, for $\mu_2 = 200.0$, $\sigma_2 = 150.0$, and $U = 547$. Table 3 presents the mean, the standard deviation, and three quantiles of W and V_U, for these parameter values and U.

	mean	std. dev.	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.9}$
W	200.0	150.0	98.8	200.0	392.2
V_U	222.8	170.4	128.1	215.2	392.1

Table 3: Characteristics of a normal random variable W, and of V_U as defined in Equation (32), for $\mu_2 = 200.0$, $\sigma_2 = 150.0$, and $U = 547$.

4.4 Testing Hypothesis

Our basic assumption is that $\mathcal{M}$ provides an adequate description of the processing times X, i.e. our null hypothesis H_0 is that

H_0 : *the processing times follow the mixture distribution defined by* $\mathcal{M}$.

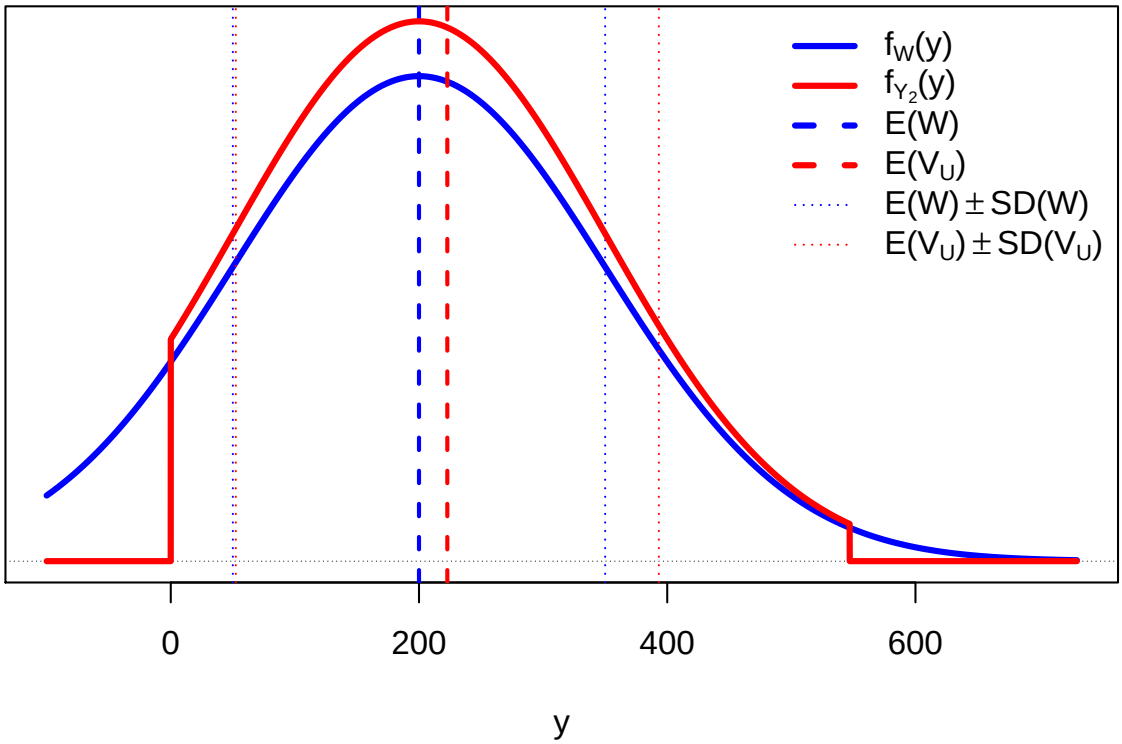

Figure 7: Densities of a normal random variable W, and of V_U as defined in Equation (32), for $\mu_2 = 200.0$, $\sigma_2 = 150.0$, and $U = 547$. The vertical lines represents some characteristics of W and V_U, and SD refer to the *standard deviation*.

This is tested against the alternative that

$$H_1: \textit{ the processing times do not follow the mixture distribution defined by } \mathcal{M}.$$

Different methods, each of which has its advantages and weaknesses, have been used when testing H_0 against the alternative H_1. In this text, we focus on the *Kolmogorov Smirnof*-test and on graphical comparisons of empirical distribution-functions to their theoretical counterparts. Also the χ^2-test has been important in the work, but it will not be presented in this text.

4.4.1 Discussion

The aim of hypothesis testing is to make an overall judgment of our assumptions, and we do not expect $\mathcal{M}$ to fit well in each cohort of our data. Occasional rejections of H_0 in some cohorts do not violate the choice of $\mathcal{M}$, as long as the overall picture is good.

4.4.2 Kolmogorov-Smirnof Test

The *Kolmogorov-Smirnof test*, named after it's inventors, is a statistical test with null-hypothesis H_0: a sample $\mathbf{X} = (X_1, \ldots, X_n)$ is realized from a continuous, one-dimensional, null-distribution F_0. The intuition is that, if H_0 holds, then the empirical distribution of the sample should be close to F_0. More specifically, let

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq x), \tag{38}$$

be the *empirical distribution function* of the sample $\mathbf{X}$. Then

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|, \tag{39}$$

measures the difference between the empirical distribution function and the hypothesized distribution function (as an example, D_n is the largest absolute distance between the black and the red curves in Figures 2 and 4). For sufficiently large n, and under H_0, the distribution of $\sqrt{n} D_n$ is well approximated by the *Kolmogorov-Smirnof*-distribution. See e.g. [11] for a more detailed description of the Kolmogorov-Smirnof test, and about the *Kolmogorov-Smirnof* distribution.

The null-hypothesis is rejected when D_n exceeds a critical value, i.e. a quantile from the Kolmogorov-Smirnof distribution. The CDF of this distribution, i.e. the probability that a random variable from the Kolmogorov-Smirnof distribution is less than say x, is

$$1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2 x^2} = \frac{\sqrt{2\pi}}{x} \sum_{k=1}^{\infty} e^{-(2k-1)^2 \pi^2 / (8x^2)}. \tag{40}$$

Equation (40) contains an infinite sum for which no explicit expression is available. A common approach is to use critical values from a table. For a

sample size $n \geq 50$, as is the case in our cohorts, we use critical values from Table 4, where

$$\mathbb{P}(\sqrt{n}D_n > D_\alpha) = \alpha.$$

The values in Table 4 are approximations that rely on D_n having a distribution function as specified in Equation (40).

α	0.02	0.05	0.1	0.15	0.2
D_α	1.51743	1.35810	1.22385	1.13795	1.07275,

Table 4: Critical values for the Kolmogorov Smirnof-test, for samples of size greater than or equal to 50.

One crucial observation when computing D_n is that the supremum in Equation (39) is always found at one of the n discontinuities of F_n, i.e. at an instance in the sample. Therefore, given that the data is ordered, we only have to perform $\mathcal{O}(n)$ computations (we have to consider $F_0(x_i) - (i-1)/n$ and $i/n - F_0(x_i)$ for $i = 1 \ldots n$).[8] Pseudocode 1, where $F_0(x)$ is the hypothesized distribution function of X, describes how the Kolmogorov-Smirnof statistic is computed.

Pseudocode 1: **Computation of the Kolmogorov-Smirnof Statistic**

```
input: parameters θ, n data points  x = (x_1, ..., x_n)
output: D_n

    sort x in ascending order;
    D <- −∞;
    for (i=1..n):
        if (F_0(x[i]) − (i − 1)/n > D) then
            D <- F_0(x[i]) − (i − 1)/n;
        if(i/n − F_0(x[i]) > D) then
            D <- i/n − F_0(x[i]);

    return D;
```

[8]Sorting the data has complexity $\mathcal{O}(n \log n)$, and this is computationally "cheap" in the present context. Therefore it is not problematic to assume that data is ordered.

4.4.3 Kolmogorov-Smirnof Test and Composite Hypotheses

One drawback of the *Kolmogorov-Smirnof test* is that the asymptotic distribution of D_n, presented in Equation (39), is invalid for composite null hypotheses. More specifically, for composite null hypothesis, the Kolmogorov-Smirnof statistic is conservative, i.e. the actual significance level of the test is smaller than the nominal one (α in Table 4). In our case this means that the computed tail-probabilities are too generous after estimating θ on the same cohort as we conduct the test on.

Monte-Carlo simulation strategies to estimate the critical region for composite hypotheses have been proposed, but this will not discussed further here. See [4] for a treatment of the Kolmogorov-Smirnof test for composite hypotheses.

Discussion

Despite the fact that critical values from Table 4 are invalid, and that the belief in a good fit based on those might become too optimistic, a low value of D_n still indicates a better fit than a higher one. Therefore, the Kolmogorov-Smirnof statistic may serve as a "relative" goodness of fit measure, and it will be presented in some of the experiments in Section 6.

4.5 Parameter Estimation in a Mixture Distribution

In this section a description of the algorithm that is used to estimate the paramter vector θ that defines $\mathcal{M}$ (see Section 2.2), is given. First follows a general description of a strategy for parameter estimation in a mixture model.

4.5.1 A General Description

One could think of a mixture distribution as describing a population where each observation comes from one of two subpopulations. More specifically, the mixture distribution corresponds to

$$X = ZY + (1 - Z)V, \tag{41}$$

where Z is a *Bernoulli* random variable with parameter $\beta \in [0, 1]$, i.e.

$$\mathbb{P}(Z = 1) = \beta = 1 - \mathbb{P}(Z = 0), \tag{42}$$

and where Y and V are random variables defined by parameters θ_1 and θ_2 respectively.

Denote by $\theta = (\theta_1, \beta, \theta_2)$ the parameters that define X. When estimating θ from a sample $\boldsymbol{x} = (x_1, \ldots, x_N)$ of X, we do not know if a given x_i is a realization of Y or V, and the log-likelihood

$$l(\theta|\boldsymbol{x}) = \sum_{n=1}^{N} \log\left(\beta f_Y(x_n|\theta) + (1 - \beta) f_V(x_n|\theta)\right) \tag{43}$$

is intractable. However, suppose that we know the "source" of each x_i, i.e. we "know"

$$\boldsymbol{z} = (z_1, \ldots, z_N),$$

with $z_i \in \{0, 1\}$ for $i = 1, \ldots, N$, the outcomes of the Bernoulli random variables. Then, for each n, one of the summands in Equation (43) cancels out and we could consider an alternative,

$$\begin{aligned} & l(\theta|\boldsymbol{x}, \boldsymbol{z}) \\ = & \sum_{n=1}^{N} \left(z_n[\log\beta + \log f_Y(x_n|\theta)] \ + \ (1 - z_n)[\log(1 - \beta) + \log f_V(x_n|\theta)]\right). \end{aligned} \tag{44}$$

We use this idea in an iterative *Expectation Maximization* (EM) algorithm to be described. See e.g. [2] or [6] for a more detailed treatment.

Given values $\theta^{(i)}$, from the ith iteration, we replace z_n in Equation (44) by

$$r_n = \frac{\beta^{(i)} f_Y(x_n|\theta^{(i)})}{\beta^{(i)} f_Y(x_n|\theta^{(i)}) + (1-\beta^{(i)}) f_V(x_n|\theta^{(i)})}. \tag{45}$$

The resulting expression. i.e. that in Equation (44) with the z_n's replaced by the r_n's, can be factored into different components corresponding to the parameters θ_1, θ_2. Hence we maximize

$$\sum_{n=1}^{N} r_n \log f_Y(x_n|\theta_1), \tag{46}$$

and

$$\sum_{n=1}^{N} (1-r_n) \log f_V(x_n|\theta_2), \tag{47}$$

separately w.r.t. θ_1 and θ_2, and

$$\sum_{n=1}^{N} (r_n \log \beta + (1-r_n)\log(1-\beta))$$

$$= \sum_{n=1}^{N} \left(r_n \log(\frac{\beta}{1-\beta}) \right) + N \log(1-\beta), \tag{48}$$

w.r.t. β, by setting

$$0 = \frac{d}{d\beta} \left(\sum_{n=1}^{N} \left(r_n \log(\frac{\beta}{1-\beta}) \right) + N \log(1-\beta) \right) \Longrightarrow$$

$$0 = \sum_{n=1}^{N} \frac{r_n}{\beta(1-\beta)} - \frac{N}{1-\beta} \Longrightarrow$$

$$\beta = \frac{1}{N} \sum_{n=1}^{N} r_n. \tag{49}$$

To get the algorithm going, θ has to be initialized (e.g. to parameters that are typical for the phenomenon at hand). Then it is iterated until some criterion of "convergence" is met.

More details on how this algorithm is adapted to estimate the parameter vector θ that defines our mixture distribution $\mathcal{M}$, and about initialization and convergence in this case, is given in Section 4.5.2.

4.5.2 Parameter Estimation - $\mathcal{M}$

If we consider $X_U = X|X < U$ instead of X as in Section 4.5.1, the density $f_X(\cdot)$ is multiplied by a factor $1/\mathbb{P}(X < U)$, and the term term

$$-\log \mathbb{P}(X < U)$$

is to be added to the log-likelihood presented in Equation (43). However, the assumed "knowledge" about which sub-population each observation comes from, i.e. the information encoded in the z_n's, should effect the likelihood, and instead of the expression in Equation (44) we get

$$l(\theta|\boldsymbol{x}, \boldsymbol{z}) =$$

$$\sum_{n=1}^{N} \left(z_n \left[\log \beta + \log f_Y(x_n|\theta) - \log \mathbb{P}(Y < U|\theta)\right] \mathbb{1}(x_n < U)\right)$$

$$+ \sum_{n=1}^{N} \left((1 - z_n) \left[\log(1 - \beta) + \log f_W(x_n|\theta) - \log \mathbb{P}(0 < W < U|\theta)\right] \mathbb{1}(x_n < U)\right), \tag{50}$$

where the indicator functions ensure that each x_i satisfies

$$0 < x_i < U,$$

and where W is a (non-truncated) $\mathcal{N}(\mu_2, \sigma_2)$ distributed random variable.

With f_Y being the density of a *lognormal* random variable, parametrized by μ_1 and σ_1, we have

$$\log f_Y(x|\theta) = -\log(2\pi)/2 - \log \sigma_1 - \log x - \frac{(\log x - \mu_1)^2}{2\sigma_1^2}. \tag{51}$$

and with f_W being the density of a normal random variable, parametrized by μ_2 and σ_2, we have

$$\log f_W(x|\theta) = -\log(2\pi)/2 - \log\sigma_2 - \frac{(x-\mu_2)^2}{2\sigma_2^2}. \tag{52}$$

The updating equation for r_n, presented in Equation (45), becomes

$$r_n = \frac{\beta^{(i)}\dfrac{f_Y(x_n|\theta^{(i)})}{\mathbb{P}(Y<U|\theta^{(i)})}}{\beta^{(i)}\dfrac{f_Y(x_n|\theta^{(i)})}{\mathbb{P}(Y<U|\theta^{(i)})} + (1-\beta^{(i)})\dfrac{f_W(x_n|\theta^{(i)})}{\mathbb{P}(0<W<U|\theta^{(i)})}}$$

$$= \frac{\beta^{(i)} f_Y(x_n|\theta^{(i)})}{\beta^{(i)} f_Y(x_n|\theta^{(i)}) + (1-\beta^{(i)}) f_W(x_n|\theta^{(i)})\dfrac{\mathbb{P}(Y<U|\theta^{(i)})}{\mathbb{P}(0<W<U|\theta^{(i)})}}, \tag{53}$$

where the CDF's also depends on the most recent value of the parameter vector θ (with $\theta^{(i)}$ as in (53), we would get the $(i+1)$th r_n estimate). The mixing parameter β is not effected by the truncation and it is updated according to Equation (49). Given the r_n's from Equation (53), we maximize

$$\sum_{n=1}^{N} \left(r_n \left[\log f_Y(x_n|\mu_1,\sigma_1) - \log \mathbb{P}(Y<U|\theta)\right] \mathbb{1}(x_n<U)\right) =$$

$$-\sum_{n=1}^{N}\left(r_n\left[\log(2\pi)/2 + \log\sigma_1 + \log x_n + \frac{(\log x_n-\mu_1)^2}{2\sigma_1^2}\right]\mathbb{1}(x_n<U)\right)$$
$$-\sum_{n=1}^{N} r_n \log \mathbb{P}(Y<U|\theta)\mathbb{1}(x_n<U) \tag{54}$$

and

$$\sum_{n=1}^{N} \left((1-r_n)\left[\log f_W(x_n|\mu_2,\sigma_2) - \log \mathbb{P}(0<W<U|\theta)\right]\right) \mathbb{1}(0<x_n<U) =$$

$$-\sum_{n=1}^{N}\left((1-r_n)\left[\log(2\pi)/2 + \log\sigma_2 + \frac{(x_n-\mu_2)^2}{2\sigma_2^2}\right]\mathbb{1}(0<x_n<U)\right)$$
$$-\sum_{n=1}^{N}(1-r_n)\log[\Phi(\tfrac{U-\mu_2}{\sigma_2}) - \Phi(-\tfrac{\mu_2}{\sigma_2})]\mathbb{1}(0<x_n<U), \tag{55}$$

w.r.t. μ_1, σ_1 and μ_2, σ_2 respectively.

Pseudocode 2 describes the outlined procedure.

Pseudocode 2: **Estimation of θ in $\mathcal{M}$**

```
1  input: vector with n data points, initial θ0
2  output: θ
3
4  θ <- θ0;
5  r <- empty n-vector; // to hold ‘‘responsibilities’’
6
7  repeat until stopping-criterion is satisfied: // (see Section 4.5.3)
8      update r according to Equation (53);
9      update theta by Equation (49)
10         and to maximize Equations (54) and (55);
11
12     restrict θ to given interval; // (see e.g. Table 5)
13
14     compute log likelihood of θ given data;
15
16 if (β < 0.02) then β<-0.0;
17 if (β > 0.98) β<-1.0;
18
19 return θ;
```

Note that the procedure on line 12 in Pseudocode 2 only effects β, since the remaining parameters are constrained already when when updated on line 10 (see Section 4.5.4).

4.5.3 Initialization and Convergence

The parameter vector θ has to be initialized and we use

$$\theta_0 = (4.8, 0.8, 0.8, 280.0, 60.0).$$

Sometimes, the estimation procedure in Pseudocode 2 is used as an intermediate step of an iterative, “outer”, algorithm. In these cases, the initial values θ_0 in line 1 of Pseudocode 2, are given by the most recently updated θ from the outer algorithm. For a more detailed description of this, see Section 4.6.

The stopping-criterion in line 7 of Pseudocode 2 is based on quantifying changes of in the log likelihood of θ. More specifically, when a sequence of three consecutive θ-values result in a log likelihood that does not increase with more than some constant (this "tolerance" defaults to 0.05) we stop iterating. There is also a lower bound on the number of iterations (this defaults to eight), to make sure, or at least to it make more likely, that the θ-space is properly examined, and an upper bound on the number of iterations, to terminate the algorithm if no convergence is achieved (this defaults to 100).

4.5.4 Heuristics

Numerical methods are required to carry through the steps of the algorithm that maximize the likelihood, i.e. the step that maximize the expressions in Equations (54) and (55). This is due to the CDF in the normalizing factor.[9] Instead of finding the zeros of the differentiated expressions by numerical methods, a "robust" heuristic is used. For each parameter of θ except for β, say θ_j, the likelihood, or rather the logarithm of the likelihood, is evaluated for a number of values as this parameter varies, while keeping the remaining parameters in θ fixed. The value of θ_j that maximize the likelihood is then returned as the estimate. Pseudocode 3 exemplifies this procedure for μ_1, but it generalizes to σ_1, μ_2 and σ_2 in the obvious way. The quantity $ll(\mu)$ in Pseudocode 3 is proportional to the expression in Equation (54), but only the terms that depend on $\mu = \mu_1$ are kept. The parameter σ_1 is taken from θ_0, and the responsibilities r are given as inputs to the algorithm.

[9] A factor $\Phi(\cdot)$ is present also after differentiation of $\log \Phi(\cdot)$.

Pseudocode 3: μ_1 **heuristic**

```
1  input: data, vector with responsibilities  r, initial θ0, U
2  output: μ
3
4  μ <- μ1 from θ0;
5  LL = ll(μ);
6
7  find the direction d along the μ1-axes in θ-space
8      in which the likelihood increases;
9  while μ is in accepted range [LB, UB]:
10     depending on d, increase or decrease μ with INCR;
11     if ll(μ) ≥ LL:
12         LL = ll(μ);
13     else:
14         break;
15
16 return μ;
```

LB and UB in Pseudocode 3 are lower and upper bounds on μ_1 to make sure we stay in a range of values that are "reasonable" for the application at hand. In our case, these bounds default to the values presented in Table 5.

parameter	LB	UB
μ_1	3.0	7.0
σ_1	0.3	1.9
μ_2	90	600
σ_2	20	220

Table 5: Bounds for parameters in θ.

The size of the steps, i.e. `INCR` in line 10 of Pseudocode 3, defaults to 0.01 for μ_1 and σ_1 that control the *lognormal* part of $\mathcal{M}$, and 0.1 for μ_2 and σ_2 that control the *truncated normal* part of $\mathcal{M}$.

4.6 Estimation with Partially Observed Data

As mentioned in Section 1.1, a method to estimate the parameters that defines the distribution of the processing times, even though just a subset of the errands in data are processed, is desired. To be clear, the mixture distribution $\mathcal{M}$ is truncated from above at a positive value U, but now we face the problem of estimating θ, that defines $\mathcal{M}$, based on data in which a subset of the processing times, those that are longer than some constant $A \leq U$, cannot be observed. A general description follows.

4.6.1 General Description

Consider a sample $\boldsymbol{X} = (X_1, \ldots, X_N)$ from a distribution with density $f(\cdot|\theta)$ parametrized by θ. For convenience, and without loss of generality, assume that $\boldsymbol{X}$ is ordered in ascending order. Suppose that we cannot observe X_i if $X_i \geq A$ for a known constant $A \in \mathbb{R}$, i.e. we observe just a subset

$$\tilde{\boldsymbol{X}} = (X_1, \ldots, X_n) \subset \boldsymbol{X},$$

with $X_i < A$ for $i = 1, \ldots, n$ and $n \leq N$. The values of the $m = N - n$ data points in $\boldsymbol{X} \backslash \tilde{\boldsymbol{X}}$ are unknown, but they are all greater than, or equal to A.[10]

The idea behind the following procedure is to repeatedly generate m points, assuming that we know θ. We generate data from our "known" distribution, conditioned on the observations being larger than, or equal to A, and iteratively update θ until some criterion of convergence is satisfied.

Let θ_0 be the initial parameters, reflecting some "prior knowledge" about $\boldsymbol{X}$. Then Pseudocode 4 explains this iterative procedure for estimation of θ.

[10] In $\mathcal{M}$, we must have $A \leq U$ for this to be meaningful.

Pseudocode 4: **Parameter estimation with partially observed data**

```
1  input: initial parameters θ_0, data X̃ = (X_1, ..., X_n), size of full sample N
2  output: θ
3
4  θ = θ_0;
5  repeat until convergence: // (see Section 4.6.4)
6      generate m variables X_h = X_{n+1}, ..., X_N from f(·|X≥A, θ);
7      estimate θ̂ based on X̃ ∪ X_h;
8      θ <- w · θ + (1 − w) · θ̂; // (see Section 4.6.3)
9
10 return θ;
```

The main ideas presented in this section, as summarized in Pseudocode 4, for the special case where $w = 0$, was proposed by Celeux and Diebolt [3] under the name Stochastic EM (SEM). A similar approach, the Monte Carlo EM algorithm (MCEM), utilize and extend the same ideas, see [12] for further details.

4.6.2 Application to $\mathcal{M}$

In our setting, the density in Pseudocode 4 is $f_{X|X<U}(\cdot|\theta)$, and clearly we must have $A \leq U$. The estimation on line 7 in Pseudocode 4 invokes the ideas presented in Section 4.5.2, initialized by the most recent estimate of θ. In particular, this implies that the algorithm from Section 4.5.2, for computing $\hat{\theta}$ on line 7, is nested within another algorithm that reduces to the SEM-algorithm when $w = 0$ on line 8. See Section 4.6.3 for a discussion about the weighting parameter w.

4.6.3 Heuristics

Occasionally, running the algorithm indicates overly vivid movements in θ, due to the randomly generated data points in $\tilde{\boldsymbol{X}}$ (line 6 of Pseudocode 4). Therefore, a strategy to slow down the procedure is used. At each step, the θ is updated as a weighted average of the new, and the last estimate of θ. Default $w \in [0, 1]$ in line 8 of Pseudocode 4 is set to $2/3$, in which case a slightly higher weight is put on the previous parameter estimate, than on the one based on the latest augmented data set. This choice of w is based on empirical studies, and the resulting decrease of the variation in consecutive

parameter estimates seems to produce good results for a variety of simulated test data sets. No results from these studies will be presented in this text.

4.6.4 Initialization and Convergence

The initial parameter vector θ is default set to

$$\theta_0 = (4.8, 0.8, 0.8, 280.0, 60.0).$$

The criterion of convergence in line 5 of Pseudocode 4 is based on quantifying changes in the explicit parameter vector θ, rather than in the likelihood.[11] More specifically, let

$$C^{(i)} = \max \left\{ j = 1, \ldots, 5; \quad \left| \frac{\theta_j^{(i)} - \theta_j^{(i-1)}}{\theta_j^{(i)}} \right| \right\}, \tag{56}$$

be the maximal, relative absolute difference of two consecutive θ vectors (see the remark in Section 4.6.5 for a short treatment of the obstacle of dividing by small numbers in this expression). To deduce convergence, we require C to be smaller than some tolerance level (this defaults to 0.1) in three consecutive iterations. See Section 4.6.6 for a discussion about this.

4.6.5 Remark

With $j = 3$ in Equation (56), i.e. when concerned with the β component of θ, the denominator in (56) might be close to zero. If this is the case, a small value (default 0.001) is added in the denominator. This is done to avoid undesired, uncontrolled results due to limitations in the floating-point precision (or even division by zero).

4.6.6 Discussion

One might deduce convergence in line 5 of Pseudocode 4 based on the likelihood of the parameters, conditioned on the full sample $\boldsymbol{X_h} \cup \tilde{\boldsymbol{X}}$, instead of based on the ideas presented in Section 4.6.4. However, the likelihood

[11] A criterion of convergence based on quantifying changes in the likelihood also deduces convergence based on the parameters in θ, but through their functional relationship to the likelihood.

crucially depends on the simulated data in $\boldsymbol{X_h}$, and might not even increase through the iterations. Likewise, evaluating the likelihood of $X|X < A$, based exclusively on the "observed" part of data, could be misleading if A is small.[12] Therefore, a convergence criterion based on quantifying changes in the parameter vector θ, is a reasonable choice.

[12]This argument applies, of course, to the strategy of deducing convergence based on C in Equation (56) as well.

5 Computer Software – Implementation

A part of the work has been focused on development of a stable and "fast" computer software to handle the tasks that arise. The programming language `C++` has been used for the algorithms and for computations, while the graphics in this text are produced in `R`. A list of some of the implemented "functions" are available in Appendix A. Functions, for which corresponding, reliable open source software is available, have been bench-marked against those other implementations, to ensure efficiency and correctness.[13] No run time comparisons or similar results from these investigations are presented in this text though.

5.1 Discussion

Implementing algorithms in a "low level" language, such as `C++`, builds up understanding and intuition, not only for their own strengths and weaknesses, but also for the underlying problem that they attempt to solve. Therefore, the construction of a well performing computer software has been important also for developing the methods.

[13] Examples of such are functions available in the `R`-base, such as densities and distribution functions.

6 Experimental Results

In this section, experimental results from applying the methods that were presented in Section 4 to simulated data and $\mathcal{D}$, are presented.

6.1 Parameter Estimation in a Mixture Model.

In this section the iterative algorithm for parameter estimation that was described in Section 4.5.2 is examined. We generate a sample from the mixture distribution $\mathcal{M}$ (see Section 2.2), parametrized by

$$\theta_0 = (4.3, 0.8, 0.75, 365, 70),$$

and truncate it from above at $U = 600$. The size of the resulting sample is $n = 300$. Estimation of θ_0 gives

$$\hat{\theta} = (4.28, 0.82, 0.75, 350, 78.8),$$

after 9 iterations. Figure 8 presents a histogram from the sample, together with the densities of X and $X|X < U$, for θ_0 and $\hat{\theta}$. The Kolmogorov-Smirnof statistics, with respect to $X|X < U$, parametrized by $\hat{\theta}$ and θ_0 respectively, are

$$\sqrt{n}D_{n,\hat{\theta}} = 0.61,$$

and

$$\sqrt{n}D_{n,\theta_0} = 0.54.$$ [14]

As expected, a comparison with the critical values in Table 4 provides no evidence against the null hypothesis that the sample is realized from $X|X < U$. As addressed in Section 4.4.3, the tabulated values are invalid after estimation of θ, but the low values on D_n still indicate a good fit. Figure 9 presents the empirical distribution function and the theoretical CDF's, based on θ_0 and $\hat{\theta}$.

6.2 Parameter Estimation with Partially Observed Data

The following three experiments are conducted to test the methodology that was presented in Section 4.6. Data is generated from $\mathcal{M}$, parametrized by

[14] The statistic that is based on θ_0, i.e. D_{n,θ_0}, is included as a reference for $D_{n,\hat{\theta}}$.

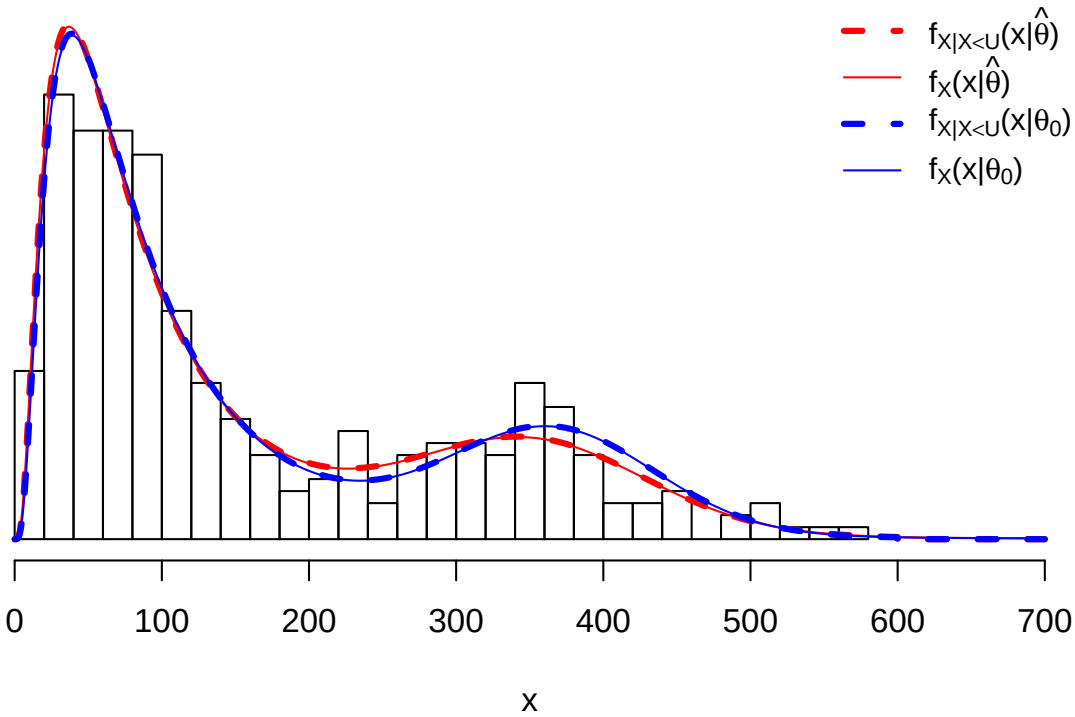

Figure 8: Histogram and densities of X and $X|X < U$, for $U = 600$, $\theta_0 = (4.3, 0.8, 0.75, 365, 70)$, and estimated parameters $\hat{\theta} = (4.28, 0.82, 0.75, 350, 78.8)$. The histogram is based on a (truncated) sample of size 300, drawn from $\mathcal{M}$, parametrized by θ_0.

θ_0,[15] and truncated from above at $U = 730$, and the parameter vector $\hat{\theta}$ is estimated for different "boundaries" between the "observed" and "hidden" parts. More specifically, for different values of A (see Section 4.6), we "observe" the subset $\tilde{\boldsymbol{X}} \subset \boldsymbol{X}$ of the generated sample $\boldsymbol{X}$, that satisfies

$$\tilde{\boldsymbol{X}} = \{x_i \in \boldsymbol{X};\ x_i < A\},$$

and consider $\boldsymbol{X} \backslash \tilde{\boldsymbol{X}}$, the simulated data points that are at least as large as A, as "hidden". We then estimate the data-generating parameters θ_0 by the methods that were described in Section 4.6. The estimate obtained after observing exclusively data points that are smaller than a boundary A will be denoted $\hat{\theta}_A$. The Kolmogorov-Smirnof statistics that are presented in the three following experiments are computed based on the full sample $\boldsymbol{X}$, regardless of the truncation point A.

[15]The value of θ_0 will be specified in each experiment separately.

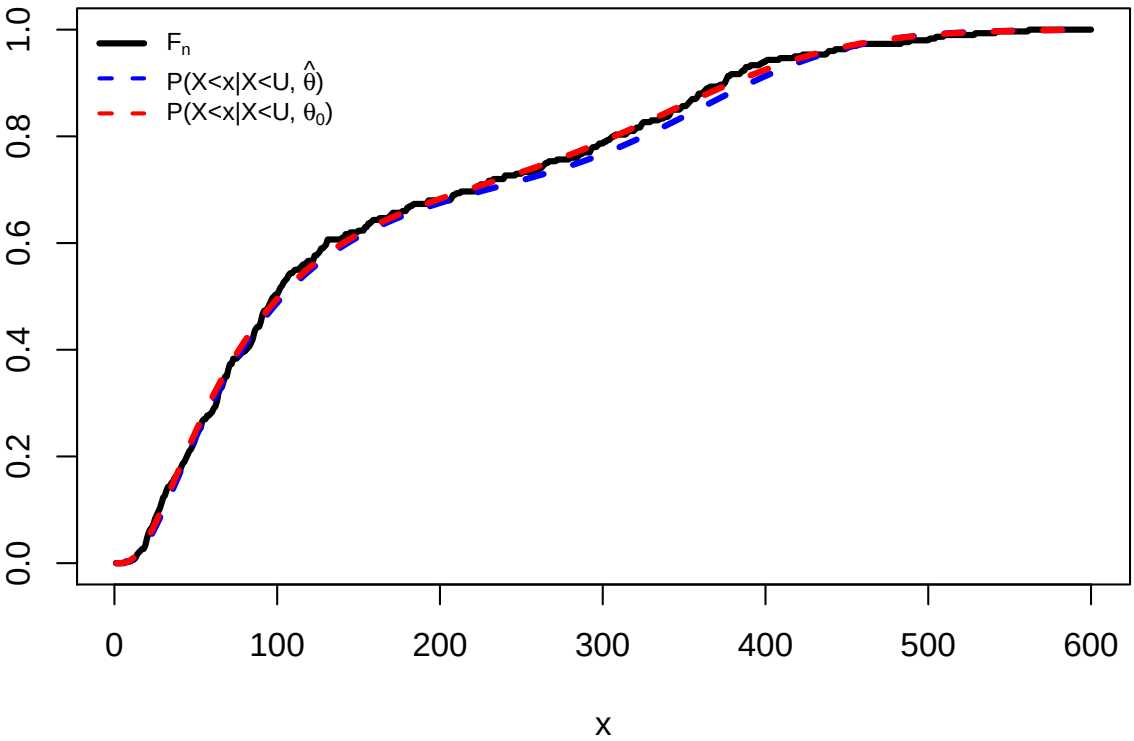

Figure 9: Black piece-wise curve: Empirical distribution function based on a sample of size 300, drawn from $\mathcal{M}$, parametrized by $\theta_0 = (4.3, 0.8, 0.75, 365, 70)$, and truncated from above at $U = 600$.
Red and blue curves: CDF's of $X|X < U$, parametrized by θ_0, and estimated parameters $\hat{\theta} = (4.28, 0.82, 0.75, 350, 78.8)$ respectively.

6.2.1 Experiment 1

We generate data points from $\mathcal{M}$, parametrized by

$$\theta_0 = (4.2, 0.55, 0.75, 260, 70),$$

and truncate them from above at $U = 730$. $N = 750$ is the size of the truncated sample. Figures 10 and 11, and Table 6 present the results. In Figure 10, the dashed, vertical, lines indicate three different values of A, beyond which data is "hidden". We use

$$A \in \{90, 140, 220\}.$$

The parameter estimates that result from each of these A-values give rise to densities that are plotted together with the histogram of the simulated data,

and the density of the mixture distribution $\mathcal{M}$, parametrized by θ_0. Figure 11 presents the corresponding distribution functions. The Kolmogorov-Smirnof statistics, D_n, in Table 6, is computed based on all N data points in the interval $(0, U)$, regardless of the truncation point A.

	$(\mu_1, \sigma_1, \beta, \mu_2, \sigma_2)$	D_n
θ_0	$(4.2, 0.55, 0.75, 260, 70)$	0.032
$\hat{\theta}_{90}$	$(4.34, 0.63, 0.91, 165, 52.93)$	3.899
$\hat{\theta}_{140}$	$(4.37, 0.67, 0.86, 243.27, 115.07)$	1.628
$\hat{\theta}_{220}$	$(4.15, 0.55, 0.71, 230.07, 71.67)$	1.029

Table 6: Estimates of the data-generating parameter vector θ_0, for different values of $A \in \{90, 140, 220\}$. See Section 6.2.1 for more information.

6.2.2 Experiment 2

We perform a similar experiment as that in Section 6.2.1. A sample is generated from the mixture distribution $\mathcal{M}$, parametrized by

$$\theta_0 = (4.7, 0.8, 0.75, 365, 100),$$

and truncated at $U = 730$. $N = 500$ is the size of the truncated sample. Figures 12 and 13, and Table 7 present the results.

	$(\mu_1, \sigma_1, \beta, \mu_2, \sigma_2)$	D_n
θ_0	$(4.7, 0.8, 0.75, 365, 100)$	0.032
$\hat{\theta}_{90}$	$(4.83, 0.75, 0.9, 244.47, 109.07)$	3.088
$\hat{\theta}_{150}$	$(4.95, 0.86, 0.9, 290.47, 98.4)$	1.464
$\hat{\theta}_{240}$	$(4.92, 0.87, 0.84, 329.6, 93.73)$	0.950

Table 7: Estimates of the data-generating parameter vector θ_0, for different values of $A \in \{90, 150, 240\}$. See Section 6.2.2 for more information.

6.2.3 Experiment 3

We perform a similar experiment as that in Section 6.2.1. A sample is generated from the mixture distribution $\mathcal{M}$, parametrized by

$$\theta_0 = (4.9, 1.2, 0.25, 365, 150),$$

and truncated at $U = 730$. $N = 500$ is the size of the truncated sample. Figures 14 and 15, and Table 8 present the results.

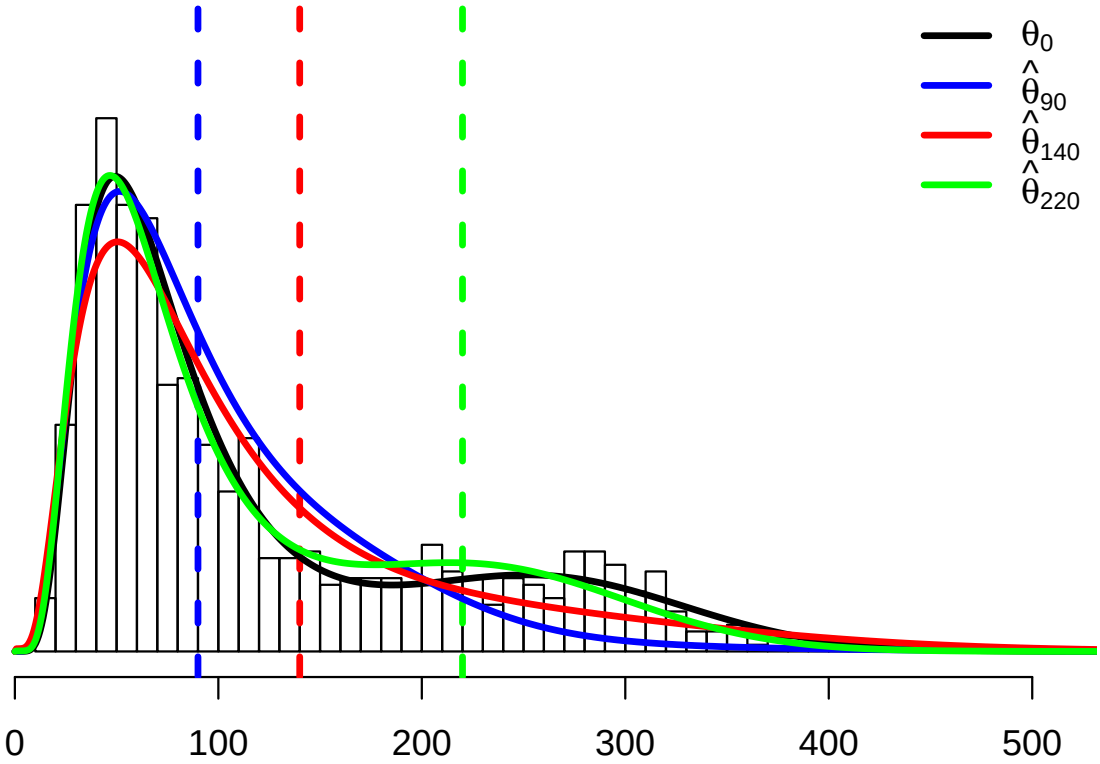

Figure 10: Histogram over data generated from the mixture distribution $\mathcal{M}$, parametrized by $\theta_0 = (4.2, 0.55, 0.75, 260, 70)$, together with densities for $\mathcal{M}$, parametrized by θ_0 and estimates of θ for three values of $A \in \{90, 140, 220\}$. See Section 6.2.1 for more information.

	$(\mu_1, \sigma_1, \beta, \mu_2, \sigma_2)$	D_n
θ_0	$(4.9, 1.2, 0.25, 365, 150)$	0.026
$\hat{\theta}_{100}$	$(6.62, 1.72, 0.51, 257.07, 98.33)$	4.730
$\hat{\theta}_{200}$	$(7, 1.66, 0.77, 314.33, 60.13)$	2.536
$\hat{\theta}_{300}$	$(6.93, 1.66, 0.73, 334.67, 71.8)$	1.953

Table 8: Estimates of the data-generating parameter vector θ_0, for different values of $A \in \{100, 200, 300\}$. See Section 6.2.3 for more information.

6.2.4 Discussion

Intuitively, it might be hard to capture a second component in the mixture model if it's mode exceeds A (the further away from the border between

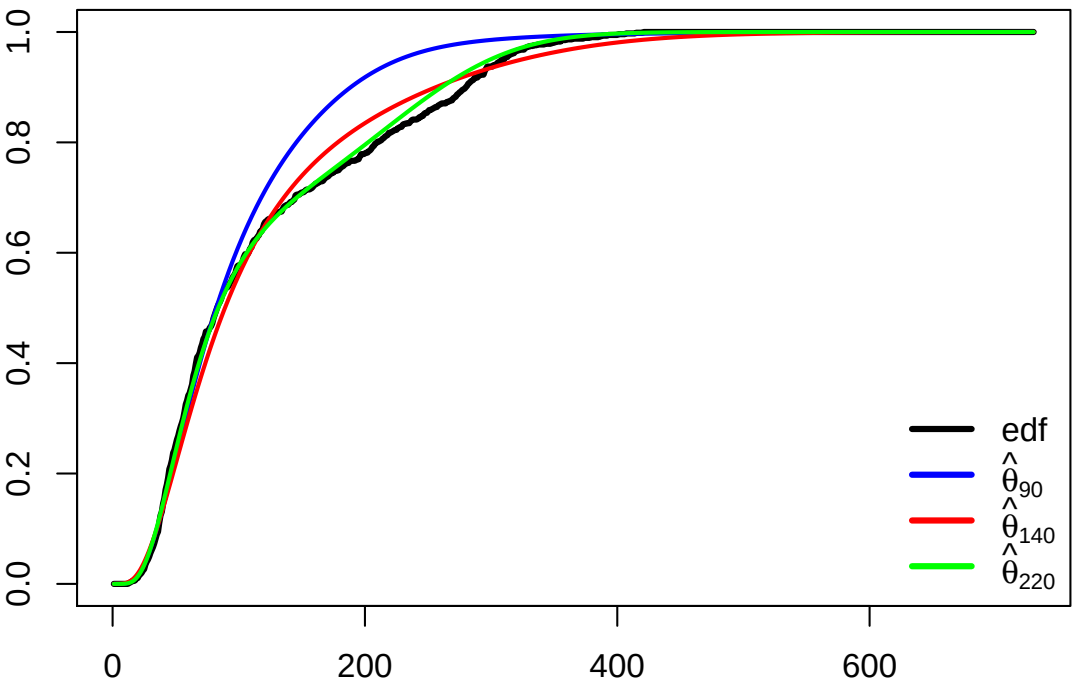

Figure 11: Empirical distribution function (edf) for data generated from the mixture distribution $\mathcal{M}$, parametrized by $\theta_0 = (4.2, 0.55, 0.75, 260, 70)$, together with CDF's for $\mathcal{M}$, parametrized by estimates of θ_0, for three values of $A \in \{90, 140, 220\}$. See Section 6.2.1 for more information.

the "observed" and the "hidden" data points that this mode is located, the harder this is). The experiments in Sections 6.2.1 through 6.2.3 confirm this, and obviously, the fit improves with increasing A. The Kolmogorov-Smirnof statistics in Tables 6 through 8 do not justify using the estimation strategy for small A-values. However, the experiments indicate that the method could be useful if the exact fit is not crucial, as is the case when we "only" desire estimates for the key values presented in Section 1.3.

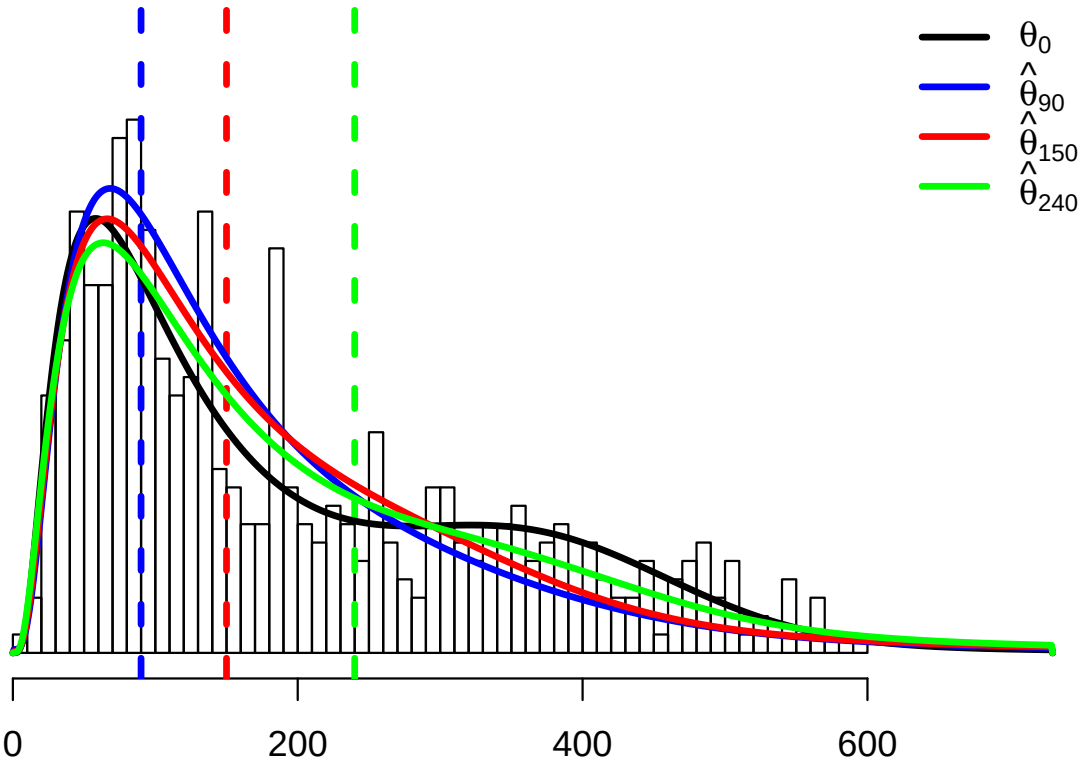

Figure 12: Histogram over data generated from the mixture distribution $\mathcal{M}$, parametrized by $\theta_0 = (4.7, 0.8, 0.75, 365, 100)$, together with densities for $\mathcal{M}$, parametrized by θ_0 and estimates of θ for three values of $A \in \{90, 150, 240\}$. See Section 6.2.2 for more information.

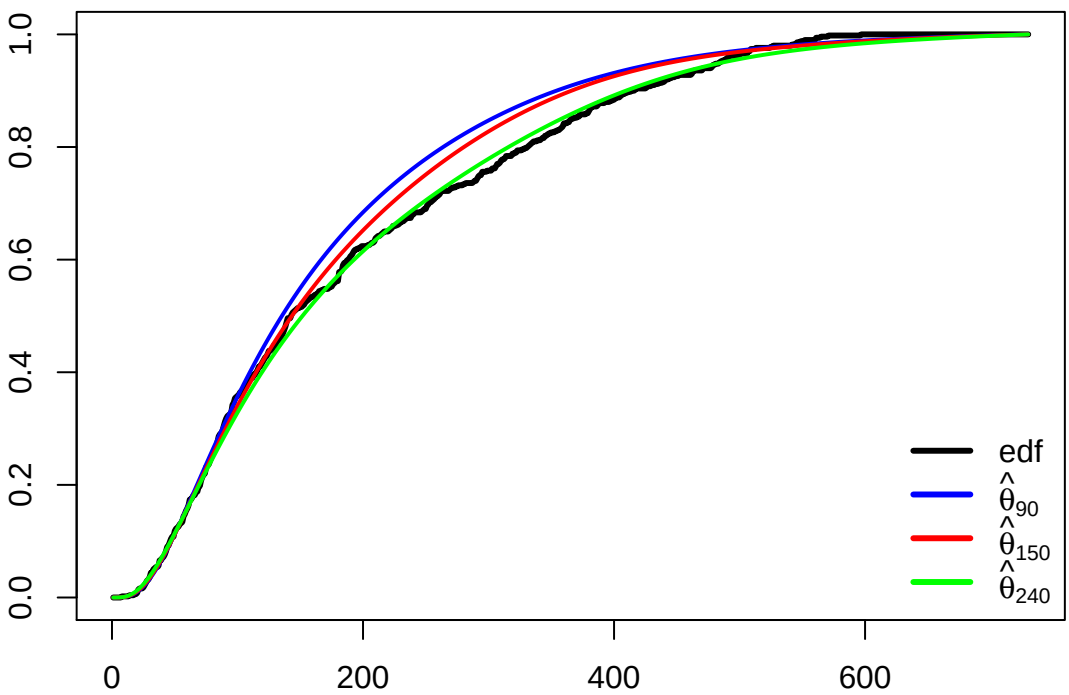

Figure 13: Empirical distribution function (edf) for data generated from the mixture distribution $\mathcal{M}$, parametrized by $\theta_0 = (4.7, 0.8, 0.75, 365, 100)$, together with CDF's for $\mathcal{M}$, parametrized by estimates of θ for three values of $A \in \{90, 150, 240\}$. See Section 6.2.2 for more information.

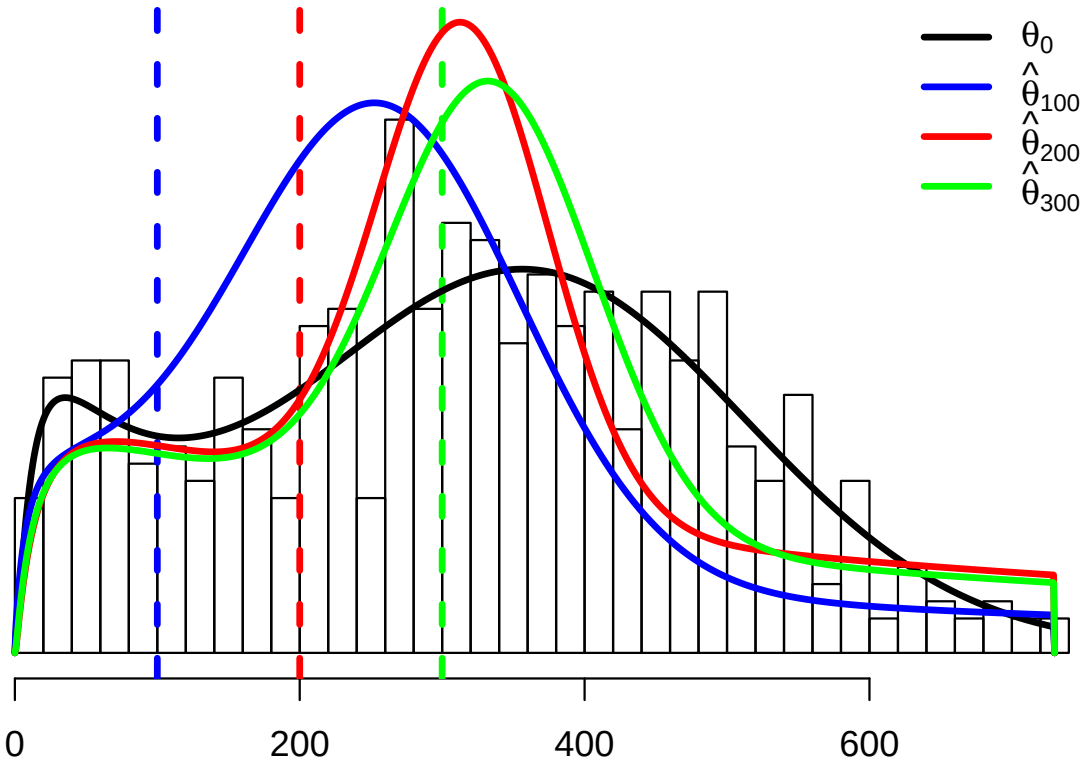

Figure 14: Histogram over data generated from the mixture distribution $\mathcal{M}$, parametrized by $\theta_0 = (4.9, 1.2, 0.25, 365, 150)$, together with densities for $\mathcal{M}$, parametrized by θ_0 and estimates of θ for three values of $A \in \{100, 200, 300\}$. See Section 6.2.3 for more information.

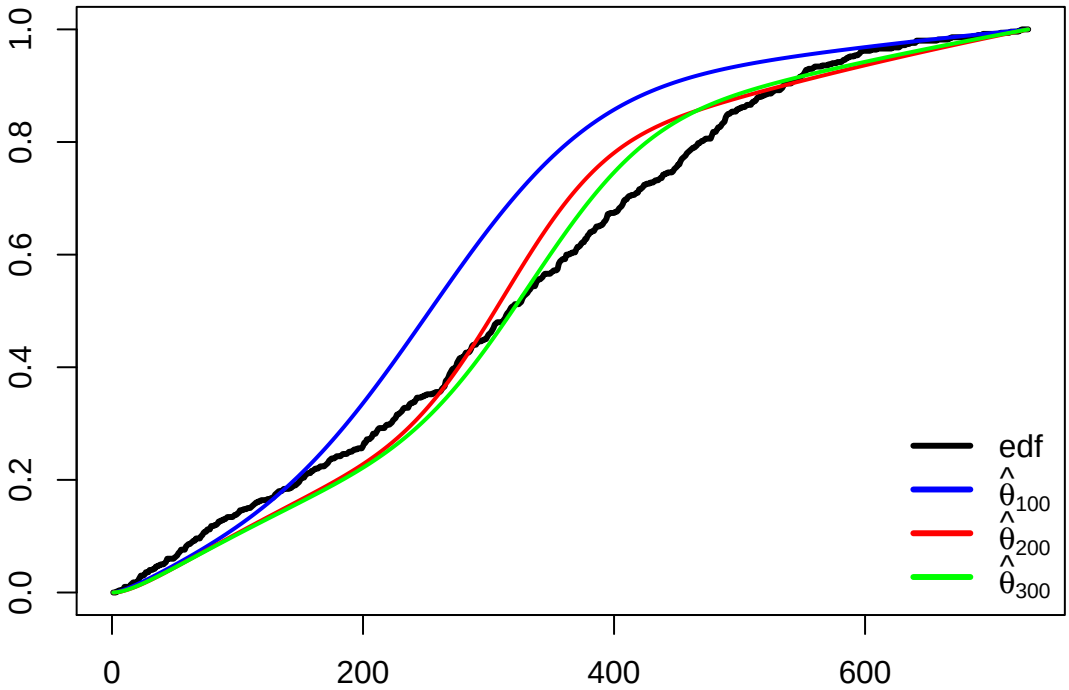

Figure 15: Empirical distribution function (edf) for data generated from the mixture distribution $\mathcal{M}$, parametrized by $\theta_0 = (4.9, 1.2, 0.25, 365, 150)$, together with CDF's for $\mathcal{M}$, parametrized by estimates of θ for three values of $A \in \{100, 200, 300\}$. See Section 6.2.3 for more information.

6.3 Real World Data - $\mathcal{D}$

In this section we will use the methods that were presented in Section 4 to characterize the distribution of the processing times in four subsets of $\mathcal{D}$, namely cohorts 3, 4, 5 and 6 that was presented in Section 3. We need to consider cohorts in which also long errands are processed by the time of the data collection. Therefore we use data with processes initiated during the first four months of 2015, and processing times that are no longer than

$$U = 1.8 \cdot 365 = 657 \text{ days.}^{16}$$

This data is partitioned into four cohorts based on the month of opening. Furthermore, the data that we consider is restricted so that `COUNTRY` is kept on one of its levels (see Section 3).[17] Again, privacy policies prevent us from giving more details about data from these cohorts.

We assume that the data is realized from the mixture distribution $\mathcal{M}$, parametrized by the unknown parameter vector θ_0 (see Section 2.2). The key values, presented in Section 1.3, obtained from the estimate $\hat{\theta}_U$ of θ_0, $\hat{\theta}_U$ itself, and the corresponding statistic for the Kolmogorov-Smirnof test, D_n, for the fours cohorts from 2015-01, 2015-02, 2015-03, and 2015-04 are presented in Tables 10 through 13. Note that $\hat{\theta}_U$ is the parameter estimate based on the complete, i.e. not just partially observed, cohort. The sample sizes N, for these four cohorts are 1344, 1189, 1326 and 1141 respectively (they are also summarized in Table 9).

When the Kolmogorov-Smirnof statistic, D_n is computed, it is based on the difference between the empirical distribution function of the full sample, and the CDF of $\mathcal{M}$, truncated at U, and parametrized by the estimated parameters.

We also assess the estimation-procedure that was developed for partially observed data, using $A \in \{225, 125\}$, by obtaining estimates $\hat{\theta}_A$ (see Section 4.6). The "observed" sample sizes, N_A, i.e. the number of data points that are smaller that A, are given in Table 9.

Figures 16 through 23 present histograms of the data, together with densities of the mixture distribution $\mathcal{M}$, based on estimated parameters, and the empirical distribution functions of data together with CDF's of $\mathcal{M}$, based on estimated parameters. The vertical lines divide data into the "observed" and the "hidden" parts.

[16] By the time of the data collection, this was the most recent data that was available and quality assured.

[17] This level on `COUNTRY` is one of the most frequent, during this period.

	2015-01	2015-02	2015-03	2015-04
N	1344	1189	1326	1141
N_{225}	510	529	566	521
N_{125}	329	379	364	312

Table 9: Sample sizes of the complete cohorts (cohorts 3 through 6), and the "observed" parts for $A \in \{225, 125\}$. See Section 6.3 for more information.

	data	$\hat{\theta}_U$	$\hat{\theta}_{225}$	$\hat{\theta}_{125}$
mean	276.25	285.15	265.6	215.19
sd	166	169.93	147.3	134.76
$Q_{0.5}$	289	300.08	279.37	217.4
$Q_{0.75}$	404	410.04	366.88	304.41
$Q_{0.9}$	493	498.19	445.56	381.95
$Q_{0.95}$	539	547.47	503.59	431.12
μ_1	NA	5.08	5.79	4.66
σ_1	NA	1.8	1.77	1.77
β	NA	0.4	0.57	0.27
μ_2	NA	355	321.2	239.53
σ_2	NA	126	86.4	110.73
$\sqrt{N}D_N$	NA	1.87	4.23	8.42

Table 10: Key values (see Section 1.3) based data from 2015-01 (cohort 3 as specified in Section 3), and the mixture model $\mathcal{M}$, parametrized by estimates of θ. See Section 6.3 for more information.

6.3.1 Discussion

The hypothesized model $\mathcal{M}$ appears to fit the examined cohorts reasonably well. The experiments also indicate that the estimation procedures performance depends on A. A threshold of $A = 225$ results in a decent fit, and key values that are reasonably close to their observed counterparts. A threshold of $A = 125$, on the other hand, appears to be far too small to obtain a good fit in terms of the Kolmogorov-Smirnof statistic, and in this case, the key values are strongly underestimated.

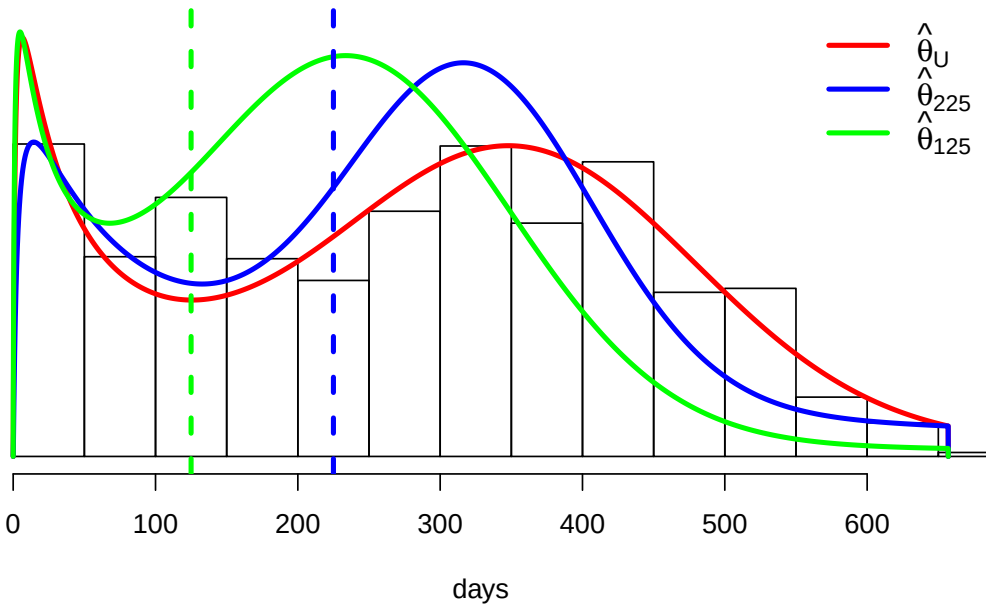

Figure 16: Data from 2015-01 (cohort 3 as specified in Section 3) together with the density curves for the mixture distribution $\mathcal{M}$, parametrized by estimates of the parameter vector θ. The vertical lines, representing A from Section 4.6, partition the data into an "observed" and a "hidden" part. See Section 6.3 for more information.

	data	$\hat{\theta}_U$	$\hat{\theta}_{225}$	$\hat{\theta}_{125}$
mean	246.41	236.09	224.26	187.29
sd	171.6	205.64	173.13	149.46
$Q_{0.5}$	245	244.63	230.95	191.79
$Q_{0.75}$	387	376.66	347.67	289.41
$Q_{0.9}$	468	483.27	442.62	372.86
$Q_{0.95}$	524.6	540.4	496.72	422.29
μ_1	NA	3.08	3.26	3.05
σ_1	NA	1.8	1.77	1.77
β	NA	0.24	0.27	0.22
μ_2	NA	294.2	282.2	220.6
σ_2	NA	172.4	142.07	127.07
$\sqrt{N}D_N$	NA	1.07	2.72	6.72

Table 11: Key values (see Section 1.3) based on data from 2015-02 (cohort 4 as specified in Section 3), and the mixture distribution $\mathcal{M}$, parametrized by estimates of θ. See Section 6.3 for more information.

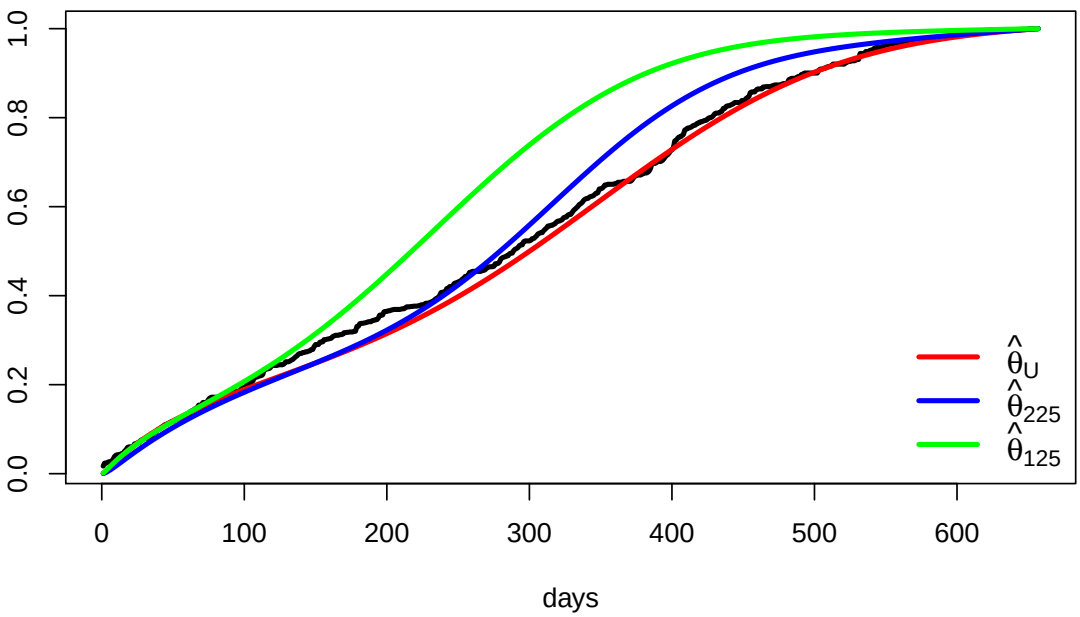

Figure 17: Empirical distribution function based on data from 2015-01 (cohort 3 as specified in Section 3), together with CDF's for the mixture distribution $\mathcal{M}$, paramterized by estimates of θ's. See Section 6.3 for more information.

	data	$\hat{\theta}_U$	$\hat{\theta}_{225}$	$\hat{\theta}_{125}$
mean	259.11	254.09	231.32	197.3
sd	165.07	187.53	167.74	151.7
$Q_{0.5}$	257	264.22	239.07	203.47
$Q_{0.75}$	399	383.72	348.22	298.4
$Q_{0.9}$	465.5	481.69	439.56	381.39
$Q_{0.95}$	534	535.6	492.31	430.73
μ_1	NA	3.55	3.43	3.05
σ_1	NA	1.8	1.77	1.77
β	NA	0.23	0.23	0.16
μ_2	NA	306.6	276.87	220.8
σ_2	NA	155.2	139.33	130.4
$\sqrt{N}D_N$	NA	1.91	4.34	7.92

Table 12: Key values (see Section 1.3) based on data from 2015-03 (cohort 5 as specified in Section 3), and the mixture distribution $\mathcal{M}$, parametrized by estimates of θ. See Section 6.3 for more information.

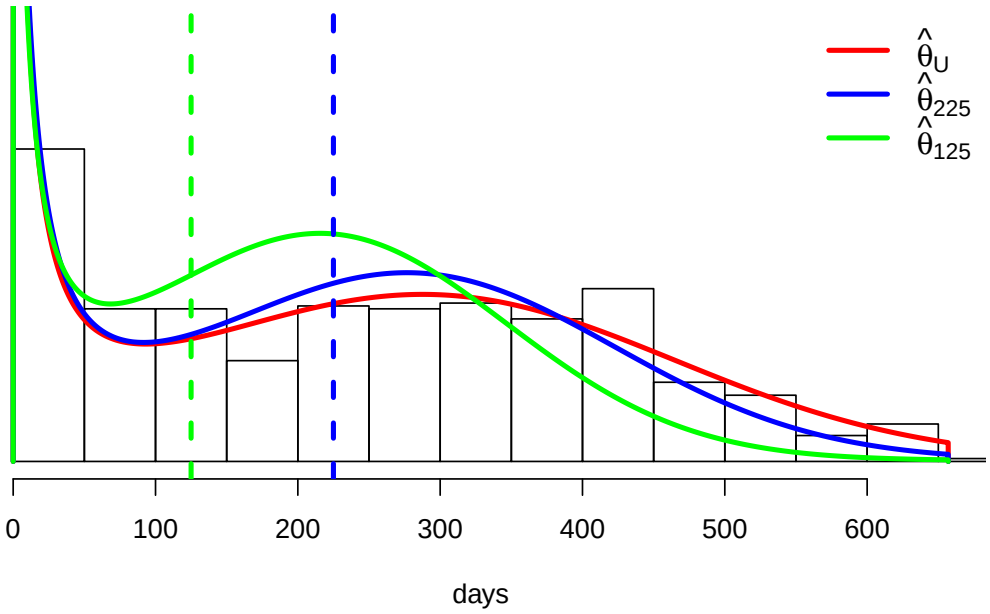

Figure 18: Data from 2015-02 (cohort 4 as specified in Section 3) together with the density curves for the mixture distribution $\mathcal{M}$, parametrized by estimates of the parameter vector θ. The vertical lines, representing A from Section 4.6, partition the data into an "observed" and a "hidden" part. See Section 6.3 for more information.

	data	$\hat{\theta}_U$	$\hat{\theta}_{225}$	$\hat{\theta}_{125}$
mean	251.01	232.03	219.79	197.23
sd	161.82	213.49	176.14	156.28
$Q_{0.5}$	240	245.2	225.87	203.9
$Q_{0.75}$	385	369.2	338.9	300.12
$Q_{0.9}$	456	475.51	435.31	385.06
$Q_{0.95}$	529	533.88	491.04	435.62
μ_1	NA	2.99	3.43	3.05
σ_1	NA	1.8	1.77	1.77
β	NA	0.13	0.21	0.14
μ_2	NA	258.2	256.27	216.53
σ_2	NA	182.6	150.53	135.27
$\sqrt{N}D_N$	NA	1.31	3.2	5.64

Table 13: Key values (see Section 1.3) based on data from 2015-04 (cohort 6 as specified in Section 3), and the mixture distribution $\mathcal{M}$, parametrized by estimates of θ. See Section 6.3 for more information.

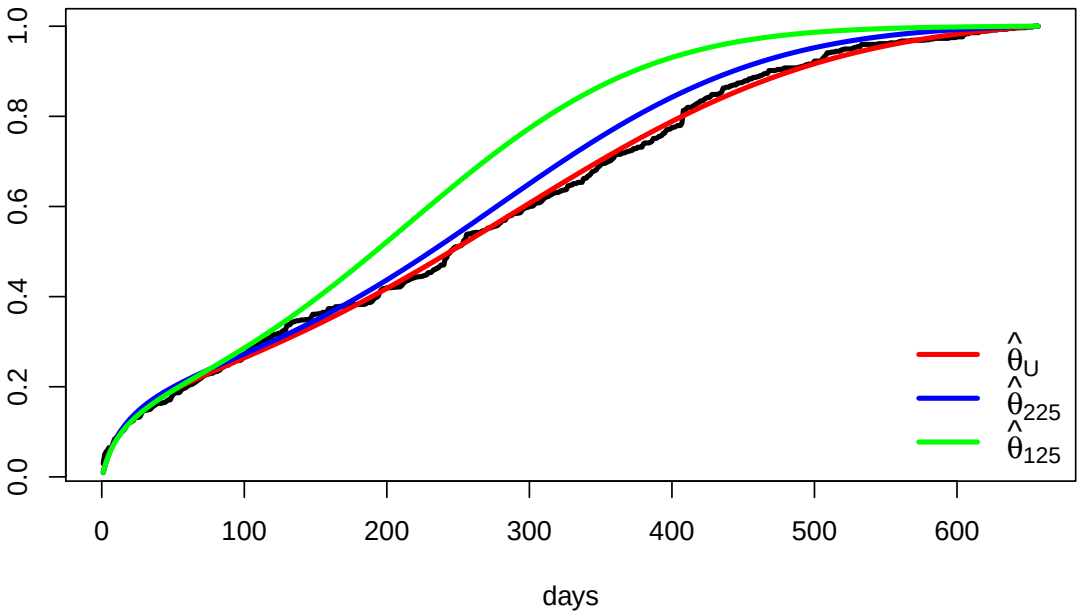

Figure 19: Empirical distribution function based on data from 2015-02 (cohort 4 as specified in Section 3), together with CDF's for the mixture distribution $\mathcal{M}$, paramterized by estimates of θ. See Section 6.3 for more information.

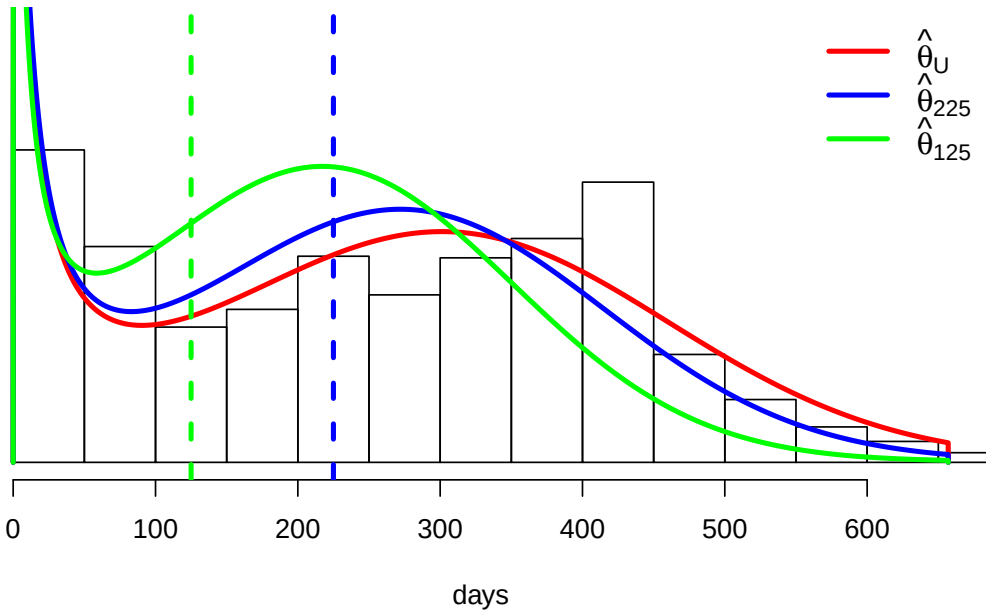

Figure 20: Data from 2015-03 (cohort 5 as specified in Section 3) together with the density curves for the mixture distribution $\mathcal{M}$, parametrized by estimates of the parameter vector θ. The vertical lines, representing A from Section 4.6, partition the data into an "observed" and a "hidden" part. See Section 6.3 for more information.

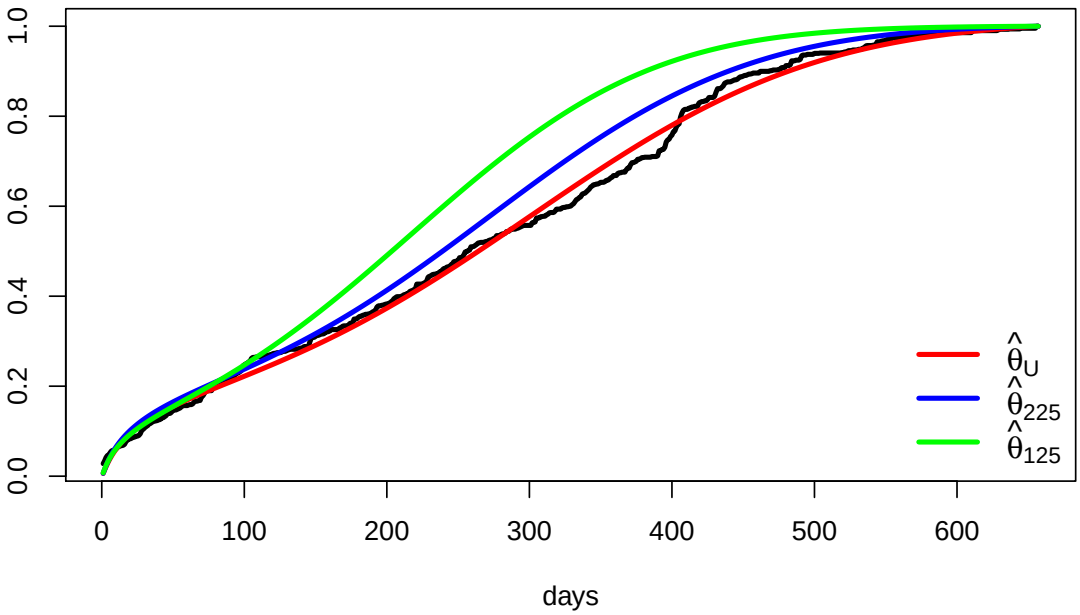

Figure 21: Empirical distribution function based on data from 2015-03 (cohort 5 as specified in Section 3), together with CDF's for the mixture distribution $\mathcal{M}$, paramterized by estimates of θ. See Section 6.3 for more information.

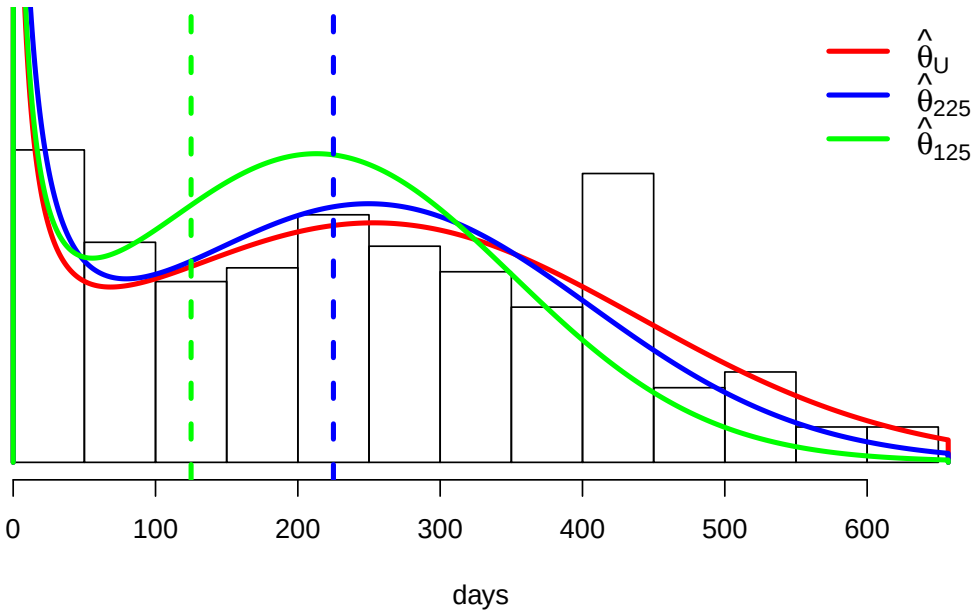

Figure 22: Data from 2015-04 (cohort 6 as specified in Section 3) together with the density curves for the mixture distribution $\mathcal{M}$, parametrized by estimates of the parameter vector θ. The vertical lines, representing A from Section 4.6, partition the data into an "observed" and a "hidden" part. See Section 6.3 for more information.

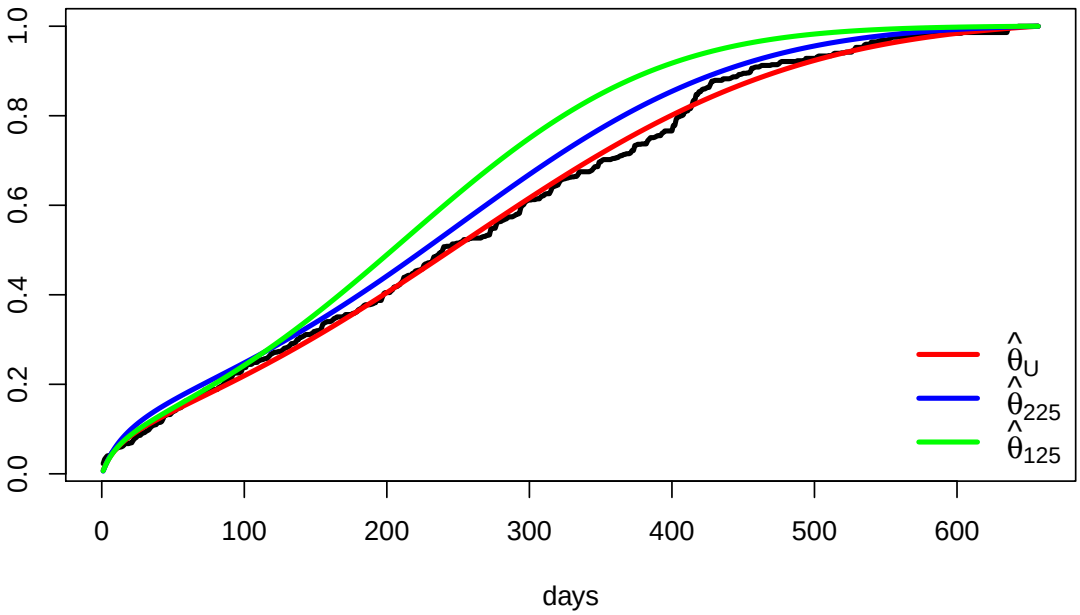

Figure 23: Empirical distribution function based on data from 2015-04 (cohort 6 as specified in Section 3), together with CDF's for the mixture distribution $\mathcal{M}$, paramterized by estimates of θ. See Section 6.3 for more information.

7 Discussion

The time frame of this work is limited and the development of a methodology for characterization of our processing times is not complete. In this section, limitations and drawbacks of the developed methodology are discussed, and an outlook towards potential improvements and further work is given. Finally, a short summary concludes the text.

7.1 Limitations and Further Work

7.1.1 Smaller Cohorts

As the experiments in Section 6.3 reveal, our mixture model $\mathcal{M}$ and estimation procedures (see Sections 2.2 and 4) provide a far from perfect characterization of our processing times. Experiments, which are not presented in this text, indicate that "better separated" cohorts, i.e. cohorts that result from partitioning the data into smaller subsets, defined by more variables and their levels, result in a better fit. According to e.g. *the law of the proportionate effect*, one could even expect that e.g. a *lognormal*, or a *power-law* distribution, on its own would provide an adequate characterization of the processing times in such cohorts, see the discussion in Section 2, [13] and [5], and the brief description of the *power-law* distributions in e.g. [14]. Note that these distributions are not flexible enough to fit the multi modal densities that our current cohorts exhibit (see Section 3.2). In conclusion, when we desire exclusively the key values presented in Section 1.3, our current framework is sufficient, but a "scheme" to partition data into more well separated cohorts could provide further improvements. Put another way, our mixture model $\mathcal{M}$ and the outlined methodology, could be used for characterization of processing times on a "more aggregated" level,[18] and as a complement to this, a more detailed characterization could be given for smaller cohorts.

7.1.2 Theoretical Motivation

Another drawback of the outlined methodology is the lack of a theoretical motivation for the mixture model $\mathcal{M}$. An attempt would be that the lognormally distributed component of our mixture model constitutes a "true, representative kernel" of a given sub-population, and that some disturbance,

[18] Like the cohorts that are considered in Section 6.3.

such as data points that really belong to other sub-populations, result in the observed displacement towards longer processing times. This latter collection of displaced processing times could, by the central limit theorem, be modeled by a normal distribution, and hence constitute the normally distributed component of $\mathcal{M}$.

However, this motivation is very loose, and a rigorous, mathematical description of the mechanisms behind the mixture model $\mathcal{M}$, would justify our methodology and give it a higher degree of credibility.

7.1.3 Estimation

As mentioned, a more sophisticated alternative to the algorithm presented in Section 4.6 is available, namely the Monte Carlo EM or MCEM procedure, see [12]. Experiments with MCEM's performance on our processing times, and adaption of this algorithm to meet our needs, could be interesting. However, lack of fit, i.e. the mixture model $\mathcal{M}$'s inability to fit our data $\mathcal{D}$, is likely not due to an inadequate performance of our estimation procedures. It is rather due to the complex patterns of the processing times in aggregated cohorts of data (see Section 7.1.1 for a discussion about partitioning $\mathcal{D}$ into smaller cohorts).

7.1.4 A Shorter Horizon

A major improvement would be if one was able to provide a good characterization of our processing times, based on smaller "observed parts" of data. More specifically, as we saw in Section 6, when the upper threshold A of the observable processing times as specified in Section 4.6, is small, we fail to fit the distribution of our processing times through parameter estimation. A good fit, based on smaller A would provide a tool to make more accurate characterizations, and better predictions of the processing times, if those vary rapidly over time. This is since, with a smaller A, more recent data would effect the parameter estimates, than if we need to use a bigger A and hence older data for estimation. Therefore further development towards a "shorter horizon" is desired.

7.2 Summary

We have characterized the processing times in our data by a mixture model, more specifically by a weighted mixture of a *lognormal* and a *truncated normal distribution*. Iterative methods, based on maximum likelihood techniques and simulation, have been adapted, developed and implemented to estimate the parameters that define this mixture distribution. Those methods were designed to obtain the parameter estimates based on only "partially observed" data, i.e. based on data in which a subset of the processing times is unknown. A study of the developed methodology on simulated data, and on subsets of our "real world data", has been conducted and presented, where goodness of fit of the proposed mixture model was assessed using e.g. the Kolmogorov-Smirnof statistic.

Appendices

A Computer Software - List of Functions

Here we present the function heads of some of the functions that are implemented as a part of this work. Each function is preceded by a descriptive comment. The results that are presented in Section 6 rely on these implementations. See Section 5 for further information.[19]

```
// density of N(mu, s)
double dNormal(double x, double mu=0.0, double s=1.0)

// P(X<x) where X in N(mu, s)
double pNormal(double x, double mu=0.0, double s=1.0)

// density of X|X>0 where X in N(mu, s)
double dNormal0(double x, double mu, double s)

// P(X<x|X>0) where X in N(mu, s)
double pNormal0(double x, double mu, double s)

// density of X|0<X<U where X in N(mu, s)
double dNormal0U(double x, double mu, double s, double U)

// P(X<x|0<X<U) where X in N(mu, s)
double pNormal0U(double x, double mu, double s, double U)

// density of lognormal(mu, s)
double dLogNormal(double x, double mu, double s)

// P(X<x) where X in lognormal(mu, s)
double pLogNormal(double x, double mu, double s)

// density of X|X<U where X in lognormal(mu, s)
double dLogNormalU(double x, double mu, double s, double U)

// P(X<x|X<U) where X in lognormal(mu, s)
double pLogNormalU(double x, double mu, double s, double U)

// mean of lognormal(mu, s)
double mLogNormal(double mu, double s)
```

[19] All functions in this list are implemented in `C++`.

```
// standard deviation of lognormal(mu, s)
double sLogNormal(double mu, double s)

// mean of X|X<U where X in lognormal(mu, s)
double mLogNormalU(double mu, double s, double U)

// E(X^2|X<U) where X in lognormal(mu, s)
double E2LogNormalU(double mu, double s, double U)

// variance of X|X<U where X in lognormal(mu, s)
double varLogNormalU(double mu, double s, double U)

// standard deviation of X|X<U where X in lognormal(mu, s)
double sLogNormalU(double mu, double s, double U)

// mean of X|0<X<U where X in N(mu, s)
double mNormal0U(double mu, double s, double U)

// E(X^2|0<X<U) where X in N(mu, s)
double E2Normal0U(double mu, double s, double U)

// variance of X|0<X<U where X in N(mu, s)
double varNormal0U(double mu, double s, double U)

// standard deviation of X|0<X<U where X in N(mu, s)
double sNormal0U(double mu, double s, double U)

// L*100th percentile of lognormal(mu, s)
double qLogNormal(double L, double mu, double s,
    double INCR=0.0001)

// L*100th percentile of X|X<U where X in lognormal(mu, s)
double qLogNormalU(double L, double mu, double s, double U,
    double INCR=0.0001)

// L*100th percentile of N(mu, s)
double qNormal(double L, double mu, double s, double INCR=0.0001)

// L*100th percentile of X|0<X<U where X in N(mu, s)
double qNormal0U(double L, double mu, double s, double U,
    double INCR=0.0001)

// density (1): beta*f1(x)+(1-beta)*f2(x),
// where f1 is density of lognormal(mu1, s1),
// f2 is density of N(mu2, s2) truncated from below at 0,
// and beta in [0,1]
double dMixture(double x, double mu1, double s1,
    double beta, double mu2, double s2)
```

```
// P(X<x) where X has density (1)
double pMixture(double x, double mu1, double s1,
    double beta, double mu2, double s2)

// P(X<x|X<U) where X has density (1)
double pMixtureU(double x, double mu1, double s1,
    double beta, double mu2, double s2, double U=HUGE_VAL)

// density of X|X<U where X has density (1)
double dMixtureU(double x, double mu1, double s1,
    double beta, double mu2, double s2, double U=HUGE_VAL)

// mean of X|X<U where X has density (1)
double mMixtureU(double mu1, double s1,
    double beta, double mu2, double s2, double U)

// E(X^2|X<U) where X has density (1)
double E2MixtureU(double mu1, double s1,
    double beta, double mu2, double s2, double U)

// variance of X|0<X<U where X has density (1)
double varMixtureU(double mu1, double s1,
    double beta, double mu2, double s2, double U)

// standard deviation of X|0<X<U where X has density (1)
double sMixtureU(double mu1, double s1,
    double beta, double mu2, double s2, double U)

// L*100th percentile of X|0<X<U where X has density (1)
double qMixtureU(double L, double mu1, double s1,
    double beta, double mu2, double s2, double U,
    double INCR=0.0001)

// compute KS-test statistic given vector of data points
// NULL-distribution: X|0<X<U where X has density (1)
// do not reject null hypothesis on NULL-distribution if
// KS < critical value from KS-distribution
double KS_Mixture(vector<double> data, double mu1, double s1,
    double beta, double mu2, double s2, double U=HUGE_VAL)

// generate n variables from X|A<X<Upper,
// where X in Normal(mu, si)
// (adaption of Box-Muller polar method)
// process aborts after MAX_ITER traverses
vector<double> generate_normal(int n, double mu=0.0, double si=1.0,
    double A=0.0, double Upper=HUGE_VAL, int MAX_ITER=500000)

// generate n variables from X|A<X<Upper,
```

```
// where X in lognormal(mu, si)
// (adaption of Box-Muller polar method)
// process aborts after MAX_ITER traverses
vector<double> generate_lognormal(int n, double mu=4.5, double si=0.8,
    double A=0.0, double Upper=HUGE_VAL, int MAX_ITER=500000)

// generate m variables from X|A<X<Upper,
// where X has density (1)
// (using generate_normal and generate_lognormal)
vector<double> generate_mixture(unsigned int m, double mu1, double s1,
    double beta, double mu2, double s2, double A,
    double Upper=HUGE_VAL)

// loglikelihood of (mu1, s1, beta, mu2, s2) for X|X<U
// where X has density (1)
double logLik0U(vector<double> data, double mu1, double s1,
    double beta, double mu2, double s2, double U)

// maximize likelihood of X|X<U
// where X has density (1),
// w.r.t. mu1, s1, mu2 and s2:
double mu1HeuristicU(vector<double> data, vector<double> r,
                     double mu1, double s1,
                     double mu_low, double mu_high, double U, double INCR)
double s1HeuristicU(vector<double> data, vector<double> r,
                    double mu1, double s1,
                    double s_low, double s_high, double U, double INCR)
double mu2HeuristicU(vector<double> data, vector<double> r,
                     double mu2, double s2,
                     double mu_low, double mu_high, double U, double INCR)
double s2HeuristicU(vector<double> data, vector<double> r,
                    double mu2, double s2,
                    double s_low, double s_high, double U, double INCR)

// defines sum that skips NA points
double sum_NA(vector<double> x)

// estimate parameters in distribution of X|X<U
// where X has density (1),
// limit parameters by param_limits_*
// INCR_* is size of increment in heuristic optimization
// perform at most LIMIT iterations
// TOL defines the stopping criterion
vector<double> emLnNMixU(vector<double> data,
    vector<double> param_limits_low,
    vector<double> param_limits_high,
    double U,
    double mu10=4.8, double s10=0.8,
    double beta0=0.8, double mu20=280.0, double s20=60.0, // initial values:
```

```
    double INCR_LN = 0.01, double INCR_N=0.2,
    unsigned int LIMIT=100, double TOL=0.15)

// estimate mixed model with partially observed data
// (we cannot observe > A)
// parameters limited by param_limits_*
// iterate no more than MAX_ITERATIONS
// TOL controls criterion of convergence.
vector<double> trunc_EM_MIX_L(vector<double> data_obs,
    int N, // size of FULL sample
    vector<double> params, // (mu1, s1, beta, mu2, s2), initial values when called
    vector<double> param_limits_low,
    vector<double> param_limits_high,
    int A, // truncation point
    double U = 600.0, // X|X<U
    double TOL=0.02, int MAX_ITERATIONS=150)

// compute chi2 test statistic (goodness of fit) of data
// w.r.t. X|X<U where X has density (1)
// exp_num is the min expected count in a bin
// take steps of size INCR when computing quantiles
vector<double> chi2_test(vector<double> data,
    unsigned int exp_num,
    double mu1, double s1, double beta, double mu2, double s2,
    double U, double INCR=0.1)
```

References

[1] Aitchison J. and Brown J.A.C., *The Lognormal Distribution with Special Reference to its Uses in Economics*, Cambridge 1966.

[2] Dempster A. P., Laird N. M. and Rubin D. B., Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society.* Series B (Methodological), Vol. 39, No. 1. (1977), pp. 1-38.

[3] Diebolt J. and Ip E.H.S., A Stochastic EM Algorithm for approximating the maximum likelihood estimate, *Technical Report* No. 301, Department of Statistics, Standford University, 1994

[4] Durbin J., Kolmogorov-Smirnov tests when parameters are estimated with applications to tests of exponentiality and tests on spacings, *Biometrika* Vol. 62. No. 1, 1975, p.5-22

[5] Gibrat R., *Les inégalités économiques*, Paris 1931.

[6] Givens G.H. and Hoeting J.A., *Computational S§tatistics*, Second edition, Hoboken 2013.

[7] Kain M., Inspektionen för socialförsäkringen, Effektiv köhantering § exemplet bostadstillägg, Rapport 2012:1.

[8] Kain M., Medelberg M., Wahlfridsson A. and Karlsson X., Inspektionen för socialförsäkringen, Effektiv köhantering § exemplet Tidig bedömning, Rapport 2010:7.

[9] Johnson N.L., Kotz S. and Balakrishnan N., *Continuous Univariate Distributions.* Volume 1, Second edition, New York 1994.

[10] McLachlan G. and Peel D., *Finite Mixture Models*, New York 2000.

[11] Sprent P. and Smeeton N. C., *Applied Nonparametric Statistical Methods*, Third Edition, 2000, Chapman & Hall/CRS

[12] Wei G.C.G and Tanner M.A., A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm, *Journal of American Statistical Association*, Vol.85, No. 411 (Sep. 1990)

[13] "Gibrats Law", *Wikipedia: The Free Encyclopedia*, 2018-01, `https://en.wikipedia.org/wiki/Gibratslaw`

[14] "Power Law", *Wikipedia: The Free Encyclopedia*, 2018-01, URL: `https://en.wikipedia.org/wiki/Power_law`