# ORNSTEIN-UHLENBECK PROCESS AND OPTIMAL SAMPLING FOR ANALYSIS OF MICROBIOME DATA

by

Junqiu Gao

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
August 2019

# Contents

**List of Tables** . . . . . **iv**

**List of Figures** . . . . . **v**

**Abstract** . . . . . **vii**

**Acknowledgements** . . . . . **viii**

**Chapter 1 Introduction** . . . . . **1**

1.1 Background . . . . . 1

1.2 Thesis Structure . . . . . 2

**Chapter 2 Fitting an Ornstein-Uhlenbeck Process to Microbiome Data** . . . . . **3**

2.1 Microbiome Data . . . . . 3

2.2 Review of Ornstein-Uhlenbeck Process . . . . . 4
2.2.1 Review of Brownian Motion . . . . . 4
2.2.2 Ornstein-Uhlenbeck Process . . . . . 5
2.2.3 Taylor Expansion of OU Process Likelihood . . . . . 7

2.3 Suitability of OU Process for Modelling Microbial Dynamics . . . . . 9
2.3.1 Applying Likelihood Ratio Test between i.i.d Normal Distribution and OU Process to Microbiome Data . . . . . 11
2.3.2 Applying the Likelihood Ratio Test between Brownian Motion and OU Process to Microbiome Data . . . . . 13

**Chapter 3 Fisher Information of OU Mean Reverting Process and Optimal Sampling** . . . . . **23**

3.1 Review of Fisher Information . . . . . 23

3.2 Fisher Information Derivation for OU Mean Reverting Process . . . . 24
3.2.1 Observed Information . . . . . 24
3.2.2 Fisher Information . . . . . 28

3.3 Determining Optimal Sampling . . . . . 30
3.3.1 Fisher Information Matrix with Equal Space Sampling . . . . 30
3.3.2 Numerical Results . . . . . 33
3.3.3 Application to Moving Picture Data . . . . . 36

3.4 Simulation . . . . . 41
3.4.1 Simulation Design . . . . . 41
3.4.2 Simulation Results . . . . . 48

**Chapter 4 Discussion . . . . . 49**

**Bibliography . . . . . 50**

## List of Tables


Table 2.1 The number of observations for each individual and body site . 10

Table 2.2 The number of abundant genera for each individual and body site 10

Table 2.3 The proportion of abundant genera which reject the null hypothesis of i.i.d. Normal distribution for each individual and body site . . . . . 13

Table 3.1 Sample size and time difference for each simulation data set . . 48

Table 3.2 Distance of Fisher information inverse matrix and covariance . 48

## List of Figures


Figure 2.1 The time series plot of each abundant gut genus for Person 1 10

Figure 2.2 Likelihood ratio test between i.i.d normal distribution and OU mean reverting process for Person 1's gut and tongue genera . 14

Figure 2.3 Likelihood ratio test between i.i.d normal distribution and OU mean reverting process for Person 1's right and left palm genera 15

Figure 2.4 Likelihood ratio test between i.i.d normal distribution and OU mean reverting process for Person 2's gut and tongue genera . 16

Figure 2.5 Likelihood ratio test between i.i.d normal distribution and OU mean reverting process for Person 2's right and left palm genera 17

Figure 2.6 Likelihood ratio test between Brownian motion and OU mean reverting process for Person 1's gut and tongue genera . . . . 20

Figure 2.7 Likelihood ratio test between Brownian motion and OU mean reverting process for Person 1's right and left palm genera . . 21

Figure 2.8 Likelihood ratio test between Brownian motion and OU mean reverting process for Person 2's gut and tongue genera . . . . 22

Figure 2.9 Likelihood ratio test between Brownian motion and OU mean reverting process for Person 2's right and left palm genera . . 22

Figure 3.1 Relation between time difference $\Delta t$ and $Var(\hat{\eta})$ for various sample sizes $n$ for true parameter values $\eta = 0.5$, $\sigma = 0.01$, $x_0 = 0$, $\mu = 0$ . . . . . 35

Figure 3.2 Relation between sample sizes $n$ and time difference $\Delta t$ for true parameter values $\eta = 0.5$, $\sigma = 0.01$, $x_0 = 0$, $\mu = 0$ . . . . . 35

Figure 3.3 Optimal time difference $\Delta t$ as function of $\eta$ . . . . . 36

Figure 3.4 Distribution of $\hat{\eta}$ over genera for Person 1 . . . . . 37

Figure 3.5 Distribution of $\hat{\eta}$ over genera for Person 2 . . . . . 38

Figure 3.6 Distribution of $\hat{\mu}$ over genera for Person 1 . . . . . 39

Figure 3.7 Distribution of $\hat{\mu}$ over genera for Person 2 . . . . . 39

Figure 3.8 Distribution of $\hat{\sigma}$ over genera for Person 1 . . . . . 40

Figure 3.9 Distribution of $\hat{\sigma}$ over genera for Person 2 . . . . . 40

Figure 3.10 Variance of $\hat{\eta}$ as function of $\Delta t$ for different genera (different curves) in gut of Person 1 . . . . . 41

Figure 3.11 Variance of $\hat{\eta}$ as function of $\Delta t$ for different genera (different curves) in tongue of Person 1 . . . . . 42

Figure 3.12 Variance of $\hat{\eta}$ as function of $\Delta t$ for different genera (different curves) in right palm of Person 1 . . . . . 43

Figure 3.13 Variance of $\hat{\eta}$ as function of $\Delta t$ for different genera (different curves) in left palm of Person 1 . . . . . 44

Figure 3.14 Variance of $\hat{\eta}$ as function of $\Delta t$ for different genera (different curves) in gut of Person 2 . . . . . 45

Figure 3.15 Variance of $\hat{\eta}$ as function of $\Delta t$ for different genera (different curves) in tongue of Person 2 . . . . . 46

Figure 3.16 Variance of $\hat{\eta}$ as function of $\Delta t$ for different genera (different curves) in right palm of Person 2 . . . . . 47

## Abstract

The Ornstein–Uhlenbeck (OU) process is a widely used model for stochastic processes, where the value drifts towards a fixed stable value. We examine how well the OU process fits the data by using likelihood ratio tests to compare models of temporal dynamics of OTUs. Then, we derive the Fisher information of the OU process and show how it can be used to maximize the temporal efficiency of sampling. We apply this to parameters estimated from real data to determine optimal sampling schemes for human microbiomes. We use simulations to show that the asymptotic theory applies to typical finite sample cases.

## Acknowledgements

I would like to express the deepest appreciation to my supervisors Dr. Hong Gu and Dr. Toby Kenney. Without their encouragement and support, this thesis would hardly have been completed.

I would also like to show my appreciation to my thesis committee, Dr. Lam Ho and Dr. Edward Susko. I'm extremely grateful to their insightful and valuable comments on this thesis.

I owe a huge debt of gratitude to my family and friends. I will never achieve that far without their immense support, love, and patience.

## List of Abbreviations and Symbols Used

| | |
|---|---|
| ***OU*** | Ornstein-Uhlenbeck |
| ***OTU*** | Operational Taxonomic Unit |
| ***BM*** | Brownian Motion |
| ***SDE*** | Stochastic Differential Equation |
| ***MLE*** | Maximum Likelihood Estimation |
| ***FIM*** | Fisher Information Matrix |

# Chapter 1

# Introduction

## 1.1 Background

A significant number of microscopic organisms live in and around the human body. Recent research was shown that human microbiome plays a significant role in human health [16] [11] [1] [14].

Technological development in DNA sequencing has permitted a more systematic study of the microbiome [4]. There has been substantial work studying the instantaneous structure of the microbiome, but the temporal dynamics of the microbiome are largely unstudied. The studies that exist suggest the microbiome is generally stable. Since the microbiome is often considered as an ecological system, it is natural to model its temporal dynamics as a stochastic process. The observed stability suggests that a mean-reverting process may be appropriate. In this thesis, we compare a mean-reverting process with both random drift and a constant state with measurement error. For this thesis, we will focus on the dynamics of a single Operational Taxonomic Unit (OTU) and ignore interactions between different OTUs [3].

In order to study temporal dynamics, it is important to collect samples at the correct frequency. Sampling too frequently may result in not covering enough time to observe the patterns, while large gaps between samples can lead to consecutive samples being uncorrelated. There are various limitations on the number of samples that can be collected, such as the cost and availability of participants. Thus determining the optimal sample size and observation time frequency is very important for studying the microbial dynamics. Despite the recent advances in sequencing technology, there still exist some problems. For instance, we have to track our participants to collect samples and there is still a significant cost to DNA sequencing analysis. If we can derive an optimal sample size and observation time frequency, it can greatly improve efficiency and reduce both economic and time costs. This can be done asymptotically by computing the Fisher information matrix of our model.

## 1.2 Thesis Structure

In this thesis, we first use likelihood ratio tests to compare models of temporal dynamics of OTUs. The results of these tests support our hypotheses that a mean-reverting stochastic process is appropriate for many microbial time series (Chapter 2). In Chapter 3, we derive the Fisher information of the OU process and show how it can be used to maximize the temporal efficiency of sampling. We apply this to parameters estimated from real data to determine optimal sampling schemes for human microbiomes. Chapter 4 concludes the thesis by discussing the implications of our results and suggesting future work.

# Chapter 2

# Fitting an Ornstein-Uhlenbeck Process to Microbiome Data

## 2.1 Microbiome Data

Trillions of symbiotic microbes are hosted by every part of the human body. They constitute the microbiome including bacteria, viruses, and fungi. The microbiome can vary significantly between individuals who have different environments, diet, and behavior, and it plays a vital role in human health and disease. Some studies have proved that gut microbes are highly related to health and some diseases such as digestion function, obesity [1] and inflammatory bowel disease [14].

Technological development in DNA sequencing has permitted a more systematic study of the microbiome [4]. Marker gene analysis is widely used to target a specific genetic region and to determine the microbial phylogenies. Operational taxonomic units (OTUs) can be used to classify different sequences by their similarity. OTUs are widely used units of microbial diversity in 16S or 18S rRNA marker gene sequence data sets.

Many studies have already shown that microbiome is generally stable. Since the microbiome is often considered as an ecological system; it is natural to model its temporal dynamics as a stochastic process. In this thesis, we study moving picture data. In the moving picture data, two healthy individuals were sampled at four body sites (gut, tongue, left, and right palms) almost daily. Researchers observed one individual for 15 months, and the other for 6 months. Samples were sequenced using PCR on the V2 region of the 16S rRNA gene [4]. To avoid sparse counts, we aggregate the data at genus level in this thesis.

## 2.2 Review of Ornstein-Uhlenbeck Process

A large number of naturally occurring stochastic processes exhibit some form of mean reversion, where the value drifts towards a fixed stable value. Examples of this behaviour come from physics [13], finance [21] and biology [19] [17]. In this section, we introduce one of the simplest and most commonly used mean reversion models, namely the Ornstein-Uhlenbeck (OU) process.

### 2.2.1 Review of Brownian Motion

Brownian motion was originally introduced by Brown in 1827 to model the fast and irregular motion exhibited by tiny particles in fluid. A thorough introduction of Brownian motion can be found in [10].

A stochastic process $X_t, t \geq 0$ with state space $R$ is said to have a *stationary increments* if the distribution of the increment $X_{s+t} - X_s$ over the interval $(s, s+t]$ depends only on the length of the interval $t$. A stochastic process $X_t, t \geq 0$ with state space $R$ is said to have an *independent increments* if the increments $X_{s+t} - X_s$ over non-overlapping intervals are independent [20].

A stochastic process $W_t, t \geq 0$ with state space $R$ is said to be a *Standard Brownian Motion* (also called a Wiener process) if $W_t, t \geq 0$ has stationary and independent increments and $W_t \sim N(0, t)$ for $t \geq 0$.

*Brownian motion* (BM) is a generalization of standard Brownian motion.
Let $W_t, t \geq 0$ be a standard Brownian motion. A stochastic process $X_t, t \geq 0$ given by

$$X_t = x_0 + \mu t + \sigma dW_t, t \geq 0$$

is called a Brownian motion with drift parameter $\mu \in R$, variance parameter $\sigma > 0$, and starting point $x_0 \in R$. We denote this BM by $BM(\mu, \sigma)$.

Brownian motion has the following properties:

1. $X_t, t \geq 0$ has stationary and independent increments.
2. $X_{s+t} - X_s \sim N(\mu t, \sigma^2 t)$, $s, t \geq 0$, and in particular $X_t \sim N(x_0 + \mu t, \sigma^2 t)$

Based on these properties, maximum likelihood can be used to estimate the drift and variance parameters. The log likelihood for a set of observed $X(t)$ values, $x_0, x_1, \ldots x_n$ corresponding to times $t_0, \cdots t_n$, is

$$l(x; \mu, \sigma) = -\frac{1}{2}\sum_{i=1}^{n} log(2\pi\sigma^2(t_i - t_{i-1})) - \sum_{i=1}^{n} \frac{[x_i - (x_{i-1} + \mu(t_i - t_{i-1}))]^2}{2\sigma^2(t_i - t_{i-1})}$$

Setting the first derivatives of this equal to 0, we get the following maximum likelihood estimates:

$$\hat{\mu} = \frac{x_n - x_1}{t_n - t_1}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{[x_i - (x_{i-1} + \hat{\mu}(t_i - t_{i-1}))]^2}{t_i - t_{i-1}}\right]$$

### 2.2.2 Ornstein-Uhlenbeck Process

An Ornstein-Uhlenbeck (OU) Process is a stochastic process. It was first introduced by Leonard Ornstein and George Eugene Uhlenbeck [22] and is widely used in the fields of physics and finance. The OU process was developed based on Brownian motion.

The OU process $X_t$ is defined by the following linear stochastic differential equation (SDE)

$$dX_t = \eta(\mu - X_t)dt + \sigma dW_t$$

Where $\eta > 0$ is the velocity of the reversion process and $\mu$ is the long-term average. When $\eta$ increases, the process will revert more quickly back to its mean value. $W_t$ is a Wiener process, which is the diffusion part of the OU process.

This stochastic differential equation is solved by Oksendal [15] with the following results

$$X_t = e^{-\eta t}X_0 + \mu(1 - e^{-\eta t}) + \sigma\int_0^t e^{\eta(s-t)}dW_s$$

Where $X_0$ is the variable status at $t = 0$ and $\int_0^t e^{\eta(s-t)}dW_s \sim N(0, \frac{1-e^{-2\eta t}}{2\eta})$

The values $X_t$ therefore follow a multivariate normal distribution with conditional expectation and variance given by the following equations [18].

$$E[X_t|X_0] = \mu + (x_0 - \mu)e^{-\eta t}$$

$$Var[X_t|X_0] = \frac{\sigma^2}{2\eta}(1 - e^{-2\eta t})$$

$$Cov[X_t, X_s] = \frac{\sigma^2}{2\eta}(e^{-\eta|t-s|} - e^{-\eta(t+s)})$$

Since we know the expectation and variance of every $X_i$ under the OU mean-reverting process, it is not hard to get the following results given $X_0 = x_0$,

$$\begin{aligned} E[X_t^2] =& Var[X_t] + (E[X_t])^2 \\ =& \frac{\sigma^2}{2\eta}(1 - e^{-2\eta t}) + (\mu + (x_0 - \mu)e^{-\eta t})^2 \end{aligned}$$

$$\begin{aligned} E[X_t X_s] =& Cov[X_t, X_s] + E[X_t]E[X_s] \\ =& \frac{\sigma^2}{2\eta}(e^{-\eta|t-s|} - e^{-\eta(t+s)}) + (\mu + (x_0 - \mu)e^{-\eta t})(\mu + (x_0 - \mu)e^{-\eta s}) \end{aligned}$$

The log likelihood function is calculated by Franco [8] and maximum likelihood can be used to estimate the three parameters $\mu$, $\eta$ and $\sigma$. The log-likelihood function is

$$\begin{aligned} l(\boldsymbol{x}; \mu, \eta, \sigma) = &- \frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{\sigma^2}{2\eta}\right) - \frac{1}{2}\sum_{i=1}^{n}\log(1 - e^{-2\eta(t_i - t_{i-1})}) \\ &- \frac{\eta}{\sigma^2}\sum_{i=1}^{n}\frac{(x_i - \mu - (x_{i-1} - \mu)e^{-\eta(t_i - t_{i-1})})^2}{1 - e^{-2\eta(t_i - t_{i-1})}} \end{aligned} \tag{2.1}$$

Using the first order condition, for fixed $\hat{\eta}$ we can estimate $\mu$ and $\sigma^2$ by estimators $\hat{\mu}$ and $\hat{\sigma}^2$ given by the following functions of $\hat{\eta}$ [8].

$$\hat{\mu} = f(\hat{\eta}) = \sum_{i=1}^{n}\frac{x_i - x_{i-1}e^{-\hat{\eta}(t_i - t_{i-1})}}{1 + e^{-\hat{\eta}(t_i - t_{i-1})}}\left(\sum_{i=1}^{n}\frac{1 - e^{-\hat{\eta}(t_i - t_{i-1})}}{1 + e^{-\hat{\eta}(t_i - t_{i-1})}}\right)^{-1} \tag{2.2}$$

$$\hat{\sigma}^2 = g(\hat{\mu}, \hat{\eta}) = \frac{2\hat{\eta}}{n} \sum_{i=1}^{n} \frac{(x_i - \hat{\mu} - (x_{i-1} - \hat{\mu})e^{-\hat{\eta}(t_i - t_{i-1})})^2}{1 - e^{-2\hat{\eta}(t_i - t_{i-1})}} \tag{2.3}$$

Therefore, plugging in the estimates $\hat{\mu}$ and $\hat{\sigma}^2$, the profile log likelihood function becomes the following function of $\eta$.

$$\begin{aligned} V(\eta) = &- \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\frac{g(f(\eta), \eta)}{2\eta}) - \frac{1}{2}\sum_{i=1}^{n}\log(1 - e^{-2\eta(t_i - t_{i-1})}) \\ &- \frac{\eta}{g(f(\eta), \eta)} \sum_{i=1}^{n} \frac{(x_i - f(\eta) - (x_{i-1} - f(\eta))e^{-\eta(t_i - t_{i-1})})^2}{1 - e^{-2\eta(t_i - t_{i-1})}} \end{aligned} \tag{2.4}$$

We can find the MLE $\hat{\eta}$ using an exhaustive grid search by (2.4), and use Equations (2.2) and (2.3) to get MLE estimates $\hat{\mu}$ and $\hat{\sigma}^2$.

### 2.2.3 Taylor Expansion of OU Process Likelihood

When $\eta$ is close to zero, the formulae (2.2)-(2.4) are numerically unstable. We therefore replace the unstable parts of them with Taylor series expansions.

Let $d_i = t_i - t_{i-1}$. We expand the first five terms of the Taylor expansions

$$e^{-\eta d_i} \approx 1 - \eta d_i + \frac{\eta^2 d_i^2}{2!} - \frac{\eta^3 d_i^3}{3!} + \frac{\eta^4 d_i^4}{4!}$$

$$1 - e^{-\eta d_i} \approx \eta d_i - \frac{\eta^2 d_i^2}{2!} + \frac{\eta^3 d_i^3}{3!} - \frac{\eta^4 d_i^4}{4!} = N_i$$

$$\frac{1 - e^{-2\eta d_i}}{2\eta} \approx d_i - \frac{2\eta d_i^2}{2!} + \frac{4\eta^2 d_i^3}{3!} - \frac{8\eta^3 d_i^4}{4!} = M_i$$

The approximations of $\hat{\mu}$ and $\hat{\sigma}^2$ are

$$\hat{\mu}_{\text{Taylor}} = \sum_{i=1}^{n} \frac{x_i - x_{i-1}e^{-\hat{\eta}d_i}}{1 + e^{-\hat{\eta}d_i}} \left( \sum_{i=1}^{n} \frac{N_i}{1 + e^{-\hat{\eta}d_i}} \right)^{-1}$$

$$\hat{\sigma}^2_{\text{Taylor}} = \frac{1}{n} \sum_{i=1}^{n} M_i^{-1} (x_i - \hat{\mu}_{\text{Taylor}} - (x_{i-1} - \hat{\mu}_{\text{Taylor}})e^{-\hat{\eta}d_i})^2$$

Therefore, the Taylor expansion of the profile log-likelihood can be expressed as:

$$V(\eta)_{\text{Taylor}} = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log\left(\frac{\hat{\sigma}^2_{\text{Taylor}}(1-e^{-2\eta d_i})}{2\eta}\right) - \frac{\eta}{\hat{\sigma}^2_{\text{Taylor}}}\sum_{i=1}^{n}\frac{(x_i - \hat{\mu}_{\text{Taylor}} - (x_{i-1} - \hat{\mu}_{\text{Taylor}})e^{-\eta d_i})^2}{1-e^{-2\eta d_i}} \tag{2.5}$$

To get the approximation for the profile log likelihood of OU process, we can examine the second and the third terms in Equation (2.5).

$$\begin{aligned}\{Second\ term\} &= -\frac{1}{2}\sum_{i=1}^{n}\log\left(\frac{\hat{\sigma}^2_{\text{Taylor}}(1-e^{-2\eta d_i})}{2\eta}\right)\\ &\approx -\frac{1}{2}\left[\sum_{i=1}^{n}\log(\hat{\sigma}^2_{\text{Taylor}}) + \sum_{i=1}^{n}\log(d_i) + \sum_{i=1}^{n}\log\left(1 - \frac{2\eta d_i}{2!} + \frac{4\eta^2 d_i^2}{3!} - \frac{8\eta^3 d_i^3}{4!}\right)\right]\end{aligned}$$

Let $P_i = -\frac{2\eta d_i}{2!} + \frac{4\eta^2 d_i^2}{3!} - \frac{8\eta^3 d_i^3}{4!}$. Taylor expansion of the last term gives $\sum_{i=1}^{n}\log(1 - \frac{2\eta d_i}{2!} + \frac{4\eta^2 d_i^2}{3!} - \frac{8\eta^3 d_i^3}{4!}) = \sum_{i=1}^{n}\log(1+P_i) \approx \sum_{i=1}^{n}(P_i - \frac{P_i^2}{2} + \frac{P_i^3}{3} - \frac{P_i^4}{4})$

So, the second term of (2.5) is:

$$\{Second\ term\} = -\frac{1}{2}\left[\sum_{i=1}^{n}\log(\hat{\sigma}^2_{Taylor}) + \sum_{i=1}^{n}\log(d_i) + \sum_{i=1}^{n}\left(P_i - \frac{P_i^2}{2} + \frac{P_i^3}{3} - \frac{P_i^4}{4}\right)\right]$$

$$\begin{aligned}\{Third\ term\} &= -\frac{\eta}{\hat{\sigma}^2_{\text{Taylor}}}\sum_{i=1}^{n}\frac{(x_i - \hat{\mu}_{\text{Taylor}} - (x_{i-1} - \hat{\mu}_{\text{Taylor}})e^{-\eta d_i})^2}{1-e^{-2\eta d_i}}\\ &= -\frac{\eta}{\hat{\sigma}^2_{\text{Taylor}}}\sum_{i=1}^{n}\frac{((x_i - x_{i-1}) + (1-e^{-\eta d_i})(x_{i-1} - \hat{\mu}_{\text{Taylor}}))^2}{1-e^{-2d_i}}\\ &= -\frac{\eta}{\hat{\sigma}^2_{\text{Taylor}}}\sum_{i=1}^{n}\left(\frac{(x_i - x_{i-1})^2}{1-e^{-2\eta d_i}}\right) - \frac{\eta}{\hat{\sigma}^2_{\text{Taylor}}}\sum_{i=1}^{n}\left(\frac{2(x_i - x_{i-1})(x_{i-1} - \hat{\mu}_{\text{Taylor}})}{1+e^{-\eta d_i}}\right)\\ &\quad - \frac{\eta}{\hat{\sigma}^2_{\text{Taylor}}}\sum_{i=1}^{n}\left(\frac{(x_{i-1} - \hat{\mu}_{\text{Taylor}})^2(1-e^{-\eta d_i})}{1+e^{-\eta d_i}}\right)\end{aligned}$$

And

$$\begin{aligned}-\frac{\eta}{\hat{\sigma}^2_{\text{Taylor}}}\sum_{i=1}^{n}\left(\frac{(x_i - x_{i-1})^2}{1-e^{-2\eta d_i}}\right) &= -\frac{1}{\hat{\sigma}^2_{\text{Taylor}}}\sum_{i=1}^{n}\left(\frac{\eta(x_i - x_{i-1})^2}{1-e^{-2\eta d_i}}\right)\\ &\approx -\frac{1}{2\hat{\sigma}^2_{\text{Taylor}}}\sum_{i=1}^{n}\left(M_i^{-1}(x_i - x_{i-1})^2\right)\end{aligned}$$

So, the third term of (2.5) is:

$$\{Third\ term\} = -\frac{1}{2\hat{\sigma}^2_{\text{Taylor}}}\sum_{i=1}^{n}\left(M_i^{-1}(x_i - x_{i-1})^2\right) - \frac{\eta}{\hat{\sigma}^2_{Taylor}}\sum_{i=1}^{n}\left(\frac{2(x_i - x_{i-1})(x_{i-1} - \hat{\mu}_{\text{Taylor}})}{1 + e^{-\eta d_i}}\right) - \frac{\eta}{\hat{\sigma}^2_{\text{Taylor}}}\sum_{i=1}^{n}\left(\frac{(x_{i-1} - \hat{\mu}_{\text{Taylor}})^2(1 - e^{-\eta d_i})}{1 + e^{-\eta d_i}}\right)$$

Therefore, the approximate form of the profile log likelihood function of the OU process can be expressed as

$$V(\eta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\left[(\sum_{i=1}^{n}\log(\hat{\sigma}^2_{\text{Taylor}}) + \sum_{i=1}^{n}\log(d_i) + \sum_{i=1}^{n}(P_i - \frac{P_i^2}{2} + \frac{P_i^3}{3} - \frac{P_i^4}{4})\right] - \frac{1}{2\hat{\sigma}^2_{\text{Taylor}}}\sum_{i=1}^{n}\left(M_i^{-1}(x_i - x_{i-1})^2\right) - \frac{\eta}{\hat{\sigma}^2_{\text{Taylor}}}\sum_{i=1}^{n}\left(\frac{2(x_i - x_{i-1})(x_{i-1} - \hat{\mu}_{\text{Taylor}})}{1 + e^{-\eta d_i}}\right) - \frac{\eta}{\hat{\sigma}^2_{\text{Taylor}}}\sum_{i=1}^{n}\left(\frac{(x_{i-1} - \hat{\mu}_{\text{Taylor}})^2(1 - e^{-\eta d_i})}{1 + e^{-\eta d_i}}\right)$$

By maximizing the above Taylor expansion of $V(\eta)$, we can get a more stable solution of $\hat{\eta}$.

### 2.3 Suitability of OU Process for Modelling Microbial Dynamics

In this section, we assess the suitability of the OU process for modelling real microbial data, in comparison with the following two alternatives based on the likelihood ratio tests.

1. Time independence

2. Brownian motion without drift

For this purpose, we use the moving picture data set [3]. This data set follows two healthy individuals over 6-month and 15-month periods respectively. Four body sites were observed: gut, tongue, right palm and left palm. Samples are not collected at completely regular time intervals. Many samples are taken at daily intervals, but many intervals of multiple days are also present. Samples were sequenced using PCR on the V2 region of the 16S rRNA gene [4].

Since the OTU counts are sparse, we aggregate them up to genus level. We take the proportion (Each observation in each genus is divided by total count of all genera) of

Table 2.1: The number of observations for each individual and body site

| | Gut | Tongue | Right Palm | Left Palm |
|---|---|---|---|---|
| Person 1 | 131 | 135 | 134 | 134 |
| Person 2 | 336 | 373 | 359 | 365 |

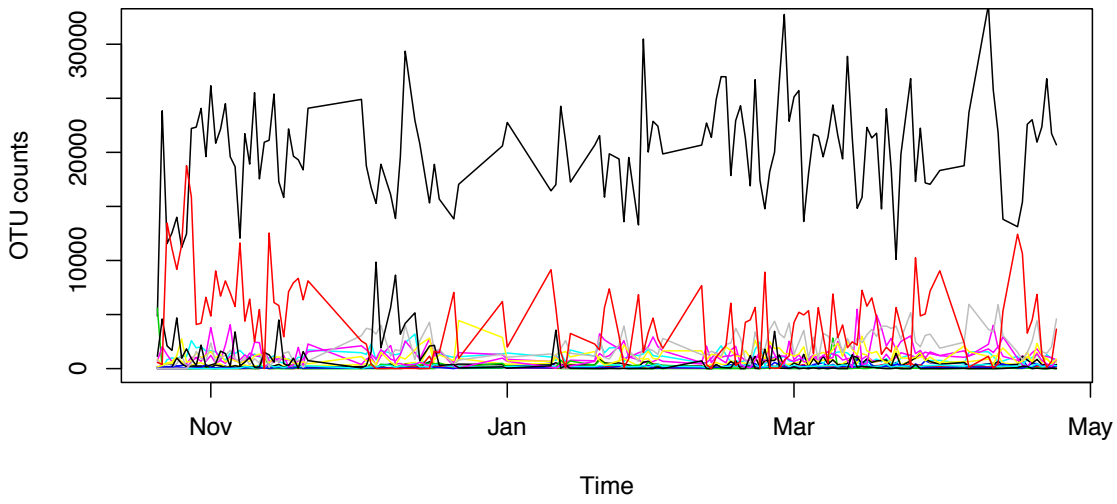


Figure 2.1: The time series plot of each abundant gut genus for Person 1

each genus to correct for variation in sequencing depth. We also focus our attention on the most abundant genera. We select all genera with total count of the genus greater than 10000 for individual 1 and 20000 for individual 2. Table 2.1 and Table 2.2 show the number of observations and abundant genera for each individual and each body site respectively. Figure 2.1 exhibits the time series plot of each abundant gut genus for Person 1.

Table 2.2: The number of abundant genera for each individual and body site

| | Gut | Tongue | Right Palm | Left Palm |
|---|---|---|---|---|
| Person 1 | 17 | 11 | 12 | 45 |
| Person 2 | 32 | 18 | 59 | 29 |

### 2.3.1 Applying Likelihood Ratio Test between i.i.d Normal Distribution and OU Process to Microbiome Data

We first check the time dependence for each person, body site and genus. The abundant genera should have time dependence to fit the OU process model. Therefore we compare the log-likelihood of an OU process model, and an i.i.d. normal model where we assume the abundances at different time points are independent and normally distributed. For the OU process, we optimize $\eta$ using a grid search in the range $0 - 50$ with the step size 0.001

The hypothesis test is

$$H_0 : X_t \ follows \ i.i.d. \ Normal \ distribution$$

$$H_1 : X_t \ follows \ an \ OU \ mean \ reverting \ process$$

First we need to prove that this log-likelihood ratio statistic is invariant under any linear transformations of the data, i. e.

$$l(\boldsymbol{x}; \mu_{ou}, \eta, \sigma_{ou}) - l(\boldsymbol{x}; \mu_N, \sigma_N) = l(\boldsymbol{y}; \mu_{ou}, \eta, \sigma_{ou}) - l(\boldsymbol{y}; \mu_N, \sigma_N)$$

Where $y$ is the linear transformation of $x$, which is $y = ax + b$. Therefore $\mu_y = a\mu_x + b$ and $\sigma_y^2 = a^2\sigma_x^2$.

We know the log-likelihood function of normal distribution is

$$l(\boldsymbol{x}; \mu_N, \sigma_N^2) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma_N^2) - \frac{1}{2\sigma_N^2}\sum_{i=1}^{n}(x_i - \mu_N)^2$$

Then,

$$
\begin{aligned}
& l(\boldsymbol{y};\mu_{ou},\eta,\sigma_{ou}) - l(\boldsymbol{y};\mu_N,\sigma_N) \\
=& -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{\sigma^2_{OU.y}}{2\eta}\right) - \frac{1}{2}\sum_{i=1}^{n}\log(1-e^{-2\eta(t_i-t_{i-1})}) \\
& - \frac{\eta}{\sigma^2_{OU.y}}\sum_{i=1}^{n}\frac{(y_i-\mu_{OU.y}-(y_{i-1}-\mu_{OU.y})e^{-\eta(t_i-t_{i-1})})^2}{1-e^{-2\eta(t_i-t_{i-1})}} \\
& + \frac{n}{2}log(2\pi) + \frac{n}{2}log(\sigma^2_{N.y}) + \frac{1}{2\sigma^2_{N.y}}\sum_{i=1}^{n}(y_i-\mu_{N.y})^2 \\
=& -\frac{n}{2}\log(\sigma^2_{OU.y}) + \frac{n}{2}log(\sigma^2_{N.y}) + \frac{n}{2}\log(2\eta) - \frac{1}{2}\sum_{i=1}^{n}\log(1-e^{-2\eta(t_i-t_{i-1})}) \\
& - \frac{\eta}{\sigma^2_{OU.y}}\sum_{i=1}^{n}\frac{(y_i-\mu_{OU.y}-(y_{i-1}-\mu_{OU.y})e^{-\eta(t_i-t_{i-1})})^2}{1-e^{-2\eta(t_i-t_{i-1})}} + \frac{1}{2\sigma^2_{N.y}}\sum_{i=1}^{n}(y_i-\mu_{N.y})^2
\end{aligned}
$$

Where

$$
\begin{aligned}
& -\frac{n}{2}\log(\sigma^2_{OU.y}) + \frac{n}{2}log(\sigma^2_{N.y}) \\
=& -\frac{n}{2}\log(a^2\sigma^2_{OU.x}) + \frac{n}{2}log(a^2\sigma^2_{N.x}) \\
=& -\frac{n}{2}\log(\sigma^2_{OU.x}) + \frac{n}{2}log(\sigma^2_{N.x})
\end{aligned}
$$

And

$$
\begin{aligned}
& -\frac{\eta}{\sigma^2_{OU.y}}\sum_{i=1}^{n}\frac{(y_i-\mu_{OU.y}-(y_{i-1}-\mu_{OU.y})e^{-\eta(t_i-t_{i-1})})^2}{1-e^{-2\eta(t_i-t_{i-1})}} + \frac{1}{2\sigma^2_{N.y}}\sum_{i=1}^{n}(y_i-\mu_{N.y})^2 \\
=& -\frac{\eta}{a^2\sigma^2_{OU.x}}\sum_{i=1}^{n}\frac{(ax_i+b-(a\mu_{OU.x}+b)-(ax_{i-1}+b-(a\mu_{OU.x}+b))e^{-\eta(t_i-t_{i-1})})^2}{1-e^{-2\eta(t_i-t_{i-1})}} \\
& + \frac{1}{2\sigma^2_{N.x}}\sum_{i=1}^{n}(ax_i+b-(a\mu_{N.x}+b))^2 \\
=& -\frac{\eta}{\sigma^2_{OU.x}}\sum_{i=1}^{n}\frac{(x_i-\mu_{OU.x}-(x_{i-1}-\mu_{OU.x})e^{-\eta(t_i-t_{i-1})})^2}{1-e^{-2\eta(t_i-t_{i-1})}} + \frac{1}{2\sigma^2_{N.x}}\sum_{i=1}^{n}(x_i-\mu_{N.x})^2
\end{aligned}
$$

Therefore,

$$
l(\boldsymbol{x};\mu_{ou},\eta,\sigma_{ou}) - l(\boldsymbol{x};\mu_N,\sigma_N) = l(\boldsymbol{y};\mu_{ou},\eta,\sigma_{ou}) - l(\boldsymbol{y};\mu_N,\sigma_N)
$$

We see that the re-scaled data will not influence the likelihood ratio test results so we can use standard normal distribution to calculate the null distribution.

Since the normal distribution is a limit case ($\eta \to \infty$, $\frac{\sigma_{OU}^2}{2\eta} \to \sigma^2$) of the OU process, the likelihood ratio statistic is not guaranteed to follow the usual $\chi^2$ distribution. We, therefore, use a simulation to estimate the null distribution. We simulate 5000 data sets using the same time points as the original data, under a standard normal distribution. The likelihood ratios for each genus, along with the null distribution for each data set, are shown in Figure 2.2 and Figure 2.3 (for Person 1) and Figure 2.4 and Figure 2.5 (for Person 2).

From the tables and figures, we see that many of the abundant genera show strong evidence of dependence between different time points, particularly in more enclosed body sites, such as the gut. More exposed body sites show less evidence of temporal dependence, which makes intuitive sense because exposure to external influences is expected to reduce the stability of the microbial ecosystem.

Table 2.3: The proportion of abundant genera which reject the null hypothesis of i.i.d. Normal distribution for each individual and body site

| | Gut | Tongue | Right Palm | Left Palm |
|---|---|---|---|---|
| Person 1 | 14/17 | 7/11 | 4/12 | 16/45 |
| Person 2 | 23/32 | 17/18 | 39/59 | 0/29 |

### 2.3.2 Applying the Likelihood Ratio Test between Brownian Motion and OU Process to Microbiome Data

In this section, we will examine whether the data have mean reversion. Therefore, we use a likelihood ratio test to determine whether the OU process fits the data better than Brownian motion. Since the data are compositional, the drift parameter for Brownian motion is set to 0, to avoid a model which will eventually violate the constraints.

Because the Brownian motion is a non-identifiable case of an OU process, the standard $\chi^2$ distribution does not apply. We, therefore, use simulation to empirically estimate the null distribution. We simulate 5000 data sets under Brownian motion with the starting point $x_0 = 0$, mean $\mu = 0$ and variance equal to the Brownian motion variance of genus T546 in Person 2's gut data. Since the Brownian motion

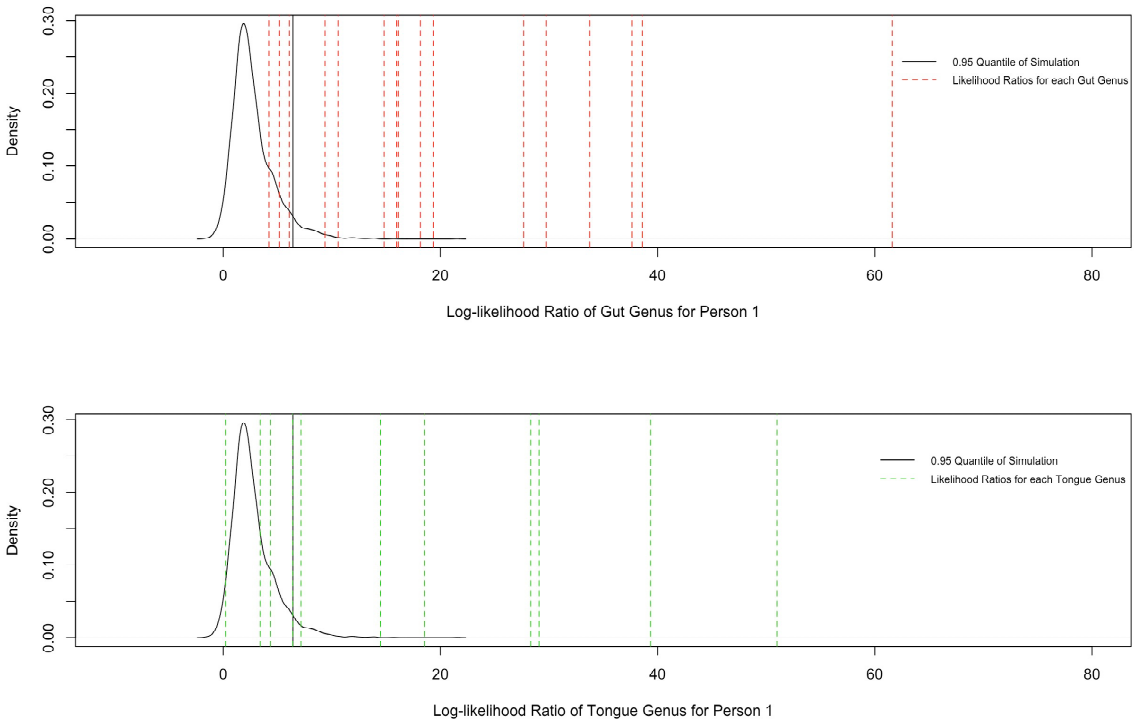


Figure 2.2: Likelihood ratio test between i.i.d normal distribution and OU mean reverting process for Person 1's gut and tongue genera

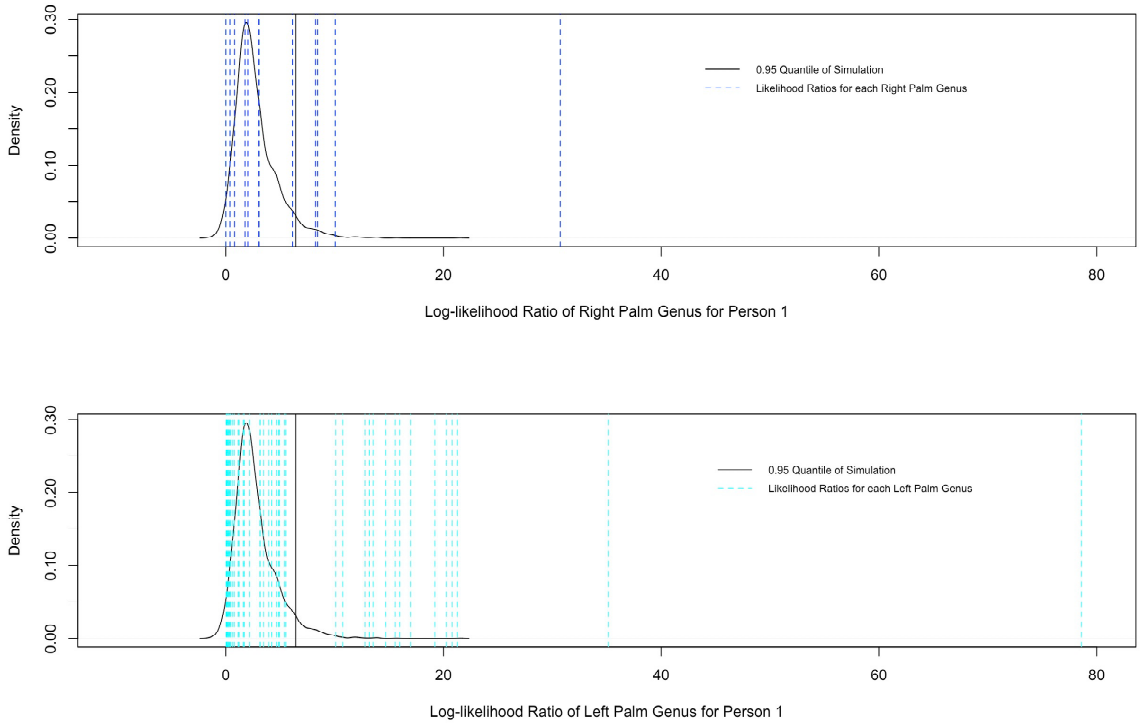


Figure 2.3: Likelihood ratio test between i.i.d normal distribution and OU mean reverting process for Person 1's right and left palm genera

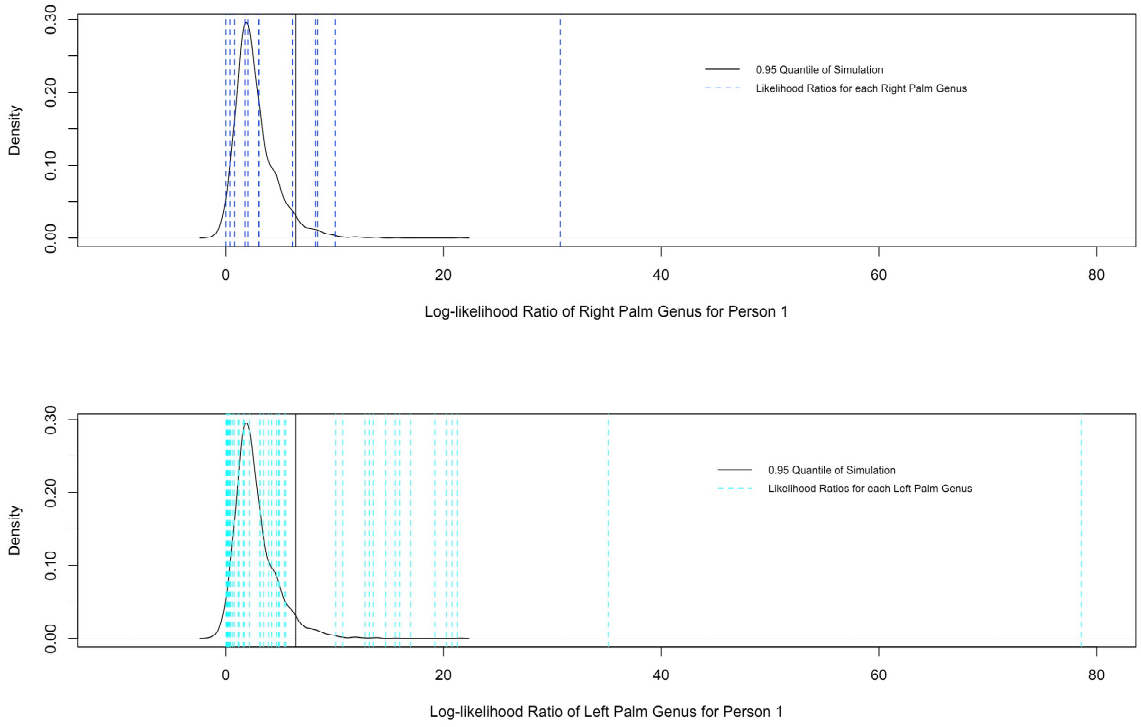


Figure 2.4: Likelihood ratio test between i.i.d normal distribution and OU mean reverting process for Person 2's gut and tongue genera

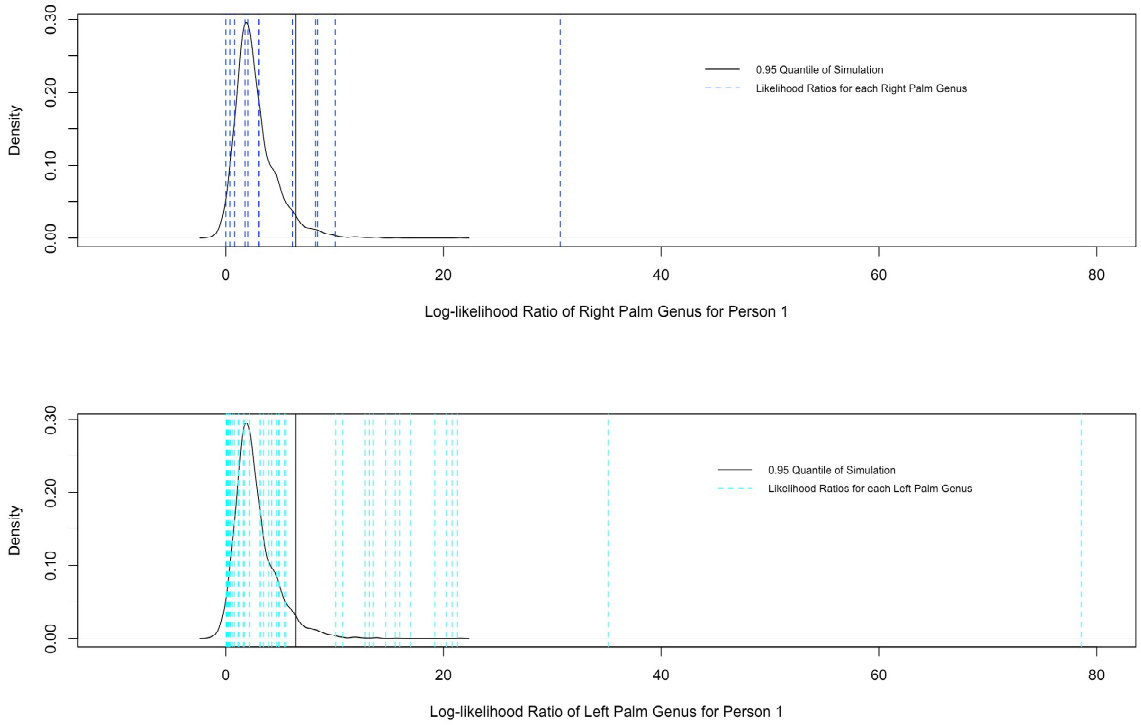


Figure 2.5: Likelihood ratio test between i.i.d normal distribution and OU mean reverting process for Person 2's right and left palm genera

and the OU process are both translation and scale distributions, the choice of the starting points and the values of $\sigma^2$ should not affect the values of the likelihood ratio statistics.

Because the data are simulated under Brownian motion, the $\eta$ estimated for the OU mean-reverting process should be close to zero. To avoid rounding errors we use the Taylor expansion approximation given in Section 2.2.3 to estimate the parameters and log-likelihood.

The null and alternative hypotheses are

$$H_0 : X_t \ follows \ Brownian \ motion \ with \ \mu = 0$$

$$H_1 : X_t \ follows \ an \ OU \ mean \ reverting \ process$$

Similarly, we need to prove that

$$l(\boldsymbol{x}; \mu_{ou}, \eta, \sigma_{ou}) - l(\boldsymbol{x}; \mu_{BM}, \sigma_{BM}) = l(\boldsymbol{y}; \mu_{ou}, \eta, \sigma_{ou}) - l(\boldsymbol{y}; \mu_{BM}, \sigma_{BM})$$

Where $y$ is the linear transformation of $x$, which is $y = ax + b$. Therefore $\mu_y = a\mu_x + b$ and $\sigma_y^2 = a^2\sigma_x^2$.

We know the log-likelihood function of Brownian motion without drift is

$$l(\boldsymbol{x}; \sigma_{BM}) = -\frac{1}{2}\sum_{i=1}^{n} log(2\pi\sigma_{BM}^2(t_i - t_{i-1})) - \sum_{i=1}^{n}\frac{(x_i - x_{i-1})^2}{2\sigma_{BM}^2(t_i - t_{i-1})}$$

$$\begin{aligned}
&l(\boldsymbol{y}; \mu_{ou}, \eta, \sigma_{ou}) - l(\boldsymbol{y}; \mu_{BM}, \sigma_{BM}) \\
=& -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{\sigma_{OU.y}^2}{2\eta}\right) - \frac{1}{2}\sum_{i=1}^{n}\log(1 - e^{-2\eta(t_i - t_{i-1})}) \\
&- \frac{\eta}{\sigma_{OU.y}^2}\sum_{i=1}^{n}\frac{(y_i - \mu_{OU.y} - (y_{i-1} - \mu_{OU.y})e^{-\eta(t_i - t_{i-1})})^2}{1 - e^{-2\eta(t_i - t_{i-1})}} \\
&+ \frac{1}{2}\sum_{i=1}^{n} log(2\pi\sigma_{BM.y}^2(t_i - t_{i-1})) + \sum_{i=1}^{n}\frac{(y_i - y_{i-1})^2}{2\sigma_{BM.y}^2(t_i - t_{i-1})} \\
=& -\frac{n}{2}\log(\sigma_{OU.y}^2) + \frac{n}{2}log(\sigma_{BM.y}^2) + \frac{n}{2}\log(2\eta) - \frac{1}{2}\sum_{i=1}^{n}\log(1 - e^{-2\eta(t_i - t_{i-1})}) \\
&- \frac{\eta}{\sigma_{OU.y}^2}\sum_{i=1}^{n}\frac{(y_i - \mu_{OU.y} - (y_{i-1} - \mu_{OU.y})e^{-\eta(t_i - t_{i-1})})^2}{1 - e^{-2\eta(t_i - t_{i-1})}} + \sum_{i=1}^{n}\frac{(y_i - y_{i-1})^2}{2\sigma_{BM.y}^2(t_i - t_{i-1})}
\end{aligned}$$

Where

$$
\begin{aligned}
&- \frac{n}{2}\log(\sigma^2_{OU.y}) + \frac{n}{2} log(\sigma^2_{BM.y}) \\
=&- \frac{n}{2}\log(a^2\sigma^2_{OU.x}) + \frac{n}{2} log(a^2\sigma^2_{BM.x}) \\
=&- \frac{n}{2}\log(\sigma^2_{OU.x}) + \frac{n}{2} log(\sigma^2_{BM.x})
\end{aligned}
$$

And,

$$
\begin{aligned}
&- \frac{\eta}{\sigma^2_{OU.y}} \sum_{i=1}^{n} \frac{(y_i - \mu_{OU.y} - (y_{i-1} - \mu_{OU.y})e^{-\eta(t_i - t_{i-1})})^2}{1 - e^{-2\eta(t_i - t_{i-1})}} + \sum_{i=1}^{n} \frac{(y_i - y_{i-1})^2}{2\sigma^2_{BM.y}(t_i - t_{i-1})} \\
=&- \frac{\eta}{a^2\sigma^2_{OU.x}} \sum_{i=1}^{n} \frac{(ax_i + b - (a\mu_{OU.x} + b) - (ax_{i-1} + b - (a\mu_{OU.x} + b))e^{-\eta(t_i - t_{i-1})})^2}{1 - e^{-2\eta(t_i - t_{i-1})}} \\
&+ \sum_{i=1}^{n} \frac{(ax_i + b - (ax_{i-1} + b))^2}{2a^2\sigma^2_{BM.x}(t_i - t_{i-1})} \\
=&- \frac{\eta}{\sigma^2_{OU.x}} \sum_{i=1}^{n} \frac{(x_i - \mu_{OU.x} - (x_{i-1} - \mu_{OU.x})e^{-\eta(t_i - t_{i-1})})^2}{1 - e^{-2\eta(t_i - t_{i-1})}} + \sum_{i=1}^{n} \frac{(x_i - x_{i-1})^2}{2\sigma^2_{BM.x}(t_i - t_{i-1})}
\end{aligned}
$$

Therefore

$$
l(\boldsymbol{x}; \mu_{ou}, \eta, \sigma_{ou}) - l(\boldsymbol{x}; \mu_{BM}, \sigma_{BM}) = l(\boldsymbol{y}; \mu_{ou}, \eta, \sigma_{ou}) - l(\boldsymbol{y}; \mu_{BM}, \sigma_{BM})
$$

We see that the re-scaled data will not influence the likelihood ratio test results so we can choose any Brownian motion to calculate the null distribution.

The likelihood ratio statistics for all abundant genera in each body site are shown in Figure 2.6 and Figure 2.7 (for Person 1) and Figure 2.8 and Figure 2.9 (for Person 2), along with the null distribution and critical values. We find that all the log-likelihood ratio tests for the real genus data reject the null hypothesis at the 5% significance level. This indicates that all abundant genera are subject to some mean reversion.

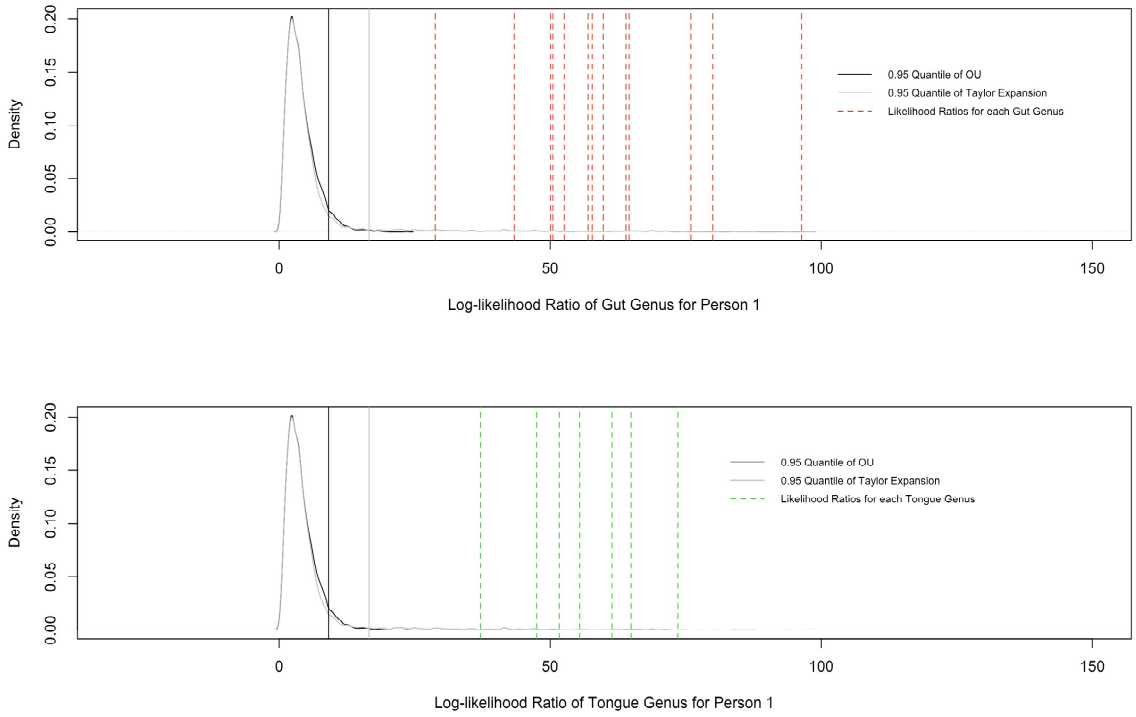


Figure 2.6: Likelihood ratio test between Brownian motion and OU mean reverting process for Person 1's gut and tongue genera

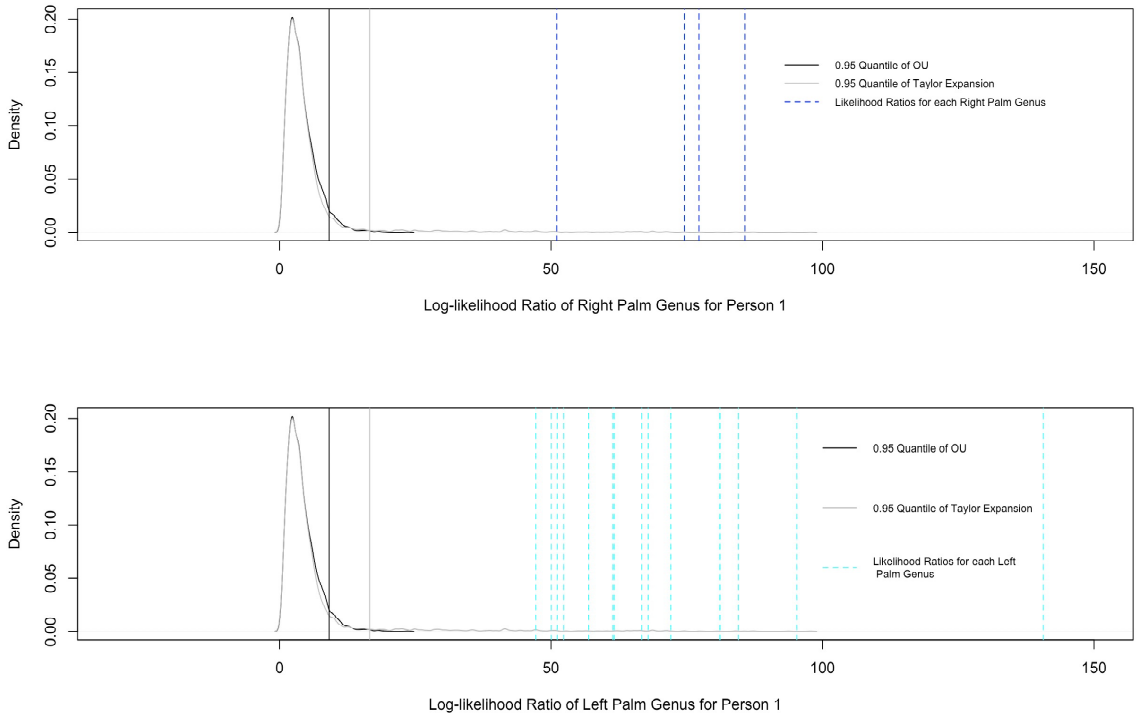


Figure 2.7: Likelihood ratio test between Brownian motion and OU mean reverting process for Person 1's right and left palm genera

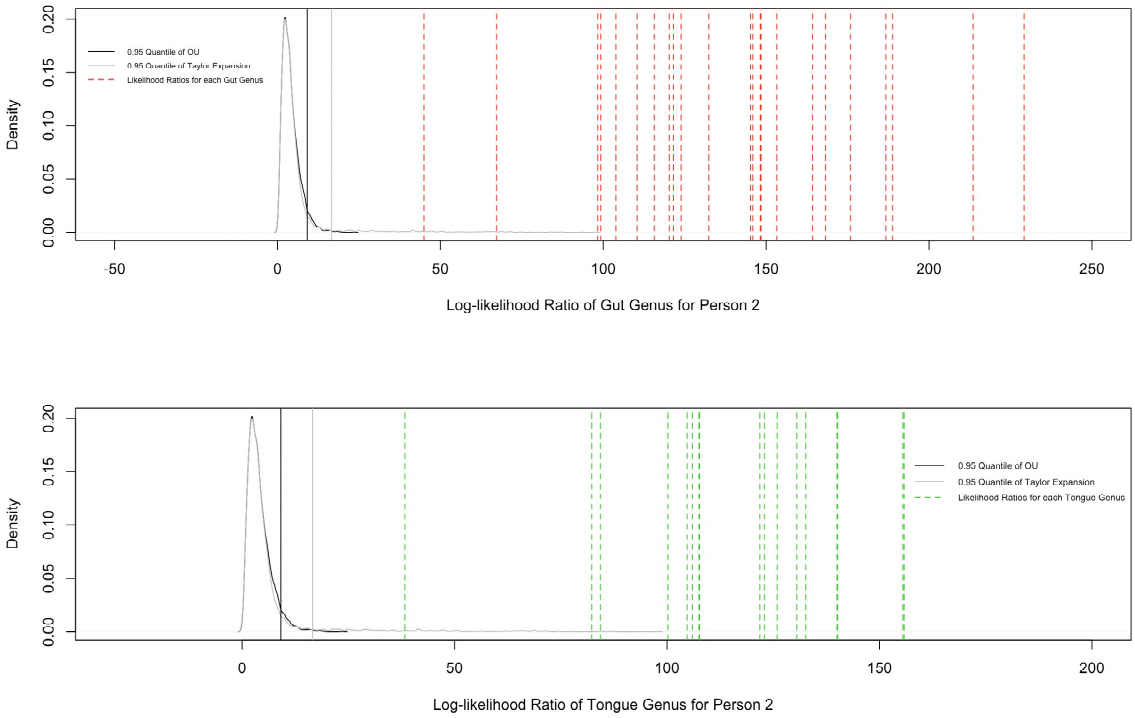


Figure 2.8: Likelihood ratio test between Brownian motion and OU mean reverting process for Person 2's gut and tongue genera

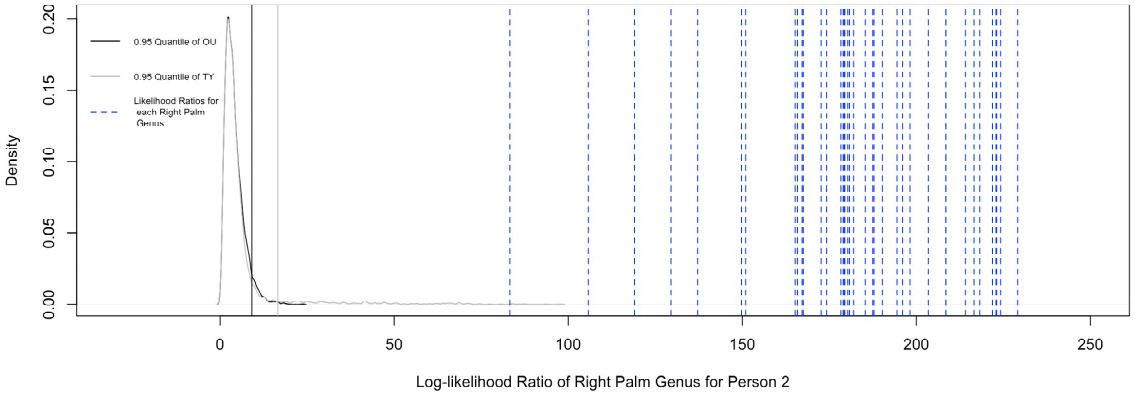


Figure 2.9: Likelihood ratio test between Brownian motion and OU mean reverting process for Person 2's right and left palm genera

# Chapter 3

# Fisher Information of OU Mean Reverting Process and Optimal Sampling

In this chapter, we consider the accuracy of the estimated parameters, and based on this, we also consider the most efficient sampling scheme for estimating the parameters with particular attention to the mean reversion velocity of the OU process. The asymptotic behaviour of MLEs for parameters is controlled by the Fisher information. We begin by reviewing the theory of Fisher information in Section 3.1. In Section 3.2 and 3.3, we derive the Fisher information matrix for an OU process and use it to determine optimal sampling frequency. In Section 3.4, we use simulations to show that the asymptotic theory applies to typical finite sample cases.

## 3.1 Review of Fisher Information

Fisher Information introduced by Ronald Fisher [12] is an important tool in asymptotic theory, used to measure the amount of information in the sample data. A review of its use can be found in [12] [9] . The main use of Fisher information for the purposes of this theory is the following theorem about the asymptotic behaviour of MLEs [2] [5].

**Theorem (Asymptotic normality of MLE)** *Let $\{f(x|\theta) : \theta \in \Omega\}$ be a parametric model, where $\theta \in \mathbb{R}$ is a single parameter. Let $X_1, X_2, \ldots, X_n$ be i.i.d. with $f(x|\theta)$ for $\theta_0 \in \Omega$ and let $\hat{\theta}$ be the MLE based on $X_1, \ldots, X_n$. Suppose certain regularity conditions hold, including:*

1. *All PDFs/PMFs $f(x|\theta)$ in the model have the same support*
2. *$\theta_0$ is an interior point (i.e., not on the boundary) of $\Omega$*
3. *The log-likelihood $l(\theta)$ is twice differentiable in $\theta$*

*4. $\hat{\theta}$ is the unique value of $\theta \in \Omega$ that score function can be 0*

*Then $\hat{\theta}$ is consistent and asymptotically normal, with*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathrm{d}} N(0, I^{-1}(\theta_0))$$

*A good explanation of this theorem and its proof can be found in [2].*

*Fisher information is a way of measuring the amount of information that a random variable $X$ contains about an unknown parameter $\theta$ for the distribution of $X$. It is the expectation of the observed information [6]. The Fisher information can be derived as*

$$I(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2} l(\boldsymbol{x};\theta)\right]$$

*Where $-\frac{\partial^2}{\partial\theta^2} l(\boldsymbol{x};\theta) = J_n(\theta)$ is the observed information which is the negative of the second derivative of the log likelihood function with respect to $\theta$, based on the sample size $n$.*

*When there are $N$ parameters, so that $\boldsymbol{\theta}$ is an $N \times 1$ vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_N]^T$, the Fisher information takes the form of an $N \times N$ matrix. This matrix is called the Fisher information matrix (FIM) and has typical element*

$$[I(\boldsymbol{\theta})]_{i,j} = -E\left[\frac{\partial^2}{\partial\theta_i\theta_j} L(\boldsymbol{x};\boldsymbol{\theta})\right]$$

*Similarly, we can derive the following results for multiple parameters.*

$$\hat{\boldsymbol{\theta}} \overset{\text{app.}}{\sim} N(\boldsymbol{\theta}, I^{-1}(\boldsymbol{\theta}))$$

## 3.2 Fisher Information Derivation for OU Mean Reverting Process

### 3.2.1 Observed Information

*To derive the Fisher information of the OU process, we have to first calculate the observed information. We compute:*

$$
\begin{aligned}
-\frac{\partial^2}{\partial\mu^2}l(\boldsymbol{x};\mu,\eta,\sigma) =& -\frac{\partial^2}{\partial\mu^2}\left[-\frac{\eta}{\sigma^2}\sum_{i=1}^{n}\frac{(x_i-\mu-(x_{i-1}-\mu)e^{-\eta d_i})^2}{1-e^{-2\eta d_i}}\right] \\
=& -\frac{\partial}{\partial\mu}\left[\frac{2\eta}{\sigma^2}\sum_{i=1}^{n}\frac{(x_i-\mu-(x_{i-1}-\mu)e^{-\eta d_i})}{(1+e^{-\eta d_i})}\right] \\
=& -\frac{2\eta}{\sigma^2}\sum_{i=1}^{n}\frac{-1+e^{-\eta d_i}}{1+e^{-\eta d_i}}
\end{aligned}
\tag{3.1}
$$

$$
\begin{aligned}
-\frac{\partial^2}{\partial\mu\partial\sigma}l(\boldsymbol{x};\mu,\eta,\sigma) =& -\frac{\partial^2}{\partial\mu\partial\sigma}\left[-\frac{\eta}{\sigma^2}\sum_{i=1}^{n}\frac{(x_i-\mu-(x_{i-1}-\mu)e^{-\eta d_i})^2}{1-e^{-2\eta d_i}}\right] \\
=& \frac{\partial}{\partial\sigma}\left[\frac{2\eta}{\sigma^2}\sum_{i=1}^{n}\frac{(x_i-\mu-(x_{i-1}-\mu)e^{-\eta d_i})}{1+e^{-\eta d_i}}\right] \\
=& \frac{4\eta}{\sigma^3}\sum_{i=1}^{n}\frac{(x_i-\mu-(x_{i-1}-\mu)e^{-\eta d_i})}{1+e^{-\eta d_i}}
\end{aligned}
\tag{3.2}
$$

*In order to simplify our expression for observed information, we can set* $(x_i-\mu-(x_{i-1}-\mu)e^{-\eta d_i}) = B_i$. *Therefore (3.2) becomes*

$$
-\frac{\partial^2}{\partial\mu\partial\sigma}l(\boldsymbol{x};\mu,\eta,\sigma) = \frac{4\eta}{\sigma^3}\sum_{i=1}^{n}\frac{B_i}{1+e^{-\eta d_i}}
$$

$$
\begin{aligned}
-\frac{\partial^2}{\partial\sigma^2}l(\boldsymbol{x};\mu,\eta,\sigma) =& -\frac{\partial^2}{\partial\sigma^2}\left[-\frac{n}{2}\log(\frac{\sigma^2}{2\eta})-\frac{\eta}{\sigma^2}\sum_{i=1}^{n}\frac{(x_i-\mu-(x_{i-1}-\mu)e^{-\eta d_i})^2}{1-e^{-2\eta d_i}}\right] \\
=& -\frac{\partial}{\partial\sigma}\left[-\frac{n}{\sigma}+\frac{2\eta}{\sigma^3}\sum_{i=1}^{n}\frac{(x_i-\mu-(x_{i-1}-\mu)e^{-\eta d_i})^2}{1-e^{-2\eta d_i}}\right] \\
=& -\frac{n}{\sigma^2}+\frac{6\eta}{\sigma^4}\sum_{i=1}^{n}\frac{(x_i-\mu-(x_{i-1}-\mu)e^{-\eta d_i})^2}{1-e^{-2\eta d_i}} \\
=& -\frac{n}{\sigma^2}+\frac{6\eta}{\sigma^4}\sum_{i=1}^{n}\frac{B_i^2}{1-e^{-2\eta d_i}}
\end{aligned}
\tag{3.3}
$$

$$
\begin{aligned}
-\frac{\partial^2}{\partial\mu\partial\eta}l(\boldsymbol{x};\mu,\eta,\sigma) =& -\frac{\partial^2}{\partial\mu\partial\eta}\left[-\frac{\eta}{\sigma^2}\sum_{i=1}^{n}\frac{(x_i-\mu-(x_{i-1}-\mu)e^{-\eta d_i})^2}{1-e^{-2\eta d_i}}\right] \\
=& -\frac{\partial}{\partial\eta}\left[\frac{2\eta}{\sigma^2}\sum_{i=1}^{n}\frac{(x_i-\mu-(x_{i-1}-\mu)e^{-\eta d_i})}{1+e^{-\eta d_i}}\right] \\
=& -\frac{2}{\sigma^2}\sum_{i=1}^{n}\frac{B_i}{1+e^{-\eta d_i}} - \frac{2\eta}{\sigma^2}\frac{\partial}{\partial\eta}\left[\sum_{i=1}^{n}\frac{B_i}{1+e^{-\eta d_i}}\right]
\end{aligned}
\tag{3.4}
$$

*We calculate*

$$
\begin{aligned}
\frac{\partial}{\partial\eta}\left[\sum_{i=1}^{n}\frac{B_i}{1+e^{-\eta d_i}}\right] =& \sum_{i=1}^{n}\frac{(d_i(x_{i-1}-\mu)e^{-\eta d_i})(1+e^{-\eta d_i})+d_ie^{-\eta d_i}B_i}{(1+e^{-\eta d_i})^2} \\
=& \sum_{i=1}^{n}\frac{d_ie^{-\eta d_i}(x_{i-1}-\mu)+d_ie^{-\eta d_i}(x_i-\mu)}{(1+e^{-\eta d_i})^2}
\end{aligned}
$$

$$
\begin{aligned}
-\frac{\partial^2}{\partial\sigma\partial\eta}l(\boldsymbol{x};\mu,\eta,\sigma) =& -\frac{\partial^2}{\partial\sigma\partial\eta}\left[-\frac{n}{2}\log(\frac{\sigma^2}{2\eta})-\frac{\eta}{\sigma^2}\sum_{i=1}^{n}\frac{(x_i-\mu-(x_{i-1}-\mu)e^{-\eta d_i})^2}{1-e^{-2\eta d_i}}\right] \\
=& -\frac{\partial}{\partial\eta}\left[-\frac{n}{\sigma}+\frac{2\eta}{\sigma^3}\sum_{i=1}^{n}\frac{(x_i-\mu-(x_{i-1}-\mu)e^{-\eta d_i})^2}{1-e^{-2\eta d_i}}\right] \\
=& -\frac{2}{\sigma^3}\sum_{i=1}^{n}\frac{B_i^2}{1-e^{-2\eta d_i}} - \frac{2\eta}{\sigma^3}\frac{\partial}{\partial\eta}\left[\sum_{i=1}^{n}\frac{B_i^2}{1-e^{-2\eta d_i}}\right]
\end{aligned}
\tag{3.5}
$$

*we calculate*

$$
\begin{aligned}
\frac{\partial}{\partial\eta}\left[\sum_{i=1}^{n}\frac{B_i^2}{1-e^{-2\eta d_i}}\right] =& 2\left[\sum_{i=1}^{n}\frac{B_i(d_i(x_{i-1}-\mu)e^{-\eta d_i})}{1-e^{-2\eta d_i}}\right] - 2\left[\sum_{i=1}^{n}\frac{d_ie^{-2\eta d_i}B_i^2}{(1-e^{-2\eta d_i})^2}\right] \\
=& 2\left[\sum_{i=1}^{n}\frac{B_id_i[(x_{i-1}-\mu)e^{-\eta d_i}-(x_i-\mu)e^{-2\eta d_i}]}{(1-e^{-2\eta d_i})^2}\right]
\end{aligned}
$$

$$
\begin{aligned}
-\frac{\partial^2}{\partial\eta^2}l(\boldsymbol{x};\mu,\eta,\sigma) =& \frac{\partial^2}{\partial\eta^2}\left[\frac{n}{2}\log(\frac{\sigma^2}{2\eta}) + \frac{1}{2}\sum_{i=1}^{n}\log(1-e^{-2\eta d_i}) + \frac{\eta}{\sigma^2}\sum_{i=1}^{n}\frac{B_i^2}{1-e^{-2\eta d_i}}\right] \\
=& \frac{n}{2\eta^2} - 2\sum_{i=1}^{n}\frac{d_i^2 e^{-2\eta d_i}}{(1-e^{-2\eta d_i})^2} + \frac{2}{\sigma^2}\frac{\partial}{\partial\eta}\left[\sum_{i=1}^{n}\frac{B_i^2}{1-e^{-2\eta d_i}}\right] \\
&+ \frac{\eta}{\sigma^2}\frac{\partial^2}{\partial\eta^2}\left[\sum_{i=1}^{n}\frac{B_i^2}{1-e^{-2\eta d_i}}\right]
\end{aligned}
\tag{3.6}
$$

*And*

$$
\begin{aligned}
&\frac{\partial^2}{\partial\eta^2}\left[\sum_{i=1}^{n}\frac{B_i^2}{1-e^{-2\eta d_i}}\right] \\
=&2\frac{\partial}{\partial\eta}\left[\sum_{i=1}^{n}\frac{B_i(d_i(x_{i-1}-\mu)e^{-\eta d_i})}{(1-e^{-2\eta d_i})}\right] - 2\frac{\partial}{\partial\eta}\left[\sum_{i=1}^{n}\frac{d_i e^{-2\eta d_i}B_i^2}{(1-e^{-2\eta d_i})^2}\right] \\
=&2\left[\sum_{i=1}^{n}\frac{(x_{i-1}-\mu)^2 e^{-2\eta d_i}d_i^2}{(1-e^{-2\eta d_i})} - \sum_{i=1}^{n}\frac{B_i d_i^2(x_{i-1}-\mu)e^{-\eta d_i}}{(1-e^{-2\eta d_i})} - 2\sum_{i=1}^{n}\frac{e^{-3\eta d_i}d_i^2 B_i(x_{i-1}-\mu)}{(1-e^{-2\eta d_i})^2}\right] \\
&- 2\left[\sum_{i=1}^{n}\frac{-2d_i^2 e^{-2\eta d_i}B_i^2 + 2d_i^2 e^{-3\eta d_i}B_i(x_{i-1}-\mu)}{(1-e^{-2\eta d_i})^2} - \sum_{i=1}^{n}\frac{4e^{-4\eta d_i}d_i^2 B_i^2}{(1-e^{-2\eta d_i})^3}\right] \\
=&2\left[\sum_{i=1}^{n}\frac{(x_{i-1}-\mu)e^{-\eta d_i}d_i^2[2(x_{i-1}-\mu)e^{-\eta d_i} - (x_i-\mu)]}{(1-e^{-2\eta d_i})}\right] \\
&- 4\left[\sum_{i=1}^{n}\frac{d_i^2 e^{-2\eta d_i}B_i[3e^{-\eta d_i}(x_{i-1}-\mu) - (x_i-\mu)]}{(1-e^{-2\eta d_i})^2}\right] + 8\left[\sum_{i=1}^{n}\frac{e^{-4\eta d_i}d_i^2 B_i^2}{(1-e^{-2\eta d_i})^3}\right]
\end{aligned}
$$

*So the observed information matrix will be*

$$
J = -\begin{bmatrix}
\frac{\partial^2}{\partial\mu^2}l(\boldsymbol{x};\mu,\eta,\sigma) & \frac{\partial^2}{\partial\mu\partial\eta}l(\boldsymbol{x};\mu,\eta,\sigma) & \frac{\partial^2}{\partial\sigma\partial\mu}l(\boldsymbol{x};\mu,\eta,\sigma) \\
\frac{\partial^2}{\partial\mu\partial\eta}l(\boldsymbol{x};\mu,\eta,\sigma) & \frac{\partial^2}{\partial\eta^2}l(\boldsymbol{x};\mu,\eta,\sigma) & \frac{\partial^2}{\partial\sigma\partial\eta}l(\boldsymbol{x};\mu,\eta,\sigma) \\
\frac{\partial^2}{\partial\sigma\partial\mu}l(\boldsymbol{x};\mu,\eta,\sigma) & \frac{\partial^2}{\partial\sigma\partial\eta}l(\boldsymbol{x};\mu,\eta,\sigma) & \frac{\partial^2}{\partial\sigma^2}l(\boldsymbol{x};\mu,\eta,\sigma)
\end{bmatrix}
$$

### 3.2.2 Fisher Information

*From Equations (3.1)-(3.6) it is straightforward to calculate the Fisher information by taking expectation for each term:*

$$\begin{aligned}
[I(\theta)]_{\mu,\mu} =& - E\left[\frac{\partial^2}{\partial\mu^2} l(\boldsymbol{x};\mu,\eta,\sigma)\right] \\
=& - E\left[\frac{2\eta}{\sigma^2}\sum_{i=1}^{n}\frac{-1+e^{-\eta d_i}}{1+e^{-\eta d_i}}\right] \\
=& \frac{2\eta}{\sigma^2}\sum_{i=1}^{n}\frac{1-e^{-\eta d_i}}{1+e^{-\eta d_i}}
\end{aligned} \tag{3.7}$$

$$\begin{aligned}
[I(\theta)]_{\sigma,\sigma} =& - E\left[\frac{\partial^2}{\partial\sigma^2} l(\boldsymbol{x};\mu,\eta,\sigma)\right] \\
=& - E\left[\frac{n}{\sigma^2} - \frac{6\eta}{\sigma^4}\sum_{i=1}^{n}\frac{B_i^2}{1-e^{-2\eta d_i}}\right] \\
=& - \frac{n}{\sigma^2} + \frac{6\eta}{\sigma^4}\sum_{i=1}^{n}\frac{E(B_i^2)}{1-e^{-2\eta d_i}}
\end{aligned} \tag{3.8}$$

*Using* $E[B_i^2] = Var[B_i] + (E[B_i])^2$*, it is not hard to derive that* $E[B_i] = 0$ *and* $Var[B_i] = \frac{\sigma^2}{2\eta}(1-e^{-2\eta d_i})$*. Therefore,* $E[B_i^2] = \frac{\sigma^2}{2\eta}(1-e^{-2\eta d_i})$*.*

$$\begin{aligned}
[I(\theta)]_{\eta,\eta} =& - E\left[\frac{\partial^2}{\partial\eta^2} l(\boldsymbol{x};\mu,\eta,\sigma)\right] \\
=& \frac{n}{2\eta^2} - 2\sum_{i=1}^{n}\frac{d_i^2 e^{-2\eta d_i}}{(1-e^{-2\eta d_i})^2} + \frac{2}{\sigma^2}\left[\sum_{i=1}^{n}\frac{E[-2d_i e^{-2\eta d_i}B_i^2]}{(1-e^{-2\eta d_i})^2}\right] \\
&+ \frac{2\eta}{\sigma^2}\sum_{i=1}^{n}\frac{(E[x_{i-1}^2] - 2E[x_{i-1}]\mu + \mu^2)e^{-2\eta d_i}d_i^2}{(1-e^{-2\eta d_i})} \\
&+ \frac{2\eta}{\sigma^2}\sum_{i=1}^{n}\frac{E[2d_i^2 e^{-2\eta d_i}B_i^2]}{(1-e^{-2\eta d_i})^2} + \frac{2\eta}{\sigma^2}\sum_{i=1}^{n}\frac{E[4d_i^2 e^{-4\eta d_i}B_i^2]}{(1-e^{-2\eta d_i})^3}
\end{aligned}$$

$$
\begin{aligned}
[I(\theta)]_{\mu,\sigma} &= -E\left[\frac{\partial^2}{\partial\mu\partial\sigma}L(\boldsymbol{x};\mu,\eta,\sigma)\right] \\
&= -E\left[-\frac{4\eta}{\sigma^3}\sum_{i=1}^{n}\frac{B_i}{1+e^{-\eta d_i}}\right] \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
[I(\theta)]_{\mu,\eta} &= -E\left[\frac{\partial^2}{\partial\mu\partial\eta}l(\boldsymbol{x};\mu,\eta,\sigma)\right] \\
&= -E\left[\frac{2}{\sigma^2}\sum_{i=1}^{n}\frac{B_i}{1+e^{-\eta d_i}}\right] \\
&\quad -E\left[\frac{2\eta}{\sigma^2}\sum_{i=1}^{n}\frac{(d_i(x_{i-1}-\mu)e^{-\eta d_i})(1+e^{-\eta d_i})+d_ie^{-\eta d_i}B_i}{(1+e^{-\eta d_i})^2}\right] \\
&= -\frac{2\eta}{\sigma^2}\left[\sum_{i=1}^{n}\frac{E[(d_i(x_{i-1}-\mu)e^{-\eta d_i})]}{(1+e^{-\eta d_i})}\right] \\
&= -\frac{2\eta}{\sigma^2}\left[\sum_{i=1}^{n}\frac{(d_i(x_0-\mu)e^{-\eta t_i})}{(1+e^{-\eta d_i})}\right] \\
[I(\theta)]_{\sigma,\eta} &= -E\left[\frac{\partial^2}{\partial\sigma\partial\eta}l(\boldsymbol{x};\mu,\eta,\sigma)\right] \\
&= -E\left[\frac{2}{\sigma^3}\sum_{i=1}^{n}\frac{B_i^2}{1-e^{-2\eta d_i}}+\frac{2\eta}{\sigma^3}\frac{\partial}{\partial\eta}\left[\sum_{i=1}^{n}\frac{B_i^2}{1-e^{-2\eta d_i}}\right]\right] \\
&= -\frac{2}{\sigma^3}\sum_{i=1}^{n}\frac{E[B_i^2]}{1-e^{-2\eta d_i}}-\frac{2\eta}{\sigma^3}\left[\sum_{i=1}^{n}\frac{-2d_ie^{-2\eta d_i}E[B_i^2]}{(1-e^{-2\eta d_i})^2}\right]
\end{aligned}
$$

*So the Fisher information matrix will be*

$$
\begin{aligned}
I &= \begin{bmatrix}
[I(\theta)]_{\mu,\mu} & [I(\theta)]_{\mu,\eta} & [I(\theta)]_{\mu,\sigma} \\
[I(\theta)]_{\mu,\eta} & [I(\theta)]_{\eta,\eta} & [I(\theta)]_{\eta,\sigma} \\
[I(\theta)]_{\mu,\sigma} & [I(\theta)]_{\eta,\sigma} & [I(\theta)]_{\sigma,\sigma}
\end{bmatrix} \\
&= \begin{bmatrix}
\frac{2\eta}{\sigma^2}\sum_{i=1}^{n}\frac{1-e^{-\eta d_i}}{1+e^{-\eta d_i}} & -\frac{2\eta}{\sigma^2}\sum_{i=1}^{n}\frac{(d_i(x_0-\mu)e^{-\eta t_i})}{(1+e^{-\eta d_i})} & 0 \\
-\frac{2\eta}{\sigma^2}\sum_{i=1}^{n}\frac{(d_i(x_0-\mu)e^{-\eta t_i})}{(1+e^{-\eta d_i})} & [I(\theta)]_{\eta,\eta} & -\frac{n}{\sigma\eta}+\frac{2}{\sigma}\sum_{i=1}^{n}\frac{d_ie^{-2\eta d_i}}{(1-e^{-2\eta d_i})} \\
0 & -\frac{n}{\sigma\eta}+\frac{2}{\sigma}\sum_{i=1}^{n}\frac{d_ie^{-2\eta d_i}}{(1-e^{-2\eta d_i})} & \frac{2n}{\sigma^2}
\end{bmatrix}
\end{aligned}
$$

*Where*

$$
\begin{aligned}
[I(\theta)]_{\eta,\eta} =& \frac{n}{2\eta^2} - 2\sum_{i=1}^{n} \frac{d_i^2 e^{-2\eta d_i}}{(1-e^{-2\eta d_i})^2} - \frac{2}{\eta}\sum_{i=1}^{n} \frac{d_i e^{-2\eta d_i}}{(1-e^{-2\eta d_i})} + 2\sum_{i=1}^{n} \frac{d_i^2 e^{-2\eta d_i}}{(1-e^{-2\eta d_i})} \\
&+ \frac{2\eta}{\sigma^2}\sum_{i=1}^{n} \frac{\left[\frac{\sigma^2}{2\eta}(1-e^{-2\eta t_{i-1}}) + (x_0-\mu)^2 e^{-2\eta t_{i-1}}\right] e^{-2\eta d_i} d_i^2}{(1-e^{-2\eta d_i})} + 4\sum_{i=1}^{n} \frac{d_i^2 e^{-4\eta d_i}}{(1-e^{-2\eta d_i})^2}
\end{aligned}
$$

## 3.3 Determining Optimal Sampling

### 3.3.1 Fisher Information Matrix with Equal Space Sampling

*Fisher information allows us to estimate the variance of our parameter estimates. This is very useful for designing experiments. To simplify this problem, we will restrict attention to equally spaced time points for all i ($d_i = t_i - t_{i-1} = \Delta t$). Fisher information will be greatly simplified. For example, we can calculate the optimal sampling scheme for given parameter values. The simplified Fisher information matrix is*

$$
\begin{aligned}
[I(\theta)]_{\eta,\eta} =& \frac{n}{2\eta^2} - 2\sum_{i=1}^{n} \frac{d_i^2 e^{-2\eta d_i}}{(1-e^{-2\eta d_i})^2} + \frac{2}{\sigma^2}\left[\sum_{i=1}^{n} \frac{E[-2d_i e^{-2\eta d_i} B^2]}{(1-e^{-2\eta d_i})^2}\right] \\
&+ \frac{2\eta}{\sigma^2}\sum_{i=1}^{n} \frac{(E[x_{i-1}^2] - 2E[x_{i-1}]\mu + \mu^2) e^{-2\eta d_i} d_i^2}{(1-e^{-2\eta d_i})} \\
&+ \frac{2\eta}{\sigma^2}\sum_{i=1}^{n} \frac{E[2d_i^2 e^{-2\eta d_i} B^2]}{(1-e^{-2\eta d_i})^2} + \frac{2\eta}{\sigma^2}\sum_{i=1}^{n} \frac{E[4d_i^2 e^{-4\eta d_i} B^2]}{(1-e^{-2\eta d_i})^3} \\
=& \frac{n}{2\eta^2} - 2n\frac{\Delta t^2 e^{-2\eta\Delta t}}{(1-e^{-2\eta\Delta t})^2} - \frac{4n\Delta t e^{-2\eta\Delta t}(\frac{\sigma^2}{2\eta})}{\sigma^2(1-e^{-2\eta\Delta t})} \\
&+ \frac{2\eta}{\sigma^2}\sum_{i=1}^{n} \frac{(\frac{\sigma^2}{2\eta}(1-e^{-2\eta t_{i-1}}) + (x_0-\mu)^2 e^{-2\eta t_{i-1}}) e^{-2\eta\Delta t}\Delta t^2}{(1-e^{-2\eta\Delta t})} \\
&+ \frac{2n\eta}{\sigma^2}\frac{2\Delta t^2 e^{-2\eta\Delta t}(\frac{\sigma^2}{2\eta})}{(1-e^{-2\eta\Delta t})} + \frac{2n\eta}{\sigma^2}\frac{4\Delta t^2 e^{-4\eta\Delta t}(\frac{\sigma^2}{2\eta})}{(1-e^{-2\eta\Delta t})^2}
\end{aligned}
\tag{3.9}
$$

*Suppose the observed process is in stationary state, we can take expectation $E[(x_0-\mu)^2] = \frac{\sigma^2}{2\eta}$. Set $\frac{1}{1-e^{-2\eta\Delta t}} = G_1(\Delta t) = G_1$ so (3.9) becomes*

$$
\begin{aligned}
[I(\theta)]_{\eta,\eta} =& \frac{n}{2\eta^2} - 2n\frac{\Delta t^2 e^{-2\eta\Delta t}}{(1-e^{-2\eta\Delta t})^2} - \frac{4n\Delta t e^{-2\eta\Delta t}(\frac{\sigma^2}{2\eta})}{\sigma^2(1-e^{-2\eta\Delta t})} \\
&+ \frac{2n\eta}{\sigma^2}\frac{\frac{\sigma^2}{2\eta}e^{-2\eta\Delta t}\Delta t^2}{(1-e^{-2\eta\Delta t})} + \frac{2n\eta}{\sigma^2}\frac{2\Delta t^2 e^{-2\eta\Delta t}(\frac{\sigma^2}{2\eta})}{(1-e^{-2\eta\Delta t})} \\
&+ \frac{2n\eta}{\sigma^2}\frac{4\Delta t^2 e^{-4\eta\Delta t}(\frac{\sigma^2}{2\eta})}{(1-e^{-2\eta\Delta t})^2} \\
=& \frac{n}{2\eta^2} - 2n\Delta t^2 G_1(G_1-1) - \frac{2n}{\eta}\Delta t(G_1-1) + 3n\Delta t^2(G_1-1) \\
&+ 4n\Delta t^2(G_1-1)^2 \\
=& \frac{n}{2\eta^2} + n\Delta t^2(G_1-1)(2G_1-1) - \frac{2n}{\eta}\Delta t(G_1-1)
\end{aligned}
\tag{3.10}
$$

$$
\begin{aligned}
[I(\theta)]_{\mu,\mu} =& -\frac{2\eta}{\sigma^2}\sum_{i=1}^{n}\frac{-1+e^{-\eta d_i}}{1+e^{-\eta d_i}} = -\frac{2n\eta}{\sigma^2}\left(\frac{-1+e^{-\eta\Delta t}}{1+e^{-\eta\Delta t}}\right) \\
=& \left[\frac{2\eta}{\sigma^2}(\frac{1-e^{-\eta\Delta t}}{1+e^{-\eta\Delta t}})\right] n
\end{aligned}
\tag{3.11}
$$

Set $\frac{1}{1+e^{-\eta\Delta t}} = G_2(\Delta t) = G_2$

$$
[I(\theta)]_{\mu,\mu} = \left[\frac{2\eta}{\sigma^2}\left(\frac{1-e^{-\eta\Delta t}}{1+e^{-\eta\Delta t}}\right)\right] n = \frac{2\eta}{\sigma^2}(2G_2-1)n \tag{3.12}
$$

$$
\begin{aligned}
[I(\theta)]_{\sigma,\sigma} =& -\frac{n}{\sigma^2} + \frac{6\eta}{\sigma^4}\sum_{i=1}^{n}\frac{E[B_i^2]}{1-e^{-2\eta d_i}} \\
=& -\frac{n}{\sigma^2} + \frac{6\eta}{\sigma^4}n\frac{(\frac{\sigma^2}{2\eta})(1-e^{-2\eta\Delta t})}{1-e^{-2\eta\Delta t}} \\
=& \frac{2n}{\sigma^2}
\end{aligned}
\tag{3.13}
$$

$$[I(\theta)]_{\mu,\sigma} = 0 \tag{3.14}$$

$$\begin{aligned} I(\theta)_{\mu,\eta} = & -\frac{2\eta}{\sigma^2}\left[\sum_{i=1}^{n}\frac{(d_i(x_0-\mu)e^{-\eta t_i})}{(1+e^{-\eta d_i})}\right] \\ = & -\frac{2\eta}{\sigma^2}\left[\sum_{i=1}^{n}\frac{(\Delta t(x_0-\mu)e^{-\eta t_i})}{(1+e^{-\eta \Delta t})}\right] \\ = & -\frac{2\eta}{\sigma^2}\frac{\Delta t(x_0-\mu)}{(1+e^{-\eta\Delta t})}\sum_{i=1}^{n}e^{-\eta t_i} \end{aligned} \tag{3.15}$$

As mentioned before, $E[x_0 - \mu] = 0$. Therefore,

$$I(\theta)_{\mu,\eta} = 0 \tag{3.16}$$

$$\begin{aligned} [I(\theta)]_{\sigma,\eta} = & -\frac{2}{\sigma^3}\sum_{i=1}^{n}\frac{E[B_i^2]}{1-e^{-2\eta d_i}} - \frac{2\eta}{\sigma^3}\left[\sum_{i=1}^{n}\frac{-2d_ie^{-2\eta d_i}E[B_i^2]}{(1-e^{-2\eta d_i})^2}\right] \\ = & -\frac{2}{\sigma^3}\sum_{i=1}^{n}\frac{\frac{\sigma^2}{2\eta}(1-e^{-2\eta d_i})}{1-e^{-2\eta d_i}} - \frac{2\eta}{\sigma^3}\left[\sum_{i=1}^{n}\frac{-2d_ie^{-2\eta d_i}\frac{\sigma^2}{2\eta}(1-e^{-2\eta d_i})}{(1-e^{-2\eta d_i})^2}\right] \\ = & -\frac{n}{\sigma\eta} + \frac{2n\Delta t}{\sigma}\frac{e^{-2\eta\Delta t}}{1-e^{-2\eta\Delta t}} \\ = & -\frac{n}{\sigma\eta} + \frac{2n\Delta t}{\sigma}(G_1 - 1) \end{aligned} \tag{3.17}$$

So the Fisher information matrix will be

$$I = \begin{bmatrix} [I(\theta)]_{\mu,\mu} & [I(\theta)]_{\mu,\eta} & [I(\theta)]_{\mu,\sigma} \\ [I(\theta)]_{\mu,\eta} & [I(\theta)]_{\eta,\eta} & [I(\theta)]_{\eta,\sigma} \\ [I(\theta)]_{\mu,\sigma} & [I(\theta)]_{\eta,\sigma} & [I(\theta)]_{\sigma,\sigma} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{2\eta}{\sigma^2}(2G_2-1)n & 0 & 0 \\ 0 & \frac{n}{2\eta^2} + n\Delta t^2(G_1-1)(2G_2-1) - \frac{2n}{\eta}\Delta t(G_1-1) & -\frac{n}{\sigma\eta} + \frac{2n\Delta t}{\sigma}(G_1-1) \\ 0 & -\frac{n}{\sigma\eta} + \frac{2n\Delta t}{\sigma}(G_1-1) & \frac{2n}{\sigma^2} \end{bmatrix}$$

### 3.3.2 Numerical Results

From the previous derivation, each element of the simplified Fisher information matrix is a function of time difference $\Delta t$, sample size $n$, mean reversion parameter $\eta$ and variance parameter $\sigma$. This means that the covariance matrix is also a function of these four parameters. Our objective is to estimate the rate of mean reversion, $\eta$ as accurately as possible. For the reason that $\eta$ is the parameter that represents the temporal dynamics and labels microbiome stability. That is we want to minimize $Var(\hat{\eta}) = [I(\theta)]^{-1}_{\eta,\eta}$.

From the former calculation, we know that

$$[I(\theta)]^{-1}_{\eta,\eta} = \frac{2n/\sigma^2}{[I(\theta)]_{\eta,\eta}[I(\theta)]_{\sigma,\sigma} - [I(\theta)]^2_{\eta,\sigma}}$$

By plugging in the results,

$$\begin{aligned} [I(\theta)]^{-1}_{\eta,\eta} &= \frac{2n/\sigma^2}{[\frac{n}{2\eta^2} + n\Delta t^2(G_1-1)(2G_2-1) - \frac{2n}{\eta}\Delta t(G_1-1)]\frac{2n}{\sigma^2} - [-\frac{n}{\sigma\eta} + \frac{2n\Delta t}{\sigma}(G_1-1))^2]} \\ &= \frac{2n/\sigma^2}{\frac{2n^2\Delta t^2}{\sigma^2}(G_1-1)(2G_2-2G_1+1)} \\ &= [n\Delta t^2(G_1-1)(2G_2-2G_1+1)]^{-1} \end{aligned} \tag{3.18}$$

Where $G_1 = \frac{1}{1-e^{-2\eta\Delta t}}$ and $G_2 = \frac{1}{1+e^{-\eta\Delta t}}$

From the above result, we know that $\sigma$ will not influence the value of $Var(\hat{\eta})$. Therefore, we need to maximize $n\Delta t^2(G_1-1)(2G_2-2G_1+1)$

Let $K = e^{-\eta\Delta t}$. $\Delta t$ becomes $\frac{-\log K}{\eta}$. The Equation (3.18) is

$$
\begin{aligned}
&[I(\theta)]^{-1}_{\eta,\eta} \\
=&[n\Delta t^2(G_1-1)(2G_2-2G_1+1)]^{-1} \\
=&\left[n\Delta t^2(\frac{1}{1-e^{-2\eta\Delta t}}-1)(\frac{2}{1+e^{-\eta\Delta t}}-\frac{2}{1-e^{-2\eta\Delta t}}+1)\right]^{-1} \\
=&\left[-n\frac{(\log K)^2}{\eta^2}\frac{K^2(K^2+2K-1)}{(1-K^2)^2}\right]^{-1}
\end{aligned}
\tag{3.19}
$$

Let $f(K) = -n\frac{(\log K)^2}{\eta^2}\frac{K^2(K^2+2K-1)}{(1-K^2)^2}$. We need to calculate the first order derivative to maximize $f(K)$

$$
\begin{aligned}
f'(K) =& -n\left[\frac{(\log K)^2}{\eta^2}\right]'\frac{K^2(K^2+2K-1)}{(1-K^2)^2} - n\frac{(\log K)^2}{\eta^2}\left[\frac{K^2(K^2+2K-1)}{(1-K^2)^2}\right]' \\
=& -\frac{2n}{\eta^2}\log K\left[\frac{K}{(1-K)^2}-\frac{2K}{(1-K^2)^2}\right] \\
&-\frac{n}{\eta^2}(\log K)^2\left[\frac{2}{(1-K)^3}-\frac{2}{(1-K)^2}-\frac{4K}{(1-K^2)^2}-\frac{8K^3}{(1-K^2)^3}\right] \\
=& -\frac{2nK}{\eta^2}\log K[(1-K^2)(K^2+2K-1)+\log K(K^3+K^2+3K-1)] \\
=&0
\end{aligned}
\tag{3.20}
$$

Using Newton's method, we find that when $K = 0.2060614$, $f'(K) \approx 0$. Then, we can find the relation between the optimal time difference $\Delta t$ and $\eta$

$$\Delta t_{Optimal} = \frac{-\log K}{\eta} = \frac{-\log(0.2060614)}{\eta} = \frac{1.579581}{\eta}$$

For fixed sample size $n$, $Var(\hat{\eta})$ is affected by time difference $\Delta t$. A number of these functions are shown in Figure 3.1. We see that for small $\Delta t$, there is not enough time to observe mean reversion so $\hat{\eta}$ is inaccurate, while for large $\Delta t$, consecutive $x_i$'s are almost independent, making the estimation of $\eta$ difficult. Figure 3.1 clearly shows that there is an optimal sampling frequency for estimating $\eta$. We can also see that the optimal time difference $\Delta t$ is the same for all sample sizes.

Figure 3.2 shows the relation between $Var(\hat{\eta})$ and sample size, for various values of $\Delta t$. As excepted, $Var(\hat{\eta})$ is inversely proportional to sample size.

Figure 3.3 shows the optimal time interval as a function of $\eta$.

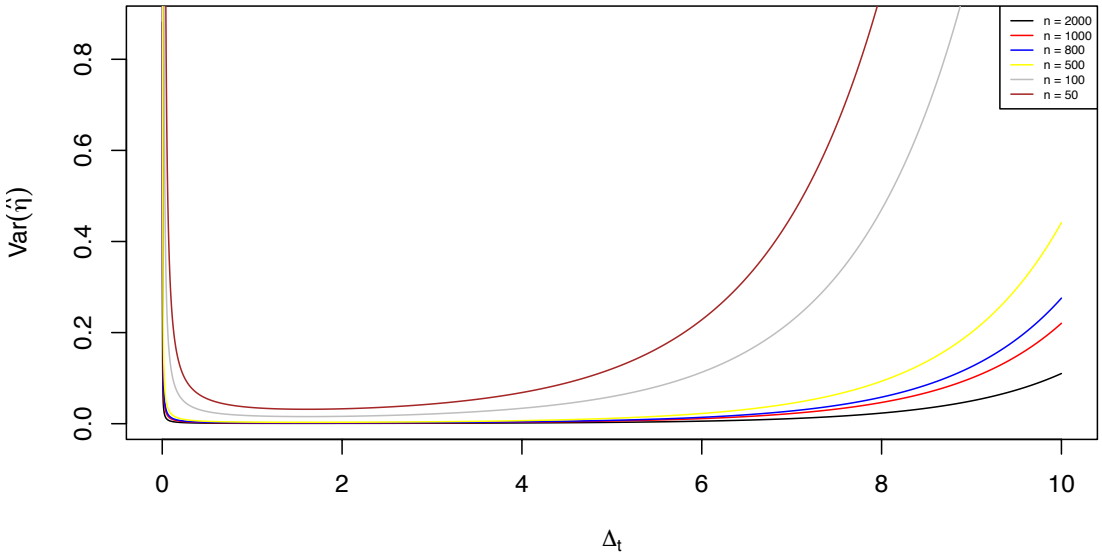


Figure 3.1: Relation between time difference $\Delta t$ and $Var(\hat{\eta})$ for various sample sizes $n$ for true parameter values $\eta = 0.5$, $\sigma = 0.01$, $x_0 = 0$, $\mu = 0$

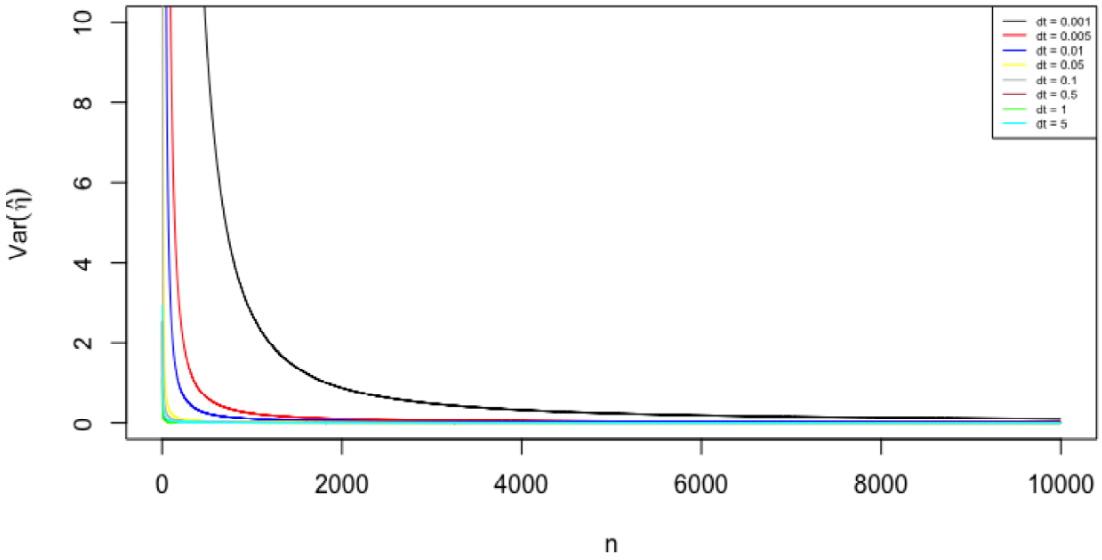


Figure 3.2: Relation between sample sizes $n$ and time difference $\Delta t$ for true parameter values $\eta = 0.5$, $\sigma = 0.01$, $x_0 = 0$, $\mu = 0$

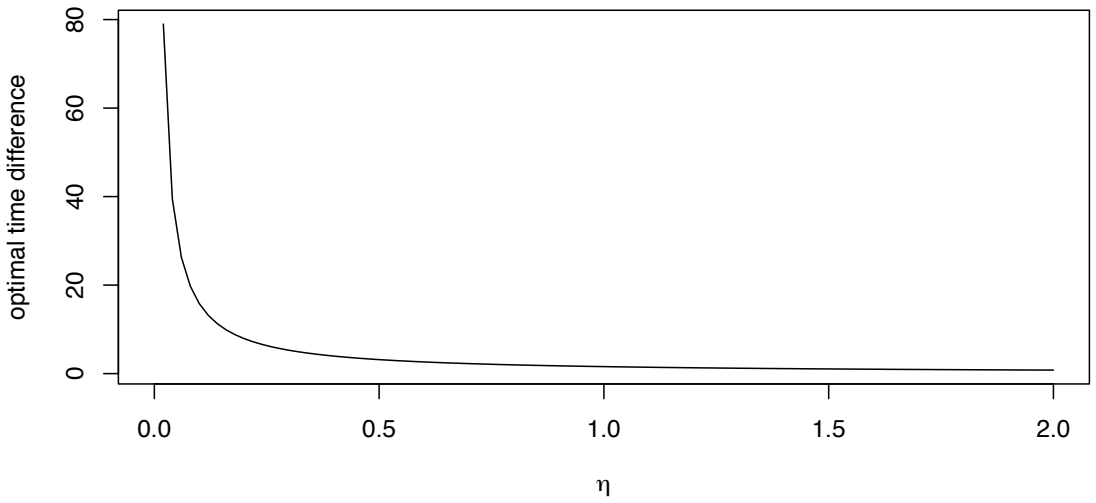


Figure 3.3: Optimal time difference $\Delta t$ as function of $\eta$

### 3.3.3 Application to Moving Picture Data

We apply this method to the moving picture data. Estimated values of $\hat{\eta}$ for all abundant genera at each body site are shown in Figure 3.4 (for Person 1) and Figure 3.5 (for Person 2). We see that for most genera, we estimate $\hat{\eta}$ in the range from 0.4 to 1.5 for Person 1 and 0.5 to 2 for Person 2. Based on our previous method, we can calculate estimated values of $\hat{\mu}$ and $\hat{\sigma}$ at each body site, which are shown in Figure 3.6 - Figure 3.9. We see that for most genera, estimated values of $\hat{\mu}$ is in the range from 0.0070 to 0.0074 for Person 1 and 0.00270 to 0.00300 for Person 2. The range of $\hat{\sigma}$ is from 0.007 to 0.014 for Person 1 and 0.001 to 0.013 for Person 2.

From the previous result, we know that it is $\eta$ that determines the optimal sampling time. For the estimated values of $\eta$, we examine the relation between $\Delta t$ and $Var(\hat{\eta})$. For Person 1, Figure 3.10-3.13 show the variance of $\hat{\eta}$ for different $\Delta t$. In the same body site with different sample sizes, the optimal time difference is similar, but with larger sample sizes, $Var(\hat{\eta})$ becomes smaller. In different body site, the optimal time difference is different as well. We find that enclosed body sites have smaller optimal sampling frequency than the external ones. Internal body sites' estimated *eta* values are relatively smaller than external body sites' estimated *eta* values, so the

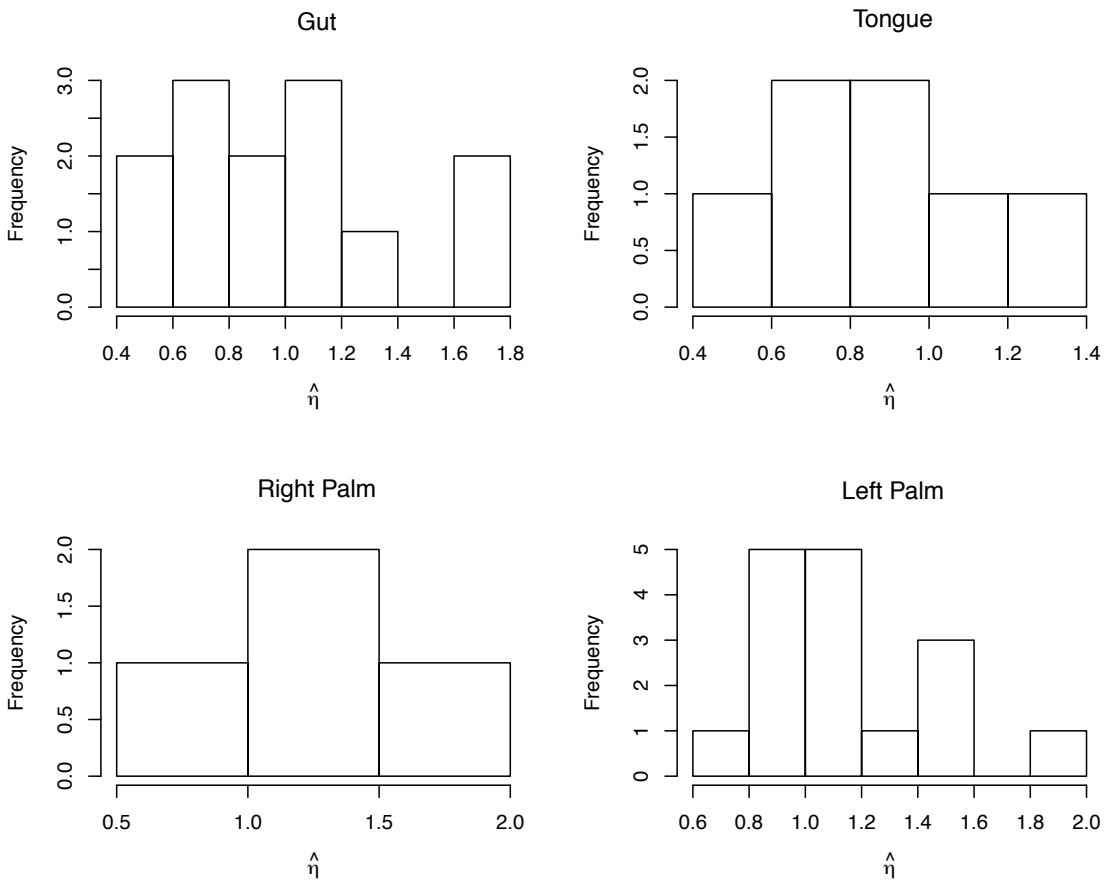


Figure 3.4: Distribution of $\hat{\eta}$ over genera for Person 1

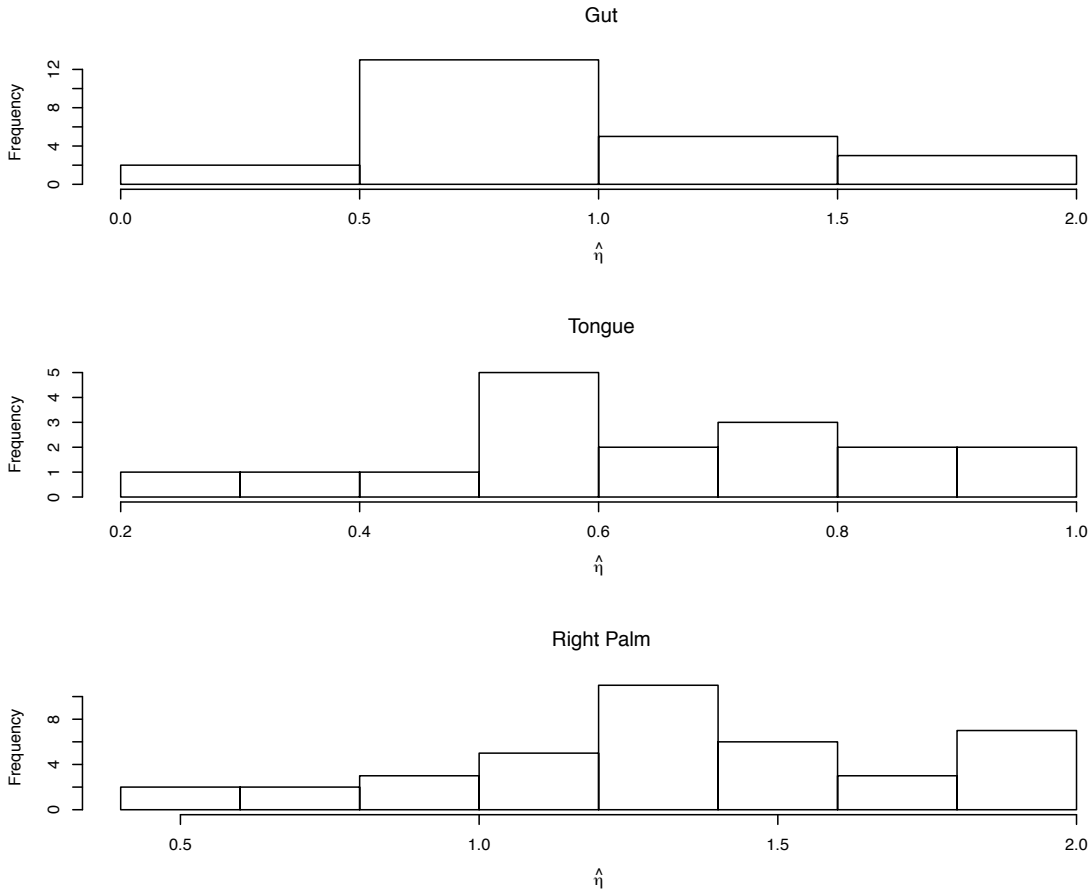


Figure 3.5: Distribution of $\hat{\eta}$ over genera for Person 2

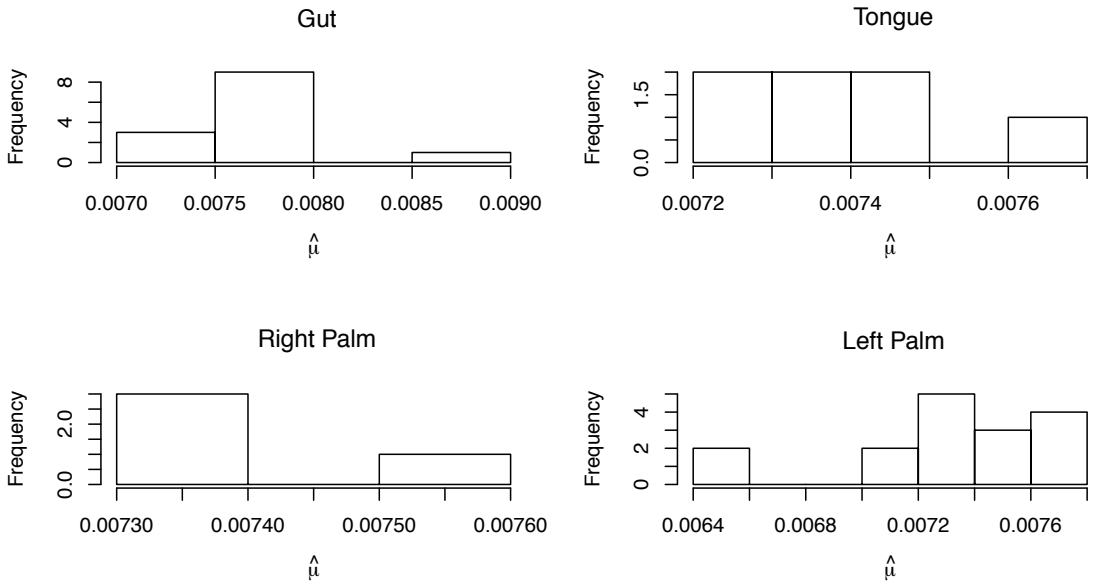


Figure 3.6: Distribution of $\hat{\mu}$ over genera for Person 1

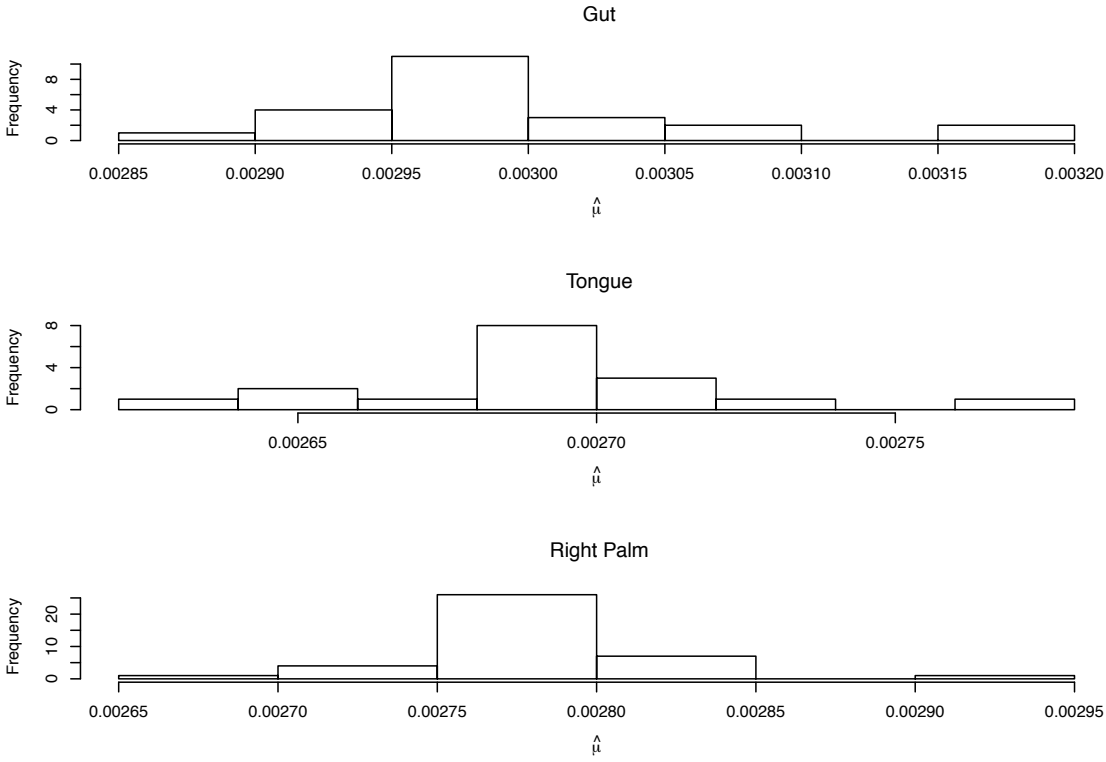


Figure 3.7: Distribution of $\hat{\mu}$ over genera for Person 2

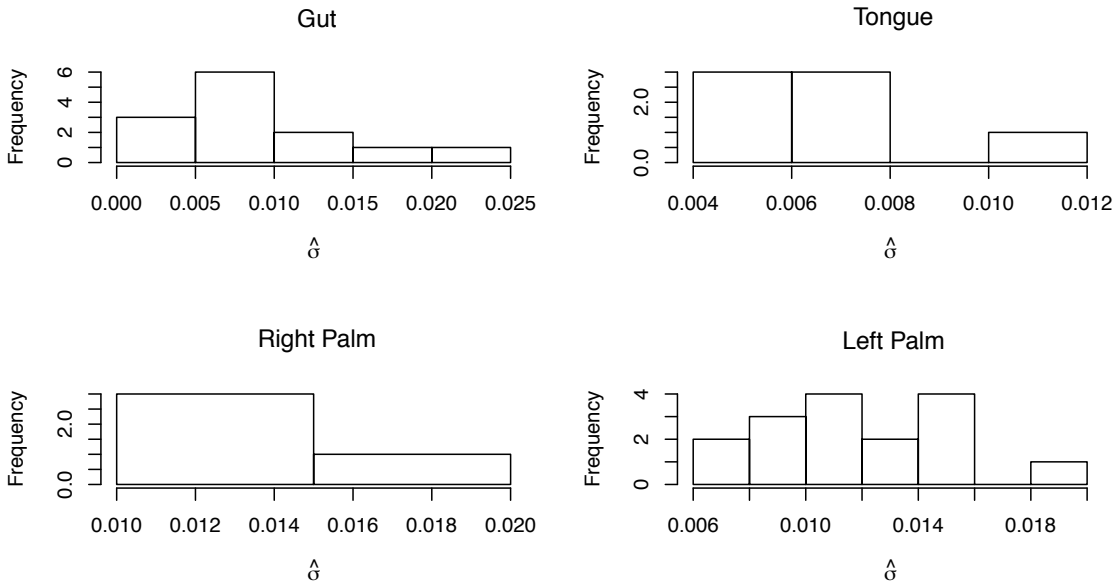


Figure 3.8: Distribution of $\hat{\sigma}$ over genera for Person 1

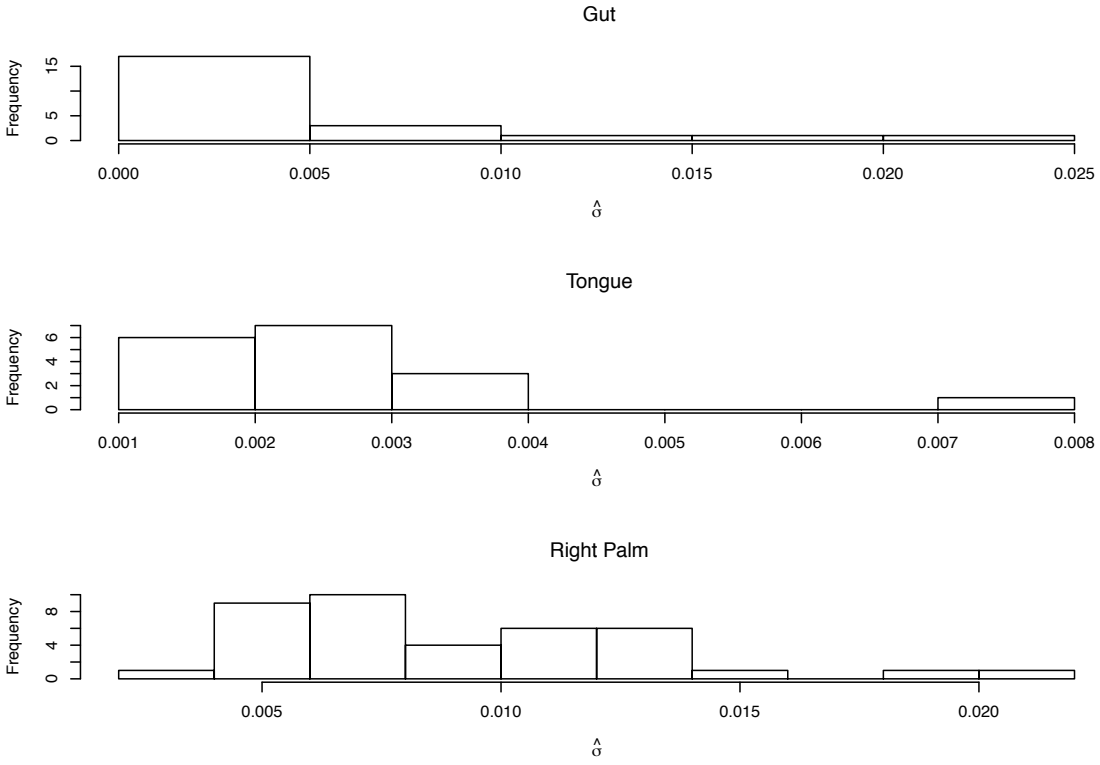


Figure 3.9: Distribution of $\hat{\sigma}$ over genera for Person 2

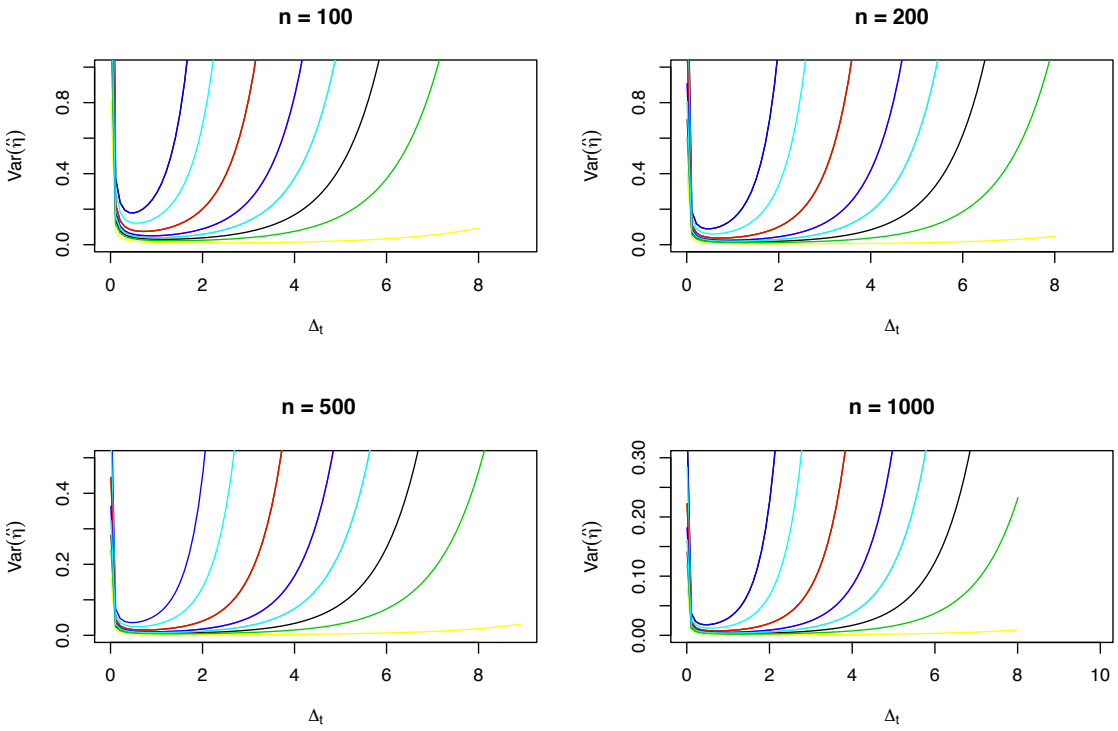


Figure 3.10: Variance of $\hat{\eta}$ as function of $\Delta t$ for different genera (different curves) in gut of Person 1

optimum sampling time intervals for internal body sites can be slightly larger than that of external body sites.

When we estimate the optimal sampling for Person 2, Figures 3.14-3.16 show similar results.

From the results, we can see that the optimal time difference is approximately 1 sample per day.

## 3.4 Simulation

### 3.4.1 Simulation Design

The asymptotic normality of MLE theorem states that for a large enough sample size, the asymptotic behaviour of MLEs can be described by the Fisher information matrix. However, it does not specify what sample size is needed for this asymptotic approximation to be reasonable. We therefore conduct a simulation study to confirm

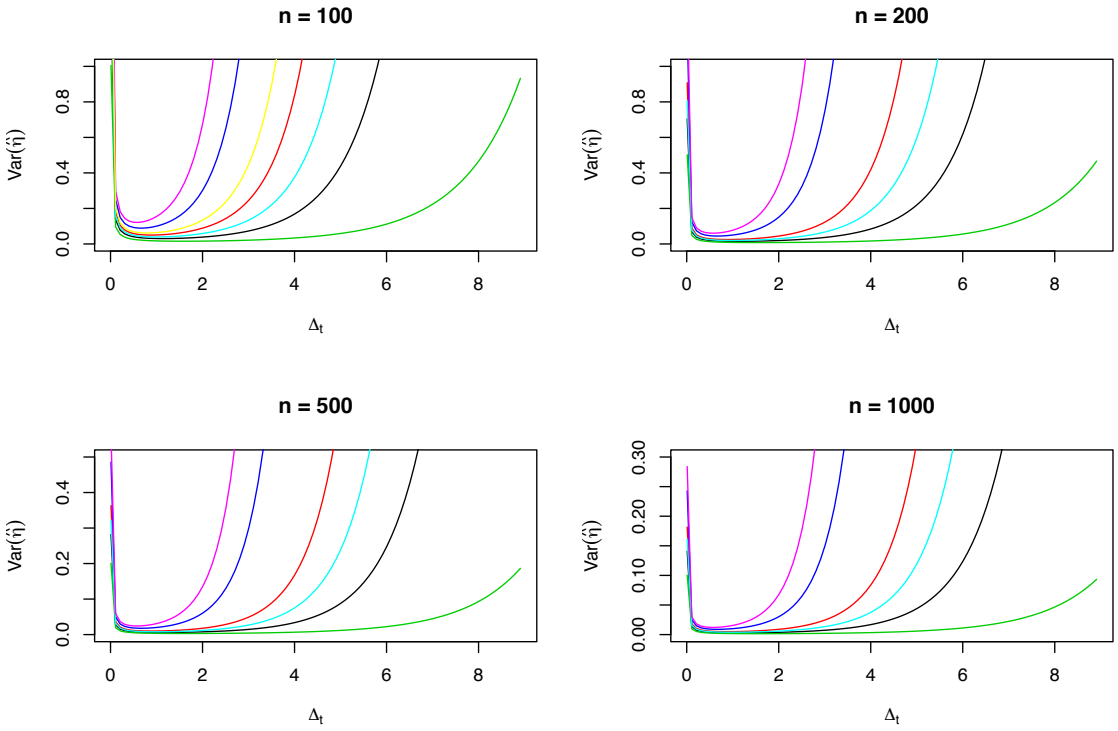


Figure 3.11: Variance of $\hat{\eta}$ as function of $\Delta t$ for different genera (different curves) in tongue of Person 1

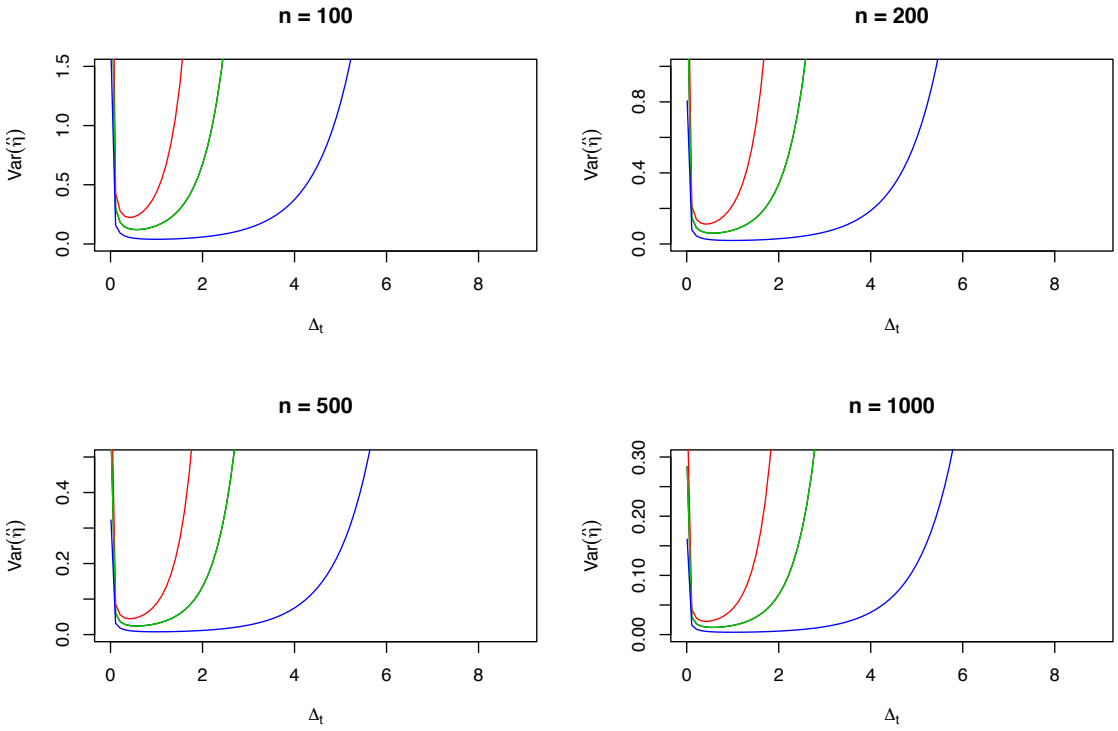


Figure 3.12: Variance of $\hat{\eta}$ as function of $\Delta t$ for different genera (different curves) in right palm of Person 1

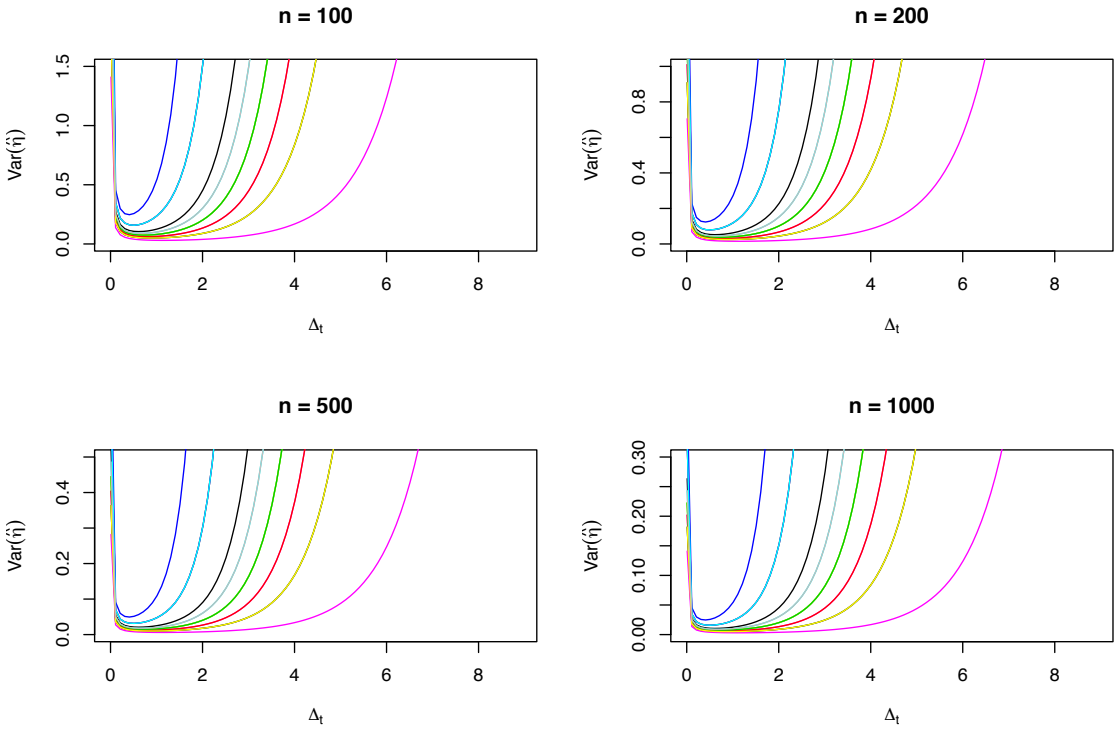


Figure 3.13: Variance of $\hat{\eta}$ as function of $\Delta t$ for different genera (different curves) in left palm of Person 1

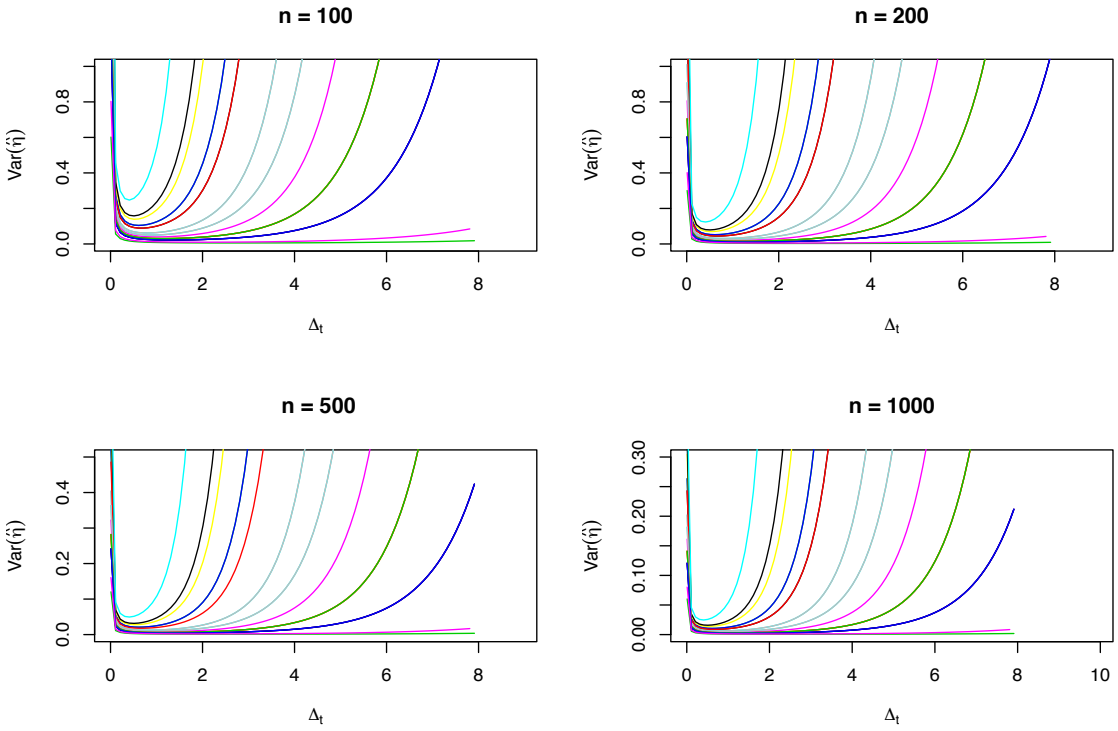


Figure 3.14: Variance of $\hat{\eta}$ as function of $\Delta t$ for different genera (different curves) in gut of Person 2

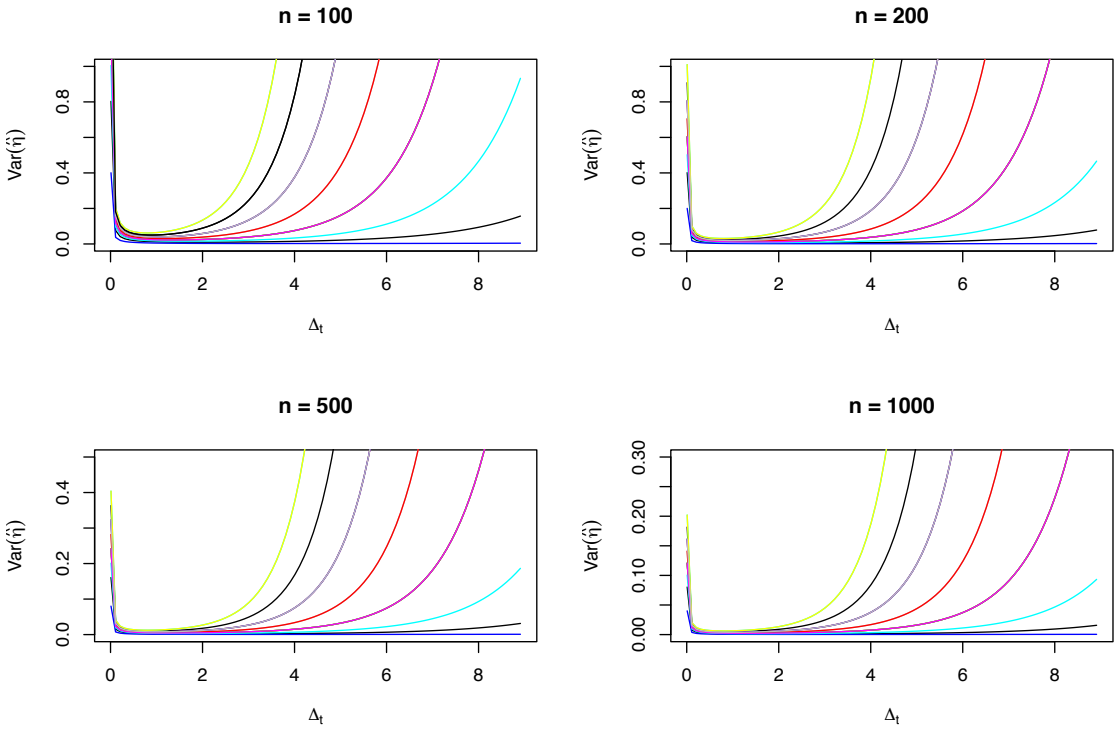


Figure 3.15: Variance of $\hat{\eta}$ as function of $\Delta t$ for different genera (different curves) in tongue of Person 2

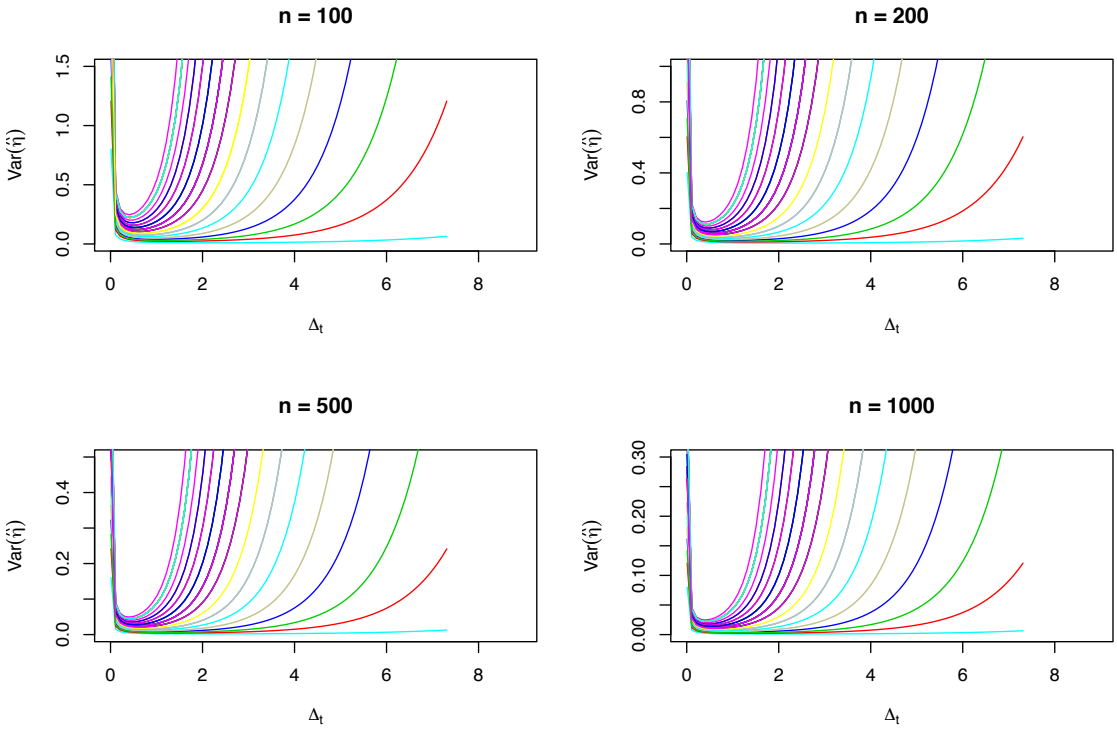


Figure 3.16: Variance of $\hat{\eta}$ as function of $\Delta t$ for different genera (different curves) in right palm of Person 2

Table 3.1: Sample size and time difference for each simulation data set

| Time difference | $\Delta t = 0.002$ | $\Delta t = 0.01$ | $\Delta t = 0.02$ |
|---|---|---|---|
| Sample size | 5000 | 1000 | 500 |
| Time difference | $\Delta t = 0.1$ | $\Delta t = 0.2$ | $\Delta t = 1$ |
| Sample size | 100 | 50 | 10 |

Table 3.2: Distance of Fisher information inverse matrix and covariance

| Time difference | $\Delta t = 0.002$ | $\Delta t = 0.01$ | $\Delta t = 0.02$ |
|---|---|---|---|
| Distance | 0.2871476 | 0.2641229 | 0.2696390 |
| Time difference | $\Delta t = 0.1$ | $\Delta t = 0.2$ | $\Delta t = 1$ |
| Distance | 0.2657307 | 4.9819356 | 6.2466051 |

that the asymptotic approximation can be used for realistic sample sizes.

In order to test our derived Fisher information matrix, we simulate 6 different OU mean reverting microbiome data sets with the same period from 0 to 10 but different sample sizes. The parameters for the simulation are $\eta = 0.8$, $\sigma = 0.01$ and $\mu = 0$. For each sample size, we compute the MLEs for $\hat{\eta}$, $\hat{\sigma}$ and $\hat{\mu}$. We compare the covariance matrices estimated over 5000 simulations, using the following widely-used matrix distance [7].

$$d^2(\boldsymbol{A}, \boldsymbol{B}) = tr(\log(\boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2}))$$

### 3.4.2 Simulation Results

For each simulation, we calculate the MLEs for 5000 simulated datasets. We estimate the covariance matrix for these MLEs, and compare with the inverse of the Fisher information matrix. The result is shown in Table 3.2. We see that Fisher information provides a good approximation when $n \geq 100$. Therefore, we consider that our use of Fisher information is appropriate, and our conclusions about optimal sampling protocols are justified.

# Chapter 4

# Discussion

## Conclusion

In this thesis, we first find evidence of temporal dependence and mean reversion in moving picture data using the likelihood ratio test. We see that for enclosed body sites, more abundant genera show strong evidence of dependence than for external ones. Furthermore, all of the abundant genera show evidence of mean reversion. Then we consider the accuracy of our estimated mean reversion velocity and the most efficient sampling scheme for estimating the parameters, particularly the mean reversion velocity $\eta$. We derive the Fisher information matrix to determine the optimal sampling frequency. For different body sites, the optimal time difference is different. The optimal sampling frequency for enclosed body sites is smaller than for exposed body sites. The results suggest that future studies be most efficient for understanding microbial dynamics if the sampling frequency is approximately 1 sample per day. Moreover, we performed simulations to confirm that the asymptotic theory applies to our finite sample cases.

## Future work

This thesis suggests many promising directions for future work. In this thesis, OU process is used to model the temporal dynamics of the moving picture data. The OU process is the simplest model with a linear velocity parameter mean reversion. We will fit a more elaborate model in the future allowing multiple stable points, such as a nonlinear mean reversion model. Moreover, we can also develop a model to incorporate interactions between OTUs for our future studies because the microbiome is driven by the interaction of many different microbes. We can also incorporate measurement error in sampling to derive more accurate estimates. This thesis currently analyzes proportional genera data sets. Another important direction for future research is to adapt our method to directly deal with count data.

## Bibliography


[1] Fredrik Bäckhed, Hao Ding, Ting Wang, Lora V Hooper, Gou Young Koh, Andras Nagy, Clay F Semenkovich, and Jeffrey I Gordon. The gut microbiota as an environmental factor that regulates fat storage. *Proceedings of the National Academy of Sciences*, 101(44):15718–15723, 2004.

[2] Peter J Bickel and Kjell A Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics, Volumes I-II Package*. Chapman and Hall/CRC, 2015.

[3] J Gregory Caporaso, Christian L Lauber, Elizabeth K Costello, Donna Berg-Lyons, Antonio Gonzalez, Jesse Stombaugh, Dan Knights, Pawel Gajer, Jacques Ravel, Noah Fierer, et al. Moving pictures of the human microbiome. *Genome biology*, 12(5):R50, 2011.

[4] Elizabeth K Costello, Christian L Lauber, Micah Hamady, Noah Fierer, Jeffrey I Gordon, and Rob Knight. Bacterial community variation in human body habitats across space and time. *Science*, 326(5960):1694–1697, 2009.

[5] Morris H DeGroot and Mark J Schervish. *Probability and statistics*. Pearson Education, 2012.

[6] Bradley Efron and David V. Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65(3):457–482, 1978.

[7] Wolfgang Förstner and Boudewijn Moonen. A metric for covariance matrices. In *Geodesy-The Challenge of the 3rd Millennium*, pages 299–309. Springer, 2003.

[8] José Carlos Garcıa Franco. Maximum likelihood estimation of mean reverting processes. *Real Options Practice*, 2003.

[9] B Roy Frieden. *Science from Fisher information: a unification*. Cambridge University Press, 2004.

[10] Ioannis Karatzas and Steven E Shreve. Brownian motion. In *Brownian Motion and Stochastic Calculus*, pages 47–127. Springer, 1998.

[11] Jeremy E Koenig, Aymé Spor, Nicholas Scalfone, Ashwana D Fricker, Jesse Stombaugh, Rob Knight, Largus T Angenent, and Ruth E Ley. Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4578–4585, 2011.

[12] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

[13] Don S Lemons and Paul Langevin. *An introduction to stochastic processes in physics*. JHU Press, 2002.

[14] Thomas T MacDonald and Sven Pettersson. Bacterial regulation of intestinal immune responses. *Inflammatory bowel diseases*, 6(2):116–122, 2000.

[15] Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

[16] Chana Palmer, Elisabeth M Bik, Daniel B DiGiulio, David A Relman, and Patrick O Brown. Development of the human infant intestinal microbiota. *PLoS biology*, 5(7):e177, 2007.

[17] Luigi M Ricciardi and Laura Sacerdote. The ornstein-uhlenbeck process as a model for neuronal activity. *Biological cybernetics*, 35(1):1–9, 1979.

[18] Sonja Rieder. Robust parameter estimation for the ornstein–uhlenbeck process. *Statistical Methods & Applications*, 21(4):411–436, 2012.

[19] Rori V Rohlfs, Patrick Harrigan, and Rasmus Nielsen. Modeling gene expression evolution with an extended ornstein–uhlenbeck process accounting for within-species variation. *Molecular biology and evolution*, 31(1):201–211, 2013.

[20] Sheldon M Ross, John J Kelly, Roger J Sullivan, William James Perry, Donald Mercer, Ruth M Davis, Thomas Dell Washburn, Earl V Sager, Joseph B Boyce, and Vincent L Bristow. *Stochastic processes*, volume 2. Wiley New York, 1996.

[21] Leung Tim Siu-tang and Li Xin. *Optimal mean reversion trading: Mathematical analysis and practical applications*, volume 1. World Scientific, 2015.

[22] G. E. Uhlenbeck and L. S. Ornstein. On the theory of the brownian motion. *Phys. Rev.*, 36:823–841, Sep 1930.