NTNU
Norwegian University of
Science and Technology

Statistical Analysis of Quantitative PCR Data

Tonje Gulbrandsen Lien

Master of Science in Physics and Mathematics
Submission date: June 2011
Supervisor: Mette Langaas, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

Problem description

The aim of this thesis is to evaluate and further develop methods for quantification of gene expression based on polymerase chain reaction (PCR) data. Emphasis is put on statistical modeling of the PCR amplification process. Both simulated and real data from two dilution experiments are used in the evaluation.

Preface

This thesis completes my Master in Industrial Mathematics at the Norwegian University of Science and Technology. I have spent five years in the wonderful city Trondheim, which has given me a fantastic time as a student. I would like to thank all my classmates for all support and happy memories.

My paper is a cross-disciplinary project, where statistical models and methods are used to analyze and interpret biological data. Biostatistics is a challenging and exciting field and I want to continue working within this topic. I would like to thank Tommy S Jørstad for his work on Jørstad, Follestad, Langaas and Bones (2008) and the Clusterin dilution-dataset, and to Christina Sæten for making the Clusterin dilution-dataset available to me.

Especially, I will give my warmest gratitude to my supervisor Mette Langaas for a deep commitment and that she always has time for a chat.

Abstract

This thesis seeks to develop a better understanding of the analysis of gene expression to find the amount of transcript in a sample. The mainstream method used is called Polymerase Chain Reaction (PCR) and it exploits the DNA's ability to replicate. The comparative CT method estimate the starting fluorescence level f_0 by assuming constant amplification in each PCR cycle, and it uses the fluorescence level which has risen above a certain threshold. We present a generalization of this method, where different threshold values can be used.

The main aim of this thesis is to evaluate a new method called the Enzymological method. It estimates f_0 by considering a cycle dependent amplification and uses a larger part of the fluorescence curves, than the two CT methods.

All methods are tested on dilution series, where the dilution factors are known. In one of the datasets studied, the Clusterin dilution-dataset, we get better estimates from the Enzymological method compared to the two CT methods.

Contents

Chapter 1

Introduction

This thesis is a cross-disciplinary project, where statistical models and methods are used to analyze and interpret biological data. The data comes from Polymerase Chain Reaction (PCR), which is a laboratory technique commonly used in molecular biology to produce many copies of any short sequence of DNA (or RNA). *Beginning with a single molecule of the genetic material DNA, the PCR can generate 100 billion similar molecules in an afternoon.* For this achievement, Kary Mullis won the 1993 Nobel Prize in Chemistry. In our analyses, the aim is to quantify the relative abundance between a selected set of DNA sequences, such as the Fold Change (FC).

In Chapter 2, the idea behind the PCR method is presented. One PCR cycle amplify selected sections of DNA for analysis. For each amplification cycle j, we measure a fluorescence level f_j, resulting in a fluorescence curve for a specific DNA sample. It is common to repeat this process for around 40 cycles. The fluorescence levels measured are assumed to be proportional to the amount of transcript in the sample. For early cycles in the process, the amplification curve is dominated by noise and uncertainty in the measurements. The amplification decreases toward the end of the PCR runs. The most popular method for estimating the Fold Change between samples, the comparative CT method, assumes constant amplification in each PCR cycle, resulting in an exponential growth in the amplification curve. This method only use the first cycle where the fluorescence level has risen above a certain threshold. We present a generalization of this method, where different threshold values can be included.

In this thesis, we will look at new methods for estimating the starting fluorescence level f_0. Chapter 3 introduces the Enzymological method, which allows cycle dependent amplification and uses a larger part of the amplification curves than the comparative CT method. The efficiency depends on the free enzyme concentration given in the Michaelis-Menton kinetics. When assuming additive, normally distributed noise the f_0 can be estimated by the maximum likelihood estimator.

The maximum likelihood is optimized numerically using the Nelder-Mead simplex method. In Chapter 4, the method is evaluated on a dataset from the Arabidopsis Thaliana plant, called the Arabidopsis dilution-dataset.

In Chapter 5 we use simulated data following the mathematical models behind the Enzymological method. We evaluate the part of the amplification curve where the model is valid and study the impact of the initial values in the Nelder-Mead simplex method. Further testing is performed on three versions of the Enzymological method.

In Chapter 6, the three methods are evaluated and compared to the two CT methods, using the two real datasets. In addition to the Arabidopsis dilution-dataset, we will analyze the Clusterin dilution-dataset with different primer pairs. Some of the primers are not optimal, leading to not perfect efficiencies. We will in this dataset see that the Enzymological method preforms better.

Chapter 2

Polymerase chain reaction data

One of the main goals in the fields of functional genomics is to measure the expression of genes in a sample. Both the absolute number of copies and the relative amount between different samples are interesting. We can for instance analyze samples from healthy individuals compared to samples from diseased individuals. A challenge is that the number of copies of DNA, called n_0 is small and thus hard to detect. A world wide used solution to this problem is the PCR method. The PCR method amplifies and simultaneously quantifies a gene, by generating thousands to millions of copies of a particular DNA sequence.

In this chapter we will present the PCR process in a mathematical perspective, and introduce methods for analyzing the PCR data. A well known method for analyzing PCR data is called the comparative CT method. We will present this method and a generalization thereof, which we call the generalized CT method. These two methods will serve as benchmark methods in the study of new methods for analyzing PCR data.

2.1 The PCR method

We have a sample with a number n_0 of copies of DNA molecules. In every living material DNA is copied, a process called replication. One PCR cycle imitates this process and make copies of all the DNA strings in our sample. For each gene we want to amplify we find optimal primer pairs, which locates the gene on the DNA string. One uses the most optimal primer available so that the amplification is as close to perfect as possible. In Schmittgen and Livak (2008) they establishes that if the PCR efficiency is not optimized, it is recommended to design new primers. Starting with n_0 DNA strings and a perfect replication, we get $n_1 = 2 \cdot n_0$ copies of the DNA strings after one PCR cycle. In PCR cycle j we get n_j number of DNA strings equal to

$$n_j = 2 \cdot n_{j-1} = 2^j \cdot n_0$$
$$n_0 = 2^{-j} \cdot n_j \tag{2.1}$$

This forms an exponential curve. In many cases the amplification is not perfect and the PCR efficiency E is not always equal to two. We do not directly count the number of DNA copies after each PCR cycle but we measure it using a fluorescence dye. The measured fluorescence level at cycle j and curve i is f_{ji}. We assume that f_{ji} is proportional to the number of DNA copies, with $f_{ji} = \gamma \cdot n_{ji}$, where the scaler factor γ is not dependent on cycle. At some cycle j the fluorescence curve is

$$f_{0i} = E^{-j} \cdot f_{ji}$$

where $1 < E < 2$. After a while the amplification rate decreases, and we normally perform around 40 PCR cycles. We often analyze many samples simultaneously, where each sample get it's own fluorescence curve. These samples have different starting fluorescence levels.

We assume that we observe this fluorescence level with a noise ε_{ji}. We observe a fluorescence level y_j after each cycle j such that we can see the results in real time. A baseline level at cycle j for curve i is called b_{ji}, and it can be estimated by various methods. We will explain the chosen method when the datasets are introduced. The observed fluorescence levels before baseline correction are called y^*_{ji}, and after baseline correction y_{ji}. The equation becomes

$$y^*_{ji} - \hat{b}_{ji} = y_{ji} = f_{ji} + \varepsilon_{ji}$$

Usually the baseline for curve i is approximately similar for each cycle. The use of an inferior baseline correction method will have higher consequences for the lower fluorescence levels in the early cycles than for late cycles. In our analysis we will not concentrate on the early cycles. The research into baseline correction methods will not be of focus in this thesis.

An example of an observed background corrected fluorescence curve y is plotted in Figure 2.1. After cycle 20 we can visually see that the number of DNA copies has increased. What looks to be the flat part of the curve from cycle 1 to 20 is called the *ground phase*. The next phase is called the *exponential phase*. Around cycle 31 we see the inflection point. Before the inflection point the fluorescence curve has a concave character and afterwards a convex character. The amplification is close to perfect for early cycles, and decreases after some PCR cycles.

In Figure 2.2 we see the log transformed fluorescence curve. We see that the exponential phase is now linear on the log scale. There are rapid fluctuations in the early cycles, where the observed fluorescence levels are small. The negative

values of y, due to the baseline correction, can not be log transformed and are thus not shown. It is not possible to see the assumed linear trend in the fluorescence level in the early cycles.

The amplification efficiency, also called PCR efficiency, is often referred to as $1 + p_j$, where p_j is the proportion of DNA strings which is perfectly duplicated in cycle j. We estimate these parameters by

$$\begin{aligned} \hat{E_j} &= y_{j+1}/y_j \\ \hat{p_j} &= y_{j+1}/y_j - 1 \end{aligned} \tag{2.2}$$

In many biological studies technical triplicates are generated. The biologist use triplicates to make their analysis robust. With technical replication they have backup curves if they have to discard a damaged sample. These three technical replicates should in principle have the same number of copies of DNA transcript. If we are not interested in the variance between the technical replicates, we will calculate the mean value of the curves. The mean value for the estimated starting fluorescence is calculated as

$$\hat{f}_0 = \frac{1}{3}\sum_{k=1}^{3} \hat{f}_{0k} \tag{2.3}$$

were k is replicate index. We often have many sets of technical triplicates. In Chapter 6 we will also look at the results using regression models, and take into account all $\hat{f_{0k}}$ including the technical triplicates.

In many studies it is common to calculate a ratio called the Fold Change (FC), based on the estimates of the starting fluorescence levels $\hat{f}_0$ between samples. In these studies there often is a group A and B which can be a case and control group. In all samples there is a PCR run for a gene of interest G, and one or more references genes. With one reference gene R and the gene of interest G in samples A and B, which are not paired, we can calculate the FC from the $\hat{f}_0$ as

$$FC = \frac{\hat{f}_0{}^{GA}/\hat{f}_0{}^{RA}}{\hat{f}_0{}^{GB}/\hat{f}_0{}^{RB}} \tag{2.4}$$

In the methods which soon will be introduced, the FC is the motivation. Later we will look at ratios between fluorescence levels, but in another setting where the ratio is known, namely dilutions series. But first, let us look at how to estimate the starting fluorescence levels.

2.2 The Comparative CT method

The comparative CT method is a mainstream method to analyze PCR data. From the fluorescence curves we choose one threshold for the fluorescence level after the

ground phase, and find the corresponding cycle called CT. The efficiency is assumed perfect, thus equal to two. The threshold is chosen early in the exponential phase to support the assumption $E = 2$. We estimate the starting fluorescence level for curve i as

$$\hat{f_{0i}} = f_{ji} \cdot E^{-j} = T \cdot 2^{-CT_i} \tag{2.5}$$

When using this method we will set one threshold for all fluorescence curves within a dataset. In Livak and Schmittgen (2001) and Schmittgen and Livak (2008), the comparative CT method is not presented as Equation (2.5) but as the FC.

2.3 Generalization of the comparative CT method

When different samples with different starting concentrations are run through the PCR process simultaneously the results are curves where the exponential phase ends at different cycles. It can be a challenge to set one threshold for all the curves. Based on this requirement we introduce a generalization of the comparative CT method.

When using the generalized CT method one threshold is chosen for each fluorescence curve within a dataset. The T_i is placed where the estimated efficiency for curve i is closest to two. We have found a method which find candidates for the CT value were the efficiency is at its highest. The threshold value is chosen as the fluorescence value corresponding to the chosen CT value. We estimate E for each cycle using Equation (2.2), and start by initializing the CT value to be at cycle m near the inflections point. We accept $CT = m - 1$ if the $\hat{E}_{m-1} > \hat{E}_m$ and smaller than 2. In general we move the CT value to cycle $k-1$ if $\hat{E}_{i,k-1} > \hat{E}_{i,k}$ and $\hat{E}_{i,k-1} < 2$. We end up with the cycle $CT_i = k$ where the efficiency is as close to perfect amplification as possible. The corresponding y_{CT_i} becomes the threshold T_i, and

$$\hat{f}_{0i} = T_i \cdot 2^{-CT_i}. \tag{2.6}$$

These two methods try to model the fluorescence curve on the interval $y_0, \ldots, y_{CT}$, using only the value $y_{CT} = T$.

The opportunity of different thresholds more suitable for each curve may improve the estimates of f_0. If the same threshold is chosen for all curve we will get the same estimates as for the comparative CT method. It is interesting to compare these two methods, which we will come back to in Chapter 6. First we will look at a method which opens up for cycle dependent efficiency and evaluation of more than one point in several cycles simultaneously.

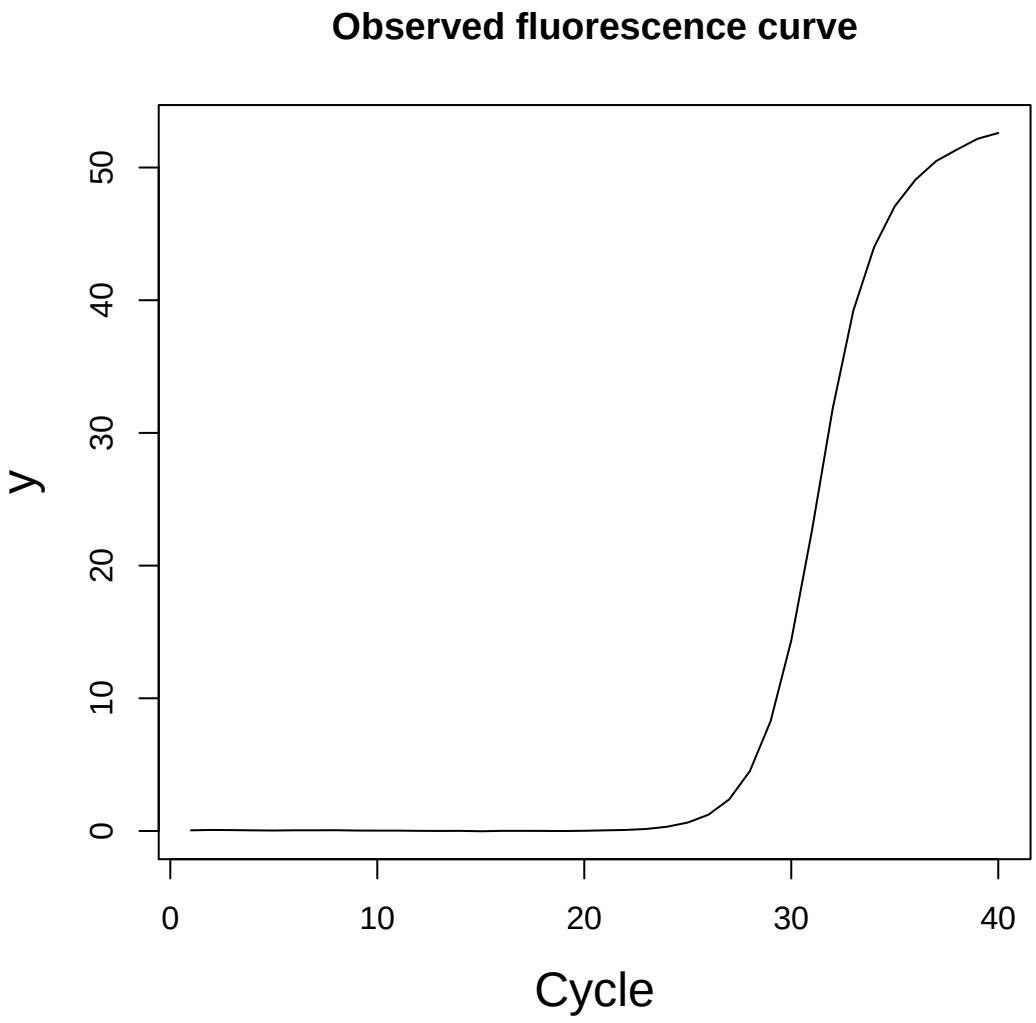

Figure 2.1: Example of an observed background corrected fluorescence curve. The curve is from a dataset called Arabidopsis dilution-dataset, which will be introduced in Chapter 4.

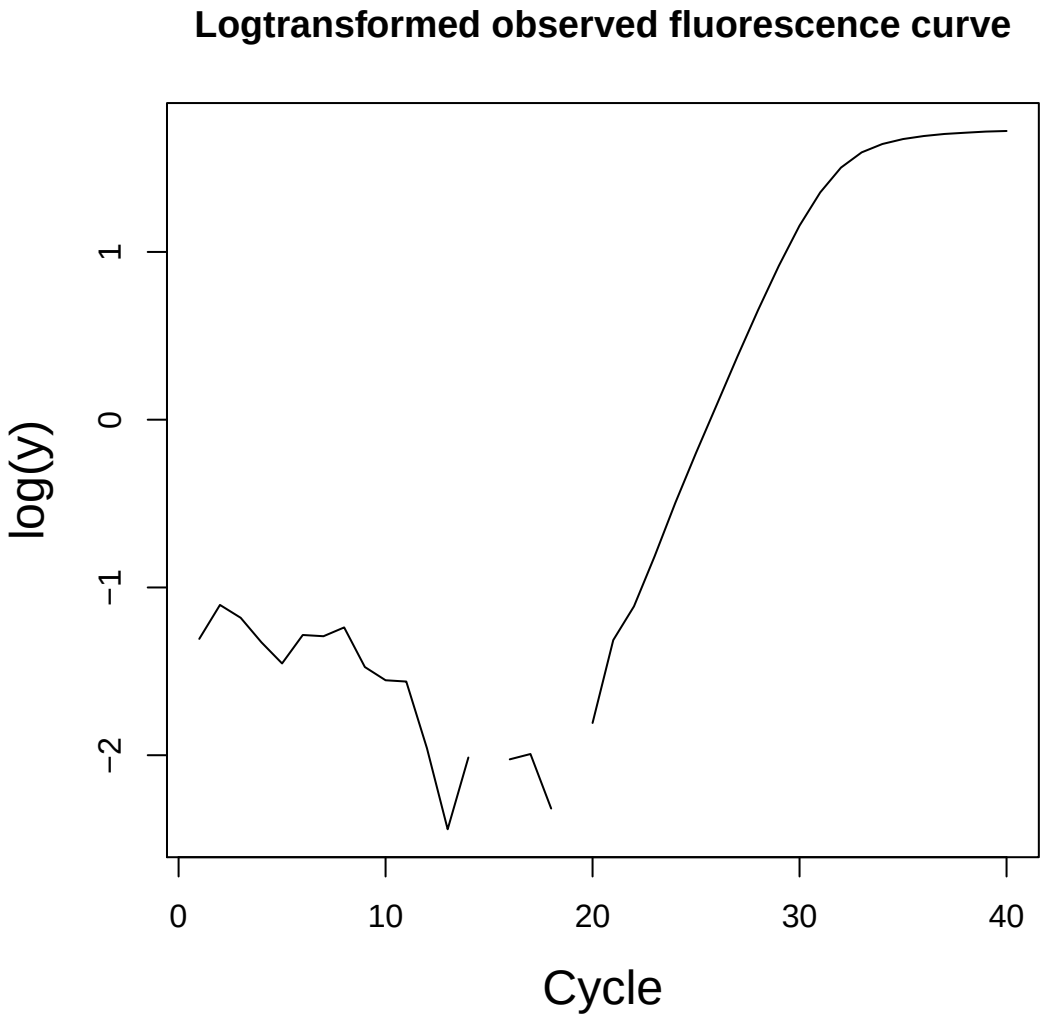

Figure 2.2: Examples of an observed log transformed baseline corrected fluorescence curve. For the early cycles we see the rapid fluctuation that can not be explained by the amplification process.

Chapter 3

The Enzymological method

In this section we will give an introduction to the new method presented in Jørstad et al. (2008). We will call this method the Enzymological method. The main goal of the method is to estimate the starting fluorescence level corresponding to a cell sample. When evaluation the results we will look at ratios between starting fluorescence levels. Jørstad et al. (2008) compares the Enzymological method to a known suitable benchmark method called MoBPA Alvarez, Vila-Ortiz, Salibe, Podhajcer and Pitossi (2007) and show that the Enzymological method outperforms this method. In this thesis we will first take a closer look at the assumptions behind the Enzymological method and evaluate its strengths and weaknesses. We will compare the Enzymological method to the comparative CT method and the generalized CT method.

3.1 A mathematical model for the PCR efficiency

The Enzymological method is presented in Jørstad et al. (2008) and uses Michaelis-Menten (enzyme) kinetics to model the PCR efficiency, see Schnell and Mendoza (1997). It is assumed that the reaction efficiency is mainly determined by the ratio of free to total enzyme concentration. For some rate constant κ the PCR efficiency E is given by the concentration of target DNA molecules x as

$$E(x) = 1 + p(x) = 1 + \frac{\kappa}{\kappa + x}, \tag{3.1}$$

where $p(x)$ is the proportion of x that is perfectly duplicated in each PCR cycle.

When x is small, almost every DNA molecule is duplicated and when x gets larger the duplication rate drops. If the PCR efficiency in cycle $j = 1, \ldots, m$ follows Equation (3.1) the concentration of target DNA at cycle j can be given by

the following deterministic model

$$x_j = x_{j-1}(1 + p(x_{j-1})) = x_{j-1}(1 + \frac{\kappa}{\kappa + x_{j-1}}), \tag{3.2}$$

where x_0 is the starting concentration of target DNA molecules.

Definition (Fluorescence level)**.** *The true (not observed) intensity level at cycle j is denoted f_j, and is assumed to be proportional to the number of target DNA molecules and thus proportional to the concentration of target DNA x. We write $f_j = \gamma x_j$ where γ is independent of cycle and sample.*

Let $\alpha = \gamma\kappa$ such that Equation (3.2) becomes

$$f_j = f_{j-1}(1 + \frac{\alpha}{\alpha + f_{j-1}}), \tag{3.3}$$

The primary goal of the Enzymological method is to estimate the starting fluorescence f_0 for each fluorescence curve.

Motivated by the analysis of experimental data in Follestad, Jørstad, Erlandsen, Sandvik, Bones and Langaas (2010) we will consider a generalized form of Equation (3.3) where we introduce parameter β

$$f_j = f_{j-1}(1 + \frac{\alpha}{\alpha + f_{j-1}^{\beta}}). \tag{3.4}$$

The proportion of DNA molecules that is duplicated for each cycle can be written

$$p(f_j) = \frac{\alpha}{\alpha + f_j^{\beta}} \tag{3.5}$$

An example of the curve from Equation (3.5) is shown in Figure 3.1. For early cycles around 95% of all DNA molecules are duplicated, and for later cycles the proportion of DNA molecules that is duplicated drops to around 50% for this example curve.

3.2 Choosing the (s_i, m_i)-interval

For each curve i we only look at data from cycles $s_i \leq j \leq m_i$. Then s_i is the start cycle and m_i is the end cycle for the part of the curve which is assumed to follow the mathematical model introduced in Equation (3.4). The assumed cycle dependent efficiencies in Equation (3.1) take in consideration the ratio of free to total enzyme concentration. However, towards the end of a PCR run we see that

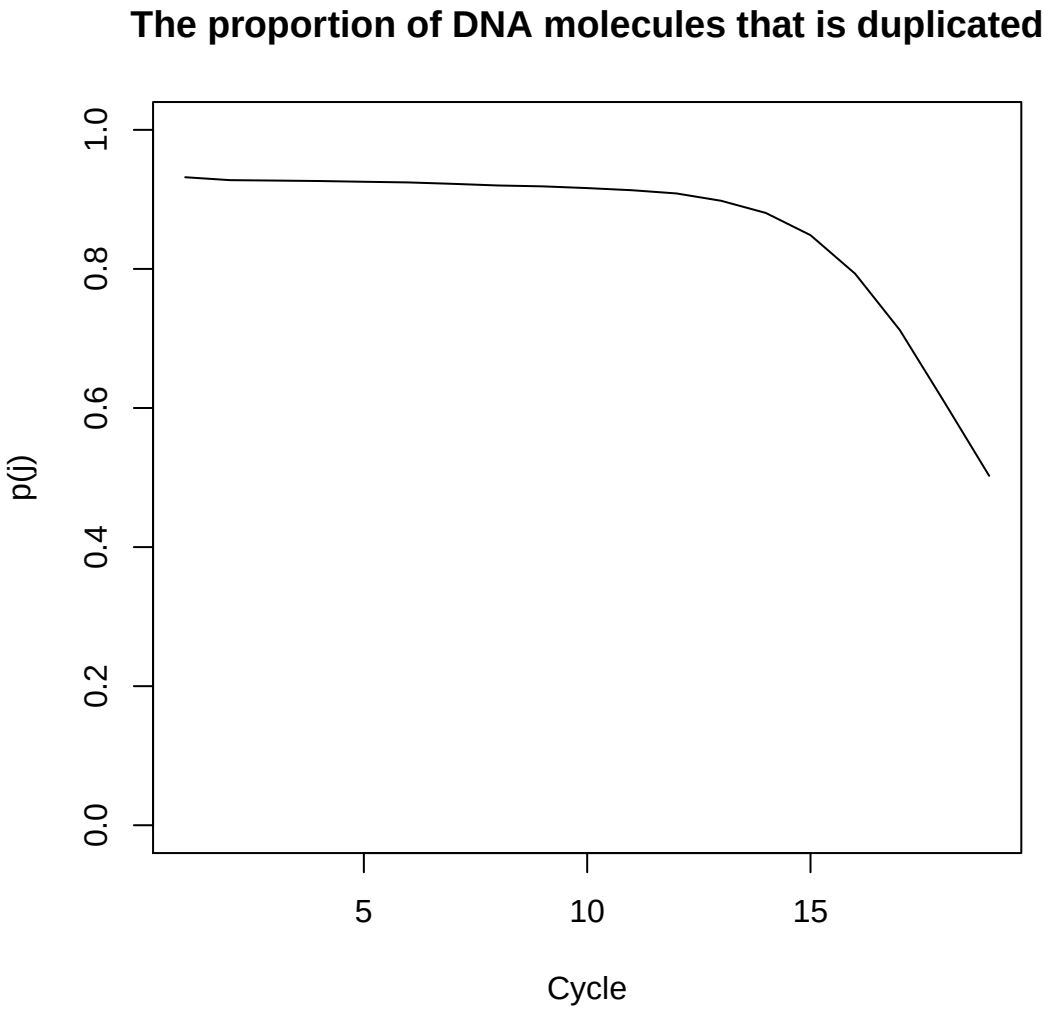

Figure 3.1: The $p(f_j)$ function presented in Equation (3.5). The curve is generated with $\alpha = 20$, $\beta = 1$ and $f_0 = 2.036e^{-06}$. All parameter values are motivated by the analysis of the Arabidopsis dilution-dataset which is introduced in Section 4. Cycles after a certain point called m (which here are cycle 19). This cycle m will be discussed in Section 3.2.

the factors such as denaturing of the enzymes, reannealing of the DNA molecules or a shortage of primers or nucleotides will eventually become important. We only fit the model in Equation (3.4) for the fluorescence curves i up to cycle m_i.

The Enzymological method models the fluorescence curve on the interval $y_{0i}, \ldots, y_{mi}$, using the interval $y_{s_i}, \ldots, y_{m_i}$. We look at the shape of the function in Equation

(3.4) to find a proper end cycle m_i. If we, in this thesis, take a closer look at the possible values for the parameter β we see that the model in Equation (3.4) has different behavior when the parameter β changes. The difference between two consecutive fluorescence levels $f_j - f_{j-1} = \frac{\alpha f_{j-1}}{\alpha + f_{j-1}^{\beta}}$ has different limiting growth. When $f_{j-1} = z \to \infty$

$$\begin{aligned}
\frac{\alpha z}{\alpha + z^{\beta}} &= \frac{\alpha}{\frac{\alpha}{z} + z^{\beta-1}} && \to 0 && \text{for } \beta > 1 \\
\frac{\alpha z}{\alpha + z} &= \frac{\alpha}{\frac{\alpha}{z} + 1} && \to \alpha && \text{for } \beta = 1 \\
\frac{\alpha z}{\alpha + z^{\beta}} &= \frac{\alpha z^{1-\beta}}{\frac{\alpha}{z^{\beta}} + 1} && \to \infty && \text{for } \beta < 1
\end{aligned}$$

For $\beta = 1$ and $\beta < 1$ we note that Equation (3.4) have a limiting growth such that the function is convex. From empirical experience, the fluorescence curve we want to model is convex before its inflection point and then concave for late cycles. In Jørstad et al. (2008) the end cycles m_i is chosen to be the cycle closest to, and smaller than the inflection point x_0 in the observed fluorescence curve. Then we avoid the limiting problem, if $\beta \leq 1$. The inflection point is determined by fitting a four-parameter sigmoid function to the observed fluorescence level on the original scale using a least squares approach. The parameterization used for the sigmoid curve is

$$y_0 + \frac{a}{(1 + e^{(x_0 - x)/b})},$$

with x_0 being the inflection point.

How can we find the start cycle s_i to be used for curve i? Why not $s_i = 1$? The mathematical model in Enzymological method assumes that the efficiency starts at the value 2 with perfect amplification from the very first cycle, and decreases later in the PCR process. From a biological perspective the mathematical model applies from cycle one. But as seen in Figure 2.2 there are rapid fluctuations in the observed fluorescence values even after baseline correction that cannot be explained by the amplification process in our mathematical model. To avoid affecting the model fit Jørstad et al. (2008) suggest to use the method discussed in Tichopad, Dilger, Schwarz and Pfaf (2003). This method considers a linear behavior for the fluorescence level in the early cycles and then tries to find the point where the exponential phase starts, thus departs from the ground phase. To make sure that we have detected a departure from the ground phase we find the third outlier from the ground phase. We will call this approach the Tichopad approach. Based on practical experiments we found this approach to be slow and to include data

points with much relative noise. We will therefore look closer at other suggestions for finding s_i in Section 5.4.

3.3 Evaluating the PCR efficiency assumption

The Enzymological method is based on the assumption that the PCR efficiency can be modeled as a function of the proportion of DNA molecules that is duplicated, using the Michaelis-Menten (enzyme) kinetics. The mathematical expression is given in Equation (3.5). Using the logit transformed model in Equation (3.5) we get a linear relationship between logit(p_j) and log(f_j)

$$logit(p(f_j)) = log(\frac{p(f_j)}{1-p(f_j)}) = log(\frac{\alpha}{f_j^{\beta}}) = log(\alpha) - \beta log(f_j). \tag{3.6}$$

We can use this relation to investigate if the model in Equation (3.4) is appropriate for $y_{si}, \ldots, y_{mi}$ in curve i. We use the following estimates for f_j and p_j

$$\hat{f}_j = y_j \tag{3.7}$$

$$\widehat{p(f_j)} = \hat{p}_j = y_{j+1}/y_j - 1 \tag{3.8}$$

We do visual inspection in a plot with $logit(\hat{p_j})$ on the y-axis and $log(y_j)$ on the x-axis. We call this plot the *p-assumption plot*.

3.4 Parameter estimation

Let y_{ji} denote the observed baseline corrected fluorescence level in sample $i = 1, \ldots, n$ and cycles $j = 1, \ldots, k$.

Definition (Number of cycles)**.** *Let the observations after each PCR cycle be* $\mathbf{y} = y_1, \ldots, y_m$*, where* y_j *is the observed fluorescence level after cycle* j*. The first observation* y_1 *is after one PCR cycle and therefore after one duplication of the cDNA. We are interested in the starting fluorescence level* y_0*.*

We only use cycle $j = s_i, \ldots, m_i$ where we assume that the PCR efficiency in Equation (3.5) holds. Following Jørstad et al. (2008) the observed fluorescence is modeled with additive noise

$$y_{ji} = f_{ji} + \varepsilon_{ji}, \tag{3.9}$$

where ε_{ji} is independent and distributed as $N(0, \sigma^2)$. The noise is relatively large compared to the fluorescence levels for early cycles, and relatively small compared to the fluorescence levels for cycles towards the end of the PCR process.

For sample $i = 1, \ldots, n$ we assume common parameters α, β and σ^2. The observed fluorescence is modeled with Equation (3.4)

$$y_{ji} = f_{j-1,i}(1 + \frac{\alpha}{\alpha + f_{j-1,i}^{\beta}}) + \varepsilon_{ji}. \tag{3.10}$$

Denoting the vector for all observed baseline corrected intensities $\{y_{ji}\}$ the log-likelihood function for $\theta' = (\sigma^2, \alpha, \beta, f_{01}, \ldots, f_{0n})$ is

$$lnL(\theta'|\mathbf{y}) = \sum_{i=1}^{n} \sum_{j=s_i}^{m_i} [-\frac{1}{2} ln(2\pi) - ln(\sigma) - \frac{1}{2\sigma^2}(y_{ji} - f_{ji})^2] \tag{3.11}$$

We observe that maximization of the likelihood can be performed separately with respect to σ^2 and $\theta = (\alpha, \beta, f_{01}, \ldots, f_{0n})$, where θ is of primary interest when finding $f_{01}, \ldots, f_{0n}$. To estimate θ we find it sufficient to minimize

$$\sum_{i=1}^{n} \sum_{j=s_i}^{m_i} (y_{ji} - f_{ji})^2.$$

This can be achieved by differentiation with respect to each parameter in θ. But gradient based methods perform poorly and the calculations are time consuming because of the recursion derived from Equation (3.4). In Jørstad et al. (2008) the estimates are found numerically using the Nelder-Mead simplex method.

The initial estimates of the parameters f_0, β and α

The initial parameters used in the Nelder-Mead simplex method are chosen as follows. The parameter β will be initialized as

$$\beta^{init} = 1, \tag{3.12}$$

because this is the theoretically correct value from Schnell and Mendoza (1997).

From our model we have found a new estimator for the amplification in each cycle

$$p_j^* = \frac{\hat{\alpha}}{\hat{\alpha} + y_j^{\hat{\beta}}}. \tag{3.13}$$

This estimator together with the estimator $\hat{p}$ in Equation (2.2) will be used in the initialization of α. First the parameter α will be initialized to $\sum_{i_1}^{n} \frac{{\alpha_i}^{init}}{n}$, where ${\alpha_i}^{init}$ is estimated for each replicate curve separately. An ${\alpha_i}^{init}$ is calculated using $\beta = 1$ and the last two observed intensities belonging to the model region, y_{i,m_i} and $y_{i,m_{i-1}}$ as

$$
\begin{aligned}
\hat{p}_{i,j-1} &= p^*_{i,j-1} \\
\frac{y_{i,m_i}}{y_{i,m_i-1}} - 1 &= \hat{\alpha}_i^{\,init}/(\hat{\alpha}_i^{\,init} + y_{i,m_i-1}) \\
\hat{\alpha}_i^{init} &= \frac{y^2_{i,m_i-1} - y_{i,m_i-1} \cdot y_{i,m_i}}{y_{i,m_i} - 2y_{i,m_i-1}} \\
\hat{\alpha}^{init} &= \sum_{i_1}^{n} \frac{\hat{\alpha}_i^{\,init}}{n}.
\end{aligned} \tag{3.14}
$$

This is motivated by that the last observed intensities y_{i,m_i} and $y_{i,m_{i-1}}$, where the model applies, are the least noisy.

The f_{0i} is initialized using Equation (3.3) in reverse by setting

$$
f_{i,m_i} = y_{i,m_i} \text{ and using the initial values of } \alpha \text{ and } \beta. \tag{3.15}
$$

Chapter 4

Using biological data to evaluate the Enzymological method

In this chapter we will use the method presented in Chapter 3 to see how the Enzymological method works in practice. The method will be applied to a biological dataset called Arabidopsis dilution-dataset, presented in Jørstad et al. (2008).

4.1 The Arabidopsis dilution-dataset

This experiment involves two genes TIP41 (gene of interest G) and PP2A (reference gene R). The genes are taken from two types of plants, the AtRAC7 knock-out mutant plants (case group A) and the wild-type plants (control group B). We call GA, RA, GB and RB the four sample types. Dilutions are made with dilution factors 1, 4, 16 and 64. In this case the ratio of two starting concentrations is known. In Table 4.1 we see an overview of the sample types and dilutions. There are triplicate observations for each of the fluorescence levels. In total there are 4 sample type times 4 dilutions time 3 replicates, which is equal to 48 starting fluorescence levels and thus 48 fluorescence curves. We have not available information about the nature of the triplicates, that is whether the triplicates are technical or biological replicates. By *Biological replicates* we mean that the samples are found from different plants but with the same sample type. By *Technical replicates* we mean that one sample is taken from a plant and divided into three. Technical replicates will be more similar than the biological replicates. Plots of the observed fluorescence curves before baseline correction are found in Figure 4.1.

The data are baseline corrected as in Jørstad et al. (2008). This is for each curve done by first ranking the fluorescence observations according to numerical value, and then finding the window of 5 consecutive data points having the smallest rank sum. In the case of multiple windows with the same minimum rank sum the

Sample type	Dilutions			
	1:1	1:4	1:16	1:64
GA	f_0^{GA}	$\frac{f_0^{GA}}{4}$	$\frac{f_0^{GA}}{16}$	$\frac{f_0^{GA}}{64}$
RA	f_0^{RA}	$\frac{f_0^{RA}}{4}$	$\frac{f_0^{RA}}{16}$	$\frac{f_0^{RA}}{64}$
GB	f_0^{GB}	$\frac{f_0^{GB}}{4}$	$\frac{f_0^{GB}}{16}$	$\frac{f_0^{GB}}{64}$
RB	f_0^{RB}	$\frac{f_0^{RB}}{4}$	$\frac{f_0^{RB}}{16}$	$\frac{f_0^{RB}}{64}$

Table 4.1: Overview of the different starting fluorescence levels organized in groups and dilution factors. There are three replicates for each of the 16 starting fluorescence levels.

window containing the smallest single rank was used. The arithmetic mean of the observations in the chosen window was used as baseline value and subtracted from all observations.

4.2 Finding the (s_i, m_i)-interval

The Enzymological method uses the Tichopad approach to find the starting cycle s_i and place the last cycle m_i right before the inflection point in the fluorescence curve, as explained in Section 3.2. In Jørstad et al. (2008) the chosen part of the curve $(y_s, \ldots, y_m)$ is evaluated by checking the assumption in Equation (3.6). The p-assumption plots in Figure 4.2 show the chosen part of the fluorescence curve for the four dilutions in group RB. The p-assumption plots for the other three groups show a similar trend. We do not see a linear relationship with slope $-\beta$. For early cycles, thus small values of $log(y_j)$, we see that the logit$(\hat{p_j})$ deviates from linearity. But for larger cycles toward the inflection point we see that the relation is more linear. In Jørstad et al. (2008) they conclude that the lines provided a reasonable explanation and a had a linear trend. Data from the earliest cycles in the plots were left out. We will in this thesis conclude otherwise. In Figure 4.2 we have included all points $(s_i, \ldots, m_i)$ and from visual inspection we think that it is clear that the assumption in Equation (3.6) is not fulfilled. The main reason for departure from the linear trend is the rapid fluctuations in the observed fluorescence level for early cycles.

When using the Enzymological method on data $(y_{s_i}, \ldots, y_{m_i})$ this method gives good estimates for the parameters f_0 even though it looks like from Figure 4.2 that s is chosen too early. Why is this? Maybe the p-assumption that Jørstad et al. (2008) ask us to check is not the right assumption to concentrate on. We will now take a closer look at how the MLE is found, thus how the MLE works. These two

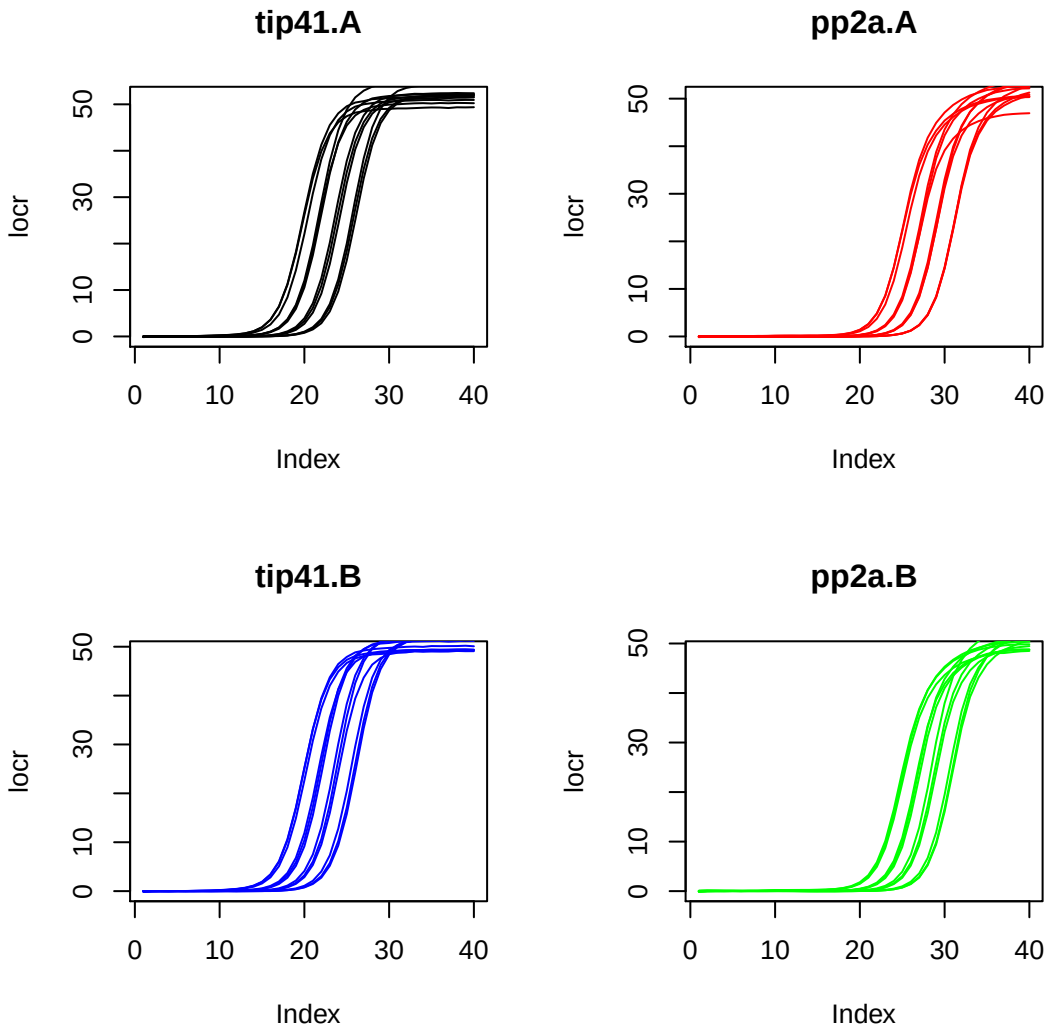

Figure 4.1: The observed fluorescence level in the Arabidopsis dilution-dataset. The four different sample types (GA, RA, GB and RB) are plotted in separate plots with different colors. In the upper left we have GA (black), upper right RA (red), lower left GB (blue) and lover right RB (green). The three curves to the left in each panel have the original concentration. The next three curves to the right have dilution factor 4, then the curves with dilution factor 16 are plotted and last the curves with dilution factor 64 to the far right. This is the fluorescence level before baseline correction.

next sections are not found in Jørstad et al. (2008).

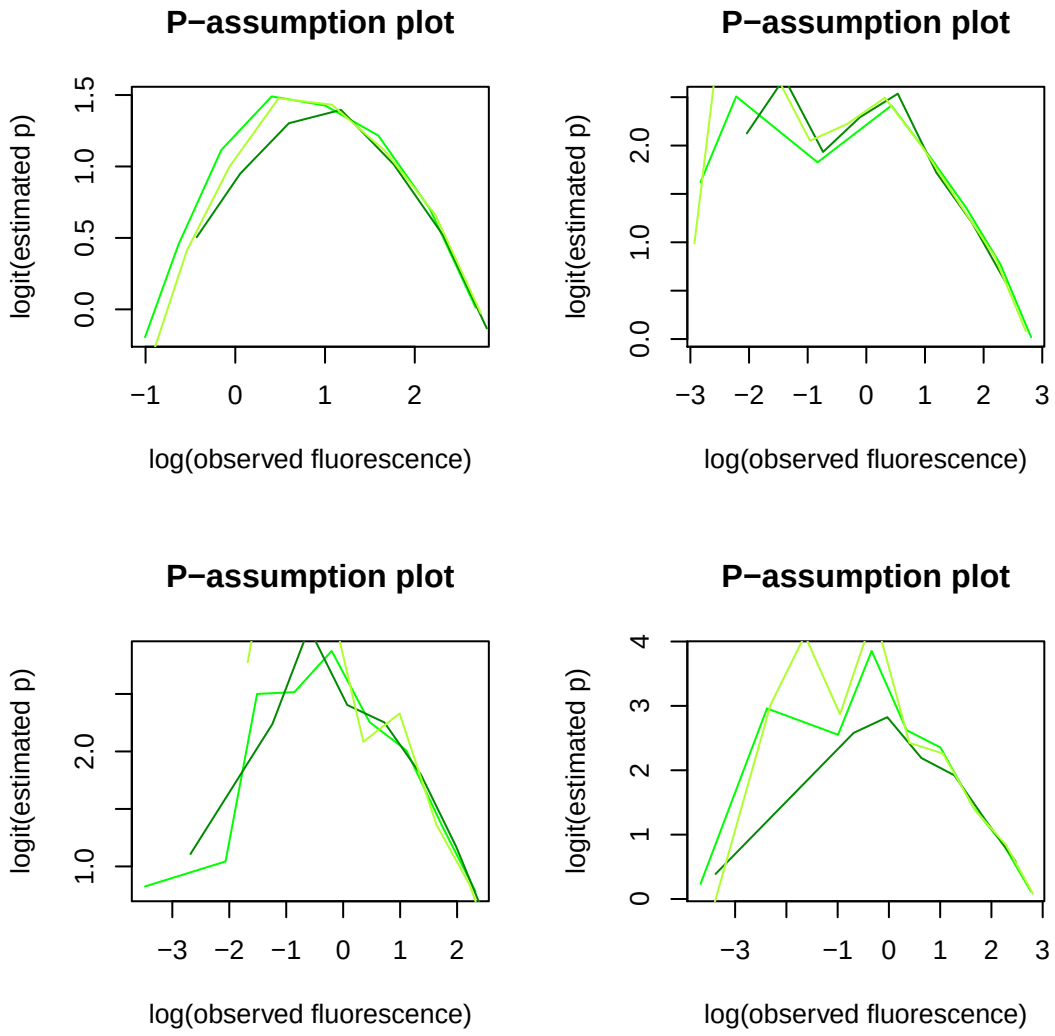

Figure 4.2: The p-assumption plot for the four dilution (1,4,16 and 64) for gene PP2A (R) from the wild type (control group B) in the Arabidopsis dilution-dataset. The three replicates are drawn in different green colors, but the difference between the replicates is not of interest here. On the x-axis we see the logarithm of the observed fluorescence level and on the y-axis we see the logit of the estimated amplification probability.

4.3 A closer look at the maximum likelihood estimator MLE

To estimate θ we maximize the likelihood in Equation (3.11) over f_0, α and β. The error term is estimated by $\hat{\varepsilon} = y - \hat{f}$, using the Enzymological method to estimate $\hat{f}$. What does $\hat{\varepsilon}$ look like? In Figure 4.3 we see a plot of the observed fluorescence level y (red lines) and the estimated fluorescence level $\hat{f}$ (blue lines) with the estimated parameters from the Enzymological method. They seem to follow the same trend. In Figure 4.4 we see $\hat{\varepsilon}$ for sample type RB for all four dilutions. In the panel to the upper left we see the original concentrated samples, where the residuals have a larger variance than for the other samples. Still, overall the residuals are mainly concentrated around zero. Towards the ending cycle (the inflection point) we see that $\hat{\varepsilon}$ increases.

The $\hat{\varepsilon}$ is calculated from $\hat{f}$ with estimated parameters $(\hat{\alpha}, \hat{\beta}, \hat{f}_0)$ such that the term $\sum_{i=1}^{n} \sum_{j=s_i}^{m_i} (y_{ji} - \hat{f}_{ji})^2$ in Equation (3.11) is minimized. What happens to the likelihood when the estimated parameters change? We will change one estimate at a time to see what happens to $\hat{\varepsilon}$. In Figure 4.5 we change $\hat{\alpha}$, and thus $\hat{f}$, such that we get another $\hat{\varepsilon}$. In Figure 4.6 we change $\hat{\beta}$. The black line is the $\hat{\varepsilon}$ with the MLE.

We see that the value for the ε in the cycle close to the inflection point has a very high influence on the resulting likelihood, and thus the estimates. We might only need the two last cycles? What are the results if we only include two points in the MLE?

MLE based on two observations

We look at $n = 1$ and $\beta = 1$ for curve i and only include cycle j and $j - 1$. The MLE solution is found by minimizing the term

$$(y_j - \hat{f}_j)^2 + (y_{j-1} - \hat{f_{j-1}})^2$$

with respect to $\hat{f}$. The solution is $\hat{f}_j = y_j$ and $\hat{f_{j-1}} = y_{j-1}$. The relationship between $\hat{f}$ and $\hat{\alpha}$ is $\hat{f}_j = \hat{f_{j-1}}(1 + \frac{\hat{\alpha}}{\hat{\alpha} + \hat{f_{j-1}}})$. Since $\hat{f} = y$ in the two observations we get $y_j = y_{j-1}(1 + \frac{\alpha}{\alpha + y_{j-1}})$. When $j = m$ we get exactly the same expression as Equation (3.14) which is the value for α used as initial values before the MLE optimization. This is an interesting result. Later in Section 5.5 we will use the initial values directly in the estimation of f_0.

We have in this section seen how the behavior of $\hat{\varepsilon}$ influence the MLE. We make an important assumption about ε, that the error is normally distributed with mean zero and variance σ^2. Can we from the $\hat{\varepsilon}$ accept the hypothesis that ε

is normally distributed? This assumption is not evaluated in Jørstad et al. (2008), therefore we will do some statistical testing on $\hat{\varepsilon}$ in the next section.

4.4 The assumption of an additive model

We want to see if the noise ε_{ji} for cycle j in curve i is independent and normally distributed with mean 0 and standard deviation σ, as assumed in Equation (3.9). Normality is tested by using the Anderson-Darling test on $(\hat{\varepsilon_1}, \ldots, \hat{\varepsilon_m})$ for each curve, thus performing 48 tests. The H_0 hypothesis is that ε_i is normally distributed, with rejection if $p < 0.05$ the significance level. We have multiple testing with some dependent samples. To account for this we calculate Bonferroni adjusted p-value for each curve. The p-value from the Anderson Darling test for each sample is multiplied by the number of samples which is 48. Meaning we accept normality if $p \cdot 48 > 0.05$ or $p > 0.05/48$ for each curve.

We will perform the Anderson-Darling test in two different situations, when we estimate α and β for each individual curve $n = 1$ and next when we estimate one α and β for each triplicate set $n = 3$. When we look at the 48 estimated residuals when α and β are estimated for each individual curve, we accept the normality for all curves. We can accept the hypothesis of normality for 31 of total 48 ε, when α and β are estimated for each triplicate.

In Figures 4.7 and 4.8 we present QQ plots of the curves where the H_0-hypothesis was rejected. We see that the departure from normality do not appears to be large. We conclude that the assumption of normality can be rejected when $n = 1$ and holds approximately for $n = 3$. This could be one of the reasons why the Enzymological method give good estimates even though the p-assumption from Equation (3.6) is not fulfilled. Instead of checking the p-assumption in new dataset we will check if we can assume normally distributed ε.

4.5 Parameter estimation for each individual curve

We start by estimating the parameters based on individual curves, and letting $n = 1$ in Equation 3.11. The results are presented as a plot over all 48 $\hat{\alpha}$ in Figure 4.9 and all 48 $\hat{\beta}$ in Figure 4.10. These two plots are also presented in Jørstad et al. (2008).

We want to see if the parameter estimates vary between gene, group and dilution. If estimates are similar we can have the same parameter for groups of curves. From the plots it looks like the two parameters vary between gene and dilutions factor. We know that the parameter estimates can also depend on the sample type used, in Jørstad et al. (2008), although this is not as evident from the

test results. We see that the two natural choices for n is either $n = 1$ or $n = 3$. Based on Figure 4.9 and Figure 4.10 Jørstad et al. (2008) concludes that α and β should be estimated based on three replicates simultaneously which have the same gene, same sample type and same dilution factor. But we observe that estimates for α are different within each triplicate. The two groups of the gene TIP2 have higher variance with respect to $\hat{\alpha}$ than the other gene. Thus we will not make any conclusions about n at this point.

Another result seen from Figure 4.10 is that all the estimated β are concentrated around one. Based on this we choose to continue to estimate beta and not force it to be equal to one. As mentioned, the motivation for including the parameter β was based on an analysis of experimental data in Follestad et al. (2010).

We now turn to estimated dilution factors, see Table 4.2. We observe that the estimates are close to the true dilution factors.

	$log_2(4/1)$	$log_2(16/1)$	$log_2(64/1)$	$log_2(16/4)$	$log_2(64/4)$	$log_2(64/16)$
True value	2	4	6	2	4	2
$\hat{ratio}$	2.19(0.13)	4.15(0.2)	6.2(0.17)	1.96(0.13)	4.01(0.08)	3.03(2)

Table 4.2: Estimated log_2 of ratios corresponding to the dilutions in the Arabidopsis dilution-dataset using the Enzymological method for each individual curve. The estimates refer to the mean estimated ratios, with standard deviation in parenthesis.

Estimating the dilution factors

We have $\hat{f}_0$ from each of the 48 samples in the Arabidopsis dilution-dataset. In this dataset we do not know if there are biological or technical triplicates. We will consider them as technical triplicates in these calculations. We take the mean value of the estimated fluorescence of a triplicate set as explained in Equation (2.3).

We will estimate dilution factors by forming ratios between dilutions within each group. For the Arabidopsis dilution-dataset we will find all ratios within each of the four groups GA, RA, GB and RB. Thus we will have 4 estimates of each of the dilution factor $4, 16$ and 64.

In general we have the dilution factors $1 : a$, $1 : b$ and $1 : c$. For our dataset we have more dilution factors, but for illustration we look at three different dilution factors. Let the starting concentration have dilution factor $1 : 1$. Then $a = 1$, thus the original concentration. In this general case we have, for each dilution series,

three estimates of the starting concentration $\hat{f}_0(D)$ corresponding to each of the three dilution factors D.

There are three different ratios R which are known, and that is b/a, c/a and c/b. For illustration we show the calculation and estimates for two of the ratios.

$$\frac{f_0(a)}{f_0(b)} = \frac{f_0(a)}{f_0(a)/b} = b$$

$$\frac{f_0(b)}{f_0(c)} = \frac{f_0(a)/b}{f_0(a)/c} = c/b$$

The estimated ratio is then

$$\hat{b} = \frac{\hat{f}_0(a)}{\hat{f}_0(b)}$$
$$\widehat{c/b} = \frac{\hat{f}_0(b)}{\hat{f}_0(c)}$$

The theoretical correct value for $\hat{b}$ is b and the theoretical correct value for $\widehat{c/b}$ is c/b. We take the logarithm

$$log(\hat{b}) = log\hat{f}_0(a) - log\hat{f}_0(b)$$
$$log(\widehat{c/b}) = log\hat{f}_0(b) - log\hat{f}_0(c)$$

We estimate the ratio for every pair of dilution D for every dilutions series l where $1 \leq l \leq L$. In the Arabidopsis dilution-dataset there are 4 different dilution series, since we assumed that the triplicates were technical. Thus $L = 4$ in the Arabidopsis dilution-dataset.

If we look at one dilution series at a time, we find the mean value for all estimates of the ratio R as

$$log\hat{R} = \frac{1}{L}\sum_{l=1}^{L}(log\hat{R}^l)$$

and the standard deviation for all estimates of the ratio R as

$$\sqrt{\frac{1}{L-1}\sum_{l=1}^{L}(log\hat{R}^l - log\hat{R})^2}$$

4.6 Parameter estimation for sets of replicate curves

We now look at sets of replicate curves ($n = 3$) in the MLE estimation in Equation (3.11). We estimate three starting concentrations f_{01}, f_{02}, f_{03}, one α and one β for each set of triplicates. We have 16 sets of estimates.

In Figure 4.11 and 4.12 we see the 16 estimates of α and β, and they still vary between gene, groups and dilutions. In Table 4.3 we see the estimated dilution factors when $n = 3$ in the MLE. The estimated ratios are closer to the true value than for $n = 3$. Most of the estimates have lower bias and lower standard deviation.

	$log_2(4/1)$	$log_2(16/1)$	$log_2(64/1)$	$log_2(16/4)$	$log_2(64/4)$	$log_2(64/16)$
True value	2	4	6	2	4	2
$\hat{ratio}$	2.08(0.06)	4.02(0.1)	6.07(0.09)	1.94(0.15)	3.99(0.12)	3.03(1.98)

Table 4.3: Estimated log_2 of ratios corresponding to the dilutions in the $Arabidopsisdilution - dataset$ using the $Enzymologicalmethod$ for each triplicate set. The estimates refer to the mean estimated ratios, with standard deviation in parenthesis.

Based on this we will from now on do parameter estimation for sets of replicate curves. We see that the estimate for the ratio between the two dilution factors 64 and 16 has a high bias, this might be due to pipetting error.

Summary

The analysis performed in this chapter indicates that the Enzymological method is a good estimation method to find the starting fluorescence level in a PCR curve. We have detected some points worth further investigation. We have seen that the p-assumption plot do not give valuable information about the fit of the model to the real data. It is more important to check the assumption of normally distributed noise. We have observed that the last cycle right before the inflection point has the most influence of the parameter estimation. When estimating the model parameters, we use triplicate sets simultaneously, thus $n = 3$.

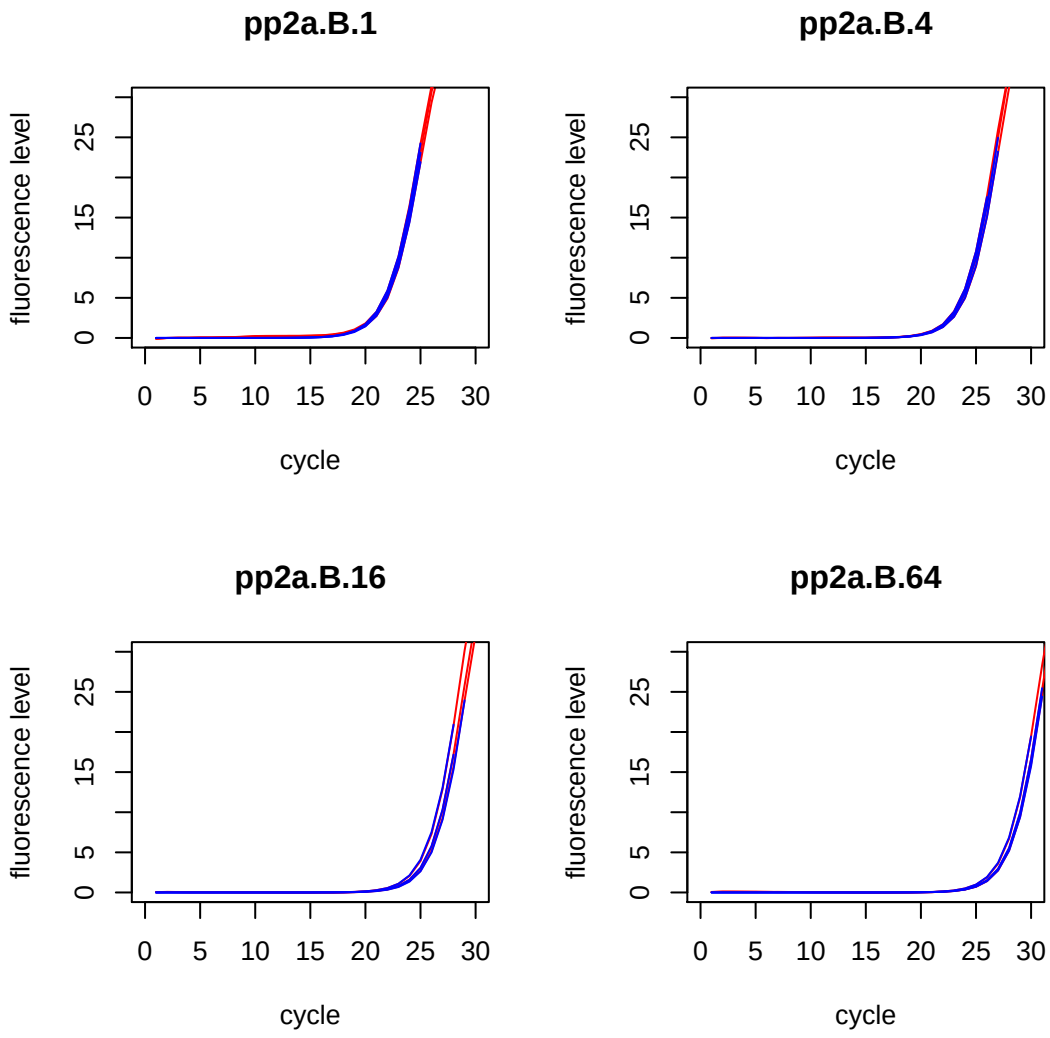

Figure 4.3: A selected section of the observed fluorescence level $(y_1, \ldots, y_{30})$ in red, and $(\hat{f}_1, \ldots, \hat{f}_m)$ in blue with estimated parameters from the Enzymological method. The corresponding triplicates are in the same panel. For illustration we show the curves belonging to PP2A (gene R) for wild-type (group B).

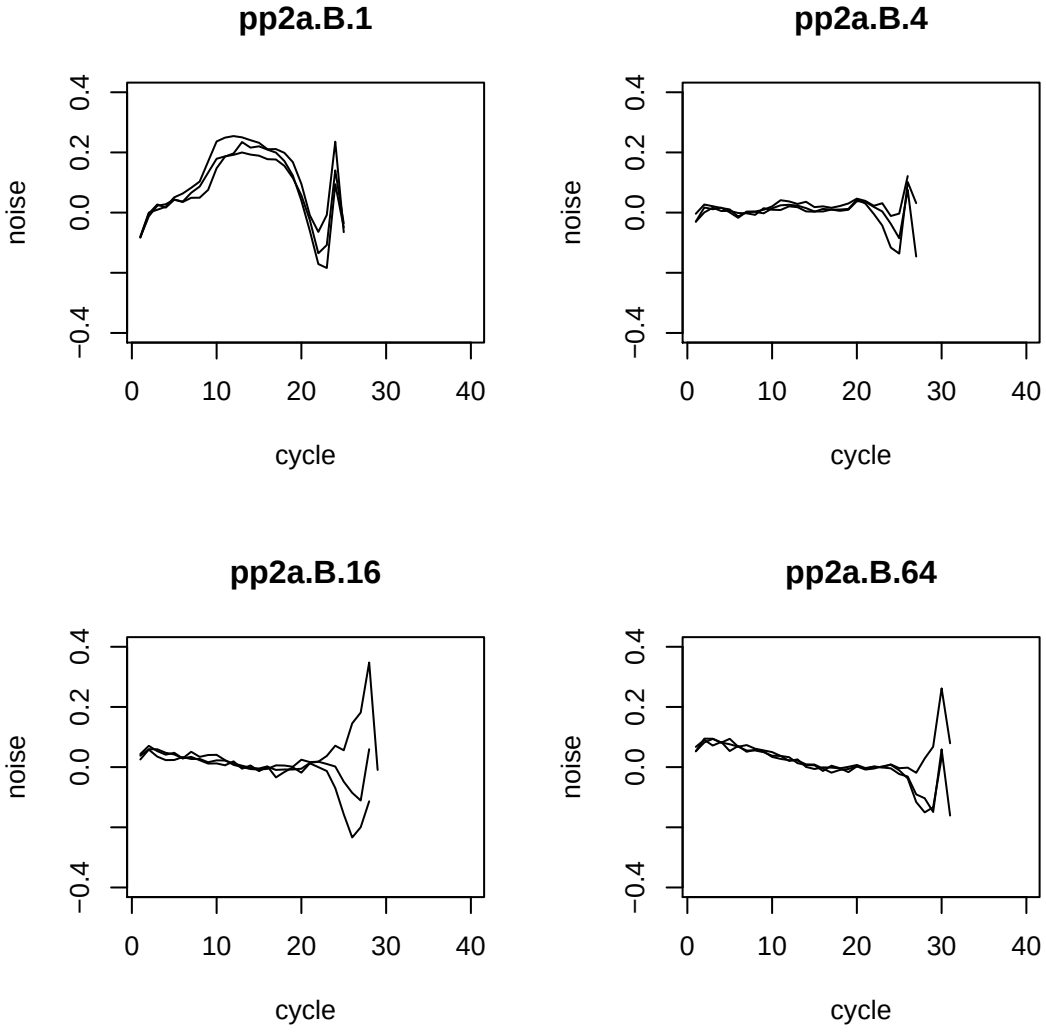

Figure 4.4: The residuals $\hat{\varepsilon} = y - \hat{f}$ where y is observed fluorescence level and $\hat{f}$ is the estimated deterministic fluorescence level from model 3.5 and estimated parameters from the Enzymological method. The corresponding triplicates plotted in the same panel. For illustration we show all dilutions for sample type RB, gene PP2A for the wild-type plant.

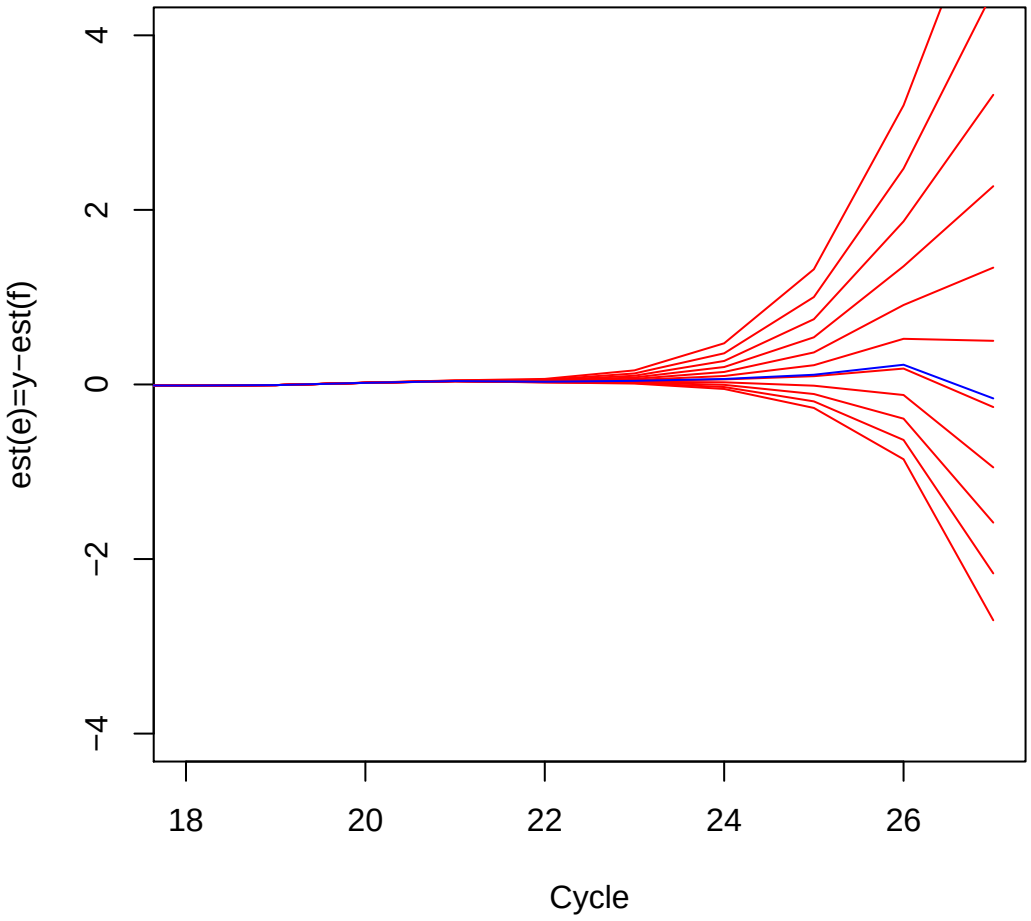

Figure 4.5: Residuals $\hat{\varepsilon}$ for different $\hat{\alpha}$ (from 10 to 30 with steps of 2). The other two estimated parameters in $\hat{f}$ are $\hat{\beta}$ and $\hat{f}_0$ and their values are held constant at 1.06 and $2.84e^{-07}$ respectively. The blue line is the residual $\hat{\varepsilon}$ where $\alpha = 21.73$ which gives the highest likelihood.

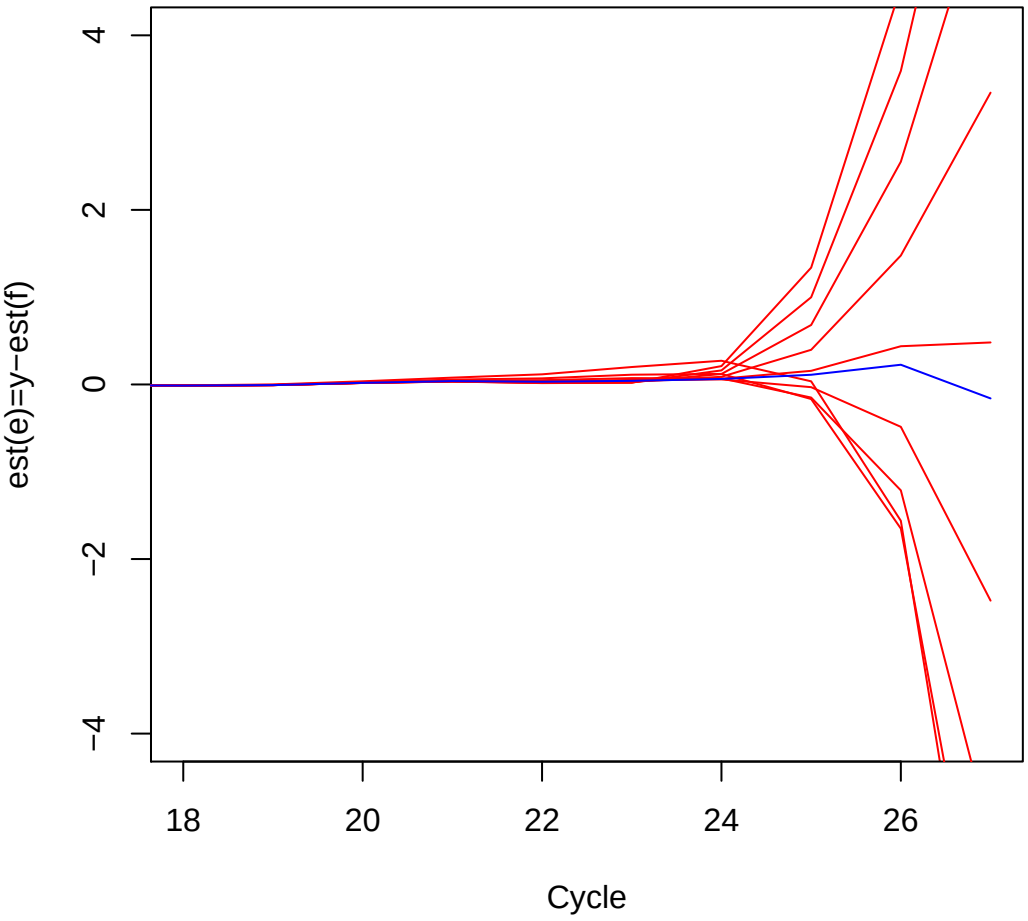

Figure 4.6: Residuals $\hat{\varepsilon}$ for different $\hat{\beta}$ (from 0.3 to 2 with steps of 0.2). The other two estimated parameters in $\hat{f}$ are $\hat{\alpha}$ and $\hat{f}_0$ and their values are held constant at 21.73 and $2.84e^{-07}$ respectively. The blue line is the estimated ε where $\beta = 1.06$ which gives the lowest likelihood.

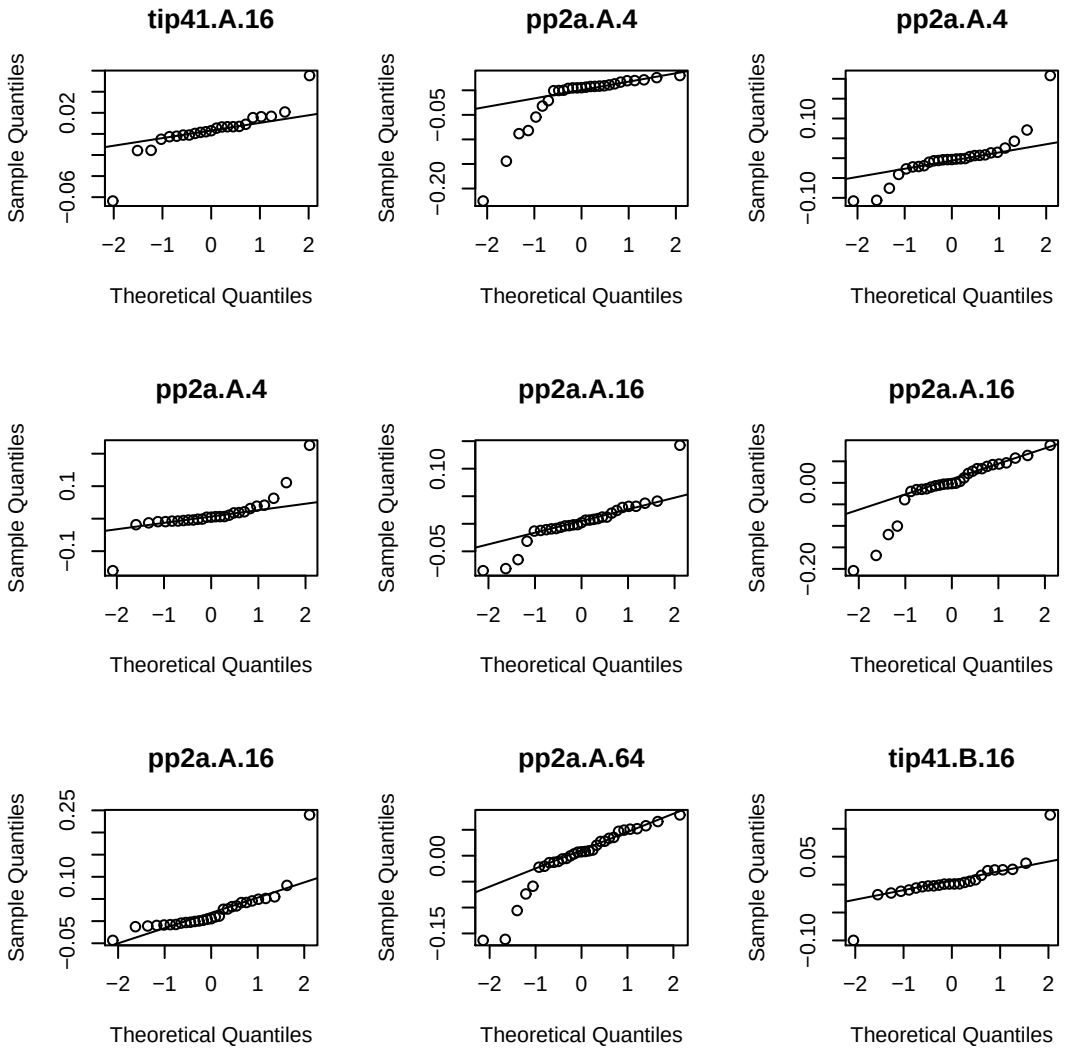

Figure 4.7: The QQ plots for 9 of the 17 curves where the hypothesis of normality in ε was rejected.

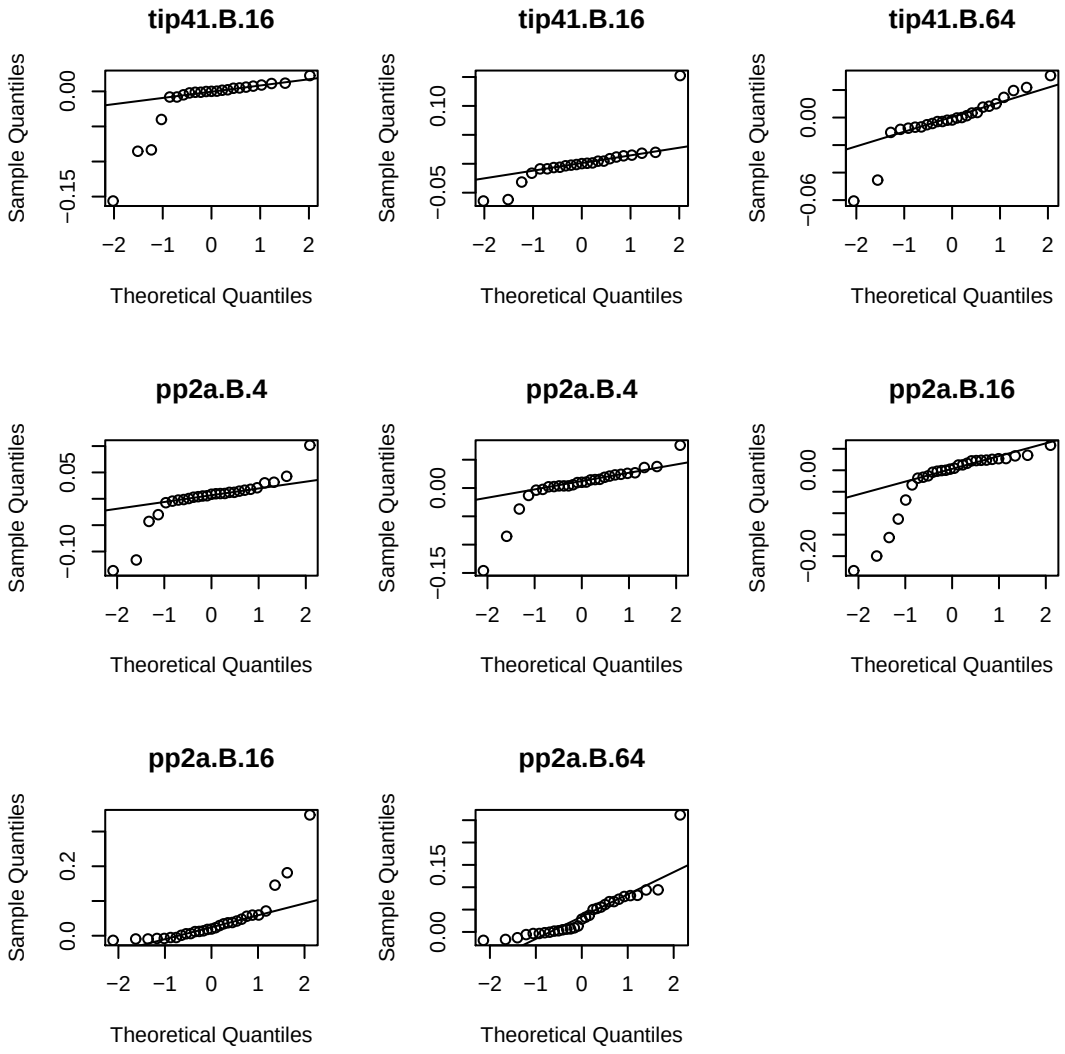

Figure 4.8: The QQ plots for the remaining 8 of the 17 curves where the hypothesis of normality in ε was rejected.

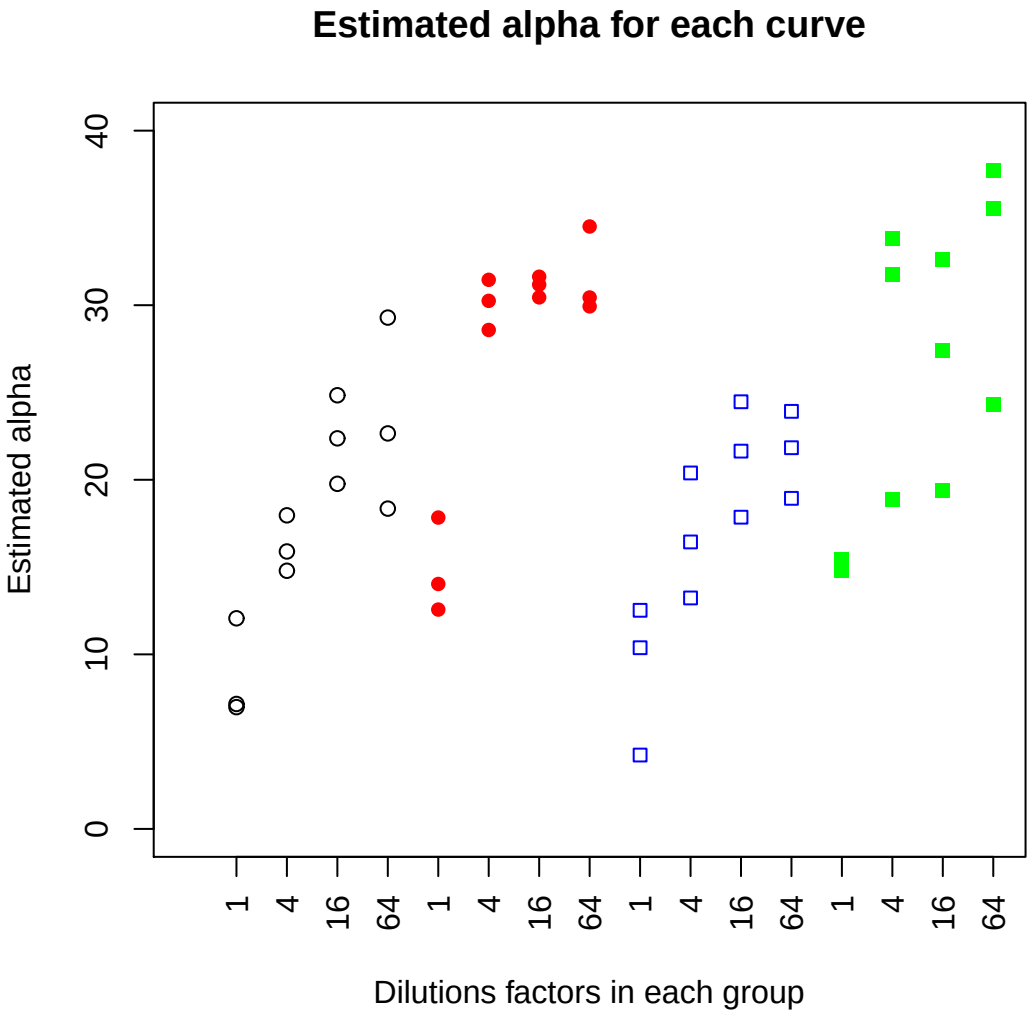

Figure 4.9: Estimation of α for each individual curve in the Arabidopsis dilution-dataset with $3 \times 16 = 48$ samples. The estimates are organized according to gene, group and dilution. Black open circles represent estimates for TIP2 (gene G) for mutant plant (case group A), red filled circles represent estimates for PP2A (gene R) for mutant plant (case group A), blue open squares represent estimates for TIP2 (gene G) for wild-type (control group B) and green filled squares represent estimates for PP2A (gene R) for wild-type (control group B). In each group the x-axis represents the true known dilutions factor in the order 1, 4, 16 and 64.

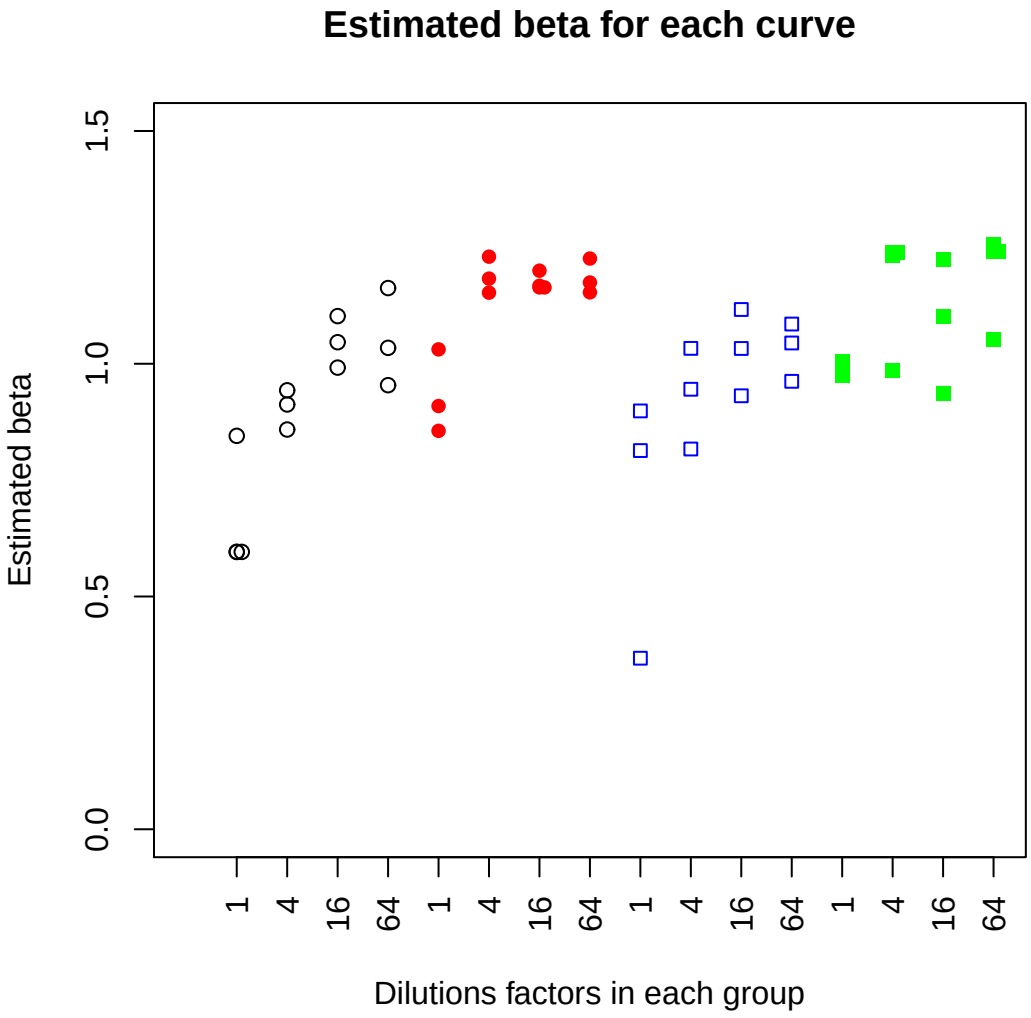

Figure 4.10: Estimation of β for each individual curve in the Arabidopsis dilution-dataset with $3 \times 16 = 48$ samples. The explanation of the plot is the same as in Figure 4.9.

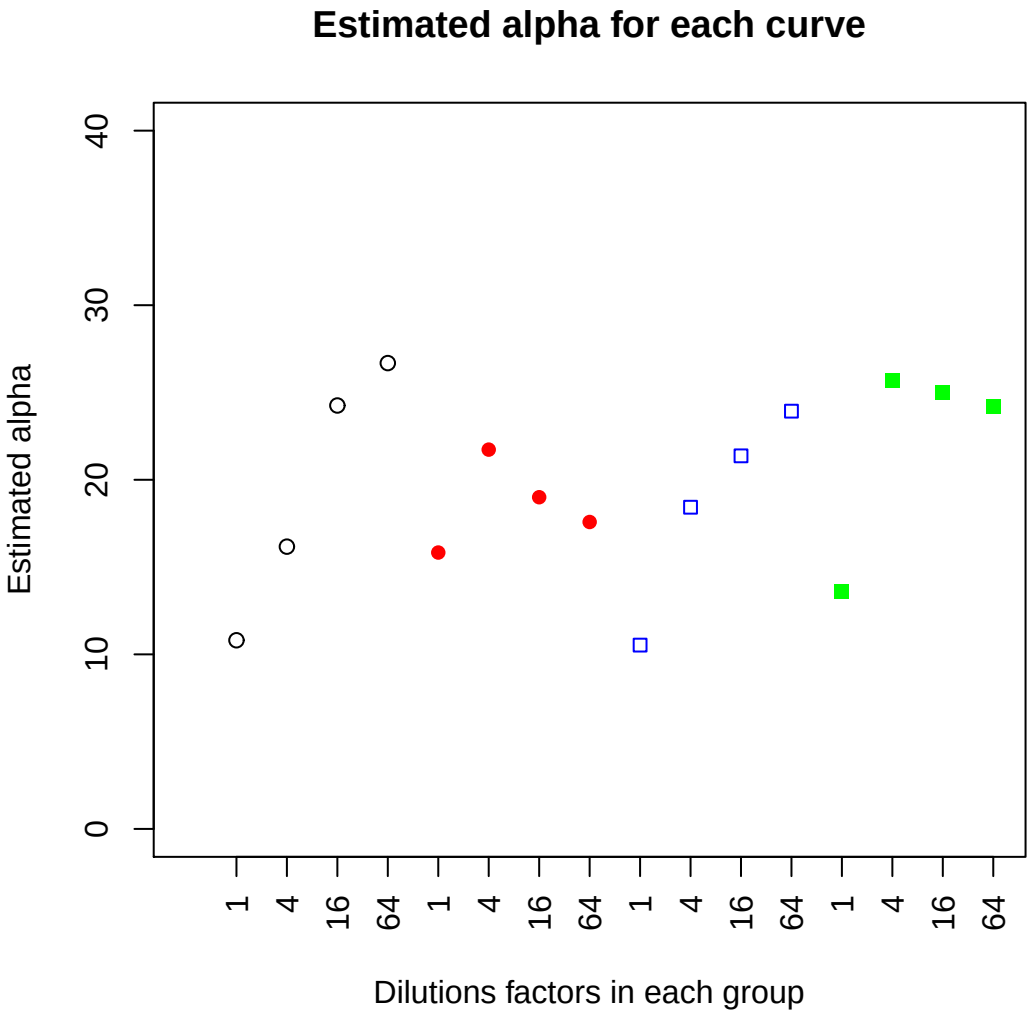

Figure 4.11: Estimation of α for each corresponding triplicate set in the Arabidopsis dilution-dataset, with 16 estimations. The estimates are organized according to gene, group and dilution. Black open circles represent estimates for TIP2 (gene G) for mutant plant (case group A), red filled circles represent estimates for PP2A (gene R) for mutant plant (case group A), blue open squares represent estimates for TIP2 (gene G) for wild-type (control group B) and green filled squares represent estimates for PP2A (gene R) for wild-type (control group B). In each group the x-axis represents the true known dilutions factor in the order 1, 4, 16 and 64.

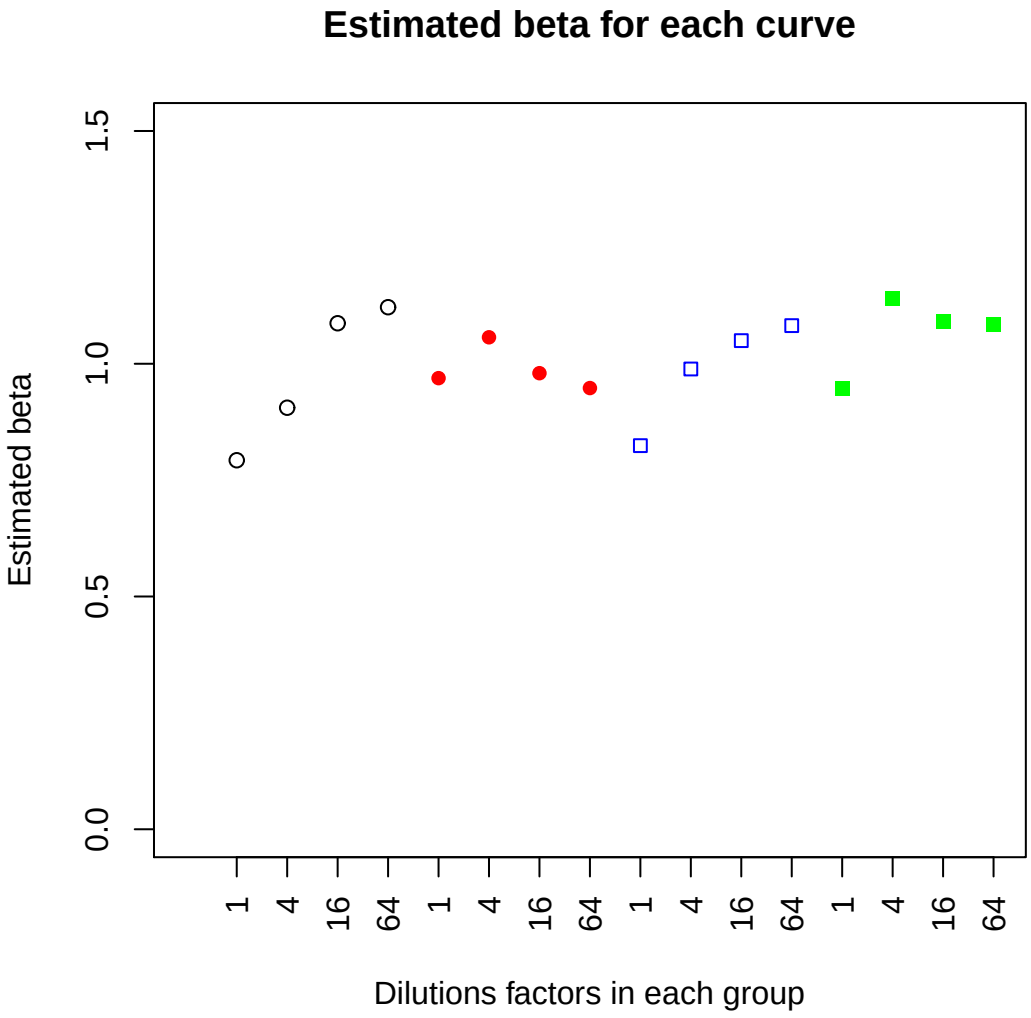

Figure 4.12: Estimation of β for each corresponding triplicate curves in the Arabidopsis dilution-dataset, with 16 estimations. The explanation of the plot is the same as in Figure 4.11.

Chapter 5

Modification of the Enzymological method

In this chapter, we will look at how the Enzymological method works on data, which we assume follows the model. The analyses in this chapter are based on simulated data, where the starting fluorescence level and model parameters are known and we can evaluate the quality of the estimated parameters. We will also suggest new approaches.

5.1 Simulated data

We simulate a dataset of fluorescence levels with parameters σ^2, α, β and f_0 chosen to match parameter estimated from the Arabidopsis dilution-dataset. From Figure 4.9, we choose α to be 20 and from Figure 4.10 we choose $\beta = 1$. To find a value for the starting fluorescence level f_0, we use the fluorescence value in the inflection point y_{mi} for a representative curve i and calculate f_0 recursively, using Equation (3.4). We do not look at cycles after the inflection point, because the mathematical model in Equation (3.4) does not apply for such late cycles. From Figure 4.1, we choose $m = 25$, and the fluorescence level at this cycle to be 30, meaning $f_{25} = 30$. After 25 calculations with Equation (3.4), we find $f_0 = 2.036 \cdot 10^6$. From these true starting parameters $\alpha = 20, \beta = 1$ and $f_0 = 2.036 \cdot 10^6$ we can generate a theoretically correct fluorescence curve i with values $f_{1i}, \ldots, f_{ji}, \ldots, f_{25,i}$ according to Equation (3.4).

The last step to calculate a simulated observed fluorescence curve y_{ji} is to add noise independently to each cycle j as explained in Equation (3.9). The noise ε_{ji} for curve i and cycle j should be normally distributed with variance σ^2.

We estimate σ^2, using the observed fluorescence levels y_{ji} from the Arabidopsis

dilution-dataset

$$\hat{\sigma^2} = \frac{1}{\nu} \sum_{i=1}^{n} \sum_{j=1}^{m_i} (y_{ji} - f_{ji})^2 \tag{5.1}$$

where ν is the number of observations y_{ji} minus the number of estimated parameters, thus $\nu = \sum_{i=1}^{n} m_i - (n+2)$. There is one estimate of f_0 for each curve i where $1 \leq i \leq n$ which adds up to n parameters, plus 2 parameters α and β. In Section 4.4, we concluded that the relation $\varepsilon_{ji} = y_{ji} - f_{ji}$ applies for $1 < j < m$ for curve i. We let $1 < i < (n = 3)$ represent technical triplicates. We calculate one $\hat{\sigma}^2$ for each technical triplicates. In the Arabidopsis dilution-dataset, there are in total 16 triplicates and thus 16 estimates of σ^2. In Table 5.1 we see a summary of these estimates. We see that all the estimates are less than 0.0239, and as a worst

	Mean	Sd	Max
Estimated σ^2	0.0071	0.007	0.0239

Table 5.1: The summary of the 16 estimated σ^2 for each triplicate in the Arabidopsis dilution-dataset from start cycle 1 to the inflection point.

case scenario we choose the true parameter $\sigma^2 = 0.025$ in our simulations.

A list of the true starting parameters in the simulated dataset is found in Table 5.2.

	σ^2	f_0	α	β
true parameters	0.025	$2.036 \cdot 10^{06}$	20	1

Table 5.2: The four true parameters used in the simulated dataset, based on the Arabidopsis dilution-dataset.

We simulate 7500 observed fluorescence curves, representing 2500 technical triplicates. All of these curves will be based on the same true α and β, but we will still let $n = 3$ to evaluate the behavior on the Enzymological method for triplicates. Since all the graphs are based on the same true starting fluorescence level, they are comparable without baseline correction.

In Figure 5.1, we see an example of a simulated fluorescence curve which follows the model from Equation (3.10) with parameters from Table 5.2. For early cycles, the fluorescence values have rapid fluctuations as for real fluorescence curves. The simulated curve ends in the inflection point and it never gets concave, because the curves follow the mathematical model in the Enzymological method which only applies up to the inflection point. We see the similarity between this simulated fluorescence curve in Figure 5.1 and a true fluorescence curve in Figure 2.1.

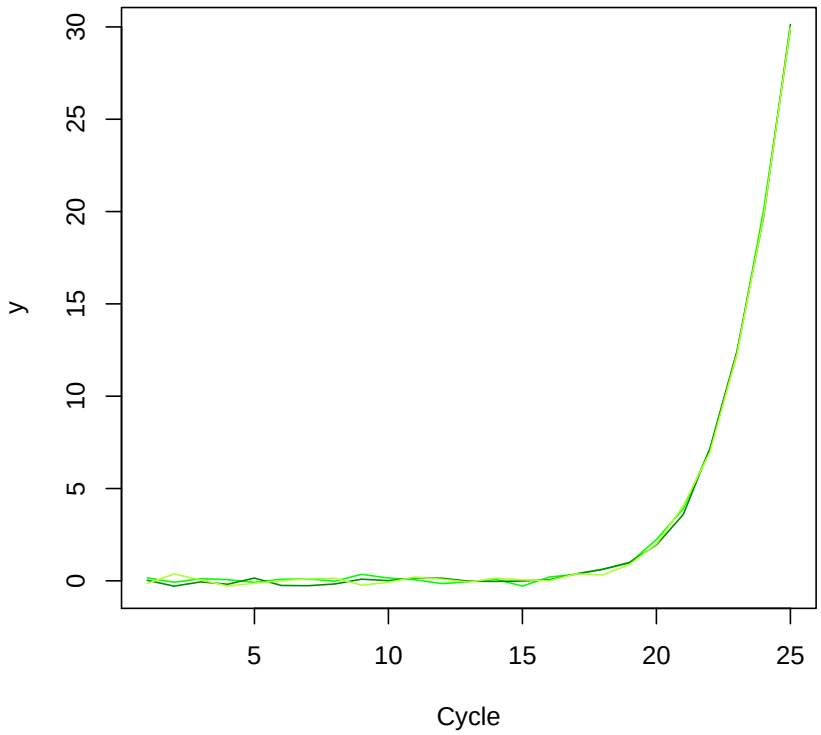

Figure 5.1: Three simulated fluorescence level y_{ji} following the model from Equation (3.10) with parameters from Table 5.2.

5.2 Results

Now we turn to parameter estimation. For each triplicate set $\{y_{ji}\}$ where $1 \leq i \leq 3$ and $s_i \leq j \leq m_i$, we estimate parameter α, β, f_{0i}. We select the part of the fluorescence curve from s_i to m_i, using the procedures explained in Section 3.2. One would think that data from all cycles would be included in the theoretically correct simulated dataset y_{ji}. However, the curve $y_{ji} = f_{ji} + \varepsilon_{ji}$ varies a lot for small j when the error ε_{ji} is normally distributed with $\sigma^2 = 0.025$ and the fluorescence level f_{ji} is less than 10^{-6}. Thus we can not see the original trend in f_{ji} for early cycle. By using the procedure from Section 3.2, the number of cycles (s_i, m_i) is around 7 or 8. The curves include from 5 to 17 cycle in the MLE.

In Table 5.3, we see the results for the 2500 estimated σ^2 from our simulated dataset and observe that the mean of the estimated σ^2 is close to the true value

$\sigma^2 = 0.025$.

	Mean	Sd	Max
Estimated σ^2	0.0261	0.005	0.0617

Table 5.3: Summary statistics for the 2500 estimated σ^2 for each triplicate in the simulated data from start cycle 1 to the inflection point, m_i.

We estimate $\hat{\alpha}$, $\hat{\beta}$ for the 2500 triplicate curves, and calculate the mean value $\hat{f}_0$ over triplicates. We choose to consider the mean value of fluorescence level in the corresponding triplicates since from a biological view these value should be identical. In Chapter 6, we will analyze triplicate samples, but in this section we focus on the mean value $\hat{f}_0$. For estimate $\hat{\theta}$ with true value θ, the results are evaluated by using estimated observed bias $\widehat{Bias(\hat{\theta})}$, the observed mean square error $MSE(\hat{\theta})$ and the coefficient of variance $CV(\hat{\theta})$

$$\widehat{Bias(\hat{\theta})} = \widehat{E(\hat{\theta})} - \theta = \frac{1}{2500}\sum_{l=1}^{2500}\hat{\theta}_l - \theta,$$

$$MSE(\hat{\theta}) = sd(\hat{\theta})^2 + \widehat{Bias(\hat{\theta})}^2 = \sqrt{\frac{l}{2500-1}\sum_{l=1}^{2500}(\hat{\theta_k} - \bar{\theta})^2} + (\widehat{E(\hat{\theta})} - \theta)^2,$$

$$CV(\hat{\theta}) = \frac{sd(\hat{\theta})}{\widehat{|E(\hat{\theta})|}}.$$

Coefficient of Variation (CV) has a value between zero and infinity. A small CV value is preferable, meaning a small standard deviation compared to the mean value. When $CV(\hat{\theta})$ is larger than 1 the standard deviation is greater than the mean value.

In Table 5.4, we see the summary of statistics for the three estimated parameters from the simulated dataset. The $\hat{\alpha}$ has a standard deviation of around 15% of the mean value. Our primary aim is to estimate the starting fluorescence level and it looks like the estimates of f_0 are more accurate. All in all the estimates are very good, but this is not surprising, since the simulated data is based on data that perfectly follow the mathematical model.

In Figure 5.2, we see a pairs plot of the estimated parameters to demonstrate the dependence between the estimates. We see that $\hat{f}_0$, $\hat{\alpha}$ and $\hat{\beta}$ are highly dependent, in an almost linear fashion.

	$\widehat{Bias(\hat{\theta})}$	$MSE(\hat{\theta})$	$CV(\hat{\theta})$
$\hat{f}_0$	3.34e-09	3.24e-15	2.79e-02
$\hat{\alpha}$	1.41e-01	9.39	1.52e-01
$\hat{\beta}$	-2.56e-03	2.62e-03	5.13e-02

Table 5.4: Summary statistics of the three estimated parameters from the simulated dataset with the Enzymological method. There are 2500 estimates of each of f_0, α and β.

5.3 Changing the initial values in the MLE

We would like to investigate the robustness of the MLE to the initial value. From Figure 5.2, we see that all the three parameters are highly correlated, thus we choose to only change the initial value for α. We set α^{init} equal to 10 and 30 which is far from the true value 20. In Table 5.5, we see the results for the estimated parameters with $\alpha^{init} = 10$ and in Table 5.6 we see the results for the estimated parameters with $\alpha^{init} = 30$.

With $\alpha^{init} = 10$, we get a mean value for $\hat{f}_0$ equal to $4.10 \cdot 10^6$ and with $\alpha^{init} = 30$ we get a mean value for $\hat{f}_0$ equal to $1.84 \cdot 10^6$. In both cases, the CV value for $\hat{f}_0$ is higher compared to Table 5.4. The bias of $\hat{f}_0$, estimated when using α^{init} from Equation 3.14, is in order of magnitude 10^{-9}. This is 1000 times better than for $\alpha^{init} = 10$ and 100 times better than for $\alpha^{init} = 30$. The MLE seems to be sensitive to the choice of α^{init}.

	$\widehat{Bias(\hat{\theta})}$	$MSE(\hat{\theta})$	$CV(\hat{\theta})$
$\hat{f}_0$	2.07e-06	6.71e-12	3.81e-01
$\hat{\alpha}$	-1.36e+01	1.98e+02	5.41e-01
$\hat{\beta}$	-4.93e-01	2.87e-01	4.15e-01

Table 5.5: Summary statistics of the three estimated parameters from the 7500 simulated curves with $\alpha^{init} = 10$

5.4 Methods for finding the starting cycle s

The end cycle m_i is placed near the inflection point on the observed fluorescence curve. We will not investigate this further. However, we want to look at the start cycle s_i and its effect on the MLE in Equation (3.11). Can we find estimates with

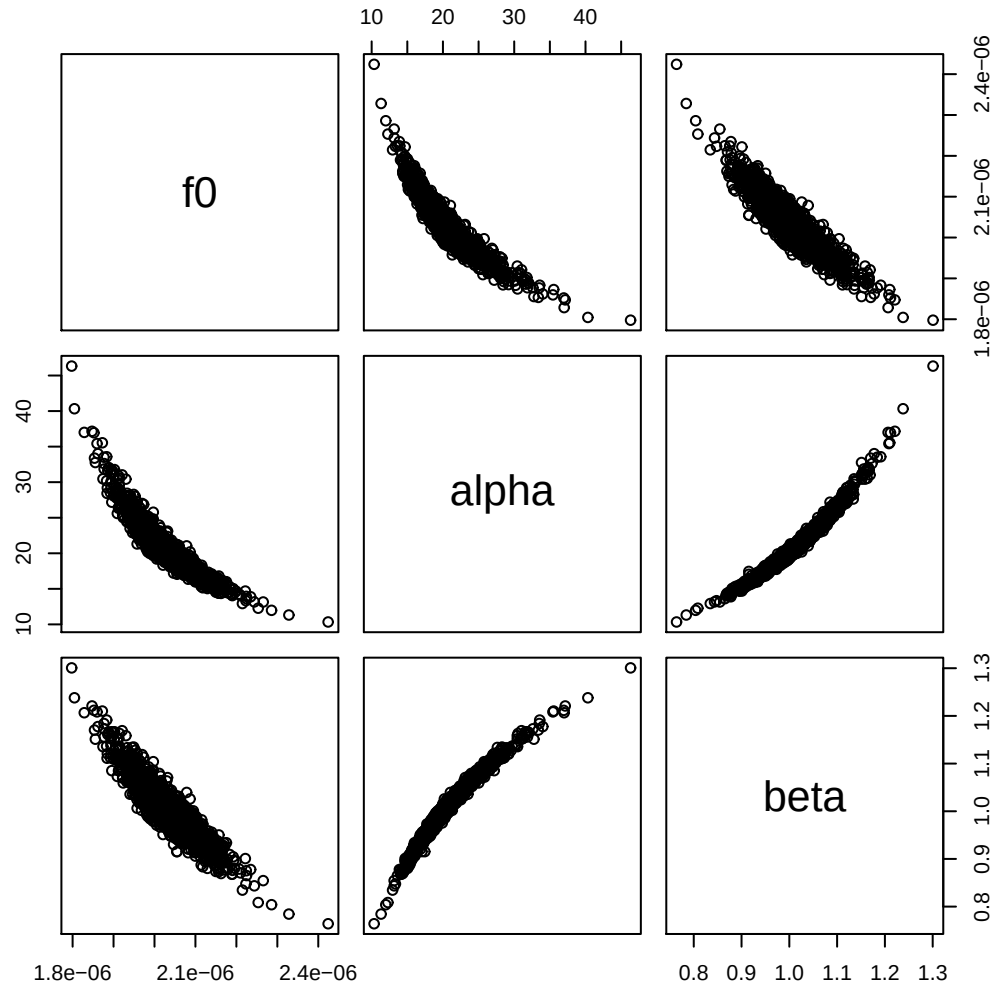

Figure 5.2: Pairs plots of the estimated parameters f_0, α and β from 2500 simulated triplicates when using the Enzymological method to calculate the estimates. There are 2500 estimates of the mean values $\hat{f}_0 = \sum_{i=1}^{3} \hat{f}_{0i}$
calculated over triplicates.

lower bias and variance if we place s later or earlier than the Tichopad approach explained in Section 3.2? What happens to the bias and variation of the estimated starting fluorescence level when we use different number of cycles in the MLE? To investigate this. we first include 2 cycles in the MLE, m and $m-1$. Since we use triplicate curves, we have six observation and five parameters. Next we include

	$\widehat{Bias(\hat{\theta})}$	$MSE(\hat{\theta})$	$CV(\hat{\theta})$
$\hat{f}_0$	-1.97e-07	4.95e-14	5.63e-02
$\hat{\alpha}$	2.36e+01	9.18e+02	4.37e-01
$\hat{\beta}$	2.41e-01	7.95e-02	1.17e-01

Table 5.6: Summary statistics of the three estimated parameters from the 7500 simulated curves with $\alpha^{init} = 30$

cycle $m-2$, such that we have three cycles for each curve. The number of cycles used is increased until 10 cycles are included for each curve. Last we show the results when including 20,21,22 and 23 cycles in each curve. The results are shown in table 5.7.

	$\widehat{Bias(\hat{\theta})}$	$MSE(\hat{\theta})$
2	1.03e-08	7.41e-14
3	2.20e-08	2.75e-14
4	6.74e-09	8.14e-15
5	4.02e-09	4.14e-15
6	9.36e-10	2.44e-15
7	5.91e-09	3.60e-15
8	-3.13e-10	2.83e-15
9	-5.23e-09	3.03e-15
10	4.94e-10	3.30e-15
20	1.45e-09	3.52e-15
21	1.26e-09	3.50e-15
22	1.34e-09	3.50e-15
23	1.38e-09	3.51e-15

Table 5.7: Summary statistics of the estimats of f_0 from the 7500 simulated curves, with different numbers of cycles in the MLE

We see that the bias of $\hat{f}_0$ decreases up to six cycles. The MSE goes from $7.41e^{-14}$ to $2.44e^{-15}$ and the bias goes from $1.03e^{-8}$ to $9.36e^{-10}$. The MSE do not get lower after six cycles even though more cycles are included. All in all the difference is very small, but it looks like there is a small advantage to include 6 or more cycles in this simulated dataset. As mentioned in Chapter 2, we saw in Figure 2.2 the rapid fluctuations in the observed fluorescence values even after baseline correction and this trend cannot be explained by our mathematical model. These rapid fluctuations can be an explanation of the irregular value of the estimated

bias after 6 cycles are included in the MLE. To avoid affecting the model fit, we want to include as few cycles as possible, but still get the lowest bias and variance as possible. In Table (5.7), it looks like six cycles may be the most optimal number.

These simulated curves have all the same starting fluorescence level and we can choose 6 cycles for all of the curves simultaneously. With real data we often have groups with different starting fluorescence level and we have to choose starting cycles individually for each curve. The Tichopad approach chooses the starting cycle by finding the endpoint of the PCR ground phase and ends up using in most cases 7 or 8 cycles. The summary of the number of cycles included in the MLE is found in Table 5.8.

(m-s+1)	5	6	7	8	9	10	11	12	13	14	15	16	17
K	57	717	3532	2679	453	40	7	1	4	2	2	4	2

Table 5.8: The number of occurrences (K) when the number of cycles (m-s+1) is used in the MLE with the Tichopad approach.

We want to see if we can find a method that find s_i for each individual curve i and overall includes fewer cycles than Tichopad approach. Another approach to finding s_i, is to start with estimating the efficiency with Equation (3.13). This estimate is close to 2 up to a certain number of cycles, where the estimated efficiency starts to drop. By fitting a linear regression, we find where this change happens. We investigated this approach and found that it ends up with almost identically starting cycles as the Tichopad approach.

Our next idea is to look at the estimated standard deviation for the noise for each triplicate curve. We call this the Sigma approach. We want to find the cycle where an observed fluorescence curve is relatively large, compared to the noise. We will find a limit where y is relatively larger than $\hat{\sigma}$. To do this, we first use the theoretically correct fluorescence curve f, which follows the mathematical model in Equation (3.3) with parameters motivated by the Arabidopsis dilution-dataset, showed in Table 5.2. Then we find the ratio between $\sigma = 0.025$ and the f curve, $\frac{\sigma}{f}$. This ratio will give us an idea of how large a normal standard deviation is compared to a normal fluorescence curve. In Table 5.9, we see that at cycle 15 the standard deviation of the noise is more than twice the fluorescence level. This indicates that the relative noise is too large for us to find a trend in the fluorescence curve. At cycle 16 we see that the standard deviation of the noise is just a little bigger than the fluorescence level. In cycle 17 the ratio is around 0.5 and the fluorescence curve do not have as highly rapid fluctuations as for early cycles. The inflection point is around cycle 25. There we see that the true standard deviation is only 0.5% of the fluorescence level.

At what ratio should we place the limit σ/f_j? We have earlier seen in this

Cycles j	16	17	18	19	20	21	22	23	24	25
σ/f_j	1.189	0.596	0.30	0.152	0.078	0.041	0.022	0.013	0.008	0.005

Table 5.9: The ratio between the $\sigma = 0.025$ and the theoretically correct fluorescence curve f for each cycle j. The ending cycle m is 25.

Number of cycles	4	5	6	7
Occurrences	89	830	6564	17

Table 5.10: The number of cycles used in the MLE with the Sigma approach

simulated dataset that 6 cycles for each curve give the best results with the MLE. When we include 6 cycles from the ending cycle m at 25, we end up with starting cycle $s = 20$. In Table 5.9 the ratio σ/f_j is 0.078 for starting cycle 20. We set a limit on the fluorescence level where $f^{limit} \cdot 0.1 = \sigma$, meaning that the fluorescence curve should be 10 times the standard deviation of the noise.

In practice, we are dealing with observed fluorescence levels $\{y_{ji}\}$ which include noise. To find the start cycle s_i, we first find the estimate $\hat{\sigma}$ for each triplicate, as explained previously in this chapter. Last we find the observed fluorescence limit $y^{limit} = \hat{\sigma}/0.1$. The starting cycle s_i for curve i is then j where $y_{i,j} > y_i^{limit}$. In Table 5.10, we see the number of cycles found with the Sigma approach. We see that the number of cycles in most cases are 6 and 5 just as expected.

In the Sigma approach compared to the Tichopad approach, the number of cycles are more concentrated around 6 points in the Sigma approach. The results for the estimates of the parameters θ with the Sigma approach can be seen in Table 5.11.

	$\widehat{Bias(\hat{\theta})}$	$MSE(\hat{\theta})$	$CV(\hat{\theta})$
$\hat{f}_0$	-3.92e-10	3.04e-15	2.71e-02
$\hat{\alpha}$	3.17e-01	9.04e+00	1.47e-01
$\hat{\beta}$	7.72e-04	2.52e-03	5.01e-02

Table 5.11: Summary statistics of the three estimated parameters from the simulated dataset with the Sigma approach. There are 2500 estimates of each of f_0, α and β.

Not unexpected the Tichopad approach and the Sigma approach gave quite similar estimates for f_0, but there are some differences.

The Tichopad approach gives good estimates of f_0. It seems like it includes unnecessary many points in the calculation. The algorithm from Tichopad et al. (2003) is very slow. It finds s directly from the observed fluorescence curve.

The Sigma approach gives slightly better estimates of f_0. It includes a suitable amount of cycles in the calculations. It is fast, but one must estimate an additional parameter σ.

5.5 Estimation with the initial values from MLE

Another interesting result is that the initial values in Equation (3.4) are good estimates. As a third method we will use these estimates and call it the Init approach. The estimates of the parameters θ with the Init approach are found in Table 5.12.

	$\widehat{Bias(\hat{\theta})}$	$MSE(\hat{\theta})$	$CV(\hat{\theta})$
$\hat{f}_0$	-8.16e-09	2.24e-15	2.30e-02
$\hat{\alpha}$	1.68e-01	4.29e-01	3.14e-02
$\hat{\beta}$	0	0	0

Table 5.12: Summary statistics of the three estimated parameters from the simulated dataset with the Init approach. There are 2500 estimates of each of f_0, α and β.

All three methods give very similar results, but the Init approach has the lowest MSE and the Sigma approach has the lowest bias. In Figure 5.3, we see boxplots of the 7500 estimated f_0 for each approach.

The $\hat{f}_0$ from the Init approach has low sample variance without using the MLE. In the calculations of all $\hat{f}_0$, we use the $\hat{\alpha}$'s and $\hat{\beta}$'s. We have seen from visual inspection in Figure 5.2 that all three estimates are highly dependent. In Figure 5.4, we see the estimated standard deviation for the $\hat{\alpha}$'s and $\hat{\beta}$'s from the Init approach as red dots, and the estimated standard deviation when we include MLE and increase the numbers of cycles used as black dots. As expected, the estimated standard deviation for the $\hat{\alpha}$'s is small and for the $\hat{\beta}$'s it is zero, since every curve is given the initial value $\beta = 1$. When the $\hat{\beta}$ have low empirical variance, this affect

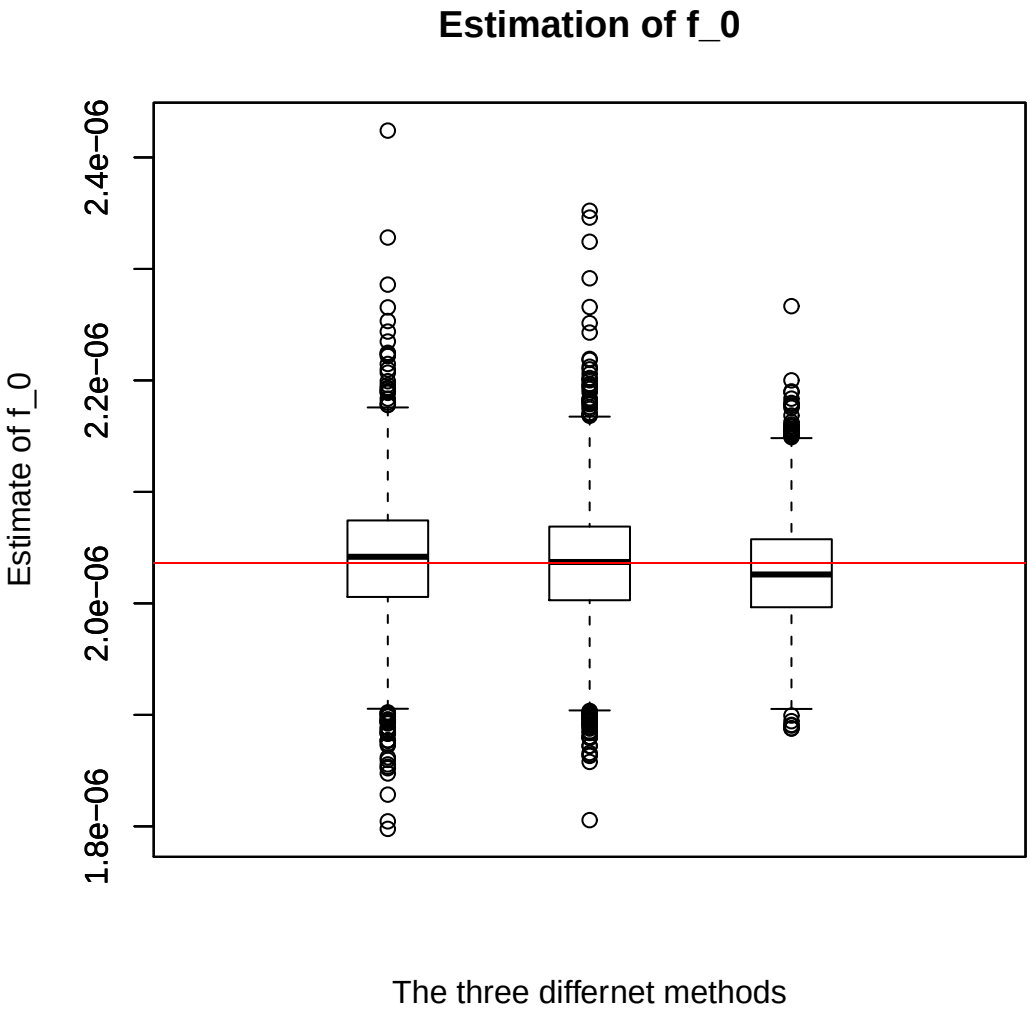

Figure 5.3: Here we see the three different methods used to estimate f_0 from 7500 simulated curves. From the left we see the Tichopad approach, in the middle the Sigma approach and to the right the Init approach. The red line is the true parameter for f_0, equal to $2.036 \cdot 10^{06}$. The boxes represent the lower and upper quartile, and the line in the middle of each box represent the median.

$\hat{\alpha}$ and $\hat{f}_0$ such that their empirical variances also become small. The estimated standard deviation increases when we use 2 cycles in each curve in MLE, but then it decreases until six cycles are included in the MLE calculations. For further inclusions, the estimated variance is stable.

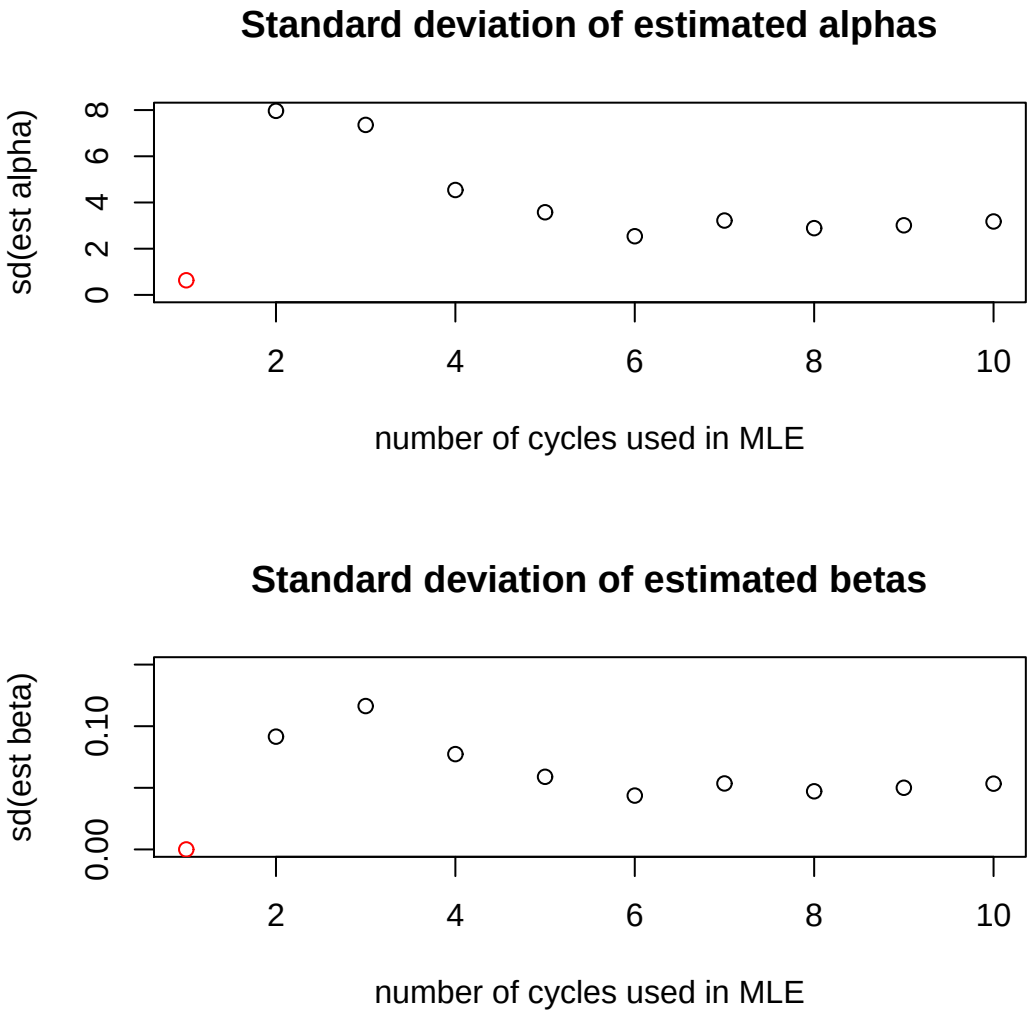

Figure 5.4: In the upper panel we see the estimated standard deviation of the 2500 estimates of α from the simulated dataset. The red dot is the value corresponding to the Init approach. The black dots are the estimated standard deviation when increasing the number of cycles used in the MLE, thus moving the start cycle s_i away from the end cycle m_i. The number of cycles used starts at 2 and are increased to 10. In the lower panel we see the same calculations of the 2500 estimates of β from the simulated dataset.

Important features of the Init approach are as follows

The Init approach finds good estimates for f_0 with low bias and MSE. It has an advantage in the simulated data because it uses additional correct information about β in this simulated dataset, and therefore $\hat{\beta}$ has zero bias and MSE. But this approach may not have this advantage for real data if the true β is not equal to 1.
It is a simple and fast method.

The Init approach used two fluorescence values from each curve y_m and y_{m-1} and the mathematical formula from Equation (3.4). We have chosen m to be the inflection point, but we can let m be other cycles. The comparative CT method and its generalization uses the cycle corresponding to a chosen threshold. Obviously this threshold can also be given variously values. We want to calculate the estimated f_0, where we vary the threshold and m on the simulated data. We know that this data follow the model that Init approach uses, thus this method should give better results. But for which cycle and to what degree is the Init approach superior? What will happen if the thresholds are chosen unwisely? In Figure 5.5, there are boxplot of the 7500 estimated f_0 where m starts at cycle 17 and are moved up to cycle 25, which is the last cycle in the simulated dataset. When $m = 25$, 24, 23 and 22, the estimates are concentrated around the red line, the true value of f_0. The standard deviation increases when m decreases. The $\hat{f}_0$ is closest to the true value when m is at the inflection point as the Init approach uses. In Figure 5.5, the generalized comparative CT method is used to estimate f_0. First the threshold is placed at the fluorescence level corresponding to cycle 17 and then it is moved up stepwise corresponding to one cycle, until cycle 25. In this simulated dataset this method works best when the threshold is placed such that the CT values are close to 19. But it is not as good as the results from the Init approach. We see that placing the threshold value lower or higher will give inferior results.

We have tried these three methods on simulated dataset with true starting parameters that corresponds to the Clusterin dilution-dataset, which will be introduced and analyzed in Section 6.3. The results for the Tichopad approach, Init approach and Sigma approach were consistent and concluded in the same manner as for the Arabidopsis dilution-dataset. This work is not complete. For further work, the bais, the MSE and the CV in simulated data with a wider span of true parameters should be analyzed. At this point we choose to investigate all three methods further on true data.

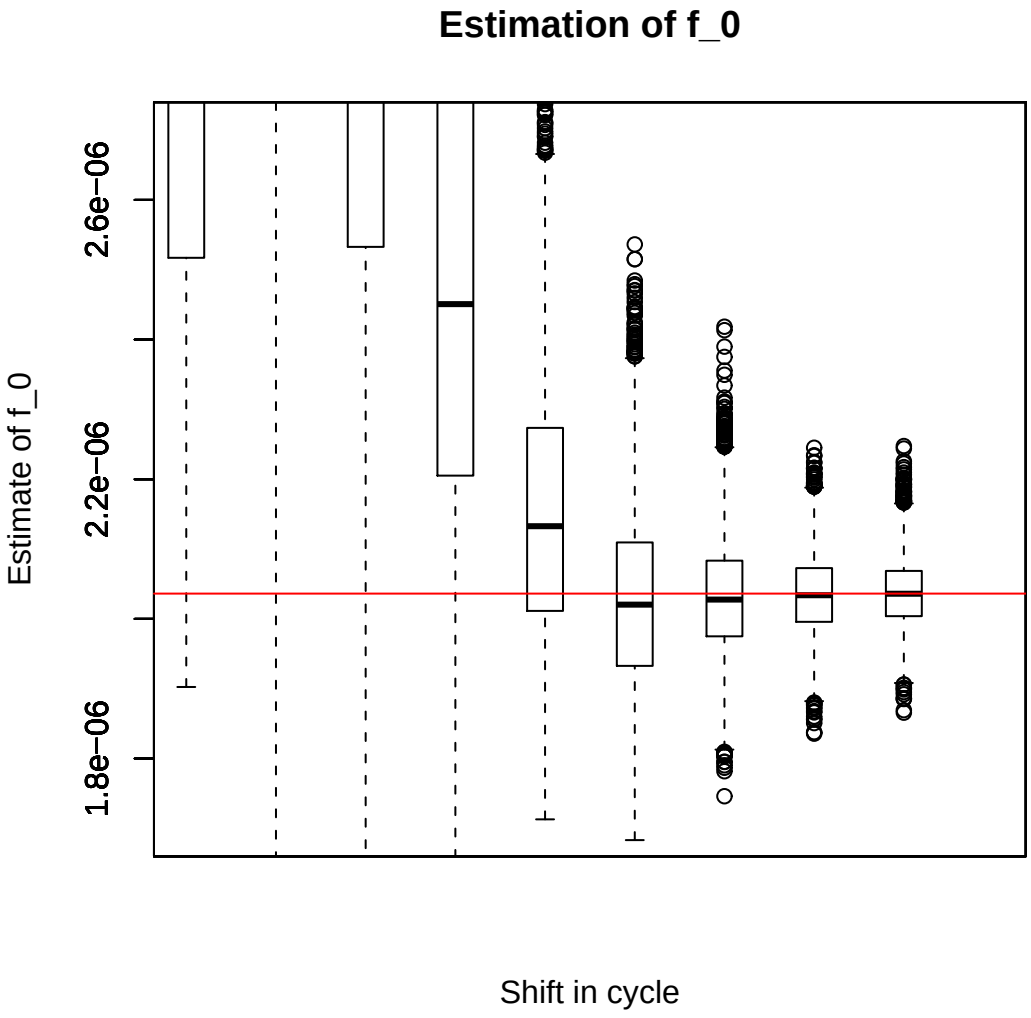

Figure 5.5: Boxplot of 7500 estimated f_0 from the simulated data. The Init approach are used where m starts at cycle 17 (the boxplot to the left) and are moved up to cycle 25 (the boxplot to the right). The red line is the true f_0 value.

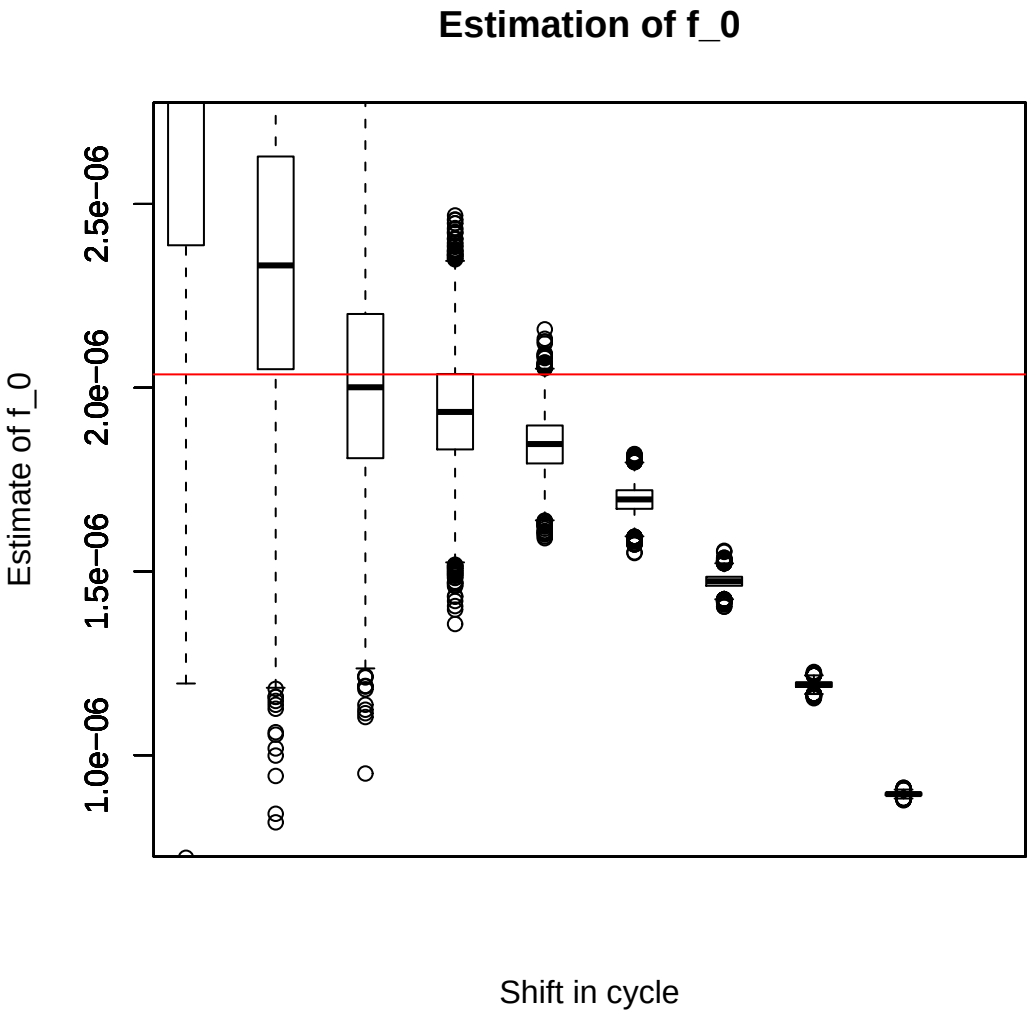

Figure 5.6: Boxplot of 7500 estimated f_0 from the simulated data. The generalized comparative CT method are used where the threshold are placed at the fluorescence level corresponding to cycle 17 (the boxplot to the left) and are moved up to the fluorescence level corresponding to cycle 25 (the boxplot to the right). The red line is the true f_0 value.

Chapter 6

Analysis of dilution datasets

This chapter presents results from five competing methods, called the Tichopad approach, the Sigma approach, the Init approach, the comparative CT method and the generalized CT method. The results of interest are the estimates of the starting concentration level f_0 and the estimated ratios between dilution factors. Two datasets will be analyzed, the Arabidopsis dilution-dataset and the Clusterin dilution-dataset.

6.1 Overview of the five competitive methods

In Tables 6.1 and 6.2, we see an overview of the five methods. The Tichopad approach, Sigma approach and Init approach are all variants of the Enzymological method with cycle dependent efficiency. There is one less parameter in the Init approach, since $\beta = 1$. The Tichopad approach and Sigma approach estimate their five parameters using MLE. The Tichopad approach finds the starting cycle s_i by identifying where the ground phase ends. The Sigma approach finds the starting cycle s_i by estimating the standard deviation of the additive error and finding the cut off when $y_{ji} > \hat{\sigma} \cdot 10$ in each triplicate set. The Init approach finds its four parameters from cycle m_i and $m_i - 1$ in three technical triplicates. The comparative CT method and generalized CT method assume exponential growth with constant efficiency equal to two, and the estimated f_0 is found from one point on the fluorescence curve corresponding to a threshold. In the generalized CT method, we choose a threshold individually for each curve.

6.2 The Arabidopsis dilution-dataset

The Arabidopsis dilution-dataset was introduced in Chapter 4. There are four sample types (GA, RA, GB and RB) and four dilution factors (1, 4, 16, 64). The

	Tichopad approach	Sigma approach	Init approach
Model	Equation (3.4)	Equation (3.4)	Equation (3.3)
PCR efficiency	Cycle dependent	Cycle dependent	Cycle dependent
Parameters in triplicates	f_{01}, f_{02}, f_{03}, α, β	f_{01}, f_{02}, f_{03}, α, β	f_{01}, f_{02}, f_{03}, α
Region on the curve y	$(y_s(thicopad) : y_m)$	$(y_s(sigma) : y_m)$	y_{m-1}, y_m
Estimation	MLE Equation (3.11)	MLE Equation (3.11)	Equation (3.4)

Table 6.1: Notation and important aspects for the three chosen methods based on the Enzymological method.

	comparative CT method	generalized CT method
Model	Exponential curve	Exponential curve
PCR efficiency	Constant = 2	Constant = 2
Parameters in triplicates	f_{01}, f_{02} and f_{03}	f_{01}, f_{02} and f_{03}
Region on the curve y	y_i = threshold T	y_i = threshold T_i
Estimation	$f_{0i} = T \cdot 2^{-CT_i}$	$f_{0i} = T_i \cdot 2^{-CT_i}$

Table 6.2: Notation and important aspects for the two methods based on CT values.

number of PCR curves is 48, where we have 16 sets of triplicates. In the Tichopad approach, the Sigma approach and Init approach we have decided to analyze triplicate curves simultaneously, thus letting $n = 3$. Since the Tichopad approach, the Sigma approach and Init approach are based on the same mathematical models, we choose to use the Tichopad approach to conclude if $n = 3$ can be used in all methods. In Section 4.5, we looked at the variation in estimated parameters α and β from the Tichopad approach for each curve, to found that it was acceptable within triplicate sets. In Section 4.4, we tested for normality in the residuals when the parameters where found with $n = 3$ in the Tichopad approach. We concluded that we could use $n = 3$ in the Tichopad approach. We will also use $n = 3$ in the Init approach and in the Sigma approach.

6.2.1 Cycles used for each method

In Table 6.3, we see summary statistics on the intervals used in the five methods. To see how the methods work the mean value over the triplicates is shown. The number of cycles used in the MLE with the Tichopad approach is mostly seven for the original concentrated samples and around ten for the samples with dilution factor 64. The number of cycles used in the MLE with the Sigma approach is mostly four for the original concentrated samples and around five for the samples

with dilution factor 64. This increasing number of cycles in the more diluted samples can be caused by the free enzymes becoming a limiting factor at a later stage. The inflection point m is placed at later cycles for the smallest concentrated samples. This is natural since the starting fluorescence level is lower. We see the same trend for the CT values in the comparative CT method. From Figure 4.1, we placed the threshold at fluorescence level four, and from linear interpretation we found the corresponding CT values. The thresholds in the generalized CT method are chosen to be where $\hat{E}$ has its value close to two, explained in Section 2.3. The CT values in the comparative CT method are often one or two cycles later than the CT values in the generalized CT method.

	No of Cycles in Thicopad-app.	No of Cycles in Sigma-app.	m	C_T (T=4)	(C_T,T)	at $\hat{E}$
GA.1	6.67	4.33	19.33	16.35	(15.33 , 2.22)	at 1.81
GA.4	9	6	21	18.22	(17 , 1.84)	at 1.91
GA.16	9.67	6.33	23.33	20.32	(19 , 1.73)	at 1.92
GA.64	9.33	6	25.33	22.29	(19.67 , 1.21)	at 1.95
RA.1	7.67	4.67	25	21.83	(20.67 , 2.03)	at 1.84
RA.4	9.67	5	27	23.86	(21.67 , 0.99)	at 1.93
RA.16	10.33	5.33	29	25.74	(24.33 , 1.83)	at 1.92
RA.64	9.67	5	31	27.8	(25.67 , 1.04)	at 1.95
GB.1	9.67	4.67	19.67	16.37	(15.33 , 2.22)	at 1.79
GB.4	9.67	6.33	21.33	18.36	(16.67 , 1.48)	at 1.89
GB.16	9	5.67	23.33	20.28	(18 , 1.07)	at 1.93
GB.64	9.67	6.67	25.33	22.33	(19.33 , 0.62)	at 1.96
RB.1	7.67	4.67	25	21.51	(20.33 , 2.12)	at 1.81
RB.4	9.33	5.67	26.67	23.49	(22 , 1.54)	at 1.92
RB.16	10.67	5	28.33	25.36	(23.33 , 1.35)	at 1.94
RB.64	11.67	5	30.67	27.43	(25 , 0.81)	at 1.97

Table 6.3: Summary statistics for the triplicates of fluorescence curves used in the calculations with the five methods. All values are means over triplicates curves. The names in the left column denote the different sample types and dilutions. In the first column we see the number of cycles used in the MLE with the Tichopad approach, the second column we see the number of cycles used in the Sigma approach, the third column shows the inflection point. The fourth column presents the CT values from the comparative CT method with thresholds equal to 4 for all curves. In the last column we see the CT values corresponding to the threshold and estimated efficiency from the generalized CT method where each curve have individual thresholds.

6.2.2 The f_0 estimates

In Figure 6.1, we see all 48 log transformed $\hat{f}_0$'s. The five methods have different colors, Tichopad approach (black), Sigma approach (red), Init approach (blue), comparative CT method (green) and generalized CT method (yellow). On the x-axis we find the four sample types (GA, RA, GB and RB). We see that the estimated f_0 from the five methods are very similar.

In Figure 6.2, we see six Mean-Difference plots of all 48 estimates of $log_2(\hat{f}_0)$ so we pairwise can compare methods. In each panel, the x-axis is the mean value $(log(\hat{f}_0)^A + log(\hat{f}_0)^B)/2$ from method A and B, and the y-axis is the difference $(log(\hat{f}_0)^A - log(\hat{f}_0)^B)$ within the same two methods A and B. If the methods give the same estimates, the difference are zero and they will follow the red line. If the difference is negative, then $(log(\hat{f}_0)^A < log(\hat{f}_0)^B)$ and method B gives higher estimates of f_0 than method A. The mean values $(log(\hat{f}_0)^A + log(\hat{f}_0)^B)/2$ are negative since the fluorescence levels are less than one. The estimated starting fluorescence levels from the Tichopad approach and the Sigma approach are very similar so we only compare the other methods to the Sigma approach in Figure 6.2. We see that the Sigma approach, the Init approach and the generalized CT method give more similar results, than the other comparison methods. This indicates that the model in the Enzymological method on the interval $(y_0, \ldots, y_m)$ fits the true fluorescence curves as good as the exponential model on the interval $(y_0, \ldots, y_{CT})$ with CT value found from the generalized CT method. In the lower panel to the right, we see that the trend is that $\hat{f}_0^{\,CT} < \hat{f}_0^{\,genCT}$. In the comparative CT method, the $T = 4$ and for the generalized CT method every threshold is places lower than 4. If the efficiency at $T = 4$ were to be close to two, these two methods would have given more similar results. If the efficiency, however, has decreased from two where the CT values are placed, the comparative CT method will give lower estimates of f_0 than the generalized CT method. This is because the comparative CT method assumes higher amplification from y_{CT} and back to the first cycle. This may indicate that the threshold in the comparative CT method is placed too high. All in all the estimated f_0 from the five methods are very similar.

To evaluate the five methods, we will perform linear regression for each method separately. The response is $\log_2(\hat{f}_{0i})^l$. At significance level 0.05 we find, with biological motivation, a significant model

$$\log_2(\hat{f}_{0i})^l = x_i^l \cdot \eta_i^l + \delta_i^l \tag{6.1}$$

for triplicate i and sample l, where $1 < i < 3$ and $1 < l < 16$. The regression coefficients are $\eta_i^l = [\eta_0, \eta_{ST2}, \eta_{ST3}, \eta_{ST4}, \eta_{D2}, \eta_{D3}, \eta_{D4}]$ for sample type RA (ST2), GB (ST3) and RB (ST4) and dilutions 4 (D2), 16 (D3) and 64 (D4), where η_0 is

the intercept. The x_i^l is a vector of 0's and 1's denoting the sample type and the dilution of each sample.

To investigate if the triplicate observations can be regarded as independent, we fit a linear mixed effects model using each of the 16 combinations of sample type and dilution as a random variable. We then calculated the intraclass correlation (ICC), which estimates the correlations between two observations from the same triplicate. The ICC was of order 10^{-5} for all methods. The intraclass correlation is small for all methods so we choose to use the linear regression model in the further analyses.

The Anderson-Darling test shows that we can accept the hypothesis of normally distributed δ's in model (6.1). We assume the error term δ is distributed as $N(0, \nu^2)$. We find the estimated variance ν^2 by the mean squared error for each method to be lowest for the comparative CT method (0.0395). The next methods are the Init approach (0.0410), the Sigma approach (0.0420), the Tichopad approach (0.0432) and the generalized CT method (0.0442). This means the model in Equation (6.1) best explains the estimates from the comparative CT method, but there are not large differences between the methods.

6.2.3 Estimation of the ratio between dilution factors

The true value of the starting fluorescence levels are unknown, but we do know the value of the ratios between dilution factors. In Table 6.4, we see the estimated log transformed ratios calculated from the mean value of the triplicates, as explained in Section 4.5. We see that all the methods give estimates close to the true value. In most cases, the generalized CT method and the Init approach have lower bias.

	Tichopad	Sigma	Init	comparativ CT	generalized CT
log(4/1)=2	2.08(0.059)	1.979(0.062)	2.032(0.067)	1.969(0.07)	2.016(0.045)
log(16/1)=4	4.022(0.1)	3.919(0.019)	3.99(0.095)	3.911(0.049)	3.956(0.072)
log(64/1)=6	6.067(0.086)	5.95(0.021)	6.035(0.053)	5.95(0.023)	5.975(0.026)
log(16/4)=2	1.942(0.147)	1.94(0.077)	1.957(0.128)	1.943(0.107)	1.94(0.105)
log(64/4)=4	3.987(0.119)	3.971(0.042)	4.003(0.094)	3.981(0.066)	3.959(0.036)
log(64/16)=2	3.031(1.985)	3.012(1.963)	3.054(2.004)	3.002(1.945)	2.987(1.984)

Table 6.4: Results for the estimated ratios between dilutions in the Arabidopsis dilution-dataset with all five methods. The calculations of the mean value and the standard deviation in paranthesis, is explained in Section 4.5.

We also estimate the ratios between dilution factors by calculating the contrasts between the dilution factors from the linear regression model in Equation (6.1).

When estimating a ratio, we look at the starting fluorescence level in a sample from the same gene and group, but with different dilutions. For instance, the estimated log transformed ratio between dilution factor number two and dilution factor number four is

$$\widehat{log_2(\hat{f}_0)}_{dil4} - \widehat{log_2(\hat{f}_0)}_{dil64} = \hat{\eta}_0 + \hat{\eta}_{D2} - (\hat{\eta}_0 + \hat{\eta}_{D4})$$

with true value

$$log_2(f_0/4) - log_2(f_0/64) = log_2(64/4) = 4.$$

Since we have no interactions between sample type and dilution, this estimate is the same for all sample types.

These estimated ratios from the Arabidopsis dilution-dataset for each of the five methods are found in the five Tables 6.5 (Tichopad approach), 6.6 (Sigma approach), 6.7 (Init approach), 6.8 (comparative CT method) and 6.9 (generalized CT method). We perform a hypothesis test that the ratio between dilution factors are equal to their true value. After using a two sided t-test we get a p-value larger than significance level 0.05, and the true bias (difference between the estimated ratios minus the true value) equal to zero lies within the 95% confidence interval for all methods. For every method, we accept the hypothesis that the ratio is equal to its true value.

	True	Estimate	Std. Error	t value	DF	Pr(>|t|)	Lower.CI	Upper.CI
D1-D2	2.00	2.07	0.08	0.88	41.00	0.39	-0.10	0.25
D1-D3	4.00	4.03	0.08	0.33	41.00	0.74	-0.14	0.20
D1-D4	6.00	6.07	0.08	0.79	41.00	0.44	-0.10	0.24
D2-D3	2.00	1.95	0.08	-0.54	41.00	0.59	-0.22	0.13
D2-D4	4.00	3.99	0.08	-0.09	41.00	0.93	-0.18	0.16
D3-D4	2.00	2.04	0.08	0.46	41.00	0.65	-0.13	0.21

Table 6.5: Estimated ratios between dilutions in the Arabidopsis dilution-dataset from the Tichopad approach with a 95% confidence interval.

In Figure 6.3, we see a boxplot of all estimated ratios between dilution factors minus their true value. It is hard to decide which method is the best.

	True	Estimate	Std. Error	t value	DF	Pr(>|t|)	Lower.CI	Upper.CI
D1-D2	2.00	2.03	0.08	0.32	41.00	0.75	-0.14	0.20
D1-D3	4.00	4.00	0.08	-0.05	41.00	0.96	-0.17	0.16
D1-D4	6.00	6.04	0.08	0.42	41.00	0.67	-0.13	0.20
D2-D3	2.00	1.97	0.08	-0.37	41.00	0.71	-0.20	0.14
D2-D4	4.00	4.01	0.08	0.10	41.00	0.92	-0.16	0.18
D3-D4	2.00	2.04	0.08	0.48	41.00	0.64	-0.13	0.21

Table 6.6: Estimated ratios between dilutions in the Arabidopsis dilution-dataset from the Sigma approach with a 95% confidence interval.

	True	Estimate	Std. Error	t value	DF	Pr(>|t|)	Lower.CI	Upper.CI
D1-D2	2.00	1.97	0.08	-0.31	41.00	0.76	-0.19	0.14
D1-D3	4.00	3.93	0.08	-0.90	41.00	0.38	-0.24	0.09
D1-D4	6.00	5.95	0.08	-0.60	41.00	0.55	-0.22	0.12
D2-D3	2.00	1.95	0.08	-0.58	41.00	0.56	-0.22	0.12
D2-D4	4.00	3.98	0.08	-0.28	41.00	0.78	-0.19	0.14
D3-D4	2.00	2.02	0.08	0.30	41.00	0.77	-0.14	0.19

Table 6.7: Estimated ratios between dilutions in the Arabidopsis dilution-dataset from the Init approach with a 95% confidence interval.

	True	Estimate	Std. Error	t value	DF	Pr(>|t|)	Lower.CI	Upper.CI
D1-D2	2.00	1.97	0.08	-0.39	41.00	0.70	-0.20	0.13
D1-D3	4.00	3.91	0.08	-1.09	41.00	0.28	-0.25	0.08
D1-D4	6.00	5.95	0.08	-0.62	41.00	0.54	-0.21	0.11
D2-D3	2.00	1.94	0.08	-0.71	41.00	0.48	-0.22	0.11
D2-D4	4.00	3.98	0.08	-0.23	41.00	0.82	-0.18	0.15
D3-D4	2.00	2.04	0.08	0.48	41.00	0.64	-0.13	0.20

Table 6.8: Estimated ratios between dilutions in the Arabidopsis dilution-dataset from the comparative CT method with a 95% confidence interval.

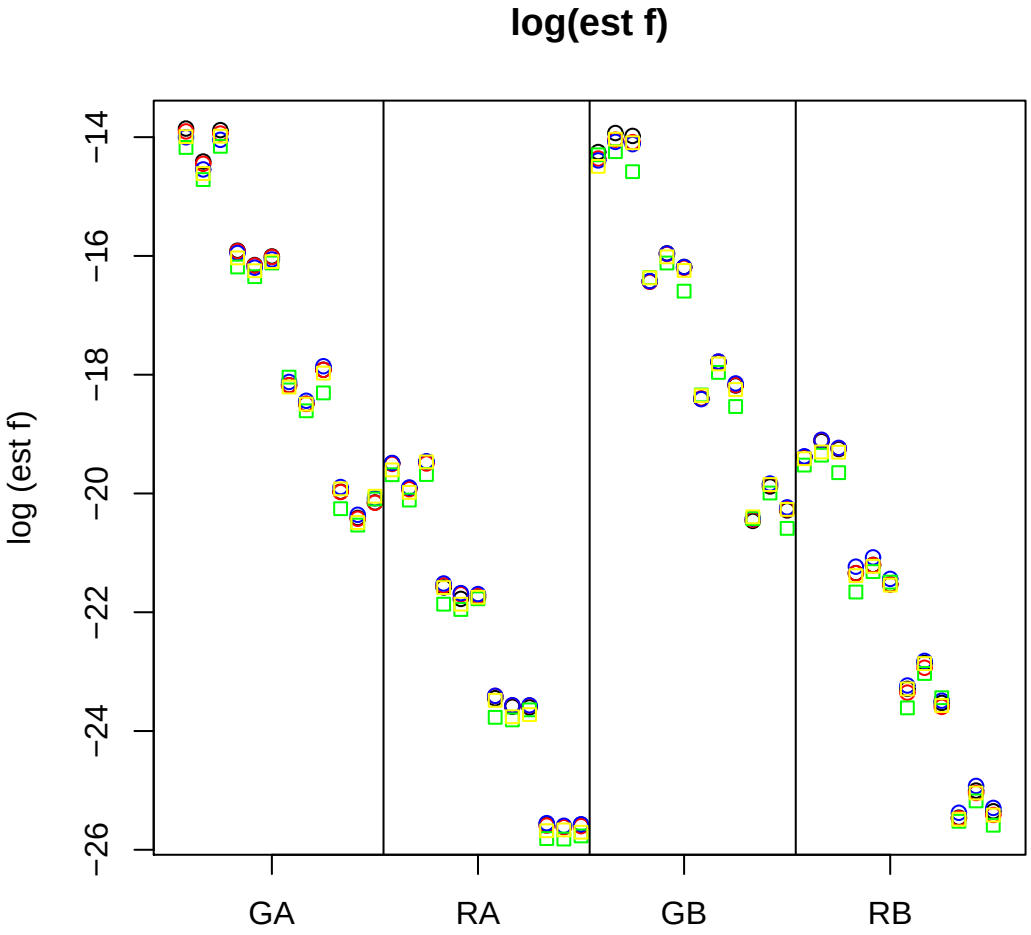

Figure 6.1: Plot of $log_2(\hat{f}_0)$ for the Tichopad approach (black circles), the Sigma approach (red circles), the Init approach (blue circles), the comparative CT method (green squares) and the generalized CT method (yellow squares). On the x-axis are the four sample types (GA, RA, GB and RB) organized by increasing dilution factor.

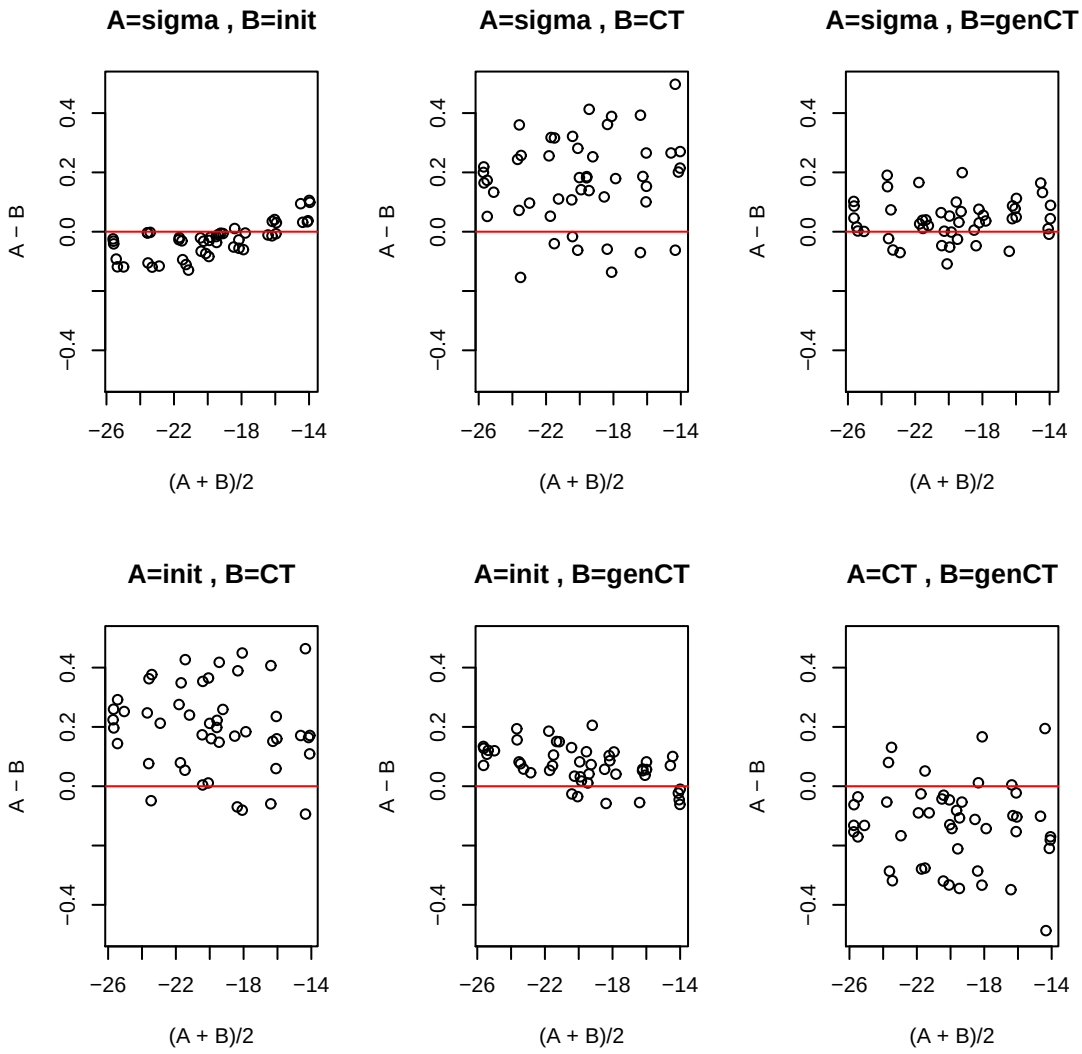

Figure 6.2: Mean-Difference plots of all 48 estimates of $log(\hat{f}_0)$ for comparing pairs of methods. In each panel we see on the x-axis the mean value $(log(\hat{f}_0)^A + log(\hat{f}_0)^B)/2$ from a method A and B, and on the y-axis we see the difference $(log(\hat{f}_0)^A - log(\hat{f}_0)^B)$ from the same two methods A and B. If the methods give the same estimates they will follow the red line.

	True	Estimate	Std. Error	t value	DF	Pr(>|t|)	Lower.CI	Upper.CI
D1-D2	2.00	2.01	0.09	0.07	41.00	0.94	-0.17	0.18
D1-D3	4.00	3.96	0.09	-0.48	41.00	0.63	-0.21	0.13
D1-D4	6.00	5.97	0.09	-0.31	41.00	0.76	-0.20	0.15
D2-D3	2.00	1.95	0.09	-0.55	41.00	0.58	-0.22	0.13
D2-D4	4.00	3.97	0.09	-0.38	41.00	0.70	-0.21	0.14
D3-D4	2.00	2.01	0.09	0.17	41.00	0.87	-0.16	0.19

Table 6.9: Estimated ratios between dilutions in the Arabidopsis dilution-dataset from the generalized CT method with a 95% confidence interval.

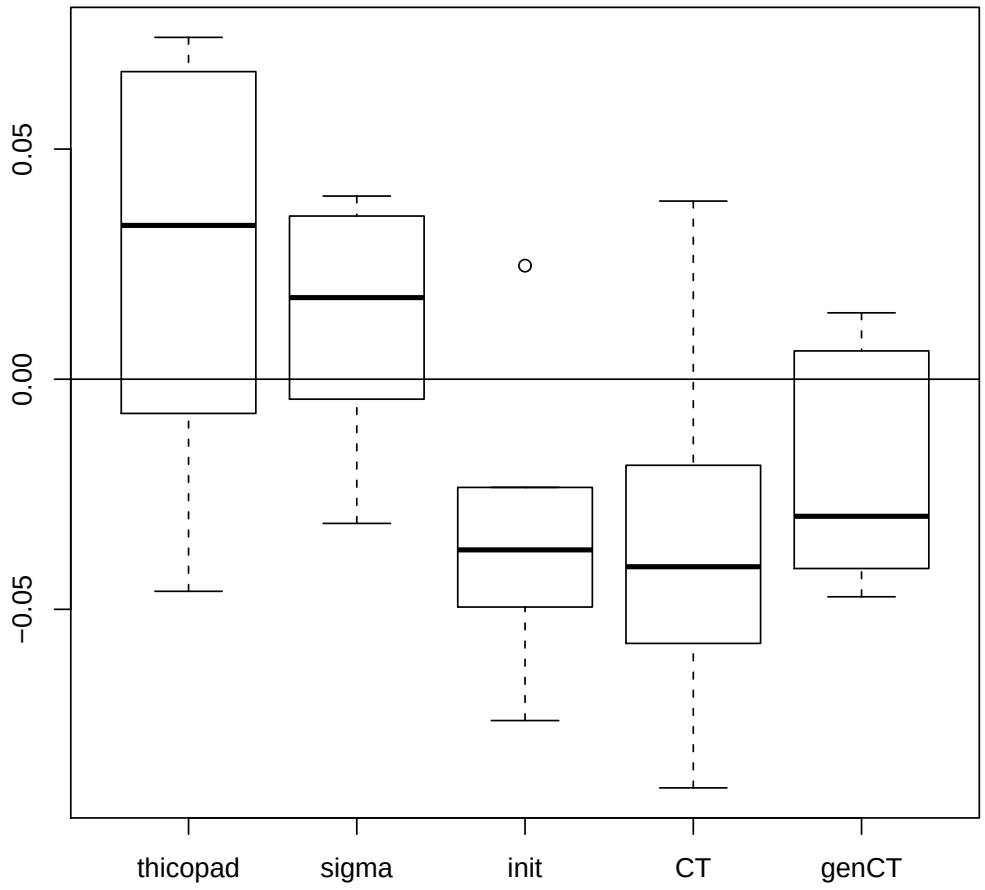

Figure 6.3: Boxplot over all estimates of the ratios minus the true value. Each boxplot is based on six estimated ratios between dilution factors.

6.3 The Clusterin dilution-dataset

6.3.1 Description of the dataset

In the Clusterin dilution-dataset, the experiments are performed on one gene called Clusterin. In the PCR runs, four different primer pairs were used called Clu, Clu1, Clu2 and Clu3, which we will call sample types. The sample types are diluted with dilution factors 1, 10, 100, 1000 and 10000. In Table 6.10, we see an overview of the sample types and the dilutions. There are technical triplicates for each of the sample types and dilution combinations. In total, there are four sample type times five dilutions time three replicates, which is equal to 60 fluorescence curves.

	Dilutions				
Sample type	1:1	1:10	1:100	1:1000	1:10000
Clu	f_0^{Clu}	$f_0^{Clu}/10$	$f_0^{Clu}/100$	$f_0^{Clu}/1000$	$f_0^{Clu}/10000$
Clu1	f_0^{Clu1}	$f_0^{Clu1}/10$	$f_0^{Clu1}/100$	$f_0^{Clu1}/1000$	$f_0^{Clu1}/10000$
Clu2	f_0^{Clu2}	$f_0^{Clu2}/10$	$f_0^{Clu2}/100$	$f_0^{Clu2}/1000$	$f_0^{Clu2}/10000$
Clu3	f_0^{Clu3}	$f_0^{Clu3}/10$	$f_0^{Clu3}/100$	$f_0^{Clu3}/1000$	$f_0^{Clu3}/10000$

Table 6.10: Overview of the different starting fluorescence levels organized in groups and dilution factors. Technical replicates are available for each of the 20 starting fluorescence levels.

All 60 baseline corrected graphs are shown in Figure 6.4. We use the baseline correction, trend, which looks for a increasing or decreasing trend in the early cycles. It is a linear trend, which is subtract from the fluorescence curve. In the Clusterin dilution-dataset, the slope of the linear curve is between ± 20. Compared to the order of magnitude for the fluorescence level at 10^4, this trend is relatively small.

6.3.2 Can we evaluate sets of technical triplicates?

Before we use the Tichopad approach, Sigma approach and Init approach we have to decide if we can let $n = 3$, meaning if we can combine triplicate curves. Since the Tichopad approach, the Sigma approach and Init approach is based on the same mathematical models, we choose to use the Tichopad approach to conclude if we can use $n = 3$ in all methods. First we perform separate estimation for all curves to look at the variation in $\hat{\alpha}$ og $\hat{\beta}$ within the triplicate sets. Secondly, we look at the normality of the residuals when $n = 3$.

A plot of the log transformed $\hat{\alpha}$ is found in Figure 6.5. They are shown on the log scale because the estimates are very different between dilutions. The values

for $\hat{\alpha}$ lies between $10^4 = 10000$ and $10^6 = 1000000$, which are substantially higher than in the Clusterin dilution-dataset. We see that within some of the triplicates the $\hat{\alpha}$ has a larger variance, but these triplicate sets are in minority. In Figure 6.6, we see a plot of the $\hat{\beta}$'s. These estimates are concentrated around one. This indicates that perhaps β should be set equal to one. From the two plots, it looks like we might be able to evaluate the sets of technical triplicates together.

Next we check for normality for cycle $1 < j < m$ in all curves when we use $n = 3$. Thus we find the 60 sets of residuals when α and β are estimated for each triplicate. Are the estimated ε normally distributed? From the Anderson-Darling test we find that for 27 curves the hypothesis of normality were accepted. This is less than half of the curves. In Figure 6.7 and 6.8, we see the QQ plot of the 18 curves with the smallest p-value within the H_0-hypotheses that were rejected. We choose to evaluate this as not substantially deviation from normality and conclude that we can evaluate sets of technical triplicates in the estimation of f_0.

6.3.3 Cycles used for each method

In Table 6.11, we see summary statistics for the cycles used for each method. We see the same trend as for the Arabidopsis dilution-dataset in Table 6.3. The threshold value for the comparative CT method were set by the biologist at 1275 for all curves.

6.3.4 The f_0 estimates

In Figure 6.1, we see a plot of all 60 log transformed $\hat{f}_0$'s. We observe that the three methods based on the Enzymological method (circles in color black, red and blue) give estimates with higher value than the two methods based on CT values (squares in color green and yellow).

The estimated f_0's from the five methods are plotted in six Mean-Difference plots in Figure 6.10. In each panel, the x-axis show the mean value $(log(\hat{f}_0)^A + log(\hat{f}_0)^B)/2$ from a method A and B, and the y-axis is the difference $(log(\hat{f}_0)^A - log(\hat{f}_0)^B)$ within the same method A and B. If the methods give the same estimates they will follow the red line. If the difference is negative, then $(log(\hat{f}_0)^A < log(\hat{f}_0)^B)$ thus method B gives higher estimates of f_0 than method A. The mean values $(log(\hat{f}_0)^A + log(\hat{f}_0)^B)/2$ are negative since the fluorescence levels are less than one. The estimated starting fluorescence levels from the Tichopad approach and the Sigma approach are very similar, so we only compare the other methods to the Sigma approach in Figure 6.10.

The estimated f_0 from the comparative CT method and the generalized CT method are very similar. In this dataset, the threshold values for the generalized

CT method are spread around $T = 1275$, which is the threshold value for the comparative CT method. We see that the Sigma approach and the Init approach give higher estimates of f_0, than both the comparative CT method and the generalized CT method. If the efficiency has decreased where the thresholds are placed for the two CT methods, these two methods will give lower estimates of f_0. Despite placing the threshold where $\hat{E}$ is close to 2, the generalized CT method find estimates of f_0 with lower value than the estimates from the three methods based on the Enzymological method.

Next we perform a linear regression analysis with $log_{10}(\hat{f_{oi}})^l$ as response separately for each of the five methods. With biological motivation, we find the significant model at significance level 0.05

$$\log_{10}(\hat{f_{0i}})^l = x_i^l \cdot \eta_i^l + u^l + \delta_i^l \tag{6.2}$$

for triplicate i and sample l, where $1 < i < 3$ and $1 < l < 20$. The regression coefficients are $\eta_i^l = [\eta_0, \eta_{ST2}, \eta_{ST3}, \eta_{ST4}, \eta_{D2}, \eta_{D3}, \eta_{D4}, \eta_{D5}]$ for sample type $Dil1$ (ST2), $Dil2$ (ST3) and $Dil3$ (ST4) and dilutions 10 (D2), 100 (D3), 1000 (D4) and 10000 (D5), where η_0 is the intercept. The x_i^l is a vector of 0's and 1's denoting the sample type and the dilution of each sample. The u^l is the random effect caused by the correlation between the technical triplicates.

In this dataset, we find a higher correlation within a triplicate set for some methods. The intraclass correlation using data from the comparative CT method is $9.489 \cdot 10^{-10}$ and in generalized CT method $8.969 \cdot 10^{-10}$. These two methods do not find a high correlation between the estimated starting fluorescence levels within triplicate sets. With the Tichopad approach we find intraclass correlation equal to 0.0987, in the Init approach 0.4959 and in the Sigma approach 0.4984. In these three methods, we have parameter α and β with the same value within triplicate sets. This can cause the higher intraclass correlation in the versions of the Enzymological method compared to the CT methods. We choose to use the mixed effects model for all of the methods to be able to compare them.

From the Anderson-Darling test, we can accept the hypothesis that all of the δ are normally distributed. We assume that δ are distributed as $N(0, \nu^2)$. We find the estimated variances ν^2 to be lowest for the CT with value $6.2840 \cdot 10^{-3}$ and for the generalized CT method with value $5.753 \cdot 10^{-3}$. The next following methods are the Sigma approach (0.0107) , the Tichopad approach (0.0114) and the Init approach (0.0132).

6.3.5 Estimation of the ratio between dilution factors

The true value of each f_0 is unknown, but we do know the true ratios within dilution series. We estimate the log_{10} transform of each estimated ratio. When we

estimate the ratios, we will look at ten ratio combinations as shown in Table 6.12.

In Table 6.13, we see the log transformed estimated ratios calculated from the mean value of the triplicates, as explained in Section 4.5. In most cases the Tichopad approach and the Sigma approach have the lowest bias.

Next we estimate the ratios by calculating the contrast between the dilution factors from the linear mixed effects regression model, where all 60 curves are taken into account simultaneously. The estimated log transformed ratios from the Clusterin dilution-dataset for each of the five methods is found in the five Tables 6.14 (Tichopad approach), 6.15 (Sigma approach), 6.16 (Init approach), 6.17 (comparative CT method) and 6.18 (generalized CT method). We perform a hypothesis test with null hypothesis that the true ratios between the dilution factors are equal to their true known value. From a two sided t-test, we find that four out ten p-values in the Sigma approach and the Init approach are smaller than significance level 0.05. We reject the null hypothesis that the ratios are equal to their true value. All four contrasts include the dilution factor 10 000, $D5$. In the Tichopad approach these same four contrast and in addition the contrast $D2 - D4$ gave rejection of the null hypothesis. In the two CT methods, eight out of ten null hypotheses were rejected.

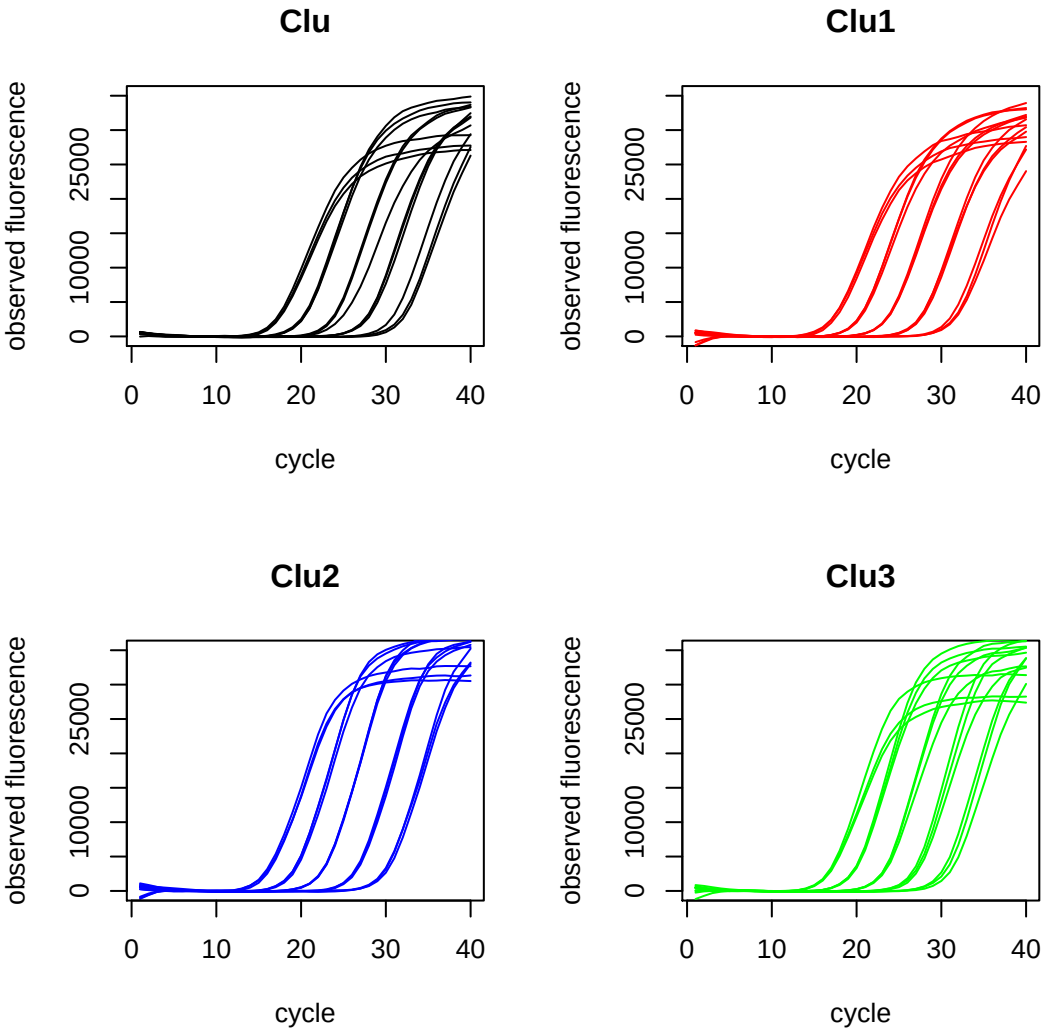

Figure 6.4: The observed fluorescence levels in the Clusterin dilution-dataset. The four different sample types (Clu, Clu1, Clu2 and Clu3) are plotted in separate plots with different colors. In the upper left we have Clu, upper right Clu1, lower left Clu2 and lover right Clu3. The curve to the left in each panel is the original concentration. The second curve from the right has dilution factor 10, then the curve with dilution factor 100 and 1000 are plotted and last the curve with dilution factor 10000 to the far right. This is the fluorescence level after baseline correction.

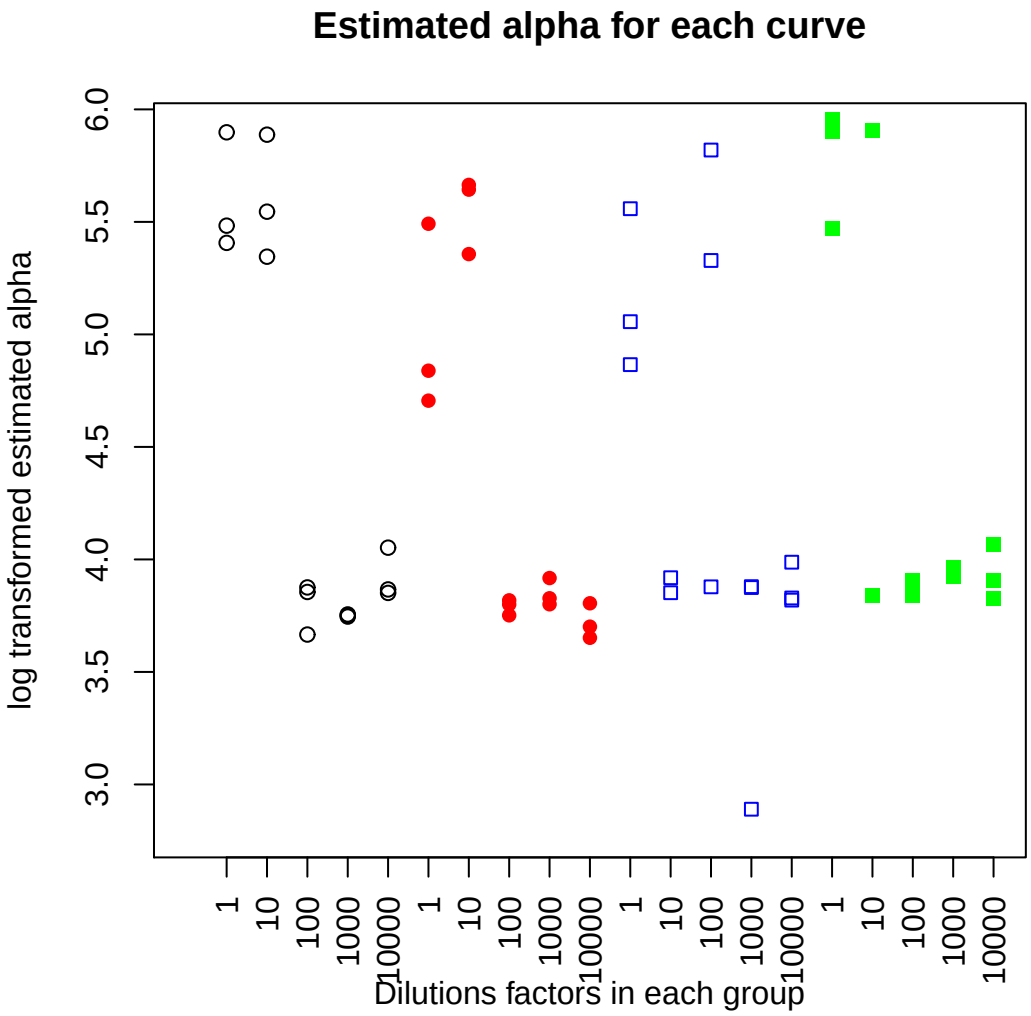

Figure 6.5: Estimation of α for each individual curve in the Clusterin dilution-dataset, in all 3×20 estimations. The estimates are organized into groups according to type of gene and sample group and dilutions. Black circles represent estimates for Clu, red circles represent estimates for Clu1, blue squares represent estimates for Clu2 and green squares represent estimates for Clu3. In each group the x-axis represents the dilutions factor in the order 1, 10, 100, 1000 and 10000.

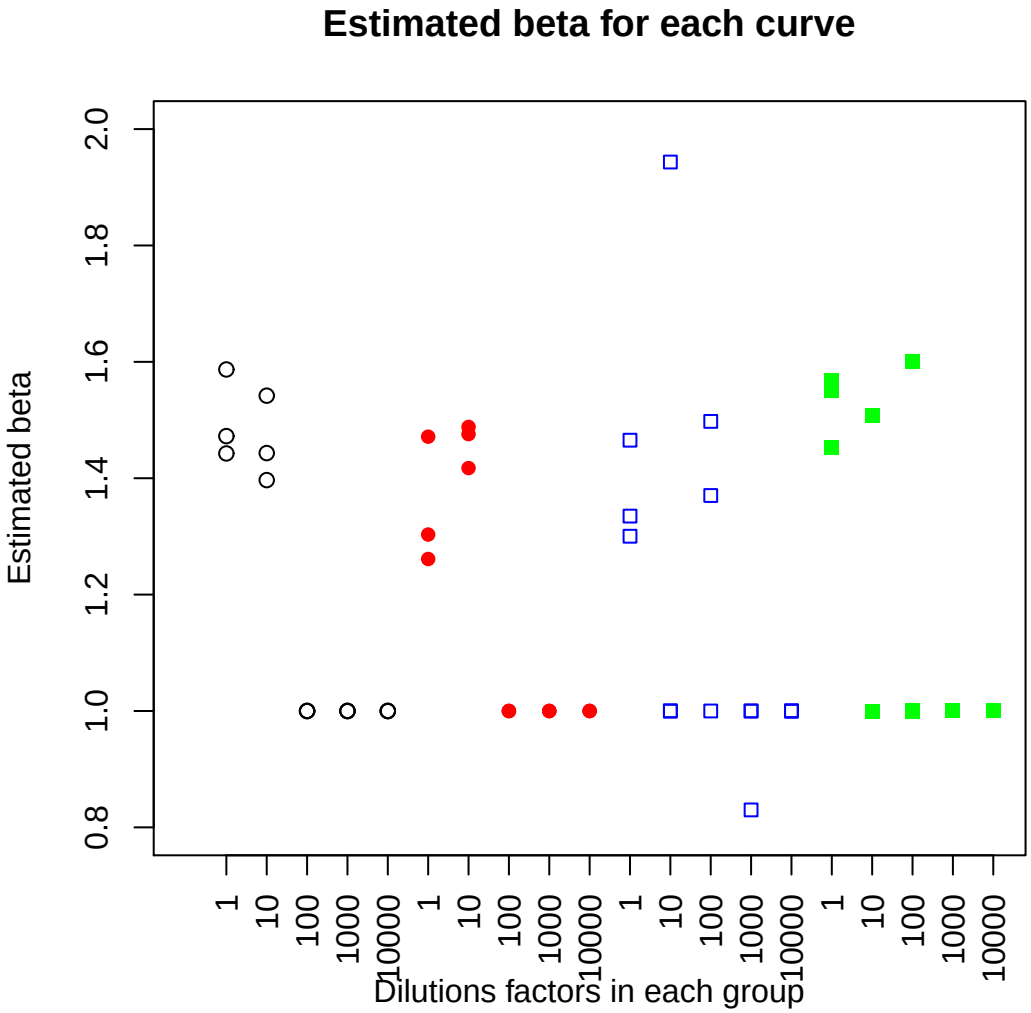

Figure 6.6: Estimation of β for each individual curve in the Clusterin dilution-dataset, in all 3×20 estimations. The explanation of the plot is found in Figure 6.5

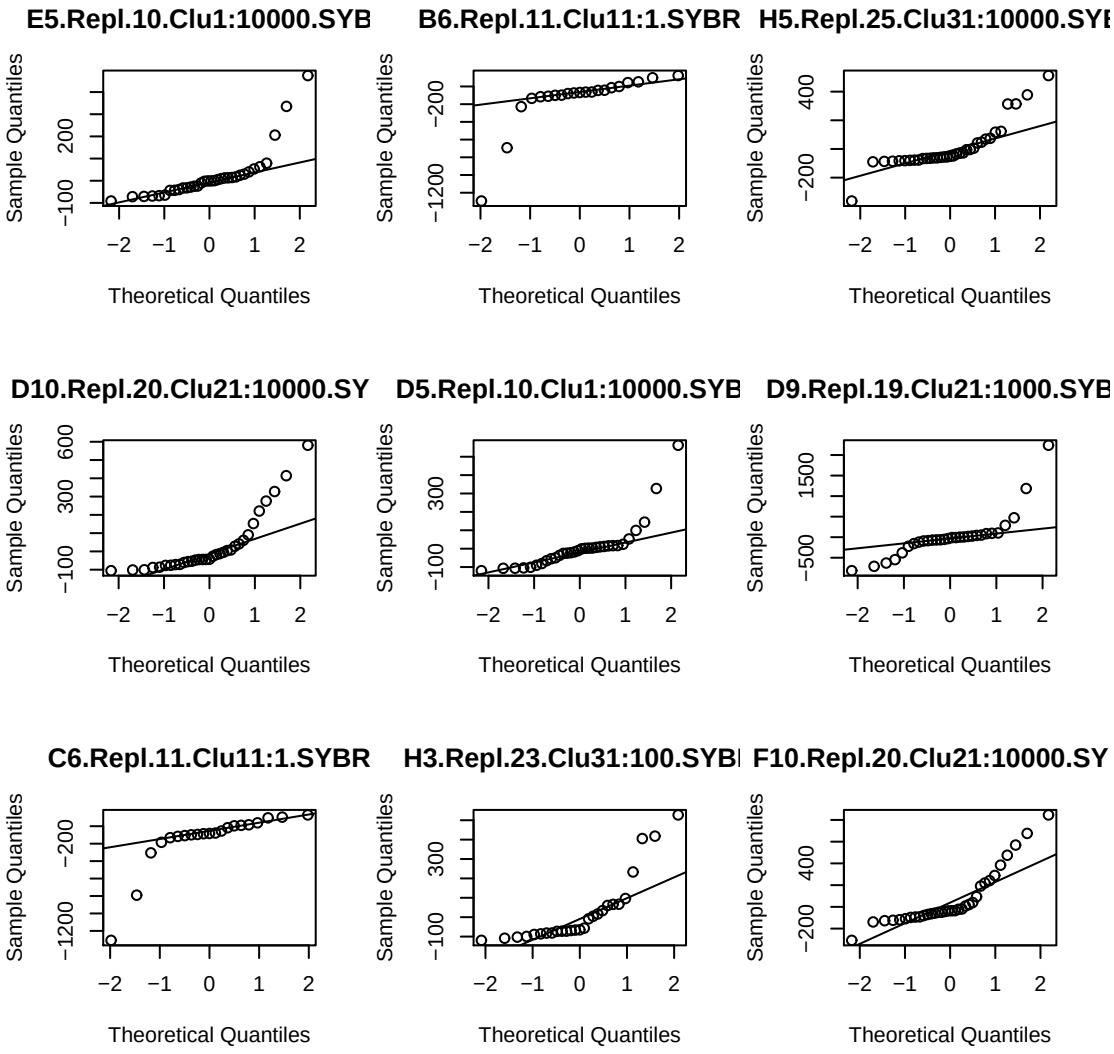

Figure 6.7: The QQ plot for 9 of the 17 sets of residuals, where the hypothesis of normally distributed error was rejected.

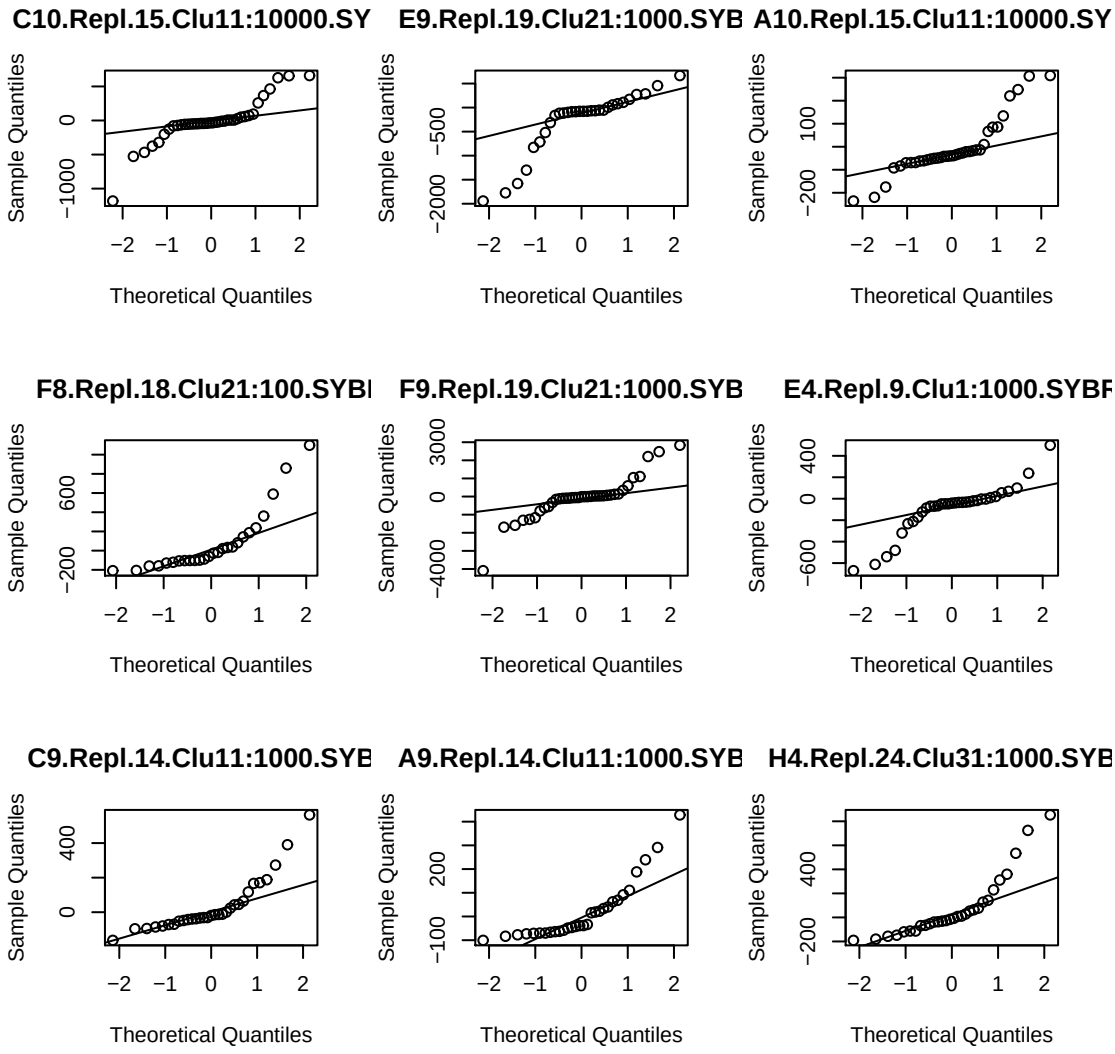

Figure 6.8: The QQ plot for the remaining 8 of the 17 sets of residuals, where the hypothesis of normally distributed error was rejected.

	No of Cycles in Thicopad-app.	No of Cycles in Sigma-app.	m	C_T (T=1275)	(C_T,T)	at $\hat{E}$
Clu1:1	8	3.33	21	16.23	(16.67, 1807)	at 1.86
Clu1:10	8.67	3	24	19.1	(19, 1207)	at 1.93
Clu1:100	9	3.33	27.67	22.97	(23, 1394)	at 1.9
Clu1:1000	13.33	4	33	26.6	(26, 970)	at 1.91
Clu1:10000	7	2.67	33.67	30.37	(29.67, 882)	at 1.93
Clu11:1	7.67	3	21	16.2	(16, 1288)	at 1.89
Clu11:10	8.33	3.33	24.33	19.23	(19.33, 1454)	at 1.95
Clu11:100	8.67	3.33	27.33	22.47	(22, 1073)	at 1.93
Clu11:1000	7.33	3.33	30.33	26.43	(26, 1158)	at 1.92
Clu11:10000	9.67	3.67	36.33	30.23	(29.67, 888)	at 1.92
Clu21:1	7.33	3	20	14.83	(15, 1645)	at 1.9
Clu21:10	7.67	3.33	22.33	17.93	(18, 1364)	at 1.96
Clu21:100	9	3.67	26.67	21.27	(21.67, 1761)	at 1.89
Clu21:1000	9.67	2.33	32.33	25.23	(25.67, 1824)	at 1.9
Clu21:10000	8.33	3	33	28.67	(28.67, 1311)	at 1.93
Clu31:1	7.33	3	20.67	15.7	(16, 1622)	at 1.93
Clu31:10	8.33	3.67	23.67	18.43	(18.33, 1246)	at 1.94
Clu31:100	9	3.33	27.33	22.03	(22, 1256)	at 1.96
Clu31:1000	8.67	3	30	25.77	(26, 1715)	at 1.92
Clu31:10000	9	3.33	33.67	29.3	(29.33, 1405)	at 1.92

Table 6.11: Summary statistics for the triplicates of fluorescence curves used in the calculations with the five methods. All values are means over triplicates curves. The names in the left column denote the different sample types and dilutions. In the first column we see the number of cycles used in the MLE with the Tichopad approach, the second column we see the number of cycles used in the Sigma approach, the third column shows the inflection point. The fourth column presents the CT values from the comparative CT method with thresholds equal to 1275 for all curves. In the last column we see the CT values corresponding to the threshold and estimated efficiency from the generalized CT method where each curve have individual thresholds.

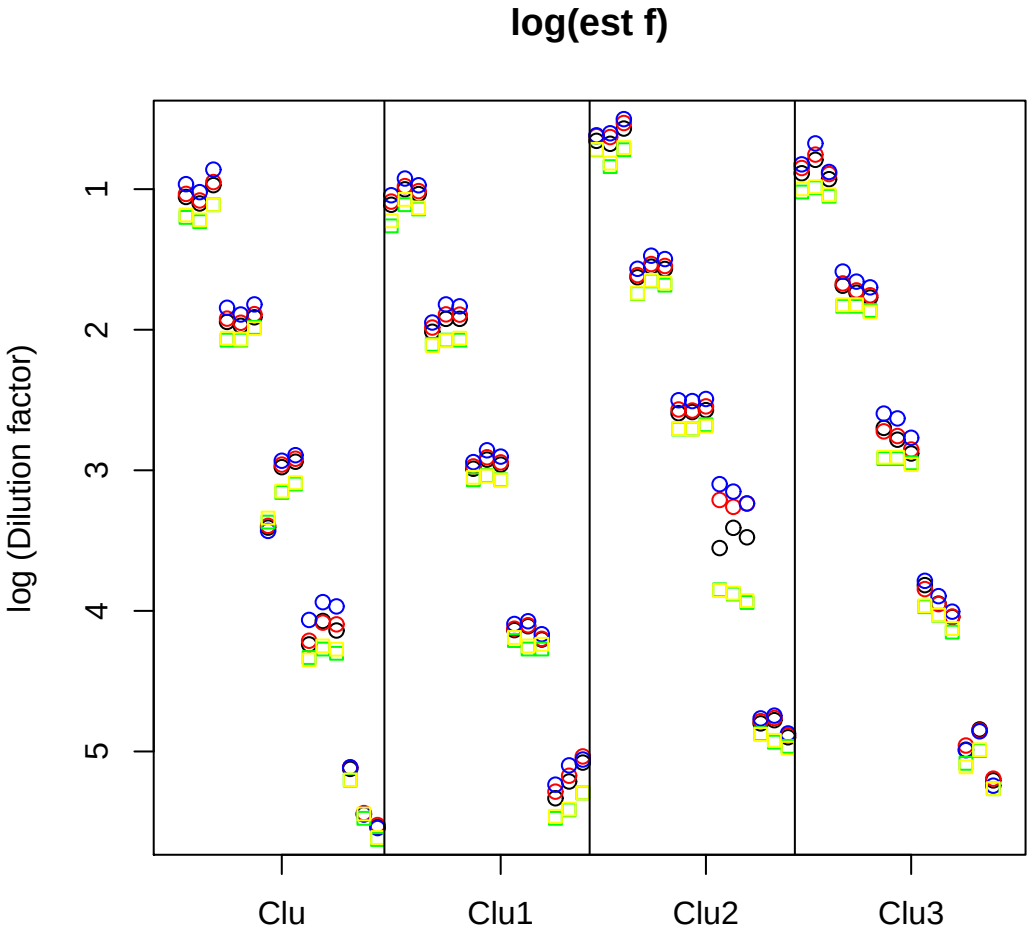

Figure 6.9: Plot of $log_{10}(\hat{f}_0)$ for the Tichopad approach (black circles), the Sigma approach (red circles), the Init approach (blue circles), the comparative CT method (green squares) and the generalized CT method (yellow squares). On the x-axis are the four sample types (Clu, Clu1, Clu2 and Clu3) organized by increasing dilution factor.

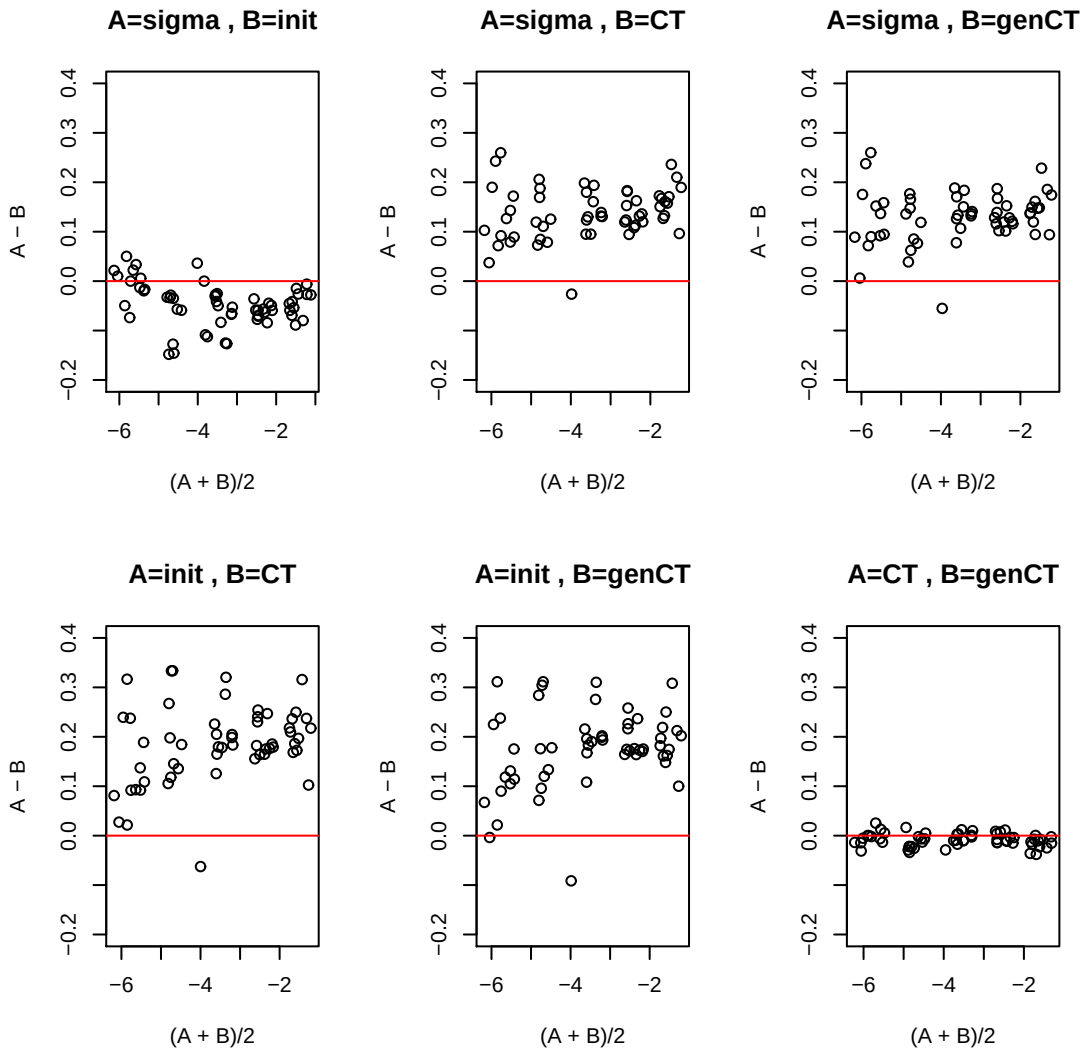

Figure 6.10: Mean-Difference plots of all 48 estimates of $log(\hat{f}_0)$ for comparing pairs of methods. In each panel we see on the x-axis the mean value $(log(\hat{f}_0)^A + log(\hat{f}_0)^B)/2$ from a method A and B, and on the y-axis we see the difference $(log(\hat{f}_0)^A - log(\hat{f}_0)^B)$ from the same method A and B. If the methods give the same estimates they will follow the red line.

ratio	log transformed true value	
f_0^{Dil1}/f_0^{Dil10}	$log_{10}(f_0) - log_{10}(f_0/10)$	$= log_{10}10 = 1$
f_0^{Dil1}/f_0^{Dil100}	$log_{10}(f_0) - log_{10}(f_0/100)$	$= log_{10}100 = 2$
$f_0^{Dil1}/f_0^{Dil1000}$	$log_{10}(f_0) - log_{10}(f_0/1000)$	$= log_{10}1000 = 3$
$f_0^{Dil1}/f_0^{Dil10000}$	$log_{10}(f_0) - log_{10}(f_0/10000)$	$= log_{10}10000 = 4$
f_0^{Dil10}/f_0^{Dil100}	$log_{10}(f_0/10) - log_{10}(f_0/100)$	$= log_{10}10 = 1$
$f_0^{Dil10}/f_0^{Dil1000}$	$log_{10}(f_0/10) - log_{10}(f_0/1000)$	$= log_{10}100 = 2$
$f_0^{Dil10}/f_0^{Dil10000}$	$log_{10}(f_0/10) - log_{10}(f_0/10000)$	$= log_{10}1000 = 3$
$f_0^{Dil100}/f_0^{Dil1000}$	$log_{10}(f_0/100) - log_{100}(f_0/1000)$	$= log_{10}10 = 1$
$f_0^{Dil100}/f_0^{Dil10000}$	$log_{10}(f_0/100) - log_{100}(f_0/10000)$	$= log_{10}100 = 2$
$f_0^{Dil1000}/f_0^{Dil10000}$	$log_{10}(f_0/1000) - log_{100}(f_0/10000)$	$= log_{10}10 = 1$

Table 6.12: Ten ratios between five dilutions 1, 10, 100, 1000 and 10000.

True value	Tichopad	Sigma	Init	comparativ CT	generalized CT
1	0.905(0.035)	0.913(0.039)	0.899(0.033)	0.324(0.713)	0.313(0.695)
2	1.95(0.052)	1.965(0.047)	1.953(0.09)	0.474(0.676)	0.486(0.662)
3	3.028(0.124)	2.995(0.234)	2.966(0.255)	1.008(0.178)	1.014(0.158)
4	4.189(0.074)	4.199(0.078)	4.242(0.1)	1.347(0.71)	1.364(0.702)
1	1.045(0.055)	1.051(0.055)	1.054(0.083)	0.151(0.511)	0.173(0.516)
2	2.123(0.152)	2.081(0.272)	2.068(0.285)	0.685(0.651)	0.701(0.649)
3	3.283(0.071)	3.286(0.078)	3.344(0.098)	1.024(0.209)	1.05(0.207)
1	1.078(0.132)	1.03(0.243)	1.014(0.267)	0.534(0.513)	0.528(0.519)
2	2.238(0.025)	2.234(0.033)	2.29(0.049)	0.873(0.32)	0.877(0.334)
1	1.161(0.139)	1.204(0.258)	1.276(0.274)	0.339(0.602)	0.35(0.613)

Table 6.13: Results for the estimated ratios in the Clusterin dilution-dataset with all five methods. The calculations of the mean value and the sample standard deviation in paranthesis is explained in Section 4.5.

	True	Estimate	Std. Error	t value	DF	Pr(>|t|)	Lower.CI	Upper.CI
D1-D2	1.0000	0.9034	0.0503	-1.9211	12.0000	0.0788	-0.2061	0.0130
D1-D3	2.0000	1.9602	0.0503	-0.7927	12.0000	0.4433	-0.1494	0.0697
D1-D4	3.0000	3.0301	0.0503	0.5995	12.0000	0.5600	-0.0794	0.1397
D1-D5	4.0000	4.2049	0.0503	4.0769	12.0000	0.0015	0.0954	0.3145
D2-D3	1.0000	1.0567	0.0503	1.1284	12.0000	0.2812	-0.0528	0.1662
D2-D4	2.0000	2.1267	0.0503	2.5206	12.0000	0.0269	0.0172	0.2362
D2-D5	3.0000	3.3015	0.0503	5.9980	12.0000	0.0001	0.1920	0.4110
D3-D4	1.0000	1.0700	0.0503	1.3923	12.0000	0.1891	-0.0395	0.1795
D3-D5	2.0000	2.2448	0.0503	4.8696	12.0000	0.0004	0.1353	0.3543
D4-D5	1.0000	1.1748	0.0503	3.4774	12.0000	0.0046	0.0653	0.2843

Table 6.14: Estimated ratios between dilutions in the Clusterin dilution-dataset from the Tichopad approach with a 95% confidence interval.

	True	Estimate	Std. Error	t value	DF	Pr(>|t|)	Lower.CI	Upper.CI
D1-D2	1.0000	0.9118	0.0839	-1.0513	12.0000	0.3138	-0.2711	0.0946
D1-D3	2.0000	1.9742	0.0839	-0.3077	12.0000	0.7636	-0.2087	0.1571
D1-D4	3.0000	2.9950	0.0839	-0.0593	12.0000	0.9537	-0.1879	0.1779
D1-D5	4.0000	4.2148	0.0839	2.5589	12.0000	0.0251	0.0319	0.3977
D2-D3	1.0000	1.0624	0.0839	0.7436	12.0000	0.4714	-0.1205	0.2453
D2-D4	2.0000	2.0833	0.0839	0.9920	12.0000	0.3408	-0.0996	0.2662
D2-D5	3.0000	3.3030	0.0839	3.6102	12.0000	0.0036	0.1202	0.4859
D3-D4	1.0000	1.0209	0.0839	0.2484	12.0000	0.8080	-0.1620	0.2037
D3-D5	2.0000	2.2406	0.0839	2.8666	12.0000	0.0142	0.0577	0.4235
D4-D5	1.0000	1.2198	0.0839	2.6182	12.0000	0.0225	0.0369	0.4027

Table 6.15: Estimated ratios between dilutions in the Clusterin dilution-dataset from the Sigma approach with a 95% confidence interval.

The dilutions factor 10000 appears to be difficult to estimate for all methods. After excluding this highest dilution, we see that the three methods based on the Enzymological method give estimates of the remaining six ratios with lower bias. In Figure 6.11, we see a boxplot of the remaining six estimated ratios minus the true value. The bias in the three methods based on Enzymological method is closer to the zero and has lower variation than the CT based methods.

	True	Estimate	Std. Error	t value	DF	Pr(>|t|)	Lower.CI	Upper.CI
D1-D2	1.0000	0.8960	0.0936	-1.1107	12.0000	0.2885	-0.3080	0.1000
D1-D3	2.0000	1.9642	0.0936	-0.3823	12.0000	0.7090	-0.2398	0.1682
D1-D4	3.0000	2.9660	0.0936	-0.3632	12.0000	0.7227	-0.2380	0.1700
D1-D5	4.0000	4.2573	0.0936	2.7477	12.0000	0.0177	0.0533	0.4613
D2-D3	1.0000	1.0682	0.0936	0.7284	12.0000	0.4803	-0.1358	0.2722
D2-D4	2.0000	2.0700	0.0936	0.7474	12.0000	0.4692	-0.1340	0.2740
D2-D5	3.0000	3.3612	0.0936	3.8584	12.0000	0.0023	0.1573	0.5652
D3-D4	1.0000	1.0018	0.0936	0.0190	12.0000	0.9851	-0.2022	0.2058
D3-D5	2.0000	2.2930	0.0936	3.1300	12.0000	0.0087	0.0891	0.4970
D4-D5	1.0000	1.2913	0.0936	3.1109	12.0000	0.0090	0.0873	0.4953

Table 6.16: Estimated ratios between dilutions in the Clusterin dilution-dataset from the Init approach with a 95% confidence interval.

	True	Estimate	Std. Error	t value	DF	Pr(>|t|)	Lower.CI	Upper.CI
D1-D2	1.0000	0.8830	0.0324	-3.6146	12.0000	0.0035	-0.1875	-0.0465
D1-D3	2.0000	1.9391	0.0324	-1.8807	12.0000	0.0845	-0.1314	0.0096
D1-D4	3.0000	3.0906	0.0324	2.7987	12.0000	0.0161	0.0201	0.1611
D1-D5	4.0000	4.1843	0.0324	5.6953	12.0000	0.0001	0.1138	0.2548
D2-D3	1.0000	1.0561	0.0324	1.7339	12.0000	0.1085	-0.0144	0.1266
D2-D4	2.0000	2.2076	0.0324	6.4133	12.0000	0.0000	0.1370	0.2781
D2-D5	3.0000	3.3013	0.0324	9.3100	12.0000	0.0000	0.2308	0.3718
D3-D4	1.0000	1.1514	0.0324	4.6794	12.0000	0.0005	0.0809	0.2220
D3-D5	2.0000	2.2452	0.0324	7.5761	12.0000	0.0000	0.1747	0.3157
D4-D5	1.0000	1.0937	0.0324	2.8966	12.0000	0.0134	0.0232	0.1643

Table 6.17: Estimated ratios between dilutions in the Clusterin dilution-dataset from the comparative CT method with a 95% confidence interval.

	True	Estimate	Std. Error	t value	DF	Pr(>|t|)	Lower.CI	Upper.CI
D1-D2	1.0000	0.8951	0.0310	-3.3863	12.0000	0.0054	-0.1723	-0.0374
D1-D3	2.0000	1.9503	0.0310	-1.6055	12.0000	0.1344	-0.1172	0.0178
D1-D4	3.0000	3.0933	0.0310	3.0118	12.0000	0.0108	0.0258	0.1607
D1-D5	4.0000	4.1969	0.0310	6.3602	12.0000	0.0000	0.1295	0.2644
D2-D3	1.0000	1.0551	0.0310	1.7809	12.0000	0.1002	-0.0123	0.1226
D2-D4	2.0000	2.1981	0.0310	6.3981	12.0000	0.0000	0.1307	0.2656
D2-D5	3.0000	3.3018	0.0310	9.7466	12.0000	0.0000	0.2343	0.3693
D3-D4	1.0000	1.1430	0.0310	4.6173	12.0000	0.0006	0.0755	0.2104
D3-D5	2.0000	2.2467	0.0310	7.9657	12.0000	0.0000	0.1792	0.3141
D4-D5	1.0000	1.1037	0.0310	3.3484	12.0000	0.0058	0.0362	0.1712

Table 6.18: Estimated ratios between dilutions in the Clusterin dilution-dataset from the generalized CT method with a 95% confidence interval.

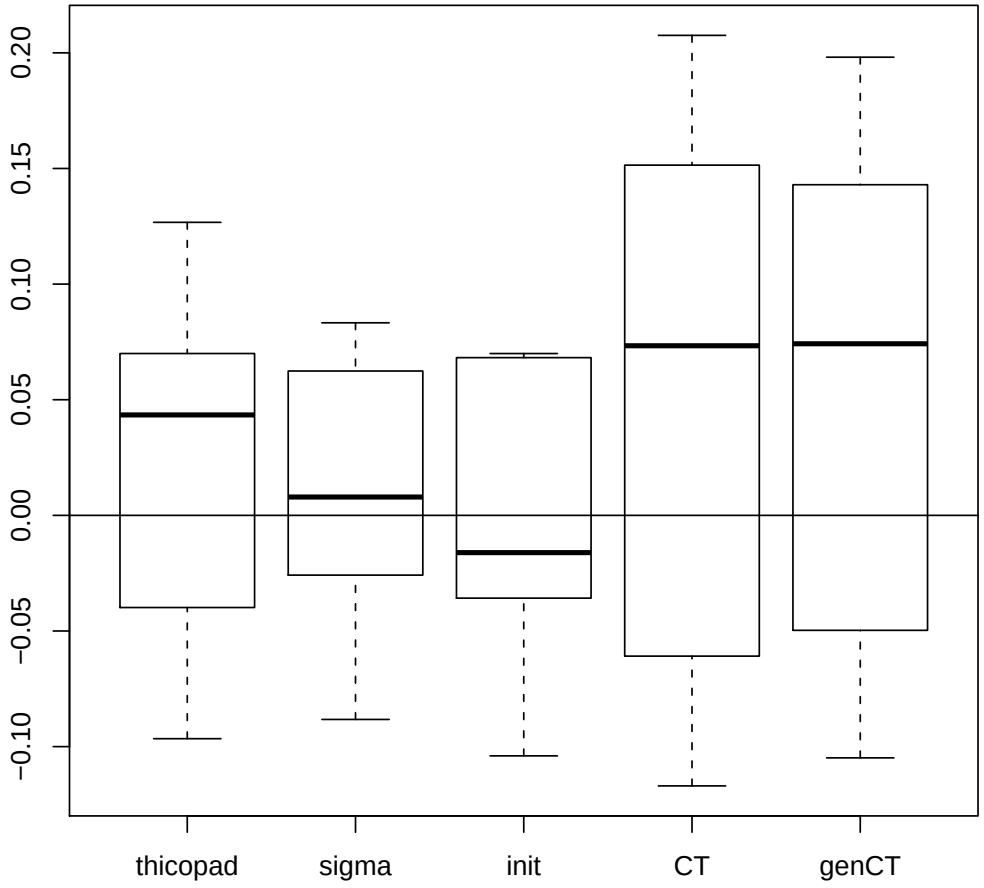

Figure 6.11: Boxplot of estimated ratios minus the true value, thus the bias. Dilution $D5$ are not included, so there are 6 estimates for each of the five methods.

Chapter 7

Discussion

For the comparison of the different methods, two datasets are analyzed. The Arabidopsis dilution-dataset have optimized primers and are a study of plants. The Clusterin dilution-dataset is a study of a cell line, where a range of primers are used with different amplification abilities. The values of the estimated parameters f_0 have different order of magnitude, which follows from the different type of technology used. The observed fluorescence levels in the Arabidopsis dilution-dataset have a maximum value at 50 in the last cycles, and in the Clusterin dilution-dataset, the maximum fluorescence levels are 30000. In the Arabidopsis dilution-dataset, the most diluted samples have dilution factor 64 as oppose to 10000 in the Clusterin dilution-dataset. Despite these differences and the additional challenge with the not perfect amplification rate in the Clusterin dilution-dataset, the Enzymological method gives good results for both datasets.

We use two regression models in the analysis of the estimated starting fluorescence levels from the two datasets. In the Arabidopsis dilution-dataset, a linear regression is used. In the Clusterin dilution-dataset we use a linear mixed effects regression model, which takes into account the correlation between technical triplicates. It seems appropriate, that a mixed effects model is used when $\hat{f}_0$ is based on the Enzymological method. This method use common $\hat{\alpha}$ and $\hat{\beta}$ for triplicates sets, which can lead to higher correlating between the $\hat{f_{0i}}$ within a triplicate set. However the mixed effects model was only found to be needed for the Clusterin dilution-dataset, and not for the Arabidopsis dilution-dataset. An explanation could be that the variance between all 60 estimates of f_0 is larger in the Clusterin dilution-dataset, because of a higher dilution factor, compared to the Arabidopsis dilution-dataset.

7.0.1 Estimation of f_0 and ratio between dilution factors

The estimates of f_0 and ratio are similar for the five methods in the Arabidopsis dilution-dataset. However, in the Clusterin dilution-dataset, there is a larger difference between the CT methods and the three methods based on Enzymological method. Why is this difference greater in the Clusterin dilution-dataset? An important difference between the CT methods and the methods based on the Enzymological method is the assumption of the efficiency.

If the efficiency over the curves is much lower than two, the difference between Enzymological method and the CT method will be larger, than if the efficiency is close to two. What are the efficiencies in the two datasets? We compare the estimated efficiencies at the inflection points for both datasets. When estimating E at the inflection point we will compare the same part of the curves for both datasets, where there are little relative noise. The estimated efficiencies at the inflection point for the Arabidopsis dilution-dataset are plotted in Figure 7.1. The mean value is the red line at 1.474. The estimated efficiencies at the inflection point for the Clusterin dilution-dataset are plotted in Figure 7.2, with mean value 1.317. This is a strong indication that the efficiencies over the curves are lower in the Clusterin dilution-dataset than in the Arabidopsis dilution-dataset. This great difference in the efficiency at the inflection point might be the main reason the Enzymological method and the CT methods perform differently in the Clusterin dilution-dataset.

7.0.2 Baseline correction

Another aspect we have not emphasized, is the fact that there are used two methods of baseline corrections. In the Arabidopsis dilution-dataset, one constant value for each cycle is subtracted from the observed fluorescence curve. We calculate this value as suggested in Jørstad et al. (2008) by ranking the observed fluorescence level according to numerical value, and then finding the window of 5 consecutive data points having the smallest rank sum. In the Clusterin dilution-dataset, a cycle dependent value is subtracted from the observed fluorescence curve. This baseline correction is called *trend* and is widely used in commercial software for PCR analysis. In the PCR research field, baseline correction is suspected to have a great influence on the results. When comparing the five estimation methods within a dataset, the same baseline correction was used, thus should the comparison between the five estimating methods still be valid.

In the analysis of the estimated efficiencies, we compared the Enzymological method in the two datasets. The influence from the baseline correction is unknown. With no baseline correction, the fluorescence levels are higher and the relative amplification is lower giving a smaller estimated efficiency. Without baseline

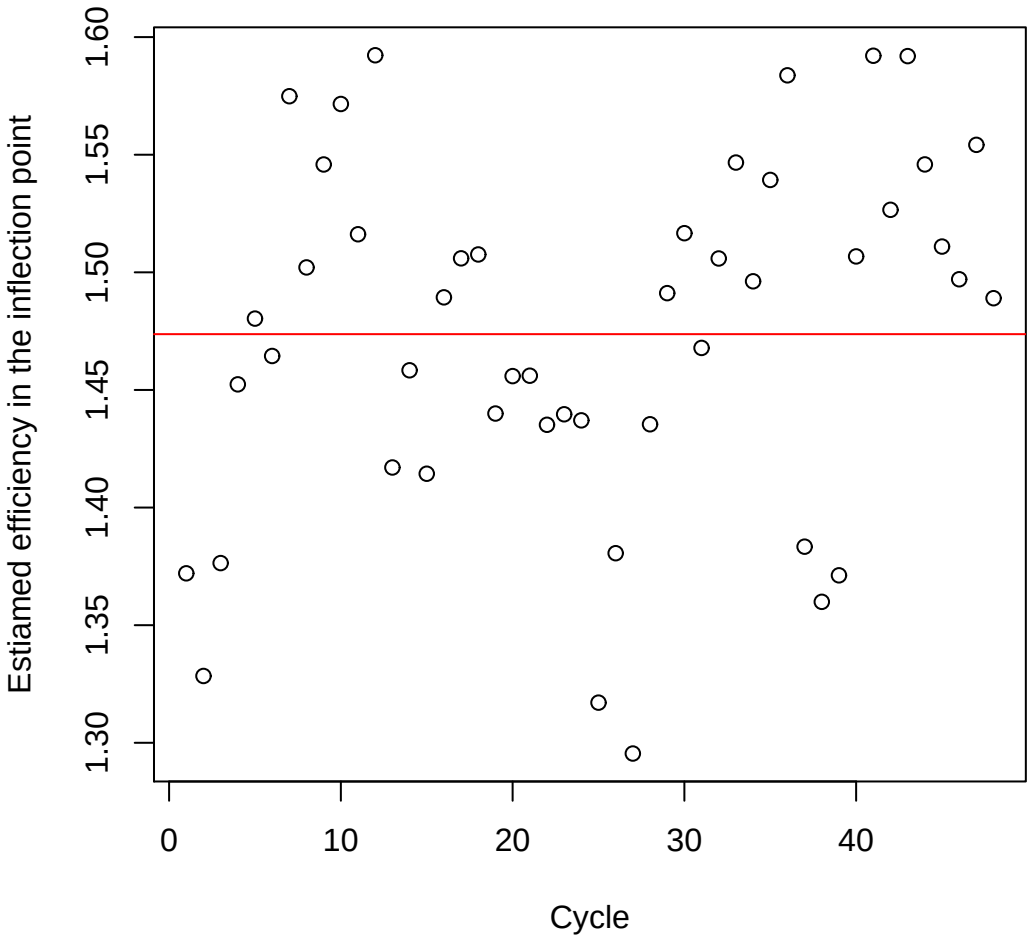

Figure 7.1: Estimated efficiencies at the inflection point for the 48 curves in the Arabidopsis dilution-dataset. The red line is the mean value.

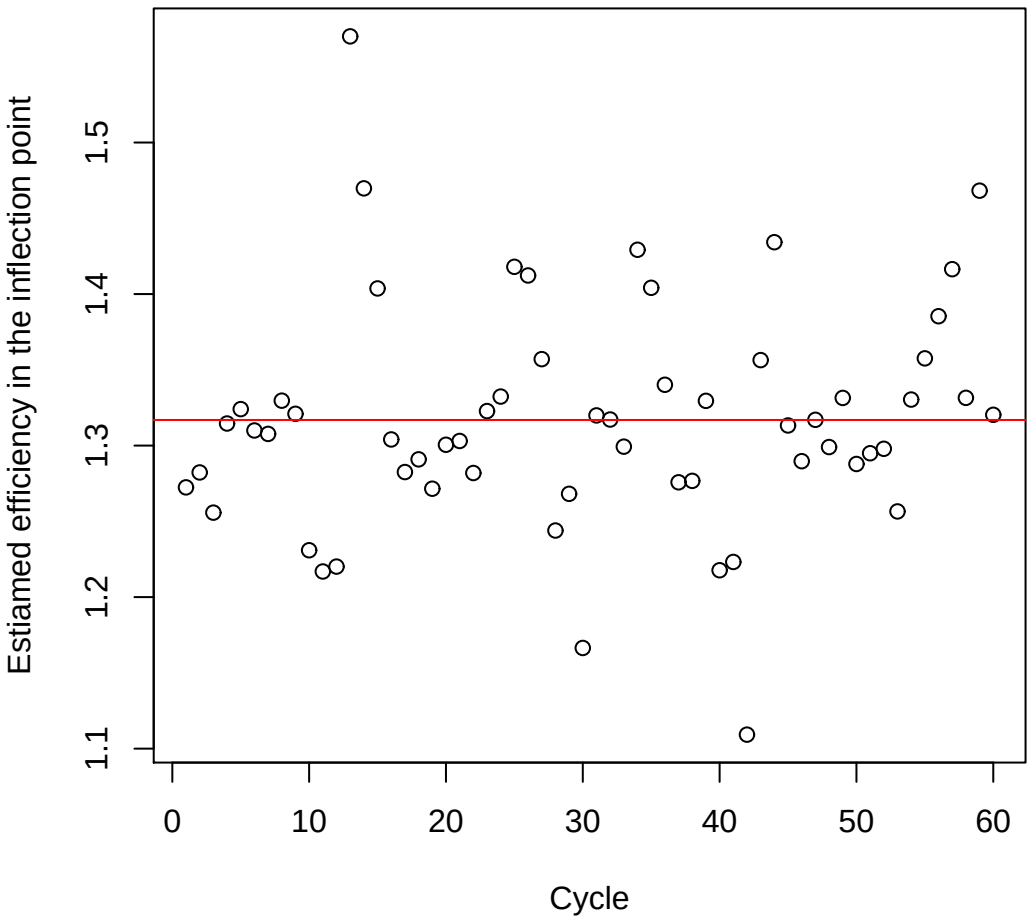

Figure 7.2: Estimated efficiencies at the inflection point for the 60 curves in the Clusterin dilution-dataset. The red line is the mean value.

correction, the efficiency still drops to a smaller level in the Clusterin dilution-dataset, than in the Arabidopsis dilution-dataset. Without baseline correction the mean value of the estimated efficiencies is 1.085 at the inflection point in the Clusterin dilution-dataset and 1.214 in the Arabidopsis dilution-dataset. This shows that the difference in the estimated efficiencies, are not due to different baseline correction.

7.1 Further evaluation of the Enzymological method

7.1.1 Assumptions and parameters

The Tichopad approach, the Init approach and the Sigma approach must be tested and analyzed on more datasets. However, they show good results for the two datasets used in this paper. Still there are aspects within each approach, which can be investigated further. We have used a mathematical model where β is a free parameter. In both datasets, we see that $\hat{\beta}$ is close to one. When setting this parameter equal to one the model gets simpler and we have one more degrees of freedom. For $\beta > 1$ the function in Equation (3.4) is concave after the inflection point, which may be useful if we want to use a larger part of the fluorescence curve for estimation. In our MLE estimation, we could have estimated only one starting fluorescence level for each triplicate sets, because the starting concentration should be the same from a biological view. Another aspect is to investigate the rapid fluctuations displayed for early cycles in many observed log transformed fluorescence curves. We have not been able to explain this trend with the amplification process. If we improve our knowledge about the fluorescence level in the early cycles, we can evaluate more of the curve. The ending cycle m_i in Figure 4.4 indicates that we can place the end cycle m_i earlier and get smaller residuals and thus a better fit.

7.1.2 Log model

In our model, we have assumed additive noise on the original scale. Another possibility is to look at the log transformed fluorescence data. When letting the observed log transformed fluorescence be modeled with additive noise, we get a model

$$\log(y_{ji}) = \log(f_{ji}) + \varepsilon_{ji}, \tag{7.1}$$

where ε_{ji} are independent and distributed as $N(0, \sigma^2)$.

In Figure 7.3, we see the estimated $\{\varepsilon_{ji}\}$ with $\hat{f}$ from Equation (7.1) in the Enzymological method, with starting cycle $s_i = m_i - 7$. With an Anderson-Darling

test, we accept the hypothesis of normally distributed noise on a significance level 5% for all curves and cycles $y_{s_i}, \ldots, y_{m_i}$.

7.2 Further evaluation of the CT methods

We have looked at the comparative CT method and the generalized CT method. In the comparative CT method, we place the common threshold T right after the ground phase. In generalized CT method, we place the individual thresholds T_i where the estimated efficiency is close to 2. These two methods give similar results for our two datasets. From our two datasets, the generalized CT method gives estimates of the ratio, which are a closer to the true value than the estimates from the comparative CT method. Still there are aspects about the efficiency and the thresholds we can investigate further.

7.2.1 Estimation of the efficiency

The CT methods find estimates of f_0, which are smaller compared to the estimates from the three other methods, as seen in Figure 6.9. The CT methods always assume efficiency equal to 2. If we implement the comparative CT method with efficiency equal to 1.9 and do the same calculations as in Section 6.3.4, we get an estimated variance for ν^2 equal to $5.388 \cdot 10^{-3}$ compared to $6.2840 \cdot 10^{-3}$ when calculating with $E = 2$. The estimated ratios are closer to the true value, and in four out of ten tests we can accept that the bias is equal to zero on a significance level 0.05 when calculating with $E = 1.9$. All these results indicate that the true efficiencies are smaller than two for the Clusterin dilution-dataset.

An improvement on the generalized CT method could be to introduce the estimated efficiency $\hat{E}$ instead of forcing it to be 2. Then we would assume that the constant efficiency from cycle CT and forward to cycle one had value $\hat{E}$. This is called the Pfaffl method, see Pfaffl (2001).

7.2.2 Threshold

A potential problem with the comparative CT method is that the threshold is set to be the same for all curves. If we change the threshold, how do the resulting estimates of f_0 and ratios change? From a small study with the Clusterin dilution-dataset, we chose ten different thresholds evenly spread out from fluorescence value 900 to 15000 and found the corresponding CT values. Using Equation $f_{0k} = T_k \cdot 2^{-CT_k}$, we find the estimated starting fluorescence level with the comparative CT method for each curve k when $1 < k < 60$. Then we get 10 sets of 60 estimated f_0's. Using the linear mixed effects model on each set of 60 $\hat{f}_0$, we

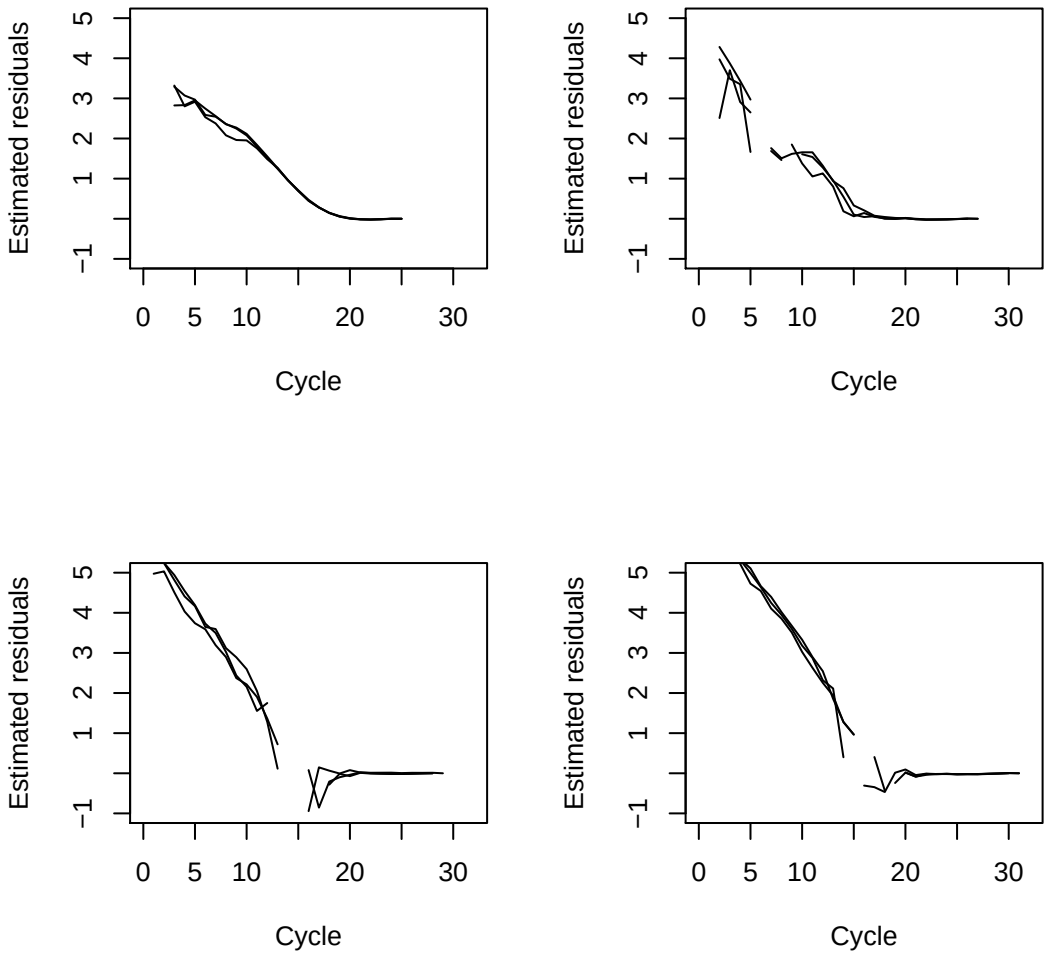

Figure 7.3: The estimated noise $\hat{\varepsilon}$ with a model where the log transformed fluorescence level have additive noise.

calculate the contrasts to find estimates of the ratios between dilution factors. The $\hat{f}_0$'s values were decreasing together with higher thresholds. This make sense, since the efficiency drops from two, while the CT assume constant efficiency equal to two. The standard deviation of the residuals increases from $7.977 \cdot 10^{-2}$ (T=900) to $1.136 \cdot 10^{-1}$ (T=15000). The standard deviation of the contrasts also increases from 0.0326 (T=900) to 0.046 (T=15000). From these results it looks like setting the threshold too late will result in estimates with greater variance and with larger bias.

Chapter 8

Conclusion

In this thesis we have presented three versions of the Enzymological method. From the simulated data, we found that including a certain number of cycles in the maximum likelihood estimation gives estimates of f_0, which are close to the true value. An interesting result is that the Init approach performs well, but including the MLE did lead to lower variance in $\hat{f}_0$. To evaluate if the Enzymological method can be used for analyzing a dataset, there is little information in the p-assumption plot. It is more important to test if the noise ε is normally distributed. Two real datasets have been analyzed, the Arabidopsis dilution-dataset with optimized primer pairs and the Clusterin dilution-dataset whit different primer pairs, meaning that some primer pairs give higher efficiency than other primer pairs.

The results from the comparative CT method and the generalized CT method give similar results when the thresholds are placed properly, but setting the threshold too late will lead to estimates of the ratio with larger variance and larger bias. From the regression model, we found that the estimated variance for the residuals in the linear regression models were very similar for all the five methods in the Arabidopsis dilution-dataset. In all methods, the estimates of all ratios between dilution factors led to acceptance of the hypothesis, that the ratio were equal to the true value. Based on our results, we can not conclude which method to prefer for the Arabidopsis dilution-dataset. In the Clusterin dilution-dataset, the estimated variance in the residuals for the linear mixed effects model were smaller for the two methods based on CT values, than for the other three competitive methods. But when looking at the estimated ratios, the versions of the Enzymological method gave lower bias than the CT methods. The estimates of the ratio between dilution factors from the version of the Enzymological method, led to acceptance of the hypothesis, that the ratio were equal to the true value. From our results, we can not conclude which of the three versions of the Enzymological method are the best.

Estimating the ratio between starting fluorescence levels are commonly calcu-

lated in the fields of functional genomics. For this practice, the versions of the Enzymological method seem to give good estimates in both of our datasets. The Enzymological method is an interesting basis for further work.

Bibliography

Alvarez, M. J., Vila-Ortiz, G. J., Salibe, M. C., Podhajcer, O. L. and Pitossi, F. J. (2007). Model based analysis of real-time PCR data from DNA binding dye protocols., *BMC Bioinformatics* **8**: 85.
URL: `http://www.ncbi.nlm.nih.gov/pubmed/17349040`

Follestad, T., Jørstad, T. S., Erlandsen, S. E., Sandvik, A. K., Bones, A. M. and Langaas, M. (2010). A Bayesian Hierarchical Model for Quantitative Real-Time PCR Data, *Statistical Applications in Genetics and Molecular Biology* **9**(1).

Jørstad, T. S., Follestad, T., Langaas, M. and Bones, A. M. (2008). A simple method for quantitating gene expression levels from quantitative real-time PCR data. Unpublished manuscript.

Livak, K. J. and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ Method., *Methods (San Diego, Calif.)* **25**(4): 402–8.
URL: `http://www.ncbi.nlm.nih.gov/pubmed/11846609`

Pfaffl, M. W. (2001). A new mathematical model for relative quantification in real-time RT-PCR., *Nucleic acids research* **29**(9): e45.

Schmittgen, T. D. and Livak, K. J. (2008). Analyzing real-time PCR data by the comparative C_T method, *Nature* **3**(6): 1101–1108.

Schnell, S. and Mendoza, C. (1997). Enzymological considerations for a theoretical description of the quantitative competitive polymerase chain reaction (QC-PCR), *Journal of theoretical biology* **184**(4): 433–40.
URL: `http://www.ncbi.nlm.nih.gov/pubmed/9082073`

Tichopad, A., Dilger, M., Schwarz, G. and Pfaf, M. W. (2003). Standardized determination of real-time PCR efficiency from a single reaction set-up, *Nucleic Acids Research* **31**(20): 122e–122.
URL: `http://www.nar.oupjournals.org/cgi/doi/10.1093/nar/gng122`

Appendix A

Notation

f_{0i} The starting fluorescence level for curve i.

j Cycle, $1 \leq j \leq J$.

$\mathbf{f_{ji}}$ True fluorescence level in cycle j in curve i. $f_{ji} = \gamma x_{ij}$ where γ is a constant and x is the number of copies of the target DNA molecules.

s_i starting cycle for a curve i for the mathematical model in Equation 3.4.

m_i ending cycle for a curve i for the mathematical model in Equation 3.4.

A Case group.

B Control group.

G Gene of interest.

R Reference gene.

Clu,Clu1,Clu2,Clu3 Primer pairs for Cluesterin.

GA,RA,GB,RB,Clu,Clu1,Clu2,Clu3 Sample types.

$\mathbf{E = 1 + p}$ The amplification in a PCR cycle, also called the PCR efficiency, with $1 \leq E \leq 2$

p The probability that a member of the target DNA population is successfully duplicated, with $0 \leq p \leq 1$

T Threshold

log Log transform with a general base (as opposed to log_2, log_{10}, $ln = log_e$ where the base is emphasized)

CT Cycle corresponding to the threshold T.

$\mathbf{b_k}$ The value used for baseline correction at cycle k.

Biological triplicate Three samples with the same sample type.

Technical triplicate Three copies from the same sample.

Ground phase The section of the PCR curve where no amplification-specific fluorescence can yet be determined.

Exponential phase The section of the PCR curve after the ground phase. This is where the generated fluorescence exceeds baseline fluorescence, but reagents have not yet begun to be limited.

Noise Normally distributed disturbances in the fluorescence measure.

Rapid fluctuations Disturbances in the fluorescence measure for early cycles, which can not be explained by the amplification process.