

LAPPEENRANTA UNIVERSITY OF TECHNOLOGY

Faculty of Technology

Department of Mathematics and Physics

Statistical Analysis of an SEIR Epidemic Model

The topic of this Master's thesis was approved by the departmental council of the Department of Mathematics and Physics on 12 November, 2008.

The examiners of the thesis were Professor Heikki Haario and PhD Matti Heiliö
The thesis was supervised by Professor Heikki Haario.

Lappeenranta, February 4, 2009

Ndanguza Rusatsi Denis
Ruskonlahdenkatu 13-15
53850 Lappeenranta, Finland
Phone: +358449480941
Denis.Rusatsi@lut.fi

Abstract

Lappeenranta University of Technology
Department of Mathematics

Ndanguza Rusatsi Denis

Statistical Analysis of an SEIR Epidemic Model

Master's thesis

2009

75 pages, 20 figures, 4 tables

Examiners : Prof Heikki Haario and PhD Matti Heiliö

Key words: posterior distribution, Markov Chain Monte Carlo, estimates, analysis, model solution

This thesis was focussed on statistical analysis methods and proposes the use of Bayesian inference to extract information contained in experimental data by estimating Ebola model parameters. The model is a system of differential equations expressing the behavior and dynamics of Ebola. Two sets of data (onset and death data) were both used to estimate parameters, which has not been done by previous researchers in [1]. To be able to use both data, a new version of the model has been built. Model parameters have been estimated and then used to calculate the basic reproduction number and to study the disease-free equilibrium.

Estimates of the parameters were useful to determine how well the model fits the data and how good estimates were, in terms of the information they provided about the possible relationship between variables. The solution showed that Ebola model fits the observed onset data at 98.95% and the observed death data at 93.6%.

Since Bayesian inference can not be performed analytically, the Markov chain Monte Carlo approach has been used to generate samples from the posterior distribution over parameters. Samples have been used to check the accuracy of the model and other characteristics of the target posteriors.

Acknowledgements

This thesis owes its existence to the help, support and inspiration of many people. I would like to express my sincere appreciation and gratitude to Prof. Heikki Haario. Working with him was a great pleasure and I am grateful to him for our discussions about Ebola modeling and its process, especially in MCMC approaches. Nothing would have been done if he didn't assist me in Matlab codes and encourage to stand for the epidemiology thesis topic.

I would like to thank the Lappeenranta University of Technology authorities who granted me scholarship for Masters program. My thanks are especially addressed to my second supervisor Matti Heiliö who initiated the cooperation between the Kigali Institute of Science and Technology (KIST) and the Lappeenranta University of Technology (LUT).

My thanks are also addressed to the Kigali Institute of Science and Technology (my institution) who granted me permission to undertake this master's program at LUT.

My sincere acknowledgements are addressed to my wife Anastasie Uwababyeyi and my two sons Cédric Ndanguza Ganza and Crispin Ndanguza Bigwi for their patience during my two years of absence at home. I'm proud for their moral support and prayers.

Finally, I would like to thank my classmates and all the staffs of the department of Mathematics and Physics at LUT. I will never forget how they familiarized me in studying and living in Lappeenranta.

Thanks a lot!

Lappeenranta, February 4, 2009

Ndanguza Rusatsi Denis

Contents

1	Introduction	1
1.1	Background of the Thesis	1
1.2	Objectives of the Thesis	2
1.3	Structure of the Thesis	3
2	Bayesian approach in parameter estimation	5
2.1	Linear and nonlinear models	5
2.2	The Bayes formula	6
2.3	Predictive distribution	7
2.4	Prior distribution	8
2.5	Maximum Likelihood and Maximum A Posteriori in estimation of parameters	8
3	Markov Chain Monte Carlo Methods (MCMC)	11
3.1	Monte Carlo Integration methods	11
3.1.1	Crude Monte Carlo	12
3.1.2	Acceptance-Rejection	15
3.1.3	Stratified Sampling	17
3.1.4	Importance Sampling	17
3.2	Markov Chain	18
3.2.1	Properties of Markov chains	19

3.2.2	Stationary distribution	21
3.3	The Metropolis-Hastings algorithm	23
3.4	The Gibbs sampler	25
3.5	Adaptive MCMC	27
3.6	Implementing MCMC	29
3.6.1	MCMC initialization	30
3.6.2	Burn-In and Thinning	33
4	Case study: Ebola Hemorrhagic Fever (EHF) in the Democratic Republic of Congo (Zaire), 1995	35
4.1	Introduction	36
4.2	Ebola data and Model description	37
4.2.1	Describing the Ebola model and transmission rate equation . .	37
4.2.2	Ebola Data	40
4.3	Parameters estimation	42
4.3.1	Chowell et al. mistake and a new model version to estimate parameters	43
4.4	Calculus of the steady points	45
4.5	The Ebola Basic Reproduction Number (R_0)	48
4.6	The model solution and discussions	51
4.6.1	Model Solution	52
4.6.2	Evaluating Model Fit	54

4.7	MCMC Results and Interpretation	56
4.7.1	MCMC parameters estimation	56
4.7.2	The Chains time-series plot	58
4.7.3	Predictive MCMC plots	60
5	Conclusion	63

List of Figures

3.1	Estimating π by Monte Carlo Integration by accept-reject method . . .	16
3.2	Example of Gibbs sampling on a $2D$ Gaussian distribution	26
3.3	Gaussian distribution of x_1	27
4.1	The transmission diagram according to Chowell et al.	38
4.2	The behavior of the transmission rate function $\beta(t)$	39
4.3	Onset data from the 1995 Ebola outbreak in the Democratic Republic of Congo (Zaïre) from March 1 (corresponding to day 1 on the x -axis) to July 21	40
4.4	Death data from the 1995 Ebola outbreak in the Democratic Republic of Congo (Zaïre) from March 1 (corresponding to day 1 on the x -axis) to July 21	41
4.5	Cumulative sum of infected and death cases for Ebola epidemics in Zaïre 1995	42
4.6	The diagram of transmission for the new model	44
4.7	Distribution of eigenvalues taking β as a vector	48
4.8	Susceptible behavior in Ebola Epidemic model	52
4.9	Exposed, Infected and Recovered behavior in Ebola Epidemic model .	53
4.10	Autocorrelation plot showing the behavior of residuals taking a num- ber of 140 lags	54
4.11	Model fitting data plots: The solution using the onset and death data cases	55
4.12	Histograms of the distribution for the basic reproduction number (R_0) and standard deviation	57

4.13	Simulated chains of unknown parameters plots	58
4.14	Simulated histogram of chains for unknown parameters	59
4.15	The pairwise scatter plots for the unknown parameters $\beta_0, \beta_1, q, 1/k,$ $1/\gamma$ and f	60
4.16	The behavior of the solution model prediction	61
4.17	The Predictive distribution plot of onset (C) cases in Ebola model . .	62

List of Tables

3.1	Table of 16 random numbers compiled in $f(x) = \frac{e^x-1}{e-1}$	15
4.1	Known Cases and Outbreaks of Ebola Hemorrhagic Fever, in Chronological Order: www.cdc.gov	36
4.2	Estimated Ebola Epidemic Model Parameters using onset and death data by least square method	44
4.3	Posterior mean and posterior standard deviation of the estimated parameters of Ebola model. In parenthesis there are the nominal 95% confidence intervals	57

1 Introduction

Mathematical modeling has emerged as an important tool for gaining understanding of the dynamics of the spread of infectious diseases. The need of accurate model describing the epidemic process is vital, because infectious diseases outbreaks disturb the host population and has financial and health consequences. In this way, statistical analysis studies the model and its fitness to observed data by considering uncertainties by means of probabilistic reasoning. For a full statistical treatment of uncertainties, all the unknown quantities are described by statistical distributions, whether they are model parameters, unknown states of the system in equation, model prediction, or prior information on the structure of the required model solutions [9].

In this section, three points will be focussed on. The first one is the background of the thesis where the reader will find what motivated the author to face the topic and environment of the research. The second point is the objectives of the thesis where the main goal and research questions are presented in an explicit way and how the solution can be found. The last point presents the structure of the thesis where the outline of it is shortly presented and each section described briefly. The point starting after few introduction is the background of the thesis presented in the following subsection.

1.1 Background of the Thesis

In order to prevent, or at least reduce an infection spread, there is a need of models that can accurately capture the main characteristics of the disease in question since understanding disease propagation is vital for the most effective reactive measures. This will be reliable if model parameters are well estimated.

This work discusses different methods used in statistical analysis of mathematical models during the estimation of unknown parameters and their distribution in models, based on measured data. The task in this work is not only to estimate the unknown model parameters, but also to estimate the distribution of the parameters. This leads to a Bayesian problem formulation where the unknown quantities in the models are thought to be random variables with certain distributions.

In this work, the application is concerned on a model with ordinary differential equations where the human population is divided into four compartments containing susceptible, infected but not yet infectious (Exposed), infectious, and removed (recovered or dead) individuals from the system. These susceptible-exposed-infected-recovered (SEIR) models are usually expressed as a system of differential equations, where the rates of flow between compartments are determined by parameters specific to the natural history of the disease.

The Case of this thesis topic is the Ebola Hemorrhagic fever in the Democratic Republic of Congo (Zaire), 1995. It was in May, 1995 when the Centers for Disease Control and Prevention (CDC) confirmed an outbreak of Ebola in Kikwit, DRC. The outbreak was of unknown magnitude and in a total of 316 cases, 256 have been killed which corresponds to 81% of fatality. It was on 9th May that Ebola was confirmed to be the outbreak in the area. Then, the intervention started with the arrival of international medical teams and doctors without borders. Data of this Ebola are available because this was the outbreak followed in deep by doctors and other health researchers who gathered them.

Analysis for epidemic models is complicated by the fact that, one or several of the model variables may be unobserved. In addition, model parameters may change over time, for instance, if interventions to control the spread of the disease were introduced during the epidemics [13]. More commonly, observed data sets are time series of counts of events that have occurred during time intervals such as a day or a week. In this work's case, observations are daily.

After the presentation of the background of the thesis, the interest is to know how to use different methods to solve this real-life problem. This will be shown briefly in the following point which is the objectives of the thesis.

1.2 Objectives of the Thesis

The objectives of this thesis is to apply the SEIR epidemic model to a real-life situation such as Ebola a terrible disease with high fatality. With the SEIR model, one of the model parameters is a function assigning the intervention process. All those parameters have to be estimated and the reader of this thesis will be able to understand methods used to estimate them and their distribution. The reader will

also be able to apply SEIR epidemic model to his/her case study following methods used in this thesis or even more.

The thesis is a case study: the SEIR model is applied to a real-life application. The goal of this thesis is to show how the statistical analysis theories and algorithms can be implemented.

The model supposed to be used was built by Chowell et al. in [1]. Because Chowell and his group made a mistake of using onset data to estimate parameters (one parameter is not affected by the data), we made a new model which allows us to estimate parameters by using both onset and death data.

The data used, were collected in [8] and were fit to a simple deterministic (discrete-time) SEIR epidemic model. The least-square fit of the model provides estimates for the epidemic parameters (optimal parameters). The estimated parameters can then be used to calculate the basic reproductive number and the quality of the disease-free equilibrium. Estimates will also quantify the impact of intervention measures on the transmission rate of the disease. These estimated parameters will help to find the distribution of the unknown parameters using MCMC methods according to the number of simulation.

The model solution is achieved by feeding it with a set of observed epidemiological data from the town of Kikwit in the Democratic Republic of Congo at a given time (t). The value of variables at any other time can be derived from these data and can be presented in the form of time trends. To find the model solution, the calculations require the assistance of some powerful software such as MATLAB.

1.3 Structure of the Thesis

The thesis begins with an introduction where its main objectives and structure are described. This introduction is followed by an introduction to the Bayesian approach in estimating parameters (section 2). In this section, some points are reviewed briefly such as the linear and non linear models, the Bayes formula, predictive and prior distribution and likelihood in estimation of parameters.

Elementary theories on Markov Chain Monte Carlo sampling methods are discussed in section 3, where, after reviewing some Markov methods in estimating the nor-

malizing integral in Bayesian formula, Markov chains and their properties as well as their distributions are discussed. The Metropolis-Hastings and Gibbs algorithms are given and the adaptive MCMC methodology justified.

In section 4, Ebola Hemorrhagic Fever is detailed by giving its history and mode of contamination, its 1995 outbreak in DRC was specified. In this section, the SEIR epidemic model with intervention was used to model the Ebola and after estimating parameters, the basic reproduction number was calculated. The way data have been collected is also given in this section and how are used to estimate unknown parameters. Using the estimated parameters, the solution of the model is given. In the same section, the use of MCMC to estimate Ebola parameter model by producing chains is made.

In the same section MCMC results are discussed by studying the characteristics of the matrix formed by posterior chains such as the mean, the standard deviation and the median. From the chains different figures will be plotted showing the distribution of posteriors, the correlation between posteriors. The accuracy of the model is shown by the MCMC predictive plot. Finally the conclusion and references are also given. To start the theoretical part, the next section passes through the Bayesian approach utilities in parameters estimation and distribution.

2 Bayesian approach in parameter estimation

In this section, the thesis will review the fundamentals of the Bayesian paradigm basic non-technical level. Generally, a model is often written in the form of

$$y = f(x; \theta) + \epsilon \quad (2.1)$$

Here y stands for the observations or measurements, x as known quantities (constants, control variables etc.), θ as unknown parameter and ϵ as measurement error. The problem remains on how the parameters θ based on the measurements y are estimated. Numerical methods for solving this problem based on random sampling are presented in the framework of Bayesian theory. The general model in 2.1 can be either linear or nonlinear according to the parameter θ .

2.1 Linear and nonlinear models

A model is said to be linear with respect to its unknown parameter θ if it can be written as $y = f(X)\theta$, where X represents the design (input) variables. The linear model can be written in extension as bellow:

$$y = \theta_0 + \theta_1 x_1 + \dots + \theta_p x_p \quad (2.2)$$

Example of a linear model;

$$y = \theta_0 + \theta_1 x$$

This is a linear model and it is easy to find the values of unknown parameters using exact analytic formulas, the problem occurs when the model is nonlinear.

No exact theory exists for estimating unknown parameters for nonlinear models [14].

As an example of a nonlinear model, take

$$y = \theta_0 (1 - e^{-\theta_1 x})$$

Finding the parameter values of this nonlinear equation is impossible by using any exact formula. So, numerical methods are required in both finding the optimum value (best estimate) and evaluating the distribution of the parameters. The approach below is one of the most used to estimate distribution of parameters for nonlinear models.

2.2 The Bayes formula

Bayesian inference, similarly to likelihood inference, requires a model that produces the likelihood, the conditional of the data given the model parameters [14]. Additionally, the Bayesian approach will place a prior distribution on the model parameters. The likelihood and the prior are then combined using Bayes' theorem to compute the posterior distribution.

The posterior distribution is the conditional distribution of the unknown quantities given the observed data and is the object from which all Bayesian inference arises [2]. By the rules of conditional probabilities, we arrive at the Bayes formula [9]:

$$\begin{aligned}\pi(\theta) = p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ &= \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta}\end{aligned}\tag{2.3}$$

The likelihood function $p(y/\theta)$ is the density function of the observations y given θ ; $p(\theta)$ is the prior distribution, $\pi(\theta)$ the posterior distribution and $\int p(y/\theta)p(\theta)d\theta$ the normalizing constant. This formula may be derived from simple probability theory. Thus the conditional probability theorem for events A and B reads as

$$\begin{aligned}Pr(A|B) &= \frac{Pr(A, B)}{Pr(B)} \\ &= \frac{Pr(B|A)Pr(A)}{Pr(B)}\end{aligned}\tag{2.4}$$

To get the Bayes formula 2.3, replace B by observations y , A by a parameter θ . In principle, the Bayes formula solves the estimation problem in a fully probabilistic sense where the peak. The MAP (Maximum A posterior) point, of the parameter distribution is found by maximizing $\pi(\theta)$. One can also determine any required portion of the probability mass $\pi(\theta)$ [5]. However, some problems may be encountered how to define the prior distribution and how to calculate the integral of the normalizing constant, especially in high dimension (dimension higher than 2 or 3) and nonlinear cases.

Numerical computations are needed for the resolution of the above problems. New approaches such as Monte Carlo methods in subsection 3.1. In section 3 we will show the ease with which full marginal densities of parameters may be obtained by modern Markov Chain Monte Carlo (MCMC).

Before starting the Monte Carlo approach in solving the integral and parameter estimation, prior knowledge about parameters and updated (posterior) knowledge about them, as well as implications for functionals and predictions which are expressed in terms of densities is required.

2.3 Predictive distribution

Let (x_i, y_i) be the observed data and θ the parameter. We simulate MCMC chain (θ^h) as a sample from $\pi(\theta)$. Let also y_{pred} be the predictive distribution which is the expectation of $p(y_{pred}|\theta^h)$ with respect to the posterior distribution $\pi(\theta)$. Then, we generate new data (x_{new}) and find $y_{pred} = f(x_{new}, \theta^h)$. When plotting we get in the same figure, the data plot, solution plot and the new function's plot as an area with, e.g, grey color. This plot exhibits the predictive distribution of the model.

Other authors such as [9] suggest that the posterior predictive distribution can be further calculated using the following expression

$$p(y_{pred}) = \int p(y_{pred}|\theta^h)p(\theta^h)d\theta^h \quad (2.5)$$

The posterior predictive distribution can be used to check whether the model is consistent with data [14]. This is not possible if the posterior distribution of the

model parameters θ is only used, as the parameters are not directly observed. The prior distribution is important to define the behavior of the posterior distribution. This is shown below.

2.4 Prior distribution

The prior distribution is a key part of Bayesian inference and represents the information about an uncertain parameter θ that is combined with the probability distribution of new data to yield the posterior distribution, which in turn is used for future inferences and decisions involving θ [3].

According to [14], the role of prior depends on the amount of data which are needed to be used. It is argued that the basic limit theorem of Bayesian probability theory is that the posterior probability converges to the actual distribution as more data arrives. This also means that the influence of the prior on the posterior distribution diminishes as more data arrives [3]. If there is no any prior knowledge about the parameters, an uninformative prior can be used. This is the case which happens in many applications, and then, state $p(\theta) = 1$ in 2.3.

2.5 Maximum Likelihood and Maximum A Posteriori in estimation of parameters

Suppose that we have a sample y from a population of measurements. The likelihood function $p(y|\theta)$ is the probability distribution of the observations when the parameter values are given [14]. However [3] argues that the likelihood is a density function with some fixed parameter estimate θ . This helps to agree that estimate θ is called a maximum likelihood estimate (ML) if it maximizes the value of the likelihood function. This can be stated in the following way:

$$\theta_{ML} = \arg \max_{\theta} p(\theta|y) \tag{2.6}$$

In MAP (Maximum A Posteriori) estimation one maximizes the posterior,

$$\max_{\theta} p(\theta|y) = \max_{\theta} p(y|\theta)p(\theta) \quad (2.7)$$

If the prior is not specified or we want to use the uninformative prior, set $p(\theta) = 1$. Then the expression (2.7) becomes

$$\max_{\theta} p(\theta|y) = \max_{\theta} p(y|\theta) \quad (2.8)$$

The least square estimate $\hat{\theta}$ often (but not always) is the maximal value or mode of the distribution $\pi(\theta)$.

As an example from [5], let $y = f(x; \theta) + \epsilon$, where the experimental error $\epsilon \sim N(0, \sigma^2 I)$. The error terms $\epsilon_i = y_i - f(x_i; \theta)$ are independent, and each normally distributed:

$$p(y_i|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i - f(x_i, \theta))^2 / 2\sigma^2} \quad (2.9)$$

By independence, the combined probability $p(y|\theta)$ is written as the product

$$p(y|\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i - f(x_i, \theta))^2 / 2\sigma^2} \quad (2.10)$$

The expression 2.10 can be written as follows

$$p(y|\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^n (y_i - f(x_i, \theta))^2 / 2\sigma^2} \quad (2.11)$$

or

$$p(y|\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{SS(\theta)}{2\sigma^2}} \quad (2.12)$$

where $SS(\theta) = \sum_{i=1}^n (y_i - f(x_i, \theta))^2$. This $SS(\theta)$ is the familiar least square (LSQ) function and maximizing the likelihood function turns out to be equivalent to min-

imizing the LSQ function. After estimation of parameters, thanks to the computational MCMC methods more complicated and realistic models can be worked out easily. The MCMC methodology allows to undertake complex Bayesian analysis without using classical numerical analysis to calculate the normalizing integral in 2.3.

3 Markov Chain Monte Carlo Methods (MCMC)

As said before, it may be difficult to compute the posterior distribution due to the integral in the denominator of the Bayes formula, especially in multidimensional cases. When this happens the easy way is to use the MCMC algorithms. Markov Chain Simulation (also called Markov Chain Monte Carlo) is a general method based on drawing values of θ from approximate distributions and then correcting those draws to better approximate the target posterior distribution, $p(\theta, y)$. For more details and discussion, see [3] for instance.

In this section, different points are presented. Before starting the algorithms such as Metropolis-Hastings and Gibbs sample implementing the MCMC methods, different analytical methods to calculate the normalizing constant in Bayes' rule may be used. Here we focus on four Monte Carlo technical methods. After these Monte Carlo methods, Markov chains will follow together with their properties and stationarity. The first point of this section is Monte Carlo Integration methods.

3.1 Monte Carlo Integration methods

The term *Monte Carlo Methods* is used to embrace a wide range of problems solving techniques which use random numbers and the statistics of probability to investigate problems. In principle any method that uses random numbers to examine some problem is a Monte Carlo Method [2].

Monte Carlo methods tend to be used when it is infeasible or impossible to compute an exact result with a deterministic algorithm and their calculations are most suited by a computer. According to [17], the Monte Carlo methods are mainly meant for solving problems that often arise in statistical analysis, which are providing a way to generate samples from a given probability distribution and giving a solution to the problem of estimating expectations of functions under some distribution.

Thus, Monte Carlo methods are a collection of different methods that all basically perform the same process. As said above, this process involves performing many simulations using random numbers and probability to get an approximation of the answer to the problem. Of course, with the increase in computer technology, computers are now able to perform millions of simulations much more efficiently and

quickly than before. This means that the technique can provide an approximate answer quickly and to a higher level of accuracy, because the more simulations that you perform, then the more accurate the approximation is [12].

One of the most important uses of Monte Carlo methods is in evaluating difficult integrals. This is especially true of multi-dimensional integrals which have few methods for computation and thus are suited to getting an approximation due to their complexity. It is in these situations that Monte Carlo approximations become a valuable tool to use, as it may be able to give a reasonable approximation in much quicker time in comparison to other formal techniques [14].

In this section, four different Monte Carlo methods will be looked at to approach the problem of integral calculation. The investigation of each method will be done by giving a basic description of its individual procedure and then attempt to give a more formal mathematic description. These four methods are Crude Monte Carlo, Acceptance-Rejection, Stratified Sampling and Importance Sampling.

Before the discussion of the first method, we bring to the reader's attention the reasons why discussing the four different Monte Carlo methods. The first point in explaining this, is to acknowledge that there are many considerations when using Monte Carlo techniques to perform approximations.

One of the main concern is to be able to get as accurate an approximation as possible. Thus, for each method the discussion of their associated error statistics (the variance is discussed). There is also the consideration of what technique is most suitable to the problem and so will get the best results. So, the study of the four methods is needed, because they all have their individual requirements and benefits.

3.1.1 Crude Monte Carlo

Suppose that $\eta_1, \eta_2, \dots, \eta_n$ are independent observations from the parent distribution with variance σ^2 and standard deviation σ . Then by the central limit theorem, an unbiased estimator of the mean of this distribution is

$$\bar{\eta} = \frac{1}{n}(\eta_1 + \eta_2 + \dots + \eta_n) \quad (3.1)$$

and it has a standard error

$$\bar{\eta}_n = \frac{\sigma}{\sqrt{n}} \quad (3.2)$$

The Crude Monte Carlo technique will be used to solve the integral I for some function f , where $x \in \chi = \mathbb{R}^D$. There are many numerical methods to do this (e.g., Simpson's rule), but they do not work well in high dimensions. Let us consider the integral

$$\theta = \int_a^b f(x)dx. \quad (3.3)$$

A basic description of this method is that, taking a number N , for each random sample s from $U(a, b)$ (in many cases $U(0, 1)$), the function value $f(s)$ can be found with expectation θ . These values can be summed, and divide by N to get the mean value of the samples, then multiply this value by the interval $(b - a)$ to get the integral. For $i = 1, \dots, N$, and x_i distributed between a and b , this can be represented as:

$$\theta = \frac{(b - a)}{N} \sum_{i=1}^N f(x_i) \quad (3.4)$$

and the mean value is

$$\bar{f} = \frac{1}{N} \sum_{i=1}^N f(x_i) \quad (3.5)$$

is unbiased estimator of θ and its variance is

$$\frac{1}{N} \int_a^b (f(x) - \theta)^2 dx = \frac{\sigma^2}{N} \quad (3.6)$$

The next part is to describe the accuracy of this approximation technique, because otherwise the answer will not be meaningful without a description of its uncertainty. For this reason, the sample variance equation is given as:

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (f(x_i) - \bar{f})^2 \quad (3.7)$$

The standard error of \bar{f} is thus:

$$\sigma_{\bar{f}} = \frac{\sigma}{\sqrt{N}}. \quad (3.8)$$

Finally, we refer to \bar{f} as then crude Monte Carlo estimator of θ .

Using this information, the confidence interval of the model can be found and determine how accurate the answer is.

As an example from [7], take

$$f(x) = \frac{e^x - 1}{e - 1} \quad (3.9)$$

Finding the values of θ and σ in $U(0,1)$;

$$\theta = \int_0^1 \frac{e^x - 1}{e - 1} dx \quad (3.10)$$

then, $\theta = 0.418$ and $\sigma = 0.286$, the standard deviation of $U(0,1)$. Take 16 random numbers and evaluate \bar{f} with the interval confidence of the model.

In this case, it can be found that $\bar{f} = 0.4185$, so that $|\bar{f} - \theta| = 0.1325$, while the theoretical standard error is $\frac{\sigma}{4} = 0.072$. The calculation is set out in table 3.1.

Finding the variance from the formula

$$s^2 = \frac{1}{n-1} \sum_1^n (f(x_i) - \bar{f})^2 \quad (3.11)$$

giving an estimate of $s = 0.29$ for σ or 0.07 for $\frac{\sigma}{\sqrt{s}}$. The result of this calculation is as follow: $1 - \alpha$ confidence interval of f is $\hat{f} \pm z_{\alpha/2} \hat{S}$, where z_q is the q 'th quartile

Table 3.1: Table of 16 random numbers compiled in $f(x) = \frac{e^x-1}{e-1}$

i	xi	f(xi)
1	0.7666	0.6706
2	0.6661	0.5509
3	0.1309	0.0814
4	0.0954	0.0583
5	0.0149	0.0087
6	0.2882	0.1944
7	0.8167	0.7351
8	0.9855	0.9772
9	0.0174	0.0102
10	0.8194	0.7386
11	0.6211	0.5011
12	0.5602	0.4371
13	0.2440	0.1608
14	0.8220	0.7421
15	0.2632	0.1752
16	0.7536	0.6546
Average	0.4916	0.4185

of a standard $N(0, 1)$ variable. The following interval contains the value of θ , then, $\theta = 0.4185 \pm 0.07$. This result means that 0.4185 is an observation from a distribution whose mean is θ and whose standard deviation estimated at 0.07. To have good results, matlab calculations for instance when n is large are highly required. The next method, almost similar to Crude is acceptance-rejection.

3.1.2 Acceptance-Rejection

The next method for discussion is Acceptance-rejection Monte Carlo. A basic description of this technique is that, the integral is the same as in the previous point. In the (a, b) interval for any given value of x in the function, the upper limit is found. This interval is then enclosed with a rectangle that is high enough to be above the upper limit, so the truth can be that the entire function for this interval is within the rectangle [12].

Now, the process is to take random points within the rectangle and evaluate this point to see if it is below the curve or not. If the random point is below the curve, then it is treated as a successful sample. Thus, N random points should be taken and perform this check and remembering to keep count of the number of successful

samples there have been. After sampling, the integral can be approximated in the interval (a, b) and find the area of the surrounding rectangle (M). Then, multiply this area by the number of successful samples (k) over the total number of samples, and this will give the approximation of the integral for the interval (a, b) [12]. Therefore;

$$\int_a^b f(x)dx = \frac{k}{N}M(b - a) \quad (3.12)$$

To find the accuracy of the approximation, the same methods as above in crude can be used.

Consider an example of estimating the value of π from [10]. It is known that the area of a circle with radius r is πr^2 .

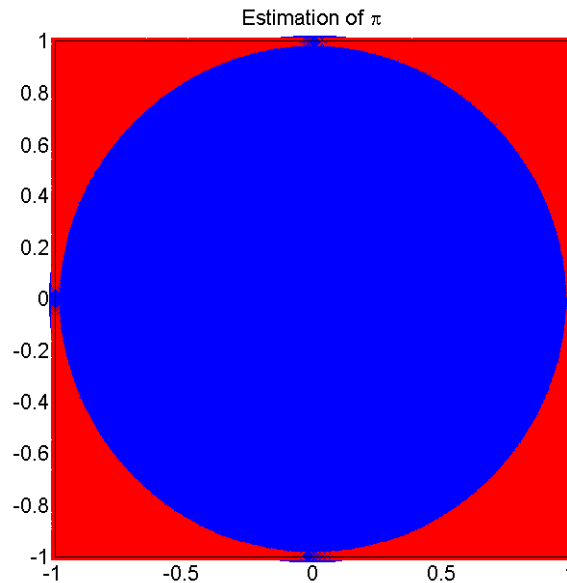


Figure 3.1: Estimating π by Monte Carlo Integration by accept-reject method

The area of a circle is given by the following equation

$$I = \int_{-r}^r \int_{-r}^r I(x^2 + y^2 \leq r^2) dx dy$$

Hence;

$$\pi r^2 = I \Rightarrow \pi = I/r^2$$

Using Matlab calculations, taking $r = 1, N = 1000000$, the following values are

found $\hat{I} = 3.1421, \hat{\pi} = 3.1421$ and standard deviation $s = 0.0016$. Points that are accepted/rejected are plotted as in figure 3.1. It is seen that the standard deviation is small when the number of simulation is high.

Blue color indicates points which are accepted (accepted points are 785564) forming the circle of radius 1 and red color stands for those which are rejected (rejected points are 214436). The third method which divides the interval in many sub-intervals and then use Crude method is the Stratified sampling which comes next.

3.1.3 Stratified Sampling

The basic principle of this technique is to divide the interval (a, b) up into sub-intervals. You then perform a crude Monte Carlo approximation on each sub-interval. The reason to use this method is that, now instead of finding the variance in one big interval, it is found by adding up the variances of each sub-interval. This may sound like doing a long-winded performance of the Crude Monte Carlo algorithm.

The advantage of the stratified sampling method, is to split the curve into parts that could have certain advantageous properties when evaluating them on their own $P(x)$ [7]. Mathematically, the integration is represented as

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx \quad (3.13)$$

Note that this equation is when the interval (a, b) has been broken into two sub-intervals (a, c) and (c, b) . The last method to be inserted is the importance sampling, its use and utility are described below.

3.1.4 Importance Sampling

This method attempts to do more samples at the regions of the function that are more important. The way it does this, is by bringing in a probability distribution function (pdf) that shows which areas (having a higher probability) of the function in the interval should get more samples. The integral can be written as follow:

$$\begin{aligned}
\theta &= \int_a^b f(x)dx = \int_a^b \frac{f(x)}{p(x)}p(x)dx \\
&= \int_a^b \frac{f(x)}{p(x)}dP(x)dx,
\end{aligned}
\tag{3.14}$$

where $p(x)$ is the probability distribution function, $P(x) = \int_0^x p(y)dy$ and $\int_a^b p(x)dx = 1$. No value of x within the interval (a, b) will $p(x)$ evaluate to 0, i.e. p is restricted to be a positive-valued function.

According to [7], $P(x)$ is a distribution function for $a \leq x \leq b$, and if η is a random number sampled from the distribution P , $\frac{f(\eta)}{p(\eta)}$ has expectation θ and variance

$$\sigma_{f/p}^2 = \int_a^b \left(\frac{f(x)}{p(x)} - \theta \right)^2 dP(x)
\tag{3.15}$$

The object in important sampling is to concentrate the distribution of the sample points in the part of the interval that are of most 'importance' instead of spreading them out evenly [7]. Some interesting examples are found in [17].

All the Monte Carlo methods above are used to estimate the integral in Bayes formula while estimating distribution of the unknown model parameters. Another way is to find the posterior distribution using a chain of parameters without calculating the normalizing constant of the Bayes' rule. This chain must have some specific properties to be used in estimating parameters. Next, we discuss Markov chains more in detail.

3.2 Markov Chain

Before introducing the Metropolis-Hastings algorithm and the Gibbs sampler, a few introductory comments on Markov chains are in order.

Definition 1. *The Markov Chain is a process where the outcome of a given experiment can affect the outcome of the next experiment.*

According to [3], a Markov chain is a sequence of random variables $X^1, X^2, \dots,$

for which, for any t , the distribution of X^t , given all the previous values of X depends only on the most recent value, X^{t-1} . In other words, samples are drawn sequentially, with the distribution of the sampled draws depending on the last value drawn. Formally,

$$P(X^{t+1} = s_{t+1} | X_0 = s_0, X^1 = s_1, \dots, X^t = s_t) = P(X^{t+1} = s_{t+1} | X^t = s_t) \quad (3.16)$$

where s_i denotes the state of the chain at time i . In a finite dimension case, a Markov chain may be described as follow: There is a set of state, $S = \{s_1, s_2, \dots, s_r\}$. The process starts in one of these states and moves successively from one state to another. Each move is called a step. If the chain is currently in state s_i , then it moves to state s_j at the next step with a probability denoted by p_{ij} , and this probability does not depend upon which states the chain was in before the current state.

The probabilities p_{ij} are called *transition probabilities*. The process may remain in the state it is in with probability p_{ii} . If the state space is discrete, the transition probability matrix can be defined as $P = [p_{ij}]$ and $\sum_j p_{ij} = 1$ for all i .

3.2.1 Properties of Markov chains

Define the probability of going from state i to state j in n time steps as

$$p_{ij}^{(n)} = Pr(X_n = j | X_0 = i) \quad (3.17)$$

and the single-step transition as

$$p_{ij} = Pr(X_1 = j | X_0 = i) \quad (3.18)$$

The *Chapman-Kolmogorov equations* [16] define the transition probability for n step and for any k such that $0 < k < n$,

$$p_{ij}^{(n)} = \sum_{r \in S} p_{ir}^{(k)} p_{rj}^{(n-k)} \quad (3.19)$$

Definition 2. The **marginal distribution** $Pr(X_n = x)$ is the distribution over states at time n and the initial distribution is $Pr(X_0 = x)$.

The evolution of the process through one time step is described by

$$\begin{aligned} Pr(X_n = j) &= \sum_{r \in S} p_{rj} Pr(X_{n-1} = r) \\ &= \sum_{r \in S} p_{rj}^{(n)} Pr(X_0 = r) \end{aligned} \quad (3.20)$$

Definition 3. A state j is said to be **accessible** from state i (written $i \rightarrow j$) if there exists an n such that

$$Pr(X_n = j | X_0 = i) > 0 \quad (3.21)$$

Definition 4. A state i is said to **communicate** with state j (written $i \leftrightarrow j$) if it is true that both i is accessible from j and that j is accessible from i .

Another Markov chain property needed in MCMC implementation is the *periodicity* of chain. A state i has period k if any return to state i must occur in some multiple of k time steps and k is the largest number with this property. Formally, the period of a state is defined as

$$k = \gcd \{n : Pr(X_n = i | X_0 = i) > 0\} \quad (3.22)$$

where \gcd is the greatest common divisor. If the chain is not periodic is said to be *aperiodic* and $\gcd \{n : Pr(X_n = i | X_0 = i) > 0\} = 1$.

Definition 5. A state i in a Markov chain is said to be **transient**, if the probability of not returning to i is non-zero. Let the random variables T_i be the next return time to state i

$$T_i = \min \{n : X_n = i | X_0 = i\} \quad (3.23)$$

then i is transient if there exists a finite T_i such that $Pr(T_i < \infty) < 1$. If the state is not transient is *recurrent* (the probability of returning to i is 1). In addition, if the expectation of the return time is positive, the state is said to be *positive recurrent*. Joining two properties, a state i is said to be *ergodic* if it is aperiodic and positive recurrent. If all states in a Markov chain are ergodic, then the chain is said to be *ergodic*. A state i is called *absorbing* if it is impossible to leave this state, i.e. $p_{ii} = 1$ and $p_{ij} = 0$ for $i \neq j$.

3.2.2 Stationary distribution

Let $\pi_j(t)$ denote the probability that the state is j and $\pi(t) = \{\pi_j(t), j = 1, \dots, k\}$ probabilities for all states at time t . That is, $\pi_t = P(X_t = s_j)$. The vector π is a *stationary distribution* (equilibrium distribution or invariant measure) if $\sum_{i \in S} \pi_i$ and satisfy

$$\pi_j = \sum_{i \in S} \pi_i p_{ij} \quad (3.24)$$

An irreducible chain has a stationary distribution if all of its states are positive-recurrent. In that case, π is unique and is related to the expected return time:

$$\pi_j = \frac{1}{M_j} \quad (3.25)$$

If the chain is both irreducible and aperiodic, then for any i and j ,

$$\lim_{n \rightarrow \infty} p_{ij}^{(n)} = \frac{1}{M_j} \quad (3.26)$$

If the state space is finite, the transition probability distribution can be represented by a matrix called the transition matrix, with the (i, j) 'th element of P equal to

$$p_{ij} = Pr(X_{n+1} = j | X_n = i) \tag{3.27}$$

and the stationary distribution π is a vector which satisfies the equation

$$\pi = \pi P \tag{3.28}$$

If the Markov chain is irreducible and aperiodic, then there is a unique stationary distribution π . In addition, P^k (the k 'th power of the transition matrix) converges to a rank-one matrix in which each row is the stationary distribution

$$\lim_{k \rightarrow \infty} P^k = \mathbf{1}\pi \tag{3.29}$$

where $\mathbf{1}$ is the column vector with all entries equal to 1. This is known as Perron-Frobenius theorem and it means that as time goes by, the Markov chain forgets where it began (its initial distribution) and converges to its stationary distribution [16].

Definition 6. *The Markov chain is said to be **reversible** with respect to a distribution π , if the detailed balance condition holds. That is*

$$\pi_i p_{i,j} = \pi_j p_{j,i}. \tag{3.30}$$

Summing over i gives

$$\sum_i \pi_i p_{i,j} = \sum_i \pi_j p_{j,i} = \pi_j \sum_i p_{j,i} = \pi_j$$

This can be written as $\pi P = \pi$.

A general and useful approach to create a Markov chain is the Markov chain Monte Carlo (MCMC) method by using random sampling so that the created chain has the posterior distribution as its unique stationary distribution (limiting distribution) [17], i.e. the MCMC methods produce ergodic Markov chains.

Markov chain simulation is used when it is not possible to sample θ directly from the posterior $\pi(\theta)$; instead the sample is done iteratively in such a way that at each step of the process there is an expectation to draw from a distribution that becomes closer and closer to $\pi(\theta)$ [14]. The most useful algorithm for drawing samples from Bayesian posterior distributions is the Metropolis-Hastings algorithm which will be introduced in the following point.

3.3 The Metropolis-Hastings algorithm

One problem with applying Monte Carlo integration is in obtaining samples from some complex probability distribution $p(x)$ and attempts to solve this problem are the roots of MCMC methods [14]. As explained in [[5],[14], [17]], the Markov Chain Monte Carlo (MCMC) algorithms generate a sequence of parameters values $\theta_1, \theta_2, \dots$ whose empirical distribution, in the histogram sense, asymptotically approaches the posterior distribution.

It can be argued that the Metropolis-Hastings algorithm is an adaptation of random walk that uses an acceptance/rejection rule to converge to the specified target distribution. In fact, the MCMC algorithm produces a chain of values in which each value can depend on the previous value in the sequence and the generation of the vectors in the chain θ_n is done by random numbers.

According to [3], the key to the method's success, however, is not the Markov property but rather that the approximate distributions are improved at each step in the simulation, in the sense of converging to the target distribution. Given a condition density $q(\theta'|\theta)$, the algorithm generates a Markov chain (θ_n) through the following steps:

1. Start with an arbitrary value θ_0
2. Update from θ_n to $\theta_{n+1}(n = 0, 1, \dots)$ by
 - Generate $\xi \sim q(\xi|\theta_n)$
 - Evaluate $\alpha(\theta_n, \xi) = \min \left(1, \frac{\pi(\xi)q(\xi, \theta_n)}{\pi(\theta_n)q(\theta_n, \xi)} \right)$
 - Set

$$\theta_{n+1} = \begin{cases} \xi & \text{with probability } \alpha, \\ \theta_n & \text{otherwise.} \end{cases}$$

The distribution $\pi(\theta)$ is often called the *target* distribution whereas the distribution with density $q(.|\theta)$ is the *proposal* distribution. If the symmetric proposal distribution holds; i.e. $q(\xi, \theta_n) = q(\theta_n, \xi)$, a particular case of the Metropolis-Hastings is found called Metropolis algorithm. The probability for the move is

$$p(\theta_n, \xi) = q(\theta_n, \xi)\alpha(\theta_n, \xi). \quad (3.31)$$

The important thing to check is the detailed balance equation

$$\pi(\theta_n)p(\theta_n, \xi) = \pi(\xi)p(\xi, \theta_n) \quad (3.32)$$

which shows that π is a stationary (invariant) distribution of the chain in the following expression:

$$\int \pi(\theta_n)p(\theta_n, \xi)d\theta_n = \pi(\xi), \quad (3.33)$$

combined with the Markov chain theory, this proves that the sampling theoretically produces correct results.

The required selection of an appropriate proposal density makes the Metropolis-Hastings algorithm more involved, but it has the advantage of being more general, and is particularly helpful for sampling parameters that lack closed, easily recognizable forms for their full conditional distributions [10].

In this particular case, the sequence of iterations $\theta^1, \theta^2, \dots$ converges to the target distribution in two steps: The first, as said in point 3.2 it is known that if the Markov chain is irreducible, aperiodic, and not transient, this simulated sequence has a unique stationary distribution. Second, it is shown that the stationary distribution equals this target distribution. And the convergence to the target distribution is proved in the same way for the Metropolis algorithm.

For other representation of the algorithm refer to [10], [3], [17] and [9]. Remind that in Bayesian computation, the posterior density $\pi(\theta)$ must be known up to a normalizing factor, but in Metropolis-Hastings, only the ratio $\frac{\pi(\xi)q(\theta_n|\xi)}{\pi(\theta_n)q(\xi|\theta_n)}$ needs to be computed. This is a key feature of this interesting algorithm. See [17] for examples of applying the Metropolis-Hastings algorithm. The second MCMC method which is popular is the Gibbs sampler, its features are seen in the following point.

3.4 The Gibbs sampler

Gibbs sampling is a special case of Metropolis-Hastings sampling where the sampled value is always accepted ($\alpha = 1$) [3]. The task remains to specify how to construct a Markov chain whose values converge to the target distribution. Suppose the parameter θ has been divided into p components or sub-vectors, $\theta = (\theta_1, \theta_2, \dots, \theta_p)$.

The idea of Gibbs algorithm is to reduce the sampling from the joint posterior distribution $\theta^1, \dots, \theta^p$ to one-dimensional full conditional distributions π_1, \dots, π_p (univariate conditional distributions) [5]. Thus, each sub-vector θ_j is updated conditional on the latest value of the other components of θ , which are the iteration t values for the components already updated, the iteration $t - 1$ values for the others, and so on. According to [5], this may be done, if the one-dimensional or conditional distribution $\pi(\theta_i | \theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_p)$ are known (the distribution of any of the parameters is known if the values of the rest of the components of θ) are fixed. The algorithm presented in [17] reads as

1. Specify an initial value $\theta^0 = (\theta_1^{(0)}, \dots, \theta_p^{(0)})$
2. For $j = 1, \dots, N$ (number of chains)
3. For $i = 1, \dots, p$ (number of parameters)
 - Sample $\theta_j^i \sim \pi(\theta_i | \theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_p)$
 - Set $chain_i(j) = \theta_j^i$
4. Return the values $(\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(N)})$

As a demonstration of Gibbs sampling (Example from [10]), try to sample from a multivariate Gaussian

$$\mathcal{N}(\vec{x} | \vec{\mu}, C) =_{def} \frac{1}{(2\pi)^{n/2} |C|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T C^{-1}(\vec{x} - \vec{\mu})\right) \quad (3.34)$$

using Gibbs sampling, take $n = (x_1, x_2)$ and its parameters

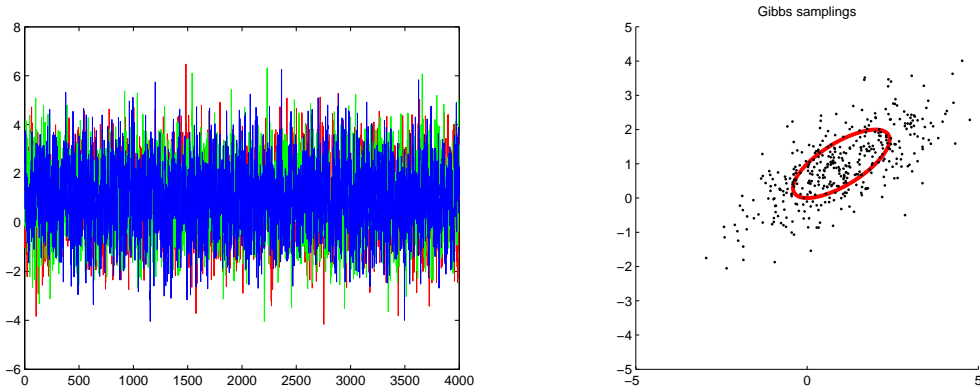


Figure 3.2: Example of Gibbs sampling on a $2D$ Gaussian distribution

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

and

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

From these expressions, $p(x_1|x_2)$ can be computed as follows:

$$\begin{aligned} p(x_1|x_2) &= \mathcal{N}(x_1; \mu_{1|2}, C_{1|2}) \\ \mu_{1|2} &= \mu_1 + C_{11}C_{22}^{-1}(x_2 - \mu_2) \\ C_{1|1} &= C_{11} - C_{12}C_{22}^{-1}C_{21}. \end{aligned}$$

Gibbs sampling is found in figure 3.2.

Figure 3.2 and 3.3 were plotted taking the number of simulation as 5000;

$$\mu = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

,

$$C = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

and burn-in=1000;

Even if it has been said above that the point taken from one-dimensional distribution is always accepted, but the creation of that one-dimensional may be difficult. If the conditional distribution for θ_k is not known in analytical form, it must be

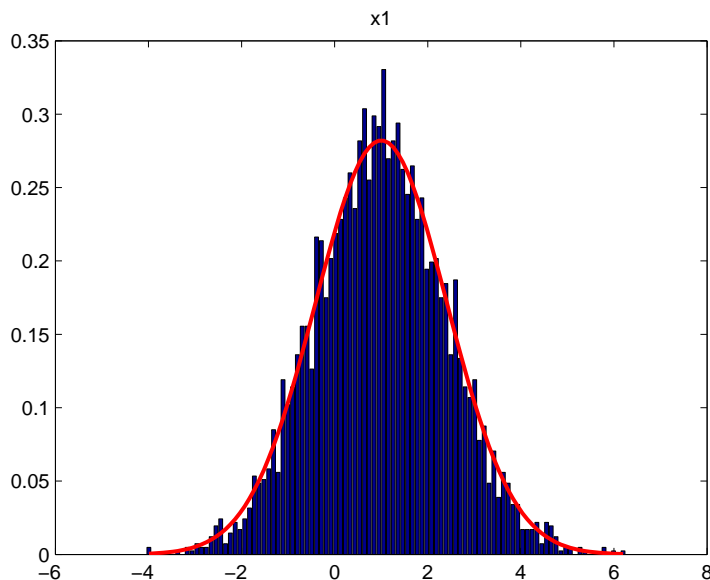


Figure 3.3: Gaussian distribution of x_1

approximatively created by evaluating the target distribution $\pi(\theta)$ with respect to the coordinate i a given number of times. Even though, the inverse CDF method requires several iterations, it will be used in the empirical distribution where the new value for θ_k is sampled.

3.5 Adaptive MCMC

In Metropolis-Hastings algorithm, the proposal distribution must be well chosen so that the sampling becomes effective. Then, the problem is to know if really this proposal matches the target distribution. According to [5], there are many methods used to improve the proposal during the run, and one simple way is to compute the covariance matrix of the chain and use it as the proposal. This process is no longer Markovian because a new point requires the knowledge of the earlier history of the chain not on the previous point [6]. Then, the crucial point of the adaptive methods (AM) adaptation is how the covariance of the proposal distribution depends on the history of the chain [6].

If $n-1$ is a time with sampled states X_0, X_1, \dots, X_{n-1} , where X_0 is the initial state, the proposal distribution $q_n(\cdot | X_0, \dots, X_{n-1})$ used is a Gaussian distribution with

mean at the current point X_{n-1} and covariance

$$C_n = s_d \text{Cov}(X_0, \dots), X_{n-1} + s_d \epsilon I_d,$$

where s_d is a parameter depending on the dimension d of the sampling space, and ϵ a small positive number [5]. Starting the adaptation procedure, C_0 an arbitrary strictly positive initial covariance matrix is defined as well as the length of the initial non-adaptation period n_0 . Then,

$$C_n = \begin{cases} C_0 & n \leq n_0 \\ s_d \text{Cov}(X_0, \dots, X_{n-1}) + s_d \epsilon I_d & n > n_0. \end{cases}$$

The empirical covariance matrix determined by points $X_0, \dots, X_k \in \mathbb{R}^d$ is defined as follow:

$$\text{Cov}(X_0, \dots, X_k) = \frac{1}{k} \left(\sum_{i=0}^k X_i X_i^T - (k+1) \bar{X}_k \bar{X}_k^T \right) \quad (3.35)$$

where $\bar{X}_k = \frac{1}{k+1} \sum_{i=0}^k X_i$.

According to [6, 5] the covariance C_n satisfies the recursive formula as follow where ϵ prevents it to be a singular matrix.

$$C_{n+1} = \frac{n-1}{n} C_n + \frac{S_d}{n} \left(n \bar{X}_{n-1} \bar{X}_{n-1}^T - (n+1) \bar{X}_n \bar{X}_n^T + X_n X_n^T + \epsilon I_d \right) \quad (3.36)$$

This formula allows the calculation of C_n without much computational cost and the mean \bar{X}_n has also a recursive formula as follows:

$$\begin{aligned}
\bar{X}_n &= \frac{X_0 + X_1 + \dots + X_n}{n+1} \\
&= \frac{n}{n+1} \frac{X_0 + X_1 + \dots + X_n}{n} + \frac{X_n}{n+1} \\
&= \frac{n}{n+1} \bar{X}_{n-1} + \frac{X_n}{n+1} \\
&= \bar{X}_{n-1} + \frac{n}{n+1} (X_n - \bar{X}_{n-1}).
\end{aligned} \tag{3.37}$$

By extending the one-step mean formula to k-step,

$$\bar{X}_{n+k-1} = \frac{n}{n+k} \bar{X}_{n-1} + \frac{1}{n+k} \sum_{i=n}^{n+k-1} X_i \tag{3.38}$$

The recursive formula for calculating C_{n+k} is given as follow:

$$C_{n+k} = \frac{n-1}{n+k-1} C_n + S_d \left(\frac{n}{n+k-1} \bar{X}_{n-1} \bar{X}_{n-1}^T - \frac{n+k}{n+k-1} \bar{X}_{n+k-1} \bar{X}_{n+k-1}^T + \frac{1}{n+k-1} X_{new} X_{new}^T \right).$$

where

$$X_{new} = (X_n, X_{n+1}, \dots, X_{n+k-1})$$

including new points and $C_{n+k} = f(C_n, \bar{X}_n, X_{new})$ [17]. If the target is a Gaussian distribution, with covariance matrix C , the scaling parameter usually is taken as $S_d = 2.4^2/d$.

3.6 Implementing MCMC

As said above, the main idea behind MCMC is to generate a Markov chain which has as its unique limiting distribution the posterior distribution of interest. As known in different literature, there is no guarantee, no matter how long you run the MCMC algorithm for, that it will converge to the posterior distribution. To run a MCMC algorithm initialization of parameters is needed, the way it is done is seen in the following point.

3.6.1 MCMC initialization

Using the random Walk Metropolis algorithm with a Gaussian proposal distribution, we get the Covariance Matrix C to be used as the proposal covariance. It is known that to have a credible convergence, it is advised to choose the starting point θ_0 correctly [5], e.g.,

$$\theta_0 = \min_{\theta} \sum_{i=1}^n (y_i - f(x_i, \theta))^2.$$

By this formula we start with point that suits the data in LSQ sense. The covariance matrix C is obtained by linearization method (function $l(\theta)$) based on Taylor expansion around the estimated point.

$$l(\theta) = l(\hat{\theta}) + \nabla l(\hat{\theta})^T (\theta - \hat{\theta}) + \frac{1}{2} (\theta - \hat{\theta})^T H(\theta - \hat{\theta}) + \dots \quad (3.39)$$

where H denotes the Hessian matrix of the second derivatives of $l(\theta)$. Considering the LSQ function as follows:

$$l(\theta) = \sum_{i=1}^n (f(x_i, \theta) - y_i)^2$$

calculating the derivative of $l(\theta)$,

$$\frac{\partial l(\theta)}{\partial \theta_p} = 2 \sum_{i=1}^n (f(x_i, \theta) - y_i) \frac{\partial f(x_i, \theta)}{\partial \theta_p}$$

The second derivative of $l(\theta)$ at $\theta = \hat{\theta}$ is as follow:

$$\frac{\partial^2 l(\hat{\theta})}{\partial \theta_p \partial \theta_q} = 2 \sum_{i=1}^n \frac{\partial f(x_i, \hat{\theta})}{\partial \theta_p} \frac{\partial f(x_i, \hat{\theta})}{\partial \theta_q} + 2 \sum_{i=1}^n \left(f(x_i, \hat{\theta}) - y_i \right) \frac{\partial^2 f(x_i, \hat{\theta})}{\partial \theta_p \partial \theta_q}$$

Assuming that the residuals $f(x_i, \hat{\theta}) - y_i$ are very small, they will be omitted and

approximate the Hessian matrix as follow:

$$H_{pq} = \frac{\partial^2 l(\hat{\theta})}{\partial \theta_p \partial \theta_q} \simeq 2 \sum_{i=1}^i \frac{\partial f(x_i, \hat{\theta})}{\partial \theta_p} \frac{\partial f(x_i, \hat{\theta})}{\partial \theta_q} \quad (3.40)$$

The Jacobian matrix J of the first derivative is defined by

$$J_{ip} = \left. \frac{\partial f(x_i, \theta)}{\partial \theta_p} \right|_{\theta=\hat{\theta}}$$

Writing the above expression in matrix form $H \simeq 2J^T J$ and substituting it to a form of the Taylor expansion,

$$l(\theta) \simeq l(\hat{\theta}) + (\theta - \hat{\theta})^T J^T J (\theta - \hat{\theta}) \quad (3.41)$$

with three terms only.

In linear case, the function

$$l(\theta) = \|X\theta - y\|^2 = \theta^T X^T X \theta - 2y^T X \theta + y^T y$$

which is a quadratic polynomial in θ . Writing it as a quadratic polynomial of $\theta - \hat{\theta}$, the expression becomes

$$l(\theta) = (\theta - \hat{\theta})^T X^T X (\theta - \hat{\theta}) + D \quad (3.42)$$

where D does not depend on θ . Substituting $\theta = \hat{\theta}$ yields $D = l(\hat{\theta})$ and

$$D = l(\hat{\theta}) = \|X(X^T X)^{-1} X y - y\| \quad (3.43)$$

and in the linear case

$$l(\theta) = l(\hat{\theta}) + (\theta - \hat{\theta})^T X^T X (\theta - \hat{\theta}) \quad (3.44)$$

Comparing these two values of $l(\theta)$ it is seen that in the linear approximation, the Jacobian matrix assumes the role of design matrix X of a linear model and

$$C = Cov(\hat{\theta}) \simeq \sigma^2 (J^T J)^{-1} \quad (3.45)$$

Using the residuals of the fit to calculate the variance by assuming that residuals are equal to measurement error,

$$\sigma^2 \approx \sigma_{MSE}^2 = \frac{RSS}{n-p} = \frac{\sum_{i=1}^n (y_i - f(x_i, \theta))^2}{n-p} \quad (3.46)$$

where RSS (Residual Sum of Squares) is the fitted value of the least squares objective function, and MSE (Mean Square Error) is computed as an average value of residual squares, n is the degree of freedom and p the number of parameters. The following pseudo code found in [17] shows the implementation of Random Walk Metropolis algorithm:

1. Initialization

- Choose N_{simu}
- Set $\theta_1 = \min_{\theta} \sum_{i=1}^n (y_i - f(x_i, \theta))^2$
- Calculate $MSE = \frac{RSS}{(n-p)}$, where n is the number of measurements and p the length of θ .
- Set $SS_{old} = SS_{\theta_1}$
- Calculate Jacobian J
- Calculate $C = (J^T J)^{-1} * MSE$
- Calculate R so that $C = R^T R$ (Cholesky decomposition)

2. Simulation loop

for $i = 2$ to N_{simu}

- Sample $z = Z_i, i = 1, \dots, p$, and $z_i \sim N(0, 1)$

- Set $\theta_{new} = \theta_{old} + Rz$
- Sample u_α from $U[0, 1]$
- Calculate SS_{new}
- Calculate $\alpha = \min\left(1, e^{\frac{1}{2\sigma^2}(SS_{new} - SS_{old})}\right)$

if $u_\alpha < \alpha$

- Set $\theta_i = \theta_{new}$
- Set $\theta_{old} = \theta_{new}$
- $SS_{old} = SS_{new}$

else

- Set $\theta_i = \theta_{old}$

endif

endfor

In fact, there are two practical issues that need investigation to establish the reliability of the chain outcome. In the first issue, there will be a way of determining the number of iterations needed for a given level of precision in a MCMC algorithm.

This method can help to diagnose lack of convergence or slow convergence due to bad starting values. This first of these two is the *burn-in*, i.e. the number of iterations that need to be discarded from the output. The second practical issue of monitoring the convergence is that after the bur-in, some *thinning* of the chain is required.

3.6.2 Burn-In and Thinning

As said in previous text, the Markov chains produced by the proposal distribution are ergodic, i.e. the distribution of (θ_n) converges as n tends to infinity, to $\pi(\cdot|y)$ for every starting value (θ_0) [14]. To diminish the effect of the starting distribution, a part of each sequence will be discarded and focuss attention on the second part of the chain.

The assumption here is that the distribution of the simulated values θ_n for large enough n , are close to the target distribution $\pi(\theta)$. Therefore, burn-in is the practice

of discarding early iterations in Markov chain simulation. However, the speed of convergence varies depending on the posterior state-space and the sampler used [14]. Thus, for n large enough, the resulting θ_n is an approximate sample from $\pi(\theta|y)$ (for the Law of Large Numbers, Central Limit Theorem are found in [14]). The real problem in MCMC algorithm is to determine what a large n mean. For the case of this thesis, the number of iterations is 1000000 with burn-in 1000 and all the results seem to converge.

Another issue that sometimes arises, once approximate convergence has been reached, is whether to thin the sequences by keeping every k th simulation draw from each sequence and discarding the rest. Whether or not the sequences are thinned, if the sequences have reached approximate convergence, they can be directly used for inferences about the parameters θ and other quantities of interest [3].

All these theories will be applied on a real-life situation, Ebola Hemorrhagic Fever. This application will be based on modelization of Ebola epidemic where parameters have been estimated through observed data.

4 Case study: Ebola Hemorrhagic Fever (EHF) in the Democratic Republic of Congo (Zaire), 1995

Ebola remains a serious public health risk in some African regions such as the Democratic Republic of Congo (DRC) where several outbreaks have been observed since the first appearance of the epidemic in 1972. Many strategies have been implemented by government and local health organizations to prevent further occurrences but it is still difficult to maintain the exact measures of control to eradicate Ebola disease.

As said in the introduction of this thesis, the modelling of different infectious diseases improves the understanding of the dynamics of the spread of them. The application of this work is based on Ebola Hemorrhagic Fever (EHF), a disease which kills many people in Africa, especially in Democratic Republic of Congo.

The statistical analysis of infectious disease data usually requires the knowledge of development of the epidemic in question. It is difficult to observe the entire infection process, it is why the incidence of an infectious disease consist of only the final number of infected individuals. The outbreak process of a disease is complicated because even if data contain the time that the symptoms occur, it is not easy to observe the actual infection times and who infects who is not observed either.

In this section, the history of Ebola will be briefly presented followed by the SEIR model with intervention rate equation. Of course, in order to accurately analyze outbreak data, a model describing the infection pathway is needed. It will be shown that the onset data alone will not identify the parameters, as mistakenly reported by Chowell and al. in [1]. To correct this mistake we will make a new model and combine two sets of data to estimate parameters.

With the new model, the estimation of parameters have been done and then used to conclude the behavior of disease-free equilibrium (E_0) and estimation of basic reproduction number (R_0). The last point will be the use of MCMC to check the accuracy of Ebola model. First, the introduction of Ebola is debated below where its contamination and history are established.

4.1 Introduction

The country of Democratic Republic of Congo (DRC) is susceptible to a vast array of outbreaks from many diseases including Ebola fever. Lack of sanitation, indoor air pollution, inadequate hygiene and insufficient water suppliers increase the risk for ill health [18]. An estimated 1200 people die each day as a result of conflict-related causes such as poverty, disease and gender-based violence. There are four million orphaned children and one-fifth (1/5) does not reach the age of 5 years. 1.1 million are living with HIV/AIDS, 60% of whom are women, and 100,000 deaths annually are caused by AIDS [18].

Definition 7. *Ebola* is the common term for a group of viruses belonging to genus *Ebola virus*, family *filoviridae*, and for the disease which they cause, *Ebola hemorrhagic fever* [20].

Chowell defined Ebola hemorrhagic fever as a highly infectious and lethal disease named after a river in the Democratic Republic of Congo (formerly Zaire) where it was first identified in 1976 with a high case fatality range lying between 50% and 90% [1]. This is confirmed in the table 4.1 which shows the Ebola outbreaks from 1976-2007 in the Democratic Republic of Congo (former Zaire) [18].

Table 4.1: Known Cases and Outbreaks of Ebola Hemorrhagic Fever, in Chronological Order: www.cdc.gov

Years	Cases	Death	Percentage (%)
1976 (September-October)	318	280	88
1977 (June)	1	1	100
1995 (April-June)	316	256	81
2001-2002 (October-March)	59	44	75
2002-2003 (December-April)	143	128	90
2003-2004 (November-January)	35	29	83
2007	249	183	78

Ebola reached the DRC at the first time in 1976, and its first outbreak happened in Cameroon in 1972. Its origin is still unknown, many assumptions say that its viruses are transmitted to humans from discrete life cycles in animals or insects, but regardless of the original source, person-to-person transmission is the means by which Ebola outbreaks and epidemics progress [1].

Ebola is transmitted by physical contact with body fluids, secretions, tissues or semen from infected persons [19]. It is why many health care workers are infected

while attending patients. Transmission also occurs through preparation of the dead body for burial arrangements.

The incubation period is from 2 to 21 days (5 -12 days in most cases) [19]. The disease begins with acute fever, diarrhea that can be bloody and vomiting followed by headache, nausea, and abdominal pain which are common [1]. Because there is no specific treatment or vaccine, infected individuals only receive limited care to try their recovery. Most infected people die within 10 days of their initial infection [1].

From the table 4.1, this thesis application was based on the 1995 outbreak in Kikwit town. The way data have been found and how the model has been built is described next.

4.2 Ebola data and Model description

This part contains the description of Ebola outbreak during the year 1995 in Kikwit and giving the way data have been found and also presenting the data in time series plots. Another point is the construction of the model describing the process and evolution of Ebola through differential equations compartments with parameters expressing the rates of immigration. The following point will describe mathematically the Ebola process. This is the contribution of mathematics in real-life problems, where mathematical results are interpreted according to some well known epidemiological theories.

4.2.1 Describing the Ebola model and transmission rate equation

In this section, there is a description of a simple model for the transmission of infectious diseases where the population is assumed to be closed (the population does not contain the demographic changes). Hence, the assumption is that during the outbreak of the epidemic no births or natural deaths occur.

The total effective population size N is divided in four compartments: Susceptible individuals in class S after being contacted with the virus enter the exposed class E at the per-capita rate $\frac{\beta I}{N}$, where β is the transmission rate per person per day, $\frac{I}{N}$ is the probability that a contact is made with an infectious individual. Exposed

individuals undergo an average incubation period of $1/k$ days before progressing to the infectious class I . Infectious individuals move to the R class (death or recovered) at the per-capita rate γ [11]. In Ebola case, the compartment R is called removed because individuals reaching it will never have chance to join the process.

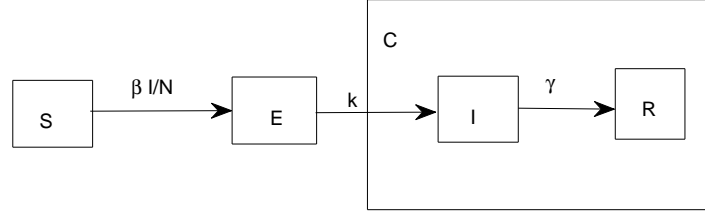


Figure 4.1: The transmission diagram according to Chowell et al.

The transition process shown in figure 4.1 is modeled by the following system of differential equations [1, 11, 13]

$$\begin{aligned}
 \frac{dS(t)}{dt} &= -\beta(t) \frac{S(t)I(t)}{N} \\
 \frac{dE(t)}{dt} &= \beta(t) \frac{S(t)I(t)}{N} - kE(t) \\
 \frac{dI(t)}{dt} &= kE(t) - \gamma I(t) \\
 \frac{dR(t)}{dt} &= \gamma I(t) \\
 \frac{dC(t)}{dt} &= kE(t)
 \end{aligned} \tag{4.1}$$

where $S(t)$, $E(t)$, $I(t)$ and $R(t)$ denote the number of susceptible, exposed, infected and removed at time t respectively. $C(t)$ is not a compartment but serves to keep track of cumulative number of Ebola cases from the time of onset of symptoms [1]. Epidemic models of this kind, where an individual is allowed to be in any of four states, susceptible, exposed, infective or removed, are often called *SEIR* epidemic model. It is seen that N is constant because $\frac{d(S+E+I+R)}{dt} = 0$, and $S(t) + E(t) + I(t) + R(t) = N \Rightarrow \frac{dN}{dt} = 0$, which shows that N is a constant. The model 4.1 has been used by Chowell [1] to calculate the basic reproduction number.

On the other hand, the transmission rate β is a function of time ($\beta(t)$). In order to account for the control intervention, assume that the transmission parameter $\beta(t)$ is constant ($\beta(t) = \beta_0$) at time between $t = 0$ and $t \leq \tau$, and after that, at $t > \tau$ it decays exponentially to be constant at $\beta = \beta_1$. This can be formulated as

$$\beta(t) = \beta_0 + (\beta_1 - \beta_0) / (1 + e^{-q(t-\tau)}) \quad (4.2)$$

where τ is the time at which interventions start and $q > 0$, is the rate of decays from β_0 to β_1 . The effect of intervention is to reduce the transmission rate β from β_0 to $\beta_1 < \beta_0$ [1]. This transmission rate is explained by figure 4.2.

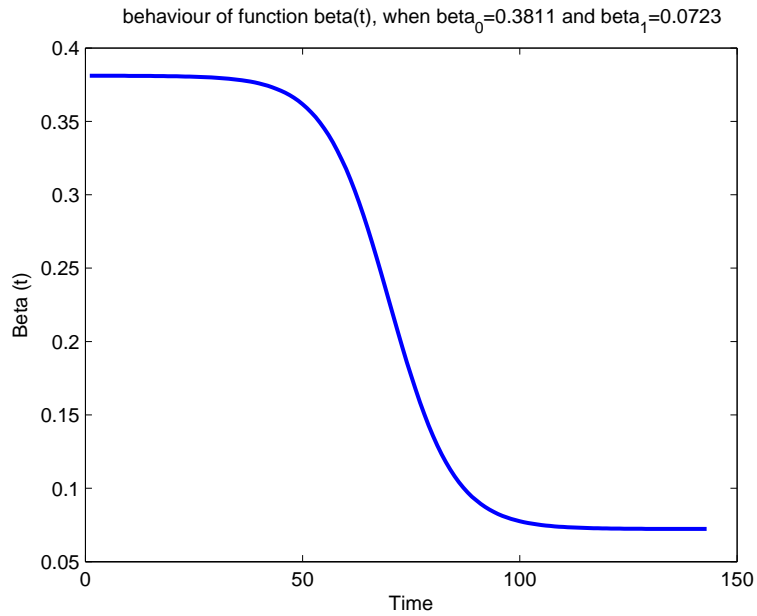


Figure 4.2: The behavior of the transmission rate function $\beta(t)$

In [1], the transmission rate has been formulated differently than in 4.2. In [1], the transmission rate was assumed to decrease gradually from β_0 to β_1 too, but $\beta(t)$ was equal to β_0 from the beginning of the onset to the intervention time τ and then decreases to β_1 . For this thesis case, the decrease starts before reaching the intervention time (check figure 4.2) because some other preliminary measures have been taken before the confirmation of Ebola virus on 9th May 1995 [8].

Before this confirmation, Ebola was treated as typhoid-associated abdominal perforation [8]. Quarantine (of exposed individuals) and isolation (of infectious individuals) were two of the most commonly used control measures before the identification

of the disease as Ebola [8]. Other measures are creation of new laboratories, education and doctors were alerted to identify this disease. It is on 9th May, when tests confirmed an Ebola virus that international teams intervened and made a team of surveillance. Other details will be discussed in the following point concerning data.

4.2.2 Ebola Data

The 1995 Kikwit Ebola outbreak in the democratic Republic of Congo is one of the most well studied epidemics due to the intervention of international teams and centers for disease control and prevention. This outbreak began in the Bandundu region, especially in Kikwit town located on the banks of the Kwilu River, situated at about 500 km south-east of the DRC capital, Kinshasa. The first identified case-patient was a 42 years old male charcoal worker and farmer who became ill on 6 January 1995 and died on 13 January 1995 [8]. Once exposed, individuals go through a latent period of 6.3 days after which they become infectious for a period between 3.5 and 10.7 days [1].

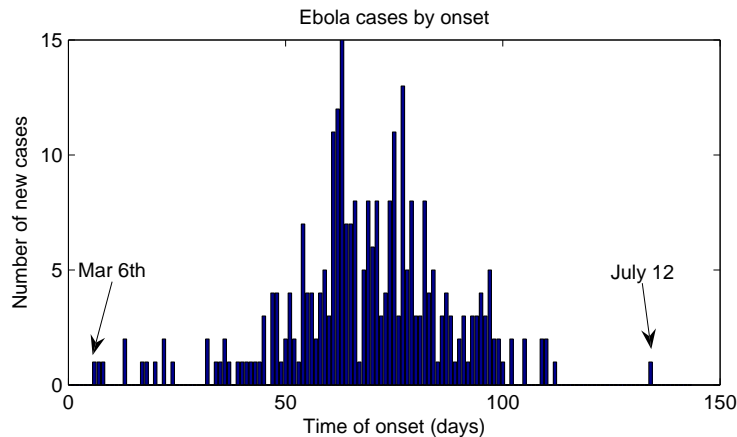


Figure 4.3: Onset data from the 1995 Ebola outbreak in the Democratic Republic of Congo (Zaire) from March 1 (corresponding to day 1 on the x -axis) to July 21

The application of this thesis is based on data from the 1995 Ebola outbreaks in the Democratic Republic of Congo. These data were collected from surveillance case report forms by Khan [8]. The data consist of two time series represented in figure 4.3 and figure 4.4 recorded from March 1 to July 21 (143 days), namely Ebola cases by onset and Ebola death cases respectively. The figure 4.3 accounts a total of 291 cases and the figure 4.4 accounting a total of 236 deaths.

As said above, the first case became ill on January 6, 1995, the last case died on

July 16, 1995 and a total of 316 cases were identified resulting in a rate of 81% fatality [13]. The Ebola virus was not identified as the causative agent until May 9 [1]. At that time an international team implemented a control plan that involved active surveillance (identification of cases), education programs for infected people and their relatives, and the use of protective clothing. Between May 10 and 19, nine international medical teams, including the World Health Organization (WHO), doctors without borders and the Centers for the diseases Control and Prevention (CDC) arrived bringing suppliers and knowledge [15].

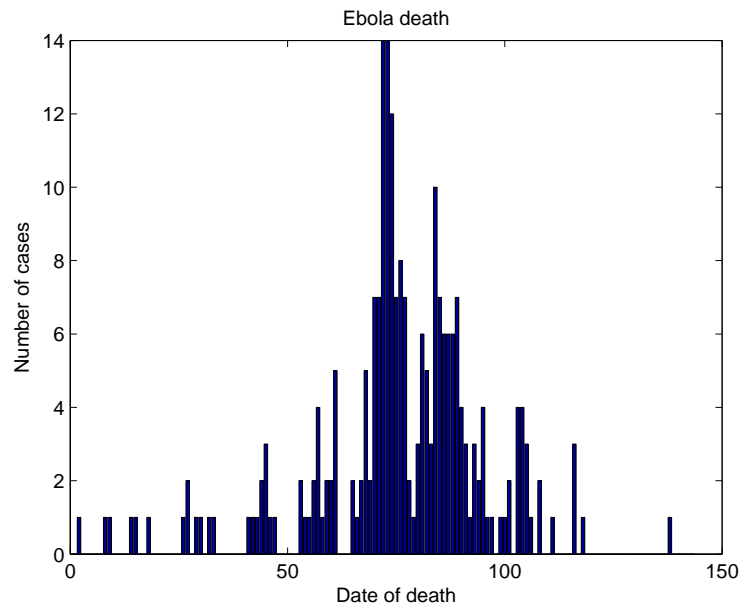


Figure 4.4: Death data from the 1995 Ebola outbreak in the Democratic Republic of Congo (Zaïre) from March 1 (corresponding to day 1 on the x -axis) to July 21

By May 11, 164 reported cases of Ebola were identified, of which 134 individuals had died, where 63 of them were health care workers (38 %) and 47 died [15]. The epidemics lasted about 200 days (from the first case on 6th January to July 24) but the recorded data starting time and evolution of the epidemic prior to March 1st. Otherwise, some data are not reported, because from the total number of 316 identified cases, 25 cases were not reported and from 256 of death cases, 20 cases were not reported in the given time series. Overall, 20-25% of the victims were health care workers [8, 15].

For all the hospital staffs infected, physicians had the highest rate of infection at 31% (4/13), followed by technicians/room attendants at 11% (7/62) and nurses at 10% (22/212) [15]. This is an important data set because the diagnoses were verified by laboratory tests. The next point will be based on estimating immigration rates

between compartments.

4.3 Parameters estimation

First, some initial values were needed in order to estimate parameters. The purpose is to find a group of parameters that can fill perfectly all the features required in the model described by 4.1. Among those features there are, to make the model more accurate to fit the observed data and improve the convergence of MCMC algorithms used in this epidemic model. At the time of the 1995 Ebola outbreak, the whole population of Congo was 60,000,000. The region of Bandundu had an effective population size of $N = 5,364,500$ [1, 13].

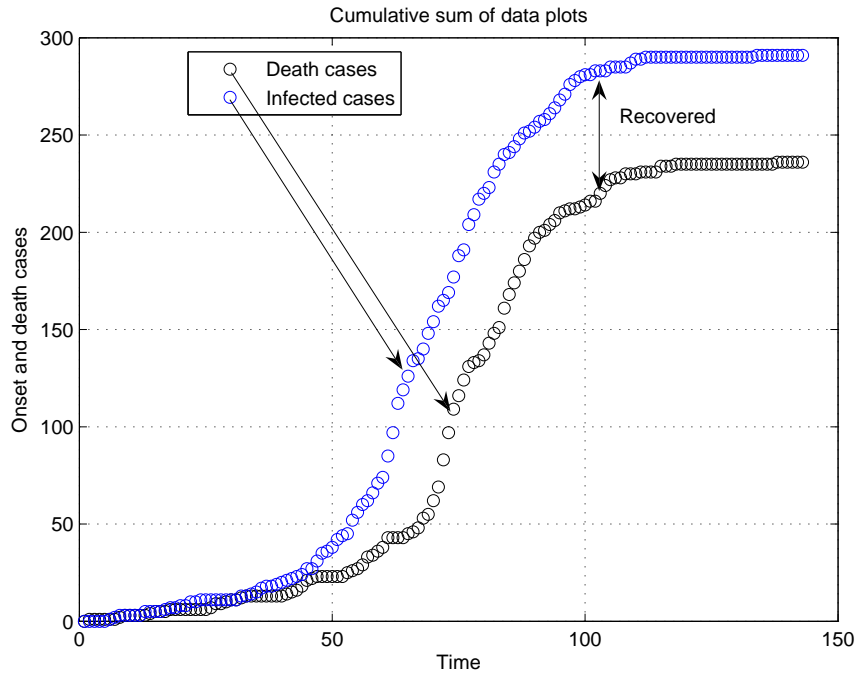


Figure 4.5: Cumulative sum of infected and death cases for Ebola epidemics in Zaïre 1995

The population consists of N individuals out of which E_0 are initially infected and they are able to have close contacts, i.e. contacts that result in infection, with other individuals of the population. The remaining $N - E_0$ individuals are assumed to be initially susceptible and can be potentially infected by the E_0 initial infective. The reported data are (t_i, y_i) and (t_i, z_i) , $i = 1, \dots, n$ where t_i denotes the i th reporting time of a new case and y_i, z_i denote the cumulative number of infectious cases and death cases respectively from the beginning of the outbreak to time t_i .

These cumulative data are plotted in figure 4.5 and the area between the two data plots is the cumulative number of individuals recovered (R_i) where $R_i = y_i - z_i$ for every t_i .

Assuming that there were one exposed individual $E_0 = 1$, means that $S_0 = 5,364,499$, $I_0 = 0$, $R_0 = 0$ and $C_0 = 0$ and the intervention started on 9th May, 2005 which corresponds to $\tau = 70$ from 1st March at the beginning of data observation. From [1, 13], initial parameter values are set following these constraints $0 < \beta < 1$; $0 < q < 100$; $1 < 1/k < 21$ and $3.5 < 1/\gamma < 10.7$. Then, model parameters to be estimated are $\theta = (\beta_0, \beta_1, k, q, \gamma)$.

4.3.1 Chowell et al. mistake and a new model version to estimate parameters

The mistake made by Chowell et al. in [1] is to use the onset data to estimate model parameters. The model in equations 4.1 and the figure 4.1 suggest that data do not affect all model parameters. For example the parameter γ is not affected by data, because infected are counted immediately after the latent period, i.e. from the class E. Another mistake made is the estimation of the basic reproduction number (R_0) using the estimates got using the model in equations 4.1 and onset data. In this Ebola model, the parameter γ does not depend on the data, therefore the basic reproduction number (R_0) calculated in 4.5 can not be estimated. In MCMC results the uncertainty about the parameter γ is not affected by latent data ($E(t)$), because the component of the likelihood involving this parameter depends only on $I(t)$.

To avoid the same mistake made by Chowell et al. in [1], we make another model by splitting the compartment R (removed) in two parts. The two compartments are the recovered (R) and the died (D). And then, because we have two sets of data (onset data and death data), we will use both two to estimate unknown parameters.

Using the diagram in figure 4.6 where the system of equations 4.3 is expressing the new model.

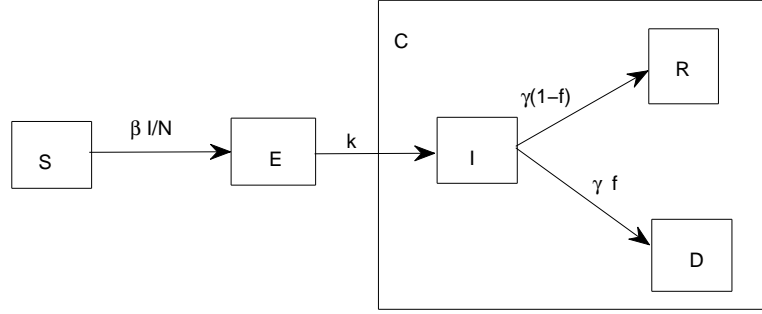


Figure 4.6: The diagram of transmission for the new model

$$\begin{aligned}
 \frac{dS(t)}{dt} &= -\beta(t) \frac{S(t)I(t)}{N} \\
 \frac{dE(t)}{dt} &= \beta(t) \frac{S(t)I(t)}{N} - kE(t) \\
 \frac{dI(t)}{dt} &= kE(t) - \gamma I(t) \\
 \frac{dR(t)}{dt} &= \gamma(1-f)I(t) \\
 \frac{dD(t)}{dt} &= \gamma f I(t) \\
 \frac{dC(t)}{dt} &= kE(t)
 \end{aligned} \tag{4.3}$$

The transmission rate $\beta(t)$ remains the same as 4.2. For this model, parameters to be estimated are $\theta = (\beta_0, \beta_1, k, q, \gamma, f)$, where f is the death probability and $(1-f)$, the recovery probability. These parameters were estimated using the onset and death data by least-squares fit and observed variables are C and D in 4.3. This consists of minimizing the errors sum of squares. Matlab (fminsearch) software was used and the following parameters gathered in table 4.2 were found.

Table 4.2: Estimated Ebola Epidemic Model Parameters using onset and death data by least square method

Parameters	Definition	Initial values	Estimates
β_0	Rate before intervention ($days^{-1}$)	0.394	0.3011
β_1	Rate after intervention ($days^{-1}$)	0.055	0.0268
q	Rate from β_0 to β_1	0.165	0.1997
$1/k$	Rate of incubation (days)	1.82	1.8604
$1/\gamma$	Rate to Removed (days)	5.02	7.8285
f	Probability of death ($\times 100\%$)	0.81	0.8243

According to the table 4.2, the transmission rate before intervention must be decreased to the transmission rate after intervention at the rate of 16.2% and the incubation period is almost 2 days. Another result is that, an individual already infectious will die or recover after 6 days but the probability of dying is 80%.

When a mathematical model is used to study disease control, it is common to consider certain quantities (derived from the model) that can provide information about the effect of control measures on disease prevalence. These quantities include the (control) reproductive number R_0 , the quality of the disease-free equilibrium. Below is the calculation of the steady point in which parameters will be substituted to qualify the disease-free equilibrium.

4.4 Calculus of the steady points

For the easy calculation of the steady point, it could be better if variables are changed in the following manner: Introduction of new variables by assuming $S = Ns$, $E = Ne$, $I = Ni$, $R = Nr$, $D = Nd$ and $C = Nc$; then the derivatives become

$$\frac{ds}{dt} = \frac{1}{N} \frac{dS}{dt}; \frac{de}{dt} = \frac{1}{N} \frac{dE}{dt}; \frac{di}{dt} = \frac{1}{N} \frac{dI}{dt}; \frac{dr}{dt} = \frac{1}{N} \frac{dR}{dt}; \frac{dd}{dt} = \frac{1}{N} \frac{dD}{dt}$$

and

$$\frac{dc}{dt} = \frac{1}{N} \frac{dC}{dt}.$$

Replacing these new variables in 4.3 , the system becomes:

$$\begin{aligned} N \frac{ds(t)}{dt} &= -\beta(t) \frac{Ns(t)Ni(t)}{N} \\ N \frac{de(t)}{dt} &= \beta(t) \frac{Ns(t)Ni(t)}{N} - kNe(t) \\ N \frac{di(t)}{dt} &= kNe(t) - \gamma Ni(t) \\ N \frac{dr(t)}{dt} &= \gamma N(1-f)i(t) \\ N \frac{dd(t)}{dt} &= \gamma f Ni(t) \\ N \frac{dc(t)}{dt} &= kNe(t) \end{aligned} \tag{4.4}$$

after simplification the system 4.4 changes to

$$\begin{aligned}
\frac{ds(t)}{dt} &= -\beta s(t)i(t) \\
\frac{de(t)}{dt} &= \beta s(t)i(t) - ke(t) \\
\frac{di(t)}{dt} &= ke(t) - \gamma i(t) \\
\frac{dr(t)}{dt} &= \gamma(1-f)i(t) \\
\frac{dd(t)}{dt} &= \gamma fi(t) \\
\frac{dc(t)}{dt} &= ke(t)
\end{aligned} \tag{4.5}$$

A suitable domain is

$$D = \{(s, e, i, r, d) \in [0, 1]^4 : s \geq 0, e \geq 0, i \geq 0, r \geq 0, s + e + i + r + d \leq 1\}$$

and $(s(0), e(0), i(0), r(0), d(0), c(0)) \in D$.

Removing the equations of r , d and c from 4.5, the disease-free equilibrium can be found by solving the system

$$\begin{aligned}
\frac{ds(t)}{dt} &= -\beta s(t)i(t) = 0 \\
\frac{de(t)}{dt} &= \beta s(t)i(t) - ke(t) = 0 \\
\frac{di(t)}{dt} &= ke(t) - \gamma i(t) = 0
\end{aligned} \tag{4.6}$$

The first steady point where $i(t) = e(t) = 0$ and $s(t) = 1$ corresponds to the situation with no infection present ($\beta \approx 0$) and the entire population is susceptible. This point is written $E_0 = (1, 0, 0)$ and is called *disease-free equilibrium* and usually the analysis is centered on determining the stability properties of it. This model does not admit an endemic equilibrium because equality in the above equations will hold if $i = 0$, i.e. Ebola can't be endemic. The disease will disappear in the population when

$$\lim_{t \rightarrow \infty} i(t) = 0, \lim_{t \rightarrow \infty} s(t) = s_\infty$$

and the number that has been infected (final size of the epidemic) is $s_0 - s_\infty$.

To study the stability of the disease free-equilibrium, first find the Jacobian matrix. Assuming that

$$f(t) = \beta s(t)i(t), g(t) = \beta s(t)i(t) - ke(t)$$

and

$$h(t) = ke(t) - \gamma i(t)$$

from 4.6

$$J = \begin{pmatrix} \frac{\partial f}{\partial s} & \frac{\partial f}{\partial e} & \frac{\partial f}{\partial i} \\ \frac{\partial g}{\partial s} & \frac{\partial g}{\partial e} & \frac{\partial g}{\partial i} \\ \frac{\partial h}{\partial s} & \frac{\partial h}{\partial e} & \frac{\partial h}{\partial i} \end{pmatrix}$$

After differentiation, the linearized Jacobian matrix becomes

$$J = \begin{pmatrix} -\beta i & 0 & -\beta s \\ \beta i & -k & \beta s \\ 0 & k & -\gamma \end{pmatrix}$$

Replacing the disease free equilibrium $E_0 = (1, 0, 0)$ in J ,

$$J_0 = \begin{pmatrix} 0 & 0 & -\beta \\ 0 & -k & \beta \\ 0 & k & -\gamma \end{pmatrix}$$

Finding the eigenvalues of the above matrix, the characteristic equation corresponding to J_0 is a third-degree polynomial with a characteristic equation given by

$$-\lambda (\lambda^2 + (k + \gamma)\lambda + (\lambda - \beta)k) = 0$$

and roots are $\lambda_1 = 0$ and $\lambda^2 + (k + \gamma)\lambda + (\lambda - \beta)k = 0$. The solution of this last quadratic equation is

$$\lambda_{1,2} = \frac{-(k + \gamma) \pm \sqrt{(k + \gamma)^2 - 4k(\gamma - \beta)}}{2}$$

Estimating the initial eigenvalues, take $\beta = \beta_0$ in table 4.2; $\lambda_2 = 0.1216 \text{ days}^{-1}$ and $\lambda_3 = -0.8347 \text{ days}^{-1}$. This is shown by the figure 4.7. Among these three eigenvalues found due to the disease-free equilibrium E_0 , the first is zero, the second is negative and the third is positive.

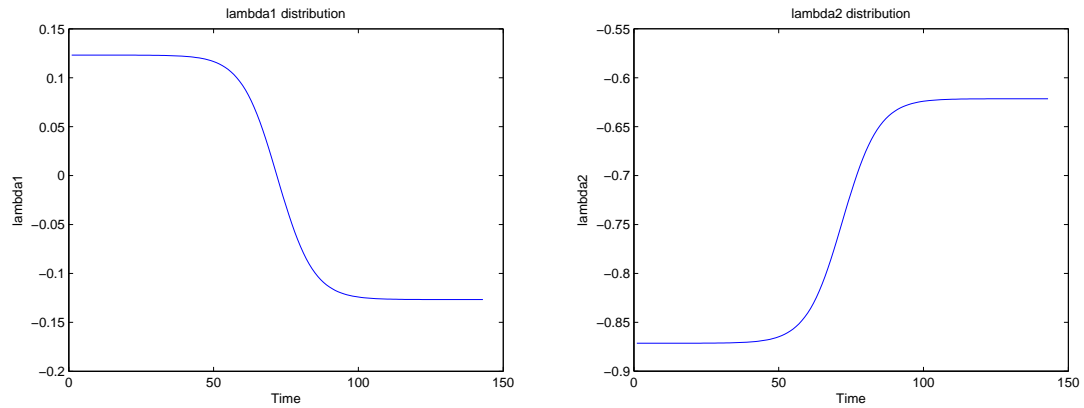


Figure 4.7: Distribution of eigenvalues taking β as a vector

According to the eigenvalues signs, the point E_0 is unstable, which means that the disease is able to invade the susceptible population if no means are taken. In other words, the endemic point if it existed could have been stable and the disease is epidemic. This is one of the ways used to know the behavior of the disease, the second one is the study of the basic reproduction number which describes the asymptotic behavior of the system considered starting from initial conditions.

4.5 The Ebola Basic Reproduction Number (R_0)

The important quantity that must be determined is the basic reproduction ratio (R_0), as this provides the key to transmission dynamics, the ease by which major epidemics may be prevented and prospects for the eradication of an infection. From many definitions that have been established by different authors, the following has been chosen here.

Definition 8. *The basic reproduction number R_0 at the beginning of an epidemic is the average number of persons a single sick individual ("patient zero") will infect [11].*

The basic reproduction number R_0 is called a threshold parameter since the value of R_0 determines whether or not an epidemic can occur. If $R_0 < 1$ then the disease will eventually die out since each person is replaced by less than one other infected person i.e. in an infinite population only a finite number of individuals will ultimately become infected. If at any time, R_0 gets smaller than 1, the disease eventually disappears from the population, and the disease-free equilibrium E_0 is

globally stable in the feasible region (see the section 4.4), because on average, each infected person cannot ensure transmission of the infectious agent to one susceptible person. Mathematically,

$$R_0 \leq 1 \Rightarrow \lim_{t \rightarrow \infty} (s(t), e(t), i(t), r(t), d(t)) = (1, 0, 0, 0, 0) = E_0$$

Conversely, the disease spreads if $R_0 > 1$ [11]. There is a positive probability that an infinitely large number of individuals will contact the disease in question. If R_0 gets greater than 1, a unique endemic equilibrium (E^*) is globally asymptotically stable in the interior of the feasible region and the disease will persist at the endemic equilibrium if it is initially present, i.e. the epidemic builds up [11]. Mathematically,

$$R_0 \geq 1 \Rightarrow \lim_{t \rightarrow \infty} (s(t), e(t), i(t), r(t), d(t)) = E^*$$

On the other hand, if R_0 equals to 1, the disease remains endemic as one infectious person transmits the infectious agent to one susceptible person on average. Thus to eradicate a disease we need to reduce R_0 to be less than 1. In this Ebola model, [1] argues that the mass action used is $(\beta(t)SI/N)$, which makes the model parameters independent of N . The basic reproduction number (R_0) can be estimated in the following way:

Consider the disease transmission model consisting of initial conditions and the following of system of equations:

$$\dot{x}_i = f_i(x) = \mathcal{F}_i(x) - \mathcal{V}_i, i = 1, \dots, 5$$

where $\dot{x}_i = f_i(x)$ represents the system 4.5 and

$$\mathcal{F} = \begin{pmatrix} \beta_0 s i \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \mathcal{V} = \begin{pmatrix} k e \\ -k e + \gamma i \\ \beta_0 s i \\ -\gamma(1-f)i \\ -\gamma f i \\ -k e \end{pmatrix}$$

Let E_0 denote the disease-free equilibrium of (4.5) and define $DF(E_0)$ and $DV(E_0)$ to be jacobian matrices of F and V at the point E_0 respectively as follows

$$DF(E_0) = \begin{pmatrix} 0 & \beta_0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and

$$DV(E_0) = \begin{pmatrix} k & 0 & 0 \\ -k & \gamma & 0 \\ 0 & \beta_0 & 0 \end{pmatrix}$$

Consider F and V being two 2×2 matrices consisting of the first rows and columns of $DF(E_0)$ and $DV(E_0)$, respectively. The basic reproduction number is given by the largest eigenvalue of FV^{-1} :

$$FV^{-1} = \begin{pmatrix} 0 & \beta_0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} k & 0 \\ -k & \gamma \end{pmatrix}^{-1} = \frac{1}{k\gamma} \begin{pmatrix} 0 & \beta_0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \gamma & 0 \\ k & k \end{pmatrix}$$

which implies that

$$FV^{-1} = \frac{1}{k\gamma} \begin{pmatrix} \beta_0 k & \beta_0 k \\ 0 & 0 \end{pmatrix}$$

and then;

$$FV^{-1} = \begin{pmatrix} \frac{\beta_0}{\gamma} & \frac{\beta_0}{\gamma} \\ 0 & 0 \end{pmatrix}$$

There are two eigenvalues, $\lambda_1 = 0$ and $\lambda_2 = \frac{\beta_0}{\gamma}$. The largest eigenvalue is:

$$\rho(FV^{-1}) = R_0 = \frac{\beta_0}{\gamma} \tag{4.7}$$

By replacing in equation 4.7 the estimated parameters from table 4.2, and its distribution (histogram) see the figure [4.12], $R_0 = \frac{\beta_0}{\gamma} = \frac{0.3011}{1/7.8243} = 2.3559$. Because R_0 is greater than 1, this disease is capable of invading susceptible population if effective measures are not taken. Anyway, the estimated value of R_0 is relatively low, which means that the epidemic can be controlled and the disease will die out quickly.

Another interpretation is that the number of individuals infected by a single infected during his or her infectious period is greater than one. An interesting point here is that R_0 reflects not only the behavior of the disease but may also reflect features of the endemic equilibrium. Therefore, comparatively with what has been found in studying the disease-free equilibrium, if $R_0 > 1$, the disease-free equilibrium is unstable and the endemic equilibrium is locally asymptotically stable [4]. The theorem explaining this is as follows:

Theorem: *Assume that $R_0 < 1$, then the trivial stationary point (E_0) of the studied epidemic model is locally asymptotically stable [4].*

The basic Reproduction number R_0 is calculated to know the spread of the disease and implement control measures. For this case of Ebola some measures have been taken to stop its spread such as isolation and quarantine. Because no special treatment for Ebola, the only control measures that were most successful were limiting contact between people in hospitals and decreasing the number of contact between people inside and outside the hospital [8].

The basic reproduction number predicts the behavior of the spread of the disease but it can't predict the number of people who will be infected and when the disease will die out. The good way is to find the model solution and check if it fits the data used to estimate parameters. The following point will contain the model solution with interpretations and discussions on found results.

4.6 The model solution and discussions

The full Ebola model is composed by variables (compartments) and parameters or constants that define the form and content of the relationship. After estimating the parameters, the model represented by a system of six differential equations has been solved using Matlab software. The validity of the model can be tested by feeding data into the model, solving it through calculations, and finally comparing the results with actual observations.

This model will show how the variables (compartments) change over a period of time or the trend of the variables. The behavior of each compartment has been briefly discussed accordingly. The epidemic is initiated at time $t = 0$ (1st March) by a small number of individuals while the rest of the population is initially susceptible.

As soon as susceptible candidates are infected, they immigrate to the compartment of Exposed. This transfer of candidates from susceptible to Exposed explains the decreasing of susceptible people. It will be the same for candidates from exposed to infected. The full explanation is coming next.

4.6.1 Model Solution

Every time, the susceptible compartment is losing candidates when they are infected. Firstly, the number of susceptible population is decreasing exponentially from the beginning at time $t = 0$ (1st March) to $t = 100$ (7th June). This situation is explained by figure 4.8.

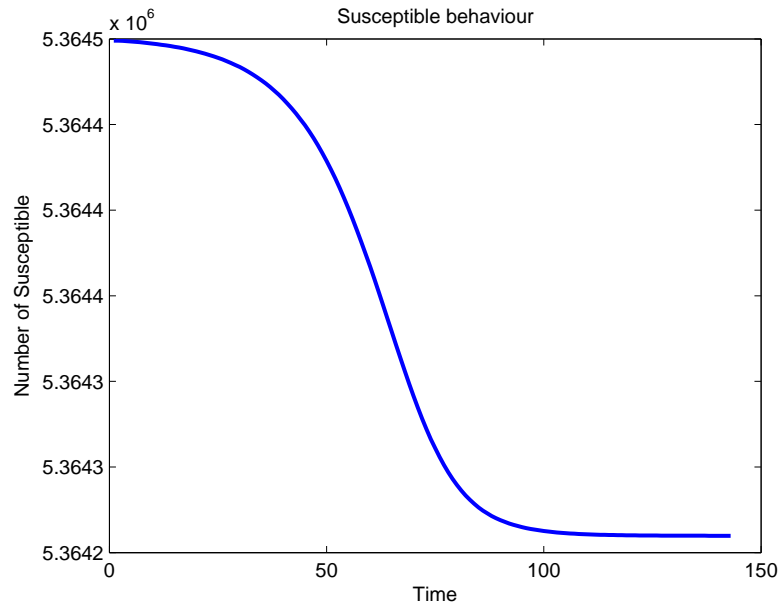


Figure 4.8: Susceptible behavior in Ebola Epidemic model

This decreasing is due to the high infection rate β , because when people are infected they leave the susceptible compartment to move to exposed class. On the other hand, due to this immigration, the exposed compartment is gaining individuals to be passed to the infected class and finally die or recover in R or D compartment.

Secondly, for the case of exposed and infected, when the number of susceptible people decreases, the number of exposed and infected increases from the beginning of the epidemic outbreak. It is seen that, when the intervention starts at $t = 70$, the number of exposed decreases to vanish at $t = 143$. In other words, when the

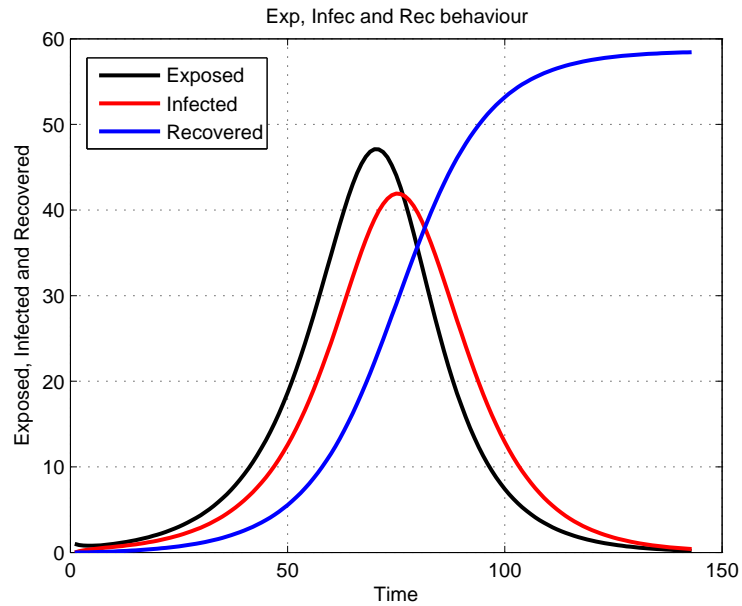


Figure 4.9: Exposed, Infected and Recovered behavior in Ebola Epidemic model

infection rate is constant ($\beta = \beta_0$), the number of exposed and infected is increasing and when $\beta \rightarrow \beta_1$, the number decreases.

At $\tau \geq 70$, the transmission rate β decreases from β_0 to β_1 , which means that the number of exposed and infected decreases exponentially while that for recovered increasing with maximum 58.43 recovered. Remind that from the data, the number of recovered is 59 (315-256). This phenomena is explained by figure 4.9. It is seen that, the infectious periods of different infected cases are assumed to be independent and identically distributed according to the distribution of a random variable I .

Finally, assume that $t \rightarrow \infty$ the infection rate $\beta \rightarrow \beta_1$, and β_1 must be near zero, which means that there is no infection at all. This situation will correspond with the disease-free equilibrium E_0 . At the end of its infectious period the individual is removed (dies or recovers) and plays no further role in the epidemic spread. The epidemic ceases when there are no infectious present in the population. The model solution is found but the problem remains to know if it fits the observed data. The following point will be focussed on checking whether the model solution answers fitness questions.

4.6.2 Evaluating Model Fit

Before studying the parameter uncertainties from sample estimates, it would be better to decide just how well the model fits the data. A first approach is to determine the coefficient of fitness starting by the following formulae:

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (4.8)$$

where Y_i are observed data and \bar{Y} the mean value. The formula 4.8 represents the variance and its numerator is called the *Total Sum of Squares* (TSS). Another quantity to determine is the *Residual Sum of Squares* (SSE). It is the discrepancy between the data and the estimated model.

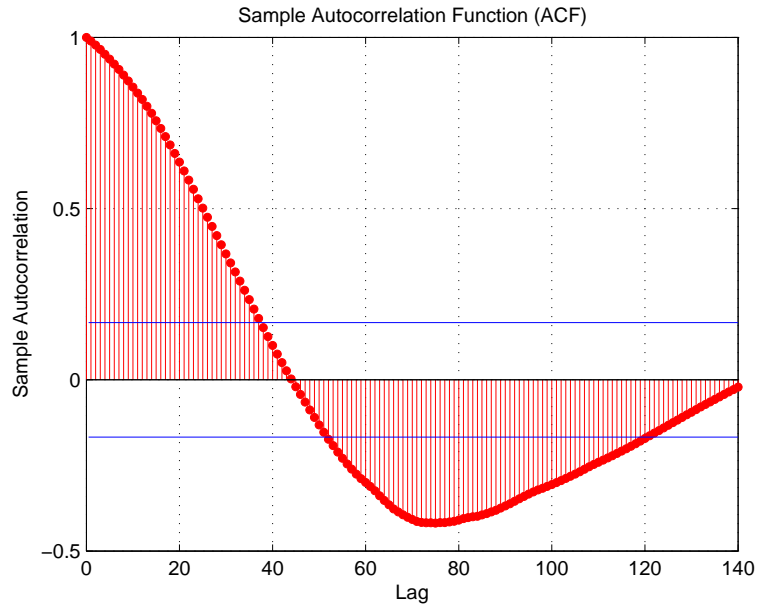


Figure 4.10: Autocorrelation plot showing the behavior of residuals taking a number of 140 lags

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (4.9)$$

where $\hat{Y} = f(x_i, \hat{\theta})$. This formula has been used to estimate model parameters and if the discrepancy is small, the estimation is qualified as good. The figure 4.10 shows how residuals vary according to their discrepancy. The variation of discrepancy is between $[-0.4, 1]$.

The difference between these two is the *Regression Sum of Squares* (RSS), it can be presented as:

$$RSS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \quad (4.10)$$

Some basic algebra reveals that, $RSS + SSE = TSS$. A measure of model fit can be constructed by taking the ratio of the explained variance to the total variance:

$$R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} \quad (4.11)$$

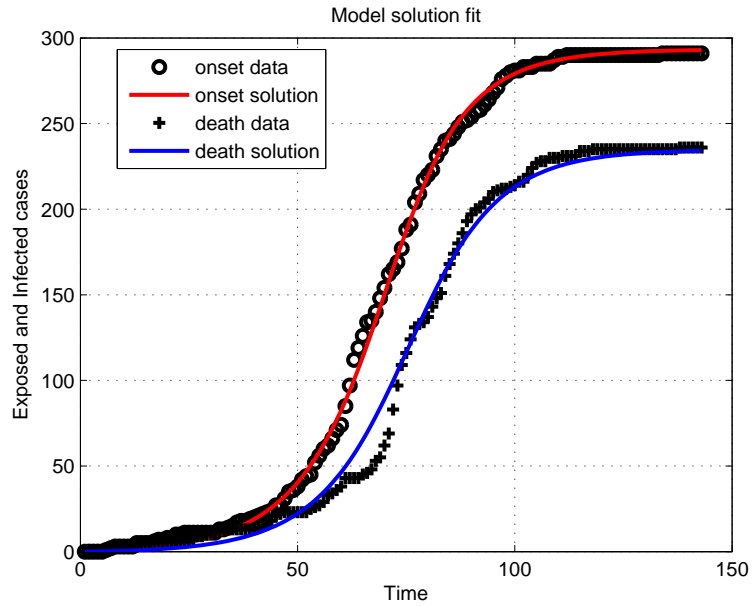


Figure 4.11: Model fitting data plots: The solution using the onset and death data cases

This measure ranges from 0 to 1. For a poor fitting model, the error (SSE) will be large (possibly equal to the TSS), making RSS small. If $\hat{Y} = \bar{Y}$, then the relationship between variables does not exist and in this case, RSS would be 0 as would be R^2 . For a perfect model on the other hand, $RSS = TSS$, so $R^2 = 1$. In this thesis case, the R^2 when using onset data was 0.9895 (98.95%), indicating a good model fit. Using death data; $R^2 = 93.6\%$.

The model seems to fit the data as shown by figure 4.11. Graphical measures are more useful than numerical measures because they allow to view the entire set of data at once and show relationship between the model and the data.

Even though it is seen that the model fits the available data perfectly, it may provide wrong answers to the question of interest than an imperfectly fitting model. This is due to the parameter estimation uncertainties and the use of MCMC in studying the model will improve its acceptability. With MCMC it is possible to examine the distribution of unknown parameters in nonlinear models which is not the approach used in the point 4.3.

4.7 MCMC Results and Interpretation

The Metropolis-Hastings algorithm discussed in subsection 3.3 will be applied to the Ebola data set. It will make chains of unknown parameters (θ) in which each value can depend on the previous value (Markov chain) in the chain. The MCMC produced is a matrix of number of simulations \times number of parameters distributed. For this thesis application's results, the number of parameters is 6 and number of simulation is 1000000, i.e. the matrix will be of size 1000000×6 .

With the help of this matrix, some statistical analysis of the model and parameters can be done. Inidentifiability of model parameters, which results from the parameter correlation, nonlinearity of the model and from the characteristics of the data is also revealed by the MCMC chain.

4.7.1 MCMC parameters estimation

As said previously, the Markov Chain Monte Carlo (MCMC) method consists of sampling from probability distributions based on constructing Markov chains converging to the posterior distribution. It means there can be a simulation of the entire joint posterior distribution of the unknown quantities and obtain simulation-based estimates of posterior parameters that are of interest. After getting the MCMC matrix chain, the posterior mean, posterior standard deviation and posterior median have been reported in table 4.3 as well as their 95% confidence intervals.

Referring to table 4.3 and table 4.2, the posterior means are relatively well in agreement with the least square estimates obtained from the complete data (onset and death data). As can be expected, the posterior standard deviations show that the level of uncertainty will increase since all parameters are effected by data. Expect

Table 4.3: Posterior mean and posterior standard deviation of the estimated parameters of Ebola model. In parenthesis there are the nominal 95% confidence intervals

Parameters	Posterior mean	Posterior STD	Post. median
β_0	0.2856 (0.2849 - 0.2860)	0.0100 (0.0087 - 0.0100)	0.2852
β_1	0.0334 (0.0330 - 0.0335)	0.0047 (0.0046 - 0.0047)	0.0334
q	0.1787 (0.1783 - 0.1790)	0.0156 (0.0140 - 0.0161)	0.1777
$1/k$	1.787 (1.8681 - 1.8686)	0.0507 (0.0505 - 0.0508)	1.7114
$1/\gamma$	7.8289 (7.8285 - 7.8291)	0.3307 (0.3301 - 0.3312)	7.8263
f	0.8244 (0.8242 - 0.8244)	0.0044 (0.0043 - 0.0045)	0.8244

γ and f which are affected by a part of data (death data) only.

The mean of the basic reproduction number (R_0) is 2.3622 (2.3618 - 2.3626) with standard deviation 0.1454 (0.1451 - 0.1457) and varies between 1.9531 and 2.9969. The figure 4.12 explains better the distribution of the basic reproduction number.

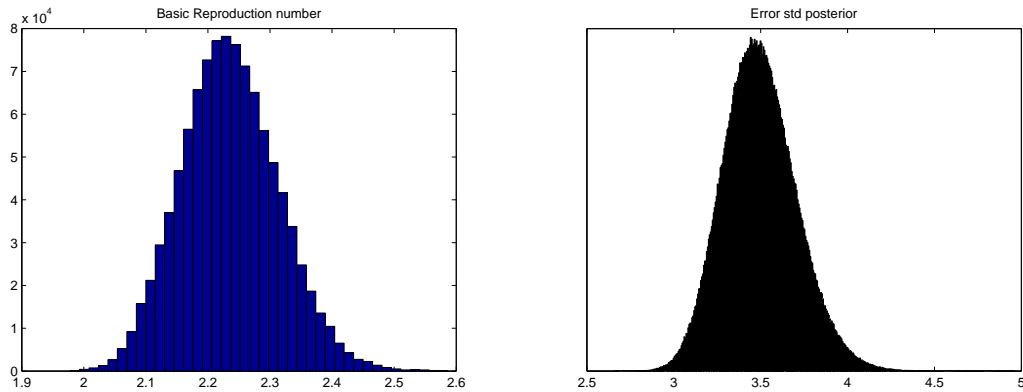


Figure 4.12: Histograms of the distribution for the basic reproduction number (R_0) and standard deviation

The standard deviation plot in figure 4.12 is a gaussian distribution between 2.6 and 4.8.

Apart statistical calculations found in table 4.3, graphical display is an important component of the MCMC process. It provides the visual display of MCMC output for checking the shape and size of the distribution of parameters.

4.7.2 The Chains time-series plot

Through the MCMC figures, it is easy to get information related to correlation , uncertainty, identifiability of parameters, convergence of Markov chain to the target distribution etc. The figure 4.13 represents the plot of the time-series for each parameter.

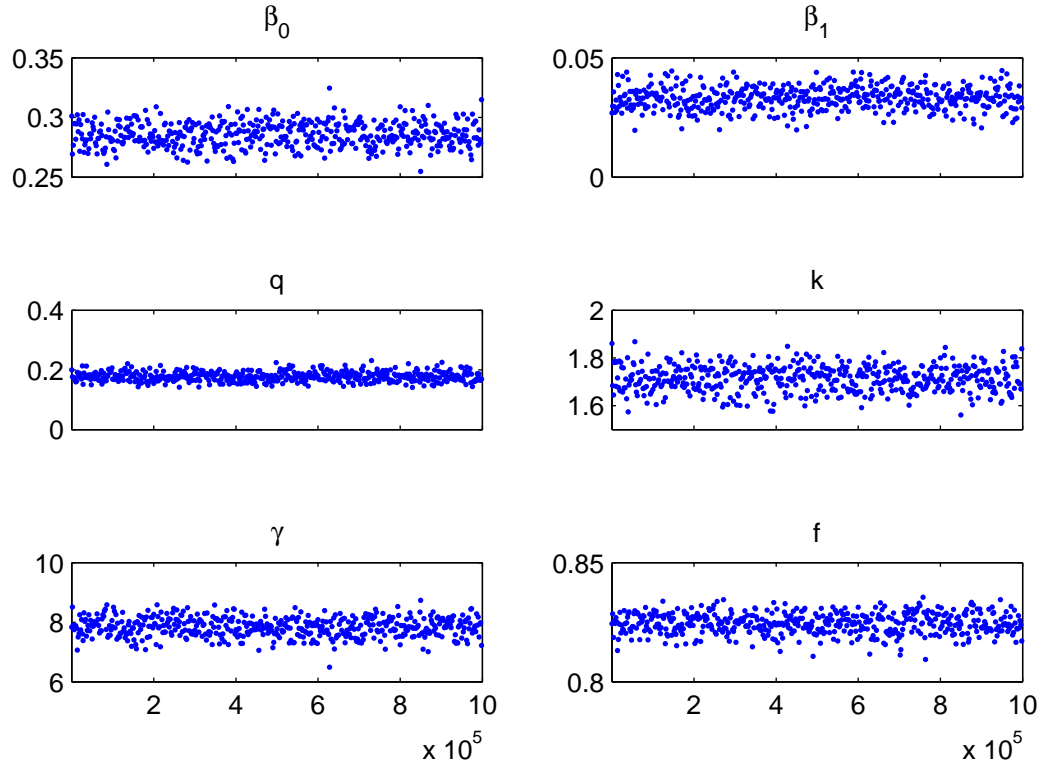


Figure 4.13: Simulated chains of unknown parameters plots

According to the figures 4.13 and 4.14, the transmission rate before intervention (β_0) in figure 4.13 is distributed between 0.2527 and 0.3404 and after intervention (β_1), the transmission rate varies between 0.005 and 0.0558. This explains the benefits of intervention which decreases the transmission rate from β_0 to β_1 . The rate q is a gaussian distribution in the interval $[0.1190, 0.2565]$. The incubation period ($1/k$) is a gaussian distribution between 1.5114 et 1.9231 days and the recovery or death rate (γ) is distributed between 6.4350 and 9.0195 days. The probability of dying f lies between 80.61% and 84.43%.

In the plot 4.15, every parameter is plotted against all other parameters and one-dimensional kernel density estimates and confidence regions (50% and 90%, for ex-

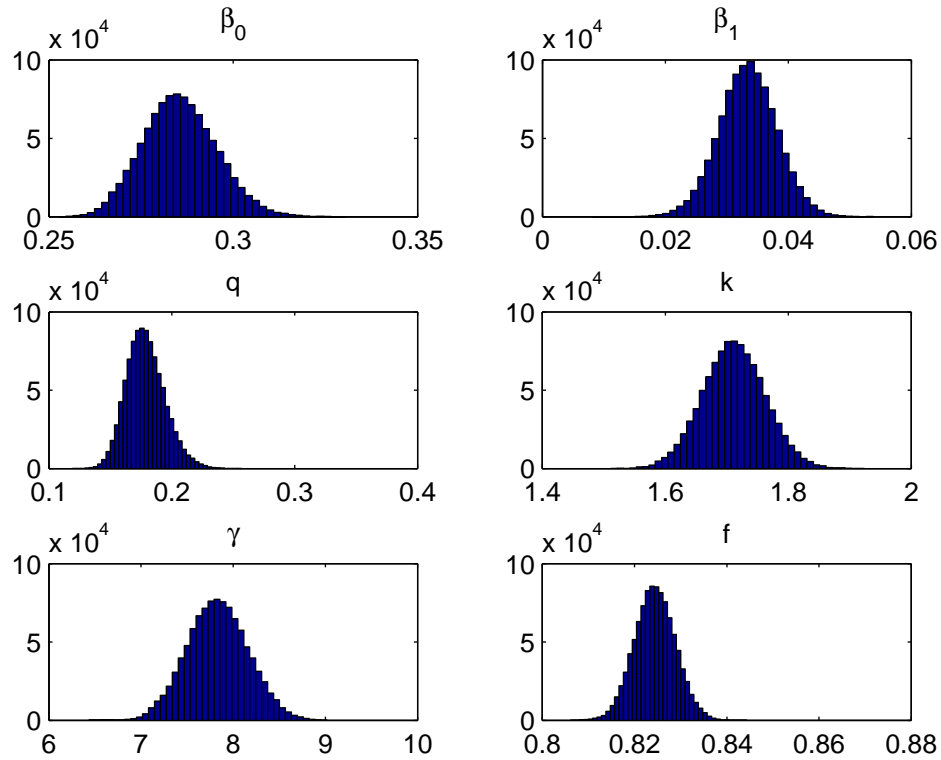


Figure 4.14: Simulated histogram of chains for unknown parameters

ample) based on two-dimensional kernel density estimation was also plotted.

The scatter plot is a graphical representation between two variables and controls the orientation of the plot, it also controls the information shown on the axes. On the figure 4.15, the scatter plot of β_0 and β_1 shows that when β_0 increases/decreases, β_1 also increases/decreases (of course, with some outliers which are not in confidence regions). They are correlated and this relationship is due to the manner of intervention with objectives of reducing β_0 to β_1 . The interpretation is the same for other scatter plots. Other parameters which are somewhat correlated are β_0 with γ and β_0 with f .

Those which are not correlated are β_1 and f , β_1 and γ . The posterior f has no correlation with many of the remaining posteriors. On the figure 4.15, the dots show the points in MCMC chain from which the distribution contour lines (50% and 95% regions of the distribution). The MCMC chains can also be used for prediction purpose by introducing an area where observation data and model solution lie.

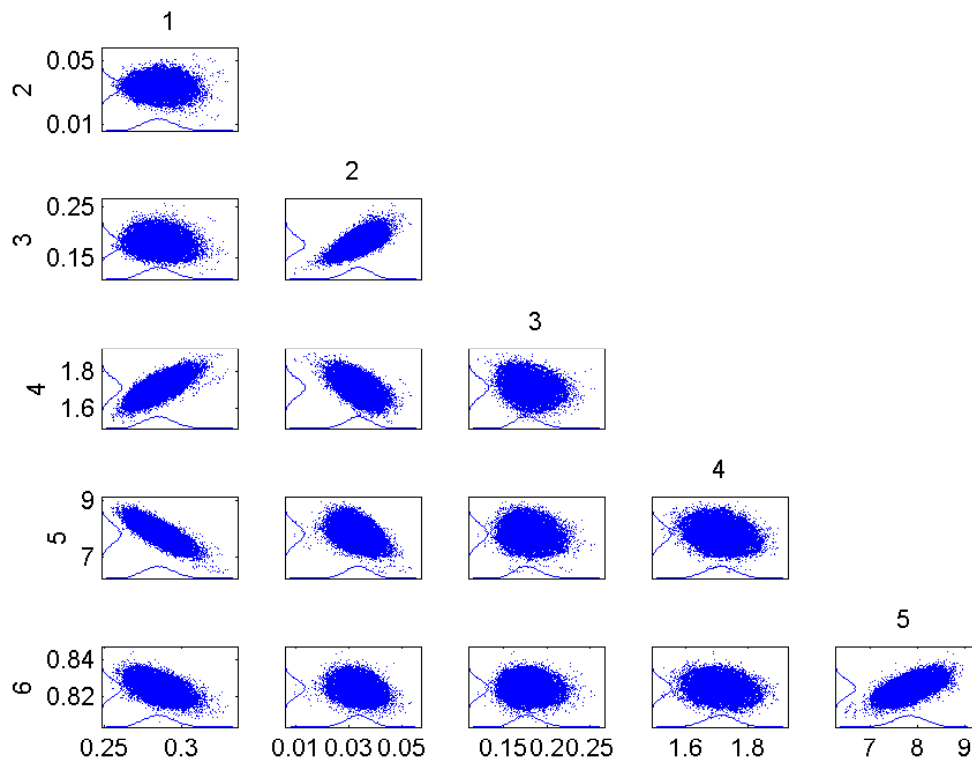


Figure 4.15: The pairwise scatter plots for the unknown parameters $\beta_0, \beta_1, q, 1/k, 1/\gamma$ and f

4.7.3 Predictive MCMC plots

The MCMC approach can also be used to check the accuracy of the model through prediction plots. It is also possible to get a plot representing an area where the model prediction lies with certain probability. In this kind of plot, the area is plotted with darker grey color.

The prediction is done by drawing samples from the posterior distributions via MCMC described in previous section and then uses these samples to draw samples of future values.

From the plot 4.16, the high variance is observed in the Exposed compartment labeled as E . Variances are small in S, I, D and C . The plot 4.17 shows that the

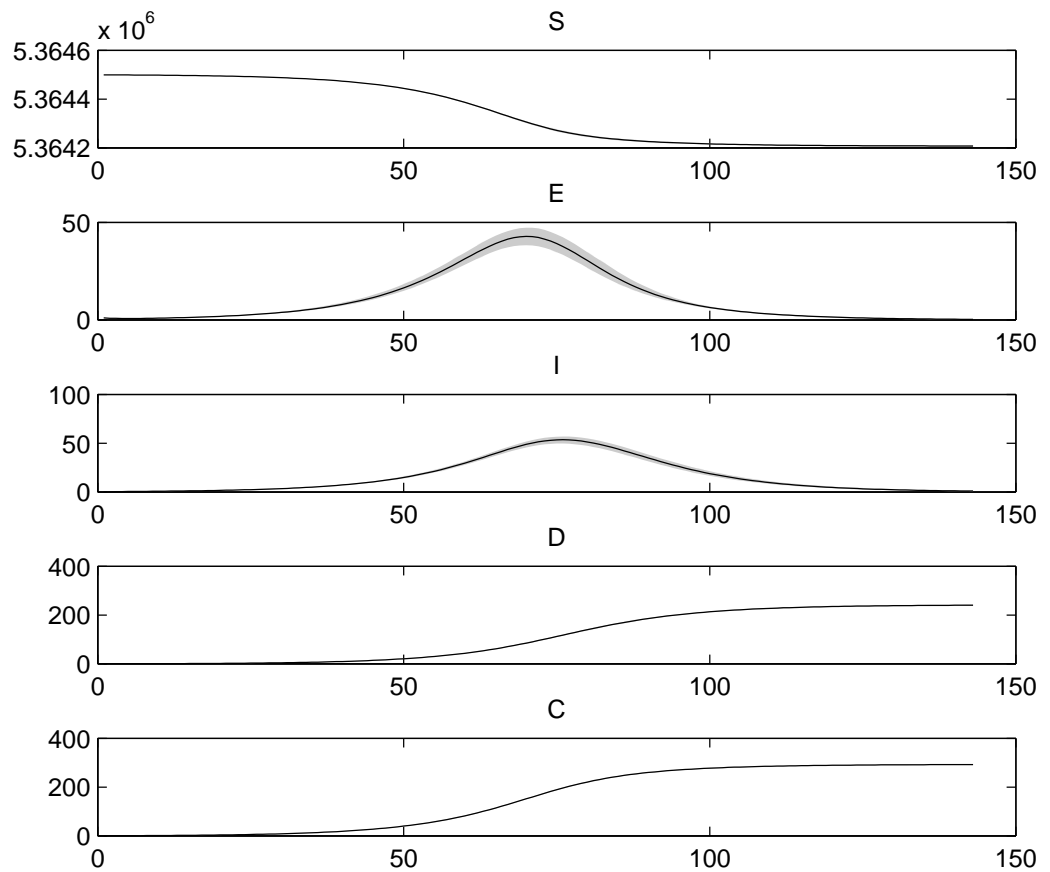


Figure 4.16: The behavior of the solution model prediction

model predicted the number of onset and number of death at 95% posterior limits which are shown by gray area around the model solution.

The variance of the predictive distribution reflects the predictive accuracy of the model. If the variance is large, it will be due to the uncertainties in the model or noise in measurements. It can be seen from figure 4.17 that the model fits the rather noisy data sufficiently well and the fit is almost perfect. The predictive limits of the observations cover the data reasonably.

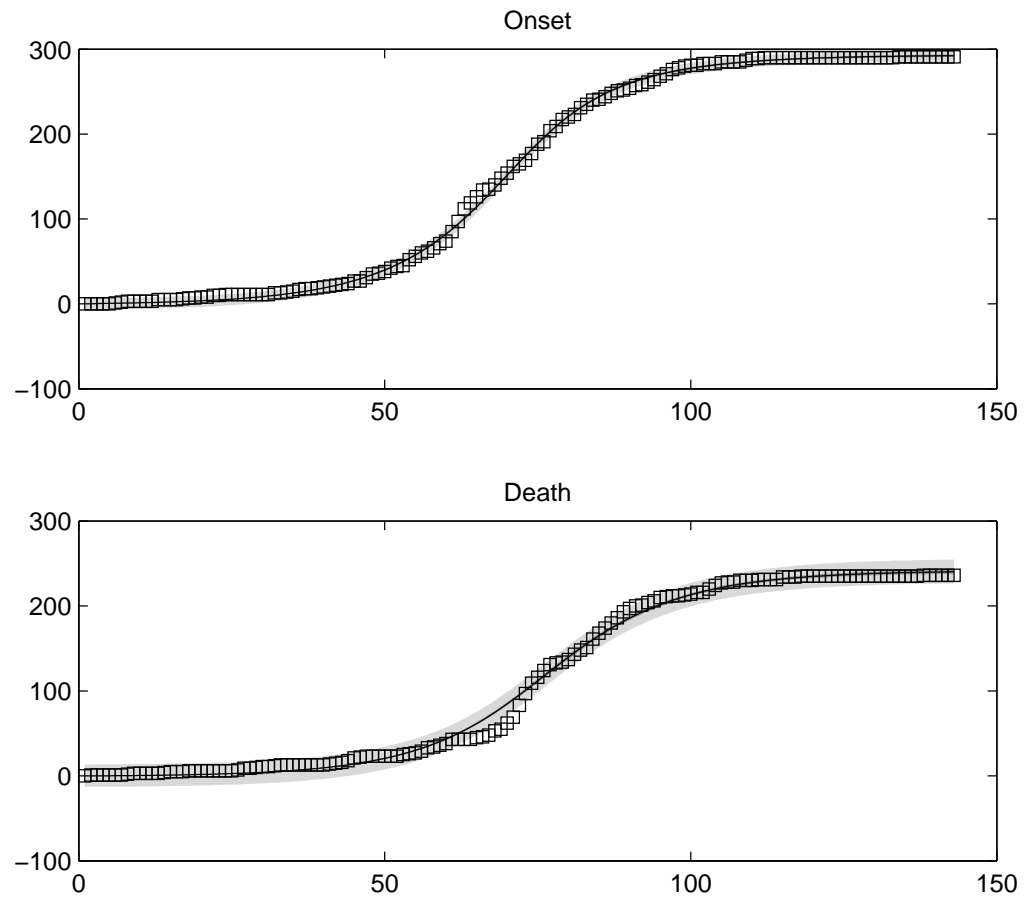


Figure 4.17: The Predictive distribution plot of onset (C) cases in Ebola model

During the plot results, we varied the MCMC chains sample size and it has been seen that figures are not nice when we take a small number of points in the posterior chain. To have nice MCMC figures it is required to consider a high number of simulation (long chains). Long chains also allows us to assess the inaccuracies in the model.

5 Conclusion

The present thesis shows that it is possible and fruitful to develop the model approach by applying mathematical statistical methods. When studying the Ebola model built by Chowell et al. in [1], it has been shown that the parameter γ does not have any influences on the onset data. It was a mistake made by Chowell and his group to use that model with the onset data only in estimating model parameters. To correct the mistake, a new SEIR epidemic model has been constructed by splitting the removed compartment in two (Recovered and Died). Model parameters have been estimated by applying least squares estimation with the aim of fitting SEIR differential equations to both observed daily cases (onset) and daily mortality data. Results show that the model fits onset data at 98.95% and death data at 93.6%.

The basic reproductive number has been calculated and found that it is greater than one, with mean value 2.4, see figure 4.12. This means that the disease was capable to invade susceptible population but the intervention measures were successful in controlling the disease. Another way of predicting the spread of Ebola was the study of disease-free equilibrium. It has been found that it was unstable which means that the Ebola could attack the susceptible population.

The Markov Chain Monte Carlo (MCMC) method has been used to estimate unknown parameters by producing chains of samples. Different tables and plots have been found which helped in interpretation of the model solution and prediction. The predictive distribution reflected the accuracy of the model.

To have good studies on how to model an epidemics, the understanding of its dynamic process is required, which means that the cooperation between epidemiologists and statistician is needed since an intimate knowledge of the epidemiology as well as mathematics methods are essential.

Those interested to further study this Ebola model, can add birth and death process in the diagram 4.6. Because Ebola has no medicine or vaccine, it will be interesting to build a model with vaccination and study the effect of vaccination and the behavior of compartments compared to that one studied above in 4.3. Another approach for the future research can be to take just a part of the data, to study to what extend the outcome of the disease can be predicted during the epidemic.

As we have seen that the intervention is important, we suggest the Democratic Republic of Congo authorities to establish a permanent team to intervene early when there is a disease outbreak in case of waiting for external support. Doctors on outbreak ground have to collect data properly and keep them seriously for future researches.

References

- [1] Chowell G., Hengartner N.W., Castillo-Chavez C., Fenimore P.W., Hyman J.M. 2004. *The basic reproductive number of Ebola and the effects of public health measures: the case of Congo and Uganda*, Journal of Theoretical Biology, Elsevier, Los Alamos, USA.
- [2] Demiris N. 2004. *Bayesian Inference for Stochastic Epidemic Models using Markov Chain Monte Carlo Methods*, thesis for the degree of Doctor of Philosophy, Nottingham.
- [3] Gelman A., Carlin J., Stern H., Rubin D. 1996. *Bayesian Data Analysis*, Second Edition, London Great Britain: Chapman Hall.
- [4] Guardiola J. and Vecchio A. 2003. *The basic reproduction number for infections dynamics models and the global stability of stationary points*, Napoli, Italy.
- [5] Haario, H. 2007. *Statistical Analysis in Modelling: MCMC methods*. Lecture Material, Lappeenranta, Finland: Lappeenranta University of Technology.
- [6] Haario H., Saksman E., Tamminen J. 2001. *An adaptive Metropolis algorithm*. Bernoulli. (Vol. 7), pp. 223-242.
- [7] Hammersley J.M., Handscomb D.C. 1975. *Monte Carlo Methods*, Oxford University Institute of Econometrics and Statistics, London: Methuen and CO LTD.
- [8] Khan, A.S., Tshioko .K., Heymann, D.L., LeGuenzo B., Nabeth P., Kerstiens D.L., Fleerackers Y., Kilmarx P.H., Rodier G.R., Nkulu O., Rollin P.E., Sanchez A., Zaki S.R., Swanepoel R., Tomori O., Nichol S.T., Peters C.J., Muyembe-Tamfum J.J., Ksiazek T.G. 1999. *The re-emergence of Ebola hemorrhagic fever, Democratic Republic of the Congo, 1995*, Journal of Infectious diseases, Atlanta.
- [9] Laine M. 2008. *Adaptive MCMC Methods with Applications in Environmental and Geophysical Models*, thesis for the degree of Doctor of Philosophy, Helsinki, ISBN 978-951-697-662-7.
- [10] Morgan B. J. T. 2000. *Applied Stochastic modelling*, Oxford University Press Inc., New York, ISBN 0340740418.
- [11] Murray J.D. 2002. *Mathematical Biology: I. An introduction*, Third Edition Springer, 576p, chapter 10 (315-393), ISBN 0-387-95223-3.

- [12] Pengelly J. 2002. *Monte Carlo Methods*, Lecture notes, (<http://www.cs.otago.ac.nz/cosc453/student-tutorials/monte-carlo.pdf>), Accessed on October, 10th, 2008.
- [13] Phenyó E. and Bärbel F. 2006. *Statistical Inference in a Stochastic Epidemic SEIR Model with Control Intervention: Ebola as a case study*, Biometrics, Warwick, Coventry CV4 7AL, U.K. (<http://ite.gmu.edu/klaskey/OR680/MSSEORProjectsSpring07>); accessed on November, 23rd 2008.
- [14] Robert P., Casella G. 2004. *Monte Carlo Statistical Methods*, New York, USA: Springer-Verlag, 645p, chapters 1-10. ISBN0-387-21239-6.
- [15] Ryan C.W. Hall, M.D., Richard C.W. Hall, M.D., Marcia J. Chapman 2008. *The 1995 Kikwit Ebola outbreak: lessons hospitals and physicians can apply to future viral epidemics*, journal of General Hospital Psychiatry, Elsevier, Cleveland.
- [16] Sheldon M.R. 1997. *Introduction to Probability Models*, sixth Edition. Academic Press, San Diego London Boston.
- [17] Solonen A. 2006. *Monte Carlo Methods in Parameter Estimation of Nonlinear Models*, LUT, Master Thesis, Lappeenranta.
- [18] The World Health Organization (WHO), 2007. *The world health report, 2007*, (<http://www.who.int/whr/2007/whr07-en.pdf>), accessed on November 2nd, 2008.
- [19] The World Health Organization Report, 2004. *Ebola Outbreak Chronology*, (<http://www.who.int/mediacentre/factsheets/fs103/en/index1.html>); accessed on November, 4th , 2008.
- [20] Wikipedia, 2008. *Ebola*. (<http://en.wikipedia.org/wiki/Ebola>); accessed on November, 4th, 2008.