UNIVERSITY OF CALIFORNIA
RIVERSIDE

Parameter Estimation in Differential Equation Based Models

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Zhen Xiao

August 2014

Dissertation Committee:

Dr. Xinping Cui, Chairperson
Dr. Daniel Jeske
Dr. Zhenbiao Yang

The Dissertation of Zhen Xiao is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I am most grateful to my advisor, Dr. Xinping Cui, for her enthusiasm, dedication and encouragement. Without her guidance, this dissertation would not exist.

I would also like to thank Dr. Nicolas Brunel, for his insightful and constructive comments and suggestions on the mathematical work presented in this thesis, and Dr. Zhenbiao Yang, for the biological motivation, experimental data and many helpful suggestions.

To my parents.

ABSTRACT OF THE DISSERTATION

Parameter Estimation in Differential Equation Based Models

by

Zhen Xiao

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, August 2014
Dr. Xinping Cui, Chairperson

There is a long history for differential equations to be utilized to model dynamic processes in many disciplines such as physics, engineering, computer science, Finance, Biology, *etc.*. Most original efforts have been devoted to simulating the dynamic process for given parameters that characterize the differential equations. In recent years, more and more attention has been given by scientists, especially statisticians who considered the problem inversely, that is, using the experimental data to recover the values of parameters that specifically describe the experimental process trajectory.

In this dissertation, we introduced a general Integro-ordinary differential equation that describes a reaction diffusion process for the tip growth of pollen tubes and proposed a constrained nonlinear mixed effects model for the dynamic response. Accordingly, we developed two estimation procedures to estimate this model, Constrained Method of Moments (CMM) and Constrained Restricted Maximum Likelihood (CREML). The advantages and disadvantages of the two procedures were investigated. Simulation studies and a real data analysis demonstrated that both estimation procedures could provide accurate estimates for the parameters.

As an extension, the parameter estimation for Integro-partial differential equation of multiple dimension would also be discussed.

Contents

List of Figures

List of Tables

Part I

First Part

Chapter 1

Introduction

1.1 Tip growth of pollen tube, a motivating example

Tip growth is an extreme form of polarized growth of living cells that results in an elongated cylindrical cell morphology with a rounded tip where the growth activity takes place. Tip growth occurs in many organisms such as fungi (hyphae) [32], neuron axon (animal) [53] and plants (e.g. root hairs and pollen tube) [3, 85]. One common theme of tip growth is the formation of the polar cap on plasma membrane via polarized Cdc42/Rho-family GTPase localization, determining the growing site and coordinating multiple signaling events including cytoskeleton organization, vesicle trafficking (including exocytosis) and ion dynamics [32, 53, 85]. Another related theme is the localized cell surface extension via cell wall material renewal [17, 25]. To understand the tip growth mechanism, it's crucial to study the coupling mechanism that links the two common themes.

Pollen tube, which demonstrates the importance of tip growth for the plant fertilization, provides us an ideal model to study the tip growth mechanism for the formation of the polar cap and how the dynamics of the polar cap modulates tip growth.

Specifically, when a pollen tube receives a certain internal or external stimulus, it will enter the first stage of ROP1 Rho GTPase polarity establishment, where through exocytosis mediated positive feedback loop, exocytosis mediated negative feedback loop and lateral diffusion of ROP1 GTPase on plasma membrane, active ROP1 GTPases polarity can be established and maintained on the plasma membrane. Once maintained, the active ROP1 triggers and regulates the exocytosis of a cell surface material called Pectin, via which controls the cell wall mechanics and eventually determines the shape of pollen tubes. Many previous theoretical biomechanical modeling attempts had been made to simulate the tip growth of pollen tube [18, 25, 47]. These models focused more on the cell wall mechanics and shape formation of pollen tubes. However, they found out at the end of the first stage, the activated ROP1 GTPases on the membrane are not uniformly distributed but rather appear in clusters, and these asymmetrically distributed ROP1 GTPases are the one that leads to a polarized growth of the cell, which results in an elongated cylindrical cell morphology. Therefore, modeling the distribution of ROP1s on the membrane is the key to understand the tip growth of pollen tube.

To understand the ROP1 GTPase polarity establishment, Yang *et.al*. proposed a model based on two versatile functions of exocytosis. First, they hypothesized that exocytosis participates positive feedback regulations of polarized ROP1 GTPase at tip growing domain. The following evidence supports that exocytosis mediated postive feedback of ROP1 GTPase

1. it has been observed that exocytosis is essential for pollen tube growth and renewal of plasma membrane proteins [28,48]. Given that ROP1 and its activator complex GEF/PRK2 is PM associated [1,7,27,39], it is believed that exocytosis is important for activation of ROP1 at apex of pollen tubes.

2. it has been shown recently that active ROP1 locally direct exocytosis at pollen tube tip region via control of F-actin dynamics [ref], thus forming a positive feedback loop.

Second, they hypothesized that exocytosis participates negative feedback regulation of polarized ROP1 GTPase cap. REN1 RopGAP is an important negative regulator of ROP1 activity and the phenotype of RopGAP weak knock down mutant is greatly enhanced by further inhibition of F-actin dependent exocytosis [46], supporting the existence of an exocytosis mediated negative regulation.

Based on the two key functions of exocytosis, Yang *et.al.* proposed a Intrego-Partial Differential Equation (IPDE) model (1.1), in which the redistribution of ROP1 GTPase is determined by the rates of four fundamental transport mechanisms including (1) positive feedback regulation, *i.e.*, recruitment of cytoplasmic molecules to the location of the signaling molecules on the membrane with rate k_{pf} ; (2) the negative feedback regulation, *i.e.*, random disassociation of signaling molecules from the membrane with rate k_{nf}; (3) lateral diffusion of molecules along the membrane with rate D; (4) spontaneous association of cytoplasmic molecules to random locations on the membrane with rate k_{on}. The spontaneous association is too weak to be considered, therefore they assume that the first three processes together lead to ROP1 polarity formation.

$$\begin{cases} \frac{\partial R(x,t)}{\partial t} = k_{pf} R(x,t)^{\alpha} (1 - \frac{\int_x R(x,t)dx}{R_{tot}}) - k_{nf} R(x,t) + D \frac{d^2 R(x,t)}{dx^2} \\ \text{where} \qquad \{x,t\} \in [-L_0, L_0] \times [0, \infty] \\ R(-L_0, t) = R(L_0, t) = 0 \end{cases} \tag{1.1}$$

In model (1.1), $R(x,t)$ denotes the ROP1 intensity in position x on the membrane at time t, and R_{tot} denotes the total free ROP1. Throughout this paper, R_{tot}, D and $\alpha > 1$ are assumed to be known constants. This model is similar as the PDE model described in [4] except in their model spontaneous association is included and α is assumed to be 1,

and the integral of $R(x,t)$ is considered as a unknown function of t that doesn't involve $R(x,t)$.

At static time t_0, the ROP1 density won't change with respect to time. That is, $\frac{\partial R(x,t)}{\partial t}|_{t=t_0} = 0$. Let's denote $R(x,t_0)$ as $R(x)$, the IPDE model (1.1) then degenerates to the following IDE model

$$\begin{cases} D\frac{d^2R}{dx^2} = k_{nf}R - k_{pf}R^{\alpha}(1 - \frac{\int_x Rdx}{R_{tot}}) \quad x \in [-L_0, L_0] \\ R(-L_0) = R(L_0) = 0 \end{cases} \tag{1.2}$$

Therefore, in order to recover the OP1 Rho GTPase polarity establishment and further the tip growth, we have to establish the related statistical methodology based on the IDE model (1.2) and estimate the parameter in the IDE model (1.2). This is the ultimate goal of this dissertation work. In the next section, several types of DEs will be reviewed.

1.2 A review of Differential Equations

A differential equation is a mathematical equation relating some function of one or more variables with its derivatives. It arises when a deterministic relation of some continuously varying variables (modeled by functions) and their rates of change in space and/or time (modeled by derivatives) is known or postulated. Such relations are so general and commonly observed that differential equations play an important role in disciplines including physics and engineering. Moreover, in recent years we have seen a dramatic increase in using differential equations in disciplines such as biology, chemistry, economics and computer sciences particularly that is related to image processing. Throughout this section, we are going to introduce several types of differential equations that are widely used in practice.

1.2.1 Ordinary Differential Equation

An ordinary differential equation (ODE) is a differential equation in which the dependent variable is a function of a single independent variable. Let u be a dependent variable and x be an independent variable, so that $u = u(x)$ is a function of x. From now on, we are going to use two distinct notations for differentiation. One is the Leibniz's notation $(\frac{du}{dx}, \frac{d^2u}{dx^2}, \cdots, \frac{d^nu}{dx^n})$ for one dimension case and $(\frac{\partial u}{\partial x}, \frac{\partial^2 u}{\partial x^2}, \cdots)$ for high dimension case. The other is Newton's and Lagrange's notation $(u', u'', \cdots, u^{(n)})$ for one dimension case and $(u_x, u_{xx}, u_{xy}, \cdots)$ or $(\nabla x, \triangle x, \cdots)$ for high dimension case. The general definition of an ODE can be as follows.

Definition 1 *Let F be a given function of x, u and derivatives of u. Then an equation of the form $u^{(n)} = F(x, u, u', \cdots, u^{(n-1)})$ is called an explicit ODE. On the other hand, an equation of the form $F(x, u, u', \cdots, u^{(n-1)}, u^{(n)}) = 0$ is called an implicit ODE.*

ODE can be further classified according to the order of an ODE, which is defined to be the order of the highest derivative of the dependent variable with respect to the independent variable appearing in the equation. For example, the equation $\frac{du}{dx} = u^2 + x$ is a first-order ODE, and $\frac{d^2u}{dx^2} = \frac{du}{dx} + u^2 + x$ is a second-order ODE. An ODE is said to be linear if F can be written as a linear combination of the derivatives of u as follows,

$$u^{(n)} = \sum_{i=0}^{n-1} a_i(x)u^{(i)}(x) + r(x)$$

where $a_i(x)$ and $r(x)$ are continuous functions in x. Otherwise, the ODE is nonlinear. The function $r(x)$ that is called the source term leads to a further classification. When $r(x) = 0$, the ODE is homogeneous. Consequently it has a zero solution $u(x) = 0$, and its nonzero solution is called a complementary function, denoted by $u_c(x)$. When $r(x) \neq 0$, the ODE is non homogeneous. A particular solution to the non homogeneous

ODE is denoted by $u_p(x)$. The general solution to the non homogeneous ODE is $u(x) = u_c(x) + u_p(x)$.

The n^{th}-order ODE $u^{(n)} = F(x, u, u', ..., u^{(n-1)})$ can be expressed as a system of n 1^{st}-order ODEs. Defining a set of dependent variables $u_i(x) = u^{(i-1)}(x)$ for $i = 1, 2, \cdots, n$, the n^{th}-order ODE can be transformed into the following

$$\begin{cases} u_1'(x) = u_2(x) \\ u_2'(x) = u_3(x) \\ \cdots \\ u_{n-1}'(x) = u_n(x) \\ u_n'(x) = F(x, u_1(x), u_2(x), \cdots, u_n(x)) \end{cases}$$

or in a vector form as follows

$$\boldsymbol{u}(x) = \boldsymbol{F}(x, \boldsymbol{u})$$

where $\boldsymbol{u}(x) = (u_1(x), \cdots, u_n(x))^T$ and $\boldsymbol{F}(x, u_1, \cdots, u_n) = (u_2, u_3, ..., F(x, u_1, ..., u_{n-1}))$

1.2.2 Partial Differential Equation

A partial differential equation (PDE) is a differential equation in which the dependent variable is a function of multiple independent variables. The general definition of a PDE can be as follows.

Definition 2 *A PDE for the function $u(x_1, x_2, ..., x_n)$ is an equation of the form*

$$F(x_1, \cdots, x_n, \frac{\partial u}{\partial x_1}, \cdots, \frac{\partial u}{\partial x_n}, \frac{\partial^2 u}{\partial x_1 \partial x_1}, \cdots, \frac{\partial^2 u}{\partial x_1 \partial x_n}, \cdots) = 0$$

The order is defined similarly to the case of ODE. If F is a linear function of u and it's derivatives, then the PDE is linear. Otherwise, the PDE is nonlinear. Common examples of linear PDEs include the heat equation

$$\frac{\partial u}{\partial t} - \alpha \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) = 0$$

the wave equation,

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0$$

Laplace's equation,

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0$$

Poisson's equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = f(x, y, z)$$

Euler-Tricomi equation, *etc.*.

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = 0$$

The second-order linear PDE in two independent variables that is of utmost importance in practice has the following form

$$L[u] = Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = G \tag{1.3}$$

where the coefficients A, B, C, D, E, F and G are given functions that may depend on x and y, and $A^2 + B^2 + C^2 > 0$ over a region of the $x - y$ plane. The operator $L_0[u] = Au_{xx} + 2Bu_{xy} + Cu_{yy}$ is called the principal part of L. It turns out that many fundamental properties of the solution of (1.3) are determined by its principal part, and more precisely, by the sign of the discriminant δ_L of the PDE given by $B^2 - AC$. We classify the equation according to the sign of δ_L.

Definition 3 *Equation (1.3) is said to be hyperbolic at a point (x, y) if $\delta_L(x, y) = B(x, y)^2 - A(x, y)C(x, y) > 0$. It is said to be parabolic at (x, y) if $\delta_L(x, y) = 0$, and it is said to be elliptic at (x, y) if $\delta_L(x, y) < 0$. Let Ω be a domain in $\Re^2$, the equation is hyperbolic (parabolic, elliptic) in Ω, if it is hyperbolic (parabolic, elliptic) at all points $(x, y) \in \Omega$.*

It deserves to emphasis that when a transformation of coordinates $(\xi, \eta) = (\xi(x, y), \eta(x, y))$ is such that the Jacobian $J = \xi_x \eta_y - \xi_y \eta_x$ of the transformation does not vanish at any point (x, y) in Ω, the type of a second-order linear PDE won't change. In the previous examples, it is easy to tell that the wave equation is hyperbolic, the heat equation is parabolic and the Laplace equation is elliptic. When A, B, C are not constant, it is possible that the PDE will not belong to any of the these categories but rather be of mixed type. The simple example is the Euler-Tricomi equtaion, which is elliptic in the region of $x < 0$, hyperbolic in the region of $x > 0$, and parabolic on the line of $x = 0$. If there are n independent variables $x_1, x_2, \cdots, x_n$, a general linear PDE of second order has the form

$$L[u] = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \text{lower order term} = 0$$

The classification depends on the signature of the eigenvalues of the coefficient matrix of A_{ij}. When the eigenvalues are all positive or all negative, the PDE is elliptic. When the eigenvalues are all positive or all negative except that one is zero, the PDE is parabolic. When the eigenvalues are all postive except one is negative, or the eigenvalues are all negative except one is positive, the PDE is hyperbolic. Otherwise, the PDE is ultrahyperbolic. A ultrahyperbolic example is

$$\frac{\partial^2 u}{\partial x_1^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2} - \frac{\partial^2 u}{\partial y_1^2} - \cdots - \frac{\partial^2 u}{\partial y_n^2} = 0$$

1.2.3 Other types of Differential Equation

In practice, there are many other types of differential equations that are widely used, including Stochastic Differential Equation (SDE), Delayed Differential Equation (DDE), Integro-Differential Equation (IDE).

A SDE is a differential equation in which at least one term is a stochastic process so that its solution is also a stochastic process. The general form of a linear SDE is

$$\frac{du(t)}{dt} = A(t, u(t)) + B(t, u(t))\frac{W(t)}{dt}$$

where $W(t)$ is a known stochastic process such as Brownian motion, Wiener process, jump process *etc.*. SDEs were utilized to model diverse system such as fluctuating stock price, physical system subject to thermal fluctuation and so on.

A DDE is a differential equation for a function of a single variable usually called time, in which the derivative of the known function at a certain time is given in terms of the values of the function at earlier times. A general form of a DDE is

$$\frac{du(t)}{dt} = f(t, u(t), u(t-\tau))$$

Examples of using DDE are the time related dynamic processes with delayed effects.

An IDE is a differential equation that involves not only the derivatives of the dependent variable, but also the integral of the dependent variable. The equations (1.1) and (1.2) in section 1.1 are all examples of IDE.

1.3 Application of Differential Equation

Differential equations are widely used to describe the dynamic processes in a broad scientific fields such as physics, biology, chemistry, engineering, economics and computer sciences. In this section, we are going to review some typical examples.

1.3.1 Differential Equation in Physics

A very simple but famous example of differential equation in physics is the Newton's second law of motion which describes the relationship between the position u

and the time t of the object of constant mass m under the force F for the motion of a particle of constant mass m. In general, F depends on the position $u(t)$ of the particle at time t, and so the unknown function $x(t)$ appears on both sides of the differential equation, as is indicated in the notation $F(u(t))$.

$$m\frac{d^2u(t)}{dt^2} = F(u(t))$$

Another famous example is the heat equation which is used to model the dynamic change in the temperature function $u(x, y, z)$ over time t. Where $u(x, y, z)$ is a function of location (x, y, z) that change over time as heat spreads throughout space. The equation is

$$\frac{\partial u}{\partial t} - \alpha(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}) = 0$$

where α is a positive constant representing the rate of heat diffusion. Such equation suggests that the maximum value of $u(x, y, z)$ is either earlier in time in the region of concern or on the edge of the region of concern.

The heat equation is of utmost importance in diverse areas. Firstly, it is a typical parabolic partial differential equation in mathematics. Secondly, it is connected to the Brownian motion by the Kolmogorov forward equation in probability theory. Thirdly, it can be used to solve the Black-Scholes partial differential equation in financial mathematics. Last, it can be generalized to the diffusion equation, which can be used to study the chemical diffusion process in chemistry.

1.3.2 Differential Equation in Neurology

FitzHugh [33] and Nagumo *et.al.* [57] proposed a ODE based model to capture the behavior of spike potentials in the giant axon of squid neurons

$$\frac{dV(t)}{dt} = \alpha(V - \frac{V^3}{3} + R) + u(t)$$
$$\frac{dR(t)}{dt} = -\frac{1}{\alpha}(V - \beta + \gamma R)$$

The above ODE describe the reciprocal dependences between the voltage V on an exon membrane and an outward current R as time goes by. An extra variable $u(t)$ is included to represent some other external effects.

1.3.3 Differential Equation in Ecology

Differential equation is widely used to model the population growth process of many organisms in population ecology. Based on the assumption that organisms have Brownian random motion, the rate of which is invariant in time and space, Okubo [60] proposed a diffusion model

$$\frac{\partial u(x,y,t)}{\partial t} = D\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)$$

where, $u(x,y,z)$ is the population density of organisms at spatial coordinates x, y at time t, and D is the diffusion coefficient that measures the dispersal rate when animals disperse at any two-dimensional coordinates. Furthermore, incorporating the fact that the external stimuli like the wind, sunshine, water *etc.* will guide the animal's travelling orientation, Helland *et.al.* [41] proposed a model by adding some terms of outside stimuli effects into the diffusion model

$$\frac{\partial u(x,y,t)}{\partial t} = D\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) - w_x\frac{\partial u}{\partial x} - w_y\frac{\partial u}{\partial y}$$

where w_x and w_y are the effect of the outside stimuli. Similarly, Skellam [71] proposed a reaction diffusion model to study the expansion of muskrat in Europe

$$\frac{\partial u(x,y,t)}{\partial t} = D\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + ru\left(1 - \frac{u}{k}\right)$$

Where r is the intrinsic rate of growth of muskrat and k is the carrying capacity. More complicated ODE models have been built in order to capture certain properties like spatial correlation, multiple species competition *etc.*, see [43] for more details.

1.3.4 Differential Equation in Finance

Bernt Oksendal [59] introduced an optimal portfolio problem. Suppose that one has two possible investments

1. A risky investment such as a stock investment, where the price function $u_1(t)$ follows a SDE

$$\frac{du_1(t)}{dt} = (\alpha + \sigma\frac{dW(t)}{dt})u_1(t) \tag{1.4}$$

   where α and σ are positive constants. $\frac{dW(t)}{dt}$ denotes a noise stochastic process.

2. A safe investment such as a bond investment, where the price function $u_2(t)$ follows a ODE

$$\frac{du_2(t)}{dt} = \beta u_2(t) \tag{1.5}$$

   where β is a positive constant with $0 < \beta < \alpha$.

At time instant t, the person can place a portion $p(t)$ of his total fortune $X(t)$ into the risky investment, and therefore place the other portion $(1 - p(t))X(t)$ into the safe investment. Given an utility function $U(t)$ and a terminal time T, the problem is to find the optimal portfolio $p(t)$ that maximizes the expected utility of the corresponding terminal fortune $X(T)$, i.e., $max_{p(t)\in[0,1]}E[(X(T)]$.

Chapter 2

Literature Review of Paramter Estimation in ODE models

As discussed in Chapter 1, differential equations have been widely used to model physical system, engineering phenomenon, economic dynamics, and biological processes. Great effort has been made to simulating the results of output or state or dependent variables for a given differential equation with given parameters, which is known as a forward problem. However, little effort has been made to estimating the parameters in a differential equation by using the measurements of the output or state or dependent variables, which is known as an inverse problem. Especially for nonlinear ODE models without closed form solutions, the parameter estimation problem is a virgin. Since 2008, however, this parameter estimation problem began to receive attention from statistical perspective and several statistical estimation procedures were proposed to solve it. In this chapter, we are going to review those procedures and discuss their advantages and disadvantages. Since a higher order ODE can be degenerated to a system of first order

ODEs, it does not lose generality to focus on the first order ODE as follows

$$\begin{cases} \boldsymbol{x}'(t) = \boldsymbol{F}(t, \boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{\theta}) \\ \boldsymbol{x}(0) = \boldsymbol{x}_0 \\ \boldsymbol{z}(t) = \boldsymbol{h}(t, \boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{\theta}) \end{cases} \tag{2.1}$$

where, $\boldsymbol{x}(t) = (x_1(t), \cdots, x_p(t))^T$ is a vector of p state variables (or dependent variables) with initial value $\boldsymbol{x}_0$. $\boldsymbol{z}(t)$ is the measurement or output vector, $\boldsymbol{u}(t) \in \Re^p$ is a known input vector, and $\boldsymbol{\theta} \in \Theta \subset R^q$ is the parameter vector. $\boldsymbol{F}$ is a vector field from $\Re^p$ to $\Re^p$, Θ is a subset of the q dimensional Euclidean space. In most cases in practice where $\boldsymbol{x}(t)$ can be observed directly, $\boldsymbol{h}(t, \boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{\theta}) = \boldsymbol{x}(t)$ and (2.1) can be simplified to be

$$\begin{cases} \boldsymbol{x}'(t) = \boldsymbol{F}(t, \boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{\theta}) \\ \boldsymbol{x}(0) = \boldsymbol{x}_0 \end{cases} \tag{2.2}$$

The inverse problem is to estimate $\boldsymbol{\theta}$ based on the observed data $\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n$ collected at n distinct time instants $t_1, t_2, \cdots, t_n$, where $\boldsymbol{y}_i = \boldsymbol{x}(t_i) + \boldsymbol{\epsilon}(t_i)$ and $\boldsymbol{x}(t)$ satisfies (2.2), and $\{\boldsymbol{\epsilon}(t_i)\}_{i=1}^n$ follow a certain distribution.

In the following sections, we are going to review all the estimation approaches in details. Those estimation approaches can be roughly divided into two groups, parametric approaches including numerical nonlinear least square approach and numerical maximum likelihood approach, and nonparametric approaches including smoothing approach, two step approach and orthogonal condition approach.

2.1 Identifiability of ODE models

Before tackling the parameter estimation problem, we have to make sure the ODE in model (2.1) or (2.2) has at least one solution, ideally an unique solution for given parameters. This is known as an identifiability problem. Normally, solution to

the identifiability problem will provide answers to the following questions. (1), Are the parameters in the ODE models identifiable based on the experimental measurements of dependent variables or their functions? (2), If not all parameters are identifiable, are some of them identifiable? (3), How many measurements in total are necessary in order to identify the parameters? In general, the concept of identifiability is defined as follows.

Definition 4 *The dynamic system modeled by ODEs (2.1) and (2.2) is identifiable if* $\boldsymbol{\theta}$ *in the ODEs can be uniquely determined by the measurable system output* $\boldsymbol{y}(t)$ *for any admissible system input* $\boldsymbol{u}(t)$*. Otherwise, it is said to be unidentifiable.*

Many other relevant concepts of identifiability were also proposed such as global identifiability and local identifiability by Ljung and Glad [52], locally strongly identifiability (or $\boldsymbol{x}_0$-identifiable) by Tunali and Tarn [74], and structurally identifiability by Xia and Moog [84].

Definition 5 *Globally identifiable: A system structure is said to be globally identifiable if for any admissible input* $\boldsymbol{u}(t)$ *and any two parameter vectors* $\boldsymbol{\theta}_1$ *and* $\boldsymbol{\theta}_2$ *in the parameter space* Θ*,* $\boldsymbol{y}(u, \boldsymbol{\theta}_1) = \boldsymbol{y}(\boldsymbol{u}, \boldsymbol{\theta}_2)$ *holds if and only if* $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$*.*

Definition 6 *Locally identifiable: A system structure is said to be locally identifiable if for any* $\boldsymbol{\theta}$ *within an open neighborhood of some point* $\boldsymbol{\theta}^*$ *in the parameter space,* $\boldsymbol{y}(\boldsymbol{u}, \boldsymbol{\theta}_1) = \boldsymbol{y}(\boldsymbol{u}, \boldsymbol{\theta}_2)$ *holds if and only if* $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$*.*

Definition 7 *Locally strongly identifiable (or* $\boldsymbol{x}_0$*-identifiable): For an admissible input function* $\boldsymbol{u}(t)$ *in the studied time interval* $[t_0, T]$ *and a given initial state* $\boldsymbol{x}_0 = \boldsymbol{x}(t_0)$ *which is independent of* $\boldsymbol{\theta}$ *and not a static point, if there exists an open set* $\Theta_0 \subset \Theta$ *such that for any* $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \in \Theta_0$*, the solutions* $\boldsymbol{x}(\boldsymbol{u}, \boldsymbol{\theta})$ *exist on* $[t_0, t_0 + \epsilon]$ $(0 < \epsilon < T - t_0)$ *for both* $\boldsymbol{\theta}_1$ *and* $\boldsymbol{\theta}_2$*, and* $\boldsymbol{y}(\boldsymbol{u}, \boldsymbol{\theta}_1) \neq \boldsymbol{y}(\boldsymbol{u}, \boldsymbol{\theta}_2)$ *on* $[t_0, t_0 + \epsilon]$*, the system structure is said to be locally strongly identifiable.*

Definition 8 *Structurally identifiable: denote* $C_{\boldsymbol{u}}^{N[t_0,T]}$ *to be the function space expanded by all input functions on* $[t_0, T]$ *which are differentiable up to the order* N*, and let* M *denote an open set of initial values of the state vector. The system structure is said to be structurally identifiable if there exist open and dense subsets* $M^0 \subset M$*,* $\Theta_0 \subset \Theta$*, and* $U_0 \subset C_{\boldsymbol{u}}^{N[t_0,T]}$ *such that the system is locally strongly identifiable at* $\boldsymbol{\theta}$ *given* $\boldsymbol{u}$ *for any* $\boldsymbol{x}_0 \in M^0$*,* $\boldsymbol{\theta} \in \Theta_0$*, and* $\boldsymbol{u} \in U_0$*.*

The concept of structural identifiability was initially introduced by Bellman and Astrom [10]. The structural identifiability problem for linear ODE has been well solved by many techniques such as Laplace transformation method of Ballman and Astrom [10], power series expansion method of Pohjanpalo [65], and similarity transformation method of Walter and Lecourtier [81]. Some of these approaches have been extended by Vajda [75], Chappel and Godfrey [21] to nonlinear ODE models. However, the extension only works for a few simple nonlinear ODE models. For general nonlinear ODE models, the identifiability problem can be solved by direct test method by Denis and Blanchard [24] and Walter [80], differential algebra method by Ljung and Glad [52], and implicit function theorem method of Xia and Moog [84]. A good review of these methods can be found in Wu [56]. It should be pointed out that the identifiability problem will be simple when we have the analytical solution whereas complicate when we don't have the analytical solution. In practice, when the parameter θ is time varying or of high dimension, the identifiability problem will be very difficult and several methods may have to be combined to solve it.

Once the ODE models are identifiabile, we can formally tackle the parameter estimation problem. In the next session, we are going to review the data fitting criteria based which several parametric estimation approaches were proposed.

2.2 Estimation (Data Fitting) Criteria

Suppose $\{\boldsymbol{\epsilon}(t_i)\}_{i=1}^n$ are independent distributed with mean zero and variance covariance matrix Σ_i, then the estimation criterion can be the error sum of squares

$$C(\boldsymbol{\theta}, \Sigma_i|\boldsymbol{y}) = \sum_{i=1}^{n} \boldsymbol{\epsilon}(t_i)^T \Sigma_i^{-1} \boldsymbol{\epsilon}(t_i) = \sum_{i=1}^{n} [\boldsymbol{y}_i - \boldsymbol{x}(t_i)]^T \Sigma_i^{-1} [\boldsymbol{y}_i - \boldsymbol{x}(t_i)] \tag{2.3}$$

where Σ_i^{-1} is the normalizing weight that rescales the observation $\boldsymbol{y}_i$.

Suppose $\{\boldsymbol{\epsilon}(t_i)\}_{i=1}^n$ are independent distributed from a distribution $G(\cdot)$, then the estimation criterion can be the negative log likelihood

$$C(\boldsymbol{\theta}, \Sigma_i|\boldsymbol{y}) = -\sum_{i=1}^{n} logG(\boldsymbol{\epsilon}(t_i)|\boldsymbol{\theta}, \Sigma_i) = -\sum_{i=1}^{n} logG(\boldsymbol{y}_i - \boldsymbol{x}(t_i)|\boldsymbol{\theta}, \Sigma_i) \tag{2.4}$$

The estimation of the parameter $\boldsymbol{\theta}$ can be obtained by solving the following ODE constrained minimization problem

$$\begin{aligned} &\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta},\Sigma}\, C(\boldsymbol{\theta}, \Sigma_i|\boldsymbol{y}) \\ &s.t. \begin{cases} \boldsymbol{x}'(t) = \boldsymbol{F}(t, \boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{\theta}) \\ \boldsymbol{x}(0) = \boldsymbol{x}_0 \end{cases} \end{aligned} \tag{2.5}$$

If the error sum of squares (2.3) is used, the corresponding estimators are called least square estimators. On the other hand, if the negative log likelihood (2.4) is used, the corresponding estimators are called maximum likelihood estimators. When $G(\cdot)$ is a normal distribution function, the two estimation criteria are the same. In the following, we illustrate the use of these criteria through a simple example as follows.

Example 9 *Suppose we have independent observations* $2.95, 6.84, 20.65, 54.57$ *for the state variable* $x(t)$ *at time instant* $1, 2, 3, 4$ *from a normal distribution with standard deviation* 1*, where* $x(t)$ *is such that*

$$\begin{cases} x'(t) = \theta x(t) \\ x(0) = 1 \end{cases} \tag{2.6}$$

Provide an estimation of θ.

Solution 10 *First, let us find the solution of (2.6)*

$$x'(t) = \theta x(t)$$

$$\frac{dx(t)}{x(t)} = \theta dt$$

$$logx(t) = c_1 + \theta t$$

$$x(t) = ce^{\theta t}$$

Since $x(0) = ce^{0t} = c = 1$, the solution of (2.6) is $x(t) = e^{\theta t}$. Then applying the least square criterion, the estimation of θ is

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} [y_i - x(t_i)]^T [y_i - x(t_i)]$$

$$= \underset{\theta}{\operatorname{argmin}} (2.95 - e^{\theta})^2 + (6.84 - e^{2\theta})^2 + (20.65 - e^{3\theta})^2 + (54.57 - e^{3\theta})^2$$

$$= 1.00038$$

2.3 Numerical Approaches

In the subsection 2.2, the estimation criteria indicate that a general parameter estimation procedure involves two steps.

1. Given a set of parameter values and an initial conditions, the ODE is solved for a solution.

2. The solution given by the first step is plugged in (2.3) or (2.4) and minimizing which will lead to the estimations.

If a closed form solution to (2.3) or (2.4) is available, the problem in step 2 is essentially a standard nonlinear regression, which can be solved by many standard statistical tools. However, in practice most of ODEs, especially with a nonlinear function $\boldsymbol{F}$ do not have

a closed form solution. In this case, numerical methods such as the Euler method, the trapezoidal rule, the first order exponential integrator method, the Runge-Kutta algorithm, the shooting methods have to be used to approximate the solution of the ODEs. A good review of these methods can be found in the textbook of Gordon C. Everstine [29]. Consequently, the two criteria (2.3) or (2.4) can be minimized for the estimates of the unknown parameters $\boldsymbol{\theta}$. Specifically, let $t_0 < t_1 < \cdots < t_{n-1} = T$ be a set of time instants on the interval $[t_0, T]$, $h_j = t_j - t_{j-1}$ be the mesh size and $h = \max_{1 \le j \le m-1} h_j$, the numerical solution of $\boldsymbol{x}(t)$ at $(t_1, \cdots, t_{n-1})$ can be obtained by

$$\boldsymbol{x}_{j+1} = \boldsymbol{x}_j + h_j \boldsymbol{\Phi}(t_j, \boldsymbol{x}_j, \boldsymbol{x}_{j+1}, h) \tag{2.7}$$

where the specific form of $\boldsymbol{\Phi}$ depends on the specific numerical method used.

The numerical approach was first proposed by mathematicians in 1970s including Bard [9], Van Domselaar and Hemker [76], and Benson [11]. Xue *et.al.* [40] established the asymptotical properties for the numerical estimates of using the nonlinear least square criterion (2.3) as follows.

Theorem 11 *Let's assume*

A1. $\boldsymbol{\theta} \in \Theta$*, where* Θ *is a compact subset of* $\Re^p$ *with a finite diameter* $R_{\boldsymbol{\theta}}$

A2. $\Omega = \{\boldsymbol{x}(t, \boldsymbol{\theta}) : t \in (t_0, T), \boldsymbol{\theta} \in \Theta\}$ *is a closed and bounded convex subset of* $\Re^p$

A3. There exist two constants $-\infty < \boldsymbol{c}_1 < \boldsymbol{c}_2 < +\infty$ *such that* $\boldsymbol{y}(t, \boldsymbol{\theta})$.

A4. All partial derivatives of $\boldsymbol{F}$ *with respect to* t *and* $\boldsymbol{x}$ *up to order* p *exist and are continuous.*

A5. The numerical method for solving ODEs is of order p.

A6. For any $\boldsymbol{\theta} \in \Theta$, $E_t[\boldsymbol{x}(t, \boldsymbol{\theta} - \boldsymbol{x}(t, \boldsymbol{\theta}_0)]^2 = 0$ *if and only if* $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

A7. The first and second partial derivatives, $\frac{\partial \boldsymbol{x}(t,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ *and* $\frac{\partial^2 \boldsymbol{x}(t,\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$ *exist and are continous and uniformly bounded for all* $t \in (t_0, T)$ *and* $\boldsymbol{\theta} \in \Theta$.

A8. For the ODE numerical solution satisfies the same condition in A7.

A9. For random design points, $t_1, \cdots, t_n$ *are i.i.d.. The joint density function of* t *and* $\boldsymbol{y}$*,* $f(t, \boldsymbol{y})$ *is bounded for all* $t \in (t_0, T)$ *and* $\boldsymbol{y} \in (\boldsymbol{c}_1, \boldsymbol{c}_2)$*.*

A10. The true value of parameter $\boldsymbol{\theta}$*,* $\boldsymbol{\theta}_0$ *is interior in* Θ*.*

A11. $\tilde{\boldsymbol{\theta}} = \operatorname{argmin} E_0[(\boldsymbol{y} - \tilde{\boldsymbol{x}}(t, \boldsymbol{\theta}))^T (\boldsymbol{y} - \tilde{\boldsymbol{x}}(t, \boldsymbol{\theta}))]$ *is interior in* Θ*, where* $\tilde{\boldsymbol{x}}(t, \boldsymbol{\theta})$ *is the numerical solution of* $\boldsymbol{x}(t, \boldsymbol{\theta})$ *and* E_0 *is the expectation with respect to the joint probability distribution of* $(t, \boldsymbol{y})$ *at the true value* $\boldsymbol{\theta}_0$*.*

A12. The matrix $\boldsymbol{V}_1 = \left[E(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}_0} \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}_0^T})\right]^{-1} E(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}_0} \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}_0^T}) \left[E(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}_0} \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}_0^T})\right]^{-1}$ *is positive definite, where* E *is expectation with respect to* t*.*

A13. The matrix $\tilde{\boldsymbol{V}} = \left[E(\frac{\partial \tilde{\boldsymbol{x}}}{\partial \tilde{\boldsymbol{\theta}}} \frac{\partial \tilde{\boldsymbol{x}}}{\partial \tilde{\boldsymbol{\theta}}^T})\right]^{-1} E_0 \left[\frac{\partial \tilde{\boldsymbol{x}}}{\partial \tilde{\boldsymbol{\theta}}} (\boldsymbol{y} - \tilde{\boldsymbol{x}}(t, \boldsymbol{\theta}))\right]^{\otimes 2} \frac{\partial \tilde{\boldsymbol{x}}}{\partial \tilde{\boldsymbol{\theta}}} \left[E(\frac{\partial \tilde{\boldsymbol{x}}}{\partial \tilde{\boldsymbol{\theta}}} \frac{\partial \tilde{\boldsymbol{x}}}{\partial \tilde{\boldsymbol{\theta}}^T})\right]^{-1}$ *is positive definite.*

Under the above assumptions, We can conclude that

1. *Under the assumptions* $A1 - A10$*, if there exist a* $\lambda > 0$ *such that* $h = O(n^{-\lambda})$*, then* $\boldsymbol{\theta}$ *converges to* $\boldsymbol{\theta}_0$ *almost surely.*

2. *For* $h = O(n^{-\lambda})$ *with* $0 < \lambda \leq \frac{1}{min(p,4)}$ *where* p *is the order of the numerical method, under assumptions* $A1 - A10$ *and* $A12$*, we have that* $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \boldsymbol{V}_1)$*.*

3. *For* $h = O(n^{-\lambda})$ *with* $0 < \lambda \leq \frac{1}{max(p,4)}$*, under assumption* $A1 - A9$*,* $A11$ *and* $A13$*, we have that* $\sqrt{n}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) \xrightarrow{d} N(0, \tilde{\boldsymbol{V}}_1)$ *with* $\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O(h^{\frac{min(p,4)}{2}})$ *and* $\|\tilde{\boldsymbol{V}}_{\boldsymbol{1}} - \boldsymbol{V}_1\| = O(h^{\frac{min(p,4)}{2}})$

A detailed proof is given by Xue *et.al.* [40]. Though the numerical estimates are proved to have a good asymptotic property, it has many drawbacks in practice. Firstly, this approach is computationally prohibitive since it requires to solve the ODE numerically repeatedly for each update of parameters. Secondly, it has a strict requirement on the step size for the appropriate convergence, which results in a much longer computing

time. Moreover, the asymptotical variance of the numerical estimators has no closed form which prevents their usage in further statistical inferences. Thirdly, the accuracy of the numerical approximation in step 1 is questionable, especially when the dependent vector cannot be observed continuously. Last, the minimization involved in this approach is always problematic, where the conventional gradient based methods such as Gauss-Newton method and quasi-Newton may fail to converge or converge to local minima if the iteration starts from a badly chosen initial values. Moreover, Ascher [6] pointed out this NLS approach can not deal with a case where the fundamental matrix of the ODE (2.2) has exponentially increasing and decreasing modes or, more generally, has non-trivial dichotomy.

An alternative solution is the embedding approach proposed by Bock [12]. However, Osborne [61] pointed out the embedding approach requires additional information about the solution structure and suitable boundary conditions. Also, the embedding approach has to choose an appropriate embedding which can be difficult if no prior information about the dynamical system is known, and sometimes the embedding may not be available. Moreover, like the NLS approach, embedding approach requires the $i.i.d.$ assumption of data, which is too strong to be satisfied in real application since the data in time series always depend on time t.

Gelman [35] proposed a Bayesian approach which avoids the problems of local minima in the numerical approaches. The author modeled observations $\boldsymbol{y}_i$ by a density centered on the numerical solution, $\boldsymbol{x}_i = \boldsymbol{x}(t_i|\boldsymbol{\theta})$, for example, $\boldsymbol{y}_i \sim N\left[\boldsymbol{x}(t_i|\boldsymbol{\theta}), \sigma^2\right]$. Then a posterior likelihood of $\boldsymbol{\theta}$ given $\boldsymbol{y}_i$ can be obtained based on the Bayesian equation. And since the posterior likelihood has no closed form, MCMC techniques such as Gibbs sampler or Metropolis-Hastings sampler or other sampler have to be used to estimate

the parameter $\boldsymbol{\theta}$. However, the Bayesian approach also encounters other problems of the numerical approaches.

In the aforementioned numerical methods, the initial values of the state vector $\boldsymbol{x}(t_0)$ have to be known. Li [86] proposed a simultaneous approach in order to handle the unknown initial condition $\boldsymbol{x}(t_0)$. His approach breaks the ODE constraint in (2.5) into a fixed number of equality constraints by transforming the ODE into some difference equations, and moreover, it treats both the parameters $\boldsymbol{\theta}$ and the state vector $\boldsymbol{x}(t_i)$ as unknown parameters. Specifically, it discretizes the ODE (2.2) into the following difference equalities

$$\boldsymbol{x}_{i+1} - \boldsymbol{x}_i = h_i \boldsymbol{F}\left(t_{i+\frac{1}{2}}, \boldsymbol{u}(t_{i+\frac{1}{2}}), \frac{\boldsymbol{x}_i + \boldsymbol{x}_{i+1}}{2}, \boldsymbol{\theta}\right) \qquad i = 1, 2, \cdots, n-1$$

As a result, the ODE constrained minimization problem is converted into

$$\begin{aligned} &\min m(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) = \textstyle\sum_{i=1}^{n} [\boldsymbol{y}_i - \boldsymbol{x}_i]^T [\boldsymbol{y}_i - \boldsymbol{x}_i] \\ &s.t. \quad \boldsymbol{x}_{i+1} - \boldsymbol{x}_i = h_i \boldsymbol{F}\left(t_{i+\frac{1}{2}}, \boldsymbol{u}(t_{i+\frac{1}{2}}), \tfrac{\boldsymbol{x}_i + \boldsymbol{x}_{i+1}}{2}, \boldsymbol{\theta}\right) \qquad i = 1, 2, \cdots, n-1 \end{aligned} \tag{2.8}$$

Define

$$\begin{aligned} &\boldsymbol{z} = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \cdots, \boldsymbol{x}_n^T, \boldsymbol{\theta})^T \\ &\boldsymbol{C}_i(\boldsymbol{x}_i, \boldsymbol{x}_{i+1}, \boldsymbol{\theta}) = \boldsymbol{x}_{i+1} - \boldsymbol{x}_i - h_i \boldsymbol{F}\left(t_{i+\frac{1}{2}}, \boldsymbol{u}(t_{i+\frac{1}{2}}), \frac{\boldsymbol{x}_i + \boldsymbol{x}_{i+1}}{2}, \boldsymbol{\theta}\right) \\ &\boldsymbol{C}(\boldsymbol{z}, \boldsymbol{\theta}) = (\boldsymbol{C}_1(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{\theta}), \cdots, \boldsymbol{C}_{n-1}(\boldsymbol{x}_{n-1}, \boldsymbol{x}_n, \boldsymbol{\theta}))^T \end{aligned}$$

then the Langrangian function associated with (2.8) is

$$l(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = m(\boldsymbol{z}) + \boldsymbol{\lambda}^T \boldsymbol{C}(\boldsymbol{z}, \boldsymbol{\theta})$$

where $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \cdots, \boldsymbol{\lambda}_{n-1})^T$. Li developed a new algorithm that combines the the sequential quadratic programming (SQP) Gauss-Newton approximation and SQP Newton approximation together to approximate the Hessian information of the Lagrangian function and finally solve the estimation problem. The contribution of Li's approach is that

it breaks the ODE into some pieces of difference equations. However, it also encounters many problems in practice. First of all, the discretization of ODE will decrease the accuracy of the estimates. Second, the discretization will bring in too many unknown parameters and too many constraints on the least square minimization, which results in numerous local minima and an intensive computation. When the sample size n gets big, the two problems become very troublesome, since the number of parameters as well as that of constraints increase linearly as n increases.

2.4 Smoothing Approaches

Ramsay *et.al.* [66] proposed an estimation approach to solve ODEs model (2.1) or (2.2) which is based on a data smoothing technique along with a generalization of profile estimation method.

On one hand, Ramsay expressed the $j^{th}(j = 1, \cdots, p)$ state variable $x_j(t)$ in $\boldsymbol{x}(t)$ in terms of a basis function expansion

$$x_j(t) = \sum_{k=1}^{K_j} \beta_{jk}\phi_{jk}(t) = \boldsymbol{\beta}_j^T \boldsymbol{\phi}_j(t) \tag{2.9}$$

where the number K_j of basis functions are chosen so that we have the flexibility to capture the variation in $x_j(t)$ and its derivative $x_j'(t)$, and $\boldsymbol{\beta}_j$ stands for the only unknown part of $x_j(t)$. The problem of obtaining $x_j(t)$ is then transformed into that of estimating the coefficients $\boldsymbol{\beta}_j$. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_p)^T$ be the vector of length $K = \sum_{j=1}^{p} K_j$. Let $\boldsymbol{\Phi}_j = (\boldsymbol{\phi}_j(t_1)^T, \boldsymbol{\phi}_j(t_2)^T, \cdots, \boldsymbol{\phi}_j(t_n)^T)^T$ be a N_j by K_j matrix, and let $\Phi = Diag\{\Phi_1, \cdots, \Phi_p\}$ be the $N = \sum_{j=1}^{n} N_j$ by K matrix. Based on this notation, $\boldsymbol{x}(t)$ can be expressed as

$$\boldsymbol{x}(t) = \boldsymbol{\Phi}(t)\boldsymbol{\beta} \tag{2.10}$$

The expression (2.9) or (2.10) of $\boldsymbol{x}(t)$ is then used to replace the state vector $\boldsymbol{x}(t)$ in the estimation criteria (2.3) or (2.4) for minimization.

On the other hand, Ramsay expressed each equation in (2.2) as a differential operator equation

$$L_{j,\boldsymbol{\theta}}(x_j) = \frac{dx_j(t)}{dt} - f_j(t, \boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\theta}) = 0 \quad j = 1, \cdots, p$$

Therefore, the goodness of fit to (2.2) for any given estimating function $\boldsymbol{x}_j$ can be measured under Euclidean norm by

$$\text{PEN}_j(\boldsymbol{x}) = \int_{t_0}^{T} [L_{j,\boldsymbol{\theta}}(x_j(t))]^2 dt$$

or under other norm such as L^1 by

$$\text{PEN}_j(\boldsymbol{x}) = \int_{t_0}^{T} |L_{j,\boldsymbol{\theta}}(x_j(t))| dt$$

And the overall goodness of fit for any given estimating function $\boldsymbol{x}$ can be measured by

$$\text{PEN}(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \sum_{j=1}^{p} \lambda_j \text{PEN}_j(\boldsymbol{x}) \tag{2.11}$$

where the multipliers $\lambda_j > 0$ are weights that control the relative emphasis on the data fitting for each state variable. As a result, the ODE constraint is converted into a penalty term that can be added into the estimation criteria (2.3) or (2.4), so that the ODE constrained minimization problem (2.5) is converted to the following penalized minimization problem (2.12) which includes three classes of parameters: (1) the system structural parameters $\boldsymbol{\theta}, \Sigma$, (2) the coefficients $\boldsymbol{\beta}$ defining the basis functions and (3) the tuning parameter $\boldsymbol{\lambda}$ defining the weights in the penalty term.

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}, \Sigma, \boldsymbol{\beta}, \boldsymbol{\lambda}}{\text{argmin}}\, J(\boldsymbol{\theta}, \Sigma, \boldsymbol{\beta}, \boldsymbol{\lambda} | \boldsymbol{y}) = \underset{\boldsymbol{\theta}, \Sigma, \boldsymbol{\beta}, \boldsymbol{\lambda}}{\text{argmin}}\, C(\boldsymbol{\theta}, \Sigma_i | \boldsymbol{y}) + \text{PEN}(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\lambda}) \tag{2.12}$$

Then, fixing the value of $\boldsymbol{\lambda}$, Ramsay proposed a profile approach to estimate $\boldsymbol{\theta}, \Sigma$ and $\boldsymbol{\beta}$ which works as follows. First, for any given value $\boldsymbol{\theta}, \Sigma$, we minimize (2.12) to

get the estimation of $\boldsymbol{\beta}$ in terms of $\boldsymbol{\theta}, \Sigma$, say $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}, \Sigma)$. After that, we plug $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}, \Sigma)$ into (2.3) or (2.4), by minimizing which we obtain the estimation of $\boldsymbol{\theta}$ and Σ. In general, the first minimization in the profile approach will require numerical solvers, but in the least square case and linear ODEs, Ramsay addressed that it is possible to express $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}, \Sigma)$ analytically. Moreover, Ramsay has also discussed the choice of the tuning parameter $\boldsymbol{\lambda}$, and investigated the large sample behavior of the parameter estimates as $\boldsymbol{\lambda} \to \infty$.

Further, Cao and Xu [44] proposed a robust smoothing approach which extends Ramsay's method to deal with outliers. In his approach, he defined a new residual measurement $f(r) = exp(-\rho(r))$ rather than the square residual, where $\rho(r)$ is a convex function and symmetrical about zero, quadratic in the neighborhood of zero and increasing at a rate slower than r^2. Accordingly, the estimation criteria (2.3) and (2.12) will be

$$C^*(\boldsymbol{\theta}, \Sigma_i|\boldsymbol{y}) = -\sum_{k=1}^{p}\sum_{i=1}^{n} \rho(\frac{y_k(t_{ki}) - x_k(t_{ki})}{\sigma_i}) \tag{2.13}$$

$$J(\boldsymbol{\theta}, \Sigma, \boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{y}) = \underset{\boldsymbol{\theta}, \Sigma, \boldsymbol{\beta}, \boldsymbol{\lambda}}{\operatorname{argmin}} C^*(\boldsymbol{\theta}, \Sigma_i|\boldsymbol{y}) + \mathrm{PEN}(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\lambda}) \tag{2.14}$$

The advantage of using the new residual measurement $f(r)$ is that it will downweight the extreme residuals in comparison with squared residuals. In practice, a common choice of $\rho(r)$ could be the family of Huber functions

$$\rho(r) = \begin{cases} r^2 & \text{if } |r| \le \kappa \\ 2\kappa|r| - \kappa^2 & \text{if } |r| > \kappa \end{cases} \tag{2.15}$$

Ramsay's smoothing approach brings a new idea to convert the ODE constraint into a penalty term so that the ODE constrained minimization problem is converted into a penalized minimization problem. Moreover, this approach avoids numerically solving the ODE and does not require to know the initial value of the state vector $\boldsymbol{x}$. However, this approach encounters several problems. First of all, this approach

introduces numerical errors which reduce the accuracy of the estimates, since it cannot compute the integration in the penalty analytically. Second, $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}, \Sigma)$ has no close form and cannot be expressed analytically in the maximum likelihood case or nonlinear ODEs case, which indicates the profile approach can only be helpful in the linear least square case. Last, this approach does not provide a feasible way to choose the tuning parameter $\boldsymbol{\lambda}$.

2.5 Two step Approach

An alternative approach that also does not require to solve the ODE was called "Two Step (TS) approach" proposed by Brunel [13]. Unlike the previous approaches that estimate the parameter directly via a parametric constrained minimization or penalized minimization process, the Two Step approach essentially divides the estimation problem into a nonparametric estimation step and a least square minimization step as described below

1. Consistent estimations of the state vector $\boldsymbol{x}(t)$ as well as its derivative $\boldsymbol{x}'(t)$, say $\hat{\boldsymbol{x}}(t)$ and $\hat{\boldsymbol{x}}'(t)$ are obtained by fitting the observations with flexible methods.

2. An estimation criterion can be defined by plugging $\hat{\boldsymbol{x}}(t)$ and $\hat{\boldsymbol{x}}'(t)$ into the ODE and then taking the difference between the two sides of the ODE

$$R^q_{n,w}(\boldsymbol{\theta}) = \|\boldsymbol{x}'(t) - \boldsymbol{F}(t, \boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{\theta})\|_{q,w} \tag{2.16}$$

where $\|f(t)\|_{q,w} = \left(\int_{t_0}^{T} |f(t)|^q w(t) dt\right)$ is a measure on $L^q(w)$ for any function $f(t)$ on $[t_0, T]$ w.r.t. the weight function $w(t)$, which is continuous and positive on $[t_0, T]$. Moreover, $q = 2$ and equation weight function $w(t_i) = \frac{1}{n}$ will lead to the standard nonlinear least square criterion of L^2 norm.

3. The TS estimators are such that the estimation criterion (2.16) is minimized.

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, R^q_{n,w}(\boldsymbol{\theta}) \tag{2.17}$$

The principle of the TS estimators is motivated by the fact that it is rather easy to construct consistent estimators of $\boldsymbol{x}(t)$ and $\boldsymbol{x}'(t)$ based on the observed data than solving them directly from the ODE. This approach was initialized by Varah [77] who suggested to use the least square splines in the first step. Other smoothing techniques were also suggested such as cubic splines by Madar [55], local polynomial regression by S. Ellner [15] and Wu [83]. or neural networks by E.O. Voit [78].

Brunel [13] established the asymptotic properties for the TS estimators when using least square splines in the first step, where he proved that under certain conditions, the TS estimators will be $\sqrt{n}$ consistent and asymptotically normal. Besides, Wu [22] established the large sample properties for the TS estimators when using local polynomial regression in the first step. With the asymptotic distribution, the interval estimates for $\boldsymbol{\theta}$ can be obtained and statistical inference can be further conducted. The TS approach provides new direction for ODE modeling since it is conceptually straightforward and doesn't require to know the solution to the ODE as well as the initial conditions. However, it has its own problem. First of all, the performance of TS approach heavily depends on the goodness of the nonparametric estimates $\hat{\boldsymbol{x}}(t)$ and $\hat{\boldsymbol{x}}'(t)$. However, in practice it's very hard to obtain such good estimates when the data are of small size or of bad local quality. In other words, the TS approach is very sensitive to the local behavior of the experimental data. The accuracy of TS estimator will be reduced dramatically with small change of data. Second, the TS estimator achieves the optimal $\sqrt{n}$ consistency only if the weight function $w(t)$ in the second step satisfies certain constraints. With a bad $w(t)$, the TS estimator may converge much slower. Therefore, the choice of

weight function is crucial. However, in practice there is no principle that guides how to choose it.

Motivated by a viral dynamics model for infectious disease, Chen and Wu [22] studied the following ODE models characterized by time-varying parameters

$$\begin{cases} \boldsymbol{x}'(t) = \boldsymbol{F}(t, \boldsymbol{x}(t), \boldsymbol{\theta}(t)) \\ \boldsymbol{x}(0) = \boldsymbol{x}_0 \end{cases} \tag{2.18}$$

He extended the TS approach to estimate the time-varying parameters as follows

1. A local polynomial regression is utilized to estimate the state vector $\boldsymbol{x}(t)$ as well as its derivative $\boldsymbol{x}'(t)$, say $\hat{\boldsymbol{x}}(t)$ and $\hat{\boldsymbol{x}}'(t)$. Specifically, the state vector can be approximated around a studied point t^*

$$x_k(t_i) \approx x_k(t^*) + (t_i - t^*)x_k^{(1)}(t^*) + \cdots + \frac{(t_i - t^*)^A x_k^{(A)}(t^*)}{A!} = \boldsymbol{z}_{i,A}^T \boldsymbol{\beta}_{k,A}(t^*) \tag{2.19}$$

where $\beta_{k,v(t^*)} = \frac{x_k^{(v)}(t^*)}{v!}$, $v = 0, \ldots, A$, and $\boldsymbol{z}_{i,m} = [1, (t_i - t^*), \ldots, (t_i - t^*)^m]^T$, and A is the order of the local polynomial. Then, the estimation $\hat{\boldsymbol{x}}_A(t^*) = [\hat{\beta}_{1,0}(t^*), \ldots, \hat{\beta}_{A,0}(t^*)]^T$ and $\hat{\boldsymbol{x}}'_A(t^*) = [\hat{\beta}_{1,1}(t^*), \ldots, \hat{\beta}_{A,1}(t^*)]^T$ can be obtained by minimizing the locally weighted least square criterion,

$$\sum_{k=1}^{p} \sum_{i=0}^{n} (y_k(t_i) - \boldsymbol{z}_{i,A}^T \boldsymbol{\beta}_{k,A}(t^*))^2 K_{h_k}(t_i - t^*) \tag{2.20}$$

with respect to β's. Here $K_{h_k}(\cdot) = K(\cdot/h_k)/h_k$ is a re-scaling function based on a kernel function $K(\cdot)$ with bandwidth $h_k > 0$.

2. A local linear method is used to approximate the parameter vector $\boldsymbol{\theta}(t)$ around t^*

$$\theta_h(t_i) = \alpha_{h,0}(t^*) + \alpha_{h,1}(t^*)(t_i - t^*) \quad h = 1, \cdots, q \tag{2.21}$$

3. Then, the parameter $\boldsymbol{\theta}(t)$ at time instant t^*, $\boldsymbol{\theta}(t^*) = (\theta_1(t^*), \cdots, \theta_q(t^*))^T$ can be estimated via minimizing the locally weighted least square criterion

$$\sum_{k=1}^{p} \sum_{i=0}^{n} (\hat{x}'_{k,A}(t_i) - f_k(\hat{\boldsymbol{x}}_A(t_i), (I_m \otimes z_{i,1}^T)\boldsymbol{\alpha}(t^*)))^2 K_b(t_i - t^*)$$

where $\otimes$ denotes the Kronecker product, I_m denotes an m-dimensional identity matrix, and $\boldsymbol{\alpha}(t^*) = (\boldsymbol{\alpha}_1^T(t^*), \cdots, \boldsymbol{\alpha}_q^T(t^*))^T$ with $\boldsymbol{\alpha}_j(t^*) = (\alpha_{j,0}(t^*), \alpha_{j,1}(t^*))^T$. $K_b(\cdot)$ is a kernel function with b being a properly selected bandwidth. Then, the local estimator of $\boldsymbol{\theta}(t^*)$ is $\hat{\boldsymbol{\theta}}(t^*) = (\hat{\alpha}_{1,0}(t^*), \cdots, \hat{\alpha}_{q,0}(t^*))^T$.

Chen and Wu [22] also established the asymptotical results for the proposed estimators and obtained closed form asymptotical biases and variances under certain conditions, where the bandwidth selection was suggested by the asymptotical results.

2.6 Orthogonal Condition Approach

To overcome the drawback of TS approach, Brunel [14] proposed an alternative approach called "Orthogonal Condition" (OC) approach. On one hand, the OC approach preserves all the advantages of the TS approach. On the other hand, it can not only avoid the choice of the weight function $w(t)$ but also achieve the $\sqrt{n}$ consistency. The most markable feature of the OC approach is the usage of orthogonal condition to characterize the solution of the ODE. The way of the characterization is so general that it does not impose any constraints into the estimation process. In this section, we review this approach in detail and demonstrated its usage in our own pollen tube growth example. However, it can be simply extended into other linear or nonlinear differential equations. Recall that the ODE (1.2) in the pollen tube growth example is

$$\begin{cases} D\frac{d^2R}{dx^2} = k_{nf}R - k_{pf}R^{\alpha}(1 - \frac{\int_x Rdx}{R_{tot}}) & x \in [-L_0, L_0] \\ R(-L_0) = R(L_0) = 0 \end{cases}$$

The OC estimation approach involves three main stages as follows.

2.6.1 OC Estimation Procedure

Stage I: nonparametric estimation of the state variable and its derivatives

In stage I, a nonparametric model is fitted to estimate the state variable $R(x)$ as well as its derivatives $R'(x)$ and $R''(x)$. In this dissertation, regression spline linear model is used in order to obtain closed form bias and variance estimates. Further, B-spline basis functions is used. Specifically, a set of $l = k + m$ B-spline basis functions of order m, denoted by $\mathbf{B}_l(x) = (B_1(x), B_2(x), \cdots, B_l(x))^T$, is firstly created with a proper chosen set of knots $\{t_0, t_1, \cdots, t_{k+1} = T\}$ that partitions $[t_0, T]$. These basises are then used to fit the data $\mathbf{Y} = (y(x_1), y(x_2), \cdots, y(x_n))^T$ to a simple linear regression model as follows

$$\begin{pmatrix} y(x_1) \\ y(x_2) \\ \vdots \\ y(x_n) \end{pmatrix} = \begin{pmatrix} B_1(x_1) & B_2(x_1) & \cdots & B_l(x_1) \\ B_1(x_2) & B_2(x_2) & \cdots & B_l(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ B_1(x_n) & B_2(x_n) & \vdots & B_l(x_n) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_l \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

that is, $\mathbf{Y} = \mathbf{B}_{n\times l}\mathbf{c} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is the vector of measurement error with mean $\mathbf{0}$ and variance matrix $\sigma^2\mathbf{I}_{n\times n}$. Hence,the estimated coefficients based on this model are

$$\hat{\mathbf{c}} = \left(\mathbf{B}_{n\times l}^T\mathbf{B}_{n\times l}\right)^{-1}\mathbf{B}_{n\times l}^T\mathbf{Y}$$

And the estimated state variable as well as its i^{th} order derivative are

$$\text{for } i = 0, 1, 2,\ \hat{R}^{(i)}(x) = \mathbf{B}_l^{(i)}(x)^T\hat{\mathbf{c}} = \mathbf{B}_l^{(i)}(x)^T\left(\mathbf{B}_{n\times l}^T\mathbf{B}_{n\times l}\right)^{-1}\mathbf{B}_{n\times l}^T\mathbf{Y} \tag{2.22}$$

where $\mathbf{B}_l^{(i)}(x) = (B_1^{(i)}(x), B_2^{(i)}(x), \cdots, B_l^{(i)}(x))^T$.

We now recall the asymptotic properties of the estimator $\hat{R}^{(i)}(x)$ derived by S. Zhou and Douglas A. Wolfe [87]:

Condition 12 ***(Condition on the Knots)***

$$\max_{1\le j\le k} |\frac{h_{j+1}}{h_j} - 1| = o(1) \quad \textit{and} \quad \frac{h}{\min_{1\le j\le k+1} h_j} \le d_1$$

where $h_j = t_j - t_{j-1}$ *is the mesh size,* $h = \max_{1\le j\le k+1} h_j$*,* $o(1) \to 0$ *as* $k \to 0$*, and* d_1 *is a pre-determined constant.*

Condition 13 ***(Condition on design points)***

$$\sup_{x\in X} |Q_n(x) - Q(x)| = o(\frac{1}{k})$$

where $Q_n(x)$ *is the empirical distribution function for the design points* $\{x_j\}_{j=1}^n$ *and* $Q(x)$ *is the corresponding distribution with density* $q(x)$*. Furthermore, if* $\{x_j\}_{j=1}^n$ *and* $q(x)$ *are such that*

$$\textit{for } j = 1, 2, \cdots, n, \quad \int_{x_j}^{x_{j+1}} q(x)dx = \frac{1}{n}$$

then (12) holds if $\frac{k}{n} \to 0$*.*

Theorem 14 *Under Conditions (12) and (13), if* $R \in C^m[0,1]$ *with* $m > 2$ *and* $\frac{k}{n} \to 0$*, and suppose* $\vec{\epsilon}$ *is the vector of measurement error with mean* $\boldsymbol{0}$ *and variance matrix* $\sigma^2 \boldsymbol{I}_{n\times n}$*, then*

$$\frac{\hat{R}^{(i)}(x) - R^{(i)}(x) - b_i(x)}{\sqrt{Var(\hat{R}^{(i)}(x))}} \xrightarrow{d} N(0,1)$$

where the bias is

$$b_i(x) = \frac{g^{(m)}(x) h_{j+1}^{m-i}}{(m-i)} B_{m-i}(\frac{x - t_j}{h_{j+1}}), \qquad \textit{if } t_j < x < t_{j+1}, j = 0, \ldots, k,$$

with $B_{m-i}(\cdot)$ *being a Bernoulli polynomial,*

and the variance is

$$Var(\hat{R}^{(i)}) = \frac{\sigma^2}{n} \boldsymbol{B}_{n\times l-i}^T G^{-1}(q) (D^{(i)})^T \boldsymbol{B}_{n\times(l-i)} + o(h^{-2i-1} n^{-1})$$

with

$$G(q) = \int_a^b \boldsymbol{B}_{n\times(l-i)} \boldsymbol{B}_{n\times(l-i)}^T q(x)dx$$

and $\boldsymbol{B}_{n\times(l-i)}$ *consisting of basis functions of order* $m-i$.

Based on Theorem 14, the bias of $\hat{R}^{(i)}(x)$ is $O(h^{m-i})$ and the variance is $O(h^{\frac{1}{2i-1}}n^{-1})$. As a result, the optimal mesh size $h_{opt} = O(n^{\frac{1}{2m+1}})$ gives rise to the minimum square error $\text{MSE}(\hat{R}^{(i)}(x)) = b_i^2(x) + \text{Var}(\hat{R}^{(i)}(x)) = O(n^{\frac{2(m-i)}{2m+1}})$. Therefore, under the conditions (12) and (13), and suppose $h = h_{opt}$, the nonparamtric estimates can achieve their optimal $\sqrt{n}$ convergence and thus are $\sqrt{n}$ consistent when $m \to \infty$.

Stage II: Construction of Orthogonal Conditions

In stage II, we construct orthogonal conditions to characterize the fidelity of estimates $\hat{R}(x)$ to the IDE. Notice that $R(x)$ can be uniquely identified by the condition that for any integrable function $\varphi(x)$ where $x \in [a, b]$,

$$\int_a^b D\frac{d^2R}{dx^2}\varphi(x)dx + \int_a^b (-k_{nf}R + k_{pf}R^{1.2}(1 - \frac{\int_x Rdx}{R_{total}}))\varphi(x)dx = 0$$

we define a weak solution of the IDE by relaxing the above requirement only upon the following set of functions.

Condition 15 *Let* $(\varphi_\ell)_{\ell\geq 1}$ *be a sequence of functions that are second continuously differentiable where* $x \in [a, b]$ *and satisfy the boundary conditions* $\varphi_\ell(a) = \varphi_\ell(b) = 0$ *and* $\varphi'_\ell(a) = \varphi'_\ell(b) = 0$. *The space spanned by* $(\varphi_\ell)_{\ell\geq 1}$ *is denoted by* $\mathcal{F}$.

Statement 2.1(a) A weak solution of the IDE is a function $R(x)$ such that for any function $\varphi_\ell(x)$ in Condition (15),

$$\int_a^b D\frac{d^2R}{dx^2}\varphi_\ell(x)dx + \int_a^b (-k_{nf}R + k_{pf}R^{1.2}(1 - \frac{\int_x Rdx}{R_{total}}))\varphi_\ell(x)dx = 0$$

Thanks to the nice boundary property of $\varphi_\ell(x)$, the first term vanishes when we integrate $\int_a^b D\frac{d^2R}{dx^2}\varphi_\ell(x)dx$ by parts, that is,

$$\int_a^b R''(x)\varphi_\ell(x)dx = R'(b)\varphi_\ell(b) - R'(a)\varphi_\ell(a) - \int_a^b R'(x)\varphi'_\ell(x)dx = -\int_a^b R'(x)\varphi'_\ell(x)dx$$
$$= -(R(b)\varphi'_\ell(b) - R(a)\varphi'_\ell(a)) + \int_a^b R(x)\varphi''_\ell(x)dx = \int_a^b R(x)\varphi''_\ell(x)dx$$

As a result, we have an equivalent statement as follows.

Statement 2.1(b) A weak solution of the ODE is a function $R(x)$ such that for any function $\varphi_\ell(x)$ in Condition (15),

$$\int_a^b DR\varphi''_\ell(x)dx + \int_a^b (k_{nf}R - k_{pf}R^{1.2}(1 - \frac{\int_x Rdx}{R_{total}}))\varphi_\ell(x)dx = 0$$

Denote $\vec{\theta} = (k_{nf}, k_{pf})^T$, we then define a condition $e_\ell(\hat{R}, \vec{\theta})$ for each $\varphi_\ell(x)$ as follows

$$e_\ell(\hat{R}, \vec{\theta}) = \int_a^b D\hat{R}\varphi''_\ell(x)dx + \int_a^b (k_{nf}\hat{R} - k_{pf}\hat{R}^{1.2}(1 - \frac{\int_x \hat{R}dx}{R_{total}}))\varphi_\ell(x)dx$$

And we call the conditions $e_1(\hat{R}, \vec{\theta})$ and $e_2(\hat{R}, \vec{\theta})$ are orthogonal with each other, if $\varphi_1(x)$ and $\varphi_2(x)$ are orthogonal functions, that is, $\int_a^b \varphi_1(x)\varphi_2(x)dx = 0$. Obviously, a reasonable estimate of $\vec{\theta}$ will minimize $e_\ell(\hat{R}, \vec{\theta})$ for $(\varphi_\ell)_{\ell\geq 1}$.

Stage III: least square optimization

In stage III, we choose sufficient orthogonal conditions to identify the estimate of $\vec{\theta}$. Recall in stage II, it's possible to obtain an infinite number of conditions. In such a case, there will be far more conditions than parameters so that the parameters will be over identified. One possible way to solve this problem is to use a set of sufficient good conditions instead of all the conditions for parameter identification. In this paper, we used a fixed number L ($L > 2$) conditions, where the magnitude of L is related to the statistical property of the final OC estimators. By "good", we mean all the conditions

should be orthogonal with each other, that is, they are constructed from a number of orthogonal functions $(\varphi_\ell(x))_{\ell=1}^{L}$. For the second order DE, B-spline basis functions are a good choice for $(\varphi_\ell(x))_{\ell=1}^{L}$ since they are nearly orthogonal. Other basis functions, such as polynomial basis, sine basis can also be used as $(\varphi_\ell(x))_{\ell=1}^{L}$ in many other DEs.

Let $\mathbf{e}_L(\hat{R},\vec{\theta})$ be the vector of L orthogonal conditions $(e_\ell(\hat{R},\vec{\theta}))_{\ell=1}^{L}$, where L is sufficiently large, then the orthogonal condition (OC) estimate of $\vec{\theta}$ is defined by minimizing a least square criterion $Q_{n,L}(\vec{\theta})$ as follows

$$\hat{\vec{\theta}}_{n,L} = \arg\min_{\vec{\theta}} Q_{n,L}(\vec{\theta}) = \arg\min_{\vec{\theta}} \left\| \mathbf{e}_L(\hat{R},\vec{\theta}) \right\|^2 \tag{2.23}$$

As mentioned above, we only use a portion of an infinite number of conditions to estimate $\vec{\theta}$, which may cause a loss of certain information. Nevertheless, we expect the OC estimate can also minimize all the orthogonal conditions, which gives rise to the next condition.

Condition 16 *$\hat{\vec{\theta}}_{n,L}$ is also a global minimizer of $Q_n(\vec{\theta}) = \left\| \boldsymbol{e}_{\mathcal{F}}(\hat{R},\vec{\theta}) \right\|^2$, where $\boldsymbol{e}_{\mathcal{F}}(\hat{R},\vec{\theta})$ is the vector consisting of all the infinite orthogonal conditions $(e_\ell(\hat{R},\vec{\theta}))_{\ell=1}^{\infty}$.*

2.6.2 Asymptotic Properties of OC estimators

In this section we derive the asymptotical representation for the OC estimator $\hat{\vec{\theta}}_{n,L}$, and linked its asymptotic property with that of $\hat{R}(x)$ through the orthogonal condition vector $\mathbf{e}_L(\hat{R},\vec{\theta})$. We further derive the closed form for the asymptotical variance of $\hat{\vec{\theta}}_{n,L}$ when using B-spline regression linear model in the first stage. The asymptotic property provides guidance of choosing the best L in stage III.

Asymptotical Representation

Denote $\vec{\theta}^*$ to be the true value of $\vec{\theta}$. Then the linearizaiton of $Q_{n,L}(\vec{\theta})$ gives rise to an asymptotical representation of the OC estimator, $\hat{\vec{\theta}}_{n,L}$, as follows:

$$\hat{\vec{\theta}}_{n,L} - \vec{\theta}^* = -M_L^* \mathbf{e}_L(\hat{R}, \vec{\theta}^*) + o_p(1) \tag{2.24}$$

where $M_L^* = [J_{\vec{\theta},L}(\hat{R}, \vec{\theta}^*)^T J_{\vec{\theta},L}(\hat{R}, \vec{\theta}^*)]^{-1} J_{\vec{\theta},L}(\hat{R}, \vec{\theta}^*)^T$, and $J_{\vec{\theta},L} = \frac{\partial \mathbf{e}_L(\hat{R}, \vec{\theta}^*)}{\partial \vec{\theta}^*}$ is the jacobian matrix of $\mathbf{e}_L(\hat{R}, \vec{\theta}^*)$ with respected to $\vec{\theta}^*$.

Proof. Recall $\mathbf{e}_L(\hat{R}, \vec{\theta}) = (e_1(\hat{R}, \vec{\theta}), e_2(\hat{R}, \vec{\theta}), \cdots, e_L(\hat{R}, \vec{\theta}))^T$. The OC estimator $\hat{\vec{\theta}}_{n,L}$ is defined by minimizing $Q_{n,L}^2(\vec{\theta})$ as follows:

$$Q_{n,L}^2(\vec{\theta}) = \mathbf{e}_L(\hat{R}, \vec{\theta})^T \mathbf{e}_L(\hat{R}, \vec{\theta}) = \left\| \mathbf{e}_L(\hat{R}, \vec{\theta}) \right\|^2$$

$$\hat{\vec{\theta}}_{n,L} = \arg\min_{\vec{\theta}} Q_{n,L}^2(\vec{\theta})$$

By tylor expansion, linearizing $Q_{n,L}^2(\vec{\theta})$ around $\vec{\theta}^*$ yields

$$\begin{aligned} Q_{n,L}^2(\vec{\theta}) &= Q_{n,L}^2(\vec{\theta}^*) + \frac{dQ_{n,L}^2(\theta)}{d\vec{\theta}^T}|_{\vec{\theta}=\vec{\theta}^*}(\vec{\theta} - \vec{\theta}^*) \\ &\quad + \frac{1}{2}(\vec{\theta} - \vec{\theta}^*)^T \frac{d^2 Q_{n,L}^2(\vec{\theta})}{d\vec{\theta} d\vec{\theta}^T}|_{\vec{\theta}=\vec{\theta}^*}(\vec{\theta} - \vec{\theta}^*) + o(\left\| \vec{\theta} - \vec{\theta}^* \right\|^2) \\ &= Q_{n,L}^2(\vec{\theta}^*) + 2J_{n,L}(\hat{R}, \vec{\theta}^*)^T \mathbf{e}_L(\vec{\theta} - \vec{\theta}^*) + \frac{1}{2} \times 2(\vec{\theta} - \vec{\theta}^*)^T \\ &\quad [(J_{n,L}(\hat{R}, \vec{\theta}^*)^T J_{n,L}(\hat{R}, \vec{\theta}^*) + H_{n,L}(\hat{R}, \vec{\theta}^*)^T \mathbf{e}_L](\vec{\theta} - \vec{\theta}^*) + o(\left\| \vec{\theta} - \vec{\theta}^* \right\|^2) \end{aligned}$$

where $H_{n,L}(\hat{R}, \vec{\theta}^*) = \frac{\partial^2 \mathbf{e}_L(\hat{R}, \vec{\theta}^*)}{\partial \vec{\theta}^* \partial [\vec{\theta}^*]^T}$ is the hessian matrix of $\mathbf{e}_L(\hat{R}, \vec{\theta}^*)$ with respected to $\vec{\theta}^*$. Therefore, to achieve the first order optimality, we require $\frac{dQ_{n,L}^2(\theta)}{d\vec{\theta}^T}|_{\vec{\theta}=\vec{\theta}^*} = 0$, and $\frac{d^2 Q_{n,L}^2(\vec{\theta})}{d\vec{\theta} d\vec{\theta}^T} > 0$, which is equivalent to the next condition

Condition 17 *The L orthogonal conditions are such that $H_{n,L}(\hat{R}, \vec{\theta}^*)^T \boldsymbol{e}_L = 0$ and $J_{n,L}(\hat{R}, \vec{\theta}^*)$ is of full rank around $\vec{\theta}^*$.*

Under condition (17), the derivative on both sides of the above equation with respected to $\vec{\theta}$ at $\hat{\vec{\theta}}_{n,L}$ yields,

$$\mathbf{0} = 2J_{n,L}(\hat{R}, \vec{\theta}^*)^T \mathbf{e}_L + 2 \times \frac{1}{2} \times 2[J_{n,L}(\hat{R}, \vec{\theta}^*)^T J_{n,L}(\hat{R}, \vec{\theta}^*)](\hat{\vec{\theta}}_{n,L} - \vec{\theta}^*) + o_p(1)$$

Therefore,

$$\hat{\vec{\theta}}_{n,L} - \vec{\theta}^* = -M_L^* \mathbf{e}_L(\hat{x}, \vec{\theta}^*) + o_p(1)$$

This completes the proof. ■

The asymptotical representation of the OC estimator provides an insight of its asymtotics, which will be studied in the following two subsections.

Asymptotical Consistency and Normality

From subsection 4.1, $\hat{\theta}_{n,L}$ can be asymptotically represented by:

$$\hat{\vec{\theta}}_{n,L} = \vec{\theta}^* - M_L^* \mathbf{e}_L(\hat{R}, \vec{\theta}^*) + o_p(1)$$

Thus, the asymptotics of $\hat{\theta}_{n,L}$ is related to the behavior of $\mathbf{e}_L(\hat{R}, \vec{\theta}^*)$, which is a continuous function of the estimate $\hat{R}$. By delta method, $\hat{\theta}_{n,L}$ will be asymptotical normal. Such results were proven under an appropriate condition by Brunel [14].

Asymptotical Variance and Interval Estimate

In this subsection we describe in detail about the steps of computing the asymptotical variance of $\hat{\vec{\theta}}_{n,L}$ and further use it to construct interval estimates. We start the description by considering the problem of computing the variance of $\mathbf{e}_L(\hat{R}, \vec{\theta}^*)$. Since $\mathbf{e}_L(\hat{R}, \vec{\theta}^*)$ is a function of $\hat{R}$, the variance of $\mathbf{e}_L(\hat{x}, \vec{\theta}^*)$ can be obtained by multivariate delta method as follows.

For any functions $f_1(x)$ and $f_2(x)$, denote $< f_1(x), f_2(x) >$ to be $\int_a^b f_1(x)f_2(x)dx$.

Then, $\mathbf{e}_L(\hat{R}, \vec{\theta}^*)$ is

$$\mathbf{e}_L(\hat{R}, \vec{\theta}^*) = \begin{pmatrix} e_1(\hat{R}, \vec{\theta}^*) \\ e_2(\hat{R}, \vec{\theta}^*) \\ \vdots \\ e_L(\hat{R}, \vec{\theta}^*) \end{pmatrix}$$

where

$$\begin{cases} e_1(\hat{R}, \vec{\theta}^*) =< D\hat{R}'', \varphi_1 > - < k_{nf}^*\hat{R} - k_{pf}^*\hat{R}^{1.2}(1 - \frac{\int_x \hat{R}dx}{R_{total}}), \varphi_1 > \\ e_2(\hat{R}, \vec{\theta}^*) =< D\hat{R}'', \varphi_2 > - < k_{nf}^*\hat{R} - k_{nf}^*\hat{R}^{1.2}(1 - \frac{\int_x \hat{R}dx}{R_{total}}), \varphi_2 > \\ \vdots \\ e_L(\hat{R}, \vec{\theta}^*) =< D\hat{R}'', \varphi_L > - < k_{nf}^*\hat{R} - k_{pf}^*\hat{R}^{1.2}(1 - \frac{\int_x \hat{R}dx}{R_{total}}), \varphi_L > \end{cases}$$

which is equal to

$$\begin{cases} e_1(\hat{R}, \vec{\theta}^*) =< \hat{R}, D\varphi_1'' - k_{nf}^*\varphi_1 > + < \hat{R}^{1.2}, k_{pf}^*(1 - \frac{\int_x \hat{R}dx}{R_{total}})\varphi_1 > \\ e_2(\hat{R}, \vec{\theta}^*) =< \hat{R}, D\varphi_2'' - k_{nf}^*\varphi_2 > + < \hat{R}^{1.2}, k_{pf}^*(1 - \frac{\int_x \hat{R}dx}{R_{total}})\varphi_2 > \\ \vdots \\ e_L(\hat{R}, \vec{\theta}^*) =< \hat{R}, D\varphi_L'' - k_{nf}^*\varphi_2 > + < \hat{R}^{1.2}, k_{pf}^*(1 - \frac{\int_x \hat{R}dx}{R_{total}})\varphi_L > \end{cases}$$

Recall in stage I of the OC estimation procedure, $\hat{R}(x) = \mathbf{B}(t)^T\hat{\mathbf{c}}$, where the variance of $\hat{\mathbf{c}}$ is

$$Var(\hat{\mathbf{c}}) = \sigma^2(\mathbf{B}_{n\times l}^T\mathbf{B}_{n\times l})^{-1}$$

and the variance of $\hat{R}(x)$ is

$$Var(\hat{R}(x)) = \sigma^2\mathbf{B}(x)^T(\mathbf{B}_{n\times l}^T\mathbf{B}_{n\times l})^{-1}\mathbf{B}(x)$$

Now denote $a_\ell(x) \triangleq D\ddot{\varphi}_\ell - k^*_{nf}\varphi_\ell$ and $b_\ell(x) \triangleq k^*_{pf}\varphi_\ell$. Rewrite $e_\ell(\hat{R},\theta)$ as

$$\begin{aligned} e_\ell(\hat{R},\vec{\theta}^*) =& < \hat{R}, D\ddot{\varphi}_\ell - k^*_{nf}\varphi_\ell > + < \hat{R}^{1.2}, k^*_{pf}(1 - \frac{\int_x \hat{R}dx}{R_{total}})\varphi_\ell > \\ =& < \hat{R}, a_\ell(x) > + < \hat{R}^{1.2}, b_\ell(x)(1 - \frac{\int_x \hat{R}dx}{R_{total}}) > \\ =& \int \hat{R}(x)a_\ell(x)dx + \int \hat{R}^{1.2}(x)b_\ell(x)dx \left(1 - \frac{1}{R_{total}}\int \hat{R}(x)dx\right) \end{aligned}$$

Since $\hat{R}(x) = \mathbf{B}(t)^T\hat{\mathbf{c}}$,

$$e_\ell(\hat{R},\vec{\theta}^*) \triangleq f_\ell(\hat{\mathbf{c}}) = \left(\int \mathbf{B}(x)^T a_\ell(x)dx\right)\hat{\mathbf{c}} + \int (\mathbf{B}(x)^T\hat{\mathbf{c}})^{1.2}b_\ell(x)dx \left(1 - \frac{1}{R_{total}}\int \mathbf{B}(x)^T\hat{\mathbf{c}}dx\right)$$

Then, the derivative $f'_\ell(\hat{\mathbf{c}}) = \frac{df_\ell}{d\hat{\mathbf{c}}}$ is

$$\begin{aligned} f'_\ell(\hat{\mathbf{c}}) =& \int \mathbf{B}(x)a_\ell(x)dx + \int 1.2 \times (\mathbf{B}(x)^T\hat{\mathbf{c}})^{0.2}\mathbf{B}(x)b_\ell(x)dx \left((1 - \frac{1}{R_{total}})\int \mathbf{B}(x)^T\hat{\mathbf{c}}dx)\right) \\ &+ \int (\mathbf{B}(x)^T\hat{\mathbf{c}})^{1.2}(x)b_\ell(x)dx \left((-\frac{1}{R_{total}})\int \mathbf{B}(x)dx\right) \\ =& \int \mathbf{B}(x)a_\ell(x)dx + \int 1.2 \times \hat{R}(x)^{0.2}\mathbf{B}(x)b_\ell(x)dx \left((1 - \frac{1}{R_{total}}\int \hat{R}(x)dx)\right) \\ &- \frac{1}{R_{total}}\left(\int \hat{R}^{1.2}(x)b_\ell(x)dx\right)\int \mathbf{B}(x)dx \end{aligned}$$

By univariate delta method, the variance of $e_\ell(R,\vec{\theta}^*) = f_\ell(\hat{\mathbf{c}})$ is

$$Var(e_\ell(\hat{R},\vec{\theta}^*)) = (f'_\ell(\hat{\mathbf{c}})Var(\hat{\mathbf{c}})(f'_\ell(\hat{\mathbf{c}}))^T \tag{2.25}$$

Further, Let

$$\mathbf{f}'(\hat{\mathbf{c}}) = \frac{d\mathbf{f}}{d\hat{\mathbf{c}}^T} = \begin{pmatrix} f'_1(\hat{\mathbf{c}}) \\ f'_2(\hat{\mathbf{c}}) \\ \vdots \\ f'_L(\hat{\mathbf{c}}) \end{pmatrix}$$

By multivariate delta method, the variance of $\mathbf{e}_L(\hat{R},\vec{\theta}^*)$ is

$$Var(\mathbf{e}_L(\hat{R},\vec{\theta}^*)) = (\mathbf{f}'(\hat{\mathbf{c}})Var(\hat{\mathbf{c}})(\mathbf{f}'(\hat{\mathbf{c}}))^T \tag{2.26}$$

In the next step, we compute the variance of $\hat{\vec{\theta}}_{n,L}$. Based on the asymptotical representation $\hat{\vec{\theta}}_{n,L} = \vec{\theta}^* - M_L^* \mathbf{e}_L(\hat{R}, \vec{\theta}^*) + o_p(1)$, the asymptotical variance of $\hat{\vec{\theta}}_{n,L}$ is

$$Var(\hat{\vec{\theta}}_{n,L}) = M_L^* Var(\mathbf{e}_L(\hat{R}, \vec{\theta}^*)) M_L^{*T} \tag{2.27}$$

Finally, let's construct an interval estimate for $\vec{\theta}$. As mentioned above, $\hat{\vec{\theta}}_{n,L}$ is asymptotical normally distributed with mean $\vec{\theta}^*$ and variance $Var(\hat{\vec{\theta}}_{n,L})$. Therefore, a $100(1-\alpha)\%$ confidence interval for $\vec{\theta}$ will be

$$\text{for any } i = 1, 2,\ C.I.(\theta_i, 1-\alpha) = \left[\hat{\theta}_i \pm z_{\frac{\alpha}{2}} \left(Var(\hat{\vec{\theta}}_{n,L}) \right)^{\frac{1}{2}}_{ii} \right]$$

Notice that $\vec{\theta}^*$ in $Var(\hat{\vec{\theta}}_{n,L})$ is unknown, thus in practice $\hat{\vec{\theta}}_{n,L}$ will be used to obtain an useable interval estimates. Therefore, the final interval estimate will be

$$\text{for any } i = 1, 2,\ C.I.(\theta_i, 1-\alpha) = \left[\hat{\theta}_i \pm z_{\frac{\alpha}{2}} \left(\widehat{Var(\hat{\vec{\theta}}_{n,L})} \right)^{\frac{1}{2}}_{ii} \right] \tag{2.28}$$

Chapter 3

Literature Review of Mixed Effects Model

In the pollen tube example introduced in Chapter 1, the experimental data are observed upon multiple positions of multiple pollen tubes. Such data is known as "repeated measures data" in Statistics. The multiple measurements per pollen tube generally result in the correlated errors that are not negligible. Mixed effects model that allows a wide variety of error correlation pattern to be modeled provides a general and flexible estimation approach in this situation. Later in Chapter 4, we will see that our ODE model will eventually be transformed into a constrained nonlinear mixed effects model. In this section, we are going to review mixed effects model and the relevant estimation approaches.

3.1 Linear Mixed Effects Model

A general form of linear mixed effect model can be written under the individual level as follows

$$\begin{aligned} Y_{ij} =& \beta_0 + \beta_1 X_{ij1} + \cdots + \beta_p X_{ijp} \\ & + b_{i1} Z_{ij1} + \cdots + b_{iq} Z_{ijq} + \epsilon_{ij} \\ b_{ik} \sim& N(0, g_{kk}), Cov(b_k, b_k') = g_{kk'} \\ \epsilon_{ij} \sim& N(0, r_{ijj}), Cov(\epsilon_{ij}, \epsilon_{ij'}) = r_{ijj'} \end{aligned}$$

where

- Y_{ij} is the response variable for the j^{th} observation ($j = 1, \cdots, n_i$) on the i^{th} of m subjects.
- $\beta_1, \cdots, \beta_p$ are the fixed effect coefficients, which are identical for all subjects. $X_{ij1}, \cdots, X_{ijp}$ are the corresponding design matrices.
- $b_{i1}, \cdots, b_{iq}$ are the random effect coefficients for the i^{th} subject that are assumed to follow a multivariate normal distribution. $Z_{ij1}, \cdots, Z_{ijq}$ are the corresponding design matrices.
- ϵ_{ij} is the error for the observation in i^{th} subject, which is assumed to follow a multivariate normal distribution, and uncorrelated with the random effect b_i's.

Alternatively but equivalently, the linear mixed model can be written in the matrix form under the subject level as follows

$$\begin{aligned} \boldsymbol{Y}_i =& \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{\epsilon}_i \\ \boldsymbol{b}_i \sim& N_q(0, G) \\ \boldsymbol{\epsilon}_i \sim& N_{n_i}(0, R_i) \end{aligned} \tag{3.1}$$

where

- $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{in_i})^T$ is response vector for the observations in the i^{th} subject with variance $V_i = var(\boldsymbol{Y}_i) = R_i + \boldsymbol{Z}_i G \boldsymbol{Z}_i^T$.

- $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is the $p \times 1$ vector of the fixed effects. $\boldsymbol{X}_i$ is the $n_i \times p$ coefficient matrix for the fixed effects for observation in the i^{th} subject. The $(k_1, k_2)^{th}$ element of $\boldsymbol{X}_i$ is $X_{ik_1k_2}$.

- $\boldsymbol{b}_i = (b_{i1}, \cdots, b_{iq})^T$ is the $q \times 1$ vector of the random effects associated with the i^{th} subject. $\boldsymbol{Z}_i$ is the $n_i \times q$ coefficient matrix associated with $\boldsymbol{b}_i$. The $(k_1, k_3)^{th}$ element of $\boldsymbol{Z}_i$ is $X_{ik_1k_3}$.

- $\boldsymbol{\epsilon}_i$ is the error for the observation in the i^{th} subject, which is assumed to follow a multivariate normal distribution, and uncorrelated with the random effect b_i's.

- G is the covariance matrix for the random effects, and R is the covariance matrix for the error.

Let $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \cdots, \boldsymbol{Y}_m^T)^T$, $\boldsymbol{X} = (\boldsymbol{X}_1^T, \cdots, \boldsymbol{X}_m^T)^T$, $\boldsymbol{b} = (\boldsymbol{b}_1^T, \cdots, \boldsymbol{b}_m^T)^T$, $\boldsymbol{Z} = diag\{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_m\}$ and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^T, \cdots, \boldsymbol{\epsilon}_m^T)^T$, the linear mixed model can be further written in a matrix form under the population level as follows

$$\begin{aligned} \boldsymbol{Y} =& \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon} \\ \boldsymbol{b} \sim& N_q(0, \Sigma) \\ \boldsymbol{\epsilon} \sim& N_n(0, R) \end{aligned} \tag{3.2}$$

where $\Sigma = diag\{G, \cdots, G\}$ and $R = diag\{R_1, \cdots, R_m\}$, and the variance of $\boldsymbol{Y}$ is $V = Cov(\boldsymbol{Y}) = \boldsymbol{Z}\Sigma\boldsymbol{Z}^T + R$

Common approaches to estimate the linear mixed effects model (3.2) include Maximum Likelihood (ML) method, REstricted Maximum Likelihood (REML) method and Expectation-Maximization (EM) algorithm.

3.1.1 Maximum Likelihood method

In the Maximum Likelihood (ML) method, the fixed effect $\boldsymbol{\beta}$ and variance components Σ and R are estimated simultaneously by maximizing the log likelihood function of model (3.2)

$$l = \frac{n}{2}log(2\pi) - \frac{1}{2}|V| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X\beta})^T V^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta}) \tag{3.3}$$

Therefore, the estimations are solutions to

$$\frac{\partial L}{\partial V} = -\frac{1}{2}V^{-1} - \frac{1}{2}(V^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta})(\boldsymbol{Y} - \boldsymbol{X\beta})^T V^{-1}) = 0$$
$$\frac{\partial L}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}V^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta})^T V^{-1}) = 0$$

or equivalently,

$$\hat{V} = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\hat{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{V}^{-1}\boldsymbol{Y}$$

$\hat{\boldsymbol{\beta}}$ in this equation is called the best linear unbiased estimate (BLUE) of $\boldsymbol{\beta}$. The random effects vector $\boldsymbol{b}$ can be predicted by the best linear unbiased predictor (BLUP)

$$\hat{\boldsymbol{b}} = G\boldsymbol{Z}^T\hat{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

Henderson [42] overcame the difficulty of solving the inverse of V in the ML approach by proposing the mixed model equation (3.4)

$$\begin{bmatrix} \boldsymbol{X}^T\hat{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\hat{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^T\hat{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\hat{R}^{-1}\boldsymbol{Z} + \hat{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{b}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\hat{R}^{-1}\boldsymbol{Y} \\ \boldsymbol{Z}^T\hat{R}^{-1}\boldsymbol{Y} \end{bmatrix} \tag{3.4}$$

which only requires to solve the inverse of G and R. Further, he proved that the solution to these mixed model equations are exactly the BLUE for $\boldsymbol{\beta}$ and BLUP for $\boldsymbol{b}$. Indeed, the inverse of R and G is easier to obtain than the inverse of V so that less iterations are needed to numerically solving (3.4) for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}$.

3.1.2 Restricted Maximum Likelihood method

Patterson and Thompson [62] proposed a REstricted Maximum Likelihood (REML) approach to estimate the model (3.2). After appropriate linear transformations, the data are divided into two parts, with distinct log likelihood l_1 and l_2.

$$
\begin{aligned}
l_1 =& -\frac{n}{2}log(2\pi) - \frac{1}{2}log|A^TVA| - \frac{1}{2}(A^T\boldsymbol{Y})^T(A^TVA)^{-1}A^T\boldsymbol{Y} \qquad (3.5)\\
=& -\frac{n}{2}log(2\pi) - \frac{1}{2}log|V| - \frac{1}{2}log|\boldsymbol{X}^TV^{-1}\boldsymbol{X}| - \frac{1}{2}\boldsymbol{Y}^T(V^{-1} - V^{-1}\boldsymbol{X}(\boldsymbol{X}^TV^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^TV^{-1})\boldsymbol{Y}\\
l_2 =& -\frac{n}{2}log(2\pi) - \frac{1}{2}log|B^TVB| - \frac{1}{2}(BT\boldsymbol{Y} - B\boldsymbol{X}\boldsymbol{\beta})^T(B^TVB)^{-1}(B\boldsymbol{Y} - B\boldsymbol{X}\boldsymbol{\beta})\\
=& -\frac{n}{2}log(2\pi) - \frac{1}{2}log|\boldsymbol{X}^TV^{-1}\boldsymbol{X}| - \frac{1}{2}(T\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^TV^{-1}\boldsymbol{X}(\boldsymbol{X}^TV^{-1}\boldsymbol{X})^{-1}X^TV^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})
\end{aligned}
$$

where l_1 is the log likelihood of the transformed data $A^T\boldsymbol{Y}$ with $A = I - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and l_2 is the log likelihood of the transformed data $B^T\boldsymbol{Y}$ with $B = \boldsymbol{X}^TV^{-1}$. The estimates of the variance component V is such that l_1 is maximized, and the BLUE of $\boldsymbol{\beta}$ is such that l_2 is maximized. Therefore,

$$\frac{\partial l_2}{\partial \boldsymbol{\beta}} = 2(\boldsymbol{X}^TM^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}^TM^T\boldsymbol{Y}) = 0$$

where $M = V^{-1}\boldsymbol{X}(\boldsymbol{X}^TV^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^TV^{-1}$. As a result,

$$
\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\hat{M}\boldsymbol{X})^{-1}\boldsymbol{X}^TV^{-1}\boldsymbol{Y}\\
&= (\boldsymbol{X}^TV^{-1}\boldsymbol{X}(\boldsymbol{X}^TV^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^TV^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^TV^{-1}\boldsymbol{X}(\boldsymbol{X}^TV^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^TV^{-1}\boldsymbol{Y}\\
&= (\boldsymbol{X}^TV^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^TV^{-1}\boldsymbol{Y}
\end{aligned}
$$

3.1.3 Expectation-Maximization algorithm

Lindstorm and Bates [50] applied the Expectation-Maximization (EM) algorithm to estimate the model (3.2) by treating the random effects $\boldsymbol{b}$ as missing values. Let $\boldsymbol{\theta} \doteq \{\boldsymbol{\beta}, G, R\}$ be the parameter vector. The Expectation (E) step creates a Q function as the expectation of the log likelihood with respect to $\boldsymbol{b}$ evaluated at the current estimates of the parameters $\boldsymbol{\theta}$ and the observed data $\boldsymbol{Y}$. The Maximization (M) step computes the parameters by maximizing the Q function defined on the E step and uses them to update the parameters for the next E step. The estimates of $\boldsymbol{\theta}$ will be obtained after a desired number of iterations between E step and M step. In this subsection, we are going to review the EM algorithm in detail. For simplicity, the error variance are assumed to be $R = \sigma^2 I$ and the number of observations on all the subjects are assumed same, *i.e.*, $n_1 = \cdots = n_m = n$.

Instead of the likelihood of $\boldsymbol{Y}$, the EM algorithm focuses on the joint likelihood of the observable data $\boldsymbol{Y}$ and the missing values $\boldsymbol{b}$ (sometimes called latent variables)

$$L = \prod_{i=1}^{m} L(\boldsymbol{Y}_i|\boldsymbol{b}_i; \boldsymbol{\beta}, \sigma^2) L(\boldsymbol{b}_i; G) \tag{3.6}$$

where

$$L(\boldsymbol{Y}_i|\boldsymbol{b}_i; \boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i)^T(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i)\right\}$$

$$L(\boldsymbol{b}_i; G) = \left(\frac{1}{\sqrt{2\pi}}\right)^q |G|^{-\frac{1}{2}} exp\{-\frac{1}{2}\boldsymbol{b}_i^T G^{-1}\boldsymbol{b}_i\}$$

Therefore, the joint log likelihood is

$$l = logL = \sum_{i=1}^{m}(logL(\boldsymbol{Y}_i|\boldsymbol{b}_i; \boldsymbol{\beta}, \sigma^2)) + logL(\boldsymbol{b}_i; G) = \sum_{i=1}^{m}(l_1 + l_2)$$

where

$$l_1 = const - \frac{mn}{2} log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i)^T (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i)$$

$$l_2 = const - \frac{mn}{2} log|G| - \frac{1}{2} \sum_{i=1}^{m} \boldsymbol{b}_i^T G^{-1} \boldsymbol{b}_i$$

E-Step

At the E step, the Q function is defined as the expectation of l with respect to $\boldsymbol{b}$ given $\boldsymbol{Y}$ and current estimates of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_0$. That is, $Q(\boldsymbol{\theta}|\boldsymbol{Y}; \boldsymbol{\theta}_0) = E(l) = E(l_1) + E(l_2)$. Since $\boldsymbol{b}_i$ and $\boldsymbol{Y}_i$ are jointly normally distributed with mean and variance

$$E \begin{bmatrix} \boldsymbol{b}_i \\ \boldsymbol{Y}_i \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{X}_i\boldsymbol{\beta} \end{bmatrix} \qquad \text{and} \qquad Var \begin{bmatrix} \boldsymbol{b}_i \\ \boldsymbol{Y}_i \end{bmatrix} = \begin{bmatrix} G & G\boldsymbol{Z}_i^T \\ \boldsymbol{Z}_i G & V_i \end{bmatrix}$$

Then

$$\begin{bmatrix} I & G\boldsymbol{Z}_i^T V_i^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_i \\ \boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_i - G\boldsymbol{Z}_i^T V_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \\ \boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \end{bmatrix}$$

is normally distributed with

$$\text{mean} \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{and variance} \begin{pmatrix} G + G\boldsymbol{Z}_i^T V_i^{-1} \boldsymbol{Z}_i G & 0 \\ 0 & V_i \end{pmatrix}$$

As a result, $\boldsymbol{b}_i - G\boldsymbol{Z}_i^T V_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$ and $\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}$ are mutually independent. Therefore, the conditional expectation and conditional variance of $\boldsymbol{b}_i$ given $\boldsymbol{Y}_i$ and $\boldsymbol{\theta}_0$ are

$$\hat{\boldsymbol{b}_i} = E(\boldsymbol{b}_i|\boldsymbol{Y}_i; \boldsymbol{\theta}) = G_0 \boldsymbol{Z}_i^T V_0^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_0)$$

$$\hat{S}_i = var(\boldsymbol{b}_i|\boldsymbol{Y}_i; \boldsymbol{\theta}_0) = G_0 - G_0 \boldsymbol{Z}_i^T V_0^{-1} \boldsymbol{Z}_i G_0$$

Therefore,

$$
\begin{aligned}
E(l_1) =& const - \frac{mn}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{m} E[(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i)^T(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i)|\boldsymbol{Y}_i;\boldsymbol{\theta}_0] \\
=& const - \frac{mn}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{m} Tr(var(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i|\boldsymbol{Y}_i;\boldsymbol{\theta}_0)) \\
& - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(E(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i|\boldsymbol{Y}_i;\boldsymbol{\theta}_0))^T(E(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i|\boldsymbol{Y}_i;\boldsymbol{\theta}_0)) \\
=& const - \frac{mn}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{m} Tr(\boldsymbol{Z}var(\boldsymbol{b}_i|\boldsymbol{Y}_i;\boldsymbol{\theta}_0)\boldsymbol{Z}^T) \\
& - \frac{1}{2\sigma^2}\sum_{i=1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_iE(\boldsymbol{b}_i|\boldsymbol{Y}_i;\boldsymbol{\theta}_0))^T(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_iE(\boldsymbol{b}_i|\boldsymbol{Y}_i;\boldsymbol{\theta}_0)) \\
=& const - \frac{mn}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\sum[Tr(\boldsymbol{Z}_i\hat{S}_i\boldsymbol{Z}_i^T) + (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i)^T(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i)]
\end{aligned}
$$

and

$$
\begin{aligned}
E(l_2) =& const - \frac{mn}{2}log|G| - \frac{1}{2}\sum_{i=1}^{m} E(\boldsymbol{b}_i^T G^{-1}\boldsymbol{b}_i|\boldsymbol{Y}_i;\boldsymbol{\theta}_0) \\
=& const - \frac{mn}{2}log|G| - \frac{1}{2}\sum_{i=1}^{m} E(Tr(\boldsymbol{b}_i^T G^{-1}\boldsymbol{b}_i)|\boldsymbol{Y}_i;\boldsymbol{\theta}_0) \\
=& const - \frac{mn}{2}log|G| - \frac{1}{2}\sum_{i=1}^{m} E(Tr(G^{-1}\boldsymbol{b}_i\boldsymbol{b}_i^T)|\boldsymbol{Y}_i;\boldsymbol{\theta}_0) \\
=& const - \frac{mn}{2}log|G| - \frac{1}{2}\sum_{i=1}^{m} Tr(G^{-1}E(\boldsymbol{b}_i\boldsymbol{b}_i^T|\boldsymbol{Y}_i;\boldsymbol{\theta}_0)) \\
=& const - \frac{mn}{2}log|G| - \frac{1}{2}\sum_{i=1}^{m}[Tr(G^{-1}(var(\boldsymbol{b}_i|\boldsymbol{Y}_i;\boldsymbol{\theta}_0) + E(\boldsymbol{b}_i|\boldsymbol{Y}_i;\boldsymbol{\theta}_0)^T E(\boldsymbol{b}_i|\boldsymbol{Y}_i;\boldsymbol{\theta}_0)))] \\
=& const - \frac{mn}{2}log|G| - \frac{1}{2}\sum_{i=1}^{m}[Tr(G^{-1}(\hat{S}_i + \hat{\boldsymbol{b}}_i^T\hat{\boldsymbol{b}}_i))]
\end{aligned}
$$

Therefore, the Q function $Q(\boldsymbol{\theta}|\boldsymbol{Y};\boldsymbol{\theta}_0) = E(l) = E(l_1) + E(l_2)$ is

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{Y};\boldsymbol{\theta}_0) =& const - \frac{1}{2\sigma^2}\sum[Tr(\boldsymbol{Z}_i\hat{S}_i\boldsymbol{Z}_i^T) + (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i)^T(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i)] \\
& - \frac{mn}{2}log(\sigma^2) - \frac{mn}{2}log|G| - \frac{1}{2}\sum_{i=1}^{m}[Tr(G^{-1}(\hat{S}_i + \hat{\boldsymbol{b}}_i^T\hat{\boldsymbol{b}}_i))]
\end{aligned}
$$

M-Step

At M step, the Q function is maximized over $\boldsymbol{\theta}$ in order to update the estimate of $\boldsymbol{\theta}$.

To estimate $\boldsymbol{\beta}$, one have to solve the derivative of Q with respect to $\boldsymbol{\beta}$. Therefore,

$$\begin{aligned}
&\frac{\partial Q}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}\sum_{i=1}^{m} 2(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i)^T(-\boldsymbol{X}_i) = 0 \\
&\Leftrightarrow \sum_{i=1}^{m}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i)^T\boldsymbol{X}_i = 0 \\
&\Leftrightarrow \sum_{i=1}^{m}(\boldsymbol{Y}_i - \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i)^T\boldsymbol{X}_i = \boldsymbol{\beta}^T\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{X}_i \\
&\Leftrightarrow \boldsymbol{\beta}^T = \sum_{i=1}^{m}(\boldsymbol{Y}_i - \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i)^T\boldsymbol{X}_i(\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1} \\
&\Leftrightarrow \hat{\boldsymbol{\beta}} = (\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1}\sum_{i=1}^{m}\boldsymbol{X}_i^T(\boldsymbol{Y}_i - \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i)
\end{aligned}$$

To estimate G, one have to solve the derivative of Q with respect to G. Therefore,

$$\begin{aligned}
\frac{\partial Q}{\partial G} &= -\frac{mn}{2}G^{-1} - \frac{1}{2}\sum_{i=1}^{m}\frac{\partial}{\partial G}[Tr(G^{-1}(\hat{S}_i + \hat{\boldsymbol{b}}_i^T\hat{\boldsymbol{b}}_i))] \\
&= -\frac{mn}{2}G^{-1} - \frac{1}{2}\sum_{i=1}^{m}(-G^{-1}(\hat{S}_i + \hat{\boldsymbol{b}}_i^T\hat{\boldsymbol{b}}_i)G^{-1}) = 0 \\
&\Leftrightarrow mnG^{-1} = G^{-1}\sum_{i=1}^{m}(\hat{S}_i + \hat{\boldsymbol{b}}_i^T\hat{\boldsymbol{b}}_i)G^{-1} \\
&\Leftrightarrow \hat{G} = \frac{1}{mn}\sum_{i=1}^{m}(\hat{S}_i + \hat{\boldsymbol{b}}_i^T\hat{\boldsymbol{b}}_i)
\end{aligned}$$

To estimate σ^2, one have to solve the derivative of Q with respect to σ^2. Therefore,

$$\frac{\partial Q}{\partial \sigma^2} = -\frac{mn}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{m}[Tr(\boldsymbol{Z}_i\hat{S}_i\boldsymbol{Z}_i^T) + (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i)^T(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i)] = 0$$

$$\hat{\sigma}^2 = \frac{1}{mn}\sum_{i=1}^{m}[Tr(\boldsymbol{Z}_i\hat{S}_i\boldsymbol{Z}_i^T) + (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i)^T(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i)]$$

At the end of M step, $\hat{\boldsymbol{\theta}} = \{\hat{\boldsymbol{\beta}}, \hat{G}, \hat{\sigma}^2\}$ is used to update $\boldsymbol{\theta}_0$ in the next E step so that the EM algorithm can iterate between the two steps until the convergency of $\hat{\boldsymbol{\theta}}$.

3.2 Nonlinear Mixed Effects Model

A general form of nonlinear mixed effects model is

$$\begin{aligned} \boldsymbol{Y}_i =& \boldsymbol{f}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i\boldsymbol{b}_i) + \boldsymbol{\epsilon}_i \\ \boldsymbol{b}_i \sim& N_q(0, G) \\ \boldsymbol{\epsilon}_i \sim& N_{n_i}(0, R_i) \end{aligned} \tag{3.7}$$

where $\boldsymbol{f}$ is a real-valued nonlinear function of the fix factors $\boldsymbol{\beta}$ and the random factors $\boldsymbol{b}_i$.

The literature on the estimation of nonlinear mixed effects models is very extensive. The important references include Sheiner and Beal [67–69], Lindstrom and Bates [49], Al-Zaid [2], Lu and Meeker [54] and Walker [79]. Specifically, Sheiner and Beal [67–69] proposed a NONMEM approach which first linearizes $\boldsymbol{f}$ with Taylor approximation at the current estimates of the fixed effects and zero of the random effects, and then fits the resulting linearized mixed effects model by maximum likelihood approach. Lindstrom and Bates [49] proposed a more accurate approximation to $\boldsymbol{f}$ with Taylor expansion at the current estimates of the fixed effects and current predictors of random effects, and they used Newton's method to maximize the restricted likelihood associated with the linearized mixed effects model. Al-Zaid [2] modified the Lindstrom and Bates' approach by suggesting EM algorithm instead of Newton's method to estimate the linearized mixed effects model. Moreover, Lu and Meeker [54] proposed a two-stage estimation approach which is essentially a method of moments approach. Walker [79] utilized the EM algorithm to estimate the nonlinear mixed effects model directly, where he applied Monte Carlo simulation to evaluate the integrals involved in the E-step. In the following subsections, we are going to review some of these procedures in detail.

3.2.1 Lu and Meeker's approach

Lu and Meeker [54] proposed the method of moments (MM) approach to estimate the nonlinear mixed effects model (3.7) as follows.

1. For each subject i, obtain the estimates $\hat{\boldsymbol{\beta}}_i$ and $\hat{\boldsymbol{b}}_i$ by minimizing the least square

$$\{\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{b}}_i\} = \arg \min_{\{\boldsymbol{\beta}, \boldsymbol{b}_i\}} \sum_{j=1}^{n_i} (Y_{ij} - f(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i\boldsymbol{b}_i))^2$$

2. Estimate the fixed effects $\boldsymbol{\beta}$ by

$$\hat{\boldsymbol{\beta}} = \frac{1}{m} \sum_{i=1}^{m} \hat{\boldsymbol{\beta}}_i$$

3. Estimate σ^2 by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - f(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i\boldsymbol{b}_i))^2}{\sum_{i=1}^{m} n_i - m(p+q)}$$

4. Let

$$\bar{\boldsymbol{b}} = \frac{1}{m} \sum_{i=1}^{m} \hat{\boldsymbol{b}}_i \quad \text{and} \quad T_i = \left[\frac{\partial \boldsymbol{f}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i\boldsymbol{b}_i)}{\partial(\boldsymbol{\beta}, \boldsymbol{b}_i)^T}\right]^T \left[\frac{\partial \boldsymbol{f}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i\boldsymbol{b}_i)}{\partial(\boldsymbol{\beta}, \boldsymbol{b}_i)^T}\right] \Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \boldsymbol{b}_i=\hat{\boldsymbol{b}}_i}$$

   Then, estimate G by

$$\hat{G} = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\boldsymbol{b}}_i - \bar{\boldsymbol{b}})(\hat{\boldsymbol{b}}_i - \bar{\boldsymbol{b}})^T - \frac{\hat{\sigma}^2}{m} \sum_{i=1}^{m} A_i^{-1}$$

   where A_i is the lower right $(q \times q)$ submatrix of T_i .

5. Noticed $\hat{G}$ is not always positive definite, Amermiya [5] suggested a modification of $\hat{G}$ for the estimate of G

$$\tilde{G} = \begin{cases} \hat{G} & \text{if } \hat{G} \text{ is positive definite} \\ \hat{G}_+ & \text{if } \hat{G} \text{ is not positive definite} \end{cases}$$

   where $\hat{G}_+ = E\Psi_+E'$, in which Ψ_+ is a diagonal matrix whose diagonal elements $\Psi_{ii} = max(\psi_i, 0)$ where ψ_i is the eigenvalue of G, and E is a $q \times q$ matrix whose i^{th} columns is the eigenvector $\boldsymbol{e}_i$ associated with ψ_i.

Luke and Meeker [54] also established the asymptotic properties for the MM approach, where the estimators were proven to be consistent and asymptotically normally distributed. However, Al-Zaid [2] addressed that in practice when the sample size is small, the MM approach performs badly. And even when the sample size is fairly large, the MM approach is still not very accurate.

3.2.2 Lindstrom and Bates' approach

Lindstrom and Bates [49] proposed an iterative two step approach to fit the nonlinear mixed effects model (3.7) which essentially works as follows

1. Given current estimates $(\boldsymbol{\beta}^{(t)}, \boldsymbol{b}_i^{(t)})$ for $(\boldsymbol{\beta}, \boldsymbol{b}_i)$ and let $\boldsymbol{X}_{ij} = (X_{ij1}, \cdot, X_{ijp})^T$ and $\boldsymbol{Z}_{ij} = (Z_{ij1}, \cdot, Z_{ijq})^T$, use Taylor expansion to linearize $f(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}, \boldsymbol{Z}_{ij}^T\boldsymbol{b}_i)$

$$f(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}, \boldsymbol{Z}_{ij}^T\boldsymbol{b}_i) = f(\boldsymbol{X}_{ij}\boldsymbol{\beta}^{(t)}, \boldsymbol{Z}_{ij}\boldsymbol{b}_i^{(t)}) + \frac{\partial f}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}, \boldsymbol{b}_i=\boldsymbol{b}_i^{(t)}}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) + \frac{\partial f}{\partial \boldsymbol{b}_i}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}, \boldsymbol{b}_i=\boldsymbol{b}_i^{(t)}}(\boldsymbol{b}_i - \boldsymbol{b}_i^{(t)})$$

   As a result, the original data $Y_{ij} = f(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}, \boldsymbol{Z}_{ij}^T\boldsymbol{b}_i) + \epsilon_{ij}$ can be expressed in terms of Y_{ij}^* as

$$Y_{ij}^* = \frac{\partial f}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}, \boldsymbol{b}_i=\boldsymbol{b}_i^{(t)}}(\boldsymbol{\beta}) + \frac{\partial f}{\partial \boldsymbol{b}_i}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}, \boldsymbol{b}_i=\boldsymbol{b}_i^{(t)}}(\boldsymbol{b}_i) \tag{3.8}$$

   where, $Y_{ij}^* = Y_{ij} - f(\boldsymbol{X}_{ij}\boldsymbol{\beta}^{(t)}, \boldsymbol{Z}_{ij}\boldsymbol{b}_i^{(t)}) + \frac{\partial f}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}, \boldsymbol{b}_i=\boldsymbol{b}_i^{(t)}}\boldsymbol{\beta}^{(t)} + \frac{\partial f}{\partial \boldsymbol{b}_i}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}, \boldsymbol{b}_i=\boldsymbol{b}_i^{(t)}}\boldsymbol{b}_i^{(t)}$. Therefore, the original nonlinear mixed effects model (3.7) becomes a linear mixed effects model (LMM).

2. Fit the LMM (3.8) and update $(\boldsymbol{\beta}^{(t)}, \boldsymbol{b}_i^{(t)})$ by the BLUE of $\boldsymbol{\beta}$ and the BLUP of $\boldsymbol{b}_i$.

3. Iterate the above two steps until convergence.

To fit the LMM (3.8), Lindstrom and Bates applied the Newton's method to conduct the maximization of the restricted likelihood of Y_{ij}^*. On the other hand, Al-Zaid [2]

suggested to use EM algorithm, since via simulation he observed that by using EM algorithm one can obtain more accurate estimates. Since Bate's approach is an iterative approach, the choice of initial values is crucial. Al-Zaid [2] suggested the estimates of the MM approach to be the initial values. A necessary convergence criterion can be defined as the differences between the current estimates and the previous estimates being less than some prespecified cutoff. The asymptotic properties of the estimators based on this approach were discussed very briefly by Lindstrom and Bates.

3.2.3 Walker's approach

Walker [79] utilized the EM algorithm to fit a general nonlinear random effects model, where he proved that the estimates in the M step have analytical forms and the Q function in the E step can be evaluated by Monte Carlo integration. In this subsection, we are going to review it in detail.

Consider a general nonlinear random effects model

$$\begin{aligned} \boldsymbol{Y}_i =& \boldsymbol{f}(\boldsymbol{X}_i, \boldsymbol{b}_i) + \boldsymbol{\epsilon}_i \\ \boldsymbol{b}_i \sim& N_q(\boldsymbol{b}, G) \\ \boldsymbol{\epsilon}_i \sim& N_{n_i}(0, \sigma^2 I_{n_i}) \end{aligned} \tag{3.9}$$

where $\boldsymbol{f}(\boldsymbol{X}_i, \boldsymbol{b}_i) = (f(\boldsymbol{X}_{i1}, \boldsymbol{b}_i), \cdots, f(\boldsymbol{X}_{in_i}, \boldsymbol{b}_i))^T$ is a known $n_i \times 1$ vector. When $\boldsymbol{b}_i = A_i\boldsymbol{b} + \boldsymbol{\eta}_i$ with A_i a known design matrix and $\boldsymbol{\eta}_i$ $i.i.d.$ $\sim N(0, G)$, the nonlinear random effects model (3.9) will become the nonlinear mixed effects model (3.7).

Like in linear mixed effects model, the EM algorithm in nonlinear mixed effects model works in two steps alternatively. Let $\boldsymbol{\theta} = (\boldsymbol{b}, G, \sigma^2)^T$. In the E step, the Q function at the current estimates $\boldsymbol{\theta}_0$ is

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) = E[l(\boldsymbol{Y}, \boldsymbol{b}; \boldsymbol{\theta})|\boldsymbol{Y}, \boldsymbol{\theta}_0]$$

where

$$l(\boldsymbol{Y}, \boldsymbol{b}; \boldsymbol{\theta}) = const - \frac{\sum_{i=1}^{m} n_i}{2} log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{m} \|\boldsymbol{Y}_i - \boldsymbol{f}(\boldsymbol{X}_i, \boldsymbol{b}_i)\|^2$$
$$- \frac{1}{2} \sum_{i=1}^{m} (\boldsymbol{b}_i - \boldsymbol{b})^T G^{-1} (\boldsymbol{b}_i - \boldsymbol{b}) - \frac{m}{2} log|G|$$

In the M step, the estimates $\hat{\boldsymbol{\theta}}$ is such that

$$Q(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}_0) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) \quad \text{for all } \boldsymbol{\theta} \in \Omega$$

where Ω is the feasible region of $\boldsymbol{\theta}$. Or equivalently,

$$\frac{\partial}{\partial \boldsymbol{\theta}} Q(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}_0)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0 \tag{3.10}$$

The solutions to (3.10) have analytical forms

$$\hat{\boldsymbol{b}} = \frac{1}{m} \sum_{i=1}^{m} E(\boldsymbol{b}_i|\boldsymbol{Y}, \boldsymbol{\theta}_0)$$
$$\hat{G} = \frac{1}{m} \sum_{i=1}^{m} E\{(\boldsymbol{b}_i - \hat{\boldsymbol{b}})(\boldsymbol{b}_i - \hat{\boldsymbol{b}})^T|\boldsymbol{Y}, \boldsymbol{\theta}_0\}$$
$$\hat{\sigma}^2 = \frac{1}{\sum_{i=1}^{m} n_i} \sum_{i=1}^{m} E\{\|\boldsymbol{Y}_i - \boldsymbol{f}(\boldsymbol{X}_i, \boldsymbol{b}_i)\|^2|\boldsymbol{Y}, \boldsymbol{\theta}_0\}$$

Introducing the notation $\bar{\boldsymbol{b}}_i = E(\boldsymbol{b}_i|\boldsymbol{Y}, \boldsymbol{\theta}_0)$, $\Theta_i = var(\boldsymbol{b}_i|\boldsymbol{Y}, \boldsymbol{\theta}_0)$, $\bar{\boldsymbol{f}}_i = E(\boldsymbol{f}(\boldsymbol{b}_i)|\boldsymbol{Y}, \boldsymbol{\theta}_0)$ and $\Psi_i = var(\boldsymbol{f}(\boldsymbol{b}_i)|\boldsymbol{Y}, \boldsymbol{\theta}_0)$, then

$$\hat{\boldsymbol{b}} = \frac{1}{m} \sum_{i=1}^{m} \bar{\boldsymbol{b}}_i$$
$$\hat{G} = \frac{1}{m} \sum_{i=1}^{m} \{(\bar{\boldsymbol{b}}_i - \hat{\boldsymbol{b}})(\bar{\boldsymbol{b}}_i - \hat{\boldsymbol{b}})^T + \Theta_i\}$$
$$\hat{\sigma}^2 = \frac{1}{\sum_{i=1}^{m} n_i} \sum_{i=1}^{m} \{\|\boldsymbol{Y}_i - \bar{\boldsymbol{f}}_i\|^2 + tr(\Psi_i)\}$$

Therefore, in order to obtain the values of $\hat{\boldsymbol{b}}_i$, $\hat{G}$ and $\hat{\sigma}^2$, the quantities $\bar{\boldsymbol{b}}_i$, Θ_i, $\bar{\boldsymbol{f}}_i$ and Ψ_i have to be evaluated at each iterations. Walker suggested to use Monte Carlo simulation technique. Specifically, to obtain $\bar{\boldsymbol{b}}_i = E(\boldsymbol{b}_i|\boldsymbol{Y}, \boldsymbol{\theta}_0) = \int \boldsymbol{b}_i p(\boldsymbol{b}_i|\boldsymbol{Y}_i, \boldsymbol{\theta}_0) d\boldsymbol{b}_i$, where

$$p(\boldsymbol{b}_i|\boldsymbol{Y}_i, \boldsymbol{\theta}_0) = \frac{p(\boldsymbol{Y}_i|\boldsymbol{b}_i, \boldsymbol{\theta}_0) p(\boldsymbol{b}_i|\boldsymbol{\theta}_0)}{\int p(\boldsymbol{Y}_i|\boldsymbol{b}_i, \boldsymbol{\theta}_0) p(\boldsymbol{b}_i|\boldsymbol{\theta}_0) d\boldsymbol{b}_i}$$

and then,

$$\bar{\boldsymbol{b}}_i = \frac{\int \boldsymbol{b}_i p(\boldsymbol{Y}_i|\boldsymbol{b}_i, \boldsymbol{\theta}_0) p(\boldsymbol{b}_i d\boldsymbol{b}_i|\boldsymbol{\theta}_0)}{\int p(\boldsymbol{Y}_i|\boldsymbol{b}_i, \boldsymbol{\theta}_0) p(\boldsymbol{b}_i|\boldsymbol{\theta}_0) d\boldsymbol{b}_i}$$

Therefore, by Monte Carlo simulation $\bar{\boldsymbol{b}}_i$ is

$$\bar{\boldsymbol{b}}_i = \frac{\sum_{l=1}^{T} \boldsymbol{b}^l p(\boldsymbol{Y}_i|\boldsymbol{b}^l, \boldsymbol{\theta}_0)}{\sum_{l=1}^{T} p(\boldsymbol{Y}_i|\boldsymbol{b}^l, \boldsymbol{\theta}_0)}$$

where $\boldsymbol{b}^1, \cdots, \boldsymbol{b}^T$ is $i.i.d$ generated from $N(\boldsymbol{b}_0, G_0)$. Θ_i, $\bar{\boldsymbol{f}}_i$ and Ψ_i are obtained by

$$\begin{aligned}
\Theta_i &= \frac{\sum_{l=1}^{T} \boldsymbol{b}^l (\boldsymbol{b}^l)^T p(\boldsymbol{Y}_i|\boldsymbol{b}^l, \sigma_0^2)}{\sum_{l=1}^{T} p(\boldsymbol{Y}_i|\boldsymbol{b}^l, \sigma_0^2)} - \bar{\boldsymbol{b}}_i \bar{\boldsymbol{b}}_i^T \\
\bar{\boldsymbol{f}}_i &= \frac{\sum_{l=1}^{T} \boldsymbol{f}(\boldsymbol{b}^l, \boldsymbol{X}_i) p(\boldsymbol{Y}_i|\boldsymbol{b}^l, \sigma_0^2)}{\sum_{l=1}^{T} p(\boldsymbol{Y}_i|\boldsymbol{b}^l, \sigma_0^2)} \\
\Psi_i &= \frac{\sum_{l=1}^{T} \boldsymbol{f}(\boldsymbol{b}^l, \boldsymbol{X}_i) \boldsymbol{f}(\boldsymbol{b}^l, \boldsymbol{X}_i)^T p(\boldsymbol{Y}_i|\boldsymbol{b}^l, \sigma_0^2)}{\sum_{l=1}^{T} p(\boldsymbol{Y}_i|\boldsymbol{b}^l, \sigma_0^2)} - \bar{\boldsymbol{f}}_i \bar{\boldsymbol{f}}_i^T
\end{aligned}$$

Once $\hat{\boldsymbol{b}}$, $\hat{G}$, $\hat{\sigma}^2$ are obtained, they can be used to update $\hat{\boldsymbol{\theta}}_0$ in the next E step. Walker also discussed the time of computation for this method via simulations. More details about the simulation results can be found in his paper.

3.3 Constrained Mixed Effects Model

In many practical settings, the fixed effect $\boldsymbol{\beta}$ in linear mixed effects model (3.2) or nonlinear mixed effects model (3.7) or the nonlinear random effects model (3.9) may have to satisfy certain constraints. These constraints typically reflect the prior knowledge about the values of $\boldsymbol{\beta}$ that people have. For example, if the response variable is known to be increasing with one explanatory variable by fixing the other, then the regression coefficient of the fixed effect associated with that particular variable must be positive.

Formally, the constraints can be written as $\boldsymbol{c}(\boldsymbol{\beta}) = (c_1(\boldsymbol{\beta}), \cdots, c_k(\boldsymbol{\beta})) \geq 0$. When $\boldsymbol{c}(\boldsymbol{\beta}) = A\boldsymbol{\beta}$, the constraints are linear. Otherwise, the constraints are nonlinear. The constraints are inequality constraints unless the sign '>' doesn't hold at all.

Accordingly, the estimation procedure for the constrained mixed effects models has to incorporate the constraints. In essence, we have to maximize the likelihood (3.3) in ML approach or restricted likelihood (3.5) in the REML approach or the joint likelihood (3.6) in EM algorithm under the constraints. Literature for the constrained linear mixed effects model includes Shin [70], Edward [26], Peddada [63], Pilla [64], Fang [31], Cai [16] and Rosen [23]. However, all these references are only applicable to linear mixed model with linear constraints. To the best of our knowledge, the estimation problem of linear mixed effects model with nonlinear constraints and nonlinear mixed effects model with linear/nonlinear constraints haven't been solved yet. In the next chapter, we will discuss this problem with a particular example and propose several approaches accordingly.

Chapter 4

A Constrained Mixed Effects Model based on Differential Equation for Cell Polarity Signaling in Tip Growth of Pollen Tube

In chapter 2, we reviewed all the existing estimation approaches that can be utilized to estimate the parameters in ordinary differential equations. All these methods can be applied to the parameter estimation problem in the example of tip growth of pollen tube introduced in chapter 1. Nevertheless, after exploration of the ordinary differential equation relevant to this particular example, we in this chapter will propose several new estimation methods which are based on constrained maximum likelihood or constrained maximum restricted likelihood. Such methods will be demonstrated to have better performances than the existing methods via simulation studies and real pollen tube data analysis. Specially, we will first establish the results for the existence and uniqueness

of solution to IDE (1.2) and derive the solution over an admissible parameter space. As a result, the IDE model for multiple pollen tubes can be reparametrized into a nonlinear mixed effects model with linear constraints whose estimation problem was discussed in chapter 3. Then, we will propose a new method by extending REML estimation approach to embrace the constraints. Alternatively, we will modify the Method of Moments (MM) approach of Lu and Meeker [54] by minimizing the moment criterion under constraints.

This chapter is organized as follows. In section 4.1, we give sufficient and necessary conditions for existence and uniqueness of a positive solution to the IDE model, and derive a tractable generic expression for solutions of the IDE. In section 4.2, we introduce the IDE based nonlinear statistical model with linear constraint for a single subject and the adapted Constrained Nonlinear Least Square (CNLS) estimation procedure. In section 4.3, we will extend the Two Step estimation approach and Orthogonal Condition estimation approach to fit the IDE model nonparametrically. In section 4.4, we extend the statistical IDE model to multiple subjects as a nonlinear mixed effects model (with linear constraints). We propose two estimators in this setting: the Constrained Method of Moments (CMM) and Constrained REML(CREML). The asymptotic properties of these estimators are also discussed in their own sections. We examine the performance of the proposed estimation procedures through simulation studies in section 4.5 as well as real pollen tube data analysis in section 4.6. We summarize this chapter in section 4.7.

4.1 Existence and Uniqueness of Solution for Second Order Semilinear Integro Differential Equation

In this section, we first introduce the semilinear elliptic equation (4.1) and prove in Proposition 1 that it has an unique solution $\sigma_0(x)$.

$$\begin{cases} -\partial_x^2 u = -u + u^\alpha \quad x \in [-c, c] \\ u(-c) = u(c) = 0 \end{cases} \tag{4.1}$$

We then link it to the semlinear elliptic equation (4.2) and prove in Proposition 2 that the equation (4.2) has an unique solution $R_{\lambda'}(x)$.

$$\begin{cases} -D\partial_x^2 u = -k_{nf} u + \lambda' k_{pf} u^\alpha \quad x \in [-L_0, L_0] \\ u(-L_0) = u(L_0) = 0 \end{cases} \tag{4.2}$$

Proposition 18 *For all $c \in (0, \infty]$, there exists an unique position solution $\sigma_0(x)$ to (4.1) with Dirichlet conditions on $[-c, c]$. Moreover, $\sigma_0(x)$ is positive, even and increasing at $[-c, 0]$ and decreasing at $[0, c]$.*

Proposition 19 *For all $\lambda^{'} > 0$, there exists an unique positive solution $R_{\lambda'}(x)$ to (4.2) with Dirichlet conditions on $\Omega = [-L_0, L_0]$, where $L_0 > 0$ can be $+\infty$. Moreover, if $\sigma_0(x)$ is the unique positive solution to (4.1) defined on $\Omega' = \left[-L_0\sqrt{\frac{k_{nf}}{D}}, L_0\sqrt{\frac{k_{nf}}{D}}\right]$, then*

$$R_{\lambda'}(x) = \left(\frac{k_{pf}}{k_{nf}}\lambda'\right)^{\frac{1}{1-\alpha}} \sigma_0(\sqrt{\frac{k_{nf}}{D}}x) \doteq R_{\lambda,\mu}(x) = \lambda\sigma_0(\mu x). \tag{4.3}$$

Where $\lambda = (\frac{k_{pf}}{k_{nf}}\lambda^{'})^{\frac{1}{1-\alpha}}$, and $\mu = \sqrt{\frac{k_{nf}}{D}}$.

It is easy to see that if there exists an unique positive solution $R_{\lambda'}(x)$ to equation (4.2) such that $\lambda^{'} = 1 - \frac{\int_x R_{\lambda'}(x)dx}{R_{tot}}$, then $R_{\lambda'}(x)$ is also a solution to equation (1.2). In the following, Theorem 20 provides a sufficient condition under which certain solutions to equation (4.2) can also be solutions to equation (1.2), Theorem 21 provides necessary

conditions that all the solutions to equation (1.2) must also be solutions to equation (4.2).

Theorem 20 *Let σ_0 be the positive solution to (4.1) defined on $\Omega' = \left[-L_0\sqrt{\frac{k_{nf}}{D}}, L_0\sqrt{\frac{k_{nf}}{D}}\right]$. Consider the family of function $R_{\lambda,\mu}(x) = \lambda\sigma_0(\sqrt{\frac{k_{nf}}{D}}x)$ as defined in (4.3), for the discriminant function*

$$\Lambda(k_{nf}, k_{pf}, D, R_{tot}, \sigma_0) = \frac{k_{nf}}{k_{pf}} - \frac{1}{\alpha}\left(\frac{\alpha-1}{\alpha}\sqrt{\frac{k_{nf}}{D}}\frac{R_{tot}}{\|\sigma_0\|_1}\right)^{\alpha-1} \tag{4.4}$$

1. *If $\Lambda(k_{nf}, k_{pf}, D, R_{tot}, \sigma_0) > 0$, then there is no solution to (1.2) that can be found in the family of function $R_{\lambda,\mu}(x)$.*

2. *If $\Lambda(k_{nf}, k_{pf}, D, R_{tot}, \sigma_0) = 0$, then there is one solution to (1.2) that can be found in the family of function $R_{\lambda,\mu}(x)$.*

3. *If $\Lambda(k_{nf}, k_{pf}, D, R_{tot}, \sigma_0) < 0$, then there are two solutions to (1.2) that can be found in the family of function $R_{\lambda,\mu}(x)$.*

Proof. Recall from equation (4.3), $R_{\lambda'}(x) = R_{\lambda,\mu}(x) = \lambda\sigma_0(\sqrt{\frac{k_{nf}}{D}}x)$ is a solution to (4.2), where σ_0 is the positive solution to (4.1) defined on $\Omega' = \left[-L_0\sqrt{\frac{k_{nf}}{D}}, L_0\sqrt{\frac{k_{nf}}{D}}\right]$. $R_{\lambda'}(x)$ is also a solution to (1.2) if

$$\lambda' = \frac{k_{nf}}{k_{pf}}\lambda^{1-\alpha} = \left(1 - \frac{1}{R_{tot}}\|R_{\lambda'}\|_1\right).$$

where

$$\begin{aligned}
\|R_{\lambda'}\|_1 &= \int_{-L_0}^{L_0} R_{\lambda,\mu}(x)dx \\
&= \frac{\lambda}{\mu}\int_{-\mu L_0}^{\mu L_0} \sigma_0(y)dy \\
&= \lambda\sqrt{\frac{D}{k_{nf}}}\|\sigma_0\|_1
\end{aligned}$$

Denote $g(\lambda) \doteq \frac{k_{nf}}{k_{pf}} - \lambda^{\alpha-1} + \frac{1}{R_{tot}}\lambda^{\alpha}\sqrt{\frac{D}{k_{nf}}}\left\|\sigma_0\right\|_1$, then
$g'(\lambda) = \lambda^{\alpha-2}\left(-(\alpha-1) + \frac{\alpha}{R_{tot}}\sqrt{\frac{D}{k_{nf}}}\left\|\sigma_0\right\|_1\lambda\right)$. The root λ_c of $g'(\lambda)$ is $\lambda_c = \frac{\alpha-1}{\alpha}\sqrt{\frac{k_{nf}}{D}}\frac{R_{tot}}{\left\|\sigma_0\right\|_1}$.
And $g(\lambda)$ is decreasing in $[0, \lambda_c]$ and increasing in $[\lambda_c, +\infty]$. Notice that $g(0) = \frac{k_{nf}}{k_{pf}}$,
$\lim_{+\infty} g = +\infty$, and

$$\begin{aligned} g(\lambda_c) &= \frac{k_{nf}}{k_{pf}} - \left(\frac{\alpha-1}{\alpha}\sqrt{\frac{k_{nf}}{D}}\frac{R_{tot}}{\left\|\sigma_0\right\|_1}\right)^{\alpha-1} + \sqrt{\frac{D}{k_{nf}}}\frac{\left\|\sigma_0\right\|_1}{R_{tot}}\left(\frac{\alpha-1}{\alpha}\sqrt{\frac{k_{nf}}{D}}\frac{R_{tot}}{\left\|\sigma_0\right\|_1}\right)^{\alpha} \\ &= \frac{k_{nf}}{k_{pf}} - \frac{1}{\alpha}\left(\frac{\alpha-1}{\alpha}\sqrt{\frac{k_{nf}}{D}}\frac{R_{tot}}{\left\|\sigma_0\right\|_1}\right)^{\alpha-1} \end{aligned}$$

1. When $g(\lambda_c) > 0$, $g(\lambda) > 0$, no solution to (1.2) can be found from the family of solutions to (4.2).

2. When $g(\lambda_c) = 0$, $g(\lambda) > 0$ for $\lambda \neq \lambda_c$, therefore one solution $R_{\lambda_c,\mu}(x)$ to (1.2) can be found from the family of solutions to (4.2).

3. When $g(\lambda_c) < 0$, there exist $\lambda_1 \in [0, \lambda_c]$ and $\lambda_2 \in [\lambda_c, \infty]$ such that $g(\lambda_1) = 0$ and $g(\lambda_2) = 0$, therefore two solutions $R_{\lambda_1,\mu}(x)$ and $R_{\lambda_2,\mu}(x)$ to (1.2) can be found from the family of solutions to (4.2).

■

Theorem 21 *Any solution to (1.2) can be written in the form* $R_{\lambda,\mu}(x) = \lambda\sigma_0(\mu x)$.

Proof. To prove Theorem 21, it's only necessary to show that for any solution R of (1.2) on $[-L_0, L_0]$, there exist $\lambda, \mu > 0$ such that $\sigma_0(x) = \frac{1}{\lambda}R(\frac{x}{\mu})$ is a solution to (4.1) on $[-L_0/\mu, L_0/\mu]$. Denote $\bar{\lambda} = \frac{1}{\lambda}$, $\bar{\mu} = \frac{1}{\mu}$, we have
$\frac{\partial\sigma_0(x)}{\partial x} = \bar{\lambda}\bar{\mu}\frac{\partial R(\bar{\mu}x)}{\partial(\bar{\mu}x)}$ and $\frac{\partial^2\sigma_0(x)}{\partial x^2} = \bar{\lambda}\bar{\mu}^2\frac{\partial^2 R(\bar{\mu}x)}{\partial(\bar{\mu}x)^2}$. $\sigma_0(x)$ is a solution to (4) on $[-L_0\bar{\mu}, L_0\bar{\mu}]$

if and only if

$$-\frac{\partial^2\sigma_0(x)}{\partial x^2} = -\sigma_0(x) + \sigma_0^\alpha(x)$$

$$-\bar{\lambda}\bar{\mu}^2 R''(\bar{\mu}x) = -\bar{\lambda}R(\bar{\mu}x) + \bar{\lambda}^\alpha R^\alpha(\bar{\mu}x)$$

$$-\bar{\lambda}\bar{\mu}^2\frac{k_{nf}R(\bar{\mu}x)}{D} + \bar{\lambda}\bar{\mu}^2\frac{k_{pf}R^\alpha(\bar{\mu}x)}{D}\left(1 - \frac{\int_{-L_0}^{L_0}R(x)dx}{R_{tot}}\right) = -\bar{\lambda}R(\bar{\mu}x) + \bar{\lambda}^\alpha R^\alpha(\bar{\mu}x)$$

$$-\bar{\mu}^2\frac{k_{nf}R(\bar{\mu}x)}{D} + \bar{\mu}^2\frac{k_{pf}R^\alpha(\bar{\mu}x)}{D}\left(1 - \frac{\int_{-L_0}^{L_0}R(x)dx}{R_{tot}}\right) = -R(\bar{\mu}x) + \bar{\lambda}^{\alpha-1}R^\alpha(\bar{\mu}x)$$

when $\bar{\mu} = \sqrt{\frac{D}{k_{nf}}}$, $\bar{\lambda}$ can be obtained by solving the following equality:

$$\frac{k_{pf}}{k_{nf}}R^\alpha(\bar{\mu}x)\left(1 - \frac{\int_{-L_0}^{L_0}R(x)dx}{R_{tot}}\right) = \bar{\lambda}^{\alpha-1}R^\alpha(\bar{\mu}x)$$

for which

$$\bar{\lambda} = [\frac{k_{pf}}{k_{nf}}\left(1 - \frac{\int_{-L_0}^{L_0}R(x)dx}{R_{tot}}\right)]^{\frac{1}{\alpha-1}}$$

■

Remark 22 *Theorem 20 and Theorem 21 provide a way to obtain the solution $R(x)$ of (1.2) as following when the values of (k_{nf} and k_{pf}) are given*

1. *Solve the semilinear elliptic equation (4.1) on $\Omega' = \left[-L_0\sqrt{\frac{k_{nf}}{D}}, L_0\sqrt{\frac{k_{nf}}{D}}\right]$*
2. *Compute $\|\sigma_0\|_1$ and the discriminant function $\Lambda(k_{nf}, k_{pf}, D, R_{tot}, \sigma_0)$*
3. *If $\Lambda(k_{nf}, k_{pf}, D, R_{tot}, \sigma_0) = 0$, find the positive roots λ^* of $g(\lambda)$ and compute the solution $R_{\lambda^*.\mu}(x) = \lambda^*\sigma_0(\sqrt{\frac{k_{nf}}{D}}x)$*
4. *If $\Lambda(k_{nf}, k_{pf}, D, R_{tot}, \sigma_0) = 0$, find the positive roots λ_1^* and λ_2^* of $g(\lambda)$, and compute the solutions $R_{\lambda_1^*.\mu}(x) = \lambda_1^*\sigma_0(\sqrt{\frac{k_{nf}}{D}}x)$ and $R_{\lambda_2^*.\mu}(x) = \lambda_2^*\sigma_0(\sqrt{\frac{k_{nf}}{D}}x)$*

In practice, when there exist two solutions, we should choose the one that is closer to the experimental data.

Remark 23 *For $\lambda > 0$, the solution $R_{\lambda,\mu}(x)$ to (1.2) is a positive and even function. Moreover, it increases at $[-L_0, 0]$ and decreases at $[0, L_0]$, and the maximum $R_{\lambda,\mu}(0) = \max_{x\in\Omega} R_{\lambda,\mu}(x) > \lambda$. The proof is provided in the Appendix. It will be shown later in section 4.6 that the ROP data reflect the above qualitative properties.*

Remark 24 *From Theorem 20, we see that the solution of (1.2) is parametrized by μ and λ. Hence, (1.2) is not over parametrized by k_{nf} and k_{pf}.*

4.2 Single subject and constrained nonlinear fixed effects model

Suppose for a single subject, an observation of ROP1 intensity in position X_j (X_j is randomly selected from known distribution $F(x)$) on the membrane at static time is denoted by

$$Y_j = R(X_j; k_{nf}, k_{pf}) + \epsilon_j \qquad j = 1, 2, \cdots, n. \tag{4.5}$$

where $R(X;\cdot)$ is the solution of (1.2) and ϵ_j are *iid* from a certain distribution f with mean 0 and variance σ^2. As shown in Theorem 20 and Theorem 21 of section 4.1, $R(X;\cdot)$ exists if and only if the discriminant function $\Lambda(\cdot)$ is non-positive. Therefore, the above IDE based model is subject to the constraint

$$\begin{cases} \Lambda(k_{nf}, k_{pf}, D, R_{tot}, \sigma_0) = \frac{k_{nf}}{k_{pf}} - \frac{1}{\alpha}(\frac{\alpha-1}{\alpha}\sqrt{\frac{k_{nf}}{D}}\frac{R_{tot}}{\|\sigma_0\|_1})^{\alpha-1} \leq 0 \\ k_{nf} > 0 \\ k_{pf} > 0 \end{cases} \tag{4.6}$$

Theorem 25 *The above constrained nonlinear model can be reparametrized into the following model*

$$Y_j = \lambda\sigma_0(\mu X_j) + \epsilon_j. \tag{4.7}$$

with μ and λ subject to the constraint

$$\begin{cases} \Lambda^*(\mu,\lambda) = \mu R_{tot} - \lambda \left\|\sigma_0\right\|_1 > 0 \\ \mu > 0 \\ \lambda > 0 \end{cases}$$

where $\mu = \sqrt{\frac{k_{nf}}{D}}$ and λ is the root of $g(\lambda)$ defined in section 4.1. The choice of λ is guided in Remark 22.

Based on the observations $\{y_j\}_{j=1}^n$ at positions $\{x_j\}_{j=1}^n$ in the biological experiment, we propose a nonparametric estimation procedure called Constrained Nonlinear Least Square (CNLS) as follows

1. Compute $\sigma_0(x)$ from DE (4.1)

2. Estimate μ and λ by minimizing least squares

$$(\hat{\lambda},\hat{\mu}) = \arg\min_{\lambda,\mu} \sum_{j=1}^{n} (y_j - \lambda\sigma_0(\mu x_j))^2$$

   under the constraint

$$\begin{cases} \Lambda^*(\mu,\lambda) > 0 \\ \mu > 0 \\ \lambda > 0 \end{cases}$$

3. Convert $\hat{\mu}$ and $\hat{\lambda}$ to $\hat{k}_{nf}$ and $\hat{k}_{pf}$

$$\begin{cases} \hat{k}_{nf} = D\hat{\mu}^2 \\ \hat{k}_{pf} = \frac{D\hat{\mu}^2}{\hat{\lambda}^{\alpha-1} - \frac{\hat{\lambda}^{\alpha}\|\sigma_0\|_1}{\hat{\mu}R_{tot}}} \end{cases}$$

4. Estimate σ^2 by $\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{\lambda}\sigma_0(\hat{\mu}x_j)\right)^2$

In the first step of CNLS, the solution of σ_0 involves a boundary value problem in an ordinary differential equation, which can be solved by many methods including shooting method [72], [73], mono-implicit Runge-Kutta (MIRK) method [19] and collocation method [8] in R package "bvpSolve". The optimization in the second step is subject to one linear constraint and two box constraints. When there is no constraint, the optimization can be tackled by many gradient based methods such as the Newton method, the BFGS method, the Gauss-Newton method, etc. which require the objective function to be differentiable. On the other hand, the simplex method of Nelder, J. A. and Mead, R. [58] that directly searches the optimum allows the objective function to be not differentiable. To apply simplex method, we first incorporate the constraints into the objective function by defining

$$f(\mu,\lambda) = \begin{cases} \sum_{j=1}^{n} \left(y_j - \lambda\sigma_0(\mu x_j)\right)^2 & \text{if } \Lambda^*(\mu,\lambda) > 0 \text{ and } \mu > 0 \text{ and } \lambda > 0 \\ +\infty & o.w. \end{cases}$$

For the general Nonlinear Least Square (NLS) estimator, the asymptotic properties has been established by Jennrich in 1969 [45]. For the general Constrained NLS (CNLS) estimator, the asymptotic properties has been established by Wang in 1996 [82]. Below we present the asymptotic property of the CNLS estimator proposed in this paper. The proof is provided in the Appendix.

Theorem 26 *Denote $\boldsymbol{\theta} = (\mu,\lambda)^T$ to be the parameter, $\boldsymbol{\theta}_0$ to be the true value of the parameter, and $\hat{\boldsymbol{\theta}}_n$ to be the CNLS estimator with n sample points. Let $R(X;\boldsymbol{\theta}) = \lambda\sigma_0(\mu X)$, then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \sigma K^{-\frac{1}{2}} \boldsymbol{Z}$$

where $\boldsymbol{Z} \sim N(0, I_2)$ is a bivariate normal vector, $K = E_X[\nabla_{\boldsymbol{\theta}} R(X;\boldsymbol{\theta}_0) \nabla_{\boldsymbol{\theta}} R(X;\boldsymbol{\theta}_0)^T]$, and $\nabla_{\boldsymbol{\theta}} R(X;\boldsymbol{\theta}_0)$ is the gradient vector of $R(X;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Moreover,

$\hat{\sigma}^2$ proposed in the above procedure is a consistent estimator of σ^2, therefore by slutsky's Theorem

$$\frac{\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)}{\hat{\sigma}} \xrightarrow{d} K^{-\frac{1}{2}}\boldsymbol{Z}$$

4.3 Single subject and nonparametric estimation approaches

As we discussed in section 4.2, the problem is to estimate the parameters k_{nf} and k_{pf} in IDE model (1.2)

$$\begin{cases} D\frac{d^2R}{dx^2} = k_{nf}R - k_{pf}R^{\alpha}(1 - \frac{\int_x Rdx}{R_{tot}}) \quad x \in [-L_0, L_0] \\ R(-L_0) = R(L_0) = 0 \end{cases}$$

under the constraint (4.6) based on the observations of $\{\boldsymbol{Y}_j\}_{j=1}^n$ in model (4.5). In this section, we pretend we don't know the form of the solution $R(x)$ and apply the TS estimation approach and OC estimation approach to estimate the two parameters nonparametrically.

Recall that both TS approach and OC approach firstly obtain the estimates of the response variable in model (1.2) as well as its derivatives by fitting a nonparametric model to the data $\{\boldsymbol{Y}_j\}_{j=1}^n$. The estimates are denoted to be $\hat{R}(x)$ and $\hat{R}''(x)$ respectively, which are then used to develop the parameter estimation criteria (2.16) or (2.23) in the least square sense as follows

for TS approach criterion,

$$R_{n,w}^q(k_{nf}, k_{pf}) = \|D\hat{R}''(x) - k_{nf}\hat{R} + k_{pf}\hat{R}^{\alpha}(1 - \frac{\int_{-L_0}^{L_0}\hat{R}dx}{R_{tot}})\|_{q,w}$$

for OC approach criterion,

$$Q_{n,L}(\vec{\theta}) = \left\|\mathbf{e}_L(\hat{R}, \vec{\theta})\right\|^2$$

where $\mathbf{e}_L(\hat{R}, \vec{\theta})$ was defined in section 2. By minimizing the criteria over the parameters, $\hat{k}_{nf}$ and $\hat{k}_{pf}$ are obtained.

However in order for the IDE (1.2) to have a solution, the parameters must satisfy the constraint (4.6). As a result, the minimization in TS or OC approach must be conducted under such a constraint. Furthermore, from Remark 23 we know that $R(x)$ is a positive and even function which is increasing at $[-L_0, 0]$ and decreasing at $[0, L_0]$. Therefore, we need to impose several constraints onto the nonparametric model involved in the TS or OC approach so that the estimates $\hat{R}(x)$, $\hat{R}''(x)$ can preserve the nice shape. This is very helpful and highly recommended in practice because the constraints on the shape can help correct the potentially bad local behavior of the nonparametric model, especially when the data is of bad local quality. As a result, the estimates $\hat{R}(x)$, $\hat{R}''(x)$ and further $\hat{k}_{nf}$, $\hat{k}_{pf}$ will be robust against the local behavior of the data. From these two points of view, the new estimation approaches is called constrained TS approach and constrained OC approach.

There is one problem has not been solved yet, *i.e.*, how to impose the constraints onto the nonparametric model in order for $\hat{R}(x)$to keep the proper shape. Recall that in the nonparametric model introduced in section 2.6, we first created a set of B-spline basis functions $\mathbf{B}_l(x) = (B_1(x), B_2(x), \cdots, B_l(x))^T$ and then used these B-spline basis functions to run a regression against the response $\{\boldsymbol{Y}_j\}_{j=1}^n$, based on which the estimates $\hat{R}(x) = \hat{c}_1 B_1(x) + \hat{c}_2 B_2(x) + \cdots + \hat{c}_l B_l(x)$ and $\hat{R}'(x) = \hat{c}_1 B_1'(x) + \hat{c}_2 B_2'(x) + \cdots + \hat{c}_l B_l'(x)$ are obtained. Now suppose we purposely create a symmetrical set of B-spline basis functions by choosing a symmetrical set of knots and further control the regression

coefficients $c_1, \cdots, c_n$ to be symmetrically positive. That is, if l is even, we let

$$\begin{array}{ll} B_1(x) = B_l(-x) & c_1 = c_l \geq 0 \\ B_2(x) = B_{l-1}(-x) & c_2 = c_{l-1} \geq 0 \\ \vdots & \vdots \\ B_{\frac{l}{2}}(x) = B_{\frac{l}{2}+1}(-x) & c_{\frac{l}{2}} = c_{\frac{l}{2}+1} \geq 0 \end{array} \tag{4.8}$$

and if l is odd, we let

$$\begin{array}{ll} B_1(x) = B_l(-x) & c_1 = c_l \geq 0 \\ B_2(x) = B_{l-1}(-x) & c_2 = c_{l-1} \geq 0 \\ \vdots & \vdots \\ B_{\frac{l-1}{2}}(x) = B_{\frac{l+1}{2}}(-x) & c_{\frac{l-1}{2}} = c_{\frac{l+1}{2}} \geq 0 \end{array} \tag{4.9}$$

Then, the resulting estimate $\hat{R}(x)$ will be positive and symmetrical about the origin. Therefore, in the first step of our constrained approaches, we have to estimate the following model under constraints (4.8) or (4.9)

$$\begin{pmatrix} y(x_1) \\ y(x_2) \\ \vdots \\ y(x_n) \end{pmatrix} = \begin{pmatrix} B_1(x_1) & B_2(x_1) & \cdots & B_l(x_1) \\ B_1(x_2) & B_2(x_2) & \cdots & B_l(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ B_1(x_n) & B_2(x_n) & \vdots & B_l(x_n) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_l \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

or equivalently in a vector form,

$$\mathbf{Y} = \mathbf{B}_{n \times l}\mathbf{c} + \vec{\epsilon}$$

The fit to such a constrained linear model can be naturally transformed into a constrained least square minimization problem which can be solved by the simplex method in section 4.2 or by the dual method of Goldfarb and Idnani [37, 38] in R packages "quadprog".

4.4 Multiple subjects and constrained nonlinear random effect model

Suppose for multiple subjects, ROP1 intensity observed for subject i at position X_{ij} on the membrane at static time can be described by

$$Y_{ij} = R_i(X_{ij}) + \epsilon_{ij} \qquad i = 1, 2, \cdots, m; j = 1, 2, \cdots, n_i \tag{4.10}$$

where i indexes the subject, and j indexes the position. ϵ_{ij} are iid from a certain distribution f with mean 0 and variance σ^2. $R_i(x)$ is the solution of IDE (1.2) for the i^{th} subject. Following section 4.3, we have

$$\begin{cases} R_i(X_{ij}) = \lambda_i \sigma_0(\mu_i X_{ij}) \\ \mu_i = \sqrt{\frac{k_{n f_i}}{D}} \\ \lambda_i = \text{root of } g(\cdot) \end{cases} \tag{4.11}$$

We further assume that

$$\begin{pmatrix} \mu_i \\ \\ \lambda_i \end{pmatrix} \sim \text{NID} \left(\begin{pmatrix} \mu \\ \\ \lambda \end{pmatrix}, \Sigma \right) \tag{4.12}$$

As a result, σ^2 measures within subject variation and Σ measures between subject variation. As mentioned before, the paramters are subject to three constraints. This is a nonlinear mixed model with parameters subject to linear inequality constraints.

Denote $\boldsymbol{\theta}_i = (\mu_i, \lambda_i)^T$ and $\boldsymbol{\theta} = (\mu, \lambda)^T$, with the experimental data $\{y_{ij}\}_{i=1,j=1}^{i=m,j=n}$ and $\{x_{ij}\}_{i=1,j=1}^{i=m,j=n_i}$ we first extend the CNLS procedure and propose a new estimation procedure called Constrained Method of Moment (CMM), which works as follows:

1. Compute $\sigma_0(x)$ from equation (4.1)

2. For each subject i, estimate $\boldsymbol{\theta}_i$ by minimizing least squares

$$\hat{\boldsymbol{\theta}}_i = \arg\min_{\boldsymbol{\theta}_i} \sum_{j=1}^{n_i} \left(y_{ij} - \lambda_i \sigma_0(\mu_i x_{ij})\right)^2$$

under the constraint

$$\Lambda^*(\boldsymbol{\theta}_i) > 0 \text{ and } \boldsymbol{\theta}_i > 0$$

3. Estimate $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}} = \frac{\sum_{i=1}^{m} \hat{\boldsymbol{\theta}}_i}{m}$

4. Estimate σ^2 by $\hat{\sigma}^2 = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(y_{ij} - \hat{\lambda}_i \sigma_0(\hat{\mu}_i x_{ij})\right)^2}{\sum_{i=1}^{m} n_i - 2}$

5. Estimate Σ by $\hat{\Sigma} = \sum_{i=1}^{m} \frac{(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}})^T}{m-1} - \hat{\sigma}^2 \sum_{i=1}^{m} \frac{T_i^{-1}}{m}$, where $T_i = \left[\frac{\partial \boldsymbol{R}_i}{\partial \boldsymbol{\theta}_i^T}\right]^T \left[\frac{\partial \boldsymbol{R}_i}{\partial \boldsymbol{\theta}_i^T}\right] \Big|_{\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i}$ and $\boldsymbol{R}_i = (R(x_{i1}; \boldsymbol{\theta}_i), R(x_{i2}; \boldsymbol{\theta}_i), \cdots, R(x_{in_i}; \boldsymbol{\theta}_i))^T$

6. Modify the estimator of Σ by

$$\tilde{\Sigma} = \begin{cases} \hat{\Sigma} & \text{if } \hat{\Sigma} \text{ is positive definite} \\ \hat{\Sigma}_+ & \text{if } \hat{\Sigma} \text{ is not positive definite} \end{cases}$$

   where $\hat{\Sigma}_+ = Q \Psi_+ Q'$, in which Ψ_+ is a diagonal matrix whose diagonal elements $\Psi_{ii} = max(\psi_i, 0)$ where ψ_i is the eigenvalue of Σ, and Q is a 2×2 matrix whose i^{th} columns is the eigenvector q_i associated with ψ_i.

7. Convert $\hat{\boldsymbol{\theta}}$ to $\hat{k}_{nf}$ and $\hat{k}_{pf}$

This procedure is motivated by the Method of Moment (MM) proposed by Lu and Meeker [54] in attempt to fit a general nonlinear mixed model. Our contribution is to extend it to constrained case by adding a constraint in the second step. The CMM procedure is a natural extension of the CNLS procedure, because $\boldsymbol{\theta}$ is estimated by the average of $\boldsymbol{\theta}_i$ of CNLS procedure.

The asymptotic property of the MM estimator was discussed by Lu and Meeker [54]. Here, we extend to the case of constraint and show through the following theorem that the proposed CMM estimators also have good asymptotic properties. The proof is provided in the Appendix.

Theorem 27 *Assume that*

1. *the sample size from each subjects are equal, i.e.,* $n_1 = n_2 = \cdots = n_m = n$

2. *both* n *and* m *tend to* $+\infty$

Then, we have the following large sample properties for $\hat{\boldsymbol{\theta}}$

1. $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$

2. $\sqrt{m}\tilde{\Sigma}^{-\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{Z}$, *where* $\boldsymbol{Z} \sim N(0, I_2)$, *and* $\tilde{\Sigma} = \Sigma + \sigma^2 E_{\boldsymbol{\theta}}[(nK_i)^{-1}]$ *with* $K_i = E_X[\nabla_{\boldsymbol{\theta}_i} R(X; \boldsymbol{\theta}_i) \nabla_{\boldsymbol{\theta}_i} R(X; \boldsymbol{\theta}_i)^T]$.

Moreover, if $\hat{\sigma}^2$ *is a consistent estimator of* σ^2, *then*

1. $\hat{\Sigma} \xrightarrow{p} \Sigma$

2. $\sqrt{m}\hat{\tilde{\Sigma}}^{-\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{Z}$, *where* $\hat{\tilde{\Sigma}} = \hat{\Sigma} + \hat{\sigma}^2 E_{\boldsymbol{\theta}}[(nK_i)^{-1}]$.

If ϵ_{ij} are *iid* normal, we can convert the nonlinear mixed model to a linear mixed model by Taylor expansion, and thereafter propose an alterative procedure called Constrained Restricted Maximum Likelihood method (CREML), which works as follows:

1. Given current Best Linear Unbiased Predictors (BLUP) $(\hat{\mu}_i^{(t)}, \hat{\lambda}_i^{(t)})$ for (μ_i, λ_i), use Taylor expansion to express $R_i(\mu_i, \lambda_i; x)$ as

$$R_i(\mu_i, \lambda_i; x) \approx R_i(\hat{\mu}_i^{(t)}, \hat{\lambda}_i^{(t)}; x) + \frac{\partial R_i}{\partial \mu_i}|_{\mu_i = \hat{\mu}_i^{(t)}}(\mu_i - \hat{\mu}_i^{(t)}) + \frac{\partial R_i}{\partial \lambda_i}|_{\lambda_i = \hat{\lambda}_i^{(t)}}(\lambda_i - \hat{\lambda}_i^{(t)})$$

   As a result, the original expression of data $y_{ij} = R_i(\mu_i, \lambda_i, x_{ij}) + \epsilon_{ij}$ can be re-written as

$$y_{ij}^* = \frac{\partial R_i}{\partial \mu_i}|_{\mu_i = \hat{\mu}_i^{(t)}} \mu_i + \frac{\partial R_i}{\partial \lambda_i}|_{\lambda_i = \hat{\lambda}_i^{(t)}} \lambda_i. \tag{4.13}$$

   where, $y_{ij}^* = y_{ij} - R_i(\hat{\mu}_i^{(t)}, \hat{\lambda}_i^{(t)}; x_{ij}) + \frac{\partial R_i}{\partial \mu_i}|_{\mu_i = \hat{\mu}_i^{(t)}} \hat{\mu}_i^{(t)} + \frac{\partial R_i}{\partial \lambda_i}|_{\lambda_i = \hat{\lambda}_i^{(t)}} \hat{\lambda}_i^{(t)}$. And our original model becomes a Constrained Linear Mixed Effect Model (CLMM).

2. Fit the CLMM (4.13) under the constraint of $\Lambda^*(\mu, \lambda) > 0$, $\mu > 0$ and $\lambda > 0$. Such a constraint at the population level can be easily embranced by the existing LMM fitting techniques.

3. Update $(\hat{\mu}_i^{(t)}, \hat{\lambda}_i^{(t)})$ by the Best Linear Unbiased Predictors (BLUP) based on the Best Linear Unbiased Estiamtes (BLUE) $(\hat{\mu}, \hat{\lambda}, \hat{\Sigma}, \hat{\sigma}^2)$ of the CLMM (4.13) obtained from step 2.

4. Iterate the above three steps until convergence.

This procedure is motivated by the iterative procedure of Lindstrom and Bates [49]. Our contribution is to extend it to constrained case by adding a constraint on Step 2 and to use a simple way to update $(\hat{\mu}_i^{(t)}, \hat{\lambda}_i^{(t)})$ in the iteration process.

The convergence behavior of the CREML procedure depends on the starting value. With a bad starting value, a bad approximation will occur at the second step, so that the resulting CLMM will be hard to fit. The singularity of the variance convariance matrix will imply such a problem. On the other hand, with a good starting value the procedure converges quite fast. A good choice of starting value could be the estimates obtained from the CMM procedure. The convergence rule can be set as the difference of estimates in current step from that in previous step being smaller than a prespecified positive threshold. The model in step 2 is a CLMM. When there is no constraint, the model can be fitted by many existing approaches such as Maximum Likelihood (ML), Restricted Maximum Likelihood (REML) and Expectation-Maximization (EM) algorithm. In this paper, we consider REML and extend it to fit the model with constraint. Note that the likelihood in the first step of REML only involves in the variance component parameters Σ and σ^2, therefore their estimates won't be affected by the constraint. On the other

hand, the likelihood in the second step of REML involves the population parameters μ and λ. So their estimates should be obtained by maximizing the reduced likelihood under the constraint. And this constrained optimization problem was discussed in the previous section of single subject case.

It deserves to point out the CMM procedure controls the constraints at the individual level whereas the CREML procedure controls at the population level. Since constraints satisfied at the individual level will be automatically satisfied at the population level, the former is more strict than the latter. In many cases of real world application especially when the number of subjects, m is large, constraint at the population level is enough and more desirable.

4.5 Simulation study

In this section, simulation studies were conducted for the single subject case and the multiple subject case respectively. All the estimation procedures were implemented in R. From the proof of Remark 23, we know $\sigma_0(x)$ is an positive and even function that achieves its maximum at 0. Further, we know $\sigma_0(x)$ is $\sim \frac{1}{2}$ when $|x| = 5$ and ~ 0 when $|x| \geq 15$. Therefore, when $\mu = 1$, $R(x) = \lambda\sigma_0(\mu x)$ is ~ 0 when $|x| \geq 15$. Therefore, in the simulation the data of $R(x)$ when $\mu = 1$ were generated from $|x| < 15$. The values of α, D and R_{tot} used in the simulations were set to be 1.2, 0.1 and 797 respectively, which were obtained empirically from real data.

4.5.1 Single subject

The first simulation example is presented below to evaluate the performance of the CNLS procedure. The data were generated based on Remark 22, where the true

values used for the simulation were $k_{nf} = 0.1$, $k_{pf} = 0.1125$, $\sigma = 4, 8, 16$. Therefore, $\mu = 1$ and $\lambda = 34.1883$. For different σ, we generated 10000 data sets of size $n = 301$, and applied the CNLS procedure on each of them. Estimates of the parameters k_{nf} and k_{pf} were obtained for each of the 10000 data sets, based on which the relative bias, standard deviation and mean square error (MSE) were computed. From Table 4.1, we could see the CNLS procedure works quite well and it's quite robust against noise when the size of data is fairly large.

	$\hat{k}_{nf}$			$\hat{k}_{pf}$		
σ	bias	sd	MSE	bias	sd	MSE
4	7.4×10^{-5}	0.0028	7.8×10^{-6}	7.9×10^{-5}	0.0023	5.2×10^{-6}
8	1.2×10^{-4}	0.0057	3.2×10^{-5}	2.2×10^{-4}	0.0047	2.2×10^{-5}
16	7.7×10^{-4}	0.0114	1.3×10^{-4}	0.0008	0.0093	8.7×10^{-5}

Table 4.1: The estimates of the CNLS procedure for single subject data set

The second simulation example is presented below to compare the performance of the CNLS procedure and the constrained OC procedure. Specifically, we impose the constraints on OC procedure in three different ways: constraints of positiveness only ($c_i \geq 0$), constraints of symmetry only ($c_i = c_{l-i}$) or constraints of both ($c_i = c_{l-i} \geq 0$). The parameter setting and data generation process are same as that in the first simulation example. The results are shown in Table 4.2, 4.3 and 4.4. From these tables, we can see that the CNLS procedure outperforms the OC procedure with or without constraints, and adding constraints improves the OC procedure a lot.

	$\hat{k}_{nf}$			$\hat{k}_{pf}$		
$\sigma = 4$	bias	sd	MSE	bias	sd	MSE
CNLS	7.4×10^{-5}	0.0028	7.8×10^{-6}	7.9×10^{-5}	0.0023	5.2×10^{-6}
OC	3.0×10^{-3}	0.0058	4.3×10^{-5}	-3.2×10^{-3}	0.0063	5.0×10^{-5}
OC^1	-3.3×10^{-5}	0.0038	1.5×10^{-5}	9.7×10^{-4}	0.0031	1.1×10^{-5}
OC^2	-8.3×10^{-4}	0.0046	2.2×10^{-5}	8.8×10^{-4}	0.0047	2.3×10^{-5}
OC^3	3.9×10^{-4}	0.0039	1.5×10^{-5}	1.1×10^{-3}	0.0031	1.1×10^{-5}

Table 4.2: The estimates for single subject data set based on procedures of CNLS, OC without constraint, OC^1 with positiveness constraints only, OC^2 with symmetry constraint only and OC^3 with both constraints when $\sigma = 4$

	$\hat{k}_{nf}$			$\hat{k}_{pf}$		
$\sigma = 8$	bias	sd	MSE	bias	sd	MSE
CNLS	1.2×10^{-4}	0.0057	3.2×10^{-5}	2.2×10^{-4}	0.0047	2.2×10^{-5}
OC	-1.5×10^{-2}	0.0178	5.5×10^{-4}	-1.6×10^{-3}	0.0199	6.6×10^{-4}
OC^1	-1.4×10^{-3}	0.0073	5.6×10^{-5}	6.9×10^{-4}	0.0062	3.9×10^{-5}
OC^2	-6.5×10^{-3}	0.014	2.3×10^{-4}	-6.9×10^{-3}	0.015	2.7×10^{-4}
OC^3	2.3×10^{-4}	0.0075	5.6×10^{-5}	1.2×10^{-3}	0.0061	3.9×10^{-5}

Table 4.3: The estimates for single subject data set based on procedures of CNLS, OC without constraint, OC^1 with positiveness constraints only, OC^2 with symmetry constraints only and OC^3 with both constraints when $\sigma = 8$

4.5.2 Multiple subject

The aim of the simulation study in this subsection is to evaluate and compare the performance of the CMM procedure and the CREML procedure. The data for each of m subjects were generated based on Remark 22 after the assocaited (μ_i, λ_i) were simulated from the normal distribution (4.12). The true values of parameters used for the simulation were $\mu = 1$, $\lambda = 34.1883$, $\sigma_\mu = 0.6$, $\sigma_\lambda = 0.2$ and $\sigma = 4$. We generated 3000 data sets and fit each of them with both procedures to obtain the estimates of μ, λ, σ_μ, σ_λ and σ. The relative bias, standard deviation and MSE for these estimates were computed. We assumed $m = 10$ and considered different n by simulating data at

	$\hat{k}_{nf}$			$\hat{k}_{pf}$		
$\sigma = 16$	bias	sd	MSE	bias	sd	MSE
CNLS	7.7×10^{-4}	0.0114	1.3×10^{-4}	0.0008	0.0093	8.7×10^{-5}
OC	-4.6×10^{-2}	0.0350	3.4×10^{-3}	-4.9×10^{-2}	0.0393	4.0×10^{-3}
OC^1	-4.7×10^{-3}	0.0131	1.9×10^{-4}	-1.3×10^{-4}	0.0114	1.3×10^{-4}
OC^2	-2.5×10^{-3}	0.0325	1.6×10^{-3}	-2.5×10^{-2}	0.0359	1.9×10^{-3}
OC^3	-1.8×10^{-3}	0.0140	2.0×10^{-4}	1.3×10^{-3}	0.0117	1.4×10^{-4}

Table 4.4: The estimates for single subject data set based on procedures of CNLS, OC without constraint, OC^1 with positiveness constraints only, OC^2 with symmetry constraints only and OC^3 with both constraints when $\sigma = 16$

		$\hat{\mu}$	$\hat{\lambda}$	$\hat{\sigma}_\mu$	$\hat{\sigma}_\lambda$	$\hat{\sigma}$
CMM	bias	0.0052	0.0109	-0.0101	-0.0688	-0.0046
	sd	0.0617	0.2711	0.0451	0.2981	0.1294
	MSE	0.0038	0.0736	0.0021	0.0936	0.0168
CREML	bias	0.00712	0.0079	-0.0228	-0.1432	-0.0008
	sd	0.0610	0.2698	0.0412	0.3035	0.1267
	MSE	0.0038	0.0728	0.0022	0.1126	0.0160

Table 4.5: Comparison of the estimation procedures in Case 1 where x is equally spaced between -5 and 5 with increament 0.2 and $m = 10$, $n = 51$

different x. In case 1, $x = (-5, -4.8, \cdots, -0.2, 0, 0.2, \cdots, 4.8, 5)$ and $n = 51$. Whereas in case 2, $x = (-5, -1, -0.2, 0.2, 1, 5)$ and $n = 6$.

From Table 4.5, we can see that when n is large, the estimates of μ, λ, σ_μ, σ_λ and σ are about the same for the CMM procedure and the CREML procedure. It indicates that both procedures perform equally well when the sample size is large. When n is small, Table 4.6 shows a slightly better performance of the CREML procedure than the CMM procedure. Similar results were also observed by Munther Al-Zaid (2001) [2]. Furthermore, CMM sometimes can produce non-positive definite estimates for the variance components if data are taken from inappropriate x. As a result, the estimates will be much worse than that of the CREML procedure. In Table 4.7 where $\sigma_\mu = 0.6$ and no data were taken from $-1 < x < 1$, ~ 150 out of 3000 simulations had the above

		$\hat{\mu}$	$\hat{\lambda}$	$\hat{\sigma}_\mu$	$\hat{\sigma}_\lambda$	$\hat{\sigma}$
CMM	bias	0.0158	0.0394	0.0103	0.0281	0.0126
	sd	0.0715	0.4809	0.0795	0.6271	0.6231
	MSE	0.0054	0.2328	0.0064	0.3940	0.3884
CREML	bias	-0.0050	-0.0167	-0.0333	-0.1806	-0.0796
	sd	0.0641	0.4554	0.0496	0.5013	0.4086
	MSE	0.0041	0.2077	0.0036	0.2839	0.1733

Table 4.6: Comparison of the estimation procedures in Case 2 where $x = (-5, -1, -0.2, 0.2, 1, 5)$ and $m = 10$, $n = 6$

		$\hat{\mu}$	$\hat{\lambda}$	$\hat{\sigma}_\mu$	$\hat{\sigma}_\lambda$	$\hat{\sigma}$
CMM	bias	0.245	-0.146	0.762	0.122	-0.024
	sd	0.146	2.974	2.796	0.797	0.449
	MSE	0.082	8.868	8.400	0.650	0.202
CREML	bias	0.224	-0.062	-0.205	-0.060	-0.100
	sd	0.140	0.657	0.101	0.713	0.420
	MSE	0.070	0.436	0.052	0.512	0.186

Table 4.7: Comparison of the estimation procedures in Case 3 where $x = (-5, -3, -1, 1, 3, 5)$ and $m = 10$, $n = 6$

issue. This is because when n is small, $\hat{\lambda}_i$ can be 0 and $\hat{\mu}_i$ can be very large, which make T_i in the CMM procedure non-invertible.

4.6 Pollen tube data study

The real data of ROP1 intensities $R(x)$ were collected from 12 pollen tubes of Arabidopsis. For each pollen tube, data were taken from one of its oblique planes at positions every 0.1205 μm from -10 μm to 10 μm. Therefore, $m = 12$ and $n = 173$. Each data point consists of two items, the position x and the associated observed intensity value $y(x)$. The ROP1 intensities in different pollen tubes are believed to have identical distributions. Therefore, it was necessary to first normalize the raw data. In this paper, quantile normalization method was utilized. Furthermore, the data were standardized into a fixed range by regression based method described below. Outliers contained in

the data sets were removed before estimation procedures were executed. The values for D, R_{tot} and α used in the study were 0.2, 30 and 1.2, respectively.

We first pooled the data sets together and fit the CNLS procedure described in section 4.2 and the constrained OC procedure described in section 4.3. We also fit these three procedures to each individual data set. We then fit the CMM procedure described in section 4.4 and the CREML procedure described in section 4.5. In a summary, the data analysis were executed in the following steps:

1. Use quantile normalization method to normalize the ROP1 intensities of different tubes, and remove possible outliers

2. Pool the data sets together and fit the pooled data nonparametrically to obtain $\hat{R}(x)$, and set the background noise to be the smallest value of $\hat{R}(x)$

3. Subtract the background noise from $\hat{R}(x)$ and all the data points

4. Standardize $\hat{R}(x)$ to range from 0 to 1. All the data points would be standardized accordingly.

5. Fit the CNLS procedure and constrained OC procedure to the pooled data sets and the individual data set respectively

6. Fit the CMM procedure and the CREML procedure to the data sets

Figure 4.1 and Figure 4.2 display the scatter plot of data of 12 tubes w/o normalization. Figure 4.3 displays the scatter plot of pooled data w/o outliers after normalization. Figure 4.4 displays the fitted curve $\hat{R}(x)$ based on pooled dataset after background noise deduction and further after standardization. Table 4.8 displays the estimates of k_{nf} and k_{pf} for the CNLS procedure, OC procedure with constraints and without constraints.

Table 4.9 displays the estimates of $(\mu, \lambda, \Sigma, \sigma)$ as well as that of (k_{nf}, k_{pf}) for the CMM procedure and the CREML procedure.

In Table 4.8, estimates of k_{nf} and k_{pf} based on CNLS from individual tube and pooled tubes are pretty close to each other. This is because that the sample size of each individual is sufficiently large ($n = 175$). This also indicates that the variation between tubes is not too large for the CNLS procedure to handle. On the other hand, estimates of k_{nf} and k_{pf} based on OC procedure vary among different tubes, especially when the constraints on the shape of $R(x)$ are not imposed. This is what we expect since the OC procedure depends more heavily on the data quality in comparison to the CNLS procedure. Nevertheless, when pooling the data together, both two procedures provide similar estimates which indicates that pooling the data can help recover the shape of $R(x)$. Especially for the OC procedure, pooling the data will greatly improve the local quality of the data so that the estimates are more accurate and closer to that of CNLS procedure. In Table 4.9, estimates of all parameters are close between the CMM procedure and the CREML procedure. This is also because the data size is large enough ($m = 12, n = 175$). The estimates of k_{nf} and k_{pf} are consistent among the three procedures. Moreover, there is a large positive correlation among μ and λ, which can be explained by the fact that the positive feedback process and negative feedback process in the first stage of tip growth process has an intrinsic connection since the strength of them both depend on the intensities of active ROP1 on the plasmic membrane. The standard deviation of the parameters μ and λ are smaller in the CREML procedure than in the CMM procedure, which implies the CREML procedure provides more accuracy.

	$\hat{k}_{nf}$			$\hat{k}_{pf}$		
	CNLS	OC	OC^3	CNLS	OC	OC^3
Tube 1	0.1866	0.1920	0.2328	0.2925	0.2975	0.3602
Tube 2	0.2278	0.5411	0.2963	0.3337	0.9001	0.4335
Tube 4	0.1814	0.1523	0.1532	0.2854	0.2399	0.2425
Tube 5	0.2205	0.1638	0.2573	0.3265	0.2526	0.3801
Tube 6	0.1788	0.1552	0.1278	0.2748	0.2526	0.2077
Tube 7	0.1892	0.2114	0.1622	0.2925	0.3263	0.2498
Tube 8	0.1917	0.1975	0.2085	0.2977	0.3022	0.3204
Tube 9	0.2121	0.1780	0.1969	0.3188	0.2719	0.2978
Tube10	0.1976	0.2019	0.2094	0.3053	0.3071	0.3218
Tube11	0.1939	0.1787	0.1804	0.3011	0.2692	0.2715
Tube14	0.1694	0.1748	0.2331	0.2766	0.2812	0.3702
Tube15	0.1809	0.1399	0.1474	0.2810	0.2230	0.2333
pooled	0.1930	0.1872	0.1940	0.2979	0.2867	0.2969

Table 4.8: Estiamtes of k_{nf} and k_{pf} of CNLS procedure, OC procedure without constraint and OC^3 procedure with both constraints. And $D = 0.2$, $R_{tot} = 30$ and $\alpha = 1.2$.

4.7 Summary

In this chapter, we have studied the parameter estimation problem for a particular example of tip growth of pollen tubes. We proposed a constrained nonlinear model with one estimation procedure, CNLS to fit single pollen tube data set and a constrained nonlinear mixed effect model with two estimation procedures, CMM and CREML to fit multiple pollen tubes data sets. We also proposed a Constrained OC approach by imposing two different types of constraints on the original OC approach. We evaluated the performance of all the estimation procedures via simulations and a pollen tube data analysis, and found out that in single pollen tube case, the CNLS approach outperforms the Constrained OC approach which outperforms the original OC approach without any constraint, especially when the data have a bad local quality. In multiple pollen tubes case, the CMM approach and CREML approach perform equally well when the sample size is sufficiently large, whereas CREML outperforms CMM when the sample size is small. We used a simple strategy to incorporate the constraint into the objective func-

tion before applying the simplex method to solve the constrained optimization problem in the estimation procedures, which works quite well. Other optimization methods such as dual method of Goldfarb and Idnani [37, 38] can be utilized. It deserves to mention that the simplex method after the modification we did can also tackle the constrained optimization problem with a nonlinear constraints, which could also be solved by methods such as Sequential Quadratic Programming (SQP).

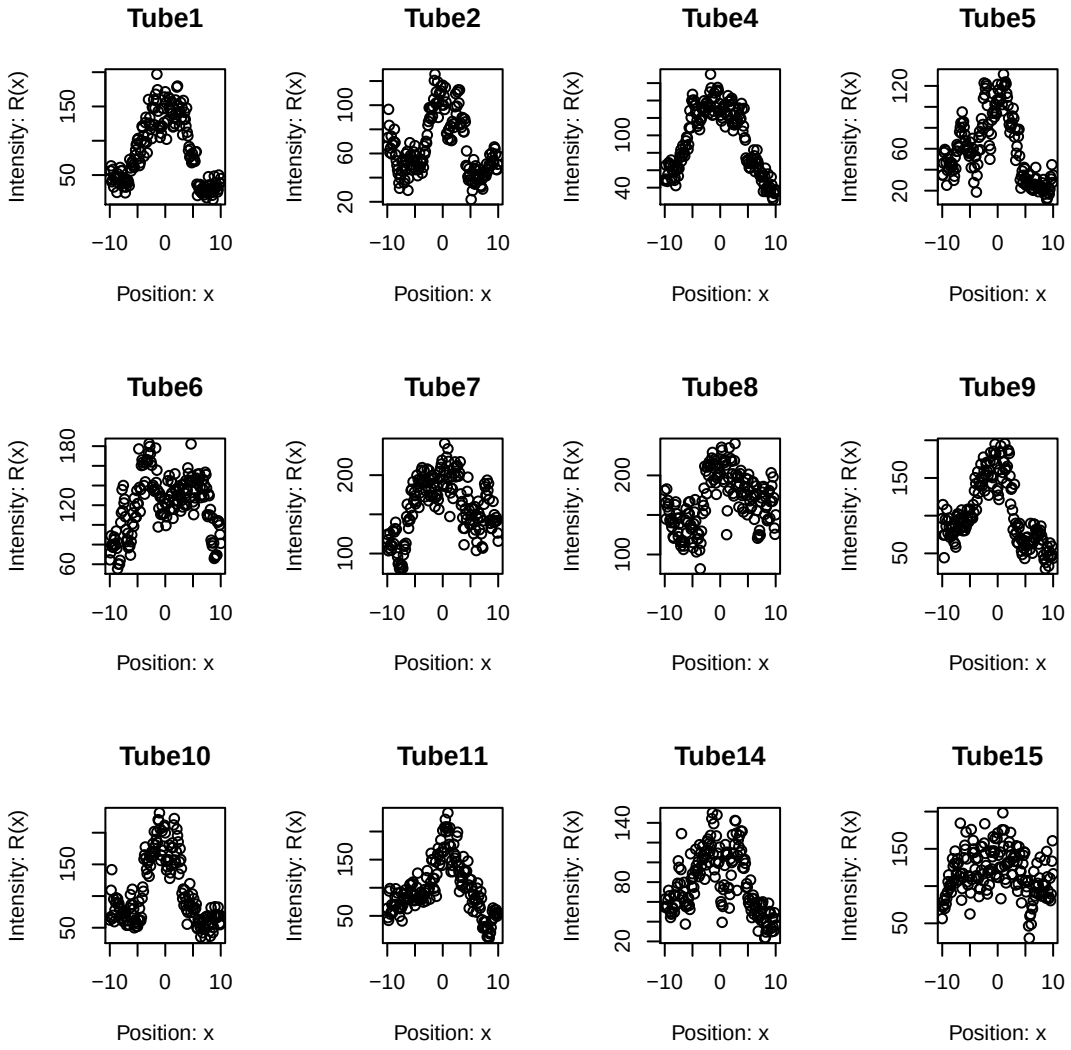

Figure 4.1: The figure is the scatterplot of the ROP1 intensities from 12 tubes of arabidopsis before normalization.

	CMM	CREML
$\hat{\mu}$	0.9708	0.9648
$\hat{\lambda}$	0.6477	0.6487
$\hat{\sigma}_\mu$	0.0789	0.0709
$\hat{\sigma}_\lambda$	0.0393	0.0258
$\hat{\rho}$	0.737	0.838
$\hat{\sigma}$	0.2064	0.2267
$\hat{k}_{nf}$	0.1942	0.1862
$\hat{k}_{pf}$	0.2987	0.2873

Table 4.9: Estimations of μ, λ, Σ and σ of CMM and CREML procedures with $D = 0.2$, $R_{tot} = 30$ and $\alpha = 1.2$.

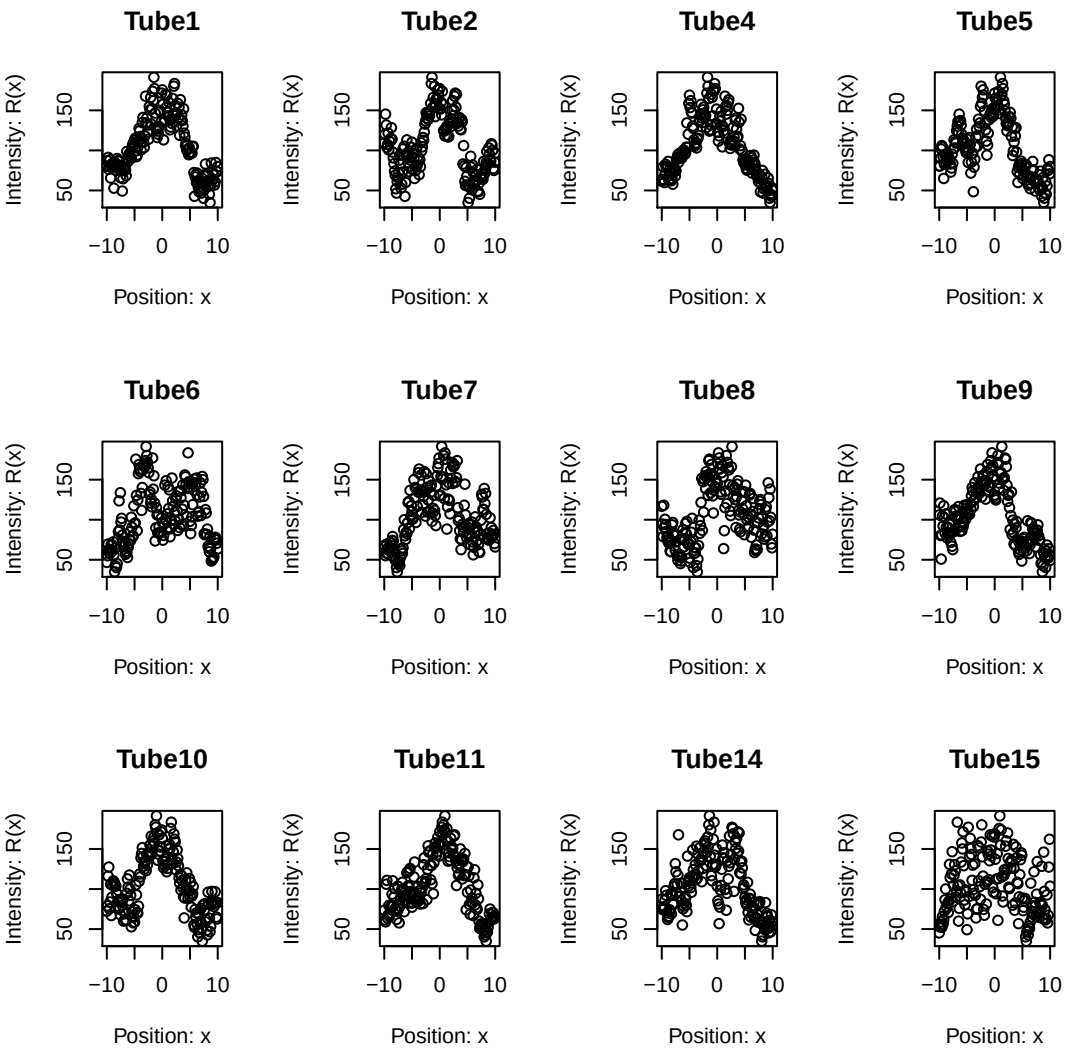

Figure 4.2: The figure is the scatterplot of the ROP1 intensities from 12 tubes of arabidopsis after normalization.

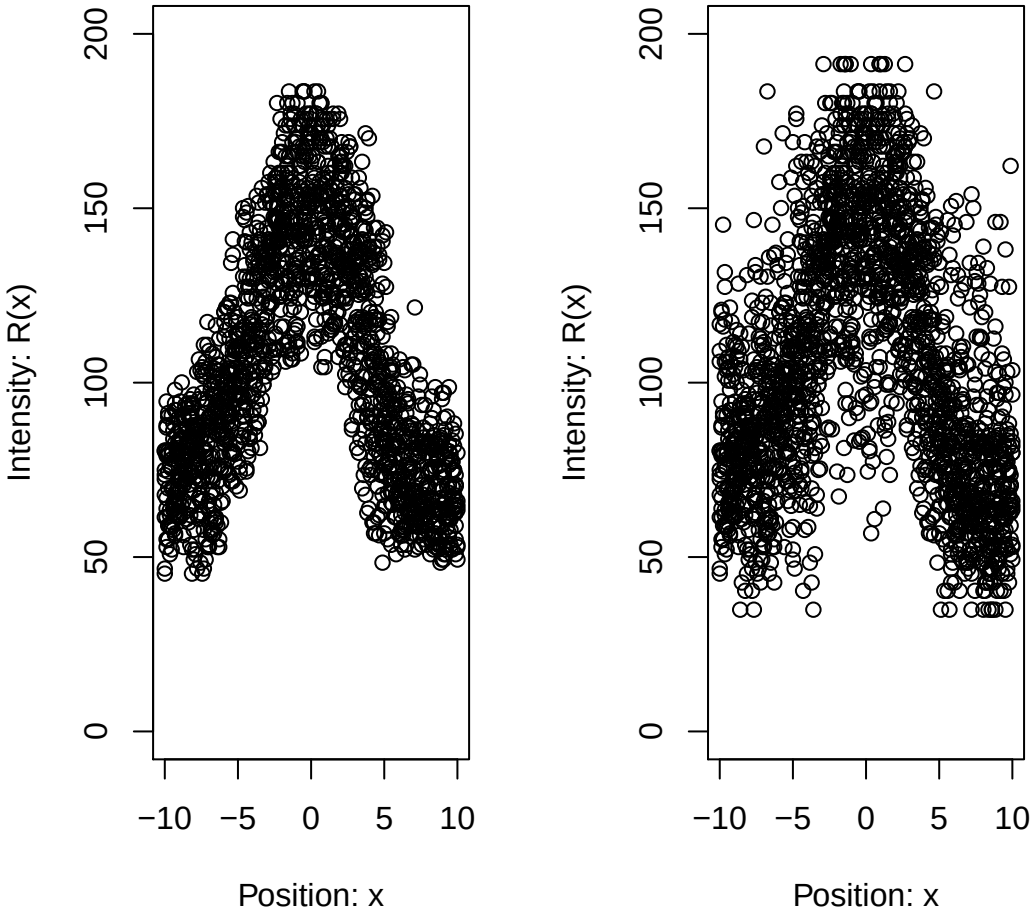

Figure 4.3: The two plots in the figure from left to right are the scatterplots of the pooled dataset of arabidopsis with and without removing outliers, respectively.

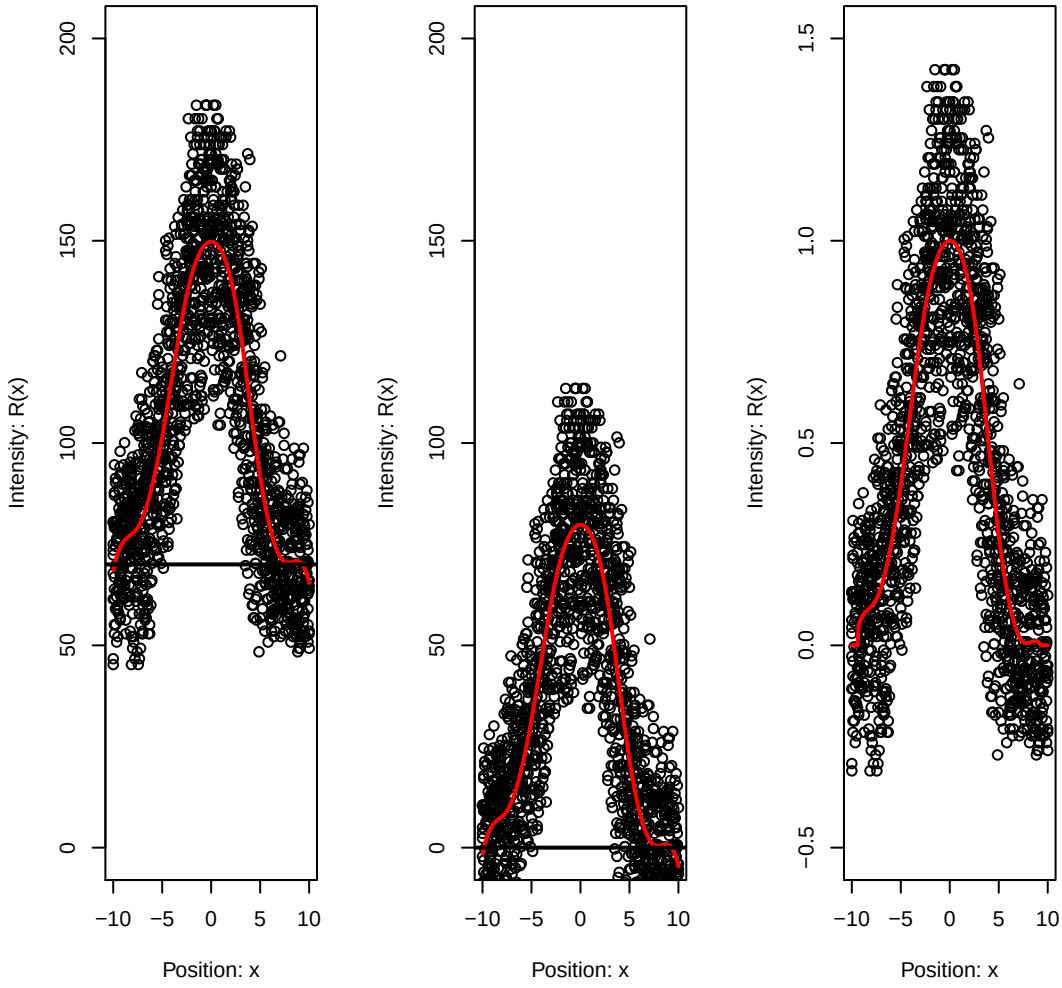

Figure 4.4: The three plots in the figure from left to right display $\hat{R}(x)$ that is before background noise deduction, after background noise deduction, and further after standardization, respectively. The black line in the leftmost plot denotes the background noise, which is 70.

Chapter 5

Parameter Estimation in PDE models

In chapter 4, we have studied the parameter estimation problem in the example of tip growth of pollen tube based on the observed ROP1 density at static time, $R(x)$, which is a solution to ODE (1.2)

$$\begin{cases} D\frac{d^2R}{dx^2} = k_{nf}R - k_{pf}R^{\alpha}(1 - \frac{\int_x Rdx}{R_{tot}}) & x \in [-L_0, L_0] \\ R(-L_0) = R(L_0) = 0 \end{cases}$$

Specifically, we proved the ODE (1.2) has a solution $R(x) = \lambda\sigma_0(\mu x)$ under the constraints of $\Lambda^*(k_{nf}, k_{pf}) < 0$, $k_{nf} > 0$ and $k_{pf} > 0$. As a result, the ODE model was transformed into a Constrained Nonlinear Model for single pollen tube and a Constrained Nonlinear Mixed Effects Model for multiple pollen tubes based on which k_{nf} and k_{pf} were estimated.

In recent years, with more advanced technology and more sophisticated equipments, the ROP density $R(x,t)$ to the PDE (1.1) can be measured over time t.

$$\begin{cases} \frac{\partial R(x,t)}{\partial t} = k_{pf} R(x,t)^{\alpha}(1 - \frac{\int_x R(x,t)dx}{R_{tot}}) - k_{nf} R(x,t) + D\frac{d^2 R(x,t)}{dx^2} \\ where \qquad \{x,t\} \in [-L_0, L_0] \times [0,\infty] \\ R(-L_0,t) = R(L_0,t) = 0 \\ R(x,0) = f(x) \end{cases} \tag{5.1}$$

As a result, one might want to consider the estimation problem based on the PDE (1.1) model rather than the ODE (1.2). Indeed, it deserves to study such a PDE model since with more observations, more accurate estimates for k_{nf} and k_{pf} can be obtained. In this chapter, we are going to focus on the PDE model and study the corresponding estimation problem.

5.1 Numerical solution to PDE (1.1)

Before devoting efforts to finding the solution to the PDE (1.1), one thing has to be made sure is that the PDE (1.1) has solutions, which is known as the identifiability problem introduced in section 2.1 of chapter 2. The identifiability problem in a general PDE was discussed by Avner Friedman [34]. Regarding to the PDE (1.1), we have established the following results.

1. for any k_{nf} and k_{pf}, the PDE (1.1) has a local solution at $[0, t^*]$, where t^* will depend on the values of k_{nf} and k_{pf}.

2. for some k_{nf} and k_{pf}, the PDE (1.1) may have a global solution at $[0, T]$, where T denotes the static time.

However, until now we are not able to identify

1. the feasible set of k_{nf} and k_{pf} analytically that can provide a global solution to the PDE (1.1) at $[0, T]$

2. the analytical form of the solution to the PDE (1.1)

Since there is no closed form solution available so far, a numerical solution to the PDE (1.1) for given values of k_{nf} and k_{pf} will be needed. In general, numerical methods for solving PDE were discussed in the book of Gordon C. Everstine [30]. As introduced in section 1.2.2 of chapter 1, the second-order PDE (1.3)

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = G$$

can be divided into three main types Parabolic, Elliptic, and Hyperbolic, depending on the sign of the discriminant $B^2 - AC$. For different types of PDE, Gordon C. Everstine suggested different numerical methods. By checking the discriminant, the PDE (1.1) belongs to the Parabolic type. Therefore according to the suggestion of Gordon C. Everstine, the appropriate numerical methods include the Explicit Finite Difference Method and the Crank-Nicolson Implicit Method. In the following, we are going to apply them to the PDE (1.1).

Explicit Finite Difference Method

Suppose the PDE (1.1) has a solution at $[0, T]$, let $0 = t_0 < t_1 < \cdots < t_{n_t-1} = T$ be a set of time instants on the interval $[0, T]$ with mesh size $\Delta t_j = t_{j+1} - t_j$. Let $-L_0 = x_0 < x_1 < \cdots < x_{n_x-1} = L_0$ be a set of positions on the interval $[-L_0, L_0]$ with mesh size $\Delta x_i = x_{i+1} - x_i$. Let $R_{i,j}$ be the numerical approximation to $R(x_i, t_j)$. $\frac{\partial R(x,t)}{\partial t}$ is approximated with the forward finite difference

$$\frac{\partial R(x,t)}{\partial t} = \frac{R_{i,j+1} - R_{i,j}}{\Delta t_j}$$

and $\frac{d^2R}{dx^2}$ is approximated with the central finite difference

$$\frac{d^2R}{dx^2} = \frac{R_{i+1,j} - 2R_{i,j} + R_{i-1,j}}{(\Delta x_i)^2}$$

and $\int_{-L_0}^{L_0} R(x,t)dx$ is approximated with the numerical integration

$$\int_{-L_0}^{L_0} R(x,t)dx = \sum_{i=1}^{n_x} R_{i,j}\Delta x_i$$

The finite difference approximation to the PDE (1.1) is then

$$\frac{R_{i,j+1} - R_{i,j}}{\Delta t_j} = k_{pf}R_{i,j}^{\alpha}(1 - \frac{\sum_i^{n_x} R_{i,j}\Delta x_i}{R_{tot}}) - k_{nf}R_{i,j} + D\frac{R_{i+1,j} - 2R_{i,j} + R_{i-1,j}}{(\Delta x_i)^2}$$

Therefore, the PDE (1.1) becomes

$$R_{i,j+1} = R_{i,j} + \Delta t_j \left(k_{pf}R_{i,j}^{\alpha}(1 - \frac{\sum_i R_{i,j}\Delta x_i}{R_{tot}}) - k_{nf}R_{i,j} + D\frac{R_{i+1,j} - 2R_{i,j} + R_{i-1,j}}{(\Delta x_i)^2} \right)$$

This method is referred to as an explicit method since based on the above formula one unknown value $R_{i,j+1}$ at later time can be found directly in terms of the known values $\{R_{i,j}\}_{i=1;j=1}^{n_x;n_t}$ at earlier time.

Crank-Nicolson Implicit Method

The key for the Crank-Nicolson Implicit method is to write the finite difference equation at a mid-level in time $t_{j+\frac{1}{2}}$. The finite difference x derivative at time $t_{j+\frac{1}{2}}$ is computed as the average of the central difference x derivatives at time t_j and t_{j+1}.

$$\frac{d^2R}{dx^2} = \frac{1}{2}(\frac{R_{i+1,j} - 2R_{i,j} + R_{i-1,j}}{(\Delta x_i)^2} + \frac{R_{i+1,j+1} - 2R_{i,j+1} + R_{i-1,j+1}}{(\Delta x_i)^2})$$

And the finite difference t derivative at position x_i is computed in the same way as in the Explicit Finite Difference Method.

$$\frac{\partial R(x,t)}{\partial t} = \frac{R_{i,j+1} - R_{i,j}}{\Delta t_j}$$

As a result, the PDE (1.1) is numerically approximated by

$$\begin{aligned}\frac{R_{i,j+1}-R_{i,j}}{\Delta t_j} =& k_{pf}R_{i,j}^{\alpha}(1-\frac{\sum_{i=1}^{n_x}R_{i,j}\Delta x_i}{R_{tot}})-k_{nf}R_{i,j}\\ &+\frac{D}{2}\left(\frac{R_{i+1,j}-2R_{i,j}+R_{i-1,j}}{(\Delta x_i)^2}+\frac{R_{i+1,j+1}-2R_{i,j+1}+R_{i-1,j+1}}{(\Delta x_i)^2}\right)\end{aligned}$$

After putting all $j+1$ terms in the above equation on the left hand side, we get

$$\begin{aligned}&\frac{R_{i,j+1}}{\Delta t_j}-\frac{D}{2}\left(\frac{R_{i+1,j+1}-2R_{i,j+1}+R_{i-1,j+1}}{(\Delta x_i)^2}\right)\\ &=\frac{R_{i,j}}{\Delta t_j}+k_{pf}R_{i,j}^{\alpha}(1-\frac{\sum_{i=1}^{n_x}R_{i,j}\Delta x_i}{R_{tot}})-k_{nf}R_{i,j}+\frac{D}{2}\left(\frac{R_{i+1,j}-2R_{i,j}+R_{i-1,j}}{(\Delta x_i)^2}\right)\end{aligned}$$

This formula is called the Crank-Nicolson equation.

The Crank-Nicolson scheme starts at the bottom of t_0 with $j=0$ and move up to t_1 with $j=1$. Then the right side of the Crank-Nicolson equation is known, whereas the left side of the Crank-Nicolson equation is unknown. To solve for $R(x,t_{j+1})$, one write the Crank-Nicolson equation for all the positions $i=1,2,\cdots,n_x$ and therefore obtain n_x equations in a row. Then the implicit algorithm can be used to solve these equations simultaneously. Notice that the coefficient matrix of the Crank-Nicolson equation is a tridiagonal matrix and remains the same step by step, it is suggested to compute and save the LU factors of the coefficient matrix so that we don't have to repeat it at each new time step, and thereafter speed up the computation. This speedup is very significant when the PDE is of high dimension in which the coefficient matrix is not tridiagonal any more.

In this dissertation, we have implemented the Explicit Finite Difference Method in R software to solve for $R(x,t)$ and confirmed that for some k_{nf} and k_{pf}, such as $k_{nf}=0.1$ and $k_{pf}=0.1125$, the PDE (1.1) has a global solution of $R(x,t)$, whereas for many other k_{nf} and k_{pf}, $R(x,t)$ only exists at the first couple of numeration steps.

Further, based on the numerical solution, we have identified the feasible region shown in figure 5.1 in which the PDE (1.1) has a global solution at $[0, T]$.

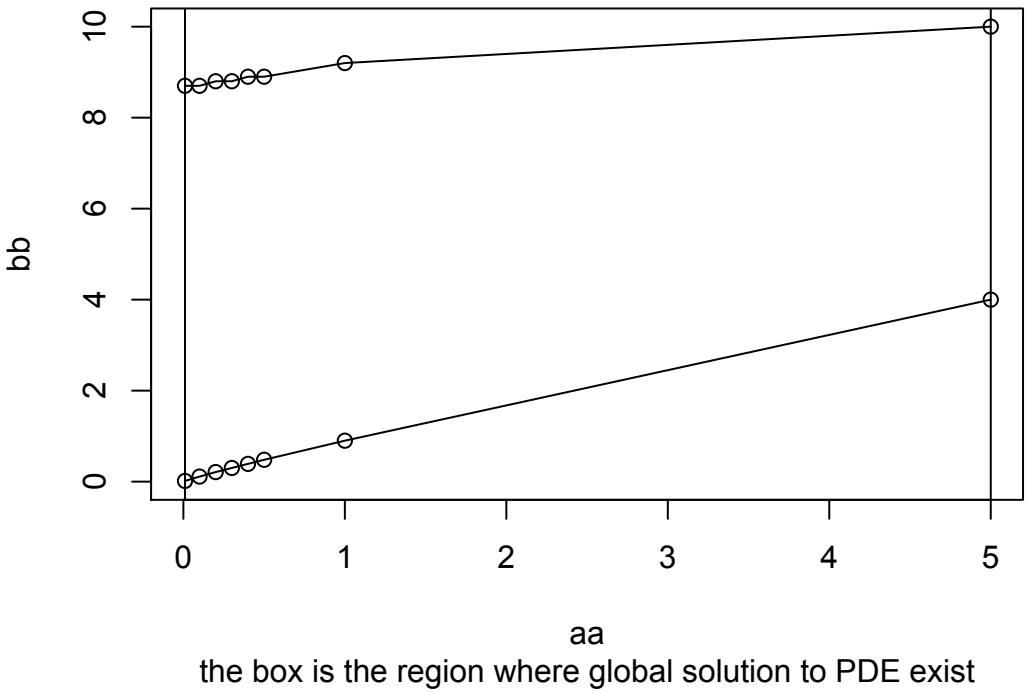

Figure 5.1: The figure whose x-axis represents k_{nf} and y-axis represents k_{pf} shows a polygon surrounded by the lines in which the PDE (1.1) has a global solution at $[0, T]$.

5.2 Parameter Estimation in PDE model

Suppose the observation of ROP1 density $R(x, t)$ at position $X_i \in [-L_0, L_0]$ and time $T_j \in [0, T]$ for single pollen tube can be written as

$$Y_{ij} = R(X_i, T_j; k_{nf}, k_{pf}) + \epsilon_{ij} \qquad j = 1, 2, \cdots, n. \tag{5.2}$$

where $R(x, t)$ is a solution to the PDE (1.1), and ϵ_{ij} is independently identically distributed from a distribution F with common mean 0 and variance σ^2. To estimate the

PDE model for $\hat{k}_{nf}$ and $\hat{k}_{pf}$, a natural approach can be proposed based on the idea of least square minimization which works as follows.

1. for given values of k_{nf} and k_{pf}, use the methods described in section 5.1 to numerically solve the PDE (1.1) for $R(x,t)$ at $[0,T]$. Notice that such a solution may not exist.

2. define an estimation criterion $Q_{n_x,n_t}(k_{nf},k_{pf})$ to be the sum of squared differences between $R(x,t)$ and $y(x,t)$ at all positions and time instants. If $R(x,t)$ does not exist, then set the sum of squared differences to be $+\infty$

$$Q_{n_x,n_t} = \begin{cases} \sum_{i=1;j=1}^{n_x;n_t}(Y_{ij} - R(X_i,T_j;k_{nf},k_{pf}))^2 & \text{if R(x,t) exists at [0,T]} \\ +\infty & \text{if R(x,t) does not exist at [0,T]} \end{cases}$$

3. minimize the estimation criterion Q_{n_x,n_t} to obtain $\hat{k}_{nf}$ and $\hat{k}_{pf}$

The estimation approach involves a two dimensional minimization problem which generally can be solved by many optimization methods such as the simplex method of Nelder and Mead [58] described in section 4.

If Q_{n_x,n_t} is a convex function of k_{nf} and k_{pf} over the feasible region, the minimization process will eventually reach to the true solution. Otherwise, the minimization process will probably converge to a wrong place or won't converge at all. Therefore, it's helpful to check the shape of Q_{n_x,n_t} before conducting the minimization process. In our simulation study, we first set $k_{nf} = 0.3$ and $k_{pf} = 1$ and solved the PDE for $R(x,t)$. The data of $R(x,t)$ were then generated by adding an $i.i.d.$ $N(0,\sigma^2 = 1)$ noise. Based on these simulated data, we first fixed $k_{nf} = 0.3$ and changed k_{pf} so that Q_{n_x,n_t} is a function of k_{pf}. Such a function is shown in figure 5.2. Likewise, we fixed $k_{pf} = 1$ and changed k_{nf} so that Q_{n_x,n_t} is a function of k_{nf}. Such a function is shown in figure 5.3. From these two figures, it's clear to see that when fixing k_{nf} or k_{pf}, Q_{n_x,n_t} is a convex

function of the other variable, which indicates that the minimization over one variable will have a good convergence behavior. Indeed, this was observed in the following simulation results. However, it was also observed that the minimization over both two variables simultaneously would converge to a wrong place with an inappropriate initial value.

1. fixing $k_{nf} = 0.3$ and treating Q_{n_x,n_t} as a function of k_{pf}, the estimation procedure will converge to 1 starting at 5.

2. fixing $k_{pf} = 1$ and treating Q_{n_x,n_t} as a function of k_{nf}, the estimation procedure will converge to 0.3 starting at 1.

3. treating the target function as a function of both k_{nf} and k_{pf}, the estimation procedure will converge to $(0.3, 1)$ starting at $(0.2, 0.2)$ or $(2, 6)$ or $(3, 5)$. On the other hand, the procedure will converge to wrong places if starting at $(5, 6)$ or $(4, 7)$ or $(3, 7)$ or $(2, 7)$.

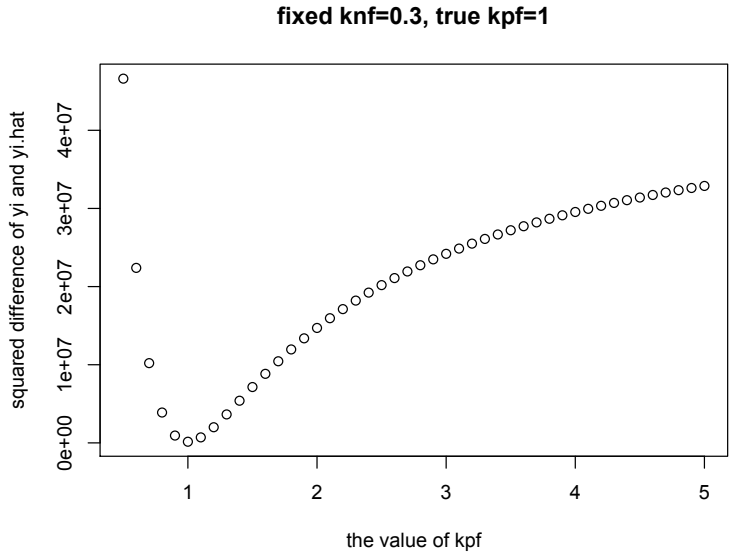

Figure 5.2: The figure shows that Q_{n_x,n_t} is a convex function of k_{pf} in (0,5).

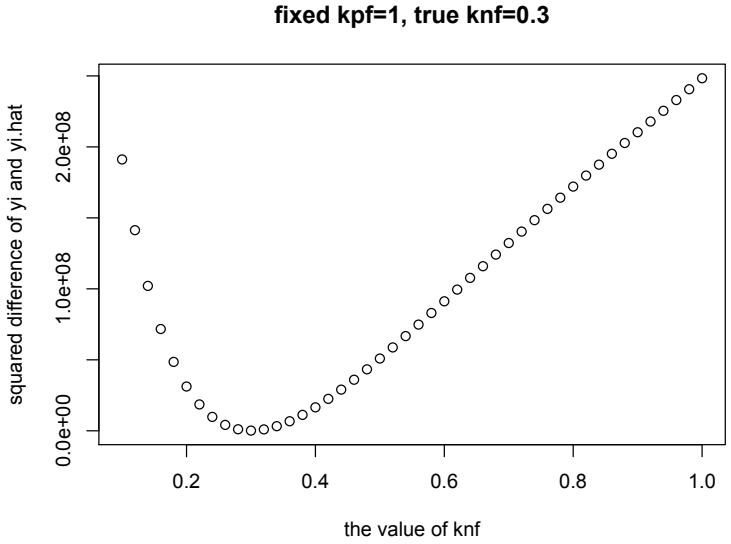

Figure 5.3: The figure shows that Q_{n_x,n_t} is a convex function of k_{nf} in (0,1).

In order to overcome the challenge of convergency due to the bad choice of initial value, we considered to conduct the minimization process with multiple initial values. And accordingly, we proposed the following minimization strategy

1. partition the feasible region into multiple lattices and for each lattice, randomly choose a point in it to be an initial value

2. conduct the minimization process with all the initial values chosen in step 1

3. identify the initial value that presents the smallest value of Q_{n_x,n_t}. The corresponding minimization solutions are the estimates we look for.

In our simulation, it was observed that this strategy successfully provided the right solution.

5.3 Future Work

So far, we specified the estimation problem for PDE models, and accordingly proposed an estimation procedure in the sense of least square minimization. We further found out that the minimization problem in the proposed estimation procedure often encounters a challenge of convergency. Without good initial values, the minimization process will either converge to a wrong place or won't converge at all. To overcome this challenge, we developed a new way to conduct the minimization process by starting it from multiple initial values simultaneously. The simulation studies showed that the new optimization strategy works well. However, there are many work left in order to solve the estimation problem completely.

First of all, more mathematical efforts should be devoted to finding the analytical solution to the PDE (1.1), or finding an accurate approximated solution, if the analytical solution does not exist. This is crucial since our estimation criterion greatly depends on the solution so that if the solution is not accurate enough, the criterion will lead to bad estimations for k_{nf} and k_{pf}.

Secondly, from our empirical experience, the minimization process for PDE models is more often to encounter the problem of convergency in comparison with the ODE model case. This problem becomes severe when the parameter dimension is high. The new optimization strategy we developed can solve this problem in principle. However, it is computationally intensive since the optimization process must be repeatedly conducted with multiple initial values. Therefore, a computationally easier optimization strategy should be discovered.

Thirdly, statistical efforts can be put to comparing the numerical methods with other nonparametric methods such as TS or OC methods in terms of their estimation

performances. This work can help identify the merits and shortcomings for each estimation methods, so that we can determine whether it deserves to focus on our numerical method or it is better to use the nonparametric methods. Besides, the large sample properties should also be established for all the methods.

Last, when the ROP1 density can be observed on multiple pollen tubes and the tube to tube variation is nonnegligible, then a PDE based mixed effects model should be formulated to incorporate such a variation, and statistical efforts can be devoted to estimating the mixed effects model.

Bibliography

[1] A. Wittinghofer A. Berken, C. Thomas. A new family of RhoGEFs activates the Rop molecular switch in plants. *Nature*, 436:1176–1180, 2005.

[2] Munther Al-Zaid and Shie-Shien Yang. An approximate em algorithm for nonlinear mixed effects models. *Biometrical Journal*, 43(7):881–893, 2001.

[3] Hen-Ming Wu Alice Y. Cheung. Structural and signaling networks for the polar cell growth machinery in pollen tubes. *nnu Rev Plant Biol.*, 59:547–572, 2008.

[4] Steven J. Altschuler, Sigurd B. Angenent, Yanqin Wang, and Lani F. Wu. On the spontaneous emergence of cell polarity. *Nature*, 454(7206):886–889, August 2008.

[5] Y. Amemiya. What Should be Done When an Estimated Between Group Covariance Matrix is not Nonnegative Definite? *The American Statistician*, 39:112–117, 1985.

[6] Uri M. Ascher, Robert M. M. Mattheij, and Robert D. Russell. *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*. Society for Industrial and Applied Mathematics, 1995.

[7] Kost B., Lemichez E., Spielhofer P., Hong Y., and Carpenter C. Tolias K. Rac Homologues and Compartmentalized Phosphatidylinositol 4, 5-Bisphosphate Act in a Common Pathway to Regulate Polar Pollen Tube Growth. *J Cell Biol.*, 145:317–330, 1999.

[8] G. Bader and U. Ascher. A new basis implementation for a mixed order boundary value ode solver. *SIAM J. Sci. Stat. Comput.*, 8(4):483–500, July 1987.

[9] Yonathan Bard. *Nonlinear parameter estimation*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1974.

[10] R. Bellman and K.J. Astrom. On structural identifiability. *Mathematical Biosciences*, 7(3â4):329 – 339, 1970.

[11] M. Benson. Parameter fitting in dynamic models. *Ecological Modelling*, 6(2):97 – 115, 1979.

[12] H.G. Bock. Recent advances in parameteridentification techniques for o.d.e. In Peter Deuflhard and Ernst Hairer, editors, *Numerical Treatment of Inverse Problems in Differential and Integral Equations*, volume 2 of *Progress in Scientific Computing*, pages 95–121. BirkhÃuser Boston, 1983.

[13] Nicolas J. B. Brunel. Parameter estimation of ODE's via nonparametric estimators. December 2008.

[14] Nicolas J.B. Brunel, Quentin Clairon, and DAlche-Buc Florence. Parametric Estimation of Ordinary Differential Equations with Orthogonality Conditions. *Journal of American Statistics Association*, page (to appear), 2013.

[15] Jost C and Ellner SP. Testing for predator dependence in predator-prey dynamics: a non-parametric approach. *Proc Biol Sci*, 267(1453):1611–1620, 2000.

[16] B. Cai and D. B. Dunson. Bayesian multivariate isotonic regression splines: applications to carcino- genicity studies. *Journal of the American Statistical Association*, 102:1158–1171, 2007.

[17] Otger Campàs and L. Mahadevan. Shape and Dynamics of Tip-Growing Cells. *Curr Biol*, 19(24):2102–2107, December 2009.

[18] Otger Campas1 and L. Mahadevan. Shape and Dynamics of Tip-Growing Cells. *Current Biology*, 19:2102–2107, 2009.

[19] J.R. Cash and F. Mazzia. A new mesh selection algorithm, based on conditioning, for two-point boundary value codes. *Journal of Computational and Applied Mathematics*, 184(2):362 – 381, 2005.

[20] Thierry Cazenave and Alain Haraux. *An introduction to semilinear evolution equations*, volume 13 of *Oxford Lecture Series in Mathematics and its Applications*. The Clarendon Press Oxford University Press, New York, 1998.

[21] Michael J. Chappell and Keith R. Godfrey. Structural identifiability of the parameters of a nonlinear batch reactor model. *Mathematical Biosciences*, 108(2):241–251, 1992.

[22] Jianwei Chen and Hulin Wu. Estimation of time-varying parameters in deterministic dynamic models. *Statistica Sinica*, 18:987–1006, 2008.

[23] Ori Davidov and Sophia Rosen. Constrained inference in mixed-effects models for longitudinal data with application to hearing loss. *Biostatistics*, 12:327–340, 2011.

[24] L. Denis-Vidal and G. Joly-Blanchard. An easy to check criterion for (un)indentifiability of uncontrolled systems and its applications. *Automatic Control, IEEE Transactions on*, 45(4):768–771, Apr 2000.

[25] Jacques Dumais, Sidney L. Shaw, Charles R. Steele, Sharon R. Long, and Peter M. Ray. An anisotropic-viscoplastic model of plant cell morphogenesis by tip growth. *The International journal of developmental biology*, 50(2-3):209–222, 2006.

[26] Lloyd J. Edwards, Paul W. Stewart, Keith E. Muller, and Ronald W. Helms. Linear Equality Constraints in the General Linear Mixed Model. *Biometrics*, 57:1185–1190, 2001.

[27] Y. Gu et. al. A Rho family GTPase controls actin dynamics and tip growth via two counteracting downstream pathways in pollen tubes. *J Cell Biol.*, 169:127–138, 2005.

[28] Sylvester T. McKenna et.al. Exocytosis Precedes and Predicts the Increase in Growth in Oscillating Pollen Tubes. *The Plant Cell*, 21:3026–3040, 2009.

[29] Gordon C Everstine. Numerical solution of.

[30] Gordon C. Everstine. *Numerical Solution of Partial Dif- ferential Equations*. 2010.

[31] H. B. Fang, G. L. Tian, X. Xiong, and M. Tan. A multivariate random-effects model with restricted parameters: application to assessing radiation therapy for brain tumors. *Statistics in Medicine*, 25:1948–1959, 2006.

[32] Reinhard Fischer, Nadine Zekert, and Norio Takeshita. Polarized growth in fungi - interplay between the cytoskeleton, positional markers and membrane domains. *Molecular Microbiology*, 68(4):813–826, May 2008.

[33] Richard FitzHugh. Impulses and Physiological States in Theoretical Models of Nerve Membrane. *Biophys J.*, 1:445–466, 1961.

[34] Avner Friedman. *Partial differential equations of parabolic type*. R.E. Krieger Pub. Co., 1983.

[35] A. Gelman, Carlin J. B., Stern H. S., and Rubin D. B. Bayesian Data Analysis. 2005.

[36] B. Gidas, Wei M. Ni, and L. Nirenberg. Symmetry and related properties via the maximum principle. *Comm. Math. Phys.*, 68(3):209–243, 1979.

[37] D. Goldfarb and A. Idnani. Dual and primal-dual methods for solving strictly convex quadratic programs. In J.P. Hennart, editor, *Numerical Analysis*, volume 909 of *Lecture Notes in Mathematics*, pages 226–239. Springer Berlin Heidelberg, 1982.

[38] D. Goldfarb and A. Idnani. A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, 27(1):1–33, 1983.

[39] Li H, Lin Y, Heath RM, Zhu MX, and Yang Z. Control of pollen tube tip growth by a Rop GTPase-dependent pathway that leads to tip-localized calcium influx. *Plant Cell*, 11:1731–1742, 1999.

[40] Xue H., Miao H., and Wu H. *Sieve Estimation of Constant and Time-Varying Coefficients in Nonlinear Ordinary Differential Equation Models by Considering Both Numerical Error and Measurement Error*, volume 38. 2010.

[41] Hoff JM Helland IS and Anderbrant O. Attraction of bark beetles (Coleoptera: Scolytidae) to a pheromone trap : Experiment and mathematical models. *J Chem Ecol.*, 10:723–752, 1984.

[42] C.R. Henderson, Oscar Kempthorne, S. R. Searle, and C. M. von Krosigk. The Estimation of Environmental and Genetic Trends from Records Subject to Culling. *Biometrics*, 15:192–218, 1959.

[43] E. E. Holmes, M. A. Lewis, J. E. Banks, and R. R. Veit. Partial differential equations in ecology: Spatial interactions and population dynamics. *Ecology*, 75(1):pp. 17–29, 1994.

[44] L. Wang J. Cao and J. Xu. Parameter estimation in ordinary differential equations. *Biometrics*, 67:1305–1313, 2011.

[45] Robert I. Jennrich. Asymptotic Properties of Non-Linear Least Squares Estimators. *The Annals of Mathematical Statistics*, 40(2):633–643, April 1969.

[46] Hwang JU, Vernoud V, Szumlanski A, Nielsen E, and Yang Z. A tip-localized RhoGAP controls cell polarity by globally inhibiting Rho GTPase at the cell apex. *Curr Biol.*, 18:1907–1916, 2008.

[47] Grant M Kroeger JH, Geitmann A. Model for calcium dependent oscillatory growth in pollen tubes. . *Journal of Theoretical Biology*, 253:363–374, 2008.

[48] YJ. Lee, A. Szumlanski, E. Nielsen, and Z. Yang. Rho-GTPase-dependent filamentous actin dynamics coordinate vesicle targeting and exocytosis during tip growth. *Cell Biol.*, 181:1155–1168, 2008.

[49] M. J. Lindstrom and D. M. Bates. Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics*, 46:673–687, 1990.

[50] Mary J. Lindstrom and Douglas M. Bates. Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data. *Journal of the American Statistical Association*, 83:1014–1022, 1988.

[51] P. L. Lions. On the existence of positive solutions of semilinear elliptic equations. *SIAM Review*, 24(4):441–467, 1982.

[52] Lennart Ljung and Torkel Glad. On global identifiability for arbitrary model parametrizations. *Automatica*, 30(2):265 – 276, 1994.

[53] Van Vactor D. Lowery LA. The trip of the tip: understanding the growth cone machinery. *Nat Rev Mol Cell Biol.*, 10:332–43, 2009.

[54] J. C. Lu and W. Q. Meeker. Using Degradation Measures to Estimate a Time-to-Failure Distribution. *Technometrics*, 35:161–174, 1993.

[55] Madar, Janos, Abonyi, Janos, Roubos, Hans, Szeifert, and Ferenc. Incorporating prior knowledge in a cubic spline approximationapplication to the identification of reaction kinetic models. *Industrial Engineering Chemistry Research*, 42(17):4043–4049, 2003.

[56] Hongyu Miao, Xiaohua Xia, Alan S. Perelson, and Hulin Wu. On identifiability of nonlinear ode models and applications in viral dynamics. *SIAM Rev.*, 53(1):3–39, February 2011.

[57] J. Nagumo, S. Arimoto, and S. Yoshizawa. An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, 50(10):2061–2070, Oct 1962.

[58] J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, January 1965.

[59] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications (Universitext)*. Springer, 6th edition, January 2014.

[60] A. Okubo, P. K. Maini, M. H. Williamson, and J. D. Murray. On the spatial spread of the grey squirrel in britain. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 238(1291):pp. 113–125, 1989.

[61] M.R. Osborne. Cyclic reduction, dichotomy, and the estimation of differential equations. *Journal of Computational and Applied Mathematics*, 86(1):271 – 286, 1997. Dedicated to William B. Gragg on the ocassion of his 60th Birthday.

[62] H.D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545, 1971.

[63] S. D. Peddada, D. B. Dunson, and X. Tan. Estimation of order-restricted means from correlated data. *Biometrika*, 93:703–715, 2005.

[64] S. R. Pilla, A. Qu, and C. Loader. Testing for order-restricted hypotheses in longitudinal data. *Journal of the Royal Statistical Society, Series B*, 68:437–455, 2006.

[65] H. Pohjanpalo. System identifiability based on the power series expansion of the solution. *Mathematical Biosciences*, 41(1â2):21 – 33, 1978.

[66] J. O. Ramsay, G. Hooker, D. Campbell, and J. Cao. Parameter estimation for differential equations: A generalized smoothing approach. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 2007.

[67] L. B. Sheiner and S. L. Beal. Evaluation of Methods for Estimating Population Pharmacokinetic Parameters, I. Michaelis-Menten Model: Routine Clinical Pharmacokinetic Data. *Journal of Pharmacokinetics and Biopharmaceutics*, 8:553–571, 1980.

[68] L. B. Sheiner and S. L. Beal. Evaluation of Methods for Estimating Population Pharmacokinetic Parameters, I. Michaelis-Menten Model: Routine Clinical Pharmacokinetic Data. *Journal of Pharmacokinetics and Biopharmaceutics*, 9:635–651, 1981.

[69] L. B. Sheiner and S. L. Beal. Evaluation of Methods for Estimating Population Pharmacokinetic Parameters, I. Michaelis-Menten Model: Routine Clinical Pharmacokinetic Data. *Journal of Pharmacokinetics and Biopharmaceutics*, 11:303–319, 1983.

[70] D. W. Shin, C. G. Park, and T. Park. Testing for ordered group effects with repeated measurements. *Biometrika*, 83:688–694, 1996.

[71] J. G. Skellam. Random dispersal in theoretical population. *Biometrika*, 38:196–218, 1951.

[72] Karline Soetaert. *rootSolve: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations*, 2009. R package 1.6.

[73] Karline Soetaert, Thomas Petzoldt, and R. Woodrow Setzer. Solving differential equations in r: Package desolve. *Journal of Statistical Software*, 33(9):1–25, 2 2010.

[74] E. Tunali and T. Tarn. New results for identifiability of nonlinear systems. *Automatic Control, IEEE Transactions on*, 32(2):146–154, Feb 1987.

[75] S. VAJDA, H. RABITZ, E. WALTER, and Y. LECOURTIER. Qualitative and quantitative identifiability analysis of nonlinear chemical kinetic models. *Chemical Engineering Communications*, 83(1):191–219, 1989.

[76] B. van Domselaar and P.W. Hemker. *Nonlinear parameter estimation in initial value problems*. MC. NW. Stichting Math. Centrum, 1977.

[77] J. Varah. A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific and Statistical Computing*, 3(1):28–46, 1982.

[78] E.O. Voit and J. Almeida. Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics*, 20(11):1670–1681, 2004.

[79] S. Walker. An EM Algorithm for Nonlinear Random Effects Models. *Biometrics*, 52:934–944, 1996.

[80] Eric Walter, Isabelle Braems, Luc Jaulin, and Michel Kieffer. Guaranteed numerical computation as an alternative to computer algebra for testing models for identifiability. In RenÃ© Alt, Andreas Frommer, R.Baker Kearfott, and Wolfram Luther, editors, *Numerical Software with Result Verification*, volume 2991 of *Lecture Notes in Computer Science*, pages 124–131. Springer Berlin Heidelberg, 2004.

[81] Eric Walter and Yves Lecourtier. Unidentifiable compartmental models: what to do? *Mathematical Biosciences*, 56(1-2):1 – 25, 1981.

[82] Jinde Wang. Asymptotics of Least-Squares Estimators for Constrained Nonlinear Regression. *The Annals of Statistics*, 24(3):1316–1326, 1996.

[83] Hulin Wu and Jin Ting Zhang. Local polynomial mixed-effects models for longitudinal data. *Journal of the American Statistical Association*, 97(459):883–897, 2002.

[84] X. Xia and C.H. Moog. Identifiability of nonlinear systems with application to hiv/aids models. *Automatic Control, IEEE Transactions on*, 48(2):330–336, Feb 2003.

[85] Zhenbiao Yang Yong Jik Lee. Tip growth: Signaling in the apical dome. *Curr Opin Plant Biol.*, 11:662–671, 2008.

[86] Michael R. Osborne Zhengfeng Li and Tania Prvan. Parameter estimation in ordinary differential equations. *IMA J Numer Anal*, 25:264–285, 2005.

[87] Shanggang Zhou and Douglas A. Wolfe. On derivative estimation in spline regression. *Statistica Sinica*, 10:93–108, 2000.

Chapter 6

Appendix

6.1 Proof of Proposition 18

Based on the classical theory of differential equation, there are potentially two solutions to the semilinear elliptical equation (4.1). As the function $u = 0$ is a solution, there is at most one non-null solution. Therefore, we have to show there exist a non-null solution σ_0 to (4.1), and further $\sigma_0 > 0$ on $[-c, c]$.

The existence of a positive solution to the semilinear elliptic equation $-\partial_x^2 u = f(u)$ is discussed thoroughly in [51]. In our case, $f(u) = u^\alpha - u$ is such that $f(0) = 0$, $f'(0) = -1 < 0$, and is superlinear[1] because $\frac{f(x)}{x} \longrightarrow \infty$. From theorem 1.1 in [51], we obtain the existence (and then uniqueness) of a positive function σ_0 in $C^2([-c, c])$ satisfying (4.1). Furthermore, when $c = \infty$, the existence and uniqueness of solution of (4.1) is also true by theorem 1.1.3 in [20], where $u_0 = (\frac{\alpha+1}{2})^{\frac{1}{\alpha-1}}$ guarantees that all the conditions are satisfied in that theorem.

[1] If λ_1 is the first eigenvalue of the operator $(-\Delta) = -\frac{\partial^2}{\partial u 2}$ on $\Omega = [-c, c]$, then f is superlinear if $\underset{+\infty}{\lim} f(u)u^{-1} > \lambda_1$. In this case, it is known that the eigenvalues are $\lambda_k = \left(\frac{\pi k}{2c}\right)^2$, $k = 1, \ldots$ with eigen functions $\varphi_k(u) = \sin\left(\frac{\pi k}{2c}(u + c)\right)$. In our case, we have $\frac{f(u)}{u} = u^{\alpha-1} - 1$.

From [36], we know that σ_0 is a positive and even function that increases at $[-c, 0]$ and decreases at $[0, c]$.

6.2 Proof of Proposition 19

We consider a family of transformed solutions $R_{\lambda,\mu}(x) = \lambda\sigma_0(\mu x)$ where $\lambda, \mu > 0$, and σ_0 is the solution to (4.2) for $c = \mu L_0$. We have $\partial_x R_{\lambda,\mu} = \lambda\mu\partial_x\sigma_0(\mu x)$ and $\partial_x^2 R_{\lambda,\mu} = \lambda\mu^2\partial_x^2\sigma_0(\mu x)$. From (4.2) , we get

$$\begin{aligned} -\partial_x^2 R_{\lambda,\mu} &= \lambda\mu^2\left(\sigma_0^\alpha(\mu x) - \sigma_0(\mu x)\right) \\ &= \lambda^{1-\alpha}\mu^2 R_{\lambda,\mu}^\alpha - \mu^2 R_{\lambda,\mu} \end{aligned}$$

Therefore, $R_{\lambda,\mu}$ satisfies the equation $-D\partial_x^2 R_{\lambda,\mu} + D\mu^2 R_{\lambda,\mu} = D\lambda^{1-\alpha}\mu^2 R_{\lambda,\mu}^\alpha$. Since $\mu, k_{nf}, k_{pf}, D > 0$, we can take $\mu = \sqrt{\frac{k_{nf}}{D}}$ and $\lambda = \left(\frac{k_{pf}}{k_{nf}}\lambda'\right)^{\frac{1}{1-\alpha}}$, then

$$-D\partial_x^2 R_{\lambda,\mu} + k_{nf} R_{\lambda,\mu} = \lambda' k_{pf} R_{\lambda,\mu}$$

Therefore, $R_{\lambda,\mu}$ is the unique positive solution to (4.2) .

6.3 Proof of Remark 23

From Proposition 18, we know that σ_0 is a positive and even function that increases at $[-c, 0]$ and decreases at $[0, c]$. Therfore, $R_{\lambda,\mu}(x) = \lambda\sigma_0(\mu x)$ has the same properties. Moreover from the proof of proposition 18, we have defined $f(x) = x^\alpha - x$, that satisfies $f(1) = 0$, and $f(x) > 0$ for $x > 1$. By theorem 3.1 of [51], $\max\sigma_0 > 1$. Therefore, $\max R_{\lambda,\mu}(x) = \lambda \times \max\sigma_0 > \lambda$.

6.4 Proof of Theorem 25

First, let's prove the following Lemma.

Lemma 28 *for any $\mu > 0$ and $\lambda > 0$, the function $h(\mu, \lambda) = \lambda^{\alpha-1} - \lambda^{\alpha}\frac{\|\sigma_0\|_1}{R_{tot}\mu} - \frac{1}{\alpha}\left(\frac{\alpha-1}{\alpha}\frac{R_{tot}\mu}{\|\sigma_0\|_1}\right)^{\alpha-1}$ is always non-positive.*

Proof. For any fixed $\mu > 0$, h is a function of λ. And

$$\frac{\partial h(\lambda)}{\partial \lambda} = (\alpha-1)\lambda^{\alpha-2} - \alpha\lambda^{\alpha-1}\frac{\|\sigma_0\|_1}{R_{tot}\mu}$$

Therefore, $\frac{\partial h(\lambda)}{\partial \lambda} = 0$ when $\lambda \doteq \lambda_c = \frac{\alpha-1}{\alpha}\frac{R_{tot}\mu}{\|\sigma_0\|_1}$. Then, we have

$$h(\lambda_c) = \lambda_c^{\alpha-1} - \lambda_c^{\alpha}\frac{\|\sigma_0\|_1}{R_{tot}\mu} - \frac{1}{\alpha}\left(\frac{\alpha-1}{\alpha}\frac{R_{tot}\mu}{\|\sigma_0\|_1}\right)^{\alpha-1} = 0$$

$$h(0) = -\frac{1}{\alpha}\left(\frac{\alpha-1}{\alpha}\frac{R_{tot}\mu}{\|\sigma_0\|_1}\right)^{\alpha-1} < 0$$

$$h(+\infty) = -\infty < 0$$

Notice that $h(\lambda)$ is a continuous function of λ, we conclude that $h(\lambda) \leq 0$ based on the above equations. Therefore, Lemma 28 holds. ■

In order to prove Theorem 25, we need to show the constraints in model 4.5 are equivalent to that in model 4.7. On one hand, if constraints in model 4.5 are satisfied, then $g(\lambda) = \frac{k_{nf}}{k_{pf}} - \lambda^{\alpha-1} + \frac{1}{R_{tot}}\lambda^{\alpha}\sqrt{\frac{D}{k_{nf}}}\|\sigma_0\|_1$ has at least one solution. As a result, $\frac{k_{nf}}{k_{pf}} = \lambda^{\alpha-1} - \frac{1}{R_{tot}}\lambda^{\alpha}\frac{1}{\mu}\|\sigma_0\|_1 > 0$, and $\mu > 0$ and $\lambda > 0$. Therfore, $\Lambda^*(\mu, \lambda) > 0$. On the other hand, if constraints in model 4.7 are satisfied, we can convert μ and λ to k_{nf} and k_{pf} by solving $k_{nf} = D\mu^2$ and $\frac{k_{nf}}{k_{pf}} = \lambda^{\alpha-1} - \frac{1}{R_{tot}}\lambda^{\alpha}\frac{1}{\mu}\|\sigma_0\|_1$. The solution of k_{nf} and k_{pf} is such that $k_{nf} > 0$, $k_{pf} > 0$ and by Lemma (1), $\Lambda(k_{nf}, k_{pf}, D, R_{tot}, \sigma_0) = \frac{k_{nf}}{k_{pf}} - \frac{1}{\alpha}(\frac{\alpha-1}{\alpha}\sqrt{\frac{k_{nf}}{D}}\frac{R_{tot}}{\|\sigma_0\|_1})^{\alpha-1} = \lambda^{\alpha-1} - \lambda^{\alpha}\frac{\|\sigma_0\|_1}{R_{tot}\mu} - \frac{1}{\alpha}\left(\frac{\alpha-1}{\alpha}\frac{R_{tot}\mu}{\|\sigma_0\|_1}\right)^{\alpha-1}$ is always non-positive.

6.5 Proof of Theorem 26

First, let's prove the following Lemmas.

Lemma 29 *Denote $A = (a_{ij})_{2\times 2}$ to be a 2×2 symmetrical matrix. Suppose all the four elements, a_{11}, a_{12}, a_{21}, a_{22}, are bounded in $[-B, B]$ for some $B > 0$, then $A \leq 2BI_2$.*

Proof. For any given vector $\boldsymbol{X} = (X_1, X_2)^T$,

$$\begin{aligned}
\boldsymbol{X}^T A \boldsymbol{X} &= a_{11}X_1^2 + 2a_{12}X_1X_2 + a_{22}X_2^2 \\
&\leq BX_1^2 + BX_2^2 + 2|a_{12}X_1X_2| \\
&\leq BX_1^2 + BX_2^2 + 2|a_{12}|(\frac{1}{2}(X_1^2 + X_2^2)) \\
&\leq 2BX_1^2 + 2BX_2^2 \\
&= 2B\boldsymbol{X}^T\boldsymbol{X}
\end{aligned}$$

Therefore, $A \leq 2BI_2$ and Lemma 29 holds. ■

Lemma 30 *Let $R(X;\boldsymbol{\theta}) = \lambda\sigma_0(\mu X)$ with $\boldsymbol{\theta} = (\mu, \lambda)^T$, then $K = E_X[\nabla_{\boldsymbol{\theta}} R(X;\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}} R(X;\boldsymbol{\theta}_0)^T]$ is a positive definite matrix, where $\nabla_{\boldsymbol{\theta}} R(X;\boldsymbol{\theta}_0)$ is the gradient vector of $R(X;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.*

Proof. Since λ, μ, and $\sigma_0(X)$ are positive, the diagonal elements in the matrix

$$K = E_X[\nabla_{\boldsymbol{\theta}} R(X;\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}} R(x;\boldsymbol{\theta}_0)^T] = \begin{pmatrix} E_X[\lambda_0^2 X^2 \sigma_0'(\mu_0 X)^2] & E_X[\lambda_0 X \sigma_0(\mu_0 X)\sigma_0'(\mu_0 X)] \\ E_X[\lambda_0 X \sigma_0(\mu_0 X)\sigma_0'(\mu_0 X)] & E_X[\sigma_0(\mu_0 X)^2] \end{pmatrix}$$

are both positive, which indicates the sum of the eigenvalues of K is greater than 0. Therefore, to prove the positive definiteness of K, we only need to show that the product of the two eigenvalues of K, i.e., the determinant of K is positive. That is,

$$E_X[\lambda_0^2 X^2 \sigma_0'(\mu_0 X)^2]E_X[\sigma_0(\mu_0 X)^2] > (E_X[\lambda_0 X \sigma_0(\mu_0 X)\sigma_0'(\mu_0 X)])^2$$

Indeed, since X is a random variable with distribution $F(x)$, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
(E_X[\lambda_0 X \sigma_0(\mu_0 X)\sigma_0'(\mu_0 X)])^2 &= (\int_{-\infty}^{+\infty} \lambda_0 X x \sigma_0(\mu_0 x)\sigma_0'(\mu_0 x) dF(x))^2 \\
&\leq \int_{-\infty}^{+\infty} (\lambda_0 x \sigma_0'(\mu_0 x))^2 dF(x) \int_{-\infty}^{+\infty} (\sigma_0(\mu_0 x))^2 dF(x) \\
&= E_X[\lambda_0^2 X^2 \sigma_0'(\mu_0 X)^2]E_X[\sigma_0(\mu_0 X)^2]
\end{aligned}$$

The equality holds if and only if the two functions $\lambda_0 x\sigma_0'(\mu_0 x)$ and $\sigma_0(\mu_0 x)$ are linearly dependent. However, since the Wronskian of such two functions,

$$W = \begin{vmatrix} \lambda_0 x\sigma_0'(\mu_0 x) & \sigma_0(\mu_0 x) \\ \lambda_0\mu_0 x\sigma_0'(\mu_0 x) + \lambda_0\sigma_0(\mu_0 x) & \mu_0\sigma_0'(\mu_0 x) \end{vmatrix}$$

$$= \lambda_0\mu_0 x\sigma_0'(\mu_0 x)^2 - \lambda_0\sigma_0(\mu_0 x)[\mu_0 x\sigma_0'(\mu_0 x) + \sigma_0(\mu_0 x)]$$

is not always zero over $x \in [-\infty, +\infty]$, the two functions are linear independent. Therefore, the equality won't be achieved, which completes our proof. ■

Following Theorem 1-6 of Wang (1996) [82], we can derive the asymptotic property of the proposed CNLS estimator. First, we need to verify that the following conditions are satisfied in our constrained model.

1. $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are iid with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = 1$

2. $R(X_i; \boldsymbol{\theta})$, $i = 1, \ldots, n$, are differentiable in $\boldsymbol{\theta}$ and there is a neighborhood W of $\boldsymbol{\theta}_0$ such that for $\boldsymbol{\theta}$ in W it holds that

$$R(X_i; \boldsymbol{\theta}) = R(X_i; \boldsymbol{\theta}_0) + \nabla_{\boldsymbol{\theta}} R(X_i; \boldsymbol{\theta}_0)^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + r_i(\boldsymbol{\theta})(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2)$$

   And $r_i(\theta)$ satisfies

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} r_i^2(\boldsymbol{\theta}) < \infty$$

   uniformly in W.

3.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} R(X_i; \boldsymbol{\theta}_0) \nabla_{\boldsymbol{\theta}} R(X_i; \boldsymbol{\theta}_0)^T = K$$

   exists and K is a positive definite matrix.

4. The inequality constraints g_i, $i = 1, \ldots, p$; and the equality constraints h_j, $j = p+1, \ldots, q$, are all continuously differentiable in W.

5. Let $I = \{i : g_i(\boldsymbol{\theta}_0) = 0, i = 1, \ldots, p\}$, then the vectors $\nabla_{\boldsymbol{\theta}} g_i(\boldsymbol{\theta}_0)$, $i \in I$; $\nabla_{\boldsymbol{\theta}} h_j(\boldsymbol{\theta}_0)$, $j = p+1, \ldots, q$ are linearly independent.

6. For each value of $\boldsymbol{\xi} \sim N(0, K)$, the minimization of

$$\begin{aligned} \min_{\boldsymbol{z}} \quad & \boldsymbol{z}' K \boldsymbol{z} - 2\boldsymbol{z}'\boldsymbol{\xi} \\ s.t. \quad & \nabla_{\boldsymbol{\theta}} g_j(\boldsymbol{\theta}_0)^T \boldsymbol{z} \leq 0 \quad i \in I \\ & \nabla_{\boldsymbol{\theta}} h_j(\boldsymbol{\theta}_0)^T \boldsymbol{z} = 0 \quad j = p+1, \ldots, q \end{aligned}$$

has an unique solution at $B(M) = \{\boldsymbol{z} : \|\boldsymbol{z}\| < M\}$ for any large M.

Based on the assumptions of our constrained model, Condition 1 is true for $\sigma^2 = 1$. For Conditions 2, since $\sigma_0(X)$ is differentiable in X, $R(X_i; \boldsymbol{\theta}) = \lambda\sigma_0(\mu X_i)$ is differentiable in $\boldsymbol{\theta}$. By Taylor expansion,

$$R(X_i;\boldsymbol{\theta}) = R(X_i;\boldsymbol{\theta}_0) + \nabla_{\boldsymbol{\theta}} R(X_i;\boldsymbol{\theta}_0)^T(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T \Delta_{\boldsymbol{\theta}} R(X_i;\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + o(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2)$$

where

$$\nabla_{\boldsymbol{\theta}} R(X_i;\boldsymbol{\theta}_0) = \begin{pmatrix} \lambda_0 X_i \sigma_0'(\mu_0 X_i) \\ \\ \sigma_0(\mu_0 X_i) \end{pmatrix}$$

$$\Delta_{\boldsymbol{\theta}} R(X_i;\boldsymbol{\theta}_0) = \begin{pmatrix} \lambda_0 X_i^2 \sigma_0''(\mu_0 X_i) & X_i \sigma_0'(\mu_0 X_i) \\ \\ X_i \sigma_0'(\mu_0 X_i) & 0 \end{pmatrix}$$

It's easy to check that both the functions $X^2\sigma_0''(X)$ and $X\sigma_0(X)$ are bounded. So there exists a constant $B > 0$ such that all the elements in $\Delta_{\boldsymbol{\theta}} R(X_i;\boldsymbol{\theta}_0)$ are within $[-B, B]$. By Lemma 29 , we have $\Delta_{\boldsymbol{\theta}} R(X_i;\boldsymbol{\theta}_0) < 2BI_2$. As a result,

$$\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T \Delta_{\boldsymbol{\theta}} R(X_i;\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + o(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2) \leq \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T 2BI_2(\boldsymbol{\theta}-\boldsymbol{\theta}_0) = B\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2$$

Therefore, $r_i(\boldsymbol{\theta}) \leq B$ and $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} r_i^2(\boldsymbol{\theta}) \leq B^2 < \infty$ holds in the whole parameter space.

For Condition 3, we can compute $\nabla_{\boldsymbol{\theta}} R(X_i;\boldsymbol{\theta}_0) = \begin{pmatrix} \lambda_0 X_i \sigma_0'(\mu_0 X_i) \\ \sigma_0(\mu_0 X_i) \end{pmatrix}$

$$\nabla_{\boldsymbol{\theta}} R(X_i;\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}} R(X_i;\boldsymbol{\theta}_0)^T = \begin{pmatrix} \lambda_0^2 X_i^2 \sigma_0'(\mu_0 X_i)^2 & \lambda_0 X_i \sigma_0(\mu_0 X_i)\sigma_0'(\mu_0 X_i) \\ \lambda_0 X_i \sigma_0(\mu_0 X_i)\sigma_0'(\mu_0 X_i) & \sigma_0(\mu_0 X_i)^2 \end{pmatrix}$$

$$\frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} R \nabla_{\boldsymbol{\theta}} R^T = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} \lambda_0^2 X_i^2 \sigma_0'(\mu_0 X_i)^2 & \frac{1}{n}\sum_{i=1}^{n} \lambda_0 X_i \sigma_0(\mu_0 X_i)\sigma_0'(\mu_0 X_i) \\ \frac{1}{n}\sum_{i=1}^{n} \lambda_0 X_i \sigma_0(\mu_0 X_i)\sigma_0'(\mu_0 X_i) & \frac{1}{n}\sum_{i=1}^{n} \sigma_0(\mu_0 X_i)^2 \end{pmatrix}$$

It's easy to check that all the functions $X^2\sigma_0'(X)^2$, $X\sigma_0(X)\sigma_0'(X)$ and $\sigma_0(X)^2$ are bounded. By Kolmogorov's Strong Law of Large Numbers (SLLN),

$$\frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} R \nabla_{\boldsymbol{\theta}} R^T \overset{a.s.}{\to} K = \begin{pmatrix} E_x[\lambda_0^2 X^2 \sigma_0'(\mu_0 X)^2] & E_X[\lambda_0 X \sigma_0(\mu_0 X)\sigma_0'(\mu_0 X)] \\ E_X[\lambda_0 X \sigma_0(\mu_0 X)\sigma_0'(\mu_0 X)] & E_X[\sigma_0(\mu_0 X)^2] \end{pmatrix}$$

That is,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} R(X_i;\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}} R(X_i;\boldsymbol{\theta}_0)^T = K = E_X[\nabla_{\boldsymbol{\theta}} R(X;\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}} R(X;\boldsymbol{\theta}_0)^T]$$

By Lemma 30 , we know that K is positive definite.

In our model, there are none equality constraint and three inequality constraints as follows

$$g_1(\boldsymbol{\theta}) = -\lambda^{\alpha-1} + \lambda^{\alpha}\frac{\|\sigma_0\|_1}{\mu R_{tot}} < 0$$

$$g_2(\boldsymbol{\theta}) = -\mu < 0$$

$$g_3(\boldsymbol{\theta}) = -\lambda < 0$$

Therefore, $p = q = 3$, $I = \{i : g_i(\boldsymbol{\theta}_0) = 0, i = 1,\ldots,p\} = \emptyset$. Obviously, the three constraint functions are all continuously differentiable in $\boldsymbol{\theta}$ in the whole parameter space,

so Condition 4 holds. Since I is an empty set and there is no equality constraints with h_j in our model, Condition 5 holds for sure and the minimization problem in Condition 6 in our case is

$$\min_{\boldsymbol{z}} \quad \boldsymbol{z}'K\boldsymbol{z} - 2\boldsymbol{z}'\boldsymbol{\xi}$$

which for any value of $\boldsymbol{\xi}$, has an unique solution $\hat{\boldsymbol{z}} = K^{-1}\boldsymbol{\xi}$ at $B(M)$ for any large $M > \|K^{-1}\boldsymbol{\xi}\|^{\frac{1}{2}}$.

So far we have proved all the conditions hold when $\sigma^2 = 1$, so we follow Theorem 6 of Wang (1996) [82] and reach our Theorem 26. In fact, we can prove that Theorem 26 holds generally for any known $\sigma^2 > 0$. As we can see in [82], $\sigma^2 = 1$ is such that $n^{-\frac{1}{2}} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} R(X_i; \boldsymbol{\theta}_0)\epsilon_i \xrightarrow{d} N(0, K)$. More generally, based on Theorem 4 and 5 of Jennrich (1969) [45], $n^{-\frac{1}{2}} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} R(X_i; \boldsymbol{\theta}_0)\boldsymbol{\epsilon}_i \xrightarrow{d} N(0, \sigma^2 K)$ for any $\sigma^2 > 0$. As a result, $F_n(\boldsymbol{\epsilon}_i, \boldsymbol{z})$ will converge in distribution to $\boldsymbol{z}'K\boldsymbol{z} - 2\boldsymbol{z}'\boldsymbol{\xi}$ where $\boldsymbol{\xi}$ now is a random vector of $N(0, \sigma^2 K)$. The rest of the proof are all same except now $\boldsymbol{\xi} \sim N(0, \sigma^2 K)$ instead of $N(0, K)$. Further, we observed that Theorem 1-6 of Wang (1996) [82] does not necessarily require σ^2 to be known. They still holds even if σ^2 is unknown based on Theorem 1-5 of Jennrich (1969) [45].

Next, we are going to prove the consistency of $\hat{\sigma}_n^2$. Before it, let's prove the following lemma

Lemma 31 *The estimator $\hat{\boldsymbol{\theta}}$ is such that*

$$\frac{1}{n}\sum_{i=1}^{n}\Big(R(X_i; \hat{\boldsymbol{\theta}}_n) - R(X_i; \boldsymbol{\theta}_0)\Big)^2 \xrightarrow{p} 0$$
$$\frac{1}{n}\sum_{i=1}^{n}\Big(R(X_i; \hat{\boldsymbol{\theta}}_n) - R(X_i; \boldsymbol{\theta}_0)\Big)\epsilon_i \xrightarrow{p} 0$$

Proof. By Taylor expansion of $R(X;\boldsymbol{\theta})$ in the neighborhood of $\boldsymbol{\theta}_0$, we have

$$R(X;\boldsymbol{\theta}) = R(X;\boldsymbol{\theta}_0) + \nabla_{\boldsymbol{\theta}} R(X;\boldsymbol{\theta}_0)^T(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + o(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|)$$

As proved earlier, $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \sigma K^{-\frac{1}{2}}\boldsymbol{Z}$. As a result, $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$. Therefore,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\Big(R(X_i;\hat{\boldsymbol{\theta}}_n) - R(X_i;\boldsymbol{\theta}_0)\Big)^2 \\
&= (R(X_i;\hat{\boldsymbol{\theta}}_n) - R(X_i;\boldsymbol{\theta}_0))^T[\frac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}} R(X_i;\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}} R(X_i;\boldsymbol{\theta}_0)^T](R(X_i;\hat{\boldsymbol{\theta}}_n) - R(X_i;\boldsymbol{\theta}_0)) \\
&+ o(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2)
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\Big(R(X_i;\hat{\boldsymbol{\theta}}_n) - R(X_i;\boldsymbol{\theta}_0)\Big)\epsilon_i \\
&= (R(X_i;\hat{\boldsymbol{\theta}}_n) - R(X_i;\boldsymbol{\theta}_0))^T\frac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}} R(X_i;\boldsymbol{\theta}_0)\epsilon_i + o(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|)
\end{aligned}$$

It has been proved that $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}} R(X_i;\boldsymbol{\theta}_0)\nabla_{\boldsymbol{\theta}} R(X_i;\boldsymbol{\theta}_0)^T = K$ is a positive definite matrix. Therefore,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\Big(R(X_i;\hat{\boldsymbol{\theta}}_n) - R(X_i;\boldsymbol{\theta}_0)\Big)^2 \\
&\to (R(X_i;\hat{\boldsymbol{\theta}}_n) - R(X_i;\boldsymbol{\theta}_0))^T K(R(X_i;\hat{\boldsymbol{\theta}}_n) - R(X_i;\boldsymbol{\theta}_0)) \xrightarrow{p} 0
\end{aligned}$$

By Theorem 4 of Jennrich 1969 [45], $\frac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}} R(X_i;\boldsymbol{\theta}_0)\epsilon_i \xrightarrow{p} 0$. Therefore,

$\frac{1}{n}\sum_{i=1}^{n}\Big(R(X_i;\hat{\boldsymbol{\theta}}_n) - R(X_i;\boldsymbol{\theta}_0)\Big)\epsilon_i \xrightarrow{p} 0$, which completes the proof of lemma 31. ■

With lemma 31, let's prove the consistency of $\hat{\sigma}_n^2$. In the CNLS procedure,

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - R(X_i;\hat{\boldsymbol{\theta}}_n)\Big)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(R(X_i;\boldsymbol{\theta}_0) + \epsilon_i - R(X_i;\hat{\boldsymbol{\theta}}_n)\Big)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(R(x_i;\hat{\boldsymbol{\theta}}_n) - R(X_i;\boldsymbol{\theta}_0)\Big)^2 - \frac{2}{n}\sum_{i=1}^{n}\Big(R(X_i;\hat{\boldsymbol{\theta}}_n) - R(X_i;\boldsymbol{\theta}_0)\Big)\epsilon_i + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2
\end{aligned}$$

By lemma 31, $\frac{1}{n}\sum_{i=1}^{n}\left(R(X_i;\hat{\boldsymbol{\theta}}_n)-R(X_i;\boldsymbol{\theta}_0)\right)^2 \xrightarrow{p} 0$ and

$\frac{2}{n}\sum_{i=1}^{n}\left(R(X_i;\hat{\boldsymbol{\theta}}_n)-R(X_i;\boldsymbol{\theta}_0)\right)\epsilon_i \xrightarrow{p} 0$. And by $SLLN$, $\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 \xrightarrow{p} \sigma^2$ Therefore, $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$.

6.6 Proof of Theorem 27

Let $\boldsymbol{\theta}_i = (\mu_i, \lambda_i)'$, $\boldsymbol{\theta} = (\mu, \lambda)'$. By Theorem 26, we can easily conclude that for large $n_i = n$, the estimators $\hat{\boldsymbol{\theta}}_i$ in step 2 has the following asymptotic distribution, conditional on $\boldsymbol{\theta}_i$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) \xrightarrow{d} MVN\left(\mathbf{0}, \sigma^2(K_i)^{-1}\right)$$

where $K_i = E_X[\nabla_{\boldsymbol{\theta}_i} R(X;\boldsymbol{\theta}_i)\nabla_{\boldsymbol{\theta}_i} R(X;\boldsymbol{\theta}_i)^T]$. Furthermore, $\{\hat{\boldsymbol{\theta}}_i : i = 1, \cdots, m\}$ are mutually independent since they were estimated from sample of different subjects.

From above, the conditional asymptotic mean and variance for $\hat{\boldsymbol{\theta}}_i$ are as follows.

$$E_\epsilon(\hat{\boldsymbol{\theta}}_i|\boldsymbol{\theta}_i) = \boldsymbol{\theta}_i$$

$$Var_\epsilon(\hat{\boldsymbol{\theta}}_i|\boldsymbol{\theta}_i) = \sigma^2(nK_i)^{-1}$$

Then taking the variability of the random effects $\boldsymbol{\theta}_i$ into account, the unconditional asymptotic mean and variance $\hat{\boldsymbol{\theta}}_i$ are as follows:

$$\begin{aligned}
E(\hat{\boldsymbol{\theta}}_i) &= E_{\boldsymbol{\theta}}[E_\epsilon(\hat{\boldsymbol{\theta}}_i|\boldsymbol{\theta}_i)] = E_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i) = \boldsymbol{\theta} \\
Var(\hat{\boldsymbol{\theta}}_i) &= Var_{\boldsymbol{\theta}}[E_\epsilon(\hat{\boldsymbol{\theta}}_i|\boldsymbol{\theta}_i)] + E_{\boldsymbol{\theta}}[Var_\epsilon(\hat{\boldsymbol{\theta}}_i|\boldsymbol{\theta}_i)] \\
&= Var_{\boldsymbol{\theta}}[\boldsymbol{\theta}_i] + E_{\boldsymbol{\theta}}[Var_\epsilon(\hat{\boldsymbol{\theta}}_i|\boldsymbol{\theta}_i)] \\
&= \Sigma + \sigma^2 E_{\boldsymbol{\theta}}[(nK_i)^{-1}] \\
&\doteq \tilde{\Sigma}
\end{aligned}$$

Since $\{\boldsymbol{\theta}_i : i = 1, \cdots, m\}$ are $i.i.d$ and $R(X;\cdot)$ is smooth, $E_{\boldsymbol{\theta}}[\sigma^2(nK_i)^{-1}]$ are same for all the subject. In other words, $E_{\boldsymbol{\theta}}[\sigma^2(nK_i)^{-1}]$ is independent on i. Therefore, $\{\hat{\boldsymbol{\theta}}_i : i = 1, \cdots, m\}$ are $i.i.d.$ with common mean and variance.

Note that in the above equations, $E_{\boldsymbol{\theta}}(\cdot)$ is referred to taking expectation with respected to $\boldsymbol{\theta}$ at the population level, and $E_\epsilon(\cdot|\boldsymbol{\theta}_i)$ is referred to taking expectation with respected to ϵ at the individual level. Same rule was applied in the variance calculation.

The CMM estimators, $\hat{\mu}$ and $\hat{\lambda}$ in step 3 can be expressed in a vector form as follows

$$\hat{\boldsymbol{\theta}} = \frac{1}{m}\sum_{i=1}^{m}\hat{\boldsymbol{\theta}}_i$$

Notice that $\{\boldsymbol{\theta}_i : i = 1, \cdots, m\}$ are $i.i.d.$ with common mean $\boldsymbol{\theta}$ and variance $\tilde{\Sigma}$, then when m goes to ∞, by SLLN we have $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$. Furthermore, by Central Limit Theorem (CLT) we have $\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \tilde{\Sigma}^{\frac{1}{2}}\boldsymbol{Z}$ with $\boldsymbol{Z} \sim N(0, I_2)$.

The CMM estimator of Σ, $\hat{\Sigma}$ in step 5 is

$$\hat{\Sigma} = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}})^T - \hat{\sigma}^2\frac{1}{m}\sum_{i=1}^{m}T_i^{-1}$$

$$\text{with} \qquad T_i = \left[\frac{\partial \boldsymbol{R}_i}{\partial \boldsymbol{\theta}_i^T}\right]^T \left[\frac{\partial \boldsymbol{R}_i}{\partial \boldsymbol{\theta}_i^T}\right] |_{\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i}$$

$$\boldsymbol{R}_i = (R(X_{i1};\boldsymbol{\theta}_i), R(X_{i2};\boldsymbol{\theta}_i), \cdots, R(X_{in_i};\boldsymbol{\theta}_i))^T$$

One one hand, since $\{\boldsymbol{\theta}_i : i = 1, \cdots, m\}$ are $i.i.d.$ with common mean $\boldsymbol{\theta}$ and variance $\tilde{\Sigma}$, by SLLN we have

$$\frac{1}{m-1}\sum_{i=1}^{m}(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}})^T \xrightarrow{p} \tilde{\Sigma}$$

On the other hand, T_i is such that

$$
\begin{aligned}
E(T_i^{-1}) &= E_{\boldsymbol{\theta}}[E_{\epsilon}(T_i^{-1}|\boldsymbol{\theta}_i)] = E_{\boldsymbol{\theta}}[E_{\epsilon}(\left(\left[\frac{\partial \boldsymbol{R}_i}{\partial \boldsymbol{\theta}_i^T}\right]^T \left[\frac{\partial \boldsymbol{R}_i}{\partial \boldsymbol{\theta}_i^T}\right] |_{\boldsymbol{\theta}_i=\hat{\boldsymbol{\theta}}_i}\right)^{-1} |\boldsymbol{\theta}_i)] \\
&\xrightarrow{p} E_{\boldsymbol{\theta}}\left[\left(\left[\frac{\partial \boldsymbol{R}_i}{\partial \boldsymbol{\theta}_i^T}\right]^T \left[\frac{\partial \boldsymbol{R}_i}{\partial \boldsymbol{\theta}_i^T}\right]\right)^{-1}\right] \\
&= E_{\boldsymbol{\theta}}\left[\left(\sum_{j=1}^{n}\left[\frac{\partial R(X_{ij};\boldsymbol{\theta}_i)}{\partial \boldsymbol{\theta}_i^T}\right]\left[\frac{\partial R(X_{ij};\boldsymbol{\theta}_i)}{\partial \boldsymbol{\theta}_i^T}\right]^T\right)^{-1}\right] \\
&\xrightarrow{p} E_{\boldsymbol{\theta}}[\big(nE_X[\nabla_{\boldsymbol{\theta}_i} R(X;\boldsymbol{\theta}_i)\nabla_{\boldsymbol{\theta}_i} R(X;\boldsymbol{\theta}_i)^T]\big)^{-1}] = E_{\boldsymbol{\theta}}[(nK_i)^{-1}]
\end{aligned}
$$

The first "$\xrightarrow{p}$" in the above equation holds since $\hat{\boldsymbol{\theta}}_i \xrightarrow{p} \boldsymbol{\theta}_i$. The second "$\xrightarrow{p}$" holds by SLLN of X. Therefore, $\{T_i^{-1} : i = 1, \cdots, m\}$ are *i.i.d.* r.v's with the same mean, by SLLN we have

$$\frac{1}{m}\sum_{i=1}^{m} T_i^{-1} \xrightarrow{p} E_{\boldsymbol{\theta}}[(nK_i)^{-1}]$$

In addition, it's assumed that $\hat{\sigma^2} \xrightarrow{p} \sigma^2$. Therefore, by *slutsky's Theorem*, $\hat{\Sigma} \xrightarrow{p} \tilde{\Sigma} - \sigma^2 E_{\boldsymbol{\theta}}[(nK_i)^{-1}] = \Sigma$.

Let $\hat{\tilde{\Sigma}} = \hat{\Sigma} + \hat{\sigma}^2 E_{\boldsymbol{\theta}}[(nK_i)^{-1}]$, then based on the asymptotical result of $\hat{\Sigma}$, we know that $\hat{\tilde{\Sigma}} \xrightarrow{p} \tilde{\Sigma}$. We also have the asymptotical result of $\hat{\boldsymbol{\theta}}$, $\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \tilde{\Sigma}^{\frac{1}{2}}\boldsymbol{Z}$. By *slutsky's Theorem*, we have $\sqrt{m}\hat{\tilde{\Sigma}}^{-\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{Z}$. Therefore, the Theorem 27 holds.