Parameter Estimation on the Partially Observed Bidimensional Ornstein-Uhlenbeck Process

by

Ziqiu Qin - 5287987

Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in

Financial Engineering

Applied Mathematics

Faculty of Electrical Engineering, Mathematics and Computer Science

Delft University of Technology

to be defended on Thursday November 3, 2022 at 11:00 AM

Supervisor: Dr. Ir. F.H. van der Meulen

Thesis Comittee: Prof. Dr. Ir. A.W. Heemink
Dr. H.N. Kekkonen

Preface

This master thesis project is precious to me because I have made many trials and breakthroughs in this process. The relevant knowledge of the thesis includes many new techniques that I have not studied systematically. During this period, I got challenged to use an unlearned programming language Julia and finish the project on my own. The process of research was tortuous and with lots of ups and downs. But when I overcame all these difficulties, I felt proud of myself and what I had done. I learned more about how to do research and how to write a scientific essay in the process of working on my thesis.

My research aimed to perform parameter estimation on a bi-dimensional Ornstein-Uhlenbeck process. As is studied, the maximum likelihood method could estimate some parameters with limitations. So, we employ a novel Bayesian approach to improve the estimation and try some tentative experiments with fewer given conditions.

I would first like to thank my supervisor, Frank van der Meulen, whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

In addition, I would like to thank my parents for their wise counsel and sympathetic ear. You are always there for me. Your company is what keeps me going.

Finally, I could not have completed this dissertation without the support of my friends, Jingya Li, Qianqian Chen, and Zhimin Cheng, who provided stimulating discussions as well as happy distractions to rest my mind outside of my research. You not only answer my questions about the graduation process, but also gives me suggestions on revision to complete my dissertation.

Ziqiu Qin
October 26, 2022

Abstract

In anti-cancer therapy, antiangiogenic treatments are applied and take effect on the vascularization of tissue. To evaluate the efficacy of treatments, we adopt two methods to solve the physiological pharmacokinetic model's parameter estimation problem, providing discrete, partial, and noisy observations of stochastic differential equations. One is to compute the exact likelihood using the Kalman filter recursion and implement numerical maximization [1]. The other is a novel Markov Chain Monte Carlo algorithm to estimate parameters using guided proposals [2] in a Bayesian setup, namely Backward Filtering with Forward Guiding algorithm [3]. The identifiability of the model and parameters [4] are established before parametric inference. We extend the BFFG algorithm to include an automatic optimal kernel finding scheme [5] for the Metropolis-Hastings-within-Gibbs sampler. In comparison, a Conjugate Gradient algorithm is applied when employing the maximum likelihood method. Besides performing parameter estimation via different methods separately, joint estimation is performed using the Bayesian approach. After that, time delay and Arterial Input Function in the statistical model are estimated via change point detection [6] and piece-wise inference. We illustrate the goodness-of-fit of estimates and advantages of the bayesian approach towards the method using the maximum likelihood.

Contents

1 Introduction

1.1 Background Information

Stochastic mathematical models are becoming an essential tool for interpreting biological and physiological problems [7]. As these models are parametric, with improving techniques for inferring the stochastic models, it becomes a significant issue to calibrate model parameters relying on experimental data. From a biological perspective, this dissertation aims to build a model for tissue microvascularization in anti-cancer therapy [1] such that it can explain the available data. We consider a bidimensional Ornstein-Uhlenbeck process to describe the microcirculation. Using a stochastic system avoids the limitations of a deterministic differential system to capture random fluctuations. So we can establish the physiologically based pharmacokinetic model we study in this thesis. This pharmacokinetic model can be defined using a bi-dimensional stochastic differential equation (SDE) by adding a Brownian motion to the deterministic model.

As a prelude to the parameter estimation experiment, we establish an identifiability test in advance since parameter estimates from a non-identifiable model may be unreliable [7]. Identifiability analysis for stochastic models is a newly-developed technique in [4], and it is well adapted to the pharmacokinetic model in this thesis. With identifiability established, we get down to parameter inference. There is abundant literature about the statistical theory of estimating parameters of ODEs and SDEs. Some methods have been proposed for estimating parameters of discretely observed SDEs, including the Maximum Likelihood Estimator and Expectation Maximization algorithm [1]. However, parametric inference based on discrete-time observations poses difficulties in the case of partially observed discrete noisy data. One problem is that actual data observed can only be obtained in discrete forms, as partial and noisy observations, whilst the targeted stochastic model is continuous. And the lack of an explicit likelihood makes these methods time-consuming and challenging to operate in high-dimensional cases. Therefore, a new approach to dealing with a delicate stochastic system with discrete, partial, and noisy data based on a low computational cost is necessitated. Another difficulty is estimating the volatility terms. Previous research succeeded in estimating the drift parameters of the pharmacokinetic model only by employing methods including the Maximum Likelihood Estimation with Kalman-filtering or Expectation Maximization schemes [1]. Therefore, estimating the diffusive parameters and the Arterial Input Function (AIF), as well as a joint estimation of all the parameters, attracts our attention.

To deal with the parameter estimation problems given discrete-time missing data without relying on discretization, we introduce a newly-proposed method called guided proposals [2] for simulating a multi-dimensional diffusion bridge. Then Bayesian estimation can be conducted on auxiliary processes [8] that are discretely observed multi-dimensional diffusion processes. An efficient algorithm to recover the paths of the diffusion process and estimate parameters is derived in [3], which is a guided proposal-dependent Backward Filtering Forward Guiding (BFFG) algorithm to sample from the exact smoothing distributions. In this manuscript, the Bayesian method using the BFFG algorithm is adapted to estimate the parameters in the pharmacokinetic model. To reduce the computational complexity, we improve the BFFG algorithm to automatically find the optimal proposal kernel in the MCMC step by including a scaled algorithm for Metropolis-Hastings-within-Gibbs sampler [5]. Furthermore, considering estimating the initial time delay and AIF, we need to detect change points to isolate experimental data. The subject of Change Point Detection has been investigated for many years. Implemented algorithms for detecting multiple change points of a given time series [9] and the segmentation method [6] is adopted for estimating the AIF in this thesis. While the Least Squares method [10] is used for detecting the time delay. We illustrate the usage and benefit of the Bayesian approach by calibrating these methods to experimental data and comparing the result obtained using the MLE method given the same setting.

1.2 Thesis Framework

The organization of this manuscript is given as follows. In Chapter 2, we describe the physiological pharmacokinetic model we study throughout the thesis and introduce all the notations used in this thesis. After that, some mathematical knowledge is recapped for easy comprehension. Then we present all the applied methodologies in chapter 3, including the identifiability analysis, parameter estimation method using Maximum Likelihood with Kalman filtering, parameter estimation method using the Bayesian approach, the scaled method for finding the optimal tuning proposals in Adaptive Markov Chains, and the Change-point Estimation Schemes for estimating Arterial Input Functions. After explaining the theories, the implementations of the above methodologies are elaborated in Chapter 4, with detailed input parameters and initial settings stated. Finally, Chapter 5 discusses the results obtained from implementing the parameter inference schemes on the model and summarizes all achieved results to make conclusions and suggestions for further research.

2 Preliminaries.

2.1 Model Description

This manuscript targets a physiological pharmacokinetic model, which was used to estimate tumor micro-circulation [11]. A bi-dimensional deterministic differential system usually models this micro-circulation. Just like the injection of a contrast agent, Vistarem is investigated by recording the evolution of the concentration of the contrast agent over time. This physiological pharmacokinetic model is applied, which describes the distribution of the contrast agent. The system includes four compartments in total, two of which in the middle can be split into two subcompartments. Firstly, the plasma compartments include arterial and venous plasma. Then comes the tumor, which can be divided into two parts, capillaries, and interstitium. And finally, we split the rest of the body into subcompartments, capillaries, and interstitium. The contrast agent pulsates in the plasma and interstitium cells, and the schematic of the model is shown in Fig.1.

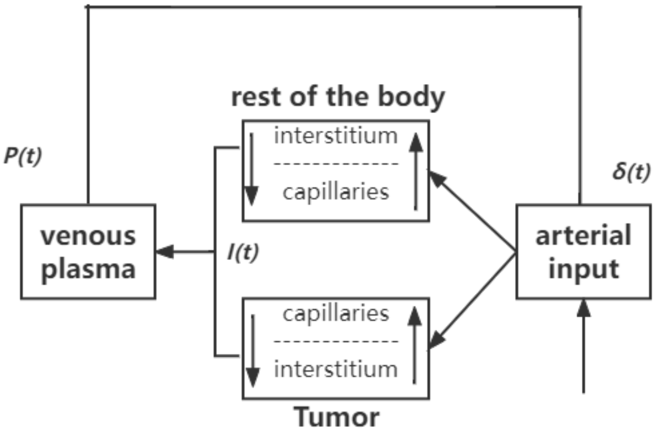


Figure 1: Description of the model

The contrast agent was introduced into the caudal vein by injection, physiologically equivalent to an injection shortly before the left ventricle with a time delay. We use an infusion in arterial blood to approximate the injection flow, with the in-and-out flows represented by arrows in Fig.1. The general framework of the contrast agent flow process is shown clearly in the above figure. It demonstrates that the contrast agent is firstly injected into the vein and then transits into the artery. Next, the contrast agent arrives in arterial plasma, with a tissue perfusion flow (denoted as F_{tp}). Eventually, the contrast agent is eliminated from venous plasma proportionally to the contrast agent in plasma with the perfusion flow F_{tp}. In the intermediate step, the exchange of flows happens. The contrast agent flows into and out of two compartments: the rest of the body and the tumor. The quantity of the contrast agent exchanged from plasma through interstitium equals the product of the contrast agent concentration in plasma and the volume transfer constant. Furthermore, the amount of contrast agent exchanged from interstitium through plasma equals the product of the contrast agent concentration and the volume transfer constant.

More precisely, we use this pharmacokinetic model to describe the kinetics of the contrast agent in the voxel with two compartments (plasma and interstitial water). The deterministic version of the pharmacokinetic model can be illustrated in the form of an ODE system defined as:

$$\begin{cases} \mathrm{d}P(t) = (\dfrac{F_{tp}}{1-h}\delta(t) - (\dfrac{F_{tp}+K_{trans}}{V_P})P(t) + \dfrac{K_{trans}}{V_I}I(t))dt, \\ \mathrm{d}I(t) = (\dfrac{K_{trans}}{V_P}P(t) - (\dfrac{K_{trans}}{V_I})I(t))dt. \end{cases} \tag{1}$$

Here $\delta(t),\ P(t),\ I(t)$ all denote the quantity of contrast agent at time t, but in the artery (namely the "Arterial Input Function"), the plasma and the interstitium respectively. h is the hematocrit rate set to $h = 0.4$, and $1-h$ is the volume of the artery. V_P and V_I denote the volume of plasma and interstitium, respectively. According to the biological constraints, we request $0 \le V_P(1-h), V_I \le 100$ and $V_P(1-h) + V_I \le 100$. F_{tp} is the perfusion flow, and K_{trans} is the volume transfer constant. The contrast agent is supposed to be injected in vein at time t_0, and the initial condition at time $t_0 = 0$ is set as $P(0) = 0,\ I(0) = 0$.

It is transformed into a stochastic differential equation, which is our targeted model, by adding a Brownian motion to the differential equation on each compartment. To get a concise form, we use reparametrization by assuming:

$$\alpha = \frac{F_{tp}}{1-h}, \quad \beta = \frac{F_{tp}}{V_P}, \quad \lambda = \frac{K_{trans}}{V_P}, \quad k = \frac{K_{trans}}{V_P} + \frac{K_{trans}}{V_I}. \tag{2}$$

This contributes to the following system of differential equations [1], which is the stochastic version of the targeted pharmacokinetic model:

$$\begin{cases} \mathrm{d}P(t) = (\alpha\delta(t) - (\lambda+\beta)P(t) + (k-\lambda)I(t))\,\mathrm{d}t + \sigma_1\,\mathrm{d}W_1(t), \\ \mathrm{d}I(t) = (\lambda P(t) - (k-\lambda)I(t))\,\mathrm{d}t + \sigma_2\,\mathrm{d}W_2(t), \end{cases} \tag{3}$$

where $P(t),\ I(t)$ represent contrast agent concentrations in compartment plasma and interstitium respectively. Considering the newly introduced parameters, $\delta(t)$ is a known input function related to the contrast agent quantity in the arterial, and $\alpha,\ \beta,\ k, \lambda$ are positive unknowns whose values are aimed to infer, with $k > \lambda$. $W_1,\ W_2$ are two independent Brownian Motions added, with diffusion terms $\sigma_1,\ \sigma_2$ respectively.

Let $S(t)$ denote the total quantity of contrast agent in the compartments at time t, i.e. $S(t) = P(t) + I(t)$. For simplicity of studying the general properties of the model later, we transform the system above to a matrix form by introducing a new matrix $X(t) = [P(t), I(t)]'$ so that the system is transformed to

$$\begin{aligned} \mathrm{d}X(t) &= \left(\begin{bmatrix} \alpha\,\delta(t) \\ 0 \end{bmatrix} + \begin{bmatrix} -(\lambda+\beta) & \beta \\ \lambda & -k \end{bmatrix} X(t) \right) \mathrm{d}t + \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} dW_1(t) \\ dW_2(t) \end{bmatrix} \\ &= \boldsymbol{b}(X(t), t, \delta(t))\,\mathrm{d}t + \boldsymbol{\sigma}\,\mathrm{d}W(t) \end{aligned} \tag{4}$$

where $\boldsymbol{b}, \boldsymbol{\sigma}$ are the matrices of drift and diffusion respectively.

We choose measure the sum of the coordinates of $X(t)$, which can also be seen as $S(t)$ using the observation model

$$\begin{aligned} y_i &= S(t_i) + \sigma \epsilon_i \\ &= HX(t_i) + \sigma \epsilon_i, \end{aligned} \tag{5}$$

where $J = [1, 1]$, $\varepsilon_i \sim \mathcal{N}(0, 1)$ represent the mutually independent Gaussian noises, and σ denotes the constant standard deviation of noise. Thus noisy and discrete observations can be obtained at times $0 = t_0 < t_1 < \cdots < t_n = T$. The measurement noise differs from the added random variations in the model due to the precision of the recording experiments and is thus an uncorrelated noise [12].

The statistical problem is to give a precise form of the bi-dimensional stochastic differential system of the pharmacokinetic model. We expect to calibrate all model parameters, defined using the set parameter θ simultaneously, given discrete, partial, and noisy observations. We will first implement parametric inference on a single parameter with other parameters fixed given experimental data, including the drift, the dispersion, and the Arterial Input Function, to achieve this goal. Then we attempt to conduct a joint estimation of all drift terms, all drift and diffusive terms, and all parameters, including AIF, one by one. Estimators differ from each other when different estimation methods are applied.

2.2 Notations Used

In this section, we summarize all the notations and abbreviations of terms used in the thesis.

Notation	Description
model	the target model throughout the thesis is the "physiological pharmacokinetic" model or pharmacokinetic model in short
SDE	the abbreviation for statistical differential equation
ODE	the abbreviation for ordinary differential equation
$\delta(t)$	the Arterial Input Function in the SDE, abbreviated to "AIF"
OU-process	the abbreviation for the bidimensional Ornstein-Uhlenbeck process is studied in this thesis
θ	the collection of all parameters to estimate, excluding the AIF. It has different forms in MLE and Bayesian setup due to reparametrization.
$\alpha,\ \beta,\ \lambda,\ k$	drfit terms among the target parameters
$\boldsymbol{b}$	drfit in matrix form
$\sigma_1,\ \sigma_2$	diffusive terms among the target parameters
$\boldsymbol{\sigma}$	dispersion in matrix form
MLE	the abbreviation for "Maximum Likelihood Estimator"
BFFG	the abbreviation for the "Backward Filtering with Forward Guiding" algorithm used in the Bayesian setup
X	the original, unconditioned diffusion process
X^o	the proposal process
$\tilde{X}$	the auxiliary process with transition densities $\tilde{p}$
GP	the abbreviation for "guided proposals", which is obtained by performing random-walk Metropolis algorithm on the original $\mathbf{X_t}$
pCN	the abbreviation for the "preconditioned Crank-Nicolson" scheme
MCMC	the abbreviation for "Markov Chains Monte Carlo" process
MH	the abbreviation for the "Metropolis-Hastings" algorithm
LS	the abbreviation for the "least squares" method
BS	the abbreviation for the "Binary Segmentation" method
Acceptance (rate)	the proportion of the number of iterations when parameters get updated over all iteration

2.3 Recaps of Mathematical Statistics

This section outlines the essential statistical notions and theories needed to understand the methods introduced and applied throughout the thesis fully. We will present the definitions without interpretation and the detail-omitted algorithms.

2.3.1 Prerequisite Statistics

Since this thesis aims to build a stochastic model using differential equations and estimate its parameters, we need to provide some basic knowledge relevant to statistical inference and stochastic processes. We first introduce the Wiener process as follows.

Theorem 2.1. A Wiener process W_t [13] is an almost surely continuous process in t satisfying the following properties:

1. $W_0 = 0$
2. $\forall t > 0,\ u \geq 0,\ W_{t+u} - W_t$ are independent of $W_s,\ s \leq t$
3. $W_{t+u} - W_t \sim \mathcal{N}(0, u)$

Then we can define a stochastic differential equation (SDE) [14] of the form used in this thesis and furthermore explain each element.

Definition 2.2. Suppose X_t is a continuous time stochastic process and W_t is a Wiener process. Then a SDE is given by

$$\mathrm{d}X_t = b(X_t, t)\,\mathrm{d}t + \sigma(X_t, t)\,\mathrm{d}W_t.$$

The function b, and σ are referred to as drift and diffusion coefficient respectively. X_t is a diffusion process which satisfies the Markov property.

When generating a Markov kernel in parameter estimation step of Bayesian method, we simply apply a Gaussian random walk.

Definition 2.3. Suppose X_t is a discrete Markov process, it is said to be a Gaussian Random Walk if it satisfies

$$X_0 = 0, \quad X_t = X_{t-1} + \epsilon_t.$$

Before arriving at the definition of identifiabilities, confidence intervals [15] need to be formally introduced.

Definition 2.4. Let θ be a random sample from a probability distribution with statistical parameter $\hat{\theta}$, which needs to be estimated. A confidence interval for the parameter $\hat{\theta}$ with confidence level γ is an interval $(u(\theta), v(\theta))$ determined by random variables $u(\theta), v(\theta)$ with the property

$$\mathrm{P}\big(u(\theta) < \hat{\theta} < v(\theta)\big) = \gamma.$$

In this thesis, we use a maximum likelihood estimator to estimate parameters. We compute the exact likelihood based on Kalman filtering, which is introduced as follows in a simplified definition [1].

Definition 2.5. Suppose (X_i) is a hidden Markov chain on $\mathbb{R}^2$, and the observations (y_i) are independent conditionally on (X_i). Let $y_{0:i} = (y_0, \dots, y_i)$ be the vector of observations until time t_i. Given the Gaussian conditional law of $X_i | y_{0:i-1}$, the Kalman filter is a procedure to give formulas for the following distributions recursively via iterations:

- (prediction step) $X_i | y_{0:i-1} \sim \mathcal{N}(\hat{X}_i^-, P_i^-)$
- (update step) $X_i | y_{0:i} \sim \mathcal{N}(\hat{X}_i, P_i)$

where

$$\hat{X}_i^- = \mathrm{E}\left[X_i | y_{0:i-1}\right], \quad P_i^- = \mathrm{E}\left[(X_i - \hat{X}_i^-)(X_i - \hat{X}_i^-)\prime\right]$$
$$\hat{X}_i = \mathrm{E}\left[X_i | y_{0:i}\right], \quad P_i = \mathrm{E}\left[(X_i - \hat{X}_i)(X_i - \hat{X}_i)\prime\right].$$

To get a system of differential equations (ODEs) from a stochastic process X_t and derive the moment equations of identifiability analysis, we need to find the differential of any time-dependent function of X_t using It$\hat{o}$'s lemma [16].

Theorem 2.6. For an It$\hat{o}$ diffusion process

$$\mathrm{d}X_t = b(X_t, t)\,\mathrm{d}t + \sigma(X_t, t)\,\mathrm{d}W_t,$$

Let $g(X_t, t)$ be a function of X_t and time t, with continuous partial derivatives,

$$\frac{\partial g}{\partial X_t}, \frac{\partial g^2}{\partial X_t^2}, \frac{\partial g}{\partial t}.$$

The differential of time-dependent function $g(X_t, t)$ also follows an It$\hat{o}$ process governed by the same Wiener process W_t,

$$\mathrm{d}g(X_t, t) = \left(\frac{\partial g}{\partial t} + b(X_t, t)\frac{\partial g}{\partial X_t} + \frac{1}{2}\frac{\partial g^2}{\partial X_t^2}\sigma^2(X_t, t)\right)\mathrm{d}t + \frac{\partial g}{\partial X_t}\sigma(X_t, t)\,\mathrm{d}W_t.$$

2.3.2 Prerequisites for Bayesian method

We then introduce preliminary knowledge of the Bayesian method's likelihood computation and parameter updates. Because the Bayesian inference algorithm needs iteratively updating the path, initial state, and parameters, the Metropolis-Hastings-within-Gibbs sampler is used. This approach samples the path $X = (X_t)$ and parameters θ via joint posterior distribution. The Metropolis-Hastings (MH) algorithm for a Markov chain (X_t) is employed for the frame of an iteration scheme. Hence we first briefly present this MH algorithm (Algorithm.1).

The Metropolis–Hastings (MH) algorithm [17] is a popular technique to build Markov chains with a given invariant distribution. Suppose we aim to sample from the distribution with density function $p(x)$. Then we generate the Markov chains using proposal distribution obtained as follows.

Algorithm 1: Metropolis-Hastings Algorithm

```
1. Initialize:
   Choose an initial state $X_0 = x_0$ and set $t = 0$.

2. Iterate from $t = 0$:

   (a) Suppose the state at time $t$ is $X_t = x_t$.
       Generate a candidate state for $X_{t+1}$ randomly from $x' \sim Q(x'|x_t)$.
   (b) Calculate the acceptance probability using the conditional probabilities
       $A(x'|x) = \min(1, \frac{\mathbb{P}(x'|x)}{\mathbb{P}(x|x')} \frac{g(x|x')}{g(x'|x)})$.
   (c) Generate a uniform random number $u \in [0, 1]$,
       • if $u \leq A(x', x_t)$, accept $X_{t+1} = x'$
       • if $u > A(x', x_t)$, keep $X_{t+1} = x_t$
```

The algorithm using the Gibbs sampling resembles a single component Metropolis-Hastings algorithm because it can convert a d-dimensional problem to d 1-dimensional problems separately. A Gibbs sampler [18] is used to used to sample from conditional distribution instead of a joint distribution given a multivariate distribution. Suppose we aim to obtain k samples of $X = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$ from a joint distribution $\pi(x_1, x_2, \dots, x_d)$. Denote the s-th sample by $X^i = (x_1^s, x_2^s, \dots, x_d^s)$. We implement the Gibbs sampling by Algorithm.2.

Algorithm 2: Gibbs sampler

```
1. Initialize:
   suppose the initial state $X^0 = (x_1^0, x_2^0, \dots, x_d^0)$ is known.

2. Iterate from $s = 0$ to $s = k$ by sampling for each dimension $i = 1, \dots, d$.

   • draw $X_1^{(s+1)} \sim \pi(x_1|x_2^{(s)}, \dots, x_d^{(s)})$
   • draw $X_2^{(s+1)} \sim \pi(x_2|x_1^{(s+1)}, \dots, x_d^{(s)})$
     ⋮
   • draw $X_i^{(s+1)} \sim \pi(x_i|x_1^{(s+1)}, \dots, x_{i-1}^{(s+1)}, x_{i+1}^{(s)}, \dots, x_d^{(s)})$
   • draw $X_{i+1}^{(s+1)} \sim \pi(x_{i+1}|x_1^{(s+1)}, \dots, x_i^{(s+1)}, x_{i+2}^{(s)}, \dots, x_d^{(s)})$
     ⋮
   • draw $X_d^{(s+1)} \sim \pi(x_d|x_1^{(s+1)}, \dots, x_{d-1}^{(s+1)})$
```

3 Methodology

This section overviews the mathematical and statistical techniques used to perform parametric inference and expatiates the critical methodologies. This chapter is split into several parts corresponding to the organization of the experiment process. By the order in which the experiment is conducted, this research can be divided into four stages.

Firstly, the method to establish identifiability analysis is explained, with two types of identifiability introduced. Then the previously studied estimation method via maximizing the likelihood is presented. After that, the novel Bayesian inference algorithm is elaborated with the scaled kernel optimization schemes. Finally, we provide approaches to detect the change points.

3.1 Identifiability

3.1.1 Overview and Setup

Identifiability analysis is established to explore model parameters that are meaningful to estimate. It should be performed in advance of the parameter estimation implementation. As the Bayesian approach using guided proposals deals with estimation for multiple parameters, including the diffusive terms simultaneously, joint estimation should be done based on the premise that all parameters are identifiable. A functional time-saving approach is introduced to detect which target parameters can be estimated successfully before the estimation experiment. We study both the pharmacokinetic model's identifiability and each model parameter's identifiability without evaluating its value, which saves computational costs during the time-consuming estimation process.

Assume n data points are given at time points $0 < t_1 < t_2 < \cdots < t_n$ as $y_1^d, y_2^2, \ldots, y_n^d$. We focus on the matrix form of the stochastic pharmacokinetic model (Eqn.4) and observations (Eqn.5). We use θ to represent the vector containing all model parameters. A stochastic differential equation and observation give the model a linear model with Gaussian noise.

$$\mathrm{d}X(t) = \boldsymbol{b}(X(t), t)\,\mathrm{d}t + \boldsymbol{\sigma}\,\mathrm{d}W(t) \tag{6}$$

$$y(t) = LX(t) + \epsilon(t) = g(t) + \epsilon(t) \tag{7}$$

Then a model describing n species concentrations x_i using a system of ODEs can be derived by integration. Allowing for a possible time delay ε before the injection, the observation model is rewritten as follows,

$$\begin{aligned} X(t) &= f(X(t), \theta, \delta(t)) \\ y(t) &= LX(t-\varepsilon) + \epsilon(t) \end{aligned} \tag{8}$$

with an externally stimulus $\delta(t)$, a set of dynamic parameters θ, an offset parameter ε, and a normally distributed noise measurement $\epsilon(t) \sim \mathcal{N}(0, \sigma^2)$.

3.1.2 Identifiability Definition

Proposition 3.1. The parameters can be estimated numerically [7] by minimizing a weighted sum of squared residuals which is used for measuring the agreement of experimental data and predicted observables. For normally distributed $\epsilon(t) \sim \mathcal{N}(0, \sigma^2)$, this corresponds to the maximum likelihood estimate (MLE) of θ

$$\hat{\theta} = \text{argmin}\, \chi^2(\theta) = \text{argmin}\left(C - 2l(\theta)\right) \tag{9}$$

where C is a constant and $l(\theta)$ is the log-likelihood, andχ^2 is a placeholder for the likelihood such that

$$\chi^2(\theta) = \sum_{i=1}^{n} \left(\frac{y_i^d - y(\theta, t_i)}{\sigma}\right)^2 \tag{10}$$

Here y_i^d denotes the data measured at time point t_i, and $y(\theta, t_i)$ denotes the observable given parameters θ at time point t_i.

Proof. The log-likelihood can be computed as

$$l(\theta) = -\frac{n}{2\log(2\pi)\sigma} + \sum_{i=1}^{n} -\frac{1}{2}\left(\frac{y_i^d - y(\theta, t_i)}{\sigma}\right)^2, \tag{11}$$

$$\Rightarrow \quad \chi^2(\theta) = \frac{n}{\log(2\pi)\sigma} - 2l(\theta)$$

□

Before defining the practical identifiability, we need to introduce one term with respect to confidence intervals.

Definition 3.2. *Likelihood-based confidence intervals* of parameters θ are defined by a confidence interval with a threshold Δ_α

$$\{\theta | \chi^2(\theta) - \chi^2(\hat{\theta}) < \Delta_\alpha\}, \quad \Delta_\alpha = \chi^2(\alpha, df) \tag{12}$$

where $\hat{\theta}$ is defined in Eqn.9, Δ_α is the α quantile of χ^2-distribution and df is degrees of freedom.

We have two choices of df: if $df = 1$, it yields a pointwise confidence interval for each parameter individually; if $df =$ number of parameters in θ, it gives joint confidence intervals for all parameters. Then we can define two types of identifiabilities [7] as follows.

Definition 3.3. If θ denotes the set of all parameters in the model, then the model and θ are said to be *structurally identifiable* if there exists a unique minimum of $\chi^2(\theta)$ with respect to θ_i for each $i = 1, .., n$

Definition 3.4. Suppose $\hat{\theta}_i$ is the estimate of the i-th parameter θ_i. Parameter θ_i is *practically identifiable* if its likelihood-based confidence region (Eqn.12) is finite.

A structurally identifiable parameter is not necessary to be practically identifiable, which manifests when the confidence interval of experimental data is infinite, and the data amount and quality are insufficient.

3.1.3 Identifiability Analysis

We first analyze the structural identifiability of the model. Establishing the structural identifiability test requires formulating equations which describes the statistical moments of the random variable X_t. To derive the system of moment equations of our model, we need to introduce the moment.

Definition 3.5. For a random variable X_t, the k-th moment is given by

$$m^{(k)}(t) = \mathrm{E}\left[X_t^k\right].$$

Specifically for high-dimensional random variable $X_t \in \mathbb{R}^N$, we denote the moment as $m_{i_1,i_2,\ldots,i_N}(t)$ such that

$$m_{i_1,i_2,\ldots,i_N}(t) = \mathrm{E}\left[\prod_{j=1}^{N} X_{j,t}^{i_j}\right], \tag{13}$$

where $J = \sum_{j=1}^{N} i_j$ indicates the order of the moment, $X_{j,t}$ denotes the j-th element of X_t, and i_j denotes the corresponding degree of component $X_{j,t}$.

In the pharmacokinetic model, $X(t) = [P(t), I(t)]' \in \mathbb{R}^2$, so $N = 2$, $X_{1,t} = P(t)$, $X_{2,t} = I(t)$. Hence we only have first moments and second moments. All possible first moments are given by

$$\begin{aligned} m_{10}(t) &= \mathrm{E}\left[X_{1,t}\right] = \mathrm{E}\left[P(t)\right], \\ m_{01}(t) &= \mathrm{E}\left[X_{2,t}\right] = \mathrm{E}\left[I(t)\right]. \end{aligned} \tag{14}$$

And all possible second moments are

$$\begin{aligned} m_{20} &= \mathrm{E}\left[X_{1,t}^2\right] = \mathrm{E}\left[P(t)^2\right], \\ m_{02} &= \mathrm{E}\left[X_{2,t}^2\right] = \mathrm{E}\left[I(t)^2\right], \\ m_{11} &= \mathrm{E}\left[X_{1,t}X_{2,t}\right] = \mathrm{E}\left[P(t)I(t)\right]. \end{aligned} \tag{15}$$

We consider moment dynamics to determine the structural identifiability of the SDE models with polynomials. Moment dynamics [19] are partial differential equations of the moments derived from a master equation that describes the time evolution of distribution.

Definition 3.6. Suppose $m_{i_1,i_2,\ldots,i_N}(t)$ is a vector containing all moments for all states, then the general moment dynamics can be defined in matrix form as

$$\frac{\partial m_{i_1,i_2,\ldots,i_N}(t)}{\partial t} = \boldsymbol{A} m_{i_1,i_2,\ldots,i_N}(t) + \boldsymbol{v},$$

where $\boldsymbol{A}$ is the matrix containing all numeric coefficients related to each moment, and $\boldsymbol{v}$ is a vector of constant terms.

Based on previous computations, only the first and second moments exist for this model. Hence, moment dynamics are derived as equations of the time derivatives of the first and second moments, respectively. To be more practical, the first-order moments correspond to the mean of each $X_{j,t}$, and the second-order moments related to the variance and covariance of each $X_{i,t}$, $X_{j,t}$, $i, j = 1, 2$. To get the explicit form of moment dynamics, we also need to compute time derivatives. We introduce a form for general time derivatives [4] as follows.

Proposition 3.7. For analytical $\boldsymbol{\sigma}$ and $\mathbf{X_t} \in \mathbb{R}^N$, an expression for the time derivative of each moment is given by:

$$\frac{\mathrm{d}m_{i_1,i_2,\dots,i_N}(t)}{\mathrm{d}t} = \mathrm{E}\left[\boldsymbol{b}(\mathbf{X_t},\boldsymbol{\theta}) \cdot \nabla(\prod_{j=1}^{N} X_{j,t}^{i_j})\right] + \frac{1}{2}\,\mathrm{tr}(\boldsymbol{\sigma}^T(\mathbf{X_t},\boldsymbol{\theta})\mathbf{H}(\prod_{j=1}^{N} X_{j,t}^{i_j})\boldsymbol{\sigma}(\mathbf{X_t},\boldsymbol{\theta}))] \tag{16}$$

where $\mathbf{H}(\cdot)$ denotes the $N \times N$ Hessian matrix of its argument and $\nabla = (\frac{\partial}{\partial X_1}, \frac{\partial}{\partial X_2}, \dots, \frac{\partial}{\partial X_N})$ denotes the gradient vector, $\boldsymbol{b}(\cdot)$ and $\boldsymbol{\sigma}(\cdot)$ represent the drift and diffusion term respectively.

Proof. Apply multivariate Itô's Lemma in differential form

$$\mathrm{d}m_{i_1,i_2,\dots,i_N}(t) = \frac{\partial m_{i_1,i_2,\dots,i_N}(t)}{\partial \mathbf{X_t}}\,\mathrm{d}\mathbf{X_t} + \frac{\partial m_{i_1,i_2,\dots,i_N}(t)}{\partial t}\,\mathrm{d}t + \frac{1}{2}\frac{\partial^2 m_{i_1,i_2,\dots,i_N}(t)}{\partial \mathbf{X_t}^2}\,\mathrm{d}\mathbf{X_t}^2$$

Then we plug in the definition of moment

$$m_{i_1,i_2,\dots,i_N}(t) = \mathrm{E}\left[\prod_{j=1}^{N} X_{j,t}^{i_j}\right],$$

and the SDE expressing $\mathbf{X_t}$

$$\mathrm{d}\mathbf{X_t} = \boldsymbol{b}(\mathbf{X_t},\boldsymbol{\theta})\,\mathrm{d}t + \boldsymbol{\sigma}(\mathbf{X_t},\boldsymbol{\theta})\,\mathrm{d}W_t$$

$$\begin{aligned}
\frac{\mathrm{d}m_{i_1,i_2,\dots,i_N}(t)}{\mathrm{d}t} &= \frac{1}{\mathrm{d}t}\left(\mathrm{E}\left[\mathbf{b}(\mathbf{X_t},\boldsymbol{\theta})\,\mathrm{d}t \cdot \nabla\left(\prod_{j=1}^{N} X_{j,t}^{i_j}\right)\right] + 0 + \mathrm{E}\left[\frac{1}{2}\,\mathrm{tr}\left(\mathbf{H}(\prod_{j=1}^{N} X_{j,t}^{i_j})\right)\boldsymbol{\sigma}^T(\mathbf{X_t},\boldsymbol{\theta})\boldsymbol{\sigma}(\mathbf{X_t},\boldsymbol{\theta})\,\mathrm{d}t\right]\right) \\
&= \frac{1}{\mathrm{d}t}\left(\mathrm{E}\left[\mathbf{b}(\mathbf{X_t},\boldsymbol{\theta})\,\mathrm{d}t \cdot \nabla\left(\prod_{j=1}^{N} X_{j,t}^{i_j}\right)\right] + 0 + \mathrm{E}\left[\frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{\sigma}^T(\mathbf{X_t},\boldsymbol{\theta})\mathbf{H}(\prod_{j=1}^{N} X_{j,t}^{i_j})\boldsymbol{\sigma}(\mathbf{X_t},\boldsymbol{\theta})\right)\mathrm{d}t\right]\right) \\
&= \mathrm{E}\left[\boldsymbol{b}(\mathbf{X_t},\boldsymbol{\theta}) \cdot \nabla\left(\prod_{j=1}^{N} X_{j,t}^{i_j}\right) + \frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{\sigma}^T(\mathbf{X_t},\theta)\mathbf{H}(\prod_{j=1}^{N} X_{j,t}^{i_j})\boldsymbol{\sigma}(\mathbf{X_t},\boldsymbol{\theta})\right)\right]
\end{aligned}$$

□

Based on the above proposition, we derive the moment dynamics for the pharmacokinetic model as:

$$\begin{cases}
\frac{\mathrm{d}m_{10}}{\mathrm{d}t} = -(\lambda+\beta)\,m_{10} + k\,m_{10}, \\
\frac{\mathrm{d}m_{01}}{\mathrm{d}t} = \lambda\,m_{10} - k\,m_{01}, \\
\frac{\mathrm{d}m_{20}}{\mathrm{d}t} = -2(\lambda+\beta)m_{20} + 2km_{11} + \sigma_1^2, \\
\frac{\mathrm{d}m_{02}}{\mathrm{d}t} = 2\lambda m_{11} - 2km_{02} + \sigma_2^2, \\
\frac{\mathrm{d}m_{11}}{\mathrm{d}t} = -\beta\,m_{11} + (\lambda+\beta)\,(m_{20}+m_{11}) - k\,(m_{11}-m_{02}),
\end{cases} \tag{17}$$

m_{10} and m_{01} denote the first moments with respect to P_t and I_t, i.e. the mean of P_t and I_t respectively. m_{20}, m_{02}, and m_{11} represent the second moments of P_t and I_t, which are variances of P_t and I_t, and their covariance respectively.

Proof. Given the moments obtained before

$$\begin{cases} m_{10} = \mathrm{E}[P(t)],\ m_{01} = \mathrm{E}[I(t)] \\ m_{20} = \mathrm{E}\left[P(t)^2\right],\ m_{02} = \mathrm{E}\left[I(t)^2\right],\ m_{11} = \mathrm{E}[P(t)I(t)] \end{cases}$$

Then we apply the equation of time derivatives in Eqn.16 to the obtained moments

$$\begin{aligned}
\frac{\mathrm{d}m_{10}}{\mathrm{d}t} &= \mathrm{E}\left[-(\lambda+\beta)\times\frac{\mathrm{d}P(t)}{\mathrm{d}t} + k\times\frac{\mathrm{d}I(t)}{\mathrm{d}t} + \sigma_1^2\times 0\right] = -(\lambda+\beta)m_{10} + km_{01} \\
\frac{\mathrm{d}m_{01}}{\mathrm{d}t} &= \mathrm{E}\left[\lambda\times\frac{\mathrm{d}P(t)}{\mathrm{d}t} - k\times\frac{\mathrm{d}I(t)}{\mathrm{d}t} + \sigma_2^2\times 0\right] = \lambda m_{10} - km_{01} \\
\frac{\mathrm{d}m_{20}}{\mathrm{d}t} &= \begin{bmatrix}-(\lambda+\beta) & k\end{bmatrix}\mathrm{E}\left[\begin{bmatrix}\dfrac{\partial P(t)^2}{\partial P(t)} \\ \dfrac{\partial 2P(t)I(t)}{\partial P(t)}\end{bmatrix}\frac{\mathrm{d}P(t)}{\mathrm{d}t}\right] + \sigma_1^2 = -2(\lambda+\beta)m_{20} + 2km_{11} + \sigma_1^2 \\
\frac{\mathrm{d}m_{02}}{\mathrm{d}t} &= \begin{bmatrix}\lambda & -k\end{bmatrix}\mathrm{E}\left[\begin{bmatrix}\dfrac{\partial 2P(t)I(t)}{\partial I(t)} \\ \dfrac{\partial I(t)^2}{\partial I(t)}\end{bmatrix}\frac{\mathrm{d}I(t)}{\mathrm{d}t}\right] + \sigma_2^2 = -2km_{02} + 2\lambda m_{11} + \sigma_2^2 \\
\frac{\mathrm{d}m_{11}}{\mathrm{d}t} &= \begin{bmatrix}-(\lambda+\beta) & k\end{bmatrix}\mathrm{E}\left[\begin{bmatrix}\dfrac{\partial - P(t)^2 + P(t)I(t)}{\partial P(t)} \\ \dfrac{\partial P(t)I(t)}{\partial P(t)}\end{bmatrix}\frac{\mathrm{d}P(t)}{\mathrm{d}t}\right] + \begin{bmatrix}\lambda & -k\end{bmatrix}\mathrm{E}\left[\begin{bmatrix}\dfrac{\partial P(t)I(t)}{\partial I(t)} \\ \dfrac{\partial - I(t)^2}{\partial I(t)}\end{bmatrix}\frac{\mathrm{d}I(t)}{\mathrm{d}t}\right] + \sigma_1\sigma_2 \\
&= [-(\lambda+\beta) + k + \lambda - k]m_{11} + (\lambda+\beta)m_{20} + km_{02} + \sigma_1\sigma_2
\end{aligned}$$

□

Then we introduce two characteristics used to describe the moment dynamics [4].

Definition 3.8. The system of the time derivatives of moment equations is said to *exactly* describe the time evolution of moments when each component of $\boldsymbol{b}$ and $\boldsymbol{\sigma\sigma}^T$ are polynomials in $\mathbf{X_t}$.

Definition 3.9. The system is said to be *closed* at order J if the expression of each moment depends only on the moments up to order J. We can truncate the system at order J and solve the exact moments directly.

Proposition 3.10. If the system of moment equations is both *closed* and *exact*, then parameters in this system are structurally identifiable.

For both continuous-time observations and discrete-time observations, identifiability analysis applies according to [4]. Because of the obtained moment dynamics (Eqn.17), the pharmacokinetic model is considered *closed and exact*. Hence, we can conclude that the model is structurally identifiable. Therefore, the parameter estimation of each parameter is reasonable.

We later combine maximum likelihood estimators (MLEs) and compute the likelihood-based confidence intervals corresponding to each parameter to analyze the practical identifiability.

If we find some parameters are practically unidentifiable, we can still estimate their values under specific conditions.

Definition 3.11. If the parameter has a detectable lower bound and is distinguished from zero, it is said to be *one-sided identifiable*.

Assume there exist some practically non-identifiable parameters. These parameters present some distinguishable trends when observed, for example, a monotonic increase. Then they are one-sided identifiable, and we can still estimate their lower bounds in this case.

3.2 Parameter estimation by maximum likelihood

Previous research has been done to estimate parameters for this partially observed Ornstein-Uhlenbeck process using maximization of the likelihood based on Kalman filtering.

3.2.1 Model transformation

We transform the model and derive a discrete-time system of ordinary differential equations via reparametrization to study the stochastic differential equations and compute the likelihood. A new matrix $U(t) = [S(t), I(t)]'$ is introduced to obtain a transformed model in matrix form:

$$\begin{cases} \mathrm{d}U(t) = \left(\begin{bmatrix} \alpha\,\delta(t) \\ 0 \end{bmatrix} + \begin{bmatrix} -\beta & \beta \\ \lambda & -k \end{bmatrix} U(t) \right) \mathrm{d}t + \begin{bmatrix} \sigma_1 & \sigma_2 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} dW_1(t) \\ dW_2(t) \end{bmatrix}, \\ y_i = JU(t_i) + \sigma\epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \end{cases} \tag{18}$$

where $J = [1, 0]$, and σ is the standard deviation of the Gaussian noise. The process $(U(t))$ is a bidimensional Ornstein-Uhlenbeck diffusion, which can be explicitly solved. Let $G = \begin{bmatrix} -\beta & \beta \\ \lambda & -k \end{bmatrix}$ and $\Gamma = \begin{bmatrix} \sigma_1 & \sigma_2 \\ 0 & \sigma_2 \end{bmatrix}$. Since all parameters are positive, the following claim holds.

Proposition 3.12. The matrix G is diagonalisable with two distinct negative eigenvalues μ_1, μ_2, with

$$d = (\beta - k)^2 + 4\beta\lambda \tag{19}$$

$$\mu_1 = \frac{-(\beta + k) - \sqrt{d}}{2} \tag{20}$$

$$\mu_2 = \frac{-(\beta + k) + \sqrt{d}}{2}. \tag{21}$$

Proof. Consider the characteristic function$f(x)$ of G,

$$\begin{aligned} f(x) &= (x + \beta)(x + k) - \beta\lambda \\ &= x^2 + (\beta + k)x + \beta(k - \lambda) \end{aligned}$$

As $\Delta = (\beta + k)^2 - 4\beta(k - \lambda) = (\beta - k)^2 + 4\beta\lambda > 0$, we can conclude that there exists two distinct eigenvalues of G, which can be computed as

$$\begin{aligned} \mu_{1,2} &= \frac{-(\beta + k) \pm \sqrt{\Delta}}{2} \\ &= \frac{-(\beta + k) \pm \sqrt{(\beta - k)^2 + 4\beta\lambda}}{2} \end{aligned}$$

$$0 < (\beta - k)^2 + 4\beta\lambda < (\beta - k)^2 + 4\beta k = (\beta + k)^2 \;\Rightarrow\; \mu_{1,2} < 0$$

□

Assumption 3.13. Assume we have observations $y_0, y_1, \dots, y_n$ at time points $0 = t_0 < t_1 < \cdots < t_n = T$. Let $y_{0:n} = (y_0, y_1, \dots, y_n)$. Define a diagonal matrix of eigenvalues and eigenvectors using Eqn.19-21 as D, P:

$$D = \begin{bmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{bmatrix}, P = \begin{bmatrix} 1 & 1 \\ \dfrac{\beta - k - \sqrt{d}}{2\beta} & \dfrac{\beta - k + \sqrt{d}}{2\beta} \end{bmatrix}.$$

Suppose $X(t) = P^{-1}U(t)$ and define the following:

$$\begin{cases} X(t+h) = e^{Dh}X(t) + B(t, t+h) + Z(t, t+h), \\ B(t, t+h) = e^{D(t+h)} \displaystyle\int_t^{t+h} e^{-Ds} P^{-1} F(s)\, ds \\ Z(t, t+h) = e^{D(t+h)} \displaystyle\int_t^{t+h} e^{-Ds} P^{-1} \Gamma\, dW_s \end{cases} \tag{22}$$

Proposition 3.14. For $t, h \geq 0$, we can derive a discrete difference equation:

$$X(t+h) = e^{Dh}X(t) + B(t, t+h) + Z(t, t+h).$$

Then, for $s \leq t$, we have

$$X(t+h)|X(s) \sim \mathcal{N}(e^{Dh}X(t) + B(t, t+h), R(t, t+h)),$$

where

$$R(t, t+h) = \left(\frac{e^{(\mu_k + \mu_l)h} - 1}{\mu_k + \mu_l} (P^{-1}\Gamma\Gamma' P^{-T})^{kl} \right)_{1 \leq\, k,l\, \leq 2}$$

If $\delta(t)$ is a constant, $(X(t))$ has a stationary Gaussian distribution.

Combining the above, we replace $(U(t))$ in the matrix form by $X(t)$ and deduce the discrete differential equations. Then the model can be transformed to the following:

$$\begin{cases} X_i = A_i X_{i-1} + B_i + \eta_i, \quad \eta_i \sim \mathcal{N}(0, R_i), \\ y_i = H X_i + \sigma_i \end{cases} \tag{23}$$

where $H = [1, 1]$. $X_i = X(t_i)$, $A_i = e^{D(t_i - t_{i-1})}$, $B_i = B(t_{i-1}, t_i)$, $R_i = R(t_{i-1}, t_i)$.

3.2.2 Likelihood Maximization

Because of the Gaussian law of $(X(t), \epsilon_i)$, maximizing the exact likelihood directly is feasible. But it requires the inversion of the covariance matrix of $(X(t_i))_{i=0}^n$ of dimension $2(n+1) \times 2(n+1)$, and the inversion can be numerically unstable, which leads to the difficulty and high-cost of computation. Therefore, we first compute the exact likelihood based on Kalman filtering. Then an alternative method to calculate the Maximum Likelihood Estimator (MLE) using the conjugate gradient is proposed.

We can compute the exact likelihood based on the Kalman filter as the data is one-dimensional. Assuming the initial value are $X_0 = [0, 0]'$, we have the following recursive formulae applying the Kalman filter to the above discrete-time equations:

$$\hat{X}_i^- = A_i \hat{X}_{i-1}^- + B_i, \quad P_i^- = A_i P_{i-1} A_i' + R_i, \quad i \geq 1,$$

$$\hat{X}_i = \hat{X}_i^- + K_i(y_i - H\hat{X}_i^-), P_i = (I - K_i H)P_i^-, \quad i \geq 0,$$

where $K_i = P_i^- H'(HP_i^- H' + \sigma^2)^{-1}$. Let $\varphi = (\theta, \sigma^2)$, and denote the mean and variance of the conditional distribution $y_i | y_{0:i-1}$ as $m_i(\varphi)$, and $V_i(\varphi)$, which are computed using the above recursive formulae by:

$$m_i(\varphi) = H\hat{X}_i^-, \quad V_i(\varphi) = HP_i^- H' + \sigma^2.$$

Then the exact likelihood of $y_{0:n}$ is given by:

$$L(\varphi, y_{0:n}) = \prod_{i=0}^{n} \frac{1}{\sqrt{2\pi V_i(\varphi)}} exp(-\frac{(y_i - m_i^-(\varphi))^2}{2V_i(\varphi)}). \tag{24}$$

The MLE is computed using a conjugate gradient method by computing the exact gradient and Hessian matrix of the log-likelihood. We use a new parametrization to simplify the computation.

Assumption 3.15. Assume $\delta(t) = c \geq 0$ is a known constant. Let the observation times are equally spaced with the time step $\Delta = t_i - t_{i-1}$, for $i = 1, \ldots, n$. Set $Z_i = X_i - M$, $m = HM$, then we have

$$A = A_i, R_i = R, B_i = B = (I - A)D^{-1}P^{-1}F = (I - A)M.$$

Then we have the following simplified discrete-time system:

$$\begin{cases} Z_i = A_i Z_{i-1} + \eta_i, & \eta_i \sim \mathcal{N}(0, R_i), \\ y_i = HZ_i + m + \sigma\epsilon_i & \end{cases} \tag{25}$$

and the exact likelihood of $y_{0:n}$:

$$L(\varphi, y_{0:n}) = \prod_{i=0}^{n} \frac{1}{\sqrt{2\pi V_i(\varphi)}} exp(-\frac{(y_i - m - H\hat{Z}_i^-(\varphi))^2}{2V_i(\varphi)}).$$

So, instead of estimating $\theta = (\alpha, \beta, \lambda, k, \sigma_1, \sigma_2)$ directly, we propose a new parametrization such that

$$\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)$$

$$A = \begin{pmatrix} \theta_1 & 0 \\ 0 & \theta_2 \end{pmatrix}, R = \begin{pmatrix} \theta_3 & \theta_5 \\ \theta_5 & \theta_4 \end{pmatrix}$$

The transformation between new parametrization and original parameters is given as follows.

$$\begin{cases} \theta_1 = e^{\mu_1 \Delta}, \quad \theta_2 = e^{\mu_2 \Delta}, \\ \theta_3 = \dfrac{(\mu_2 + \beta)^2 \sigma_1^2 + \mu_2^2 \sigma_2^2}{\dfrac{2\mu_1 d}{\theta_1^2 - 1}}, \\ \theta_4 = \dfrac{(\mu_1 + \beta)^2 \sigma_1^2 + \mu_1^2 \sigma_2^2}{\dfrac{2\mu_2 d}{\theta_2^2 - 1}}, \\ \theta_5 = \dfrac{\sigma_1 \beta k - \mu_1 \mu_2 (\sigma_1^2 + \sigma_2^2)}{\dfrac{(\mu_1 + \mu_2)(\mu_2 - \mu_1)^2}{e^{\Delta(\mu_1 + \mu_2)} - 1}}, \\ \theta_6 = m \end{cases} \tag{26}$$

To maximize the likelihood $L(\varphi, y_{0:n})$ with respect to φ, we need to derive the gradient and Hessian matrix with respect to φ. The log-likelihood is given by

$$l_{0:i}(\varphi) = l_{0:i-1}(\varphi) - \frac{1}{2} \log(2_i(\varphi)) - \frac{1}{2} \frac{W_i(\varphi)^2}{V_i(\varphi)}$$

and the gradient is

$$\frac{\partial l_{0:i}}{\partial \varphi_q}(\varphi) = \frac{\partial l_{0:i-1}}{\partial \varphi_q}(\varphi) - \frac{1}{2} \frac{1}{V_i(\varphi)} \frac{\partial V_i}{\partial \varphi_q}(\varphi) - \frac{W_i(\varphi)}{V_i(\varphi)} \frac{\partial W_i(\varphi)}{\partial \varphi_q} + \frac{1}{2} \frac{W_i(\varphi)^2}{V_i(\varphi)^2} \frac{\partial V_i(\varphi)}{\partial \varphi_q}$$

The computation of Hessian matrix is rather complicated, which is shown as

$$\begin{aligned} \frac{\partial^2 l_{0:i}}{\partial \varphi_r \partial \varphi_q}(\varphi) = \frac{\partial^2 l_{0:i-1}}{\partial \varphi_r \partial \varphi_q}(\varphi) &- \frac{1}{2} \frac{1}{V_i(\varphi)} \frac{\partial^2 V_i}{\partial \varphi_r \partial \varphi_q}(\varphi) + \frac{1}{2} \frac{1}{V_i^2(\varphi)} \frac{\partial V_i}{\partial \varphi_r}(\varphi) \frac{\partial V_i}{\partial \varphi_q}(\varphi) \\ &- \frac{1}{2} \left(2 \frac{W_i(\varphi)}{V_i(\varphi)} \frac{\partial^2 W_i(\varphi)}{\partial \varphi_r \partial \varphi_q} - \frac{W_i(\varphi)^2}{V_i(\varphi)^2} \frac{\partial^2 V_i(\varphi)}{\partial \varphi_r \partial \varphi_q} \right) \\ &- \left(\frac{\partial W_i(\varphi)}{\partial \varphi_r} \frac{\partial W_j(\varphi)}{\partial \varphi_q} \frac{1}{V_i(\varphi)} - \frac{W_i(\varphi)}{V_i(\varphi)^2} \frac{\partial W_i(\varphi)}{\partial \varphi_r} \frac{\partial V_i(\varphi)}{\partial \varphi_q} \right) \\ &+ \left(\frac{\partial W_i(\varphi)}{\partial \varphi_q} \frac{\partial V_i(\varphi)}{\partial \varphi_r} \frac{W_i(\varphi)}{V_i(\varphi)^2} - \frac{W_i(\varphi)^2}{V_i(\varphi)^4} \frac{\partial V_i(\varphi)}{\partial \varphi_r} V_i(\varphi) \frac{\partial V_i(\varphi)}{\partial \varphi_q} \right) \end{aligned} \tag{27}$$

After obtaining the exact form of likelihood and Hessian matrix, a numerical algorithm (Algorithm.3) is applied to achieve the maximum likelihood. The conjugate gradient method is chosen for this study to update the parameters and descent directions.

Algorithm 3: Maximizing the Likelihood using Conjugate Gradient

```
1. Initialize:
   Let $\varphi_0$ denote the initial value and set the descent direction as $u_0 = \varphi_0$.
2. Iterate (for $k$-th iteration):
   given $\varphi_k$ and $u_k$, update the parameter by
       $\varphi_{k+1} = \varphi_k - \frac{\langle u_k,\ \nabla l^-(\varphi_k)\rangle}{\langle u_k,\ Hess_l^-(\varphi_k)u_k\rangle} u_k,$
   update the descent direction by
       $u_{k+1} = -\nabla l^-(\varphi_{k+1}) + \frac{\|\nabla l^-(\varphi_{k+1})\|}{\|\nabla l^-(\varphi_k)\|} u_k,$
   if $\|\nabla l^-(\varphi_k)\| \geq 0.01$, and $\|\varphi_{k+1} - \varphi_k\| \geq 0.01$, iteration terminates.
```

3.3 Bayesian Method For Parameter Estimation.

The maximum likelihood method has a limitation that it cannot estimate all the parameters simultaneously. Another drawback is that due to the need of computation of the maximum likelihood, the time consumption is high for high-dimensional models [1]. Therefore, we introduce an alternative Bayesian estimation method for joint estimation. This Bayesian method involves a novel Backward Filtering with Forward Guiding (BFFG) algorithm, a likelihood-based estimation scheme proposed in [3].

3.3.1 Overview and Preliminaries

The whole parameter estimation process can be split into two parts, the backward filtering and the iterative part [3]. We define the diffusion process for backward filtering, and the guided proposal [2]. Then we discuss the choice of the auxiliary process and a corresponding auxiliary observation scheme [8].

Assume the stochastic process is partially observed on discrete time points $0 = t_0 < t_1 < ... < t_n = T$. In this thesis, we consider the simplest case that the observations satisfy the following conditional distribution:

$$V_i | X_{t_i} \sim k_i(X_{t_i}) = \mathcal{N}(LX_{t_i}, \Sigma), \tag{28}$$

where k_i denotes the conditional density and $X_{t_i} \in \mathbb{R}^2$. $k_i(X_{t_i})$ represent the density of $\mathcal{N}(LX_{t_i}, \Sigma)$-distribution, with $L = [1, 1], \Sigma = 0.1$. For $t \geq 0$, denote the set of non-past observations by $\mathcal{V}_t$, given by

$$\mathcal{V}_t = \{V_i : t_i \geq t\}, \quad \mathcal{V}_i = \mathcal{V}_{t_i}.$$

Definition 3.16. Suppose at the end time T, the conditioned diffusion $V_i | X_{t_i}$ is fixed at an endpoint $v \in \mathbb{R}^2$. Define p as the transition densities of the original diffusion process X. Further assume that for $s < \tau$,

$$P^{(s,\tau)}(X_\tau \in \mathrm{d}y) = p(s,x;\tau,y).$$

Definition 3.17. For $i \in \{1,\dots,n\}$, $t \in (t_{i-1}, t_i]$, let $X_{t_{i-1}} = x_{t_{i-1}}$, and $V_i,\dots,V_n$ are given. Let $\rho(t_i, x_i)$ denote the likelihood of x_i given the present and future observation $\mathcal{V}_i$, i.e.

$$\rho(t_i, x_i) = \pi(\mathcal{V}_i | x_i).$$

Then we can give a formula of ρ by integrating out the latent future states

$$\begin{aligned} \rho(t,x) &= \int p(t,x;t_i,\varepsilon_i) k_n(\varepsilon_n) \prod_{j=i}^{n-1} p(t_j,\varepsilon_j;t_{j+1},\varepsilon_{j+1}) k_j(\varepsilon_j)\,\mathrm{d}\varepsilon_i \dots \mathrm{d}\varepsilon_n \\ \rho(0,x) &= k_0(x)\rho(0+,x) \end{aligned} \tag{29}$$

This reveals that ρ is related to a backward filtered marginal density of X.

We employ the matrix form of the model to describe the diffusion dynamics:

$$\mathrm{d}X_t = \boldsymbol{b}(X_t,t)\,\mathrm{d}t + \boldsymbol{\sigma}\,\mathrm{d}W_t \tag{30}$$

$$X_t = \begin{bmatrix} P(t) \\ I(t) \end{bmatrix},\ \boldsymbol{b}(X_t,t) = \begin{bmatrix} \alpha\delta(t) \\ 0 \end{bmatrix} + \begin{bmatrix} -(\beta+\lambda) & k \\ \lambda & -k \end{bmatrix} X_t,\ \boldsymbol{\sigma} = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix},\ W_t = \begin{bmatrix} W_1(t) \\ W_2(t) \end{bmatrix},$$

$\boldsymbol{b}$ and $\boldsymbol{\sigma}$ represent the drift and diffusion respectively.

Proposition 3.18. Let the ratio of likelihoods be a weight w_i,

$$w_i = \frac{\pi(\mathcal{V}_i|x_i)}{\pi(\mathcal{V}_i|x_{i-1})} = \exp\log\pi(\mathcal{V}_i|x_i) - \log\pi(\mathcal{V}_i|x_{i-1}),$$

then the weight equals an exponential process [3]:

$$w_i = \exp\int_{t_{i-1}}^{t_i} \boldsymbol{\sigma}(s,X_s)'\nabla\log\rho(s,X_s)\,\mathrm{d}W_s - \frac{1}{2}\int_{t_{i-1}}^{t_i} \|\boldsymbol{\sigma}(s,X_s)'\nabla\log\rho(s,X_s)\|^2\,\mathrm{d}s$$

Therefore, we can sample from the smoothing distribution of X_i by simulating the process X with drift $\boldsymbol{b} + \boldsymbol{\sigma}\boldsymbol{\sigma}'\nabla\rho$ via

$$\mathrm{d}X_t = (\boldsymbol{b} + \boldsymbol{\sigma}\boldsymbol{\sigma}'\nabla\rho)\,\mathrm{d}t + \boldsymbol{\sigma}\,\mathrm{d}W_t$$

instead of sampling $X_i|X_{i-1},\mathcal{V}_i$ directly [3].

3.3.2 Linear Guided Proposals and Backward Filtering

From the above, the likelihood ρ is required to sample from the smoothing distribution. Since computing the function ρ is inefficient, we define a proxy $\tilde{\rho}$ to ρ and introduce guided proposals.

Definition 3.19. Let $\tilde{k}_{n+}(x)$ be the density of distribution $\mathcal{N}(0, P_{n+})$, where P_{n+} is a given (strictly positive definite) covariance matrix. Similar to ρ (defined in Eqn.29), we define $\tilde{\rho}$ by

$$\tilde{\rho}(t,x) = \int \tilde{k}_{n+}(\varepsilon_n)\tilde{p}(t,x;t_i,\varepsilon_i)\tilde{k}_n(\varepsilon_n,v_n)\prod_{j=i}^{n-1}\tilde{p}_j(t_j,\varepsilon_j;t_{j+1},\varepsilon_{j+1})\tilde{k}_j(\varepsilon_j;v_j)\,\mathrm{d}\varepsilon_i \dots \mathrm{d}\varepsilon_n$$

where $\tilde{p}$ is the transition density of a chosen auxiliary process $\tilde{X}$, and $\tilde{k}_i$ is the density of the auxiliary observation scheme.

Definition 3.20. A *guided proposal process* X^o is the solution to the SDE

$$\mathrm{d}X_t^o = \boldsymbol{b}(t, X_t^o)\,\mathrm{d}t + \boldsymbol{a}\tilde{r}(t, X_t^o)\,\mathrm{d}t + \boldsymbol{\sigma}\,\mathrm{d}W_t, \quad X_{t_i}^o = x_{t_i},$$

where $\boldsymbol{a} = \boldsymbol{\sigma}\boldsymbol{\sigma}\prime$, $\tilde{r}(t, x) = \nabla_x \log \tilde{p}(t, x)$.

We aim to sample from the guided proposal X^o, so the explicit form of $\tilde{p}$ is required. Therefore, we need to determine the auxiliary process and auxiliary observation scheme so that $\{\tilde{k}_i\}$ and $\tilde{X}$ well approximates $\{k_i\}$ and X. We first give the choice of the auxiliary observation scheme.

Definition 3.21. The auxiliary observation scheme $\tilde{k}_i$ is the density of the diffusion with

$$\tilde{k}_i(X_{t_i}) = \mathcal{N}(L_i X_{t_i}, \Sigma_i),$$

such that

$$V_i | X_{t_i} \sim k_i(X_{t_i}) = \mathcal{N}(L X_{t_i}, \Sigma) \approx \mathcal{N}(L_i X_{t_i}, \Sigma_i),$$

where L_i is a 1×2 matrix, $X_{t_i} \in \mathbb{R}^2$, Σ_i a positive number.

Remark 3.22. Due to the specific density of k_i in this thesis, we can choose L_i, Σ_i such that

$$\forall i = 0, \dots, n, \quad L_i = L, \quad \Sigma_i = \Sigma.$$

Thereafter, we decide the choice of the auxiliary process $\tilde{X}$, which gives the values of $\tilde{p}$.

Definition 3.23. An auxiliary process $\tilde{X}$ is defined by the following SDE which can well approximate the law of the target process X

$$\mathrm{d}\tilde{X}_t = (\beta(t) + B(t)\tilde{X}_t)\,\mathrm{d}t + \tilde{\sigma}(t)\,\mathrm{d}W_t.$$

Remark 3.24. As the drift is linear in X_t and diffusion is independent of X_t, we can simply choose a linear guided proposal for the auxiliary diffusion [8] $\tilde{X}_t$:

$$\beta(t) = \begin{bmatrix} -(\lambda + \beta) & \beta \\ \lambda & -k \end{bmatrix}, \quad B(t) = \begin{bmatrix} \alpha\ \delta(t) \\ 0 \end{bmatrix}.$$

Proof. Apply the choice of the combine approach in [8] by approximating $\boldsymbol{b}(t, X_t)$ with

$$\boldsymbol{b}(t, x(t)) + V(t, x(t))(X_t - x(t))$$

and $\tilde{\sigma} = \sigma$. Assume $V(t, y)_{i,j} = \dfrac{\partial b_i(t, y)}{\partial y_j}$ for $y \in \mathbb{R}^n$.

$$\boldsymbol{b}(t, x(t)) = \begin{bmatrix} -(\lambda + \beta) & \beta \\ \lambda & -k \end{bmatrix} x(t) + \begin{bmatrix} \alpha\ \delta(t) \\ 0 \end{bmatrix},$$

$$V(t, y) = \begin{bmatrix} -(\lambda + \beta)y_1 + \beta y_2 + \alpha\delta(t) \\ \lambda y_1 - k y_2 \end{bmatrix}.$$

Hence, we get

$$\begin{cases} B(t) = V(t, x(t)) = \begin{bmatrix} -(\lambda + \beta) & \beta \\ \lambda & -k \end{bmatrix} \\ \beta(t) = \boldsymbol{b}(t, x(t)) - V(t, x(t))x(t) = \begin{bmatrix} \alpha\ \delta(t) \\ 0 \end{bmatrix} \end{cases}$$

□

An ODE-system for backward filtering (information filter) [3] is used for efficiently computing the terms $\tilde{r}$, H and $\tilde{\rho}$, which are in need of implementing the guided proposals. We solve the factors for computing the likelihood from the following "HFC" recursions (Eqn.31).

Theorem 3.25. For $(t, x) \in [0, T] \times \mathbb{R}^2$, We solve $\tilde{\rho}$ by decomposing it into:

$$\log \tilde{\rho}(t, x) = -c(t) - \frac{1}{2} x' H(t) x + F(t)' x,$$

where on each interval $(t_{i-1}, t_i]$, H, F and c are solutions of the backward ODEs

$$\begin{aligned} \mathrm{d}H(t) &= (-B(t)' H(t) - H(t) B(t) + H(t) \boldsymbol{\sigma}\boldsymbol{\sigma}' H(t))\, \mathrm{d}t, \\ \mathrm{d}F(t) &= (-B(t)' F(t) - H(t) \boldsymbol{\sigma}\boldsymbol{\sigma}' F(t) + H(t) \beta(t))\, \mathrm{d}t, \\ \mathrm{d}c(t) &= (\beta(t)' F(t) + \frac{1}{2} F(t)' \boldsymbol{\sigma}\boldsymbol{\sigma}' F(t) - \frac{1}{2} \operatorname{tr}(H(t) \boldsymbol{\sigma}\boldsymbol{\sigma}'))\, \mathrm{d}t \end{aligned} \tag{31}$$

3.3.3 Backward Filtering Forward Guiding algorithm for Parameter Estimation

In this part, a Metropolis-Hastings-within-Gibbs sampler algorithm for parameter estimation is introduced, namely the Backward Filtering Forward Guiding (BFFG) algorithm. Firstly, we give the explicit joint posterior distribution to sample from in this algorithm.

Assume we can obtain guided proposals from above, then there exists a measurable map GP_θ defining the guided proposal

$$X = GP_\theta(X_0, Z),$$

given initial state X_0 and parameters θ, where Z is a Wiener process in $\mathbb{R}^2$. Assume

$$\boldsymbol{b}, \boldsymbol{\sigma}, \tilde{p}, \{\tilde{k}_i\}, \{k_i\}, \pi(x_0)$$

depend on parameters θ which admits a prior $\kappa(\theta)$. Then we can sample the latent path from samples (θ, x_0, Z) through $X = GP_\theta(X_0, Z)$ iteratively. The density of the joint posterior distribution of (θ, x_0, Z) has the specific form

$$\frac{\kappa(\theta)\pi_{(}x_0)\tilde{\rho}(0, x_0)}{\int \kappa(\theta)\pi_{(}x_0)\rho(0, x_0)\, \mathrm{d}(x_0, \theta)} \Psi(GP_\theta(x_0, Z)) \prod_{i=0}^{n} C_i(GP_\theta(x_0, Z)_{t_i}),$$

where

$$\Psi(X^o) = \exp \int_0^{t_n} G(s, X^o_s)\, \mathrm{d}s, \tag{32}$$

$$C_i(x) = \begin{cases} \dfrac{k_i(x)}{\tilde{k}_i(x)} & 1 \leq i \leq n-1 \\ \dfrac{k_n(x)}{\tilde{k}_{n+}(x)} & i = n \end{cases} \tag{33}$$

and

$$G(s, x) = \left(b(s, x) - \tilde{b}(s, x) \right)' \tilde{r}(s, x) - \frac{1}{2} \operatorname{tr} \left([\boldsymbol{a} - \tilde{\boldsymbol{a}}][H(s) - \tilde{r}(s, x) \tilde{r}(s, x)'] \right),$$

which are all proved in [3].

Then we focus on the initialization of the BFFG algorithm (Algorithm.4). Assume the initial parameter value θ and initial state X_0 are known. Fix their values by θ, and x_0, respectively. Before conducting parameter inference, smoothing needs to be undertaken on the diffusion process to obtain samples under the target law. Then we choose a tuning parameter γ and smooth the distribution.

Algorithm 4: Smoothing step using Preconditioned Crank-Nicolson for fixed θ and x

```
1. Choose a tuning parameter $\gamma \in [0, 1)$. Suppose the current state is $(\theta, x_0, Z)$ and
   $X = GP_\theta(x_0, Z)$.
2. Initialise $H(t_{n+}), F(t_{n+}), c(t_{n+})$.
3. Solve the backward ODEs and compute $H(t_i), F(t_i), c(t_i)$.
4. Sample an independent Wiener process W ans set $Z^o = \gamma Z + \sqrt{1-\gamma^2}W$. Compute
   $X^o = GP_\theta(x_0, Z^o)$
5. Compute $A = \frac{\Psi(X^o)}{\Psi(X)} \prod_{i=1}^{n} \frac{C_i(X^o_{t_i})}{C_i(X_{t_i})}$.
6. Draw $U \sim Uniform\,(0, 1)$. If $U < A$, replace $X = X^o$ and $Z = Z^o$.
```

In the iterative step for updating the initial state, path, and parameters, we define a Markov transition kernel and generate the candidate state X^o and parameters θ^o. Then the guiding terms are recomputed, and the corresponding guided proposal is computed. Similarly, the following Metropolis–Hastings-within-Gibbs algorithm (Algorithm.5) is used to determine whether to update the path and parameters or to keep the terms unchanged in the next iteration.

Algorithm 5: Updating the path, initial state and parameters in BFFG algorithm

```
1. Suppose the current iterate is $(\theta, x_0, Z)$, and $X = GP_\theta(x_0, Z)$.
2. Choose a Markov kernel $q$ and compute a new state $x_0^o \sim q_\theta(x_0)$
3. Compute the corresponding guided proposal $X^o = GP_\theta(x_0^o, Z)$.
4. Apply MH-algorithm to decide whether or not to update the path and initial state by
   computing
   $A = \frac{q_\theta(x_0|x_0^o)\pi(x_0^o)\tilde{\rho}(0, x_0^o)\Psi(X^o)}{q_\theta(x_0^o|x_0)\pi(x_0)\tilde{\rho}(0, x_0)\Psi(X)} \prod_{i=1}^{n} \frac{C_i(X_{t_i}^o)}{C_i(X_{t_i})},$
5. Draw $U \sim Uniform\ (0, 1)$. If $U < A$, update $X = X^o$ and $x_0 = x_0^o$. Otherwise, keep the
   current iterate unchanged.
6. Choose a Markov kernel $\tilde{q}$ and propose a new parameter $\theta^o$ by $\theta^o \sim \tilde{q}_\theta(\theta)$.
7. Solve the backward ODEs and recompute $H(t_i), F(t_i), c(t_i)$ with $\theta^o$.
8. Compute the corresponding guided proposal $X^o = GP_{\theta^o}(x_0, Z)$.
9. Apply MH-algorithm to decide whether or not to update the path and parameters by
   computing
   $A = \frac{\tilde{q}(\theta|\theta^o)\kappa(\theta^o)\tilde{\rho}_{\theta^o}(0, x_0^o)\Psi_{\theta^o}(X^o)}{\tilde{q}(\theta^o|\theta)\kappa(\theta)\tilde{\rho}_\theta(0, x_0)\Psi_\theta(X)} \prod_{i=1}^{n} \frac{C_i(X_{t_i}^o)}{C_i(X_{t_i})},$
10. Draw $V \sim Uniform\ (0, 1)$. If $V < A$, update $X = X^o$ and $\theta = \theta^o$. Otherwise, keep the
    current iterate unchanged.
```

3.3.4 Reparametrization for Joint Estimation

After explaining the parameter estimation scheme for a single parameter, we now attempt to accomplish joint estimation for multiple parameters. We sparkle from the maximum likelihood method and try reparametrization. Considering the biological background of the model, we call back the real meaning of each parameter:

$$\alpha = \frac{F_{tp}}{1-h}, \quad \beta = \frac{F_{tp}}{V_P}, \quad \lambda = \frac{K_{trans}}{V_P}, \quad k = \frac{K_{trans}}{V_P} + \frac{K_{trans}}{V_I}.$$

h is the hematocrit rate, and $1 - h$ is the volume of the artery. V_P and V_I denote the volume of plasma and interstitium, respectively. F_{tp} is the perfusion flow, and K_{trans} is the volume transfer constant.

Firstly, considering estimating two parameters at a time, it is reasonable to estimate the ratios of parameter pairs instead of estimating them separately. Since flows in compartments interact with each other and the direction of flow exchanges in and out of compartments are fixed, the ratios of different parameters in drift terms make sense. So, we can choose to compute some ratios, for instance,

$$\frac{\beta}{\alpha} = \frac{V_P}{1-h}, \quad \frac{\lambda}{k-\lambda} = \frac{V_I}{V_P} \quad \frac{\beta}{\alpha} + \frac{\lambda}{k-\lambda} = \frac{V_I + 1 - h}{V_P}, \tag{34}$$

and each of them should be a constant equal to the ratio of volumes of different compartments. Therefore, we use the above ratios for joint estimation and their sums. As the parameter α is the coefficient of the known function $\delta(t)$, it has no relation to the flows, and we can separate it from other parameters by isolated estimation. So, estimating the sum of the ratio above is also a good choice of reparametrization for three and four parameters. So, we attempt to perform joint estimation of all parameters, and try to estimate the above ratios if the parameter itself cannot have a good estimate.

3.4 Optimal Tuning Proposals

As explained in the last section, we employ the Metropolis-Hastings-within-Gibbs sampler in the Bayesian inference algorithm. Indeed, the Metropolis-Hastings algorithm requires an appropriate choice of proposal distributions [5]. When performing parameter inference using the BFFG algorithm in a Bayesian setup, there is an essential step to determine whether or not to update the path and parameters. Determining the choice of accepting the proposed moves or not relates to the tuning parameter γ, which is the memory parameter appearing in the preconditioned Crank–Nicolson step. The efficiency of the Bayesian estimation approach strongly depends on the choice of γ as it strongly affects the acceptance rate. Therefore, we attempt to apply optimal proposal scaling and adaptive algorithms to find good proposals automatically for the proposal of the tuning parameter γ instead of choosing it as an arbitrary value.

We can assign the memory parameter and the proposal kernel random values and get their appropriate ones by trial and error. However, it becomes particularly challenging when the data set is large and the number of iterations grows (as we need a correct value at each iteration). When we perform parameter inference on top of smoothing, different local values of parameters might have drastically different optimal proposals. To alleviate these problems, we employ a scaling method for adaptive MCMC algorithms and attempt to automatically find out the optimal values of the proposal kernel in the updating scheme and the memory parameter of the pCN scheme to ensure high efficiency and less computation.

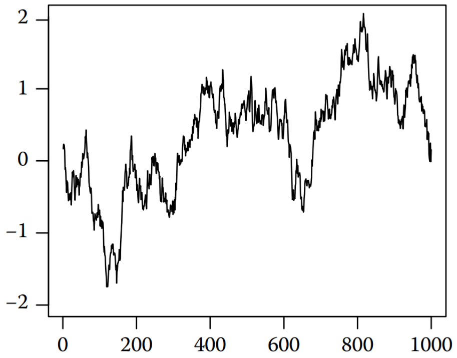
3.4.1 Optimal Scaling of Random-Walk Metropolis

We first focus on the most common case, the symmetric random-walk Metropolis algorithm (RMW) [5], where the proposal distribution is given by $X_{n+1} = X_n + Z_{n+1}$, with the increments $\{Z_i\}$ independently and identically distributed from a fixed symmetric distribution with the scaling parameter $\sigma > 0$

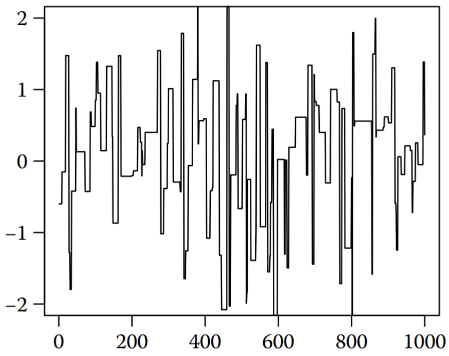
$$Z_i \sim \mathcal{N}(0, \sigma^2 I_d),$$

where d is the dimension of the random walker.

In this case, scaling the proposal, specifically the choice of σ^2 becomes a crucial issue in optimizing the algorithm. The value of scales σ strongly affects the proposed moves. We use



trace plot with small σ, large acceptance rate



trace plot with large σ, small acceptance rate

the *acceptance rate* to measure the proportion of accepted moves in all proposed moves. For σ extremely small (as in Fig.2(b)), the acceptance will be approximately 1. Hence, the algorithm will accept almost all proposed moves, and the proposed jumps will be small. So it takes a long time for the RMW algorithm to converge to a stationary distribution, and the algorithm will be highly inefficient. For σ extremely large (as in Fig.2(a))., the acceptance rate will be close to 0, and the algorithm will accept nearly no moves. So, the chain will stay fixed for large numbers of iterations, leading to poor mixing. Therefore, we manage to avoid both extremes by monitoring the acceptance rate and keeping it both far from 0 and far from 1.

Under the assumption that the proposal increments follow the distribution of the form $N(0, \sigma^2 I_d)$, we aim to find an optimal proposal scaling for σ. As we monitor the performance of σ by acceptance rate, we expect to find the optimal acceptance rate.

Proposition 3.26. For a RWM $X_{n+1} = X_n + Z_{n+1}$ with $Z_i \sim \mathcal{N}(0, \sigma^2 I_d)$, as $d \to \infty$, the precise optimal acceptance rate is 0.234. For finite dimensional situations where $d \geq 5$, the optimal acceptance rate approximates the asymptotic acceptance rate 0.234. For $d = 1$, the optimal acceptance rate is approximately 0.44 [5].

3.4.2 Adaptive Metropolis-within-Gibbs

After obtaining the optimal acceptance rate, we aim to find the appropriate proposal scaling σ to achieve the optimum, which is our goal in this section as well. One commonly used method is trial and error. That is, we reduce the proposal scaling when the acceptance rate is too high and increase the scaling when the acceptance rate is too low. This method works but is time-consuming because of its need for repeated manual intervention. Therefore, we consider alternative algorithms, called adaptive MCMC, which improve the Markov chains and convergence in their processes. As to the algorithm used in the Bayesian parameter estimation method, we apply the Metropolis-Hasting-within-Gibbs algorithm. Hence, we focus on the specific adaptive Metropolis-within-Gibbs algorithm.

Sufficient conditions to guarantee convergence is essential before implementing the adaptive algorithm to avoid the possibility of converging to wrong numbers.

Let each P_γ be a Metropolis algorithm, with $\Gamma_n = \gamma$ being the n-th proposal choice. Suppose χ is the set of all candidate states of X_n. We can guarantee the convergence assuming the following conditions [5].

Assumption 3.27. *Diminishing (Vanishing) adaptation condition:* Suppose the amount of adapting at the n-th iteration goes to 0 in probability as $n \to \infty$:

$$\lim_{n\to\infty} \sup_{x\in\chi} \| P_{\Gamma_{n+1}}(x,\cdot) - P_{\Gamma_n}(x,\cdot)\|.$$

If this condition is not satisfied, the algorithm can result in samples from a distribution different from the target distribution.

Assumption 3.28. *Containment (Bounded convergence) condition:* For $\varepsilon > 0$, $\{M_\varepsilon(X_n, \Gamma_n)\}_{n=0}^{\infty}$ is bounded in probability, where M_ε is the convergence time of the kernel P_γ beginning in state $x \in \chi$,

$$M_\varepsilon(x, y) = \inf\{n \geq 1 : \|P_\gamma^n(x,\cdot) - \pi(\cdot)\| \leq \varepsilon\}.$$

This condition is intrinsically satisfied for the adaptive Metropolis-within-Gibbs. It has been proved in [5] that when these two conditions are satisfied, Markov chains converge for the algorithms.

Proposition 3.29. Assuming the diminishing adaptation condition and containment conditions, the Markov chain (X_t) has asymptotic convergence:

$$\lim_{n\to\infty} \sup_{A\subset\chi} \|\mathrm{P}\big(X_N \in A\big) - \pi(A)\| = 0.$$

And for all bounded $g : \chi \to \mathbb{R}$, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(X_i) = \pi(g).$$

Ensuring the convergence, we explain the adaptive algorithm next. In the Metropolis-within-Gibbs algorithm, each of the variables is updated at a time using a Metropolis algorithm with a one-dimensional proposal. Furthermore, good values of σ^2 can differ from each other greatly considering updating one variable to the other. Therefore, we conclude that the adaptive method should be used to automatically compute good values of σ^2 instead of computing manually. Moreover, Metropolis-within-Gibbs algorithm essentially corresponds to $d = 1$ case, and thus the optimal acceptance rate is usually closer to 0.44 than 0.234.

For the i-th variable of the current sample iterate (θ, x_0, Z), let ls_i denote the logarithm of the standard deviation of the increment proposed to i-th variable. Suppose the i-th variable is updated using a proposal increment distribution $\mathcal{N}(0, e^{2ls_i})$. Then we aim to find optimal values of ls_i such that the acceptance rate is approximately 0.44 [5].

Let p_1, p_2 denote the minimum and maximum allowable from which the random walker can sampler from respectively, i.e. $p_1 \leq \gamma \leq p_2$. Suppose there exists a time delay ε in the start of decelerating the adaptation extent, meaning that after which the shrinking of adaptive steps is supposed to start. We first initialize $ls_i = 0, \forall i$ corresponding to the unit proposal variance.

Then we set a batch equal to 100 iterations, which means that we determine every 100 iterations whether to update the ls_i, taking away the time delay ε at the beginning. At the n-th batch of 100 iterations, we need to calculate an acceptance rate (assumed to be ar) and compare it with the optimal acceptance rate $or = 0.44$. Then, according to the given process of adaptive Metropolis-within-Gibbs (as proposed in [5]), we update ls_i by adding or subtracting $\delta(n)$, which is defined as the adaptation amount at n-th batch.

$$ls_i = \begin{cases} ls_i + \sigma(n), & ar > or \\ ls_i - \sigma(n), & ar < or \\ ls_i, & ar = or \end{cases} \tag{35}$$

where $\sigma(n) = \min(0.01, \frac{1}{\sqrt{n}})$ is the constraint satisfying the diminishing adaptation condition to guarantee the convergence.

However, given the need to compute the logarithm of the standard deviation of the variable and the log-likelihood we have calculated in the original Bayesian inference algorithm, we make minor improvements on the update step above. We first keep the adaptation amount $\sigma(n)$, and introduce the following two reciprocal functions. The *sigmoid function* $S(x)$ given by

$$S(x) = \frac{1}{1 + e^{-x}},$$

and the *logit function* $L(x)$ given by

$$L(x) = \log\left(\frac{x}{1 - x}\right).$$

Then we modify the update step by:

$$\gamma = \begin{cases} S(L(\gamma) - 2\sigma(n)), & ar < or \\ \gamma, & ar \geq or \end{cases} \tag{36}$$

To briefly summarize, the adaptive Metropolis-within-Gibbs solves the optimal proposal for a tuning parameter to automatically achieve the optimal acceptance rate when updating the path and states. And we can use this adaptive algorithm (Algorithm.6) to tune the memory parameter γ of the preconditioned Crank-Nicholson scheme.

Algorithm 6: Adaptive algorithm for Tuning γ

1. Suppose now it is the N-th iteration.
2. Compute the batch number: $n = \frac{N}{100-\varepsilon}$
3. Denote the move of γ for each iteration as $\sigma(n)$, which decreases proportional to $\frac{1}{\sqrt{N}}$ roughly. Compute m using
$$\sigma(n) = \frac{1}{\sqrt{\max(1.0, n)}}.$$
4. Compute the acceptance rate ar when the adaptation occurs by
$ar = \dfrac{accepted\ steps}{total\ proposed\ steps}.$
5. Compute γ by transforming using two reciprocal functions (chosen as the sigmoid and logit functions) and the moving distance $\sigma(n)$.
   - if $ar < or$, $\gamma = S(L(\gamma) - 2\sigma(n)) = \dfrac{1}{1 + e^{-\log(\gamma)+\log(1-\gamma)+2\sigma(n)}}$;
   - if $ar \geq or$, γ keeps unchanged.
6. Trim excessive updates by $\gamma = \max\left(\min(\gamma, p_2), p_1\right)$.

3.5 Change-point Estimation

Finally, we attempt to estimate the Arterial Input Function $\delta(t)$ in the pharmacokinetic model using a change point detection technique. Since the previous study mainly employs the Maximum Likelihood method when the Arterial Input Function is a constant [1], we expect to extend the Bayesian inference method to adaptive to models with general Arterial Input Functions. Once we estimate the Arterial Input Function, we can realize the goal of evaluating the model even if the AIF is unknown. Apart from estimating the AIF, there is still a problem with detecting the time delay before the injection. Change point detection appears to estimate the time delay, and the same approach is taken to estimate the AIF.

3.5.1 Overview and Setup

We first introduce the change point detection and address its typical applications. Change point detection identifies times when the probability distribution of a stochastic process or time series changes. Loosely speaking, analysis of change points can be equal to identifying the points within a data set where the statistical properties change [9].

Definition 3.30. Recall the observation function of the model in this thesis,

$$y(t) = LX(t) + \varepsilon(t) = g(t) + \varepsilon(t), \quad 0 \leq t \leq T$$

where $y(t)$ is the observation taken from time t, $\varepsilon(t)$ is the random error, with $\mathrm{E}[\varepsilon(t)] = 0$, and $g(t)$ is an unknown left-continuous and piecewise smooth function. Then a point t_0 satisfying $g(t_0) \neq g(t_0+)$ is said to be a *jump change point*. Else, t_0 is called a first order continuous change point, usually abbreviated to *continuous change point* [9] if it satisfies

$$\begin{cases} g(t_0) = g(t_0+), \\ \lim\limits_{t\to t_0-} \dfrac{\mathrm{d}g(t)}{\mathrm{d}t} \neq \lim\limits_{t\to t_0+} \dfrac{\mathrm{d}g(t)}{\mathrm{d}t} \end{cases}$$

Change point detection detects abrupt shifts in time series trends or shifts in the instantaneous velocity of time series. These shifts are easy to identify using human eyes, and we will list some statistical means to pinpoint them in this section.

Back to our model, we consider the existence of change points given the background. An unknown time delay may exist between the period of the injection of the contrast agent and the arrival of the contrast agent in the plasma. In addition, our model may reach some peaks at the time points when the contrast agent admits a rapid jump. All these changes can yield change points in the model. In this section, we will discuss measures to solve the estimation problem for a general case where the time series change points may exist due to the non-constant Arterial Input Function $\delta(t)$ and the time delay. Again we use the matrix form of the model in this part.

$$\mathrm{d}X(t) = \left(\begin{bmatrix} \alpha\ \delta(t) \\ 0 \end{bmatrix} + \begin{bmatrix} -(\lambda+\beta) & \beta \\ \lambda & -k \end{bmatrix} X(t) \right) \mathrm{d}t + \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} dW_1(t) \\ dW_2(t) \end{bmatrix}$$

We focus on the AIF in this section, so we define an injection term as:

$$F(t) = \begin{bmatrix} \alpha\delta(t) \\ 0 \end{bmatrix}.$$

For all AIFs, we assume the initial condition satisfies

$$\delta(t_0) = 0, \quad t_0 = 0.$$

A least square estimator inspired by [10] is used for change-point estimation separately. This method illustrates the asymptotic property of change-point estimation by assuming that the AIF is a piecewise constant function.

Assumption 3.31. The Arterial Input Functions can be split into piecewise constants in general:

$$\delta(t) = \sum_{j=1}^{n} a_j \mathbb{1}_{[(j-1)\Delta, j\Delta]}(t),$$

with $(a_j)_{j=1,\ldots,n}$ being unknown numbers whose values to be determined.

In addition, computation of the SDE likelihoods needs integration of the equations, which means that integral of AIFs is required. Intuitively, this integral cannot be computed directly for general AIFs, we use the following proposition with its piecewise constant characteristic.

Proposition 3.32. The integral of AIF can be computed using the trapezoidal method [12] by

$$\int_{t_0}^{t_i} \delta(x)\,\mathrm{d}x \approx \sum_{k=1}^{i} \frac{(t_k - t_{k-1})(\delta(t_k) - \delta(t_{k-1}))}{2}.$$

3.5.2 Estimate the Time Delay

Firstly, we deal with the time delay used in the adaptive Metropolis-within-Gibbs algorithm before. We solve the change point where the time delay occurs by redefining the injection term $F(t)$ with the delay parameter ε

$$F(t) = \begin{bmatrix} \alpha\delta(t-\varepsilon)\mathbb{1}_{[\varepsilon,\infty)}(t) \\ 0 \end{bmatrix}.$$

Given this injection term $F(t)$ with time delay, we assume for simplicity that the AIF has a linear growth [12]. We apply the transformed model with discrete-time differential equations derived in the Maximum Likelihood method. Taking the time delay into account, we get a slightly different system from Eqn.23 in MLE method:

$$\begin{cases} X_i = A_i X_{i-1} + B_i + \eta_i, \quad \eta_i \sim \mathcal{N}(0, R_i), \\ y_i = HX_i + \sigma\epsilon_i \end{cases}$$

with $H,\ A$ unchanged. But B_i, R_i adapt to satisfying:

$$R_{i+1} = \int_{(i-1)\Delta}^{i\Delta} e^{D(t_i-s)}\Gamma\Gamma' e^{D'(t_i-s)}\,\mathrm{d}s,$$

$$B_i = \begin{cases} 0 & 0 \le i\Delta \le \varepsilon, \\ \displaystyle\int_{(i-1)\Delta}^{i\Delta} e^{D(t_i-s)}P^{-1}F(s)\,\mathrm{d}s & \varepsilon < i\Delta \end{cases}$$

Then we attempt to separate the system by the time delay and adapt the Maximum Likelihood method. We isolate the change point by introducing the auto-regressive process $Z_i = AZ_{i-1} + \eta_i$, thus obtaining an equivalent system as:

$$\begin{cases} Z_i = A_i Z_{i-1} + \eta_i, \quad \eta_i \sim \mathcal{N}(0, R_i), \\ \tilde{y}_i = HZ_i + \sigma\epsilon_i, \\ y_i = c_i + \tilde{y}_i \end{cases} \tag{37}$$

with $c_i = HC_i$, and

$$C_i = AC_i + B_i = A^i \int_0^{i\Delta} e^{-Ds}P^{-1}F(s)\,\mathrm{d}s.$$

The ODE system is thus split into two parts by the change point ε. Before arriving at the change point, $c_i = 0$. Then we give the detailed assumption of the AIF with linear growth to get an explicit system.

Assumption 3.33. (A1) Linear Assumption: Assume $\delta(t)$ has a linear growth ($a_j = j\Delta$) so that c_i can be approximated by $c_i = \mu_0 + \mu_1(i - j^*)$, where μ_0 and μ_1 are two unknown parameters depending on θ and Δ.

Under $A1$, we can rewrite the system (Eqn.37) as:

$$\begin{cases} Z_i = A_i Z_{i-1} + \eta_i, & \eta_i \sim \mathcal{N}(0, R_i), \\ \tilde{y}_i = HZ_i + \sigma\epsilon_i, \\ y_i = \mu_0 + \mu_1(i - j^*)\mathbb{1}_{i>j^*} + \tilde{y}_i \end{cases} \tag{38}$$

It is proved in [12] that this system can reduce to

$$v_i = \mu_1 \mathbb{1}_{i \geq j^*} + U_i, \tag{39}$$

by introducing $v_i = y_{i+1} - y_i$ and $U_i = \tilde{y}_{i+1} - \tilde{y}_i$. So detecting the change-points of this model is equivalent to a similar problem with the process $(\tilde{y}_i)$ with a constant change of drift $\mu_0 + \mu_1 \mathbb{1}_{i>j^*}$. To reduce the computational cost, we use the Least Square method to detect the change point, where the underlying error distribution is not necessarily specified. The Least Square method for change point detection applies to a simple binary case according to [10].

Proposition 3.34. Consider an observation model, $y(t) = g(t) + \varepsilon(t), 0 \leq t \leq T$. $g(t)$ can only take two different values before and after time τ_0 so that τ_0 is the target change point,

$$g(t) = \begin{cases} g_1, & t \leq \tau_0 \\ g_2, & t > \tau_0 \end{cases} \tag{40}$$

where g_1, g_2, and τ_0 are unknown. Then the least squares estimator of τ_0 can be defined as

$$\hat{\tau} = \underset{\tau}{\text{argmin}} \left(\min_{g_1, g_2} \{ \sum_{t=1}^{\tau} (y(t) - g_1)^2 + \sum_{t=\tau+1}^{T} (y(t) - g_2)^2 \} \right).$$

As the observation model does not fit the simple binary condition, we consider the newly-developed process of increments (v_i). According to Eqn.39, $v(t) = g(t) + \varepsilon(t)$, with

$$g(t) = \begin{cases} \mu_0, & t \leq t_{j^*} \\ \mu_0 + \mu_1, & t > t_{j^*} \end{cases} \tag{41}$$

so the above least-squares change-point estimator can be applied to (v_i). Let $\overline{v}_j$ and $\overline{v}_j^*$ denote the mean of the first j observations and the last $n - j$ observations respectively.

Proposition 3.35. The least squares estimator for (v_i) is given by

$$\hat{j} = \underset{j}{\text{argmin}}\, S_j^2,$$

where S_j^2 represents the sum of squares of residuals:

$$S_j^2 = \sum_{i=1}^{j} (v_i - \overline{v}_j)^2 + \sum_{i=j+1}^{n} (v_i - \overline{v}_j^*)^2.$$

Hence, the change point can be computed by

$$\hat{\tau} = \hat{j} * \Delta,$$

where $\Delta = t_i - t_{i-1}$, and $\hat{\tau}$ denotes the estimate of the change point, which is also the time delay.

Therefore, we can estimate the time delay by implementing Algorithm.7.

Algorithm 7: Time delay estimation using the Least Squares method

```
• Initialize:
    1. let the initial observation be $y_0 = 0$ at time $t_0 = 0$
    2. suppose the time delay lie in the point of index $j$
    3. generate $n$ observations $y_1, y_2, \dots, y_n$ according to the diffusion model at time
       $t_1, t_2, \dots, t_n$
    4. compute data series $(v_i)_{i=1}^n$ by $v_i = y_i - y_{i-1}, \quad i = 1, \dots, n$
• Iterate from $j = 1$ to $n$:
    1. compute the mean of the first $j$ observations $\overline{v}_j$
    2. compute the mean of the first $j$ observations $\overline{v}_j^*$
    3. compute the sum of squares of residuals when the change point lies in index $j$:
           $$S_j^2 = \sum_{i=1}^{j}(v_i - \overline{v}_j)^2 + \sum_{i=j+1}^{n}(v_i - \overline{v}_j^*)^2.$$
    4. find the minimal $S_{j^*}$ and output the index of the time delay $j^*$
```

We can improve the Maximum Likelihood method by adding a time delay detection, shown as Algorithm.8.

Algorithm 8: MLE with change points

```
1. Compute the exact log-likelihood $L(j, \theta, y_{0:n})$ by
       $$L(j, \theta, y_{0:n}) = \sum_{i=1}^{j} \frac{(y_i - H\hat{Z}_i^-)^2}{HP_i^-H' + \sigma^2} + \log(HP_i^-H' + \sigma^2)$$
       $$- \sum_{i=j+1}^{n} \frac{(y_i - H\hat{Z}_i^- - c_j)^2}{HP_i^-H' + \sigma^2} + \log(HP_i^-H' + \sigma^2),$$
   $\hat{Z}_i^-$, $P_i^-$ are the mean and variance of the prediction distribution $p(Z_i|y_{0:i-1})$.
2. Maximize $L(j, \theta, y_{0:n})$ in $\theta$ for each $j$ to get $\hat{\theta}(j) = \text{argmax}_\theta\, L(j, \theta, y_{0:n})$.
3. The MLE is obtained by inputting the change point $\hat{j}$, i.e. $\hat{\theta} = \text{argmax}_\theta\, L(\hat{j}, \theta, y_{0:n})$.
```

3.5.3 Estimating the Arterial Input Function

In previous experiments, one assumption is that the AIF is known. In order to extend the Bayesian parametric inference approach to a more generally applicable method, we also attempt to perform an estimation with an unknown AIF. We take the injection term $\alpha\delta(t)$ as a whole and transform the problem of parameter estimation as an estimation problem with a changing parameter α'. In this part, we propose an approach to estimating the AIF based on the change point estimation. The problem of estimating the AIF is equivalent to recovering the AIF using change points, so the process of the experiment can be divided into two stages. The first stage is to detect the total change points of the SDE system, with the number and positions of change points unknown. In the second stage, estimation is done by separately measuring the part of AIF in a bin obtained by setting the change points as two endpoints of a time interval. Then the inference problem of an unknown function is split into pieces of estimation problems for parameter α' where the AIF is a constant 1.

As stated in [10], the Least Squares method only applies when there is only one change point. However, there may be change points when estimating the AIF. So we need an approach to detecting various change points instead. Several techniques are discovered to handle multiple change points via minimizing a cost function over possible numbers and locations of change points. Possible choice of cost functions includes penalized likelihood and minimum description length. In this thesis, we adopt the Binary Segmentation method (described in Algorithm.9), proposed to multiple change points.

At first, we use $\mathcal{S}$ and $\mathcal{C}$ to denote the set of segments of the data which need change point detection and the set of detected change points, respectively. Then we introduce a general test statistic $\Lambda(\cdot)$, defined by the segmentation cost, and a penalty term $\log(n)$. Similar to what is used in the single change point detection with MLE before, we choose the sum of squares as the segmentation cost here. Additionally let estimator of change point position be $\hat{\tau}(\cdot)$, and rejection threshold be $\boldsymbol{C}$. $\boldsymbol{C}$ should be chosen to avoid an expensive computational cost. We first use all data and compute the its of squares.If no change point is detected, then terminate. Otherwise, the data is split into two segments before and after the change point. And we apply the detection method to each segment. Then we repeat this procedure until the algorithm can detect no further change points.

Algorithm 9: Generic binary segmentation algorithm

```
1. Input a set of data points $y_1, y_2 \dots, y_n$.
2. Define the test statistic:
       $$\Lambda(y_{s:t}) = \frac{y_t^2 - y_s^2 - |y_t - y_s|^2}{t - s}.$$
3. Initialisation: let $\mathcal{C} = \emptyset$ and $\mathcal{S} = \{[1, n]\}$.
4. Iteration while $\mathcal{S} \neq \emptyset$:
   • Arbitrarily choose an element of $\mathcal{S}$ and denote it as $[s, t]$.
   • If $\Lambda(y_{s:t}) < \boldsymbol{C}$, remove $[s, t]$ from $\mathcal{S}$.
   • If $\Lambda(y_{s:t}) \geq \boldsymbol{C}$, then:
       (a) remove $[s, t]$ from $\mathcal{S}$;
       (b) add a new element $r = \hat{\tau}(y_{s:t}) + s - 1$ to $\mathcal{C}$;
       (c) if $r \neq s$, add a new element $[s, r]$ to $\mathcal{S}$;
       (d) if $r \neq t - 1$, add a new element $[r + 1, t]$ to $\mathcal{S}$.
5. Output the recorded change points via set $\mathcal{C}$.
```

Assume we obtain m change points with their positions after the above multiple change point detection. Suppose the indices of all these m change points are listed as $c_1, c_2, \dots, c_m$. As the whole time period we study is $[0, T]$, assuming n observations are used totally for the change point detection, we can thus split the time interval by several bins using the change points into

$$[0, \frac{c_1}{n} \cdot T], [\frac{c_1}{n} \cdot T, \frac{c_2}{n} \cdot T], \dots, [\frac{c_m}{n} \cdot T, T].$$

Furthermore, we assume $c_0 = 0, c_{m+1} = n$ for consistency. Then the problem of estimating the AIF in each bin between the change points is equivalent to estimating a new drift parameter α' when AIF is assigned a constant value 1.

When estimating the AIF, all the parameters are unknown. Therefore, we try joint estimation using the Bayesian approach and record the estimates of α' in each time bin with the AIF $\delta'(t) = 1$.

In the i-th bin $[\frac{c_i}{n} \cdot T, \frac{c_{i+1}}{n} \cdot T]$, the model can be considered with parameters

$$\alpha', \beta, \lambda, k, \sigma_1, \sigma_2, \delta'(t) = 1.$$

Hence, we perform the full Bayesian method for joint estimation of these parameters on every time interval $[\frac{c_i}{n} \cdot T, \frac{c_{i+1}}{n} \cdot T],\ i = 0, \dots, m$. Then the estimate of each α' in the i-th time bin, denoted as $\hat{\alpha}_i$ can serve as the value of $\delta(t)$ in this time interval. The procedure containing the complete Bayesian parameter estimation method can then be briefly summarized in Algorithm.10.

```
Algorithm 10: Bayesian estimation with change points
    1. Assume the true model is composed of parameters $\alpha, \beta, \lambda, k, \sigma_1, \sigma_2$ and the AIF $\delta(t)$.
    2. For the model in each bin $[\frac{c_i}{n} \cdot T, \frac{c_{i+1}}{n} \cdot T]$, $i = 0, \dots, m$, we equivalently estimate a
       system with the following unknown model parameters and a constant AIF
           $\alpha_i', \beta, \lambda, k, \sigma_1, \sigma_2, \delta'(t) = 1, \quad t \in [\frac{c_i}{n} \cdot T, \frac{c_{i+1}}{n} \cdot T].$
    3. Estimate the parameter $\alpha_i'$ using the Bayesian method for joint parameter estimation
       (proposed in Algorithm.5). Denote the estimate as $\hat{\alpha}_i$.
    4. Let $\hat{\delta}(t)$ be the estimate of $\delta(t)$. Compute $\hat{\delta}(t)$ by
           $\hat{\delta}(t) = \sum_{i=0}^{m} \hat{\alpha}_i \mathbb{1}_{\left[\frac{c_i}{n} \cdot T, \frac{c_{i+1}}{n} \cdot T\right]}.$
```

4 Experiments and Results

To illustrate the usage and benefits of the Bayesian inference method, we compare its performance with that of the maximum likelihood method by conducting simulated experiments to calibrate the pharmacokinetic model. The whole experiment procedure can be divided into two main steps. Firstly, we established an identifiability analysis of the model to ensure the feasibility of parameter estimation. Next, we use the maximum likelihood and Bayesian method separately to estimate the identifiable parameters. During the experiment using MLE, we only deals with the model parameters. While in the process of the Bayesian inference experiment, we also managed to evaluate the time delay and the Arterial Input Function via change point estimation. The Julia codes corresponding to the simulations can be found at BayesianInferenceOU, and the MATLAB codes can be found at FVBiOU.

4.1 Identifiability Analysis.

We first ensure the identifiability of the model and parameters before estimating parameters. Since the identifiability only applies to drift terms, we employ the analysis technique in [4] to the drift parameters.

First, we establish the structural identifiability test of the model. We claim that the pharmacokinetic model is structurally identifiable because the moment dynamics up to the second order (Eqn.14 15) are closed and exact. Therefore, all the drift terms are structurally identifiable, and the MLE in Eqn.9 exists. However, there exists no algorithm that analyzes the identifiability of diffusive parameters. So we will perform parameter inference on each diffusive parameter.

Then we analyze the practical identifiability of each model parameter. We compute the log-likelihood using Eqn.11 and compute the corresponding individual likelihood-based confidence intervals using Eqn.12. For simplicity, we use the credible intervals of the log-likelihood to represent the likelihood-based confidence intervals. We get the left and right endpoints of the credible intervals corresponding to each parameter, shown in the following table. We can observe from Table.1 that all the credible intervals for likelihoods are finite. Hence, all model parameters are regarded as practically identifiable.

Parameter	α	β	λ	k	σ_1	σ_2
left endpoint	-4457.83	-4764.63	-4023.36	-3939.13	-1784.86	-1783.82
right endpoint	-4040.65	-4738.65	-4020.33	-3936.16	-1781.97	-1780.73

Table 1: Credible interval for log-likelihood of each parameter

Due to the above results from analyzing the two types of identifiability, we could estimate all the parameters. It is reasonable for us to perform a multivariate estimation with all parameters estimated simultaneously using the Bayesian approach with guided proposals.

4.2 Comparative Experiments

From the above identifiability analysis, the pharmacokinetic model and the drift terms are proved identifiable. So, we could experiment to estimate all model parameters containing the diffusive terms but not the AIF. For the comparative experiment, we randomly generate two data sets where the values are the true values of the model parameters. Then we successively apply the Maximum Likelihood and Bayesian inference methods to these two data sets. We compare the performances of these two methods using different estimators, chosen as the maximum likelihood estimator (MLE) and the posterior mean, respectively.

4.2.1 General Setup

To compare and analyze the performances of the two methods for parameter estimation, we should experiment using different methods under the same setup and control the variables.

- **Δ: Time grid**. Observation times are equally spaced. We try to simulate the observations using a dense enough time grid $\Delta = t_i - t_{i-1} = 0.001s$.
- **σ: noise measurement**. We set the Gaussian noise decorated to observations as $\sigma = 1$.
- **T: period**. We set the time interval to observe the process $[0, T]$ as $[0, 10]$ seconds.
- **X_0: initial state**. We set the initial state to be $X_0 = [0, 0]'$ in every experiment.
- **n: iteration**. We experiment by $n = 10000$ iterations to implement the updating algorithm.
- **the first data set**.
  - True values of biological parameters $(\alpha, \beta, \lambda, k)$ and noises σ_1, σ_2 are given by

$$\alpha = 1.0,\ \beta = 6.0,\ \lambda = 1.0,\ k = 1.5,\ \sigma_1 = 2.0,\ \sigma_2 = 0.5. \tag{42}$$

  - Initial values are assigned by

$$\alpha_0 = 0.0,\ \beta_0 = 0.0,\ \lambda_0 = 0.0,\ k_0 = 0.0,\ \sigma_{10} = 1.0,\ \sigma_{20} = 0.1 \tag{43}$$

- **the second data set**.
  - Values of biological parameters $(\alpha, \beta, \lambda, k)$ and noises σ_1, σ_2 are given by

$$\alpha = 10.0,\ \beta = 30.0,\ \lambda = 15.0,\ k = 20.0,\ \sigma_1 = 6.0,\ \sigma_2 = 2.0. \tag{44}$$

  - Initial values are assigned by

$$\alpha_0 = 0.0,\ \beta_0 = 0.0,\ \lambda_0 = 0.0,\ k_0 = 0.0,\ \sigma_{10} = 1.0,\ \sigma_{20} = 0.1 \tag{45}$$

4.2.2 Maximum Likelihood Method

Since the computation of maximum likelihood is time-consuming, the estimation method using the MLE mainly deals with models when the Arterial Input Function is a known constant. So, we do comparison experiments using the maximum likelihood and Bayesian method within the same scenario such that $\delta(t) = 50$ [1]. To verify the superiority of the Bayesian method, we should keep the true models and the initial values consistent when implementing the maximum likelihood and Bayesian approach. However, the difficulty is that the parameters estimated are in different forms in these two methods due to reparametrization. To be specific, we estimate model parameters $\alpha,\ \beta,\ \lambda,\ k,\ \sigma_1,\ \sigma_2$ directly in the Bayesian setup, but the reparametrization of $\theta_1,\ \theta_2,\ \theta_3,\ \theta_4,\ \theta_5,\ \theta_6$ (Eqn.26) when implementing the maximum likelihood method. Therefore, we need to transform the model parameters and compute their corresponding initial values according to Eqn.26 and Eqn.43-45. Due to the limitation of the maximum likelihood approach found in [1], only five parameters among $\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6$ can be estimated simultaneously. So, it is reasonable to fix θ_5 to its true value and perform joint estimation of all other parameters. We assign initial values as for both the first and the second data set (Eqn.42-44)

$$\theta_1 = 0.5,\ \theta_2 = 0.3,\ \theta_3 = 0.2,\ \theta_4 = 0.1,\ \theta_6 = 0.0.$$

Then the results are illustrated in Table.2-3. True values of the reparametrized model are computed by transformation. The estimates represent the values resulting from the Maximum Likelihood Estimators, and the standard errors are shown in the brackets.

parameter	θ_1	θ_2	θ_3	θ_4	θ_6
true value	0.91	0.99	0.2	0.01	20
estimate	0.87(0.08)	0.94(0.12)	0.06(0.26)	0.03(0.13)	20.01(0.56)

Table 2: MLE for the first data set

parameter	θ_1	θ_2	θ_3	θ_4	θ_6
true value	0.6	0.9	0.7	0.2	20
estimate	0.64 (0.17)	0.79 (0.15)	0.78 (0.26)	0.12 (0.16)	20.01(0.38)

Table 3: MLE for the second data set

4.2.3 Bayesian Inference Method

In this part, we conduct a comparative experiment on the two data sets in a Bayesian setup. We first generate continuous data. Then we display the data using the second data set in Fig.2. Here, the blue curve y_1 denotes the value changes of a sum of coordinates of the continuous sampling, while the orange curve y_2 denotes the values of the second coordinate. And the green scatter plots are the values of the sum of coordinates of the discrete sampling.

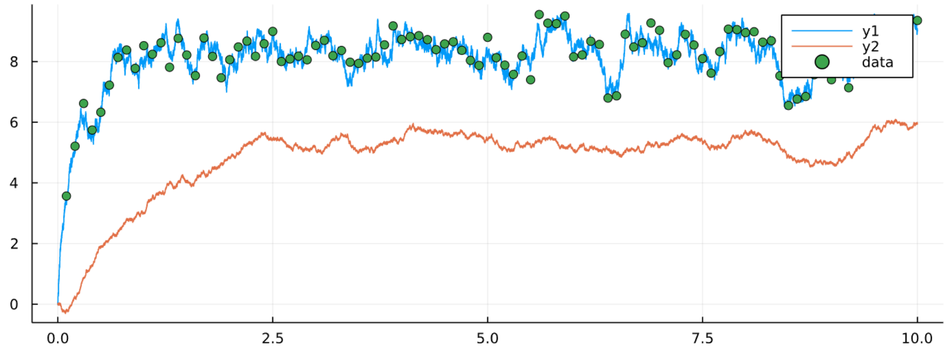


Figure 2: Data for the second data set

Then we recover the paths, reconstruct the states, and generate discrete observations consistent with the continuous observations. We decorate each observable with a Gaussian noise $\Sigma = 1$ and set $L = [1, 1]$ to observe the sum of coordinates. Next, we implement the adaptive BFFG algorithm with a tuned memory parameter. We keep the initial value of memory parameter γ as 0.5 in all experiments because the algorithm will tune the proposal automatically when estimating different parameters. To demonstrate the validity of the Bayesian approach, we estimate each parameter separately and measure its estimate, which is the posterior mean. For comparison with the performances of MLEs, we compute and record the posterior means of the parameters with their standard errors in Table.4-5 for the two data sets.

We observe that the drift terms can be well estimated. But the diffusive terms have relatively large errors in general. In addition, the standard errors of the Bayesian method are mostly smaller than those of the MLEs. Therefore, we can conclude that the Bayesian approach performs better than the maximum likelihood method.

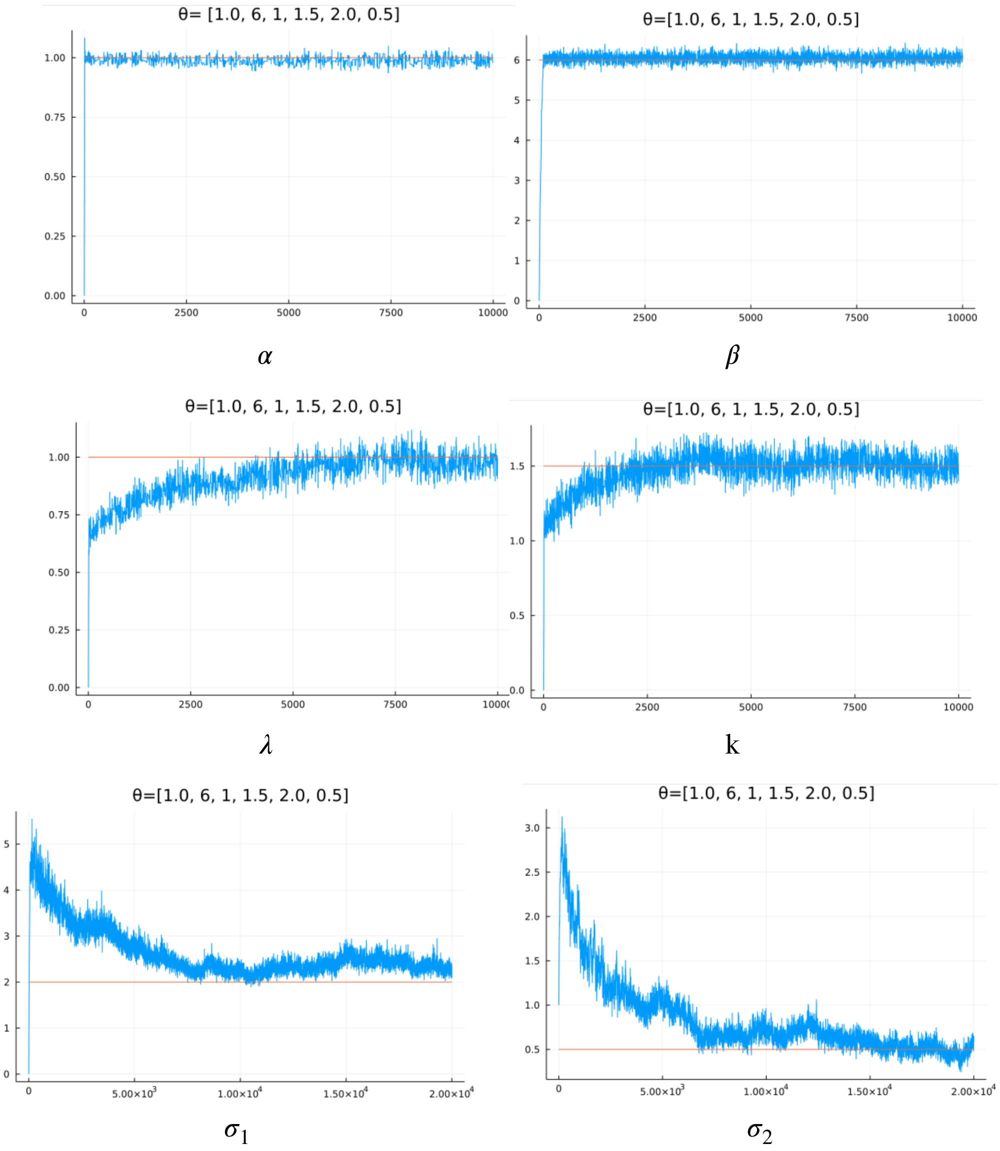
parameter	α	β	λ	k	σ_1	σ_2
true value	1	6	1	1.5	2	0.5
estimate	1.001	6.083	0.991	1.481	2.232	0.479
error	0.001	0.09	0.010	0.020	0.244	0.022

Table 4: Bayesian estimates for the first data set

parameter	α	β	λ	k	σ_1	σ_2
true value	10	30	15	20	6	2
estimate	10.054	29.897	15.059	20.092	6.141	2.091
error	0.092	0.045	0	0.096	0.252	0.095

Table 5: Bayesian estimates for the second data set

To see the convergence of estimates to their corresponding true values more clearly, we plot the trajectories of the estimates. The trace plots of the first data set are given by Fig.3(a)-3(f), and those of the second data set are given by Fig.4(a)-4(f). In each figure, the red line represents the true value of the parameter, the blue curve represents the trajectory of the estimate, and the estimated parameter is written in the caption. The title of each subfigure illustrates the true values of this model. We can see from the figures that all the parameters converge to their corresponding true values stably.



α

β

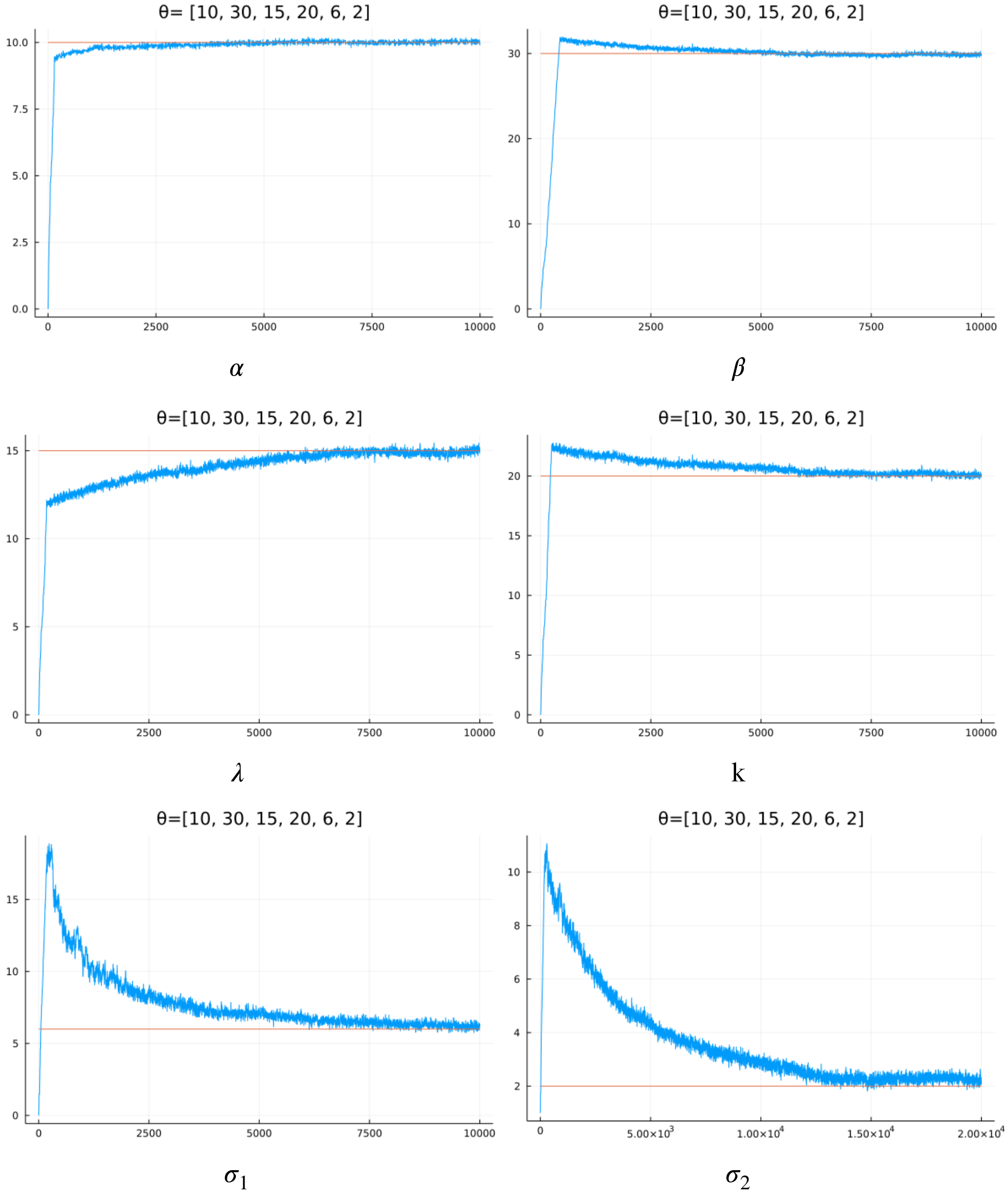
λ

k

σ_1

σ_2

Figure 3: Individual parameter estimates for the first data set



α β

λ k

σ_1 σ_2

Figure 4: Individual parameter estimates for the second data set

4.3 Tentative Experiments

In addition to comparative experiments, we carry out tentative experiments using the Bayesian approach to solve the limitations of the maximum likelihood method and improve its drawbacks. In tentative experiments, we attempt to estimate diffusive terms and multiple parameters simultaneously using the Bayesian inference method. Furthermore, we detect the time delay and estimate the AIF by employing change point estimation When implementing the Bayesian estimation method.

4.3.1 Setup

To implement the Bayesian joint estimation scheme and change point estimation, we conduct the experiments on a new data set (the third data set) under the following assumptions.

- $\Delta = t_i - t_{i-1} = 0.001s$
- $\sigma = 1$
- $[0, T] = [0, 10]s$
- $X_0 = [0, 0]'$
- $n = 10000$
- **the third data set**:
  - Values of biological parameters $(\alpha, \beta, \lambda, k)$ and noises σ_1, σ_2 are given by

$$\alpha = 116.7,\ \beta = 5.83,\ \lambda = 1.25,\ k = 2.25,\ \sigma_1 = 2.0,\ \sigma_2 = 1.0. \tag{46}$$

  - Initial values are assigned by

$$\alpha_0 = 100.0,\ \beta_0 = 0.0,\ \lambda_0 = 0.0,\ k_0 = 0.0,\ \sigma_{10} = 1.0,\ \sigma_{20} = 0.0 \tag{47}$$

- **AIF**: The AIF is fixed to be $\delta(t) = \dfrac{t}{1+t^2}$ for joint estimation. For change point estimation, we try two cases

$$\delta(t) = \frac{t}{1+t^2}, \quad \delta(t) = \frac{1}{1+t^2}$$

4.3.2 Joint Estimation

We first implement the Bayesian joint estimation scheme for all parameters using the third data set (Eqn.46). Similar as before, we can display the data in Fig.5 Differences can be observed easily due to the non-constant AIF.

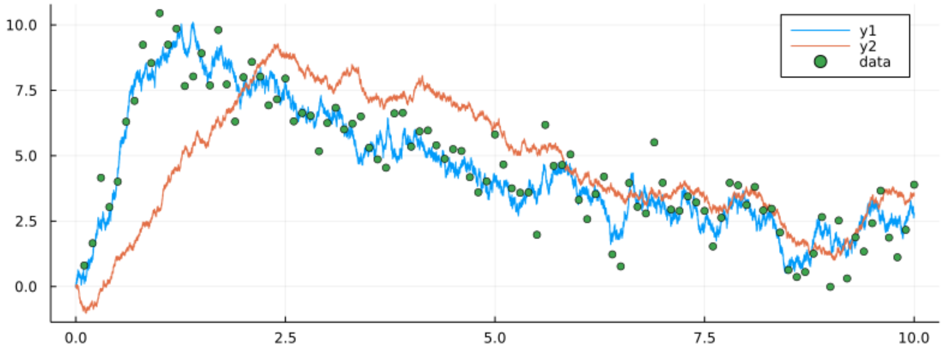


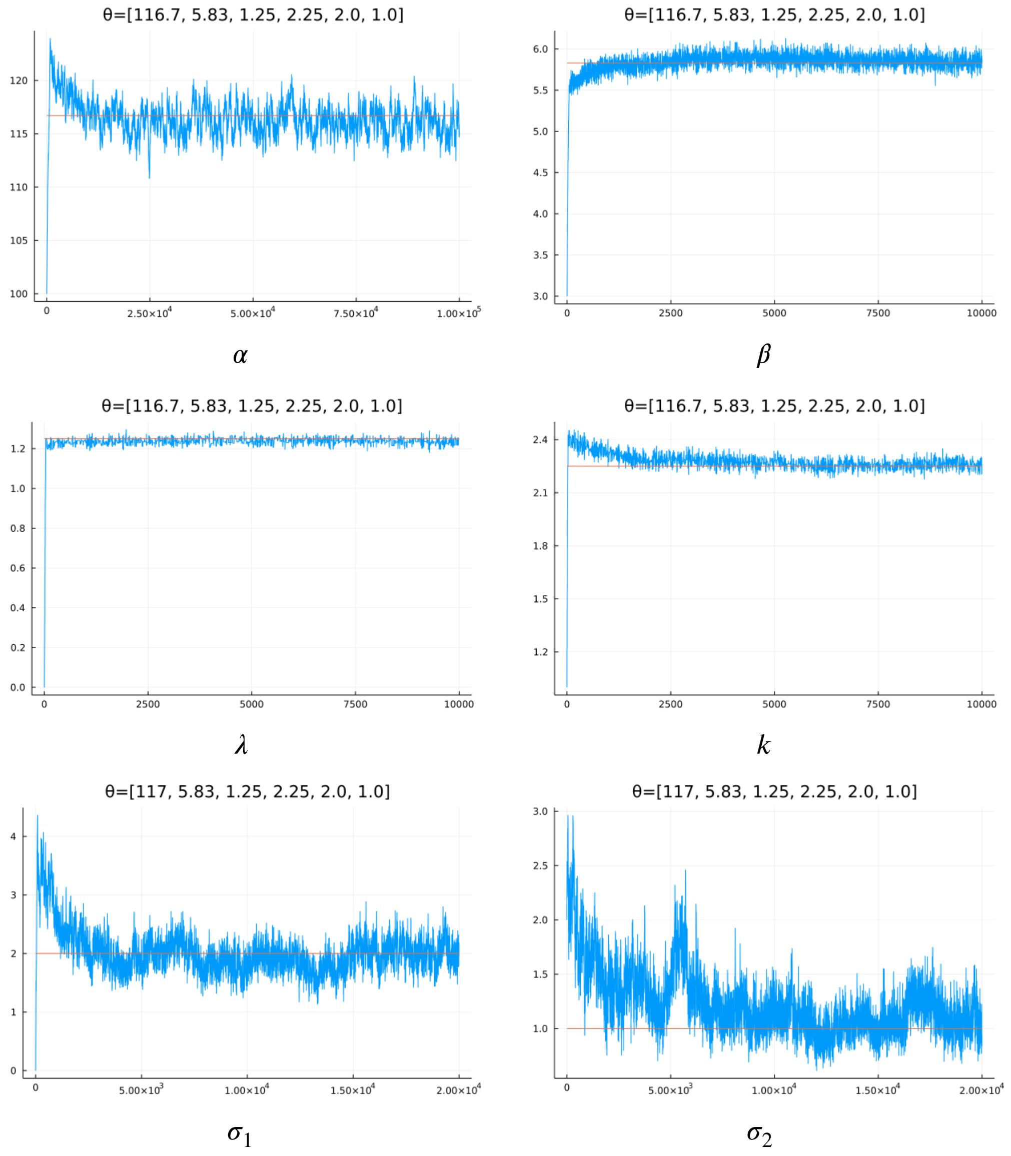
Figure 5: Data for the third data set

We aim to estimate all model parameters simultaneously, both drift and dispersion. Assuming all model parameters are unknown, we estimate each parameter one by one, in one trial, with a non-constant AIF. We conduct the joint estimation experiment and record the estimate of each parameter with the standard error. The results are given in Table.6, structured similarly as before.

parameter	α	β	λ	k	σ_1	σ_2
true value	116.7	5.83	1.25	2.25	2.0	1.0
estimate	116.685	6.04	1.247	2.246	2.211	1.121
error	0.605	0.360	0.042	0.035	0.221	0.112

Table 6: Joint estimates for the third data set

Meanwhile, we plot the trajectories of all parameters in Fig.6. In practice, we also try to estimate the ratios and sums using the example reparametrisation in Eqn.34. The trace plots of the reparametrized model are given in Fig.7.



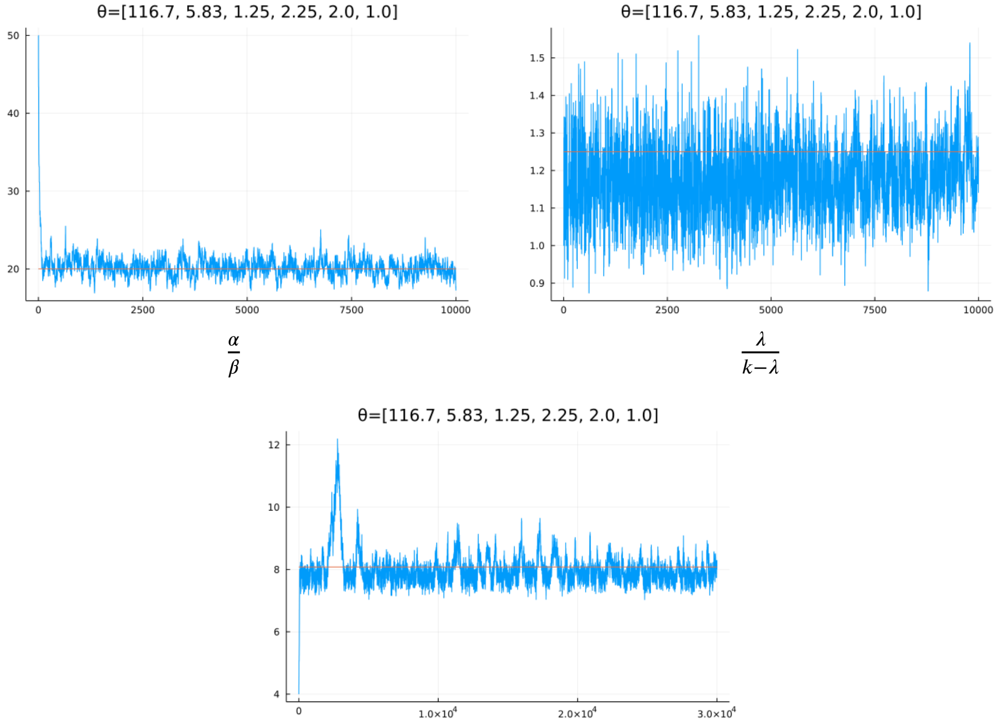
α β

λ k

σ_1 σ_2

Figure 6: Joint estimates for the third data set

From Fig.6, we can see that the estimate of each parameter using joint estimation converges to their true values within three or four orders of their magnitudes. Therefore, we can conclude that joint estimation of the pharmacokinetic model can be solved using the Bayesian parametric inference method.



$\frac{\alpha}{\beta}$ $\frac{\lambda}{k-\lambda}$

$\beta + k$

Figure 7: Joint estimates using reparametrization for the third data set

4.3.3 Change Point Estimation

In this section, We simulate estimation scheme using change point detection. We conduct change point estimation experiment twice with respect to two AIFs, $\delta(t) = \dfrac{t}{1+t^2}$, and $\delta(t) = \dfrac{1}{1+t^2}$. We sample the continuous data by generating observations every $0.0001s$ on the time interval $[0, 10]s$, the same as what we have done in joint estimation. Among all, we take one discrete observation every 0.1 second, resulting in 100 observations. Let i denote the index of the data point at time t_i.

Firstly, we estimate the time delay by implementing Algorithm.7. We can obtain indices of the time delay, 6 for $\delta(t) = \dfrac{t}{1+t^2}$ and 3 for $\delta(t) = \dfrac{1}{1+t^2}$. So we can conclude that there exists a time delay at the beginning of injection. And the time delay lies at approximately $\varepsilon = 0.6s$ or $\varepsilon = 0.3s$, corresponding to the two choices of AIFs.

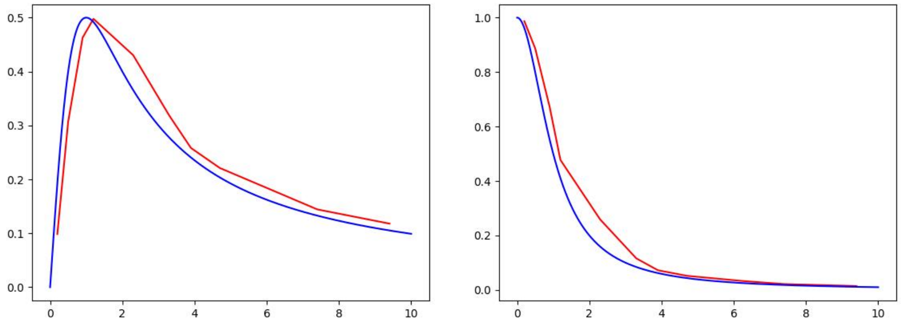
Next we turn to estimate the AIF by implementing Algorithm.9. In this case, the whole time interval is split into $n = 100$ small intervals, with a time increase $\Delta = t_i - t_{i-1} = 0.1s$. We perform multiple change points detection on the third data set with two choices of non-constant AIFs (Eqn.46), and then get a series of indices representing the positions of change points. For true AIF being $\delta(t) = \dfrac{t}{1+t^2}$, the indices of the change points are

$$cps = [0, 2, 4, 5, 7, 10, 26, 33, 42, 48, 63, 83, 93, 100],$$

For true AIF $\delta(t) = \dfrac{1}{1+t^2}$, the indices of the change points are

$$cps = [0, 1, 4, 6, 9, 14, 18, 22, 26, 42, 48, 84, 94, 100],$$

Before estimating the AIF, we can first use the change points to recover the function to support that we have the true change points in advance. The recovery of functions with their true functions same as before is shown in Fig.8.



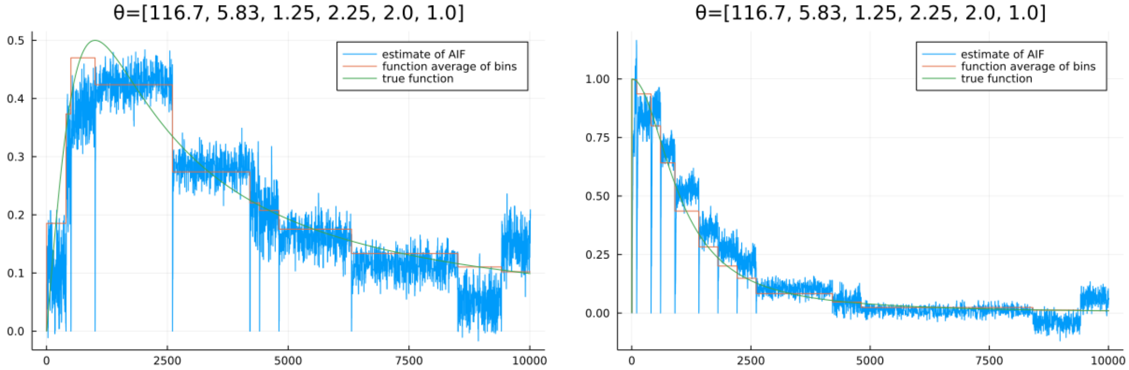
$\delta(t) = \frac{t}{1+t^2}$ $\delta(t) = \frac{1}{1+t^2}$

Figure 8: recovery of AIFs using change points

Based on the correct change points detected, we implement the estimation scheme for the two AIFs according to Algorithm.10. However, it failed to meet our expectations that we could not estimate the AIF when all parameters are unknown.

To validate our estimation scheme, we try to recover the AIF when all other model parameters are given. Then piecewise estimates of the AIFs are shown in Fig.9(a) and 9(b) respectively. In each figure, the red curve represents the ideal structure of each AIF composed of piecewise constants obtained using the posterior mean of every bin, and the green one represents the curve of true AIF.

Therefore, we can claim that the Bayesian method for estimating the AIF works only for given model parameters, but fails with all unknown parameters.



$\delta(t) = \frac{t}{1+t^2}$ $\delta(t) = \frac{1}{1+t^2}$

Figure 9: AIF Estimation

5 Conclusion and Discussion

As stated in [1], the popular maximum likelihood method to estimate model parameters of the bi-dimensional Ornstein-Uhlenbeck process has limitations. It is not only time-consuming for a high-dimensional data set but works only assuming one fixed parameter among all. Meanwhile, a novel Bayesian method for parametric inference is available, proposed in [3]. Therefore, it is vital to experiment the Bayesian approach on the pharmacokinetic model, thus validating its applicability. The overarching aim of this thesis is to calibrate the pharmacokinetic model relying on experimental data using the maximum likelihood and Bayesian method and illustrate the superiority of the Bayesian approach. Specifically, the focus is on comparing these two methods and a new trial for joint estimation with an unknown AIF.

We experiment the maximum likelihood and Bayesian method on two data sets and estimate every model parameter, respectively. By the results of comparative experiments, we conclude that the Bayesian approach generally presents better estimates because it provides fewer standard errors than the maximum likelihood method. Moreover, we can perform joint estimation using the Bayesian approach with all parameters unknown, which solves the limitation of the maximum likelihood method. In addition, we can roughly estimate the time delay of injection using the change point estimation combined with the Bayesian parameter estimation method. However, the Arterial Input Function cannot be estimated when no parameter is fixed at first. Therefore, we fall short of the expectation to solve a more general problem when an unknown injection appears randomly.

To conclude, the Bayesian parameter estimation method applies to the bi-dimensional Ornstein-Uhlenbeck process and provide little-biased estimates. This method not only deals with joint parameter estimation, but also manages to estimate the time delay. Additionally, it can work with the AIF even though we need to fix other model parameters at the beginning. Therefore, we acknowledge the Bayesian method as a practical approach for estimating parameters of SDEs.

For further study, we should continue exploring how to estimate the AIF given no fixed parameter. We could also study the ways to extend the Bayesian method to more complicated models or higher-dimensional cases.

References

[1] F. Benjamin and S. Adeline, "Parameter Estimation for a Bidimensional Partially Observed Ornstein-Uhlenbeck Process with Biological Application," *Scandinavian Journal of Statistics*, vol. 37, no. 2, pp. 200–220, 2010.

[2] M. Schauer, F. van der Meulen, and Z. van Harry, "Guided proposals for simulating multi-dimensional diffusion bridges," *Bernoulli*, vol. 23, no. 4, p. 2917–2950, 2017.

[3] M. Marcin, S. Moritz, and F. van der Meulen, "Continuous-discrete smoothing of diffusions," *Electronic Journal of Statistics*, vol. 15, no. 3, pp. 4295–4342, 2021.

[4] A. P. Browning, D. J. Warne, K. Burrage, R. E. Baker, and M. J. Simpson, "Identifiability analysis for stochastic differential equation models in systems biology," *Journal of the Royal Society, Interface*, 2020.

[5] G. O. Roberts and J. S. Rosenthal, "Optimal scaling for various metropolis-hastings algorithms," *Statistical Science*, vol. 16, no. 4, pp. 351–367, 2001.

[6] I. A. Eckley, P. Fearnhead, and R. Killick, "Analysis of changepoint models," in *Bayesian Time Series Models*, D. Barber, A. Cemgil, and S. Chiappa, Eds. Cambridge University Press, 2011.

[7] R. A, K. C, M. T, B. J, S. M, K. U, and T. J., "Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood," *Bioinformatics*, vol. 25, no. 15, pp. 1923–1929, 2009.

[8] S. Moritz and F. van der Meulen, "Bayesian estimation of discretely observed multi-dimensional diffusion process using guided proposals," *Electronic Journal of Statistics*, vol. 11, no. 1, pp. 2358–2396, 2017.

[9] P. Krishnaiah and B. Miao, "19 review about estimation of change points," vol. 7, pp. 375–402, 1988. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016971618807021X

[10] J. Bai, "Least squares estimation of a shift in linear processes," *Journal of Time*, vol. 15, no. 5, p. 453, 1994.

[11] C. Brochot, B. Bessoud, D. Balvay, C. A. Cuénod, N. Siauve, and F. Y. Bois, "Evaluation of antiangiogenic treatment effects on tumors' microcirculation by bayesian physiological pharmacokinetic modeling and magnetic resonance imaging," *Magnetic Resonance Imaging*, vol. 24, no. 8, pp. 1059–67, 2006.

[12] C. A. Cuenod, B. Favetto, V. G. Catalot, Y. Rozenholc, and A. Samson, "Parameter estimation and change-point detection from dynamic contrast enhanced mri data using stochastic differential equations," *Mathematical Biosciences*, vol. 233, pp. 68–76, 2011.

[13] "Characterisations of the wiener process," https://en.wikipedia.org/wiki/Wiener_process.

[14] "Characterisations of the wiener process," https://en.wikipedia.org/wiki/Stochastic_differential_equation.

[15] "Confidence interval," https://en.wikipedia.org/wiki/Confidence_interval.

[16] I. A. Eckley, P. Fearnhead, and R. Killick, "Ito's lemma," in *Computational finance*, D. Barber, A. Cemgil, and S. Chiappa, Eds. Cambridge University Press, 2011.

[17] "Metropolis–hastings algorithm," https://en.wikipedia.org/wiki/MetropolisHastings_algorithm.

[18] "Gibbs sampler in bayesian inference and its relation to information theory," https://en.wikipedia.org/wiki/Gibbs_sampling.

[19] M. Razo, "Moment dynamics generation," https://www.rpgroup.caltech.edu/chann_cap/software/moment_dynamics_system.html.