University of Essex

Doctoral Thesis

Sampling-Assisted non-parametric Inference of Intractable Models

Bo Zhang

Department of Mathematical Sciences

supervised by
Dr. Hongsheng Dai

April 14, 2022

Abstract

Likelihood function plays a fundamental role in most statistical analysis. Nevertheless, when the likelihood is too complex to compute in traditional ways, indirect inference or Approximate Bayesian Computation (ABC) is often adopted. This paper proposes a new approach to estimate the intractable likelihood function based on the Hermite polynomial expansion. Based on the estimated likelihood function, a maximum likelihood estimate and its large sample properties are achieved. Comparing to other existing methods, the proposed method provides more accurate estimation and are justified by asymptotic theories. Moreover, it does not require extra auxiliary models, which are necessary for indirect methods and are often not easy to find. The proposed method also has a very close link to ABC in the sense that it also simulates many pseudo data sets; however it does not need to compare the pseudo data with the raw data as what ABC does. Therefore, the proposed method does not suffer from the drawbacks of ABC; for example ABC always requires rich experience for choosing a proper value of the distance function, data summary statistics and threshold for acceptance rate.

Acknowledgement

I would like to express my sincere gratitude to my supervisor Hongsheng Dai, for giving me opportunity to do further research in statistical inference.

Besides my supervisor, my deep thanks goes to colleagues, including Daoji Li, Yanchun Bao and Jinchi Lv. They help me with the related paper writing and project research.

I would also thank the thesis committee who provide suggestions for revising.

I am extremely grateful to my parents for their love and sacrificing for education as well as guidance for my future.

Contents

Chapter 1

Introduction

The modern statistical analysis often relies on the parametric or nonparametric likelihood based on the underlying statistical models. Therefore, the likelihood function and the corresponding maximum likelihood estimate (MLE) are essential tools to draw inferences from unknown model parameters.

For simple statistical models, the closed form of likelihood functions is readily available. However, modern technologies have led to more complicated and much larger data sets in many research areas, such as genetics, finance, ecological systems, and astrophysics, which often require more sophisticated statistical modelling approaches (Pritchard et al., 1999; Van Duijn et al., 2009; Wood, 2010; Zhu et al., 2016). The statistical models for such problems may be so complicated that the likelihood functions no longer have a closed form, or subsequent calculations are too complicated to be performed. Specific examples include quantile distribution such as g-and-k model (Allingham et al., 2009) and Lotka-Volterra (LV) model (Boys et al., 2008), which do not have explicit likelihood function expressions.

In frequentists' view, one way to solve problems with such intractable likelihood functions is to use pseudo-likelihood inference to approximate the maximum likelihood estimator (Cox and Reid, 2004). Examples include analyses for spatial data

(Heagerty and Lele, 1998; Nott and Rydén, 1999), serial-correlated count data (Henderson and Shimakura, 2003), multivariate extremes (Coles, 2001, Chapter 8), pairs of censored failure times (Parner, 2001). The replacement of the true likelihood function by a pseudo-likelihood might perform well on estimation for certain problems. However, straightforwardly using it on an estimation of uncertainty and significance testing should not be recommended (Dodge, 2003). Sometimes, pseudo-likelihood even cannot avoid bias on parameter estimation compared to the maximum likelihood method (Van Duijn et al., 2009). When the data are time-series data or have strong correlations, indirect inference based on the auxiliary models may be more suitable (Gourieroux et al., 1993). However, such indirect methods may be of limited use since it requires the explicit expression on the relationship between auxiliary and original models (Rubio and Johansen, 2013).

Under a Bayesian framework, it is also extremely challenging to tackle problems with no explicit likelihood function, as it is impossible to obtain the analytical formula for the posterior function. Therefore, Markov chain Monte Carlo (MCMC) methods could not be adopted in such studies. Approximate Bayesian Computation (ABC) is an alternative approach (Drovandi et al., 2015; Karabatsos and Leisen, 2018) to overcome such likelihood-free problems. The ABC algorithm only assumes that we can simulate a pseudo data set $\boldsymbol{Y}_i^*$ from the underlying statistical models $f(\cdot|\boldsymbol{\theta})$ for any given parameter value $\boldsymbol{\theta}$, which is generated from the prior distribution $\pi(\cdot)$. See examples in Chapter 4. If the simulated pseudo data set $\boldsymbol{Y}_i^*$ is very *close* to the observed factual data set $\boldsymbol{Y}_{obs}$, then $\boldsymbol{\theta}$ will be accepted as a realisation from the posterior distribution.

In ABC Algorithms , the tolerance $\varepsilon = \|\boldsymbol{Y}_i^*, \boldsymbol{Y}_{obs}\|$ controls the distance between the pseudo data set $\boldsymbol{Y}_i^*$ and the real data $\boldsymbol{Y}_{obs}$, as a threshold for accepting a proposed sample $\boldsymbol{\theta}'$. It is easy to prove that as the tolerance $\varepsilon \to 0$, we have $\boldsymbol{Y}_i^* \to \boldsymbol{Y}_{obs}$ and the accepted Monte Carlo realisation $\boldsymbol{\theta}_i$ will be an exact sample generated from the

posterior distribution $f(\boldsymbol{Y}_{obs}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Such approximate Bayesian Computation, not focusing on evaluating the (approximate) likelihood function but the criteria of the comparison between the original and the simulated samples, is growing prosperously. For example, ABC algorithms currently can handle g-and-k, LV model, Stochastic Gompertz model and two-dimensional correlated geometric Brownian motion model (Picchini and Anderson, 2017). Note that if we choose uniform priors, the results of ABC approaches will be comparable to frequentist inference. Moreover, it shows that it performs better than other existing pseudo-likelihood approaches and indirect methods using auxiliary models (Fearnhead and Prangle, 2012).

Even though ABC methods can provide key parameter estimators when the likelihood is intractable, this method still has instability, low accuracy, and high computational complexity problems. Fearnhead and Prangle (2012) showed that for the g-and-k distribution, the original ABC models (Allingham et al., 2009) performed splendidly on the estimation for the location and scale parameters while patchily on skewness and kurtosis parameters. Although the semi-auto ABC method can improve the performance of ABC, the low acceptance probability and the stability of estimation are still troublesome (Fearnhead and Prangle, 2012). Meanwhile, there is still no hard rule on how to choose the distance function for measuring the difference between original observation $\boldsymbol{Y}_{obs}$ and simulated pseudo samples $\boldsymbol{Y}_i^*$ and on choosing the tolerance ε like a threshold determining whether a sample is accepted (Toni et al., 2009). For example, we may first work out a number of sample statistics based on the simulated data and real data, i.e. $S_1(\boldsymbol{Y}_{obs}), \cdots, S_k(\boldsymbol{Y}_{obs})$ and $S_1(\boldsymbol{Y}_i^*), \cdots, S_k(\boldsymbol{Y}_i^*)$ (Fearnhead and Prangle, 2012). However, there is no hard rule on how to choose such statistics and how many statistics shall be used. It is a compromise between the complexity of the calculation and the estimate's accuracy in ABC, which is still an open question.

This thesis proposes a more efficient parameter estimation method under the likelihood-free framework. The proposed methodology uses the Hermite Polynomial expansion to approximate the intractable likelihood and find the maximum likelihood estimator. Moreover, we can prove that our MLE based on Hermite polynomial expansion of the likelihood is still consistent and asymptotic to normal distribution. Therefore, unlike original or semi-auto ABC methods, which solely rely on users' experience or a large number of experiments to speculate a 'trade-off' between ABC posterior approximation error and MC approximation error (Fearnhead and Prangle, 2012), the proposed method could provide an accurate error estimate and computation complexity. In addition, compared to hundreds of summary statistics chosen in ABC (Allingham et al., 2009), the proposed method only requires the number of statistics no less than the dimension of the unknown parameter. Thus, the new likelihood-free approach proposed in this thesis is easy to use. Meanwhile, for the proposed method, there is no need to worry about the low acceptance rate, which is unavoidable in ABC due to the control threshold tuning parameter.

In the study, we mainly focus on the experiments for quantile distribution(g-and-k distribution) and LV models. Here g-and-k distribution is a typical case for a continuous distribution with several unknown parameters, which could be easily extended to other general distribution functions. Meanwhile, LV models consist of continuous variables (time t) and discrete states, which means our proposed methods could figure out some complex problems, even some discrete functions. Overall, if our proposed methods could solve these two types of problems, the bright future of widely used parameter inference could be foreseen while other examples provided in the thesis could be dealt with similarly.

The rest of the thesis is organized as follows. Chapter 2 provides some preliminaries for the study, such as Hermite Polynomials and Bootstrap algorithm. Chapter 3 presents several likelihood-free examples, including g-and-k models and LV models.

Chapter 4 introduces the existing methods (including Maximum Likelihood Estimate (MLE), Indirect Inference by pseudo-likelihood and ABC methods), for parameter inference. In Chapter 5, we propose the proposed method based on Hermite Polynomial expansion and establish the large sample properties of the estimate. In Chapter 6, we present the details on simulation studies which demonstrate our method performs much better than existing methods, and then we present an analysis for Covid-19 data to further illustrate the performance of the Hermite Polynomials Expansion method. Finally, chapter 7 provides a brief discussion.

Chapter 2

Preliminary

In general, we consider statistical problems or experiments in a probability space with notation $(\mathbf{\Omega}, \mathfrak{F}, \mathbb{P})$ where: $\mathbf{\Omega}$ stands for a set sometimes called outcomes or a sample space; $\mathfrak{F}$ is a σ-algebra of events and $\mathbb{P}$ is a probability measure equipped on the $\mathbf{\Omega}$. In this chapter, some types of functions in a probability space are taken for granted and some basic knowledge for Hermite Polynomials is supplied as well. Then, bootstrapping methods is one of the important sampling algorithms, which is also introduced.

2.1 Some Distribution Functions and Probability Preliminary

The distribution function is rather important in the statistical method. The main reason is that based on the summary of a random variable X, the expected value of the variable could be estimated in the statistical models. In this case, whether X is determined by a continuous or discrete distribution, the statistical inference could propose a rational and quantitative result. The cumulative distribution function (CDF) $F(x) = P(X \leq x), x \in \mathbb{R}$, one of the key distribution functions, indicate the

probability that a real-valued random variable X is no more than a value x.

In summary, CDF is regarded as a representative of the distribution function in this study. In addition, it finds that CDF above is considered as a type of left-tail distribution function. In a mirrored view, there exists a right-tail distribution function indeed, sometimes called the reliability function.

Definition 2.1. *Suppose that $F^{\complement}$ is a right-tail distribution.*

$$F^{\complement}(x) = 1 - F(x) = P(X > x), x \in \mathbb{R}.$$

For example, if a random variable $X \geq 0$, for a lifetime of a device, has a probability density function f, $F^{\complement}$ is the reliability function of X. Based on this case, the conditional probability for the lifetime function is proposed.

Definition 2.2. *A failure rate function or a hazard function $\hbar$ of X is defined by:*

$$\hbar(x) = \frac{f(x)}{F^{\complement}(x)}, x \geq 0.$$

Note that $F^{\complement}(x) = P(X > x)$. That is, $F^{\complement}(x)$ stands for the probability that the device could last more than time x. Denote that $\mathtt{d}x$ is an infinitesimal time interval. We have,

$$P(x < X < x + \mathtt{d}x | X > x) = \frac{P(x < X < x + \mathtt{d}x)}{P(T > x)} \approx \frac{f(x)\mathtt{d}x}{F^{\complement}(x)} = \hbar(x)\mathtt{d}x.$$

Thus, $\hbar(x)\mathtt{d}x$ represents for the approximate probability of failure happening within the interval $(x, x+\mathtt{d}x)$ under the situation that the device's lifetime larger than time x.

If the function f is continuous, it is a fact that $\frac{\mathtt{d}}{\mathtt{d}x}F^{\complement}(x) = -f(x)$. Hence,

$$\int_0^x \hbar(v)\mathtt{d}v = \int_0^x \frac{f(v)}{F^{\complement}(v)}\mathtt{d}v = \int_0^x \frac{-\frac{\mathtt{d}}{\mathtt{d}v}F^{\complement}(v)}{F^{\complement}(v)}\mathtt{d}v = -\ln F^{\complement}(x).$$

Therefore, $F^{\mathfrak{c}}(x) = e^{-\int_0^x h(v)\mathrm{d}v}$.

Certainly, there exists other ways to express distribution function as well, including the Empirical Distribution Function $\tilde{F}$.

Definition 2.3. *Suppose that a sequence of independent variables* $(X^{(1)}, X^{(2)}, ..., X^{(n)})$ *sampled from the same distribution of* X *with observed value* $(x^{(1)}, x^{(2)}, ..., x^{(n)})$.

$$\tilde{F}(x) = \frac{1}{n}\#\left\{i \in \{1, 2, ..., n\} : x^{(i)} \leq x\right\} = \frac{1}{n}\sum_{i=1}^{n} I_{x^{(i)} \leq x}, x \in \mathbb{R}.$$

Empirical Distribution Functions rely on the results of experiments, especially for transferring qualitative problems into quantitative analysis.

Consider that $\boldsymbol{X}$ is a multi-dimensional variable. We have a multivariate distribution or a joint CDF defined as follow,

Definition 2.4. *For* d *random variables* $X_1, ..., X_d$, *denote* $\boldsymbol{X} = (X_1, ..., X_d)$, *the joint CDF* $F_{\boldsymbol{X}}$ *given by*

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = P(X_1 \leq x_1, X_2 \leq x_2, ..., X_d \leq x_d). \tag{2.1}$$

There is often an inverse distribution function or so-called quantile function corresponding to each distribution function, due to the property of a monotonic increasing function.

Definition 2.5. *The quantile function* Q *or* F^{-1} *of* X *is defined by*

$$Q(\alpha) = F^{-1}(\alpha) = \inf\{x \in \mathbb{R} : F(x) > \alpha\}. \tag{2.2}$$

Based on the definition of quantile function, we will introduce quantile distribution in the next chapter.

2.2 Hermite Polynomials Preliminary

In the study, Hermite polynomials are adopted for approximating the likelihood function. The Hermite polynomial are a type of classic orthogonal polynomials in mathematics. Speaking of the scalar product $\int_a^b \xi_1(x)\xi_2(x)w(x)dx$ between two function $\xi_1(x)$ and $\xi_2(x)$, $w(x)$ is a non-negative function, so-called the weight function and the interval [a,b] is in the real line (sometimes $a \to -\infty$ or $b \to \infty$). Orthogonal sequences are defined as follow:

Definition 2.6. *The functions $\{\xi_m(x)\}$, not containing any null function, form the orthonormal system(a family of orthogonal functions), if:*

$$\int_a^b \xi_i(x)\xi_j(x)w(x)dx = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j, \end{cases} \tag{2.3}$$

where i and j are selected from $1, 2, 3.., m$.

In general, a pair of functions satisfying (2.3) are considered as orthogonal. Hermite Polynomials satisfy the basic property indeed.

2.2.1 Hermite Polynomials with uni-variate variable

It is supposed that Hermite polynomials, as a typical kind of orthogonal polynomials, can be distinguished from others on its special weight function, an exponential function. The simplest form of the weight function in Hermite polynomials is e^{-x^2} in the physics field, but for statistics or probability study it is proper to adopt the exponential function $e^{-\frac{x^2}{2}}$. Even though this thesis focus on the estimation of the complicated likelihood function, a typical statistical problem, here we prefer to use the former one e^{-x^2} because of the convenience in the usage of the property of the basic ones called physicists' Hermite polynomials.

Definition 2.7. The standard Hermite polynomials are defined as, for $m = 0, 1, 2, ...,$

$$\mathrm{h}_m(x) = \frac{p_m(x)}{(2^m m! \sqrt{\pi})^{\frac{1}{2}}}, \tag{2.4}$$

where

$$p_m(x) = (-1)^m e^{x^2} \cdot \frac{\mathtt{d}^m}{\mathtt{d}x^m} e^{-x^2}. \tag{2.5}$$

□

Standard Hermite polynomials are such that

$$\int_{-\infty}^{\infty} e^{-x^2} \mathrm{h}_i(x) \mathrm{h}_j(x) \mathtt{d}x = \delta_{ij} = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases} \tag{2.6}$$

Here is a fact Bonan and Clark (1990) that:

Lemma 2.1. *For a standard Hermite polynomial* $\mathrm{h}_m(x)$*, there exists positive constants* A_1*,* A_2 *and* A_3 *such that*

$$\mathrm{h}_m^2(x) exp(-x^2) \leq A_1 \sqrt{2m + 1 - x^2}\,, \text{ when } |x| \leq \sqrt{2m+1},$$

$$max_{x \in \mathbb{R}}\, \mathrm{h}_m^2(x) exp(-x^2) \leq A_2 m^{-1/6},$$

$$max_{x \in \mathbb{R}}\, \mathrm{h}_m^2(x) exp(-x^2) \geq A_3 m^{-1/6},$$

(for $m = 1, 2, 3, ...$*)*

From this lemma, we know that the upper boundary of the function $\mathrm{h}_m^2(x) exp(-x^2)$ only relies on the order of polynomials m. What's more, the function $\mathrm{h}_m^2(x) exp(-x^2)$ should be equivalent to $O(m^{-\frac{1}{6}})$ even when x rang over all real values. Further, the upper boundary of the absolute value of the Hermite funciton $\mathrm{h}_m(x) exp(-x^2)$ should be no more than $O(m^{-\frac{1}{12}})$. This property could be rather general for approximating an unknown continuous function by Hermite Expansion. Unlike the Taylor series,

there is no need to consider the convergence radius or strong condition for analytic function. This is the main reason why the Hermite polynomials are adopted instead of other polynomials for expansion in the thesis.

From Szegö (1959) we have the following lemma for Hermite expansion.

Lemma 2.2. *Suppose $f(x)$ is a piece-wise differentiable function on the interval $[a, b]$ and $\int_a^b f(x)^2 e^{-x^2} dx < \infty$. Then we have*

$$f(x) = \sum_{m=0}^{\infty} C_m \mathrm{h}_m(x) e^{-\frac{x^2}{2}}, \textit{ when } f(x) \textit{ is continuous}$$

where

$$C_m = \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} f(x) \mathrm{h}_m(x) \mathtt{d}x.$$

The result still holds if $[a, b]$ is changed to $(-\infty, \infty)$ and $\int_a^b f(x)^2 e^{-x^2} \mathtt{d}x < \infty$.

It ensures that the coefficient C_m should be a real number instead of becoming infinite due to the property of Hermite Polynomials. Due to the convergence of polynomials, we can use the Monte Carlo (MC) algorithm to estimate the value of C_m. The lemma indicates that regardless of the selection of x, Hermite polynomials could provide an explicit formula for a certain derivative function with proper coefficient. In the Taylor series, only an analytic function could be expressed at a single point with a small interval, which is a strong requirement indeed. After comparison, it is found that Hermite polynomial expansion could be more general and may provide better performance over the whole real value, which could be used for the convolution of function as well.

More details about Hermite Polynomials in Bateman (1953), Bonan and Clark (1990) and Krasikov (2004).

2.2.2 Hermite Polynomials with multi-variables

The multivariate Hermite Polynomials are in introduced in this part. Compared with uni-variate problem, the format for the multivariate Hermite Polynomials turns out to be more complex.

Definition 2.8. *For a system of polynomials, assume* $\boldsymbol{x} = (x_1, x_2, ..., x_d)$

$$\begin{aligned} \mathfrak{h}_{\boldsymbol{m}}(\boldsymbol{x}) &= \mathfrak{h}_{m_1,m_2,m_3,...,m_d}(x_1, x_2, x_3, ..., x_d) \\ &= (-1)^m e^{\varphi(\boldsymbol{x})} \frac{\partial^m}{\partial x_1^{m_1} \partial x_2^{m_2} ... \partial x_d^{m_d}} e^{-\varphi(\boldsymbol{x})}, \end{aligned} \tag{2.7}$$

where

$$m = m_1 + m_2 + ... + m_d, \boldsymbol{m} = (m_1, m_2, ..., m_d),$$

$$\varphi(\boldsymbol{x}) = (\boldsymbol{C}\boldsymbol{x}, \boldsymbol{x}) = \sum c_{jl} x_j x_l, j, l = 1, 2, 3, .., d.$$

and $\boldsymbol{C}$ *will be a fixed positive definite symmetric square matrix of real elements* c_{jl}*.*

To simplify the model in high-dimensional problem, we could let $\boldsymbol{C} = \boldsymbol{I}$ where $\boldsymbol{I}$ is the identity matrix. In this way, we have

$$\mathfrak{h}_{\boldsymbol{m}}(\boldsymbol{x}) = (-1)^m e^{\sum x_j^2} \frac{\partial^m}{\partial x_1^{m_1} \partial x_2^{m_2} ... \partial x_d^{m_d}} e^{-\sum x_j^2}. \tag{2.8}$$

In terms of standard Hermite polynomials,

$$\boldsymbol{h}_{\boldsymbol{m}}(\boldsymbol{x}) = \frac{\mathfrak{h}_{\boldsymbol{m}}(\boldsymbol{x})}{\prod_{l=1}^d (2^{m_l} m_l! \pi^{\frac{1}{2}})^{\frac{1}{2}}} = \prod_{l=1}^d \mathrm{h}_{m_l}(x_l), \tag{2.9}$$

and

$$\mathcal{H}_{\boldsymbol{m}}(\boldsymbol{x}) = \frac{e^{-\frac{\sum x_j^2}{2}} \mathfrak{h}_{\boldsymbol{m}}(\boldsymbol{x})}{\prod_{l=1}^d (2^{m_l} m_l! \pi^{\frac{1}{2}})^{\frac{1}{2}}} = \prod_{l=1}^d \mathrm{H}_{m_l}(x_l). \tag{2.10}$$

It is obvious that $\mathcal{H}_{\boldsymbol{m}}(\boldsymbol{s}) = \boldsymbol{h_m}(\boldsymbol{s})e^{-\frac{\sum s_j^2}{2}}$. As for expansion, $\boldsymbol{h_m}$ replaces h_m, and Lemma 2.2 is still tenable.

2.3 Bootstrap Introduction

Bootstrapping algorithms are one of the most common methods for re-sampling. In practice, observations may not be obtained easily or sometimes replicated experiments are hard to achieve. For instance, data for Covid-19 is provided in the study (Chapter 6 for Experiment), which cannot be repeatedly collected in the same place within a certain period. In this case, Bootstrapping algorithm could be taken as one of the option alternatives.

According to Dennis (2003), there are three types of problems that are often solved by bootstrapping: estimating standard errors; confidence intervals for unknown parameters; or test hypothesis. The one-sample problem is perfectly summarized by Efron (1979) via Bootstrap methods. A one-sample problem means that re-sample cases are selected from one distribution F for original samples $\boldsymbol{Y}_{obs}$. More details for samples from different distributions are provided in Efron (1979) as well but we do not need further study for that in the thesis. Then the simple algorithm for Bootstrapping is shown as follow.

Algorithm 1: Bootstrapping algorithm

1.Draw n independent bootstrap samples $\boldsymbol{Y}_1^*, ..., \boldsymbol{Y}_n^*$ with replacement from original observations $\boldsymbol{Y}_{obs}$;
2.Then we could calculate statistics corresponding to each bootstrap sample that $\hat{\boldsymbol{\theta}}^* = s(\boldsymbol{Y}^*), \boldsymbol{Y}^* = (\boldsymbol{Y}_1^*, ..., \boldsymbol{Y}_n^*)$, $\hat{\boldsymbol{\theta}}^*$ is an estimator for paramter $\boldsymbol{\theta}$;
3.Analyse the result of asymptotic or hypothesis testing if necessary;

Moreover, we could get the estimate error of the parameter $\boldsymbol{\theta}$ through the algorithm as well. Let $\hat{se}^{(b)} = \left\{\sum_{i=1}^n \frac{[\hat{\theta_i}^* - \hat{\theta}(\cdot)]^2}{n-1}\right\}^{\frac{1}{2}}$, where $\hat{\theta_i}^* = s(\boldsymbol{Y}_i^*), i = 1, 2, 3...n$ and

$\hat{\theta}(\cdot) = \sum_{i=1}^{n} \frac{\hat{\theta_i}^*}{n}$. As $n \to \infty$, it is clear that the limit of the estimate error $\hat{se}^{(b)}$ would turn out to be the population standard deviation.

Bootstrapping algorithm goes complex for time series problem. To estimate the parameter $\boldsymbol{\theta}$ with p-dimensional marginal distribution, we need re-sampling blocks $(X_{t+1}, ..., X_{t+l})$, corresponding to the original time series structure (Bühlmann, 2002). Let l be the length of the block. Suppose that the consecutive observations $Y_t = (X_{t-p+1}, ..., X_t), t = p, ..., n$. Then we have overlapping blocks of vectors $(Y_p, ..., Y_{p+l-1})$, $(Y_{p+1}, ..., Y_{p+l-1})$,...,$(Y_{n-l+1}, ..., Y_n)$. If we assume that $n-p+1 = k \cdot l$ where k is a natural number. We could construct k blocks independently by re-sampling with replacement. With block starting points $s_1, .., s_k$ drawn independently from $\mathcal{U}(\{p-1, .., n-l\})$, we can write k blocks that $\{Y_{s_i+1}, ..., Y_{s_i+l}\}_{i=1,2,...,k}$. In this way, we could obtain k blocks for bootstrapping.

Chapter 3

Examples

Suppose that a data set $\mathbf{y}$ follows a statistical model $f(\cdot|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the p-dimensional parameter vector of interest. Here and throughout the thesis, the data set $\mathbf{y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)'$ is an $n \times d$ matrix, where each $\boldsymbol{y}_j$ is a d-dimensional data point and n represents the total number of data points. Note that the data points $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ may be i.i.d. and they may also be correlated representing n discrete observations of a stochastic process. However, the explicit formula of f is not available in our study. There are many likelihood-free problems in the literature, such as chaotic ecological dynamic system model, quantile distributions and heteroscedastic time series models. Fearnhead and Prangle (2012) supplies many details on parameter inference for g-and-k distributions while Boys et al. (2008) provide ABC methods for LV models or prey-predator models that could be used for the simple ecological system. We present several examples here for illustration.

3.1 Quantile distribution

In this section, we focus on some examples of quantile distribution. Especially, the distribution is determined by a quantile function while the PDF could not be available for an explicit expression.

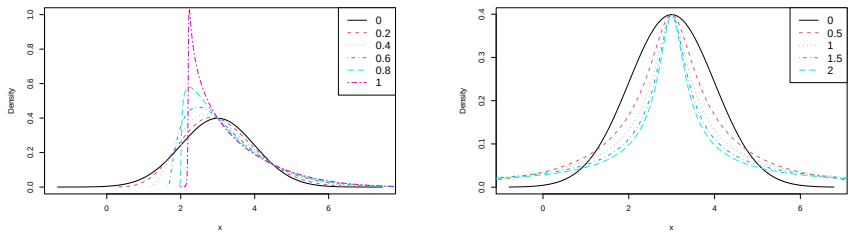

Figure 3.1: Example g-and-k densities, when $A = 3$ and $B = 1$. The first panel fixes $\mathtt{g} = 0$ and varies $\mathtt{k}$; the second panel fixes $\mathtt{k} = 0$ and varies $\mathtt{g}$.

3.1.1 The g-and-k distribution

The g-and-k distribution is defined by its quantile function (or inverse distribution function), but has no closed-form for its density function (Haynes, 1998; Allingham et al., 2009). Therefore, the corresponding likelihood can be considered intractable as its numerical calculation is rather costly (Rayner and MacGillivray, 2002; Drovandi and Pettitt, 2011; Fearnhead and Prangle, 2012). The quantile function of the g-and-k distribution is given by

$$Q_{\mathtt{gk}}(u; A, B, \mathtt{g}, \mathtt{k}) = A + B\left[1 + \mathtt{c}\frac{1 - \exp(-\mathtt{g}z_u)}{1 + \exp(-\mathtt{g}z_u)}\right] z_u(1 + z_u^2)^{\mathtt{k}}, \tag{3.1}$$

where z_u is the u-th quantile of standard normal distribution and $\mathtt{c}$ is a given value representing the level of asymmetry of the model. The parameter A is the position parameter, B is the scale parameter, $\mathtt{g}$ stands for the shape parameter that primarily affects skewness and $\mathtt{k}$ is the shape parameter that primarily affects kurtosis.

From Figure 3.1 or the expression, it is apparent that the g-and-k distribution would turn to be a normal distribution as $\mathtt{g} = 0$ and $\mathtt{k} = 0$. The first panel from Figure 3.1 provides more details that the top point of the curve rockets to 1 and the shape goes sharp as $\mathtt{k}$ increases. However, if $\mathtt{k}$ is fixed and $\mathtt{g}$ goes up, the second panel shows that the top point would remain unchanged while the convex curve would become narrow as well.

The practical problem is to estimate the unknown parameters $A, B, \mathtt{g}, \mathtt{k}$ based on i.i.d. observations $y_1, \cdots, y_n$ and a fixed value of $\mathtt{c}$, for instance $\mathtt{c} = 0.8$ (Rayner and MacGillivray, 2002). The dimension of the data point $d = 1$ and the dimension of the parameter $\boldsymbol{\theta} = (A, B, \mathtt{g}, \mathtt{k})$ is $p = 4$. The likelihood function is not available because we do not have explicit expression for f, the density of y_j.

3.1.2 The g-and-h distribution

The expression for the g-and-h distribution changes the format of parameter k from the g-and-k distribution, shown as follows:

$$Q_{\mathtt{gh}}(u; A, B, \mathtt{g}, \mathtt{h}) = A + B\left[1 + \mathtt{c}\frac{1 - \exp(-\mathtt{g}z_u)}{1 + \exp(-\mathtt{g}z_u)}\right] z_u \exp\left(\frac{\mathtt{h}z_u^2}{2}\right).$$

There is a fact that the same method could be adopted as well due to the similar format of quantile distribution if parameters from g-and-k distribution could be estimated. Therefore, a similar experiment is not displayed in this thesis.

3.1.3 The generalized λ distribution

Another example, generalized λ distribution, is proved to be estimated by ABC methods in Allingham et al. (2009), which also belongs to quantile distributions. The expression is shown as follow:

$$Q_{G\lambda D}(u; \lambda_1, \lambda_2, \lambda_3, \lambda_4) = \lambda_1 + \frac{\frac{u^{\lambda_3}-1}{\lambda_3} - \frac{(1-u)^{\lambda_4}-1}{\lambda_4}}{\lambda_2}.$$

In the distribution, the random variable u is drawn from a standard uniform distribution $\mathcal{U}(0, 1)$ as well. Similarly to parameters A and B from a g-and-k distribution, parameters λ_1 and λ_2 stands for location and scale factors respectively.

Then, parameters λ_3 and λ_4 "jointly" determine the trend of two side tails.

Even though the format of expression changes, the generalized λ distribution could also be solved when the method for a g-and-k distribution does not rely on the physical meaning of parameters. Luckily, both the ABC methods and our new proposed method cater for the requirement. This is because that regardless of a g-and-k distribution, a g-and-h distribution or a generalized λ distribution, pseudo data or samples could be obtained from generative models instead of the expression of the distributions. Once parameters from one of the three distributions are estimated by a method, the other distributions could follow in the same way. In our experiment, the g-and-k distribution is selected as an example for quantile distribution.

3.2 Stochastic kinetic networks (SKN) model

SKN models are often used for systems biology or biochemical data. For example, it is assumed that there are d chemical species in a fixed volume, which could inter-react with different transition rates (often rate constants) through each reaction channel. Then given the initial state of molecules of each species, the model to describe the trend of molecular population changing is rather important.

Traditionally, the system could be written as first-order, ordinary differential equations. It is assumed that a continuous and single-valued function $Y_j(t)(j = 1, 2, ..., d)$ stands for the number of j-th species in the volume at time t, as well as the process of chemical reactions, could be considered as a continuous transition, we

have:

$$\begin{aligned}
\frac{\mathrm{d}Y_1}{\mathrm{d}t} &= f_1(Y_1, Y_2, ..., Y_d), \\
\frac{\mathrm{d}Y_2}{\mathrm{d}t} &= f_2(Y_1, Y_2, ..., Y_d), \\
&... \\
\frac{\mathrm{d}Y_d}{\mathrm{d}t} &= f_d(Y_1, Y_2, ..., Y_d).
\end{aligned}$$

The function $f_j, j = 1, 2, 3, ..., d$ are actually determined by the system of experiments. In Algebra, there is no doubt that these equations could be solved numerically on a computer, by searching the roots. We cannot ignore the physical basis for the approach, especially when the time series is not a continuous process in a deterministic system. The apparent reason is that the molecular population must be observed as a discrete amount instead of a continuous value. Meanwhile, we cannot keep each experiment the same, for reactions between molecules could not be controlled and always happen randomly instead. Markov chains provide a good example explaining that even if the process is stable, the time series would not remain unchanged at each state for replicated experiments.

Gillespie (1977) provided a method to simulate the stochastic time evolution and set an example for molecular combination. Firstly, the "grand probability function" $P(Y_1, Y_2, ..., Y_d; t)$ is defined as the probability for the state that there are Y_1 molecules of species 1, Y_2 molecules of species 2,..., and Y_d molecules of species d in the system at time t (Gillespie, 1977).

In this case, we consider the system that the molecules are mixed perfectly in the volume, so that the inter-reaction through p different chemical reaction channels happens randomly with the transition rate $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)$. To be specific, a molecular combination go through μ-th reaction channel with a rate $\theta_\mu, \mu = 1, 2, ..., p$. These transition rates are supposed to be constants, which only depends on the properties

of the molecules in the physics and the construct of the system or environment where the temperature is a key factor and is supposed to be a constant as well over the process.

Then, the meaning of $\theta_\mu \mathtt{d}t, \mu = 1, 2, ..., p$ is explained as the "average probability" of μ-th molecular reactions within the infinitesimal time interval $\mathtt{d}t$. Here "average" indicates that in the next infinitesimal time interval $(t, t + \mathtt{d}t)$, the ratio between μ-th reactant combination occurring and total μ-th reactions at time t should be a constant θ_μ as a real value from 0 to 1. Further, the factor $\theta_\mu \mathtt{d}t$ product with the number of μ-th molecular combinations from total reactions at time t should be the probability of a μ-th reaction occurring in the volume within the time interval $(t, t + \mathtt{d}t)$.

Gillespie (1977) studied the probability for the state $(Y_1, ..., Y_d)$ at time t. As usual, the state $(Y_1, ..., Y_d)$ at time t could be given as an initial state and then the model could be simplified. We define a "reaction probability density function" $P(\tau, \mu)$ as follows:

$$P(\tau, \mu)\mathtt{d}\tau = P\Big(\big\{\text{a } \mu\text{-th reaction will occur in } (t + \tau, t + \tau + \mathtt{d}\tau)\big\} | (Y_1, ..., Y_d) \text{at time } t\Big).$$

An expression for $P(\tau, \mu)(0 \leq \tau < \infty \text{ and } \mu = 1, 2, .., p)$, is a joint probability density function. Following the assumption of the "average probability" mentioned above, the expression for $P(\tau, \mu)$ could be analytical.

In this case, we denote $a_\mu \mathtt{d}t = g_\mu \theta_\mu \mathtt{d}t$ where a function $g_\mu(\cdot)$ stands for the amount of μ-th reactant combinations, or how many μ-th reactions available within the infinitesimal interval, given the state $(Y_1, ..., Y_d)$ at time t. The relationship between $P(\tau, \mu)\mathtt{d}\tau$ and a_μ could be deduced as well:

$$P(\tau, \mu)\mathtt{d}\tau = P_0(\tau) a_\mu \mathtt{d}\tau,$$

where $P_0(\tau)$ means the probability of no reaction will occur in the time interval $(t, t + \mathtt{d}\tau)$. Meanwhile,the expression for $P_0(\tau)$ can be written as:

$$P_0(\tau) = e^{-\sum_{i=1}^{p} a_i \tau}.$$

This is because the definition of $P_0(\tau)$ provide the fact that $P_0(\tilde{\tau} + \mathtt{d}\tilde{\tau}) = P_0(\tilde{\tau})[1 - \sum_{i=1}^{p} a_i \mathtt{d}\tilde{\tau}]$. Denote $a_0 = \sum_{i=1}^{p} a_i$, we have:

$$P(\tau, \mu) = \begin{cases} a_\mu e^{-a_0 \tau} & \text{, if } 0 \leq \tau < \infty \text{ and } \mu = 1, 2, .., p, \\ 0 & \text{, otherwise.} \end{cases}$$

In the algorithm, the pair (μ, τ) should be generated randomly. If we draw two number r_1 and r_2 from two independent standard uniform distributions for a random pair (μ, τ), we have:

$$\tau = (1/a_0) \ln 1/r_1,$$

and μ satisfying:

$$\sum_{i=1}^{\mu-1} a_i < r_2 a_0 \leq \sum_{i=1}^{\mu} a_i.$$

To sum up, the algorithm for estimate samples is shown as follow.

Algorithm 2: simulate samples for SKN (Gillespie, 1977)

Set initial state $\boldsymbol{Y}(0) = (Y_1(0), ...Y_d(0))$, reaction constants or transition rates $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)$;
Calculate and save the quantities $a_1, a_2, ..., a_p$ and $a_0 = \sum_{i=1}^{p} a_i$;
Generate two random number r_1 and r_2 from standard uniform distributions, calculate τ and μ;
Using τ and μ, adjust values of $Y_i(t + \tau)$;

3.3 Lotka-Volterra or prey-predator model

Stochastic kinetic networks (SKN) models normally describe complex chemical reactions in a fixed volume containing the evaluation of different species with certain types of reactions (Golightly and Wilkinson, 2005). A special type of SKN model is the Lotka-Volterra model (or prey-predator model), where there are 2 species containing preys (Y_1) and predators (Y_2) as well as 3 reactions, as introduced in experiments for LV models.

The Lotka-Volterra model has been widely used to explore how competition influences the abundance of two competing species (Murray, 2002). There are 2 species containing preys (Y_1) and predators (Y_2) as well as three Markov transitions with transition rates θ_1, θ_2, and θ_3:

$$\begin{aligned} Y_1 &\xrightarrow{\theta_1} 2Y_1, \\ Y_1 + Y_2 &\xrightarrow{\theta_2} 2Y_2, \\ Y_2 &\xrightarrow{\theta_3} \emptyset, \end{aligned}$$

where $\emptyset$ denotes death of predators. Practically, we can only observe a sequence of observations (a time series) $\boldsymbol{Y}(t_j) = (Y_1(t_j), Y_2(t_j)), j = 1, \cdots, n$, at discrete time points $t_1, \cdots, t_n \in [0, T]$. We are interested in the unknown parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$. However, we cannot observe the continuous trajectory of $\boldsymbol{Y}(t), t \in [0, T]$ and the transition density between these discrete points is intractable. Therefore, the explicit likelihood is not available. For this example, the data point dimension $d = 2$ and the dimension of $\boldsymbol{\theta}$ is $p = 3$.

The model may also be denoted using transition probabilities over an infinitesimal

interval $(t, t+\mathtt{d}t]$, as:

$$P(\boldsymbol{Y}_{t+\mathtt{d}t}|\boldsymbol{Y}_t) = \begin{cases} \theta_1 y_1 \mathtt{d}t + o(\mathtt{d}t) & \text{if } Y_1(t+\mathtt{d}t) = y_1 + 1 \text{ and } Y_2(t+\mathtt{d}t) = y_2, \\ \theta_2 y_1 y_2 \mathtt{d}t + o(\mathtt{d}t) & \text{if } Y_1(t+\mathtt{d}t) = y_1 - 1 \text{ and } Y_2(t+\mathtt{d}t) = y_2 + 1, \\ \theta_3 y_2 \mathtt{d}t + o(\mathtt{d}t) & \text{if } Y_1(t+\mathtt{d}t) = y_1 \text{ and } Y_2(t+\mathtt{d}t) = y_2 - 1. \end{cases}$$

If there is no transition over the infinitesimal interval $(t, t+\mathtt{d}t]$, we have:

$$P(\boldsymbol{Y}_{t+\mathtt{d}t}|\boldsymbol{Y}_t) = \ 1 - (\theta_1 y_1 + \theta_2 y_1 y_2 + \theta_3 y_2)\mathtt{d}t + o(\mathtt{d}t), \quad \text{if } \boldsymbol{Y}(t+\mathtt{d}t) = (y_1, y_2).$$

Boys et al. (2008) provided the likelihood as well as the estimation of parameters $\boldsymbol{\theta}$ by Bayesian methods if the entire process $\boldsymbol{y}(t)$ was observed over the interval $[0, T]$.

However, it is impossible to fully acquire all the transition event times when collecting data. Instead, it is feasible to observe the number of species at discrete time points. For discretely-observed Markov process data, Boys et al. (2008) proposed a reversible jump (RJ) method, block updateing (BU) method and diffusion approximate (DA) method. In this thesis, we compare our proposed Hermite expansion method with those methods. All simulation studies are based on Lotka-Volterra model with true rate parameters $\theta_1 = 0.5, \theta_2 = 0.0025, \theta_3 = 0.3$ and initial state with $y_1(0) = 71$ and $y_2(0) = 79$. The process $\boldsymbol{y}_t = (y_1(t), y_2(t))$ is observed at n time points $t_1, \cdots, t_n$ in a time interval $[0, T]$ with $T = 40$. Such a setting was used in Boys et al. (2008) and the result is shown in Chapter Experiment.

Chapter 4

Existing Methodologies

In statistical inference, parameter estimation is one of the most typical problems. The maximum likelihood estimation (MLE) is a common parameter estimation method in the case of the probability distribution with a parametric family. Myung (2003) indicate that the strength of MLE includes "sufficiency, consistency, efficiency and parameterization invariance".

Definition 4.1. *Suppose probability density function (PDF)* $f(\mathbf{y}|\boldsymbol{\theta})$ *is for samples* $\mathbf{y} = (y_1, y_2, ..., y_n)$ *with an unique vector* $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$. *If* $\mathbf{y}$ *is fixed, the joint density turns to be a function for* $\boldsymbol{\theta}$, *defined as likelihood function* $L(\boldsymbol{\theta}|\mathbf{y})$, *that is,*

$$L(\boldsymbol{\theta}|\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta}). \tag{4.1}$$

Although the probability density function (PDF) mathematically equals to the likelihood function, the statistical meaning of the two functions is a bit different. Probability density function (PDF) generate some data with higher probability than other data given a stable set of parameters; while likelihood function stands for the likelihood of the parameter given the observed data $\mathbf{y}$.

As for mathematical models including diffusion models or signal detection theory, the requirement is that the proposed probability density $f(\mathbf{y}|\boldsymbol{\theta})$ for each observation

from the theoretical mechanism of generating the data is provided. Then, assuming that the observations $\{y_1, y_2, ..., y_n\}$ are independent and identically distributed, the likelihood is defined as

$$L(\boldsymbol{\theta}|Y_1 = y_1, Y_2 = y_2, ...Y_n = y_n) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}). \tag{4.2}$$

4.1 Maximum likelihood estimate

The larger value of likelihood means the more likely the parameter to approximate the true value which generates samples in the system. As usual, as the maximum value of the likelihood function stands for the most likely true value, the top point is a key for estimation in statistical inference.

Definition 4.2. *Denote observations* $\mathbf{y} = (y_1, y_2, ..., y_n)$ *with a likelihood function* $L(\boldsymbol{\theta}, \mathbf{y})$. *If there exists statistics* $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y})$ *satisfying* $L(\hat{\boldsymbol{\theta}}, \mathbf{y}) = \sup_\theta L(\boldsymbol{\theta}, \mathbf{y})$, $\hat{\boldsymbol{\theta}}$ *would be called Maximum Likelihood Estimate(MLE).*

In summary, MLE is a method to seek the maximum value of the likelihood function so that observed data seemed most likely. There are two main methods to approach MLE: one is to check all valid values, if possible, who are the MLE by definition; the other is by calculating the derivative of a likelihood function or log function. The former one is easy to determine whether the result is valid, comparing the definition straightly. The latter one requires more calculation and search a root for $\boldsymbol{\theta}$ so that $\frac{\mathtt{d}}{\mathtt{d}\boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathbf{y}) = 0$. In this case, the expression for the likelihood function should be supplied.

Here are some examples using in our experiments. As for g-and-k distribution, even we can write the likelihood function that

$$L_{\mathtt{gk}}(A, B, \mathtt{g}, \mathtt{k}|\boldsymbol{u}) = \prod_{i=1}^{n} \frac{\partial}{\partial u_i} Q_{\mathtt{gk}}^{-1} \left(\left\{ A + B \left[1 + \mathtt{c} \frac{1 - \exp(-\mathtt{g} z_{u_i})}{1 + \exp(-\mathtt{g} z_{u_i})} \right] z_{u_i} \right\} (1 + z_{u_i}^2)^{\mathtt{k}} \right).$$

Rayner and MacGillivray (2002) pointed out that even though the general expression of likelihood may not be solved, each value of the likelihood function could be obtained by the algorithm. Thus, the maximum likelihood estimate could be searched numerically, and called Numerical Maximum Likelihood Estimation(NMLE).

In terms of LV models, Boys et al. (2008) claims that

$$L_{LV}(\boldsymbol{\theta}|\mathbf{y}) = \left[\prod_{i=0}^{T-1}\prod_{j=1}^{n_i}\hbar_{\mu_{ij}}(\mathbf{y}(t_{\mu_{i,j-1}}),\theta_{\mu_{ij}})\right]\cdot e^{\int_0^T \hbar_0(\mathbf{y}(t),\boldsymbol{\theta})\mathtt{d}t}, \tag{4.3}$$

where $\hbar_\mu$ stands for the hazard function of reactions, n_i is the amount of reactions over the unit interval $(i, i+1]$ in which $t_{i,\mu}$ is one of time points, for μ-th reaction, Here $\mathbf{y} = \{\mathbf{y}(t) : t \in (i, i+1], i = 0, 1, 2, ..., T-1.\}$. And also it is assumed that

$$\hbar_\mu(\mathbf{y}, \theta_\mu) = \theta_\mu g_\mu(\mathbf{y}), \mu = 1, 2, 3. \tag{4.4}$$

Let $\hbar_0(\mathbf{y}, \theta) = \sum_{\mu=1}^3 \hbar_\mu(\mathbf{y}, \theta)$. Here $g_\mu(\cdot)$ is a known function and its properties are mentioned in SKN models (Chapter 3 for Examples).

According to the expression of likelihood, it is deduced that

$$\hat{\theta}_\mu = \frac{\sum_{i=1}^T r_{\mu i}}{\int_0^T g_\mu(\mathbf{y}(t))\mathtt{d}t} = \frac{r_\mu}{\int_0^T g_\mu(\mathbf{y}(t))\mathtt{d}t}, \mu = 1, 2, 3, \tag{4.5}$$

where $r_{\mu i}$ is the quantity of type μ reactions in a time interval $(i, i+1]$.

Boys et al. (2008) provides the evidence that the rate constants should be a independent distributed gamma variable based on the construction of the likelihood (4.5). In Bayes's theorem, if we takes θ_μ from a prior for gamma distribution $\Gamma(a_\mu, b_\mu)$, then

we have

$$\begin{aligned}
\theta_1|\mathtt{Y}_t &\sim \Gamma\left(a_1 + r_1, b_1 + \int_0^T y_1(t)\mathtt{d}t\right), \\
\theta_2|\mathtt{Y}_t &\sim \Gamma\left(a_2 + r_2, b_2 + \int_0^T y_1(t)y_2(t)\mathtt{d}t\right), \\
\theta_3|\mathtt{Y}_t &\sim \Gamma\left(a_3 + r_3, b_3 + \int_0^T y_2(t)\mathtt{d}t\right). \qquad (4.6)
\end{aligned}$$

In the Boys et al. (2008) study, it is also claimed that "maximum aposteriori (MAP) estimator" could be adopted in the situation. In addition, when $a_\mu = 1$ and $b_\mu = 0$, "MLE and MAP estimator coincide". According to the expression, it is apparent that completely observed species over the whole process are required for calculating the value $\int_0^T g_\mu(\mathbf{y}(t))\mathtt{d}t, \mu = 1, 2, 3$; otherwise, we cannot acquire a credible result through MLE methods.

4.2 Pseudo-Likelihood and Indirect Inference

Sometimes Pseudo-likelihood estimate could be called composite likelihood estimate or quasi-maximum likelihood estimate as well. In practice, it is not necessary or impossible to acquire the true PDF, which means we could not find an actual log-likelihood in general. Gourieroux et al. (1993) point out that there is no need to conduct a direct estimator of the parameter but a family of pseudo-likelihood functions should provide estimators of parameters satisfying consistence and normal asymptotic property.

The pseudo-likelihood estimate could be applied in g-and-k distribution. Even if the maximum likelihood function could be calculated in a complex format, the estimate of parameters could be computed more easily by a simple format for the combination of data. For example, position parameter estimator $\hat{A}$ follows the distribution of sample median and other parameters could be estimated in the similar as well

through their physical properties (Garcia et al., 2011; McCulloch, 1986). Then Garcia et al. (2011) extend the method from arbitrary quantiles to a pseudo-likelihood function, by Constrained Indirect Inference with a skewed-t score generator. And it is pointed out that including g-and-k distribution, the α-stable family of distributions with four parameters could follow in the same way. Moreover, Gourieroux et al. (1993) provide a more general format for Indirect Inference.

In the economic model, it is supposed that

$$y_t = G(y_{t-1}, x_t, u_t; \boldsymbol{\theta}), t = 1, 2, 3, ..., T, \tag{4.7}$$

where both y_t's endogenous variables and exogenous variables x_t's are observable whereas u_t as random errors are not observed. It is assumed that the initial value y_0 is provided and the random errors are always regarded as white noise with a known probability distribution.

To estimate the parameter $\boldsymbol{\theta}$, the auxiliary model or the criterion is introduced by a density function $f(y_t|y_{t-1}, x_t, \boldsymbol{\beta})$ where $\boldsymbol{\beta}$ is the auxiliary parameter. As usual, the expression of the function f is analytical and it is feasible to estimate parameters of the auxiliary model. We have

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{t=1}^{T} \log f(y_t|y_{t-1}, x_t, \boldsymbol{\beta}). \tag{4.8}$$

If we can generate a simulated sequence of endogenous variables $y_{t,i}(\boldsymbol{\theta}), i = 1, 2, ..., n$, we have

$$\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \arg\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \log f(y_{t,i}(\boldsymbol{\theta})|y_{t-1,i}(\boldsymbol{\theta}), x_t, \boldsymbol{\beta}). \tag{4.9}$$

The key point of the Indirect Inference is to search a proper value $\boldsymbol{\theta}$ so that $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ approximates to $\hat{\boldsymbol{\beta}}$. In the Wald approach, a quadratic form is adopted. If a matrix for measuring the distance between $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\beta}}$ is chosen, the value of $\boldsymbol{\theta}$ can be

calculated by least square method. That is,

$$\hat{\boldsymbol{\theta}}_{n,T} = \arg\min_{\boldsymbol{\theta}} \left[\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) - \hat{\boldsymbol{\beta}}\right]' \Omega_T \left[\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) - \hat{\boldsymbol{\beta}}\right], \tag{4.10}$$

where Ω_T is a sequence of positive definite matrix or weighting matrix converging to Ω. Here the simple format is $\Omega_T = \Omega$ which is a deterministic positive definite matrix. The estimator of parameter $\boldsymbol{\theta}$ could be determined in this way.

Above all, it is indicated that even though the density function f sometimes do not coincide totally with the distribution of y_t, parameters can still be estimated by the auxiliary model.

4.3 Posterior Distributions

In Bayesian Statistics, the parameter is regarded as a random variable with a certain distribution. The process of calculation, according to the real data combined with the experience of estimating the parameters of the distribution, tends to be complicated. In recent years, due to the improvement of computer calculation capability, various simulation-based computational methods have emerged, which greatly promotes Bayesian statistics. In the era of large data samples, Bayesian's has certain advantages over frequentists' inference. It is because that the method is based on sample data to estimate parameters. To sum up, the basic idea of Bayesian statistics is that the parameter posterior distribution function is proportional to a product of a prior probability and a likelihood. Based on the posterior distribution, the parameters are statistically analyzed in Bayesian views.

In Bayesian Inference, calculating or estimating these probability distributions can be divided into two aspects. Firstly, we must be able to calculate the likelihood of the data; that is, given a model with a set of parameters $\boldsymbol{\theta}$, we must specify each observation in the probability sample.

Last but not the least, the prior distribution of $\boldsymbol{\theta}$ should be considered. The parameters $\boldsymbol{\theta}$ should be derived analytically in an explicit way according to the prior information of the problem (Fearnhead and Prangle, 2012). To be simplified, parameters could be treated as random variables according to data while the estimation of the parameters is based on the probability distribution after observing some data. In this case, the prior distribution can be determined based on our previous understanding of the possibility of $\boldsymbol{\theta}$. Otherwise, a prior probability may explain the fact that there is no information supplied with $\boldsymbol{\theta}$ in prior, which means that we may use a prior probability without information, or the $\boldsymbol{\theta}$ widely dispersed within the allowable or supported range. Above all, whether the information is provided in advance, a posterior distribution should be updated through a conditional probability function of $\boldsymbol{\theta}$ given observing the data $\boldsymbol{Y}_{obs} = \{y_1, y_2, ..., y_n\}$. In a Bayes' view, we have:

$$\pi(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}) = \frac{L(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})\pi(\boldsymbol{\theta})}{\int L(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}, \tag{4.11}$$

where $\pi(\cdot)$ is the prior distribution function and $\pi(\cdot|\cdot)$ is the posterior distribution function.

In terms of analyzing the model, all information may be covered through the posterior distribution, such as "model checking and validation", prediction as well as "decision making" (Sisson et al., 2018). Providing that the likelihood function $L(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})$ or proposal function are tractable to obtain a close form, we can calculate the result directly; otherwise, Monte Carlo(MC) algorithm , as a common numerical approach, could contribute to the inference. This method utilizes the posterior distribution drawing the samples, $\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}, ..., \boldsymbol{\theta}_{(N)} \sim \pi(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})$, and based on the samples, the posterior distribution could be approximated by:

$$\pi(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}) \approx \frac{1}{N}\sum_{i=1}^{N} \delta_{\boldsymbol{\theta}_{(i)}}(\boldsymbol{\theta}), \tag{4.12}$$

where $\delta_{\boldsymbol{X}}(x)$ is defined as Dirac function, that is, $\delta_{\boldsymbol{X}}(x) = 1$ if $x \in \boldsymbol{X}$ otherwise $\delta_{\boldsymbol{X}}(x) = 0$. In other words, the samples were then filtered in such a way that those consistent with the posterior were retained otherwise were discarded. In this case, it is ensured by the law of large numbers that the distribution of filtered samples approximates the posterior distribution. That is,

$$\pi(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \delta_{\boldsymbol{\theta}_{(i)}}(\boldsymbol{\theta}).$$

In addition, the expectation of a function $g(\boldsymbol{\theta})$ under the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})$ could be obtain:

$$\begin{aligned} \mathbb{E}_{\pi}[g(\boldsymbol{\theta})] &= \int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} g(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})d\boldsymbol{\theta} \\ &\approx \frac{1}{N}\sum_{i=1}^{N} g(\boldsymbol{\theta}_{(i)}), \end{aligned} \tag{4.13}$$

where $\boldsymbol{\Theta}$ is the set of valid vectors of model parameters $\boldsymbol{\theta}$.

Through various MC methods or related sampling algorithms, it seems to be easy for Bayesian analysis by generating samples from the posterior distribution. However, Sisson et al. (2018) also point out that there are still two shortages for likelihood function numerical evaluation, including the "computationally prohibitive" and intractable format for the complete likelihood function. The former one often happens in the situation that the sufficiently large sample size requires feasible sufficient statistics. As for intractable likelihood function, quantile functions are one of the typical examples. Thus, if the target likelihood function for the model is intractable, it is invalid to draw samples from the posterior distribution in such a standard Bayesian simulation method. Contrasted with changing into a simpler model, a more effective method is to search an approximation to the original target model (Sisson et al., 2018). This posterior distribution approximation method requires replicated samples

from the model instead of the expression of the likelihood function, which is called a likelihood-free Bayesian method. Approximate Bayesian Computation (ABC), as a special case for likelihood free methods, would be introduced in the next section.

4.4 ABC algorithm

The cases in the previous chapter (Chapter Examples) are indeed difficult to solve in reality as the calculation for some of them tends to be much more impractical. In terms of the intractable likelihood problems, it requires the implementation of ABC methods since the limitation of standard Bayesian simulation methods.

The ABC method has been gradually developed in recent years and has become one of the hot spots in the Bayesian Computational method. The main idea of the method is to sample parameters given the prior distribution and then calculate the posterior distribution with these different parameters.

The ABC method, different from standard simulation methods, requires one more step for comparing the difference between observations and simulated samples.The ABC method constructs artificial data set $\boldsymbol{Y}^*$ by simulating the generated parameters in the model, as well as calculates the distance between the simulated data set $\boldsymbol{Y}^*$ and the observed data $\boldsymbol{Y}_{obs}$ on condition that the distance $\|\cdot\|$ and the tolerance value ε are set in the method. The distance $\|\cdot\|$ satisfied:

- $\|x, y\| \geq 0$, and $\|x, y\| = 0$ if and only if $x = y$;
- $\|x, y\| = \|y, x\|$;
- $\|x, z\| \leq \|x, y\| + \|y, z\|$;

If the distance is acceptable within the range of tolerance, the distribution with parameters simulating the artificial data set is considered to be similar to the true

posterior distribution. Then, the parameters can be directly estimated by the above numerical simulation process.

In a simple ABC algorithm, it is often assumed that the observation are univariate. Thus, we can set the distance function that $\|Y_{obs}, Y^*\| = Y_{obs} - Y^*$. But things go complex when observations and samples are multi-dimension. The selection of the distance expression $\|\boldsymbol{Y}_{obs}, \boldsymbol{Y}^*\|$ in the ABC method surely affects the result of the parameter estimation. More details are discussed in the next section. In this part, we start to introduce the original ABC method. Sometimes it is called Rejection ABC, which is the "oldest", and the "simplest" as there are no more steps required for sampling after rejection decisions (Nunes and Prangle, 2015; Karabatsos and Leisen, 2018). The calculation steps of the ABC method are as follows (assuming the a priori of the parameters is $\pi(\boldsymbol{\theta})$):

Algorithm 3: original ABC method (Rejection ABC)

for $i = 1$ to N do
repeat
Simulate parameters $\boldsymbol{\theta}^*_{(i)}$ from a prior distribution $\pi(\boldsymbol{\theta})$;
Generate synthetic data sets $\boldsymbol{Y}^*$ based on models and parameters $\boldsymbol{\theta}^*_{(i)}$;
If the observed data is close enough to the synthetic data set, that is,
$\|\boldsymbol{Y}_{obs}, \boldsymbol{Y}^*\| \leq \varepsilon$, then accept θ^*_i otherwise reject;
end for

The posterior distribution function in ABC satisfying:

$$\pi_{ABC}(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}) \propto \pi(\boldsymbol{\theta}) \int f(\boldsymbol{Y}|\boldsymbol{\theta}) I_\delta \left(\|\boldsymbol{Y}_{obs}, \boldsymbol{Y}\| \leq \varepsilon\right) d\boldsymbol{Y}, \tag{4.14}$$

where $f(\cdot)$ stands for the likelihood function, $I_\delta(\cdot)$ is a indicator function and the tolerance $\varepsilon > 0$.

In practice, the summary statistic $s(\cdot)$ often replace the role of samples or observations so that computational problem could be reduced especially for high dimensional $\boldsymbol{Y}_{obs}$ or large size samples. That is, the algorithm could adopt the summary statistic

$s(\boldsymbol{Y}^*)$ contrasted with $s(\boldsymbol{Y}_{obs})$, instead of $\boldsymbol{Y}^*$ with $\boldsymbol{Y}_{obs}$. Here $s(y)$ is sufficient for the parameter so that if we denote the density function as $f(y|\theta)$, then

$$f(s(y)|\theta) = \int \delta_s(s(y)) f(y|\theta) \mathtt{d}y,$$

where $\delta_s(\cdot)$ means that $I_\delta(\|s, s^*\| \leq \varepsilon)$. Due to the property that

$$\lim_{\varepsilon \to 0} \pi_{ABC}(\boldsymbol{\theta}|s(\boldsymbol{Y}_{obs})) \propto f(s(\boldsymbol{Y}_{obs})|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{4.15}$$

and

$$\pi(\boldsymbol{\theta}|s(\boldsymbol{Y}_{obs})) = \pi(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}),$$

we can draw a conclusion that samples from the distribution $\pi(\boldsymbol{\theta}|s(\boldsymbol{Y}_{obs}))$ cater for the requirement for the estimation on the posterior function $\pi(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})$. And the summary statistic can be a quantile or a moment statistic of y (expectation, variance, etc.). Sometimes $s(y)$ is non-sufficient, and convergence to the true posterior still works in ABC algorithms but the loss of information requires more computation to keep the precision of the posterior evaluation (Mengersen et al., 2013). Therefore, a sufficient summary statistic is the first choice for posterior approximation; otherwise, Sisson et al. (2018) suggests that the dimension of the summary statistic is supposed to be large enough as the informative representation for observations.

When a proper summary statistic $s(y)$ is adopted and the error value $\varepsilon \to 0$ can be tolerated, the output of the ABC algorithm should converge to the genuine posterior distribution with parameters. The evaluation sometimes coincides with the posterior distribution if ε is set as 0. In this scenario, the likelihood free method is no longer the ABC algorithm as the key idea of approximation is reduced. And the algorithm is not feasible as well especially for continuous observations, for the accept rate turns out to be zero and nearly all samples seem to be rejected before output.

Therefore, the tolerance ε, even we expect as close to 0 as possible, is often larger than 0 in ABC methods.

In practical applications, the estimation effect for the ABC method to a certain extent depends on the selection of the summary statistic $s(y)$ and the tolerable ε. $s(y)$ reflects all the information characteristics of the observation value $\boldsymbol{Y}_{obs}$, and the sufficient degree of $s(y)$ determines the size of the estimated deviation in the ABC method. The dimension of $s(y)$ is often less than that of observations or samples while the acceptance rate of the estimated parameters in the ABC method decreases with the increase of the dimension of the statistic $s(y)$; Except for the non-parametric system error of the method itself, and the error value ε controls the error range of accepting samples. In the model, the lower acceptance rate for samples means that more simulations are required to collect enough samples for inference and then time cost for computations would reduce the feasibility of the method; When the loosening of the error term will increase the acceptance rate of the parameter, the quality of the estimated output in the model, however, will reduce. Therefore, the balance between computational feasibility and simulation precision holds a critical role in the performance of the inference.

Above all, $s(y)$ should be low-dimensional and informative in the data set as well as the error term ε could be set small enough to ensure that the simulated data is close to the observed data. In this case, the ABC method is likely to have a good estimator. In other words, the choice for various tuning parameters in the ABC method, such as the summary statistics $s(y)$ and the error term ε, has always been a difficult point in the application of the ABC method.

In the application process, the ABC method requires two key choices to be carefully undertaken, containing summary statistics and tolerance (Nunes and Prangle, 2015). However, the uncertainty of the non-sufficient summary statistic and the requirement of prior information in the tolerance determination would be the main lim-

itation of the ABC methods. Moreover, sometimes even though summary statistics could solve the "curse of dimensionality problem", multi-dimensional observations or statistics still derive a new problem in which distance function is proper in each situation.

4.5 Improved ABC Algorithms

The ABC method is a kind of computational method for the likelihood of Bayesian inference based on computer data simulation rapidly developed. In recent years, this method has been widely used and has become a significant branch of the Bayesian computational method. Due to the limitation of basic ABC mentioned above, ABC methods often combine with other iterative algorithms to generate approximate samples from the evaluated posterior, including importance sampling(IS), Markov chain Monte Carlo (MCMC), sequential Monte Carlo (SMC) and other methods (Karabatsos and Leisen, 2018). These combinations are called ABC-IS, ABC-MCMC and ABC-SMC methods, respectively. In this section, we focus on the classification for improved ABC methods through the approximate likelihood instead.

4.5.1 Kernel-ABC

In the original ABC, it finds that the indicator function only determines whether to accept the simulated samples, wasting the information. To be specific, as $\varepsilon \to 0$, the marginal distribution of the parameter $\boldsymbol{\theta}$ vector is the true posterior. That is:

$$f(\boldsymbol{Y}_{obs}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \propto \lim_{\epsilon \to 0} \int I_\delta(\|s(\boldsymbol{Y}), s(\boldsymbol{Y}_{obs})\| \leq \epsilon) f(\boldsymbol{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \mathtt{d}\boldsymbol{Y}. \tag{4.16}$$

As we know, the indicator term $I_\delta(\cdot)$ only takes a value of 0 or 1. In this way, the statement for accepting samples in the algorithm can transform into a mathematical

formula with simplifying implementation. In this scenario, all accepted samples are undertaken equally in the system. That is, the indicator item cannot distinguish between two exactly different samples within the same tolerance: ones are from the estimated data set $\boldsymbol{Y}^*$ that is more close to or even equal to the observed data set $\boldsymbol{Y}_{obs}$; ones are furthest from those points in the data set $\boldsymbol{Y}_{obs}$. In other words, this qualitative judgement wastes the information in which some simulated samples satisfying $\|\boldsymbol{Y}_{obs}, \boldsymbol{Y}\| < \varepsilon$ seem to be more close to the true posterior distribution than samples that $\|\boldsymbol{Y}_{obs}, \boldsymbol{Y}\| = \varepsilon$. Thus, the former samples are more likely to be sampled from the true posterior distribution than the latter cases. The doubt whether we can make use of the information motivates the application of a smooth Kernel function replacing the indicator function.

Kernel methods could be used widely in the practical as positive definite kernels are available for any types of data sets (Hein and Bousquet, 2005).

A real-valued function $\mathcal{K}(\cdot, \cdot)$ on $\mathcal{X} \times \mathcal{X}$ can be called as a positive definite (p.d.) kernel on $\mathcal{X}$ if and only if:

$$\sum_{i,j}^{n} c_i c_j \mathcal{K}(x_i, x_j) \geq 0, \tag{4.17}$$

where for all $n \in \mathbb{N}, x_i \in \mathcal{X}, i = 1, 2, ..., n$, and for all $c_i \in \mathbb{R}, i = 1, 2, ..., n$. In the mathematical model, when the left item of the inequality equal to 0, it is implied that all $c_i = 0$. That is, "any finite matrix constructed by pairwise evaluation" should have fully positive (p.d.) or non-negative eigenvalues.

As Hein and Bousquet (2005) provides evidence for all common Hilbertian matrix on probability measures, using a smooth Kernel function to weight neighboring observations or samples, sometimes called Kernel smoother, is one branch of applications of Kernel methods. In ABC algorithms, the indicator function $I_\delta(\cdot)$ can be regarded as one of common kernel functions (uniform kernel) as well (Sisson et al., 2018). Now we adopt the standard smooth kernel function $\mathcal{K}_\epsilon(u)$, with $u = \|(Y^*), (Y_{obs})\|$

or $u = \|s(Y^*), s(Y_{obs})\|$, satisfying:

$$\mathcal{K}_\epsilon(u) = \frac{1}{\epsilon}\mathcal{K}(\frac{u}{\epsilon}), \tag{4.18}$$

where ϵ stands for the scale parameter or "bandwidth". Kernels, symmetric functions, have the properties as follows:

- $\mathcal{K}(u) \geq 0$ for all u ;
- $\int \mathcal{K}(u)du = 1$;
- $\int u\mathcal{K}(u)du = 0$;
- $\int u^2\mathcal{K}(u)du < \infty$;

Sisson et al. (2018) provides the proof of the convergence for the estimation with a kernel function. In the uni-variate case, the study even supplies the result of how accurate the simulation can be. We can propose the likelihood f_{ABC} for the K-ABC algorithm, with a fixed value of θ, that,

$$f_{ABC}(Y_{obs}|\theta) = \int K_\epsilon(\|(Y), (Y_{obs})\|)f(Y|\theta)\mathtt{d}Y. \tag{4.19}$$

In this formula, we can consider the function f_{ABC} as the ABC likelihood to approximate the true likelihood or density function $f(Y|\theta)$, for which ABC seems to be a regular Bayesian analysis with an approximated likelihood (Sisson et al., 2018). Assume that $Y, Y_{obs} \in \mathbb{R}$ and $\|u\| = |u| = |Y - Y_{obs}|$, then we can express the ABC likelihood at the point Y_{obs} through a Taylor expansion if the function $f(Y|\theta)$ is analytic. That is,

$$f_{ABC}(Y_{obs}|\theta) = f(Y|\theta) + \frac{1}{2}\epsilon^2 f''(Y_{obs}|\theta)\int u^2 K(u)\mathtt{d}u - ... \tag{4.20}$$

The expression for bias between ABC likelihood and the true likelihood for parameter θ is $b_\epsilon(y|\theta) = f_{ABC}(y|\theta) - f(y|\theta)$. The point estimation for the bias is,

$$\hat{b}_\epsilon(y|\theta) = \frac{1}{2}\epsilon^2\sigma_K f''(y|\theta), \tag{4.21}$$

where $\sigma_K = \int u^2 K(u)\mathtt{d}u$. Through the expression, it is apparent that bias reduces as the scale parameter ϵ become less. Even we can obtain the expression of bias estimation in theory while it is infeasible to calculate the second derivation of the likelihood in practice as the likelihood function is supposed to be intractable in advance.

If the same manner is adopted for the bias of ABC posterior approximation, we have,

$$\begin{aligned} b_\epsilon(Y_{obs}|\theta)\pi(\theta) &= f_{ABC}(Y_{obs}|\theta)\pi(\theta) - f(Y_{obs}|\theta)\pi(\theta) \\ &= \pi_{ABC}(\theta|Y_{obs})c_{ABC} - \pi(\theta|Y_{obs})c, \end{aligned} \tag{4.22}$$

where $c_{ABC} = \int f_{ABC}(Y_{obs}|\theta)\pi(\theta)\mathtt{d}\theta$ and $c = \int f(Y_{obs}|\theta)\pi(\theta)\mathtt{d}\theta$.

Then,

$$\begin{aligned} \tilde{B}_\epsilon(\theta|Y_{obs}) &= \pi_{ABC}(\theta|Y_{obs})c_{ABC} - \pi(\theta|Y_{obs})c \\ &= \frac{b_\epsilon(Y_{obs}|\theta)\pi(\theta)}{c_{ABC}} + \pi(\theta|Y_{obs})\left(\frac{c}{c_{ABC}-1}\right). \end{aligned} \tag{4.23}$$

It is found that the function of θ could describe the difference between ABC posterior approximation and the true posterior. With a fixed point of θ, b_ϵ turns to be 0 as $\epsilon \to 0$ and then $\tilde{B}_\epsilon$ goes to 0 as well. Therefore, to choose a proper scale parameter θ seems to be more significant than to determine which type of kernel function in the K-ABC (Sisson et al., 2018).

As for a smooth kernel for multivariate relations, let $\boldsymbol{u} = \boldsymbol{Y}^* - \boldsymbol{Y}_{obs}$, where

$\boldsymbol{u} = (u_1, ..., u_n)$, $\boldsymbol{Y}^* = (y_1^*, ..., y_n^*)$ and $\boldsymbol{Y}_{obs} = (y_{obs,1}, ..., y_{obs,n})$, so we have $u_i = y_i^* - y_{obs,i} (i = 1, 2, ..., n)$. Then we can write $\mathcal{K}_\epsilon(\boldsymbol{u}) = \prod_{i=1}^n \mathcal{K}_{\epsilon_i}(u_i)$, where the scale parameter $\mathcal{K}_{\epsilon_i}(u_i)$ of each individual kernel function can be different. If we denote that:

$$f_{ABC}(\boldsymbol{\theta}|s(\boldsymbol{Y}_{obs})) = \int \pi(\boldsymbol{\theta}) \mathcal{K}_\epsilon \left(|s(\boldsymbol{Y}) - s(\boldsymbol{Y}_{obs})|\right) f(s(\boldsymbol{Y})|\boldsymbol{\theta}) \mathtt{d}\boldsymbol{Y}.$$

then the ABC posterior is that:

$$\pi_{ABC}(\boldsymbol{\theta}|s(\boldsymbol{Y}_{obs})) \propto \pi(\boldsymbol{\theta}) f_{ABC}(\boldsymbol{\theta}|s(\boldsymbol{Y}_{obs})).$$

Similarly, the bias between ABC approximation and the true posterior could be extended from the uni-variate case mentioned above.

All in all, K-ABC as one of improved the ABC algorithms is shown as follow.

Algorithm 4: Importance sampling implementation of ABC method (Fearnhead and Prangle, 2012)

a proposal density $g(\boldsymbol{\theta})$, with $g(\boldsymbol{\theta}) > 0$ if $\pi(\boldsymbol{\theta}) > 0$
for $i = 1 : N$ do
repeat
1. Simulate parameters θ_i^* from function $g(\theta_i)$.
2. Generate artificial data sets $\boldsymbol{Y}^*$ based on models and parameters θ_i^*.
3. Calculate the statistical data $s(\boldsymbol{Y}_{obs}), s(\boldsymbol{Y}^*)$ corresponding to the observation data $\boldsymbol{Y}_{obs}$ and the artificial data set $\boldsymbol{Y}^*$.
4. If the observed data is close enough to the artificial data set, that is, with probability $K_\epsilon(s(\boldsymbol{Y}_{obs}) - s(\boldsymbol{Y}^*))$, then set $w_i = \frac{\pi(\theta_i^*)}{g(\theta_i^*)}$ otherwise $w_i = 0$.
end for

4.5.2 EL-ABC

The empirical likelihood was first proposed by Owen in 1988, who pointed out that this model had similar properties in convergence and simpler calculations compared with common likelihood (Owen, 1988). Other than K-ABC or related methods focusing on the density function f, the empirical likelihoods assume the constraints

to determine the value of parameter θ in advance. Unlike Rejection ABC models, empirical likelihood has such outstanding points: improved performance to estimate divergence on a set with large sample size; loosening the importance of a tolerance zone or a non-sufficient summary statistics; better estimation accuracy; and reduced computational complexity (Mengersen et al., 2013).

The empirical likelihood is a kind of nonparametric likelihood in which a related, interesting parameter θ, such as the moment of the distribution, is first defined and then the likelihood can be considered as a performance function. To be specific, it is assumed that the dataset $y = (y_1, y_2, ..., y_n)$ are independent and identically distributed from the unknown distribution function f. Y is the random variable following the f distribution. The likelihood can be written as:

$$L_{el}(\theta|y) = \max_{\boldsymbol{p}} \prod_{i=1}^{n} p_i, \tag{4.24}$$

where a vector $\boldsymbol{p}$ subject to the constraints $\{\boldsymbol{p} \in [0,1]^n, \boldsymbol{p} = (p_1, p_2, ..., p_n), \sum p_i = 1, \sum_i p_i g(Y, \theta) = 0\}$. Here $g(\cdot)$ is a vector-valued function, called estimating function. In terms of a vector mean, we can take $g(Y, \theta) = Y - \theta$ and then $\hat{\theta} = \bar{Y}$. In the case of one dimension, if $\theta = \mathbb{E}_f[Y]$, the empirical likelihood of θ is the maximum of the product $p_1 p_2 ... p_n$ under the condition $p_1 y_1 + ... p_n y_n = \theta$. In addition, the empirical likelihood can be performed with fast calculation by the 'emplik' package built on the Newton-Lagrange algorithom in the application of R (Mengersen et al., 2013).

Within proper constraints, the empirical likelihood has been proved to have the properties of convergence in theory. When the data is non-independent and identically distributed, some potentially independent and identically distributed structures can be built to calculate the empirical likelihood. The limit of empirical likelihood should reflect the information in characteristics of the model and ensure the likelihood convergence.

The empirical likelihood is used as a common likelihood and combined it with the ABC algorithm to obtain an improved algorithm, which is called Bayesian Computation with an empirical likelihood (BCel) or Empirical Likelihood ABC Method (EL-ABC). The main idea of the EL-ABC is shown as follow:

Algorithm 5: EL-ABC method

For $i = 1 : N$ do;
repeat;
1. Draw θ_i^* from the prior function$\pi(\theta_i)$;
2. Calculate the likelihood $L_{el}(\theta_i^*|y)$ with θ_i^* ;
3. Set a value to important sampling weight and let $w_i = L_{el}(\theta_i^*|y)$.;
end for

The sampling results of EL-ABC are N parameters with weight w_i, which is similar to the important sampling method for the parameter $\boldsymbol{\theta}$ with the function $g(\boldsymbol{\theta})$ that:

$$\hat{g}(\boldsymbol{\theta}) = \frac{\sum_{i=1}^{N}[w_i \cdot g(\theta_i)]}{\sum_{i=1}^{N} w_i}. \tag{4.25}$$

In the process, when $g(\boldsymbol{\theta})$ is the mean function of $\boldsymbol{\theta}$, the value of the parameter estimation of $\boldsymbol{\theta}$ can be obtained. The performance of the EL-ABC algorithm can get the effective sample size (ESS):

$$ESS = \frac{1}{\sum_{i=1}^{N}[\frac{w_i}{\sum_{j=1}^{N} w_j}]^2}. \tag{4.26}$$

In the calculation, ESS estimates the number of independent and identically distributed samples, whose value is usually between 1 (for unsatisfactory results) and N (for perfect performance results). It is indicated that the EL-ABC method, for posterior distribution estimate, has the advantage in which the known prior information of the parameters can be fully considered in the parameter estimation.

EL-ABC has the limitation of estimation as well. It is apparent that a proper choice of the constraint requires experience in prior. Even data cloning sometimes

reduces the requirement by replacing the constraints, trouble for poor summary statistics fitting the model is motivated as well. Mengersen et al. (2013) provides a feasible method through replicated experiments to check the approximation precision.

4.5.3 Semi-automatic ABC

In the Fearnhead and Prangle (2012) study, Semi-automatic ABC methods provided the evidence that the chosen summary statistics can be equal to ABC posterior means $\mathbf{E}(\boldsymbol{\theta}|\boldsymbol{Y})$. The algorithm is to simulate samples for parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)$, after calculating the summary statistic.

Algorithm 6: Semi-automatic ABC method (Fearnhead and Prangle, 2012)

1. a pilot run of ABC to get posterior mass;
2. simulate parameters $\boldsymbol{\theta}$ and samples $\boldsymbol{Y}^*$;
3. using the simulated parameters and samples to estimate the summary statistics $s(\boldsymbol{Y}^*)$;
4. run ABC with the selected summary statistics $s(\boldsymbol{Y}^*)$

In the regression models of the algorithm, we choose a vector-valued function or candidate summary statistics $\tilde{\boldsymbol{s}}(\boldsymbol{x})$, for example, $\tilde{\boldsymbol{s}}(\boldsymbol{x}) = (\boldsymbol{x}, \boldsymbol{x}^2, \boldsymbol{x}^3, \boldsymbol{x}^4)$ (Fearnhead and Prangle, 2012). Then, we could get transformations of the data, $\tilde{\boldsymbol{s}}(\boldsymbol{Y}^{(j)}), j = 1, .., N$, corresponding to the each parameter $\theta_i^{(j)}$ estimate. The fitting model is that,

$$\theta_i = \mathbf{E}(\theta_i|\boldsymbol{Y}) + \varepsilon_i = \beta_{0,i} + \beta_i \tilde{\boldsymbol{s}}(\boldsymbol{Y}) + \varepsilon_i, \tag{4.27}$$

where ε is the residual and β_i is a coefficient. As the constant $\beta_{0,i}$ would play no role for Bayesian inference, we focus on the term $\beta_i \tilde{\boldsymbol{s}}(\boldsymbol{Y})$ for i-th summary statistic estimate. The resulting estimate $s(\boldsymbol{Y}) = \hat{\boldsymbol{\beta}} \tilde{\boldsymbol{s}}(\boldsymbol{Y})$, where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)$, can be regarded as the summary statistics in the Rejection ABC algorithm ((Park et al., 2016)).

In this way, the algorithm constructs the summary statistics with regression mod-

els. The study of Fearnhead and Prangle (2012) suggests that the pilot experiment in advance requires summary statistics as well as explanatory variables for the linear models, to replace a arbitrary choice of summary statistics. It also points out that the approach of Semi-automatic ABC is more robust for optimal choices of summary statistics than standard ABC based on the pilot run. However, Park et al. (2016) indicates that there seems to be no evidence of calculating analytically.

Chapter 5

Hermite MC Method

As existing methods are provided before, this chapter aims to supply more details of our new proposed method on how to approximate the likelihood function and estimate parameter by Hermite Polynomials. Further, we would provide the proof to ensure the feasibility of our method, according to large sample theories. Then, the main algorithm for our method are summarised. And we could extend the conclusion to infinite dimensions space for parameters. Moreover, multi-variate Hermite expansion for likelihood free inference is available as well.

5.1 Hermite expansions for the likelihood function

Now we will present a proposed method which can provide a consistent estimator for the unknown model parameters when the likelihood function is not available. Our methodology only requires that given any model parameter value $\boldsymbol{\theta}$, a pseudo data set can be simulated from the underlying statistical model.

First, we need to construct M independent sample statistics based on the data

set $\mathbf{y}$, denoted as $S_1, \cdots, S_M$ with

$$S_i = \mathbb{T}_i(\mathbf{y}), \tag{5.1}$$

where $\mathbb{T}_i$ is the function to calculate the ith statistic. We also require that all chosen statistics S_i, $i = 1, \cdots, M$, such as sample quantiles (Fearnhead and Prangle, 2012), converge to a certain value s_i^* (the population characteristics) in probability. Such statistics are easy to find since their convergence properties are guaranteed by the large sample theorem and central limit theorem.

The independence property of the chosen statistic can be guaranteed for practical problems, by choosing appropriate $\mathbb{T}_i$ (such as a proper subset of $\mathbf{y}$) to construct each S_i. For example, S_i may depend on only $K < n$ data points of $\mathbf{y}$. Then the rate of convergence $S_i \to s_i^*$ is $O_{\mathtt{p}}\left(K^{-\frac{1}{2}}\right)$(Liu and Yang, 2012). Such independent statistics $S_i, i = 1, \cdots, M$ can even be achieved for discretely observed Markov processes by designing $\mathbb{T}_i$ appropriately. Practical details will be provided in Section 6.1. For notation simplicity, throughout this section we require that each of $S_1, \cdots, S_M$ is constructed based on $K = n/M$ different data points of $\mathbf{y}$. Further, the density for S_i will depend on the number of observations in $\mathbf{y}$ which are used to construct S_i, which is therefore denoted as $f_{i,n}(s|\boldsymbol{\theta})$.

The likelihood function based on these M statistics $(S_1, \cdots, S_M) = (s_1, \cdots, s_M)$ is given by

$$\tilde{l}(\boldsymbol{\theta}|\boldsymbol{s}) \propto \prod_{i=1}^{M} f_{i,n}(s_i|\boldsymbol{\theta}). \tag{5.2}$$

Because f is unknown, the densities $f_{i,n}(s_i|\boldsymbol{\theta})$ are not available either. In this thesis, we approximate the densities $f_{i,n}(s_i|\boldsymbol{\theta})$ via Hermite expansions. We first give some background and results on Hermite expansions.

The standard Hermite polynomials are provided in the Chapter 2 (Definition 2.7)

and the standard Hermite functions are defined as

$$\mathrm{H}_m(x) = \mathrm{h}_m(x)e^{-\frac{x^2}{2}}, \tag{5.3}$$

for $m = 0, 1, 2, \cdots$.

Standard Hermite polynomials defined above are orthogonal with respect to the weight function e^{-x^2}, that is,

$$\int_{-\infty}^{\infty} e^{-x^2}\mathrm{h}_i(x)\mathrm{h}_j(x)\mathtt{d}x = \int_{-\infty}^{\infty} \mathrm{H}_i(x)\mathrm{H}_j(x)\mathtt{d}x = \delta_{ij} = \begin{cases} 0, & i \neq j, \\ 1, & i = j\ . \end{cases} \tag{5.4}$$

It follows from Lemma 2.2 that the Hermite expansion of the function $f_{i,n}(s_i|\boldsymbol{\theta})$ is

$$f_{i,n}(s_i|\boldsymbol{\theta}) = \sum_{m=0}^{\infty} C_m^{(i)}(\boldsymbol{\theta})\mathrm{H}_m(s_i), \tag{5.5}$$

where

$$C_m^{(i)} \ := \ C_m^{(i)}(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} f_{i,n}(v|\boldsymbol{\theta})\mathrm{H}_m(v)\mathrm{d}v. \tag{5.6}$$

Note that for simplicity we use $C_m^{(i)}$, dropping the notation $\boldsymbol{\theta}$ in $C_m^{(i)}(\boldsymbol{\theta})$.

In theory, statistics $S_i, i = 1, \cdots, s_M$ can be chosen to be correlated, i.e. we may choose M correlated statistics $\boldsymbol{S} = (S_1, \cdots, S_M)$. Then equation (5.5) will become an M-dimensional Hermite polynomial expansion for the multivariate density of $\boldsymbol{S}$, $f_n(s_1, \cdots, s_M|\boldsymbol{\theta})$. We focus on a independent statistic S_i in our simulation studies and data analysis, because computationally one-dimensional Hermite expansion (5.5) is much easier to be implemented than multi-dimensional Hermite expansion. Detailed methodology results for such multi-dimensional cases is presented in the chapter above, while in the main thesis we will focus on using the independent

statistic.

5.2 Working likelihood function estimation

The likelihood $\tilde{l}(\boldsymbol{\theta}|\boldsymbol{s})$ in (5.2) and $f_{i,n}$ in (5.5) cannot be used in practice, because it is impossible to evaluate $C_m^{(i)}$ explicitly. However, if $\boldsymbol{\theta}$ is given, we can simulate N_n random realisations $v_i^{(1)}, \cdots, v_i^{(N_n)}$ from $f_{i,n}(\cdot|\boldsymbol{\theta})$, by drawing $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(N_n)}$ from a generative model of the distribution $f(\mathbf{x}|\boldsymbol{\theta})$ and then calculating a statistic $v_i^{(k)} = \mathbb{T}_i\left(\mathbf{x}^{(k)}\right)$, as discussed in Experiment chapter. Here N_n is a large integer depending on the sample size n, such that $N_n \to \infty$ as $n \to \infty$. Then the estimation for $C_m^{(i)}$ is given by

$$\widehat{C}_m^{(i)} := \widehat{C}_m^{(i)}(\boldsymbol{\theta}) = \frac{1}{N_n} \sum_{k=1}^{N_n} \mathrm{H}_m(v_i^{(k)}), \tag{5.7}$$

where the dependence on $\boldsymbol{\theta}$ is because the simulation of $v_i^{(k)}$ depends on $\boldsymbol{\theta}$. In our proof, we also drop off $\boldsymbol{\theta}$ in $\widehat{C}_m(\boldsymbol{\theta})$ for simplicity.

Therefore, an estimation for $f_{i,n}(s_i|\boldsymbol{\theta})$ is given by

$$\hat{f}_{i,n}(s_i|\boldsymbol{\theta}) \;=\; \sum_{m=0}^{q(N_n)} \widehat{C}_m^{(i)}(\boldsymbol{\theta}) \mathrm{H}_m(s_i), \tag{5.8}$$

where $q(N_n)$ is a threshold for the order of the Hermite Polynomial Expansion satisfying $q(N_n) \to \infty$, but $\frac{q(N_n)}{N_n} \to 0$ as $n \to \infty$.

Further we have the following working likelihood function as

$$\hat{l}(\boldsymbol{\theta}|\boldsymbol{s}) \propto \prod_{i=1}^{M} \hat{f}_{i,n}(s_i|\boldsymbol{\theta}). \tag{5.9}$$

This result tells us that we can actually use (5.9), instead of the original $f(\cdot|\boldsymbol{\theta})$ to find the MLE of $\boldsymbol{\theta}$. The following section provides the theoretical justification for

the MLE based on the working likelihood function (5.9).

5.3 Large sample theories

Denote by $\boldsymbol{\theta}^*$ the true parameter and $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$, a closed-ball neighbourhood of $\boldsymbol{\theta}^*$. Let $\hat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$, which is defined as $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}}\, \hat{l}(\boldsymbol{\theta}|\boldsymbol{s})$. To investigate the large-sample properties of $\hat{\boldsymbol{\theta}}$, we assume that the sample statistic $S_i, i = 1, 2, .., M$ satisfy the following condition.

Suppose that each $\boldsymbol{y}_i$ of the dataset $\mathbf{y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)$ follows distribution density $f(\cdot|\boldsymbol{\theta}^*)$ and each of the independent statistic $S_1, \cdots, S_M$ is constructed based distinct $K_n = n/M$ observations in $\mathbf{y}$ and they satisfy the following conditions.

Condition 5.1. *$S_i \to s_i^*$ almost surely and $\mathbb{E}|S_i - s_i^*|^2 \to 0$, as $n \to \infty$ and the central limit theorem holds for S_i, i.e. $\sqrt{K_n}(S_i - s_i^*) \Rightarrow \mathcal{N}(0, \sigma_i^2)$;*

Condition 5.2. *For any i, the density function $f_{i,n}(s|\boldsymbol{\theta})$ for S_i is twice differentiable with respect to s and $\boldsymbol{\theta}$ and we can write*

$$\frac{\mathtt{d} f_{i,n}(s|\boldsymbol{\theta})}{\mathtt{d}s} f_{i,n}(s|\boldsymbol{\theta})^{-1} = \rho_n \cdot g_{i,n}(s|\boldsymbol{\theta}), \tag{5.10}$$

where $\rho_n = O(K_n)$ and the function $g_{i,n}$ is such that $\sup_s |\frac{\mathtt{d}}{\mathtt{d}s} g_{i,n}(s|\boldsymbol{\theta})| = O(1)$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Condition 5.3. *For any i, the density function $f_{i,n}(s|\boldsymbol{\theta})$ such that*

$$\boldsymbol{\phi}_i(s, \boldsymbol{\theta}) := \frac{\partial f_{i,n}(s|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f_{i,n}(s|\boldsymbol{\theta})^{-1} = \rho_n \cdot \zeta_{i,n}(s|\boldsymbol{\theta}), \tag{5.11}$$

where $\zeta_{i,n}$ is such that for fixed value s, $\sup_{\boldsymbol{\theta}} |\zeta_{i,n}(s|\boldsymbol{\theta})| = O(1)$, $\sup_{\boldsymbol{\theta}} |\frac{\mathtt{d}}{\mathtt{d}s} \zeta_{i,n}(s|\boldsymbol{\theta})| = O(1)$ and $\sup_{\boldsymbol{\theta}} |\frac{\partial}{\partial \boldsymbol{\theta}} \zeta_{i,n}(s|\boldsymbol{\theta})| = O(1)$ (the supreme is taken in $\boldsymbol{\Theta}$).

The following Lemma 5.1 tells us that the quantile statistic for independent observations satisfy Condition 5.1,5.2,5.3. If M statistics are needed, we can construct an α_i-quantile statistic with $\alpha_i = \frac{i}{M+1}, i = 1, \cdots, M$ and each of them is based on K_n observations. In addition, the quantile statistic for discretely observed Markov processes also satisfy Condition 5.1,5.2,5.3, which will be proved in Lemma 5.3 in the supplementary file. Therefore, throughout this thesis, we will consider the quantile-related statistic, which were also used dominantly in the literature (see Fearnhead and Prangle (2012) and the references therein). Note that the following lemma is presented for uni-variate data points only, for notation simplicity. For higher d-dimensional data, the argument is still valid as we can always construct a statistic based on a single component of the d-dimensional data point.

Lemma 5.1. *For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, suppose that K i.i.d. observations $y_1, \cdots, y_K$ follow distribution density $f(\cdot|\boldsymbol{\theta})$ and cumulative distribution $F(\cdot|\boldsymbol{\theta})$, where f and F satisfy some mild conditions. Then the αth sample quantile statistic ($\alpha \in (0,1)$), denoted as $S = y_{(K\alpha)}$, satisfies Condition 5.1,5.2,5.3.*

Proof. For simplicity we denote $f(\cdot|\boldsymbol{\theta})$ and $F(\cdot|\boldsymbol{\theta})$ as f and F. First, the large sample properties for the sample quantile statistic (L^2 convergence, strong consistency and asymptotic normality) can be found in Del Barrio et al. (2005) and Liu and Yang (2012). Therefore, we only need to prove that quantile statistics satisfy Condition 5.2,5.3.

The αth sample quantile statistic S_α can be obtained via the corresponding ordered statistics $S_\alpha = Y_{(\omega)}$, with $\omega = \lfloor K\alpha \rfloor$. The density function of S is given by (omitting $\boldsymbol{\theta}$ for simplicity) (Rényi, 1953)

$$f^{(K)}(s) = \omega \binom{K}{\omega} F(s)^{\omega-1}(1-F(s))^{K-\omega} f(s). \tag{5.12}$$

Condition 5.4. *The distributions functions $f(x|\boldsymbol{\theta})$, $F(x|\boldsymbol{\theta})$ are such that, for any*

$\alpha \in (0,1)$,

- *for any* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\sup_x \left|\frac{f''(x)}{F(x)(1-F(x))}\right| = O(1)$; $\sup_x \left|\frac{f'(x)}{F(x)(1-F(x))}\right| = O(1)$; $\sup_x \left|\frac{f''(x)}{f(x)} - \left(\frac{f'(x)}{f(x)}\right)^2\right| = O(1)$;
- $\sup_{\boldsymbol{\theta}} \left|\frac{\frac{\partial^2 F}{\partial \boldsymbol{\theta}^2}}{F(1-F)}\right| = O(1)$; $\sup_{\boldsymbol{\theta}} \left|\frac{\frac{\partial F}{\partial \boldsymbol{\theta}}}{F(1-F)}\right| = O(1)$; $\sup_{\boldsymbol{\theta}} \left|\frac{\frac{\partial^2 f}{\partial \boldsymbol{\theta}^2}}{f} - \left(\frac{\frac{\partial f}{\partial \boldsymbol{\theta}}}{f}\right)^2\right| = O(1)$, *where the supreme is taken for any* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

It is easy to get

$$\begin{aligned}\frac{\mathtt{d} f^{(K)}(s)}{\mathtt{d}s} f^{(K)}(s)^{-1} &= (\omega-1)\frac{f(s)}{F(s)} - (K-\omega)\frac{f(s)}{1-F(s)} + \frac{f'(s)}{f(s)} \\ &= K \cdot \underbrace{\left((\alpha - K^{-1})\frac{f(s)}{F(s)} - (1-\alpha)\frac{f(s)}{1-F(s)} + K^{-1}\frac{f'(s)}{f(s)}\right)}_{:=g^{(K)}(s)}\end{aligned}$$

where $\sup_s \left|\frac{\mathtt{d}}{\mathtt{d}s} g^{(K)}(s)\right| = O(1)$ under the following mild Condition 5.4.

Similarly, if we consider f and F as functions of $\boldsymbol{\theta}$, we also have

$$\frac{\partial f^{(K)}(s|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f^{(K)}(s|\boldsymbol{\theta})^{-1} = K \cdot \underbrace{\left((\alpha - K^{-1})\frac{\frac{\partial F}{\partial \boldsymbol{\theta}}}{F} - (1-\alpha)\frac{\frac{\partial F}{\partial \boldsymbol{\theta}}}{1-F} + K^{-1}\frac{\frac{\partial f}{\partial \boldsymbol{\theta}}}{f}\right)}_{:=\zeta(s;\boldsymbol{\theta})}$$

where $\zeta(s;\boldsymbol{\theta})$ is such that for fixed value s, $\sup_{\boldsymbol{\theta}} |\zeta(s;\boldsymbol{\theta})| = O(1)$, $\sup_{\boldsymbol{\theta}} |\frac{\mathtt{d}}{\mathtt{d}s}\zeta(s;\boldsymbol{\theta})| = O(1)$ and $\sup_{\boldsymbol{\theta}} |\frac{\partial}{\partial \boldsymbol{\theta}}\zeta(s;\boldsymbol{\theta})| = O(1)$ for any fixed value s under Condition 5.4. □

First, we need to introduce the following lemmas, which will be used to show the large sample properties of the estimate later.

Lemma 5.2. *Suppose that statistics* $S_i, i = 1, \cdots, M$ *satisfy Condition 5.1,5.2,5.3 and the log-likelihood,* $n^{-1}\sum_{i=1}^{M} \log f_{i,n}(s_i|\boldsymbol{\theta})$, *is concave in a neighbourhood of* $\boldsymbol{\theta}^*$ *for large* n, *then the maximum point* $\tilde{\boldsymbol{\theta}}$ *of* $\tilde{l}(\boldsymbol{\theta}|\boldsymbol{s})$ *given in (5.2) converges to* $\boldsymbol{\theta}^*$ *in probability.*

Proof. In this proof, we write $K_n = n/M$, instead of K, to emphasize that K depends on the sample size n. It is easy to show that the following function of $\boldsymbol{\theta}$,

$$n^{-1}\sum_{i=1}^{M}\int \log f_{i,n}(s|\boldsymbol{\theta})\cdot f_{i,n}(s|\boldsymbol{\theta}^*)\mathtt{d}s \tag{5.13}$$

has derivative $\mathbf{0}$ at $\boldsymbol{\theta}^*$. Therefore (5.13) reaches the maximum at $\boldsymbol{\theta}^*$, under the concavity assumption for $n^{-1}\sum_{i=1}^{M}\log f_{i,n}(s_i|\boldsymbol{\theta})$ in a neighborhood of $\boldsymbol{\theta}^*$.

On the other hand, for any $\boldsymbol{\theta}$ and any s_i under Condition 5.1,5.2, we can apply Taylor's approximation and then,

$$\begin{aligned}
& K_n^{-1}\left(\log f_{i,n}(s_i^*|\boldsymbol{\theta}) - \int \log f_{i,n}(s|\boldsymbol{\theta})\cdot f_{i,n}(s|\boldsymbol{\theta}^*)\mathtt{d}s\right)\\
= \;& K_n^{-1}\int [\log f_{i,n}(s_i^*|\boldsymbol{\theta}) - \log f_{i,n}(s|\boldsymbol{\theta})]\cdot f_{i,n}(s|\boldsymbol{\theta}^*)\mathtt{d}s\\
= \;& K_n^{-1}\int \left[\left.\frac{\mathtt{d}f_{i,n}(s|\boldsymbol{\theta})}{\mathtt{d}s}\right|_{s=\tilde{s}_i}\cdot f_{i,n}(\tilde{s}_i|\boldsymbol{\theta})^{-1}\cdot(s_i^*-s)\right]\cdot f_{i,n}(s|\boldsymbol{\theta}^*)\mathtt{d}s\\
= \;& K_n^{-1}\cdot\int K_n g_i^{(n)}(\tilde{s}_i|\boldsymbol{\theta})\cdot(s_i^*-s)\cdot f_{i,n}(s|\boldsymbol{\theta}^*)\mathtt{d}s\\
= \;& O(1)\cdot\int\left|\frac{\mathtt{d}}{\mathtt{d}s}g_i^{(n)}(s|\boldsymbol{\theta})|_{s=\breve{s}_i}\right|\cdot(s_i^*-s)^2\cdot f_{i,n}(s|\boldsymbol{\theta}^*)\mathtt{d}s\\
= \;& O(1)\cdot\mathbb{E}|s_i^*-s_i|^2 = o(1)
\end{aligned}$$

where $\tilde{s}_i$ is a value between s_i^* and s, and $\breve{s}_i$ is a value between $\tilde{s}_i$ and s.

Under Condition 5.1,5.2 and 5.3, we also have $K_n^{-1}\log f_{i,n}(s_i|\boldsymbol{\theta}) - K_n^{-1}\log f_{i,n}(s_i^*|\boldsymbol{\theta}) = o_{\mathtt{p}}(1)$. Therefore, with $K_n = n/M$ we have

$$\begin{aligned}
& n^{-1}\sum_{i=1}^{M}\log f_{i,n}(s_i|\boldsymbol{\theta}) - n^{-1}\sum_{i=1}^{M}\int \log f_{i,n}(s|\boldsymbol{\theta})\cdot f_{i,n}(s|\boldsymbol{\theta}^*)\mathtt{d}s\\
= \;& n^{-1}\sum_{i=1}^{M}\log f_{i,n}(s_i|\boldsymbol{\theta}) - n^{-1}\sum_{i=1}^{M}\log f_{i,n}(s_i^*|\boldsymbol{\theta})\\
& + n^{-1}\sum_{i=1}^{M}\log f_{i,n}(s_i^*|\boldsymbol{\theta}) - n^{-1}\sum_{i=1}^{M}\int \log f_{i,n}(s|\boldsymbol{\theta})\cdot f_{i,n}(s|\boldsymbol{\theta}^*)\mathtt{d}s\\
= \;& o_{\mathtt{p}}(1).
\end{aligned}$$

This implies that $n^{-1}\sum_{i=1}^{M}\log f_{i,n}(s_i|\boldsymbol{\theta})$ converges to (5.13), which has $\boldsymbol{\theta}^*$ as the maximum point. Therefore, under the concavity assumption, the maximum point $\tilde{\boldsymbol{\theta}}$ of $n^{-1}\sum_{i=1}^{M}\log f_{i,n}(s_i|\boldsymbol{\theta})$ (or of $\log\tilde{l}(\boldsymbol{\theta}|\boldsymbol{s})$) also converges to $\boldsymbol{\theta}^*$ in probability. □

The following lemma is an extension of Lemma 5.1, for correlated observations.

Lemma 5.3. *Suppose that K correlated observations $y_1,\cdots,y_K$ have the same marginal density f and define joint densities for $\omega-1$ of them with labels $\boldsymbol{\iota}=\{i_1,\cdots,i_{\omega-1}\}$ as*

$$\begin{aligned} G_{\boldsymbol{\iota}}(s) &= \mathbb{P}(y_{i_1}<s,\cdots,y_{i_{\omega-1}}<s|y_k=s),\\ \bar{G}_{\boldsymbol{\iota}}(s) &= \mathbb{P}(y_{j_1}>s,\cdots,y_{j_{K-\omega}}>s|y_k=s). \end{aligned}$$

If the distribution functions f, $G_{\boldsymbol{\iota}}$ and $\bar{G}_{\boldsymbol{\iota}}$ satisfy Condition 5.5, then the αth sample quantile statistic ($\alpha\in(0,1)$), denoted as $S_\alpha=y_{(\omega)}$ with $\omega=K\alpha$, satisfies Condition 5.1,5.2,5.3.

Proof. The functions f, $G_{\boldsymbol{\iota}}$ and $\bar{G}_{\boldsymbol{\iota}}$ depend on parameter $\boldsymbol{\theta}$, which was dropped off for notation simplicity.

First, for a sequence of random variables under the strong mixing condition (5.14), the large sample properties for sample quantile statistics (L^2 convergence, strong consistency and asymptotic normality) exist for their quantile statistics (Doss et al., 2014; Wang et al., 2016). Therefore we only need to demonstrate that Condition 5.2,5.3 are satisfied.

The αth sample quantile statistic S can be obtained via the corresponding ordered statistics $S=Y_{(\omega)}$, with $\omega=\lfloor K\alpha\rfloor$. The density function of S is given by (Rényi, 1953)

$$f^{(K)}(s) = \sum_{k=1}^{K} f(s)\sum_{\iota} G_{\boldsymbol{\iota}}(s)\bar{G}_{\boldsymbol{\iota}}(s),$$

where the second $\sum$ notation means sum over all possible choices of the labels $\boldsymbol{\iota} = \{i_1, \cdots, i_{\omega-1}\}$. Therefore, we can write

$$\frac{\mathtt{d}f^{(K)}(s)}{\mathtt{d}s} f^{(K)}(s)^{-1} = Kg^{(K)}(s),$$

where

$$g^{(K)}(s) = K^{-1}\frac{\mathtt{d}f(s)/\mathtt{d}s}{f(s)} + \frac{K^{-1}\sum_{k=1}^{K}\sum_{\boldsymbol{\iota}} \frac{\mathtt{d}}{\mathtt{d}s}G_{\boldsymbol{\iota}}(s)\cdot \bar{G}_{\boldsymbol{\iota}}(s)}{\sum_{k=1}^{K}\sum_{\boldsymbol{\iota}} G_{\boldsymbol{\iota}}(s)\bar{G}_{\boldsymbol{\iota}}(s)} + \frac{K^{-1}\sum_{k=1}^{K}\sum_{\boldsymbol{\iota}} G_{\boldsymbol{\iota}}(s)\cdot \frac{\mathtt{d}}{\mathtt{d}s}\bar{G}_{\boldsymbol{\iota}}(s)}{\sum_{k=1}^{K}\sum_{\boldsymbol{\iota}} G_{\boldsymbol{\iota}}(s)\bar{G}_{\boldsymbol{\iota}}(s)}.$$

Condition 5.5.

1. *The random observations $y_1, y_2, \cdots$ in Lemma 5.3 satisfy the strong mixing condition*

$$\sup_{k\geq 1}\sup_{E_1\in\mathcal{F}_1^k, E_2\in\mathcal{F}_{k+m}^{\infty}} |\mathbb{P}(E_1\cap E_2) - \mathbb{P}(E_1)\mathbb{P}(E_2)| = O(m^{-(2+\delta)}), \quad \delta > 1/2, \tag{5.14}$$

   where event $E_1 \in \mathcal{F}_1^k$ and $E_2 \in \mathcal{F}_{k+m}^{\infty}$, for $i < j$, $\mathcal{F}_i^j = \sigma\{y_k; k = i, \cdots, j\}$.

2. *The distributions functions f and F are such that, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,*

$$\sup_s K^{-1}\left|\frac{\frac{\mathtt{d}^2}{\mathtt{d}s^2}G_{\boldsymbol{\iota}}(s)}{G_{\boldsymbol{\iota}}(s)}\right| = O(1),\ \sup_s K^{-1}\left|\frac{\frac{\mathtt{d}^2}{\mathtt{d}s^2}\bar{G}_{\boldsymbol{\iota}}(s)}{\bar{G}_{\boldsymbol{\iota}}(s)}\right| = O(1);$$
$$\sup_s K^{-1}\left|\frac{\frac{\mathtt{d}}{\mathtt{d}s}G_{\boldsymbol{\iota}}(s)}{G_{\boldsymbol{\iota}}(s)}\right| = O(1),\ \sup_s K^{-1}\left|\frac{\frac{\mathtt{d}}{\mathtt{d}s}\bar{G}_{\boldsymbol{\iota}}(s)}{\bar{G}_{\boldsymbol{\iota}}(s)}\right| = O(1);$$
$$\sup_s \left|\frac{f''(s)}{f(s)} - \left(\frac{f'(s)}{f(s)}\right)^2\right| = O(1).$$

3.

$$\sup_{\boldsymbol{\theta}} \left| \frac{\frac{\partial^2 G_{\boldsymbol{\iota}}}{\partial \boldsymbol{\theta}^2}}{G_{\boldsymbol{\iota}} \bar{G}_{\boldsymbol{\iota}}} \right| = O(1), \ \sup_{\boldsymbol{\theta}} \left| \frac{\frac{\partial G_{\boldsymbol{\iota}}}{\partial \boldsymbol{\theta}}}{G_{\boldsymbol{\iota}} \bar{G}_{\boldsymbol{\iota}}} \right| = O(1);$$
$$\sup_{\boldsymbol{\theta}} \left| \frac{\frac{\partial^2 \bar{G}_{\boldsymbol{\iota}}}{\partial \boldsymbol{\theta}^2}}{G_{\boldsymbol{\iota}} \bar{G}_{\boldsymbol{\iota}}} \right| = O(1), \ \sup_{\boldsymbol{\theta}} \left| \frac{\frac{\partial \bar{G}_{\boldsymbol{\iota}}}{\partial \boldsymbol{\theta}}}{G_{\boldsymbol{\iota}} \bar{G}_{\boldsymbol{\iota}}} \right| = O(1);$$
$$\sup_{\boldsymbol{\theta}} \left| \frac{\frac{\partial^2 f}{\partial \boldsymbol{\theta}^2}}{f} - \left(\frac{\frac{\partial f}{\partial \boldsymbol{\theta}}}{f} \right)^2 \right| = O(1),$$

where the supreme is taken for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Then under Condition 5.5 we have

$$\sup_{s} \left| \frac{\mathtt{d}}{\mathtt{d}s} g^{(K)}(s) \right| = O(1).$$

Therefore, Condition 5.2 is satisfied. Similarly, we can prove Condition 5.3 also holds.

□

Lemma 5.4. *Suppose* $f_{i,n}$ *is* r*th differentiable for some integer* $r \geq 2$. *Then for* $q(N_n) = O(N_n^{\frac{1}{r}})$, *the mean integrated square error of the estimate* $\hat{f}_{i,n}(s|\boldsymbol{\theta}), \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$, *satisfies*

$$\mathbb{E} \int \sup_{\boldsymbol{\theta}} |\hat{f}_{i,n}(s|\boldsymbol{\theta}) - f_{i,n}(s|\boldsymbol{\theta})|^2 \mathrm{d}s = O(N_n^{\frac{5}{6r}-1}) \tag{5.15}$$

and further for any value s, *we have*

$$\sup_{\boldsymbol{\theta}} |\hat{f}_{i,n}(s|\boldsymbol{\theta}) - f_{i,n}(s|\boldsymbol{\theta})| = O_{\mathtt{p}}(N_n^{\frac{5}{12r}-\frac{1}{2}}) \tag{5.16}$$

Proof. From the expressions of $\hat{f}_{i,n}$ and $f_{i,n}$ we have

$$\begin{aligned}
&\mathbb{E}\int \sup_{\boldsymbol{\theta}} |\hat{f}_{i,n}(s|\boldsymbol{\theta}) - f_{i,n}(s|\boldsymbol{\theta})|^2 \mathrm{d}s \\
\leq &\mathbb{E}\int \sup_{\boldsymbol{\theta}} \left\{ \left[\sum_{j=0}^{q(N_n)} (\hat{C}_j^{(i)} - C_j^{(i)}) \mathrm{H}_j(s) \right]^2 + 2 \sum_{j=0}^{q(N_n)} \left| (\hat{C}_j^{(i)} - C_j^{(i)}) \mathrm{H}_j(s) \right| \sum_{l=q(N_n)+1}^{\infty} |C_l^{(i)} \mathrm{H}_l(s)| \right\} \mathrm{d}s \\
&+ \int \sup_{\boldsymbol{\theta}} \left[\sum_{j=q(N_n)+1}^{\infty} C_j^{(i)} \mathrm{H}_j(s) \right]^2 \mathrm{d}s.
\end{aligned}$$

Note that $C_j^{(i)}$ and $\hat{C}_j^{(i)}$ depend on $\boldsymbol{\theta}$, but $H_j(s)$ does not depend on $\boldsymbol{\theta}$.

Using the Hermite function properties $\int \mathrm{H}_j(s)\mathrm{H}_k(s)\mathrm{d}s = 1$ if $j = k$; and $\int \mathrm{H}_j(s)\mathrm{H}_k(s)\mathrm{d}s = 0$ if $j \neq k$, we have

$$\mathbb{E}\int \sup_{\boldsymbol{\theta}} \left[\sum_{j=0}^{q(N_n)} (\hat{C}_j^{(i)} - C_j^{(i)}) \mathrm{H}_j(s) \right]^2 \mathrm{d}s = \sum_{j=0}^{q(N_n)} \mathbb{E} \sup_{\boldsymbol{\theta}} \left[(\hat{C}_j^{(i)} - C_j^{(i)})^2 \right], \tag{5.17}$$

$$\mathbb{E}\int \sup_{\boldsymbol{\theta}} \left\{ \sum_{j=0}^{q(N_n)} \left| (\hat{C}_j^{(i)} - C_j^{(i)}) \mathrm{H}_j(s) \right| \sum_{l=q(N_n)+1}^{\infty} |C_l^{(i)} \mathrm{H}_l(s)| \right\} \mathrm{d}s = 0 \tag{5.18}$$

and

$$\int \sup_{\boldsymbol{\theta}} \left[\sum_{j=q(N_n)+1}^{\infty} C_j^{(i)} \mathrm{H}_j(s) \right]^2 \mathrm{d}s = \sum_{j=q(N_n)+1}^{\infty} \sup_{\boldsymbol{\theta}} \left[C_j^{(i)} \right]^2. \tag{5.19}$$

Since $\hat{C}_j^{(i)}$ is an unbiased estimation for $C_j^{(i)}$ (sample mean based on N_n Monte Carlo samples, see equation (5.7)), we have $\sup_{\boldsymbol{\theta}} |\hat{C}_j^{(i)} - C_j^{(i)}|$ converges to 0 with $N_n \to \infty$, at rate $o(N_n^{\frac{5}{12r} - \frac{1}{2}})$, for an finite integer $r \geq 1$. Using the Hermite function result in Walter (1977),

$$|\mathrm{H}_j(x)| \leq A_\infty \cdot (j+1)^{-\frac{1}{12}}, \quad \forall x \in (-\infty, \infty), \quad j = 0, 1, \cdots, \tag{5.20}$$

where A_∞ is a constant, and using (5.6) and (5.7), we know that $|C_j^{(i)}| \leq A_\infty \cdot (j+1)^{-\frac{1}{12}}$, $\mathrm{Var}\Big(\mathrm{H}_j(v_i^{(k)})\Big) \leq A_\infty^2 (j+1)^{-\frac{1}{6}}$ and

$$\begin{aligned}\mathbb{E}\sup_{\boldsymbol{\theta}}\left(\hat{C}_j^{(i)} - C_j^{(i)}\right)^2 &= \mathbb{E}\sup_{\boldsymbol{\theta}}\left\{\left(\frac{1}{N_n}\sum_{k=1}^{N_n}\mathrm{H}_j(v_i^{(k)}) - C_j^{(i)}\right)^2\right\} \\ &\leq \frac{1}{N_n}A_\infty^2(j+1)^{-\frac{1}{6}}, \end{aligned} \tag{5.21}$$

Then, as the convex function $(x+1)^{-\frac{1}{6}}$, we have

$$\begin{aligned}&\mathbb{E}\int \sup_{\boldsymbol{\theta}} |\hat{f}_{i,n}(s|\boldsymbol{\theta}) - f_{i,n}(s|\boldsymbol{\theta})|^2 \mathrm{d}s \leq \sum_{j=0}^{q(N_n)} \mathbb{E}\sup_{\boldsymbol{\theta}}(\hat{C}_j^{(i)} - C_j^{(i)})^2 + \sum_{j=q(N_n)+1}^{\infty}\sup_{\boldsymbol{\theta}}\left[C_j^{(i)}\right]^2 \\ \leq &\frac{1}{N_n}A_\infty^2 \cdot \sum_{j=0}^{q(N_n)}(j+1)^{-\frac{1}{6}} + \sum_{j=q(N_n)+1}^{\infty}\sup_{\boldsymbol{\theta}}\left[C_j^{(i)}\right]^2 \\ \leq &A_1\frac{[q(N_n)+1]^{\frac{5}{6}}}{N_n} + \sum_{j=q(N_n)+1}^{\infty}\sup_{\boldsymbol{\theta}}\{{C_j^{(i)}}^2\},\end{aligned}$$

where A_1 is a constant.

On the other hand, Bonan and Clark (1990) provided the following properties of standard Hermite Polynomials,

$$\mathrm{h}_j'(x) = \sqrt{2j}\mathrm{h}_{j-1}(x). \tag{5.22}$$

Therefore, for integer $r \geq 1$,

$$\begin{aligned} C_j^{(i)} &= \int f_{i,n}(v)\mathrm{h}_j(v)e^{-\frac{v^2}{2}}\mathrm{d}v = \int (2j)^{-\frac{1}{2}}\mathrm{h}_{j+1}(v)\mathrm{d}(f_{i,n}(v)\cdot e^{-\frac{v^2}{2}}) \\ &= \int (2j)^{-\frac{1}{2}}(2j+2)^{-\frac{1}{2}}...(2j+2r-2)^{-\frac{1}{2}}\mathrm{h}_{j+r}(v)\mathrm{d}^r(f_{i,n}(v)\cdot e^{-\frac{v^2}{2}}) \\ &\leq (2j)^{-\frac{r}{2}}\int \mathrm{h}_{j+r}(v)\mathrm{d}^r(f_{i,n}(v)\cdot e^{-\frac{v^2}{2}}). \end{aligned} \tag{5.23}$$

Let $b_j = \int \mathrm{h}_{j+r}(v)\mathrm{d}^r(f_{i,n}(v)\cdot e^{-\frac{v^2}{2}})$, which could be regarded as the coefficients of

function $\frac{\mathrm{d}^r(f_{i,n}(v)\cdot e^{-\frac{v^2}{2}})}{\mathrm{d}v}$. The series sum $\sum b_j^2$ should be bounded above by some constant (Walter, 1977). So,

$$\begin{aligned}\sum_{j=q+1}^{\infty}\sup_{\boldsymbol{\theta}} {C_j^{(i)}}^2 &\leq \sum_{j=q+1}^{\infty}(2j)^{-r}\left[\int h_{j+r}(v)\mathrm{d}^r(f_{i,n}\cdot e^{-\frac{v^2}{2}})\right]^2 \\ &\leq \frac{1}{(2+2q(N_n))^r}\cdot\sum_{j=q+1}^{\infty}b_j^2 = O(N_n^{-1}), \end{aligned} \tag{5.24}$$

which is much less than $A_1\frac{[q(N_n)+1]^{\frac{5}{6}}}{N_n} = O(N_n^{\frac{5}{6r}-1})$. The lemma is proved. □

Now we can introduce the main results.

Theorem 5.1. *Suppose that for an integer $r \geq 2$, $N_n^{\frac{5}{12r}-\frac{1}{2}} = o(n^{-\frac{1}{2}})$, i.e. $N_n^{-1} = o(n^{-\frac{6r}{6r-5}})$ and $q(N_n) = O(N_n^{\frac{1}{r}})$, and suppose the statistics $S_i, i = 1,\cdots,M$, their densities $f_{i,n}$ and the original data density f satisfy Condition 5.1,5.2,5.3. If $S_i, i = 1,\cdots,M$ are independent and if $n^{-1}\log\hat{l}(\boldsymbol{\theta}|\boldsymbol{s})$ and $n^{-1}\log\tilde{l}(\boldsymbol{\theta}|\boldsymbol{s})$ are concave in $\boldsymbol{\Theta}$, then the maximum point $\hat{\boldsymbol{\theta}}$ of $\hat{l}(\boldsymbol{\theta}|\boldsymbol{s})$ given in (5.9) converges to $\boldsymbol{\theta}^*$ in probability.*

Proof. Lemma 5.4 immediately tells us that $\left|\hat{f}_i^{(n)}(s_i|\boldsymbol{\theta}) - f_{i,n}(s_i|\boldsymbol{\theta})\right| = o_{\mathtt{p}}\left(n^{-\frac{1}{2}}\right)$. Therefore, under the concavity assumption, the maximum point $\hat{\boldsymbol{\theta}}$ of $\hat{l}(\boldsymbol{\theta}|\boldsymbol{s})$ and the maximum point $\tilde{\boldsymbol{\theta}}$ of $\tilde{l}(\boldsymbol{\theta}|\boldsymbol{s})$ are such that $\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\| \to 0$ in probability, as $n \to \infty$. Further Lemma 5.2 immediately tell us that $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$ in probability, as $n \to \infty$. □

If we denote

$$\begin{aligned}\tilde{f}_{i,n}(s|\boldsymbol{\theta}) &:= \sum_{m=0}^{q(N_n)} C_m^{(i)}\mathrm{H}_m(s), \\ \tilde{\boldsymbol{\phi}}_i(s,\boldsymbol{\theta}) &:= \frac{\frac{\partial}{\partial\boldsymbol{\theta}}\tilde{f}_{i,n}(s|\boldsymbol{\theta})}{\tilde{f}_{i,n}(s|\boldsymbol{\theta})}, \quad \tilde{\boldsymbol{\phi}}_i^{(1)}(s,\boldsymbol{\theta}) := \frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\phi}_i(s,\boldsymbol{\theta}), \quad \tilde{\boldsymbol{\psi}}_i(s,\boldsymbol{\theta}) := \frac{\mathtt{d}}{\mathtt{d}s}\boldsymbol{\phi}_i(s,\boldsymbol{\theta}), \\ \boldsymbol{\phi}_i^{(1)}(s,\boldsymbol{\theta}) &:= \frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\phi}_i(s,\boldsymbol{\theta}), \quad \boldsymbol{\psi}_i(s,\boldsymbol{\theta}) := \frac{\mathtt{d}}{\mathtt{d}s}\boldsymbol{\phi}_i(s,\boldsymbol{\theta}), \end{aligned} \tag{5.25}$$

where $\boldsymbol{\phi}_i(s,\boldsymbol{\theta})$ is given in (5.11), then we have the following asymptotic normality result for $\hat{\boldsymbol{\theta}}$.

Theorem 5.2. *Under the same condition of Theorem 5.1, the maximum point* $\hat{\boldsymbol{\theta}}$ *of* $\hat{l}(\boldsymbol{\theta}|\boldsymbol{s})$ *given in (5.9) satisfies*

$$\begin{aligned}
&\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \to \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \\
&\boldsymbol{\Sigma} = \\
&\lim_{n\to\infty} M \left(\sum_{i=1}^{M} \boldsymbol{\phi}_i^{(1)}(s_i^*, \boldsymbol{\theta}^*)\right)^{-1} \left(\sum_{i=1}^{M} \sigma_i^2 \cdot \boldsymbol{\psi}_i(s_i^*, \boldsymbol{\theta}^*)^{\otimes 2}\right) \left(\sum_{i=1}^{M} \boldsymbol{\phi}_i^{(1)}(s_i^*, \boldsymbol{\theta}^*)\right)^{-1} \quad (5.26)
\end{aligned}$$

where $\otimes$ *is the original Kronecker product (Schafer, 1966),* σ_i^2 *is the asymptotic variance of* $\sqrt{K_n}(S_i - s_i^*)$ *and each* S_i *is constructed based on* $K_n = n/M$ *different samples.*

Proof. Using first-order Taylor expansion on $\frac{\partial \log \hat{l}(\boldsymbol{\theta}|\boldsymbol{s})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ at point $\boldsymbol{\theta}^*$ we have

$$\begin{aligned}
\frac{\partial \log \hat{l}(\boldsymbol{\theta}|\boldsymbol{s})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= M^{-1} \sum_{i=1}^{M} \left[\frac{\frac{\partial}{\partial \boldsymbol{\theta}} \hat{f}_{i,n}(s_i|\boldsymbol{\theta})}{\hat{f}_{i,n}(s_i|\boldsymbol{\theta})}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} + \frac{\partial}{\partial \boldsymbol{\theta}} \frac{\frac{\partial}{\partial \boldsymbol{\theta}} \hat{f}_{i,n}(s_i|\boldsymbol{\theta})}{\hat{f}_{i,n}(s_i|\boldsymbol{\theta})}\Big|_{\boldsymbol{\theta}=\breve{\boldsymbol{\theta}}} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) \right] \\
&= M^{-1} \sum_{i=1}^{M} \left[\frac{\frac{\partial}{\partial \boldsymbol{\theta}} \tilde{f}_{i,n}(s_i|\boldsymbol{\theta})}{\tilde{f}_{i,n}(s_i|\boldsymbol{\theta})}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} + \frac{\partial}{\partial \boldsymbol{\theta}} \frac{\frac{\partial}{\partial \boldsymbol{\theta}} \tilde{f}_{i,n}(s_i|\boldsymbol{\theta})}{\tilde{f}_{i,n}(s_i|\boldsymbol{\theta})}\Big|_{\boldsymbol{\theta}=\breve{\boldsymbol{\theta}}} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) \right] \\
&\quad + O_{\mathtt{p}}\left(N_n^{\frac{1}{2}}\right) \\
&= M^{-1} \sum_{i=1}^{M} \left[\tilde{\boldsymbol{\phi}}_i(s_i, \boldsymbol{\theta}^*) + \tilde{\boldsymbol{\phi}}_i^{(1)}(s_i, \breve{\boldsymbol{\theta}}) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) \right] + O_{\mathtt{p}}\left(N_n^{\frac{1}{2}}\right), \quad (5.27)
\end{aligned}$$

where $\breve{\boldsymbol{\theta}}$ is between $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}$.

The second equality is because, given the observations or the statistics $\boldsymbol{s}$, $\sqrt{N_n}(\tilde{f}_{i,n} - \hat{f}_{i,n}) \to \mathcal{N}(0, \eta_i^2)$ where $\eta_i^2 = \sum_{m=0}^{q(N_n)} \mathrm{var}(C_m^{(i)} - \widehat{C}_m^{(i)}) \cdot \mathrm{H}_m(s_i) < \infty$.

On the other hand, we have

$$\tilde{\boldsymbol{\phi}}_i(s_i, \boldsymbol{\theta}) = \frac{1}{\tilde{f}_{i,n}(s_i|\boldsymbol{\theta})} \sum_{m=1}^{q(N_n)} \mathrm{H}_m(s_i) \int \frac{\partial}{\partial \boldsymbol{\theta}} f_{i,n}(v|\boldsymbol{\theta}) \cdot \mathrm{H}_m(v) \mathtt{d}v. \quad (5.28)$$

We shall also further expand $\tilde{\boldsymbol{\phi}}_i(s_i, \boldsymbol{\theta})$ in terms of s_i at s_i^*, i.e.

$$\tilde{\boldsymbol{\phi}}_i(s_i, \boldsymbol{\theta}) = \tilde{\boldsymbol{\phi}}_i(s_i^*, \boldsymbol{\theta}) + \tilde{\boldsymbol{\psi}}_i(s_i^*, \boldsymbol{\theta}) \cdot (s_i - s_i^*) + O_{\mathtt{p}}((s_i - s_i^*)^2 \cdot \mathbf{1}), \tag{5.29}$$

where $\mathbf{1}$ is a vector of 1s. Using the property that as $n \to \infty$, $\sqrt{n/M}(s_i - s_i^*) \sim \mathcal{N}(0, \sigma_i^2)$ under Condition 5.1,5.2 and 5.3, we will have asymptotically

$$\sqrt{M/n}\left(\tilde{\boldsymbol{\phi}}_i(s_i, \boldsymbol{\theta}) - \tilde{\boldsymbol{\phi}}_i(s_i^*, \boldsymbol{\theta})\right) \sim \mathcal{N}\left(\mathbf{0}, \sigma_i^2 \cdot \frac{M}{n}\boldsymbol{\psi}_i(s_i^*, \boldsymbol{\theta}) \otimes \frac{M}{n}\boldsymbol{\psi}_i(s_i^*, \boldsymbol{\theta})\right) \tag{5.30}$$

where $\frac{M}{n}\boldsymbol{\psi}_i(s_i^*, \boldsymbol{\theta})$ is a vector with finite elements under Condition 5.3.

In the equation (5.27), the left-hand-side is $\mathbf{0}$ at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Therefore, asymptotically

$$\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) = -\left(\sum_{i=1}^{M} \tilde{\boldsymbol{\phi}}_i^{(1)}(s_i, \boldsymbol{\theta}^*)\right)^{-1} \sum_{i=1}^{M} \tilde{\boldsymbol{\phi}}_i(s_i, \boldsymbol{\theta}^*), \qquad \text{equal in distribution.} \tag{5.31}$$

Due to $\mathbb{E}\boldsymbol{\phi}_i(s_i, \boldsymbol{\theta}^*) = \mathbf{0}$, and from (5.29), (5.30) and (5.31), we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\Sigma} = \lim_{n\to\infty} M \left(\sum_{i=1}^{M} \boldsymbol{\phi}_i^{(1)}(s_i^*, \boldsymbol{\theta}^*)\right)^{-1} \left(\sum_{i=1}^{M} \sigma_i^2 \cdot \boldsymbol{\psi}_i(s_i^*, \boldsymbol{\theta}^*)^{\otimes 2}\right) \left(\sum_{i=1}^{M} \boldsymbol{\phi}_i^{(1)}(s_i^*, \boldsymbol{\theta}^*)\right)^{-1}. \tag{5.32}$$

The theorem is proved. □

An estimate of $\boldsymbol{\Sigma}$ can be obtained via bootstrap methods (Bühlmann, 2002; Zhu et al., 2016). Suppose that $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ are independent observations with density f, cumulative distribution F and statistics $S_1, \cdots, S_M$. Denote the relation between the estimator $\hat{\boldsymbol{\theta}}$ which maximize the working likelihood and the statistics $\boldsymbol{S}$ as function $\boldsymbol{\Theta}$, i.e. $\hat{\boldsymbol{\theta}} = \boldsymbol{\Theta}(\boldsymbol{s})$. Suppose $\tilde{\boldsymbol{y}}_1, \cdots, \tilde{\boldsymbol{y}}_n$ are i.i.d. bootstrap samples from the empirical distribution $\hat{F}$, with corresponding bootstrap statistics $S_1^{(b)}, \cdots, S_M^{(b)}$ and bootstrap

estimate $\hat{\boldsymbol{\theta}}^{(b)} = \boldsymbol{\Theta}(\boldsymbol{S}^{(b)})$.

For correlated observations $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ collected from a continuous time stochastic process, we consider m out n without replacement bootstrap, i.e. a bootstrap sample $\tilde{\boldsymbol{y}}_1, \cdots, \tilde{\boldsymbol{y}}_m$. In other words, the bootstrap sample is a subset of the original time series. We also denote the original estimator as $\hat{\boldsymbol{\theta}} = \boldsymbol{\Theta}(\boldsymbol{s})$ and the bootstrap estimator as $\hat{\boldsymbol{\theta}}^{(b)} = \boldsymbol{\Theta}(\boldsymbol{S}^{(b)})$.

Denote $\hat{\boldsymbol{\Sigma}}$ as the sample covariance matrix of the B bootstrap estimates $\hat{\boldsymbol{\theta}}^{(b)}, b = 1, \cdots, B$ and $\bar{\boldsymbol{\theta}}^{(b)} = B^{-1} \sum_{i=1}^{n} \hat{\boldsymbol{\theta}}^{(b)}$, i.e.

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{B-1} \sum_{i=1}^{B} (\hat{\boldsymbol{\theta}}^{(b)} - \bar{\boldsymbol{\theta}}^{(b)})^{\otimes 2}. \tag{5.33}$$

Lemma 5.5. *If $\kappa/K \to 1$ with $K \to \infty$ then for any x, $\sup_v |\mathbb{P}(F_n(x) < v) - \mathbb{P}(F_{n;b}(x) < v)| \to 0$, as $K \to \infty$.*

Proof. The sequence $I[y_j \leq x]$ is a strong α-mixing sequence because of the Markov property of $y_1, \cdots, y_K$. Therefore both $\sqrt{K} F_n(x)$ and $\sqrt{\kappa} F_{n;b}(x)$ will be asymptotically normal with the same mean. Their asymptotic variance will also be the same, if $\kappa/K \to 1$, as $K \to \infty$. For example, we can choose $\kappa = K - \sqrt{K}$. □

We have the following theorem

Theorem 5.3. *For either independent observations $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$, or observations from a continuous time stochastic process, the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ based on the bootstrap estimators is such that $\hat{\boldsymbol{\Sigma}}$ converges to $\boldsymbol{\Sigma}$ in probability as $n \to \infty$.*

Proof. For notation simplicity, we illustrate this using one-dimensional data with K observations, $y_1, \cdots, y_K$. The argument applies for multi-dimensional case, since we can always use one component of the multi-dimensional data to construct S_i.

First for independent observations $y_1, \cdots, y_n$, we always have that quantile statistics S and their bootstrap copies $S^{(b)}$ guarantee $\sup_s |\mathbb{P}(S < s) - \mathbb{P}(S^{(b)} < s)| \to 0$.

For dependent Markov process observations $y_1, \cdots, y_K$ and κ out K (without replacement) bootstrap sample $\tilde{y}_1, \cdots, \tilde{y}_\kappa$, we can also construct an α-quantile statistic as $S_\alpha = F_n^{-1}(\alpha) = \inf\{x; F_n(x) \geq \alpha\}$, where $F_n(x) = K^{-1}\sum_{j=1}^{K} I[y_j \leq x]$, and the bootstrap estimator $S_\alpha^{(b)} = F_{n;b}^{-1}(\alpha)$ with $F_{n;b}(x) = \kappa^{-1}\sum_{j=1}^{\kappa} I[\tilde{y}_j^{(b)} \leq x]$. Based on Lemma 5.5 we also have $\sup_s |\mathbb{P}(S < s) - \mathbb{P}(S^{(b)} < s)| \to 0$.

Now we immediately have that for any p-dimensional column vector $\boldsymbol{v}$,

$$\left|\mathbb{P}\left(\hat{\boldsymbol{\theta}} \leq \boldsymbol{v}\right) - \mathbb{P}\left(\hat{\boldsymbol{\theta}}^{(b)} \leq \boldsymbol{v}\right)\right| \to 0$$

as $K = n/M \to \infty$. This implies that the bootstrap estimates have the same asymptotic variance as the original estimate and therefore the sample bootstrap estimate $\hat{\boldsymbol{\Sigma}} \to \text{var}(\hat{\boldsymbol{\theta}}^{(b)}) = \boldsymbol{\Sigma}$. The theorem is proved. □

5.4 The main Algorithm

To find the maximum point for the working likelihood $\hat{l}$, we consider the idea of gradient descent algorithm (LeCun et al., 2012). Starting from a suitable starting point $\boldsymbol{\theta}^{(0)}$, the gradient descent algorithm searches the parameter space by moving to the next candidate point in the direction of the gradient, which will lead to fast convergence. In practice, the gradient is usually calculated via numerical differentiation. Here, at every step of the iteration, we consider both the points at the gradient direction and some points at other directions to make sure that the algorithm can always move to the best direction towards the maximum points. In summary, we have the following gradient descent algorithm to find the maximum likelihood $\hat{\boldsymbol{\theta}}$ based on equation (5.9).

Algorithm 7: Hermite Polynomials Expansion for Finding $\hat{\boldsymbol{\theta}}$

Initialisation: observation $\mathbf{y}$, a generator model f, a simple density $f_0(\boldsymbol{\theta})$, a threshold value $\epsilon > 0$, a large value N_n, $q(N_n)$;

Calculate statistics $S_i = \mathbb{T}_i(\mathbf{y})$, $i = 1, 2, .., M$;

Set a starting value $\boldsymbol{\theta}^{(0)}$, which could be sampled via $\boldsymbol{\theta}^{(0)} \sim f_0(\boldsymbol{\theta})$;

Sampling N datasets $\mathbf{x}^{(k)}, k = 1, \cdots N$ from $f(\mathbf{x}|\boldsymbol{\theta})$, and get the corresponding M statistics $v_i^{(k)}, i = 1, \cdots, M$ for each simulated dataset $\mathbf{x}^{(k)}$;

Calculate $\hat{l}(\boldsymbol{\theta}^{(0)}|\boldsymbol{s}), \boldsymbol{s} = (S_1, ..., S_M)$ according to (5.7), (5.8) and (5.9);

Find a $\boldsymbol{\theta}$ that maximizes the value of $\hat{l}(\boldsymbol{\theta}|\boldsymbol{s})$;

Output $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ as the MLE;

5.5 Likelihood free inference with divergence p

In practice, the dimension of the parameter space p may go to ∞ as $n \to \infty$. For example, an infectious disease under the LV model framework may have the same disease transmission rate which never changes, but the death rate may change over time due to improvement in medical treatments, which may result in that the parameter value θ_3 in the LV model has piece-wise changes over time but θ_1 and θ_2 are constant. The unknown model parameters become $\theta_1, \theta_2, \theta_3, \cdots, \theta_{p_n}$, where $\theta_i, 3 \leq i \leq p_n$ representing the death rate at different stages. We will show in this section that our methodology still work for a lower dimensional projection of $\hat{\boldsymbol{\theta}}$ even with $p_n \to \infty$, as that in Liang and Du (2012).

Condition 5.6. *For all i, the density function $f_{i,n}(s|\boldsymbol{\theta})$ such that*

$$\boldsymbol{\phi}_i(s, \boldsymbol{\theta}) := \frac{\partial f_{i,n}(s|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f_{i,n}(s|\boldsymbol{\theta})^{-1} = \rho_n \cdot \zeta_{i,n}(s|\boldsymbol{\theta}) \tag{5.34}$$

where $\zeta_{i,n}$ is such that for fixed value s, $\sup_i \sup_{\boldsymbol{\theta}} |\zeta_{i,n}(s|\boldsymbol{\theta})| = O(1)$, $\sup_i \sup_{\boldsymbol{\theta}} |\frac{\mathrm{d}}{\mathrm{d}s}\zeta_{i,n}(s|\boldsymbol{\theta})| =$

$O(1)$ *and* $\sup_i \sup_{\boldsymbol{\theta}} |\frac{\partial}{\partial \boldsymbol{\theta}} \zeta_{i,n}(s|\boldsymbol{\theta})| = O(1)$.

Corollary 5.1. *Suppose the parameter dimension* $p_n = O(n^\gamma)$*, where* $\gamma \in (0,1)$ *and the number of independent statistics* M_n *is such that* $M_n = O(p_n)$*. We also assume that* $f_{i,n}$ *to be* r*th differentiable (*$r \geq 2$*), and that* $n^{-1}\log \hat{l}(\boldsymbol{\theta}|\boldsymbol{s})$ *and* $n^{-1}\log \tilde{l}(\boldsymbol{\theta}|\boldsymbol{s})$ *are concave in a neighbourhood of* $\boldsymbol{\theta}^*$*. If the statistics* $S_1, \cdots, S_{p_n}$ *satisfies Condition 5.1,5.2,5.6 then we have that for any* $d \times p$ *matrix* $\boldsymbol{\Lambda}$ *with each row vector of* $\boldsymbol{\Lambda}$ *as unit vector, where* d *is a finite number, and for the maximum point* $\hat{\boldsymbol{\theta}}$ *of* $\hat{l}(\boldsymbol{\theta}|\boldsymbol{s})$ *given in (5.9),* $\boldsymbol{\Lambda}\hat{\boldsymbol{\theta}}$ *converges to* $\boldsymbol{\Lambda}\boldsymbol{\theta}^*$ *in probability and satisfies*

$$\begin{aligned}
&\sqrt{n}\boldsymbol{\Lambda}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \to \mathcal{N}\left(\mathbf{0}, \lim_{n\to\infty} \cdot \boldsymbol{\Lambda}\boldsymbol{\Sigma}_n\boldsymbol{\Lambda}^T\right), \\
&\boldsymbol{\Sigma}_n = \\
&M_n \left(\sum_{i=1}^{M_n} \boldsymbol{\phi}_i^{(1)}(s_i^*, \boldsymbol{\theta}^*)\right)^{-1} \left(\sum_{i=1}^{M_n} \sigma_i^2 \cdot \boldsymbol{\psi}_i(s_i^*, \boldsymbol{\theta}^*)^{\otimes 2}\right) \left(\sum_{i=1}^{M_n} \boldsymbol{\phi}_i^{(1)}(s_i^*, \boldsymbol{\theta}^*)\right)^{-1} \qquad (5.35)
\end{aligned}$$

Proof. The proof of this Corollary follows from the proofs of Theorem 5.1 and Theorem 5.2. We only need to show that the limiting covariance matrix has finite diagnal elements.

From definition (5.25) and Condition 5.2,5.6, we know that the elements of $\phi_i^{(1)}$ and $\boldsymbol{\psi}_i$ are in the same order. Therefore $\sup_i \left[\boldsymbol{\phi}_i^{(1)}\right]^{-1} \cdot \boldsymbol{\psi}_i$ is a finite matrix. Therefore the limit of the elements in

$$\boldsymbol{\Sigma}_n = M_n \left(\sum_{i=1}^{M_n} \boldsymbol{\phi}_i^{(1)}(s_i^*, \boldsymbol{\theta}^*)\right)^{-1} \left(\sum_{i=1}^{M_n} \sigma_i^2 \cdot \boldsymbol{\psi}_i(s_i^*, \boldsymbol{\theta}^*)^{\otimes 2}\right) \left(\sum_{i=1}^{M_n} \boldsymbol{\phi}_i^{(1)}(s_i^*, \boldsymbol{\theta}^*)\right)^{-1}$$

exists. Therefore $\lim_{n\to\infty} \boldsymbol{\Lambda}\boldsymbol{\Sigma}_n\boldsymbol{\Lambda}^T$ exists. □

5.6 Likelihood free inference with multi-variate Hermite Polynomials

Suppose that the M-dimensional statistic $\boldsymbol{s} = (s_1, \cdots, s_M)$ of the data has density function $f_n(\boldsymbol{s}|\boldsymbol{\theta})$, which can be expanded by

$$f_n(\boldsymbol{s}|\boldsymbol{\theta}) = \sum_{m=0}^{\infty} \sum_{\boldsymbol{m}} C_{\boldsymbol{m}}(\boldsymbol{\theta}) \mathcal{H}_{\boldsymbol{m}}(\boldsymbol{s}), \tag{5.36}$$

where the coefficients satisfy,

$$C_{\boldsymbol{m}} \;\; := \;\; C_{\boldsymbol{m}}(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} f_n(\boldsymbol{v}|\boldsymbol{\theta}) \mathcal{H}_{\boldsymbol{m}}(\boldsymbol{v}) \mathrm{d}\boldsymbol{v}. \tag{5.37}$$

Since each calculation requires the sum of all situations under different values of the M-dimensional vector $\boldsymbol{m}$, the complexity of computation will increase sharply with the rising of dimensions M and be much more infeasible than using independent statistics $s_1, \cdots, s_M$.

Because of the computational in-feasibility of using pdf $f_n(\boldsymbol{s}|\boldsymbol{\theta})$, for correlated $s_1, \cdots, s_M$, we here only provide the theoretical results. If $\boldsymbol{v}^{(i)}$ is sampled from the density function f_n, the estimate for $C_{\boldsymbol{m}}$ is

$$\hat{C}_{\boldsymbol{m}} = \frac{1}{N_n} \sum_{i=1}^{N_n} \mathcal{H}_{\boldsymbol{m}}(\boldsymbol{v}^{(i)}). \tag{5.38}$$

We also denote,

$$\hat{f}_n(\boldsymbol{s}|\boldsymbol{\theta}) = \sum_{m=0}^{q(N_n)} \sum_{\boldsymbol{m}} \hat{C}_{\boldsymbol{m}}(\boldsymbol{\theta}) \mathcal{H}_{\boldsymbol{m}}(\boldsymbol{s}) \tag{5.39}$$

and the working likelihood

$$\hat{l}(\boldsymbol{\theta}|\boldsymbol{s}) = \hat{f}_n(\boldsymbol{s}|\boldsymbol{\theta}). \tag{5.40}$$

Lemma 5.6. *Suppose that* $|\boldsymbol{m}| = m_1 + m_2 + ... + m_d$ *where* $\boldsymbol{m} = (m_1, m_2, ..., m_d)$. *For any* $\boldsymbol{m}$ *such that* $|\boldsymbol{m}| = i \geq 0$ *and any* $b > 1$, *we have that*

$$\sum_i \sum_{|\boldsymbol{m}|=i} \prod_{j=1}^{d} (m_j + 1)^{-b} < \infty. \tag{5.41}$$

Proof. For $1 > \rho > \frac{d-b}{d-1}$, we have

$$\begin{aligned}
&\sum_{|\boldsymbol{m}|=i} \prod_{j=1}^{d} (m_j+1)^{-b} \leq \left(\frac{i+d}{d}\right)^{-(b-\rho)} \sum_{|\boldsymbol{m}|=i} \left[\prod_{j=1}^{d} (m_j+1)^{-\rho}\right] \\
&= \left(\frac{i+d}{d}\right)^{-(b-\rho)} (i+d)^{-d\rho} \sum_{|\boldsymbol{m}|=i} \left[\prod_{j=1}^{d-1} \left(\frac{m_j+1}{i+d}\right)^{-\rho}\right] \cdot \left(1 - \sum_{j=1}^{d-1} \left(\frac{m_j+1}{i+d}\right)\right)^{-\rho} \\
&= \left(\frac{i+d}{d}\right)^{-(b-\rho)} (i+d)^{-d\rho} \sum_{|\boldsymbol{m}|=i} \left[\prod_{j=1}^{d-1} \left(\frac{m_j+1}{i+d}\right)^{-\rho}\right] \cdot \left(1 - \sum_{j=1}^{d-1} \left(\frac{m_j}{i+d}\right)\right)^{-\rho} \left(\frac{m_d+d}{m_d+1}\right)^{\rho}.
\end{aligned} \tag{5.42}$$

Consider variable transformation $u_j = \frac{m_j+1}{i+d}$ and therefore $u_j \in \{\frac{1}{i+d}, \frac{2}{i+d}, \cdots, \frac{i+1}{i+d}\}$. Condition $|\boldsymbol{m}| = i$ becomes $\lceil \boldsymbol{u} \rceil = 1$. Consider $y_j \in (u_j - \frac{1}{i+d}, u_j), j < d$ and $y_d \in (u_d, u_d + \frac{d-1}{i+d})$, we have

$$\begin{aligned}
&\sum_{|\boldsymbol{m}|=i} \prod_{j=1}^{d} (m_j+1)^{-b} \\
&\leq \left(\frac{i+d}{d}\right)^{-(b-\rho)} (i+d)^{-d\rho} \sum_{\lceil \boldsymbol{u} \rceil=1} \left[\prod_{j=1}^{d-1} u_j^{-\rho}\right] \cdot \left(u_d + \frac{d-1}{i+d}\right)^{-\rho} \left(\frac{m_d+d}{m_d+1}\right)^{\rho} \\
&\leq \left(\frac{i+d}{d}\right)^{-(b-\rho)} (i+d)^{-d\rho} \cdot \left[(i+d)^{d-1} \int_{\sum y_i=1} \left\{\prod_{j=1}^{d} y_j^{-\rho}\right\} \mathrm{d}y_1 \cdots \mathrm{d}y_{d-1}\right] \left(\frac{m_d+d}{m_d+1}\right)^{\rho} \\
&\leq d^b (i+d)^{-b+(d-1)(1-\rho)} \cdot \frac{\Gamma(1-\rho)^d}{\Gamma(d(1-\rho))}
\end{aligned} \tag{5.43}$$

Therefore, if $\rho > \frac{d-b}{d-1}$ then

$$\sum_i \sum_{|\boldsymbol{m}|=i} \prod_{j=1}^{d} (n_j+1)^{-b} \leq \sum_i \underbrace{d^b (i+d)^{-b+(d-1)(1-\rho)} \cdot \frac{\Gamma(1-\rho)^d}{\Gamma(d(1-\rho))}}_{:=\xi_i} < \infty. \tag{5.44}$$

The lemma is proved. □

Lemma 5.7. *Suppose f_n is rth differentiable for some integer $r \geq 2$. Then for $q(N_n) = O(N_n^{\frac{1}{r}})$, the mean integrated square error of the estimate $\hat{f}_n(\boldsymbol{s}|\boldsymbol{\theta}), \forall \boldsymbol{\theta}$, satisfies*

$$\mathbb{E} \int \sup_{\boldsymbol{\theta}} |\hat{f}_n(\boldsymbol{s}|\boldsymbol{\theta}) - f_n(\boldsymbol{s}|\boldsymbol{\theta})|^2 \mathrm{d}\boldsymbol{s} = O(N_n^{\frac{1}{r}-1}) \tag{5.45}$$

and further for any value $\boldsymbol{s}$, we have

$$\sup_{\boldsymbol{\theta}} |\hat{f}_n(\boldsymbol{s}|\boldsymbol{\theta}) - f_n(\boldsymbol{s}|\boldsymbol{\theta})| = O_p(N_n^{\frac{1}{2r}-\frac{1}{2}}). \tag{5.46}$$

Proof.

$$\begin{aligned}
&\mathbb{E} \int \sup_{\boldsymbol{\theta}} |\hat{f}_n(\boldsymbol{s}|\boldsymbol{\theta}) - f_n(\boldsymbol{s}|\boldsymbol{\theta})|^2 \mathrm{d}\boldsymbol{s} \\
\leq &\mathbb{E} \int \sup_{\boldsymbol{\theta}} \left\{ \left[\sum_{m=0}^{q(N_n)} \sum_{\boldsymbol{m}} \hat{C}_{\boldsymbol{m}} \mathcal{H}_{\boldsymbol{m}}(\boldsymbol{s}) \right]^2 - 2 \sum_{m=0}^{q(N_n)} \sum_{\boldsymbol{m}} \left| (\hat{C}_{\boldsymbol{m}} - C_{\boldsymbol{m}}) \mathcal{H}_{\boldsymbol{m}}(\boldsymbol{s}) \right| \sum_{k=q(N_n)+1}^{\infty} \sum_{\boldsymbol{k}} |C_{\boldsymbol{k}} \mathcal{H}_{\boldsymbol{k}}(\boldsymbol{s})| \right\} \mathrm{d}\boldsymbol{s} \\
&+ \int \sup_{\boldsymbol{\theta}} \left[\sum_{m=q(N_n)+1}^{\infty} \sum_{\boldsymbol{m}} C_{\boldsymbol{m}} \mathcal{H}_{\boldsymbol{m}}(\boldsymbol{s}) \right]^2 .
\end{aligned}$$

Note that $C_{\boldsymbol{m}}$ and $\hat{C}_{\boldsymbol{m}}$ depend on $\boldsymbol{\theta}$, but $\mathcal{H}_{\boldsymbol{m}}(s)$ does not depend on $\boldsymbol{\theta}$.

Using the Hermite function properties $\int \mathcal{H}_{\boldsymbol{j}}(\boldsymbol{s}) \mathcal{H}_{\boldsymbol{k}}(\boldsymbol{s}) \mathrm{d}\boldsymbol{s} = 1$ if $\boldsymbol{j} = \boldsymbol{k}$; and $\int \mathcal{H}_{\boldsymbol{j}}(\boldsymbol{s}) \mathcal{H}_{\boldsymbol{k}}(\boldsymbol{s}) \mathrm{d}\boldsymbol{s} = 0$ if $\boldsymbol{j} \neq \boldsymbol{k}$, we have

$$\mathbb{E} \int \sup_{\boldsymbol{\theta}} \left[\sum_{m=0}^{q(N_n)} \sum_{\boldsymbol{m}} (\hat{C}_{\boldsymbol{m}} - C_{\boldsymbol{m}}) \mathcal{H}_{\boldsymbol{m}}(\boldsymbol{s}) \right]^2 \mathrm{d}\boldsymbol{s} = \sum_{m=0}^{q(N_n)} \sum_{\boldsymbol{m}} \mathbb{E} \sup_{\boldsymbol{\theta}} \left[(\hat{C}_{\boldsymbol{m}} - C_{\boldsymbol{m}})^2 \right], \tag{5.47}$$

$$\mathbb{E}\int \sup_{\boldsymbol{\theta}} \left\{ \sum_{m=0}^{q(N_n)} \sum_{\boldsymbol{m}} \left| (\hat{C}_{\boldsymbol{m}} - C_{\boldsymbol{m}}) \mathcal{H}_{\boldsymbol{m}}(\boldsymbol{s}) \right| \sum_{k=q(N_n)+1}^{\infty} \sum_{\boldsymbol{k}} |C_{\boldsymbol{k}} \mathcal{H}_{\boldsymbol{k}}(\boldsymbol{s})| \right\} \, \mathrm{d}\boldsymbol{s} = 0 \quad (5.48)$$

and

$$\int \sup_{\boldsymbol{\theta}} \left[\sum_{m=q(N_n)+1}^{\infty} \sum_{\boldsymbol{m}} C_{\boldsymbol{m}} \mathcal{H}_{\boldsymbol{m}}(\boldsymbol{s}) \right]^2 \mathrm{d}\boldsymbol{s} = \sum_{m=q(N_n)+1}^{\infty} \sum_{\boldsymbol{m}} \sup_{\boldsymbol{\theta}} [C_{\boldsymbol{m}}]^2 . \quad (5.49)$$

Using the Hermite function result in Walter (1977),

$$|\mathcal{H}_{\boldsymbol{m}}(x)| \leq \prod_{j=1}^{d} A_{\infty} \cdot (m_j + 1)^{-\frac{1}{12}}, \;\; \forall x \in (-\infty, \infty), \;\; j = 0, 1, \cdots , \quad (5.50)$$

where A_{∞} is a constant, and using (5.6) and (5.7), we know that $|C_{\boldsymbol{m}}| \leq \prod_{j=1}^{M} A_{\infty} \cdot (m_j + 1)^{-\frac{1}{12}}$, $\mathrm{Var}\Big(\mathrm{H}_j(v_i^{(k)})\Big) \leq A_{\infty}^2 (j+1)^{-\frac{1}{6}}$

$$\begin{aligned} \mathbb{E} \sup_{\boldsymbol{\theta}} \left(\hat{C}_j - C_j \right)^2 &= \frac{1}{N_n^2} \mathbb{E} \sup_{\boldsymbol{\theta}} \left\{ \sum_{k=1}^{N_n} \left(\mathcal{H}_{\boldsymbol{m}}(v^{(k)}) - C_{\boldsymbol{m}} \right)^2 \right\} \\ &\leq \frac{2}{N_n} \prod_{j=1}^{M} A_{\infty}^2 (m_j + 1)^{-\frac{1}{6}}, \end{aligned} \quad (5.51)$$

Then, based on the Lemma 5.6, we have

$$\begin{aligned} &\mathbb{E}\int \sup_{\boldsymbol{\theta}} |\hat{f}_n(\boldsymbol{s}|\boldsymbol{\theta}) - f_n(\boldsymbol{s}|\boldsymbol{\theta})|^2 \mathrm{d}\boldsymbol{s} \leq \sum_{m=0}^{q(N_n)} \sum_{\boldsymbol{m}} \mathbb{E} \sup_{\boldsymbol{\theta}} (\hat{C}_{\boldsymbol{m}} - C_{\boldsymbol{m}})^2 + \sum_{m=q(N_n)+1}^{\infty} \sum_{\boldsymbol{m}} \sup_{\boldsymbol{\theta}} [C_{\boldsymbol{m}}]^2 \\ \leq &\frac{2}{N_n} A_{\infty}^2 \cdot \sum_{m=0}^{q(N_n)} \sum_{\boldsymbol{m}} \prod_{j=1}^{M} (m_j + 1)^{-\frac{1}{6}} + \sum_{m=q(N_n)+1}^{\infty} \sum_{\boldsymbol{m}} \sup_{\boldsymbol{\theta}} [C_{\boldsymbol{m}}]^2 \\ \leq &A_1 \frac{q(N_n)}{N_n} + \sum_{m=q(N_n)+1}^{\infty} \sum_{\boldsymbol{m}} \sup_{\boldsymbol{\theta}} \{ {C_{\boldsymbol{m}}}^2 \}, \end{aligned}$$

where A_1 are constants.

Therefore, for integer $r \geq 1$,

$$\begin{aligned} C_{\boldsymbol{m}} &= \int f_n(\boldsymbol{v}) \prod_{j=1}^{M} \left[\mathrm{h}_{m_j}(v_j) e^{-\frac{v_j^2}{2}} \mathrm{d}v_j \right] = \int \prod_{j=1}^{M} \left[(2m_j)^{-\frac{1}{2}} \mathrm{h}_{m_j+1}(v_j) \frac{\partial}{\partial v_j} \left(f_n(\boldsymbol{v}) \cdot e^{-\frac{\boldsymbol{v}\boldsymbol{v}^T}{2}} \right) \right] \\ &= \int \prod_{j=1}^{M} \left[(2m_j)^{-\frac{1}{2}} (2m_j+2)^{-\frac{1}{2}} ... (2m_j+2r-2)^{-\frac{1}{2}} \mathrm{h}_{m_j+r}(v_j) \frac{\partial^r}{\partial {v_j}^r} \left(f_n(\boldsymbol{v}) \cdot e^{-\frac{\boldsymbol{v}\boldsymbol{v}^T}{2}} \right) \right] \\ &\leq \prod_{j=1}^{M} \left[(2m_j)^{-\frac{r}{2}} \int \mathrm{h}_{m_j+r}(v_j) \frac{\partial^r}{\partial {v_j}^r} (f_n(\boldsymbol{v}) \cdot e^{-\frac{\boldsymbol{v}\boldsymbol{v}^T}{2}}) \right]. \end{aligned} \tag{5.52}$$

Let $b_{m_j} = \int \mathrm{h}_{m_j+r}(v_j) \frac{\partial^r}{\partial {v_j}^r} (f_n(\boldsymbol{v}) \cdot e^{-\frac{\boldsymbol{v}\boldsymbol{v}^T}{2}})$, which could be regarded as the coefficients of function $\frac{\partial^r}{\partial {v_j}^r} (f_n(\boldsymbol{v}) \cdot e^{-\frac{\boldsymbol{v}\boldsymbol{v}^T}{2}})$. The series sum $\sum b_{m_j}^2$ should be bounded above by some constant (Walter, 1977). So,

$$\begin{aligned} \sum_{m=q+1}^{\infty} \sum_{\boldsymbol{m}} \sup_{\boldsymbol{\theta}} {C_{\boldsymbol{m}}}^2 &\leq \sum_{m=q+1}^{\infty} \sum_{\boldsymbol{m}} \prod_{j=1}^{M} \left[(2m_j)^{-\frac{r}{2}} \int \mathrm{h}_{m_j+r}(v_j) \frac{\partial^r}{\partial {v_j}^r} (f_n(\boldsymbol{v}) \cdot e^{-\frac{\boldsymbol{v}\boldsymbol{v}^T}{2}}) \right]^2 \\ &\leq \frac{1}{(q(N_n))^r} \cdot \sum_{\boldsymbol{m}} b_{m_j}^2 = O(N_n^{-1}), \end{aligned} \tag{5.53}$$

which is much less than $A_1 \frac{q(N_n)}{N_n} = O(N_n^{\frac{1}{r}-1})$. The theorem is proved. □

We also need the following condition

Condition 5.7. *Suppose that the correlated statistic of $\boldsymbol{s} = (s_1, \cdots, s_M)$ with a fixed number M is calculated based on all n observations and it is such that*

1. *$\boldsymbol{s} \to \boldsymbol{s}^*$ almost surely and $\mathbb{E}\|\boldsymbol{s} - \boldsymbol{s}^*\|^2 \to 0$, as $n \to \infty$ and the central limit theorem holds for $\boldsymbol{s}$, i.e. $\sqrt{n}(\boldsymbol{s} - \boldsymbol{s}^*) \Rightarrow \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{s}})$; and*

2. *we can write*

$$\frac{\partial f_n(\boldsymbol{s}|\boldsymbol{\theta})}{\partial \boldsymbol{s}} f_n(\boldsymbol{s}|\boldsymbol{\theta})^{-1} = \rho_n \cdot g_n(\boldsymbol{s}|\boldsymbol{\theta}) \tag{5.54}$$

where $\rho_n = O(n)$ and $\sup_{\boldsymbol{s}} |\frac{\partial}{\partial \boldsymbol{s}} g_n(\boldsymbol{s}|\boldsymbol{\theta})| = O(1)$.

Theorem 5.4. *Under Condition 5.7, The maximum point $\hat{\boldsymbol{\theta}}$ of $\hat{l}(\boldsymbol{\theta}|\boldsymbol{s})$ given in (5.40) converges to the true parameter value $\boldsymbol{\theta}^*$ in probability and it should satisfy*

$$\begin{aligned}&\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\to\mathcal{N}\left(\mathbf{0},\boldsymbol{\Sigma}\right)\\&\boldsymbol{\Sigma}=\left(\boldsymbol{\phi}^{(1)}(\boldsymbol{s}^*,\boldsymbol{\theta}^*)\right)^{-1}\boldsymbol{\psi}(\boldsymbol{s}^*,\boldsymbol{\theta})\boldsymbol{\Sigma}_{\boldsymbol{s}}\boldsymbol{\psi}(\boldsymbol{s}^*,\boldsymbol{\theta})\left(\boldsymbol{\phi}^{(1)}(\boldsymbol{s}^*,\boldsymbol{\theta}^*)\right)^{-1}.\end{aligned}\tag{5.55}$$

Proof. By noticing a similar result as Lemma 5.2 can be achieved for the log-likelihood $\hat{l}(\boldsymbol{\theta}|\boldsymbol{s})$, the consistency of $\hat{\boldsymbol{\theta}}$ is easy to prove.

Using first-order Taylor expansion on $\left.\frac{\partial\log\hat{l}(\boldsymbol{\theta}|\boldsymbol{s})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ at point $\boldsymbol{\theta}^*$ we have

$$\begin{aligned}\left.\frac{\partial\log\hat{l}(\boldsymbol{\theta}|\boldsymbol{s})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}&=\left[\left.\frac{\frac{\partial}{\partial\boldsymbol{\theta}}\hat{f}_n(\boldsymbol{s}|\boldsymbol{\theta})}{\hat{f}_n(\boldsymbol{s}|\boldsymbol{\theta})}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}+\frac{\partial}{\partial\boldsymbol{\theta}}\left.\frac{\frac{\partial}{\partial\boldsymbol{\theta}}\hat{f}_n(\boldsymbol{s}|\boldsymbol{\theta})}{\hat{f}_n(\boldsymbol{s}|\boldsymbol{\theta})}\right|_{\boldsymbol{\theta}=\breve{\boldsymbol{\theta}}}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\right]\\&=\left[\left.\frac{\frac{\partial}{\partial\boldsymbol{\theta}}\tilde{f}_n(\boldsymbol{s}|\boldsymbol{\theta})}{\tilde{f}_n(\boldsymbol{s}|\boldsymbol{\theta})}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}+\frac{\partial}{\partial\boldsymbol{\theta}}\left.\frac{\frac{\partial}{\partial\boldsymbol{\theta}}\tilde{f}_n(\boldsymbol{s}|\boldsymbol{\theta})}{\tilde{f}_n(\boldsymbol{s}|\boldsymbol{\theta})}\right|_{\boldsymbol{\theta}=\breve{\boldsymbol{\theta}}}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\right]+O_{\mathrm{p}}\left(N_n^{\frac{1}{2}}\right)\\&=\left[\boldsymbol{\phi}(\boldsymbol{s},\boldsymbol{\theta}^*)+\boldsymbol{\phi}^{(1)}(\boldsymbol{s},\breve{\boldsymbol{\theta}})\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\right]+O_{\mathrm{p}}\left(N_n^{\frac{1}{2}}\right)\end{aligned}\tag{5.56}$$

where $\breve{\boldsymbol{\theta}}$ is between $\boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}}$.

The second equality is because, given the observations or the statistics $\boldsymbol{s}$, $\sqrt{N_n}(\tilde{f}_n-\hat{f}_n)\to\mathcal{N}(0,\sigma^2)$ where $\sigma^2=\sum_{m=0}^{q(N_n)}\sum_{\boldsymbol{m}}\mathrm{var}(C_{\boldsymbol{m}}-\widehat{C}_{\boldsymbol{m}})\cdot\mathcal{H}_{\boldsymbol{m}}(\boldsymbol{s})<\infty$.

On the other hand, we have

$$\boldsymbol{\phi}(\boldsymbol{s},\boldsymbol{\theta})=\frac{1}{\tilde{f}_n(\boldsymbol{s}|\boldsymbol{\theta})}\sum_{m=1}^{q(N_n)}\sum_{\boldsymbol{m}}\mathcal{H}_{\boldsymbol{m}}(\boldsymbol{s})\int\frac{\partial}{\partial\boldsymbol{\theta}}f_n(\boldsymbol{v}|\boldsymbol{\theta})\cdot\mathcal{H}_{\boldsymbol{m}}(\boldsymbol{v})\mathtt{d}\boldsymbol{v}.\tag{5.57}$$

We shall also further expand $\boldsymbol{\phi}(\boldsymbol{s},\boldsymbol{\theta})$ in terms of $\boldsymbol{s}$ at $\boldsymbol{s}^*$, i.e.

$$\boldsymbol{\phi}(\boldsymbol{s},\boldsymbol{\theta})=\boldsymbol{\phi}(\boldsymbol{s}^*,\boldsymbol{\theta})+\boldsymbol{\psi}(\boldsymbol{s}^*,\boldsymbol{\theta})\cdot(\boldsymbol{s}-\boldsymbol{s}^*)+O_{\mathrm{p}}(\|\boldsymbol{s}-\boldsymbol{s}^*\|^2\cdot I_{\boldsymbol{\theta}})\tag{5.58}$$

where $I_{\boldsymbol{\theta}}$ is a coefficience vector. Using the property that as $n\to\infty$, $\sqrt{n}(\boldsymbol{s}-\boldsymbol{s}^*)\sim$

$\mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{s}})$ under Condition 5.1,5.2,5.3, we will have

$$\sqrt{n}\left(\boldsymbol{\phi}(\boldsymbol{s}, \boldsymbol{\theta}) - \boldsymbol{\phi}(\boldsymbol{s}^*, \boldsymbol{\theta})\right) \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\psi}(\boldsymbol{s}^*, \boldsymbol{\theta})\boldsymbol{\Sigma}_{\boldsymbol{s}}\boldsymbol{\psi}(\boldsymbol{s}^*, \boldsymbol{\theta})\right). \tag{5.59}$$

In the equation (5.56), the left-hand-side is $\mathbf{0}$ at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Therefore, asymptotically

$$\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) = -\left(\boldsymbol{\phi}^{(1)}(\boldsymbol{s}, \boldsymbol{\theta}^*)\right)^{-1} \boldsymbol{\phi}(\boldsymbol{s}, \boldsymbol{\theta}^*), \qquad \text{equal in distribution.} \tag{5.60}$$

Due to $\mathbb{E}\boldsymbol{\phi}(\boldsymbol{s}, \boldsymbol{\theta}^*) = \mathbf{0}$, we have

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\Sigma} = \left(\boldsymbol{\phi}^{(1)}(\boldsymbol{s}^*, \boldsymbol{\theta}^*)\right)^{-1} \boldsymbol{\psi}(\boldsymbol{s}^*, \boldsymbol{\theta})\boldsymbol{\Sigma}_{\boldsymbol{s}}\boldsymbol{\psi}(\boldsymbol{s}^*, \boldsymbol{\theta}) \left(\boldsymbol{\phi}^{(1)}(\boldsymbol{s}^*, \boldsymbol{\theta}^*)\right)^{-1}. \tag{5.61}$$

□

Chapter 6

Experiment

Based on the algorithm 7 provided, we could do simulation studies in computer. This chapter provide the evidence for theories according to experiments. The first section is for simulating samples and the other one for observations from the real data.

6.1 Simulation Studies

In this section, original data or observations in our study are sampled randomly from given generative models. Thus, replicated experiments are available in the situation. Here we focus on two examples consist of g-and-k Distribution and LV models.

In the prior experiments, we find that the correlated statistics perform poorly in parametric inference, for which there is no theoretic evidence either. Thus we would only display the independent statistics instead of correlated ones, which means all data set would be split by groups in our method. As for limited data or small simple size, we have to notice that other methods, such as ABC, cannot perform well either as the data hardly provide enough information for estimate.

6.1.1 Example 1 – g-and-k Distribution

Firstly, we use g-and-k distribution as an example to compare our methods with other likelihood-free algorithms in the literature including the indirect methods (Gourieroux et al., 1993), the Approximate Bayesian computation (ABC) in Allingham et al. (2009) and Fearnhead and Prangle (2012).

We can usually choose $M = p$ as the dimension of the unknown parameter vector. In practice, one may choose many more statistics. For example (Fearnhead and Prangle, 2012) and (Allingham et al., 2009) suggested using hundreds of statistics, i.e. M is no less than 100 for $n = 10000$ observations. However, this is not recommended in our method. In our method, choosing more statistics S_i (larger M) means that each S_i will be constructed based less number of observations ($K_n = n/M$), which will result in unstable estimate for σ_i^2 in (5.26). Also choosing a larger M means that we need more computing cost which is prohibitive in many complex statistical analysis.

Following Rayner and MacGillivray (2002), we consider $B > 0$, $\mathtt{k} > 0$ and fixing $\mathtt{c} = 0.8$. We use quantile statistics for S_i in our algorithm. Since there are four unknown parameters, we only need to extract four statistics S_1, S_2, S_3, S_4 from the data, i.e. $M = 4$. We therefore choose $S_i, i = 1, \cdots, M$ as the $\frac{i}{M+1}$ quantile, i.e. the $20\%, 40\%, 60\%, 80\%$ quantiles. To ensure the independence of the sample statistics S_1, S_2, S_3, S_4, we randomly split the n samples $\mathbf{y}$ into $M = 4$ groups $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4$, each with $\frac{n}{M}$ observations. For the ith sub-sample, we construct S_i as the $\frac{i}{M+1}$-quantile of $\mathbf{y}_i$.

Simulation results – comparison with existing methods

Based on the independent quantile statistics S_i, we can construct the working likelihood (5.9) and use Algorithm 7 to find the maximum likelihood estimate for $\boldsymbol{\theta}$. To justify the performance of the estimators, following Fearnhead and Prangle

(2012), we consider "absolute errors" for the ith parameter as,

$$|err_i| = \sqrt{(\theta_i^* - \hat{\theta}_i)^2}. \tag{6.1}$$

We set the true parameter value $\boldsymbol{\theta}^* = (3, 1, 2, 0.5)$ for the g-and-k distribution and sample size $n = 10^4$. For our proposed Hermite expansion method we set MC samples $N_n = 200$, order of Hermite Polynomials $q(N_n) = 16$ (because $q(N_n) = O(N_n^{\frac{1}{r}})$ with $r = 2$), $M = 4$ and S_i as the α_i sample quantile ($\alpha_i = \frac{i}{M+1}$). Here you may have a question that $q(N_n) \approx 14$ instead of $q(N_n) = 16$ when $r = 2$. From the theory, we can find that all values around 14 is valid for the experiment. If the order of function is restricted, we would prefer the larger one as the more precise result we may obtain. However, the computation would increase as well. After balancing the precision and the time complexity, we choose the order $q(N_n) = 16$. Note that, Garcia et al. (2011) also recommended such a prospective that only four quantile statistics ($M = 4$) with pairs of symmetry at the medium point is needed in the Indirect Inference.

We compare our method with other existing likelihood-free approaches. The results are summarized in Table 6.1, which include results based on our proposed method, the result of quantile-based indirect method in Garcia et al. (2011), the ABC method using all sorted observations and results based on full sorted data and auxiliary regression model in Allingham et al. (2009), results based on the semi-automatic ABC approach (100 ordered statistics) and semi-automatic ABC with auxiliary regression model in Fearnhead and Prangle (2012), and the maximum likelihood approach. The prior distribution in the Bayesian approach is set as a uniform distribution $U[0, 10]$. Such uniform priors will provide results equivalent to likelihood inference.

As shown in the Table 6.1, Hermite Polynomial expansion method provides the best parameter estimates and even works better than the method Maximum Likelihood Estimation (based on numerical calculation of the likelihood). ABC is almost

Method;bias(variance)	A	B	g	k
Semi-auto ABC	0.015 (0.023)	0.066(0.049)	0.030(0.57)	0.31(0.40)
Indirect Inference	0.013(0.00017)	0.035(0.0012)	0.054(0.0038)	0.057(0.0025)
NMLE	0.013(0.00010)	0.023(0.00044)	0.036(0.00077)	0.037(0.00014)
Hermite Expansion*	0.00048(0.00032)	0.00063(0.0026)	0.0032(0.024)	0.0092(0.0078)

Table 6.1: Bias (MC variance) of various existing methods and Hermite expansion methods based on 50 replicated data sets with true parameters $(3, 1, 2, 0.5)$ and sample size $n = 10000$. NMLE stands for numerical maximum likelihood estimation(Rayner and MacGillivray, 2002).

Method;bias(variance)	A	B	g	k
Semi-auto ABC	0.025 (0.023)	0.070(0.11)	0.16(1.96)	0.33(0.31)
Indirect Inference	0.0086(0.0014)	0.017(0.0076)	0.032(0.043)	0.034(0.034)
NMLE	0.010(0.0010)	0.013(0.0065)	0.0047(0.012)	0.0051(0.0022)
Hermite Expansion*	0.0015(0.0031)	0.036(0.012)	0.012(0.029)	0.0047(0.012)

Table 6.2: Bias (MC variance) of various existing methods and Hermite expansion methods based on 50 replicated data sets with true parameters $(3, 1, 2, 0.5)$ and sample size $n = 1000$.

as good as the numerical MLE method. Note that, for the ABC method, the number of statistics, the choice of distance function and the value of tolerance determine its performance. Although higher number of ordered statistics and small tolerance values will lead to good estimation accuracy, they will also lead to much heavier computational costs and make ABC methods not practical. If we choose 50 or 25 ordered statistics, or change to larger tolerance values, the ABC method will provide poor results. The biases of the estimates based on other method cannot be quantified (Drovandi et al., 2015). However, our Hermite Expansion method does not require such pre-specified tuning parameters and provides much smaller mean quadratic losses by using appropriate order of the polynomials.

Compared with Table 6.2, we can find the fact the performance of methods almost go worse with smaller samples adopted, especially for semi-auto ABC. The table

shows that, with less samples in the model, the bias of Indirect Inference and NMLE slightly decrease even though the variance increase. It might be the randomness of unbiased estimate. In terms of parameter estimate for g and k, semi-auto ABC performs far worse than the other three ones. It means semi-auto ABC prefers the large simple size. Our proposed method also requires a larger data set but the bias does not that significantly raise if proper tuning parameters are set (such as increasing the order of Hermite Polynomials Expansion).

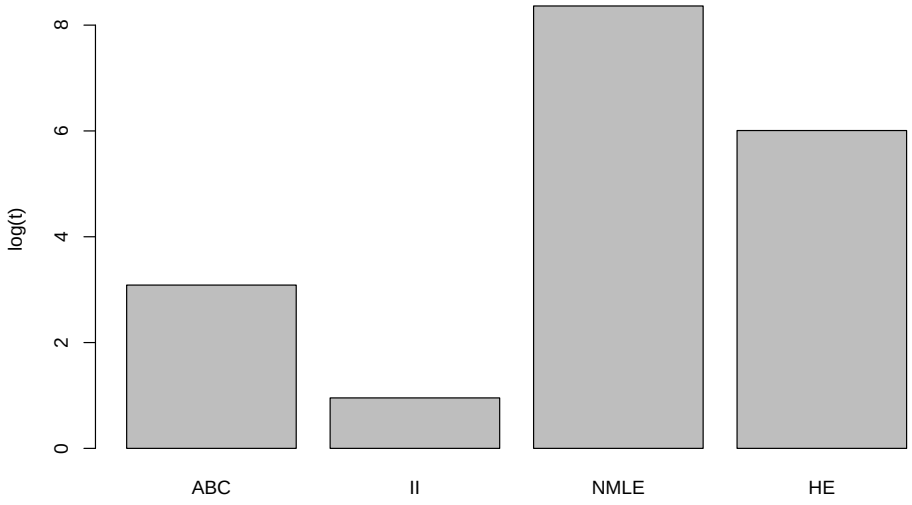

Figure 6.1: System time ($\log(t)$) of various methods based on a data set with simple size $n = 10000$.Here ABC stands for semi-auto ABC, II for Indirect Inference, NMLE for numeric maximum likelihood estimate and HE for Hermite Expansion methods

As for the computational complexity, Figure 6.1 shows how long the four methods spend obtaining the output based on a data set. Our proposed methods performs average in calculating time as we require a high level of less bias. Meanwhile, both NMLE and our proposed methods,using 'mle' or 'optim' function in R, need to map a maximum point over an area, which requires much replicated calculation at each point. There is no doubt that quantile-based Indirect Inference spends least time for estimate parameters of g-and-k distribution due to the special format for the case.

Another advantage of the proposed method is that our estimate $\hat{\boldsymbol{\theta}}$ is guaranteed to converge to the true value parameter $\boldsymbol{\theta}^*$. None of the other existing methods can provide such large sample properties. Figure 6.2 shows the changes of mean square errors of the parameter estimates, as the sample size n increases, where the simulations settings are the same as before, with $N_n = 200$, $q(N_n) = 16$ and $M = 4$. We can see from Figure 6.2 that mean square errors of all parameter estimates are decreasing in general as n increases.

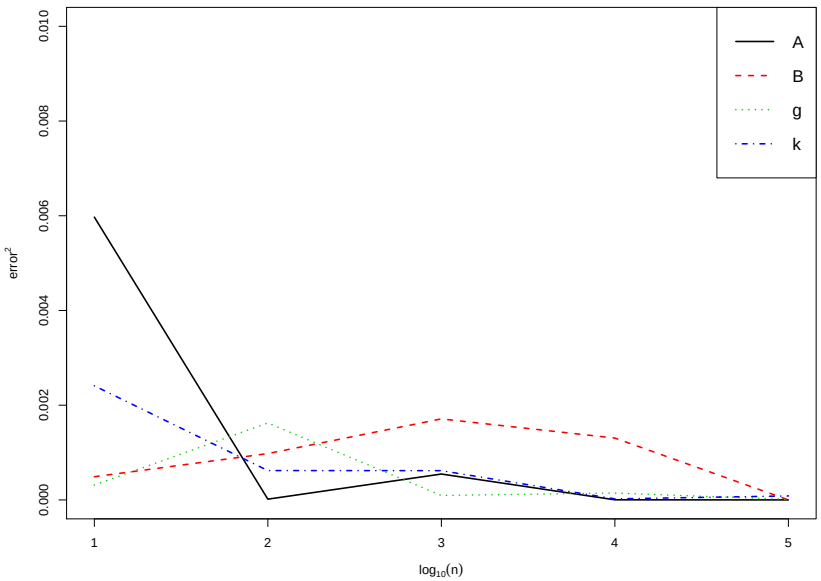

Figure 6.2: Mean squared errors between estimates and true value against $\log(n)$ for our proposed method.

Another clear superiority of the new Hermite expansion method is that it tells how many statistics we should use in order to construct the working likelihood (in the order of $O(1)$). The ABC approach uses 100 quantile statistics in their analysis and there is no solid theoretical justification on how many statistics which should be used in ABC.

Both our methods and the ABC approach need to simulate many replicate samples based on different possible parameter values. However, we have to choose small

tolerance values for ABC methods in order to achieve good estimates, which will require simulating many replicate data sets because the acceptance probability decreases to 0 as tolerance value goes down to 0. However, our new approach provides the exact order for N_n, the number of replicate data sets required.

To further justify our methods under different tuning parameter settings for N_n, $q(N_n)$ and M, we provide more detailed simulation results in the following subsection.

Simulation results for different tunning parameters

First, we study the performance of the proposed method under different values of N_n, the number of replicated Monte Carlo samples. The simulation setting is the same as before, with $n = 10,000$, $q(N_n) = 16$ and $M = 4$. The results are shown in Figure 6.3. We can see that the mean square errors for all parameter estimates do not change too much, under any value of $N_n = 10$ to $N_n = 10000$. This is clearly justified by Theorem 5.2, which tells us that the convergence rate of the estimates only depends on the actual sample size n. In other words, it only depends on the rate of each statistic s_i converging to s_i^*. The number of Monte Carlo replicates N_n is only required to be $N_n^{\frac{5}{12r}-\frac{1}{2}} = o(n^{-\frac{1}{2}})$ (see Theorem 5.1). As long as N_n satisfies this condition, when sample size n is fixed, increasing N_n will not improve much on the estimate accuracy.

Second, we study the performance of our method, when varying the number of statistics, M. The simulation setting is the same as before, with $n = 10,000$, $q(N_n) = 16$ and $N_n = 200$. From Figure **??** we can see that parameter estimates for A and B were not affected much by changing the values of M. However, as M increases (more statistics use in the algorithm), the mean square errors increases for parameter estimates of g and k. The main reason is that as M increases, the number of samples $\frac{n}{M}$ used to construct each s_i will decrease. Thus each statistic s_i converges to the true population quantile s_i^* at a lower rate, and the variance of each

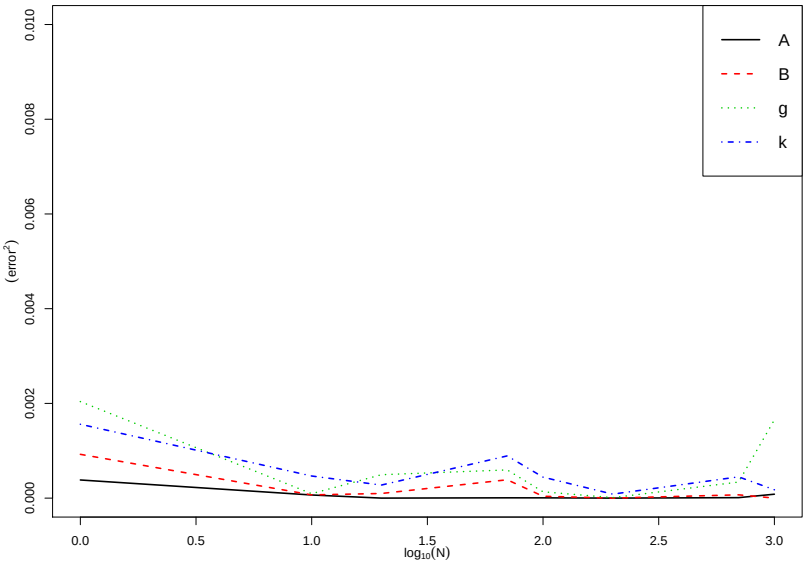

Figure 6.3: Mean squared errors between estimate parameter and true value of our proposed method with different parameter for different $log(N_n)$

s_i becomes large, which gives the main contribution on the variance of estimates (see the formula (5.32)).

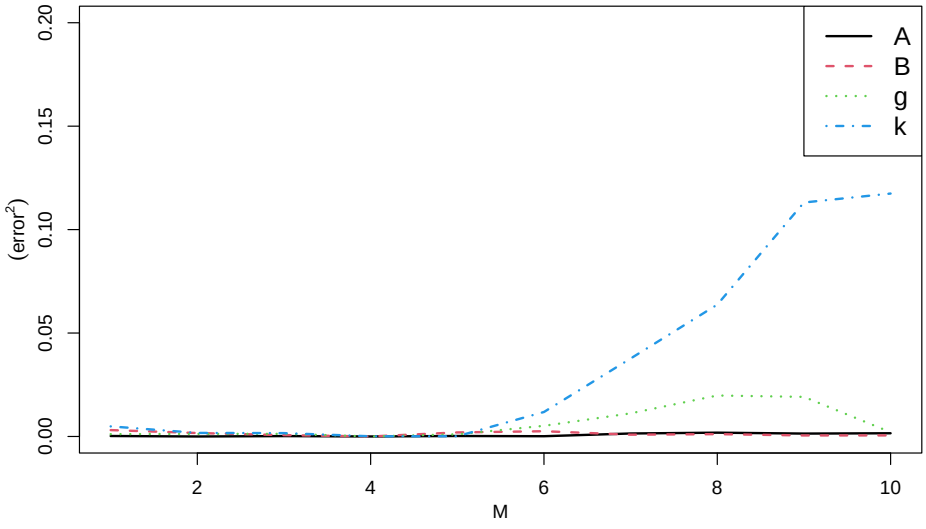

Figure 6.4: Mean square errors between estimates and true values against the number of statistics M.

In the end, we study how the order of Hermite expansion $q(N_n)$ affects the performance of the proposed method. When we chose small values of $q(N_n)$, such as $1, 2, 3$, the errors of the parameter estimates are very large. Figure 6.5 shows that bias of the parameter estimates decreases to 0 as $q(N_n)$ increases. This is indeed what we expect. On the other hand, Lemma 5.4 tells us that $q(N_n) = O(N_n^{\frac{1}{r}})$, where r is such that $f_{i,n}$ is rth differentiable. The implies that if $f_{i,n}$ is very smooth, i.e. r to be a large integer, then $q(N_n)$ takes smaller values (less computation costs needed); otherwise if r is a small integer then $q(N_n)$ takes larger values (more computation costs needed).

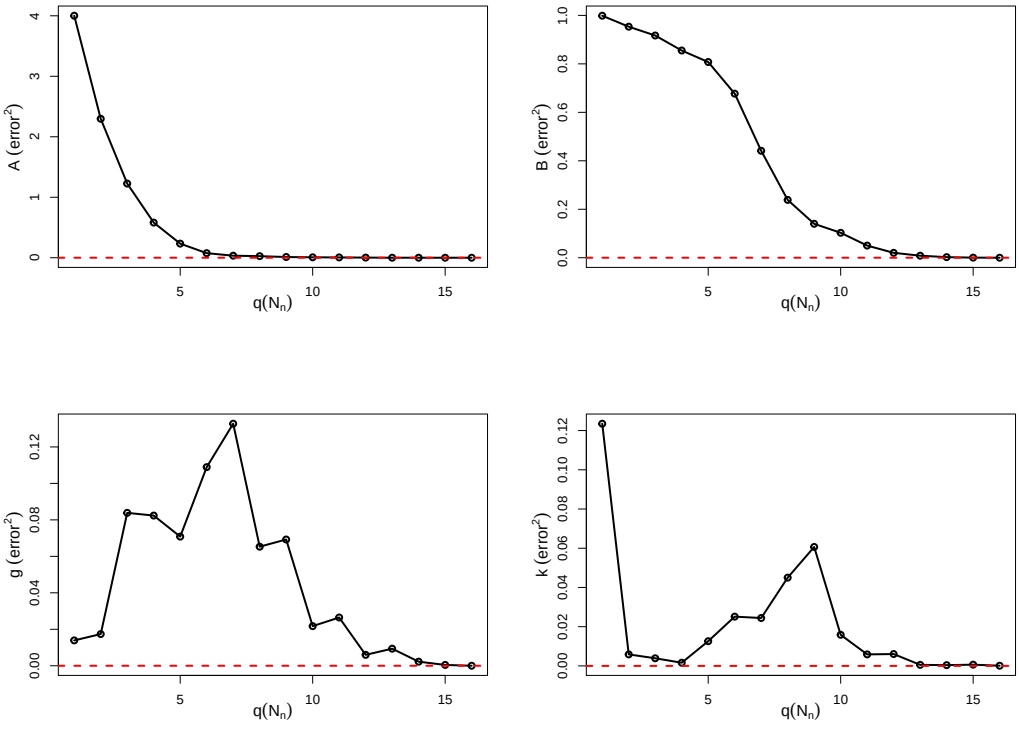

Figure 6.5: Mean square errors between estimates and true values against $q(N_n)$.

6.1.2 Example 2 – Lotka-Volterra models

Stochastic kinetic networks (SKN) models normally describe complex chemical reactions in a fixed volume containing the evaluation different species with certain types of reactions (Golightly and Wilkinson, 2005). A special type of SKN models is the Lotka-Volterra model (or prey-predator model), where there are 2 species containing preys (Y_1) and predators (Y_2) as well as 3 reactions, as introduced in Chapter Examples.

There are three unknown parameters in the model, therefore we just need to choose 3 statistics to construct the working likelihood. Similarly as before, we choose sample quantiles as well. To construct statistics s_1, s_2, s_3, we divide the whole process $(y_1(t), y_2(t))$ into three parts, in time intervals $[0, \tau_1]$, $(\tau_1, \tau_2]$ and $(\tau_2, T]$, respectively. Then given the observed data $\mathbf{y} = \{\boldsymbol{y}_{t_j}, t_j, j = 1, \cdots, n\}$, each s_i is constructed via different parts of the process, as

$$\begin{aligned} s_1 &:= \mathbb{T}_1(\mathbf{y}) = \frac{\#\{j;\ \text{such that } y_1(t_j) < a_1, t_j \in [0, \tau_1]\}}{\#\{j;\ \text{such that } t_j \in [0, \tau_1]\}}; \\ s_2 &:= \mathbb{T}_2(\mathbf{y}) = \frac{\#\{j;\ \text{such that } y_2(t_j) > a_2, t_j \in (\tau_1, \tau_2]\}}{\#\{j;\ \text{such that } t_j \in (\tau_1, \tau_2]\}}; \\ s_3 &:= \mathbb{T}_3(\mathbf{y}) = \frac{\#\{j;\ \text{such that } y_1(t_j) < a_3, t_j \in (\tau_2, T]\}}{\#\{j;\ \text{such that } t_j \in (\tau_2, T]\}} \end{aligned}$$

where $a_i, i = 1, 2, 3$ are predefined values. Therefore, each of these three statistics means the percentages of points of $y_1(t)$ (or $y_2(t)$) are below or above a certain threshold.

Remark 6.1. *Denote*

$$\begin{aligned} u_1 &= \max_j \{t_j;\ \textit{such that } t_j \in [0, \tau_1]\}, \\ u_2 &= \max_j \{t_j;\ \textit{such that } t_j \in (\tau_1, \tau_2]\}. \end{aligned}$$

Given any parameter $\boldsymbol{\theta}$, we simulate a pseudo data set $\mathbf{x} = \{\boldsymbol{x}_{t_1}, \cdots, \boldsymbol{x}_{t_n}\}$, condi-

tional on the points $\boldsymbol{x}_{t_1} = \boldsymbol{y}_{t_1}$, $\boldsymbol{x}_{u_1} = \boldsymbol{y}_{u_1}$ *and* $\boldsymbol{x}_{u_2} = \boldsymbol{y}_{u_2}$. *Then the three statistics* $s_1 = \mathbb{T}_1(\mathbf{x}), s_2 = \mathbb{T}_2(\mathbf{x}), s_3 = \mathbb{T}_3(\mathbf{x})$ *are independent, because they are based on three different pieces from a Markov chain and conditional on the boundary points* $\boldsymbol{y}_{u_1}$ *and* $\boldsymbol{y}_{u_2}$.

In our simulation, we take a_1 as the 0.33-quantile of $y_1(t_1), \cdots, y_1(t_n)$, and a_2 is the 0.67-quantile of $y_2(t_1), \cdots, y_2(t_n)$, and a_3 is the 0.67-quantile of $y_1(t_1), \cdots, y_1(t_n)$. Table 6.3 presents the results based on 50 simulations. In all simulations, our Gradient Descent Algorithm uses a starting point $\tilde{\boldsymbol{\theta}}_{\mathbf{0}}$ randomly generated from three uniform distribution: $\mathcal{U}[0.4, 0.6]$, $\mathcal{U}[0.002, 0.003]$, $\mathcal{U}[0.2, 0.4]$. The means of the estimates from the 50 simulations, as well as the standard deviation, are presented in the table. Note that the standard deviation of the parameter estimates can also be calculated via time-series bootstrap methods (Bühlmann, 2002).

Hermite Expansion performs much better on estimating the value of the parameter (much less biased) with slightly larger standard errors, comparing to the other methods. Also our method is very stable in terms of choosing different starting values. In terms of RJ methods and BU methods, the Metropolis-Hastings algorithm is adopted for estimating a new interval with a certain accepted probability (Boys et al., 2008). Meanwhile, DA methods use Gibbs sampling to approximate the pseudo-likelihood which is a joint posterior density function (Golightly and Wilkinson, 2005).

Method; true value	$\theta_1^* = 0.5$	$\theta_2^* = 0.0025$	$\theta_3^* = 0.3$
RJ	0.4799 (0.0171)	0.00255 (0.000094)	0.3075 (0.0113)
BU	0.4797(0.0170)	0.00247(0.000094)	0.3073 (0.0113)
DA	0.4800 (0.0163)	0.00254 (0.000091)	0.3067 (0.0110)
Hermite Expansion	0.5041(0.0315)	0.00247(0.000375)	0.2970(0.0421)

Table 6.3: Estimation on means (standard deviations) of rate constants using different algorithms based on 40 data points, evenly spread in the interval $[0, 40]$.

We also study the performance of our method under different tuning parameters N_n for this LV model application. The true rate parameters $\theta_1 = 0.5, \theta_2 = 0.0025, \theta_3 = 0.3$ and initial state with $y_1(0) = 71$ and $y_2(0) = 79$ in the process remain unchanged as before. With varying the number of simulate samples N_n, other tuning parameters for the simulation setting are the same as before, with $t_n = 40$, $q(N_n) = 16$ and $M = 3$. It is found that for all parameters, if we set $\log_{10}(N) \geq 2$ ($N \geq 100$), square errors for all parameters can be controlled less than 0.003. The results will not change much with much larger N_n, which correspond to the result of Theorem 5.1 that we only need N_n to be such that $N_n^{\frac{5}{12r}-\frac{1}{2}} = o(n^{-\frac{1}{2}})$. Choosing unnecessarily large N_n will only cost more computing time.

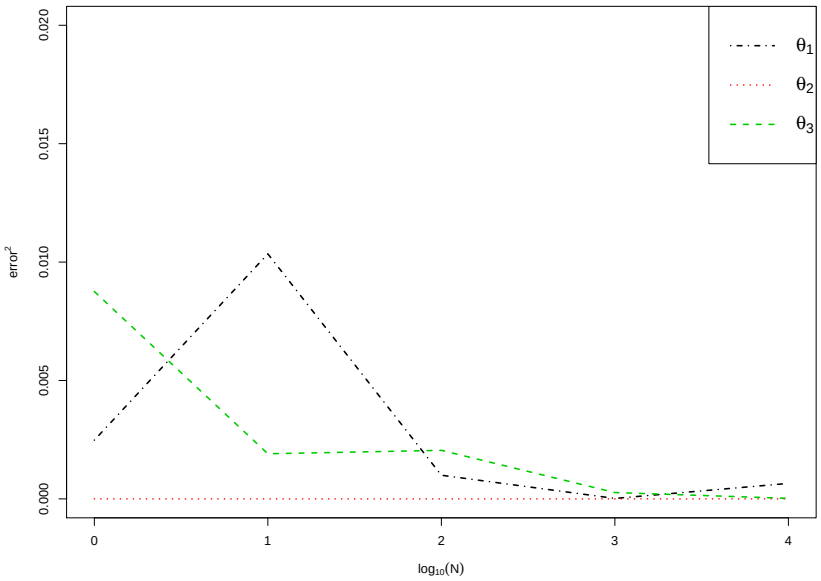

Figure 6.6: Error square between estimate parameter and true value of Hermite Expansion with different parameter for $\log_{10}(N_n)$

6.2 Real data application

As Covid-19 spread in 2020, many models for epidemics are provided for government or the public. Here SIR models, as one of most common models for epidemiol-

ogy, are adopted in our study for parameter inference.

6.2.1 Background

SIR models describe the trend of infected people with the contagious illness over time, where S stands for the number of the susceptible, I for the infectious and R for the recovered individuals (or immune or death). There has already been a vast literature on COVID-19 research based on SIR epidemiology models. In particular, the UK government's lock-down strategy were made according to such SIR model forecasting results (Ferguson et al., 2020). A crucial challenge of such forecasting research is how to estimate the model parameters properly, in particular for the basic reproduction number R_0. All existing research actually estimate R_0 via very naive approaches (Adam, 2020; Ferguson et al., 2020). We present in this section how to estimate R_0 via our proposed methods.

In the deterministic SIR model, there are three ordinary differential equations to govern the system:

$$\frac{\mathtt{d}S}{\mathtt{d}t} = -\beta SI, \qquad \frac{\mathtt{d}I}{\mathtt{d}t} = \beta SI - \gamma I, \qquad \frac{\mathtt{d}R}{\mathtt{d}t} = \gamma I,$$

where β is the infection rate, and γ is the recovery rate. Meanwhile, the epidemiological basic reproduction number R_0, a key value for the system, is given by (Kermack and Mckendrick, 1927)

$$R_0 = \frac{\beta S}{\gamma}. \tag{6.2}$$

When $R_0 < 1$, the spread of disease could be controlled over the period; when $R_0 > 1$, the epidemic would spread to the whole population. Estimating R_0 will be the key to understand the pandemic, especially in the early stages. A better estimation for R_0 will help government and authorities make the right decision.

In the early stages of Covid-19, the number of susceptible people is much larger

than infected people, the derivative of S compared with the total S, $\frac{\mathtt{d}S}{S\mathtt{d}t}$ is very tiny. Thus S can be considered as a large constant over the study period and the SIR model could reduce to the two-dimensional IR model, with $Y_t = (I_t, R_t)$ where $Y_1(t) = I_t$ and $Y_2(t) = R_t$. We have the following relationship between the above IR model and the Lotka-Volterra model, which is given by the following transition changes:

$$P(Y_{t+\mathtt{d}t}|Y_t) = \begin{cases} \theta_1 y_1 \mathtt{d}t + o(\mathtt{d}t) & \text{if } Y_1(t+\mathtt{d}t) = y_1 + 1 \text{ and } Y_2(t+\mathtt{d}t) = y_2, \\ \theta_2 y_1 \mathtt{d}t + o(\mathtt{d}t) & \text{if } Y_1(t+\mathtt{d}t) = y_1 - 1 \text{ and } Y_2(t+\mathtt{d}t) = y_2 + 1, \end{cases}$$

and

$$P(Y_{t+\mathtt{d}t}|Y_t) = \ 1 - (\theta_1 y_1 + \theta_2 y_1)\mathtt{d}t + o(\mathtt{d}t) \quad \text{if } Y_1(t+\mathtt{d}t) = y_1 \text{ and } Y_2(t+\mathtt{d}t) = y_2.$$

We have $\theta_1 = \beta S - \gamma$ and $\theta_2 = \gamma$ then $R_0 = \frac{\theta_1+\theta_2}{\theta_2}$. This model is also known as Kermack-McKendrick model (Kermack and Mckendrick, 1927).

As we mentioned in the previous simulation section, the likelihood inference for such IR models or LV models are very challenging because the we only get discretely observed data points and the likelihood function is not readily available. We will apply our method to Covid19 data under such modelling framework.

6.2.2 Data sources

We consider the Covid19 data provided in

https://github.com/mrc-ide/COVID19_CFR_submission

In particular, we use the Singapore data from Jan 22nd to March 15th, including the daily confirmed new cases and the daily death/recovered number. The raw data used are plotted in Figure 6.7. In the period under consideration (54 days ending on 15th March 2020), there are 121 confirmed individuals and 105 removed ones (death

or cured). These observed data points are very small numbers comparing to the population (5,703,600), which satisfies the assumption in the IR model.

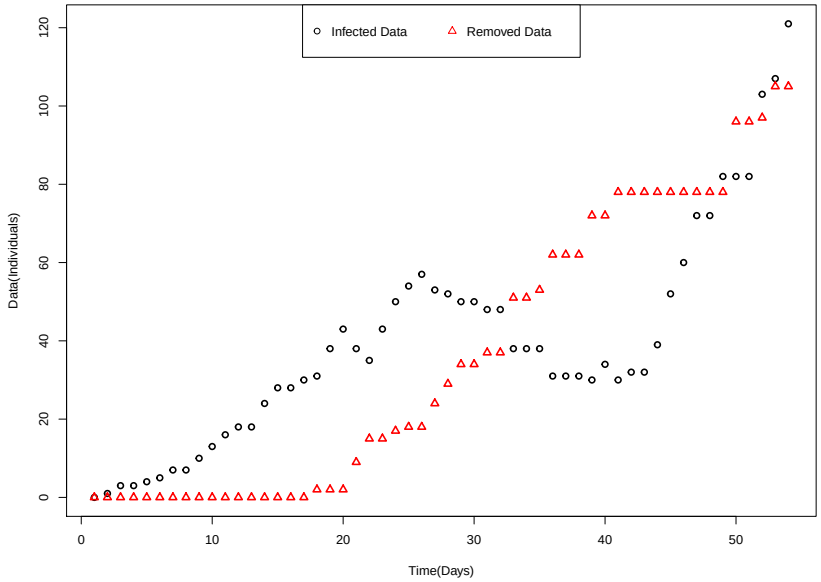

Figure 6.7: Covid-19 Data in Singapore.

6.2.3 Analysis Results

In the prior experiments, we can simulate samples from the modified IR model with an arbitrary pair of parameters. If we take 2 statistics in the model, we can find that the difference between estimate and true value is rather large. However, 3 statistics can significantly reduce the difference.

After that, it suggests that 3 statistics should be selected to construct the working likelihood, even though there are two unknown parameters (θ_1, θ_2) in the model. The same sample quantiles from the LV model are determined in the modified SIR model as well.

Here we take $M = 3$, $N_n = 100$, $q(N_n) = 16$ and $T = 44$ (from the 11th day to the 54th day) for the IR model. The initial state is $y_1(0) = 16$ and $y_2(0) = 0$ while starting point $\tilde{\boldsymbol{\theta}}_{\mathbf{0}}$ is sampled from two uniform distributions $\mathcal{U}[0.15, 0.2]$ and

$\mathcal{U}[0.05, 0.15]$.

In our gradient-decent algorithm, we also choose a starting point randomly. The estimation of parameters (θ_1, θ_2) is $(0.1812235, 0.09385114)$ which implies R_0 is 2.930967. As shown in Adam (2020), the estimation of R_0 before March 26th 2020 should be between 2.4 and 3.3.

Chapter 7

Discussion

From results of experiments, we have demonstrated that our proposed method performs no worse than other classic estimates for unknown parameters in these cases; primarily, it provides more stable and less biased estimates. Further, even a multivariate problem could be solved by multiplying uni-variable Hermite Polynomials Expansion if proper statistics could be selected with independence. In this way, the multi-dimensional problem could be dealt with straightforward. Note that, from a practical point of view, the implementation of our method relies on the requirement of using independent statistics. Although in theory, correlation statistics are still holding for our method based on the high-dimensional Hermite polynomial expansion theory (Szegö, 1959), practically it is prohibitive to use correlated statistics because the terms required in the Hermite expansion will increase exponentially, i.e. $q(N_n)$ will be too large to deal with. However, as the examples presented in the thesis, independent statistics can always be found by partition the data in a suitable way, even for time series data.

Our proposed method has the advantages of both ABC and MLE. The method, like ABC, requires simulated samples with given parameters. The difference is that instead of narrowing the distance between simulated samples and the original data in

ABC methods, our method uses these simulated samples to estimate the coefficient value $\widehat{\boldsymbol{C}}$ in the Hermite expansion to approximate the MLE. Thus, unlike qualitative asymptotic analysis in the ABC, the proposed method could provide quantitative proof on the convergence. At the same time, the proposed method has a higher universality, mainly when MLE could hardly deal with the case that there is no close form for the likelihood.

Our proposed method is helpful for both continuous functions, such as quantile and discrete problems, including LV models. Due to the restriction of a deadline, we introduce the solution of the two typical problems with experiments. However, the widespread use of the method would be extended to other fields as well. Image Identification is one of the examples. We could recognize each item from an image which builds the generative function between dependent variables and pixels in a figure. Even we could hardly work out the relationship between combinations of pixels and the figure of an item; the method seems possible to provide a likelihood approximation for statistical inference. As our method provide an approximation to a likelihood function instead of a single point from a function, the convolution of the likelihood function could not impact the efficiency in theory. Further, the link between the proposed method and complex analysis may be constructed to extend future theories' application.

I have to admit that our proposed method requires the initial state of parameters close to the true value; otherwise the resulting estimate would likely converge to a local extreme point rather than the true value. However, as the analytical calculating could provide the inference of error term, our proposed method could improve the estimate more precisely. In this case, I think the combination with other pilot experiments, such as Rejection ABC methods, may reduce the probability of a local extreme point problem.

Bibliography

D. Adam. Special report: The simulations driving the world's response to covid-19. *Nature*, 580(7803):316–318, 2020.

D. Allingham, R. A. King, and K. L. Mengersen. Bayesian estimation of quantile distributions. *Stat. Comput.*, 19(2):189–201, 2009.

H. Bateman. *Higher Transcendental Functions*, volume II. McGraw-Hill, New York, 1953.

S. S. Bonan and D. S. Clark. Estimates of the hermite and the freud polynomials. *J. Approx. Theory*, 63(2):210–224, 1990.

R. J. Boys, D. J. Wilkinson, and T. B. Kirkwood. Bayesian inference for a discretely observed stochastic kinetic model. *Stat. Comput.*, 18(2):125–135, 2008.

P. Bühlmann. Bootstraps for time series. *Statist. Sci.*, 17(1):52–72, 2002.

S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag, London, 2001.

D. R. Cox and N. Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737, 2004.

E. Del Barrio, E. Giné, and F. Utzet. Asymptotics for l_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted wasserstein distances. *Bernoulli*, 11(1):131–189, 2005.

B. D. Dennis. Introduction to the bootstrap world. *Statistical Science*, 18(2):168–174, 2003.

Y. Dodge. *The Oxford Dictionary of Statistical Terms*. Oxford University Press, 2003.

C. R. Doss, J. M. Flegal, G. L. Jones, and R. C. Neath. Markov chain monte carlo estimation of quantiles. *Electron. J. Stat.*, 8(2):2448–2478, 2014.

C. C. Drovandi and A. N. Pettitt. Likelihood-free bayesian estimation of multivariate quantile distributions. *Computational Statistics & Data Analysis*, 55(9):2541–2556, 2011.

C. C. Drovandi, A. N. Pettitt, and A. Lee. Bayesian indirect inference using a parametric auxiliary model. *Statist. Sci.*, 30(1):72–95, 2015.

B. Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1): 1–26, 1979.

P. Fearnhead and D. Prangle. Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 74(3):419–474, 2012.

N. M. Ferguson, D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunubá, and G. Cuomo-Dannenburg. Impact of non-pharmaceutical interventions (npis) to reduce covid-19 mortality and healthcare demand. *The 9th report from the WHO Collaborating Centre for Infectious Disease Modelling*, 2020. URL `https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-9`

R. Garcia, E. Renault, and D. Veredas. Estimation of stable distributions by indirect inference. *J. Econometrics*, 161(2):325–337, 2011.

ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2010.12.007. URL https://www.sciencedirect.com/science/article/pii/S0304407610002514.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81(25):2340–2361, 1977.

A. Golightly and D. J. Wilkinson. Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(3):781–788, 2005.

C. Gourieroux, A. Monfort, and E. Renault. Indirect inference. *J. Appl. Econometrics*, 8(S1):S85–S118, 1993.

M. A. Haynes. *Flexible distributions and statistical models in ranking and selection procedures with applications.* PhD thesis, Queensland University of Technology, 1998.

P. J. Heagerty and S. R. Lele. A composite likelihood approach to binary spatial data. *J. Amer. Statist. Assoc.*, (443):1099–1111, 1998.

M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In *International Workshop on Artificial Intelligence and Statistics*, pages 136–143. PMLR, 2005.

R. Henderson and S. Shimakura. A serially correlated gamma frailty model for longitudinal count data. *Biometrika*, 90(2):355–366, 2003.

G. Karabatsos and F. Leisen. An approximate likelihood perspective on abc methods. *Stat. Surv.*, 12:66–104, 2018.

W. Kermack and A. Mckendrick. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. A*, 115:700–721, 1927.

I. Krasikov. New bounds on the hermite polynomials. *arXiv preprint math/0401310*, 2004.

Y. LeCun, L. Bottou, G. Orr, and M. K.R. Efficient backprop. In G. Montavon, G. Orr, and M. K.R., editors, *Neural Networks: Tricks of the Trade*, pages 9–48. Springer, Berlin, Heidelberg, 2012.

H. Liang and P. Du. Maximum likelihood estimation in logistic regression models with a diverging number of covariates. *Electronic Journal of Statistics*, 6:1838–1846, 2012.

J. Liu and X. Yang. The convergence rate and asymptotic distribution of the bootstrap quantile variance estimator for importance sampling. *Adv. in Appl. Probab.*, 44(3):815–841, 2012.

J. H. McCulloch. Simple consistent estimators of stable distribution parameters. *Communications in Statistics - Simulation and Computation*, 15(4):1109–1136, 1986. doi: 10.1080/03610918608812563. URL `https://www.tandfonline.com/doi/abs/10.1080/03610918608812563`.

K. L. Mengersen, P. Pudlo, and C. P. Robert. Bayesian computation via empirical likelihood. *Proceedings of the National Academy of Sciences*, 110(4): 1321–1326, 2013. ISSN 0027-8424. doi: 10.1073/pnas.1208827110. URL `https://www.pnas.org/content/110/4/1321`.

J. D. Murray. *Mathematical biology: I. An introduction*, volume 17. Springer-Verlag, New York, 2002.

I. J. Myung. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1):90–100, 2003. ISSN 0022-2496. doi: https://doi.org/10.1016/S0022-2496(02)00028-7. URL `https://www.sciencedirect.com/science/article/pii/S0022249602000287`.

D. J. Nott and T. Rydén. Pairwise likelihood methods for inference in image models. *Biometrika*, 86(3):661–676, 1999.

M. Nunes and D. Prangle. abctools: An r package for tuning approximate bayesian computation analyses. *The R Journal*, 7(2):189–205, July 2015. ISSN 2073-4859.

A. B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 06 1988. ISSN 0006-3444. doi: 10.1093/biomet/75.2.237. URL `https://doi.org/10.1093/biomet/75.2.237`.

M. Park, W. Jitkrittum, and D. Sejdinovic. K2-abc: Approximate bayesian computation with kernel embeddings. *ArXiv*, abs/1502.02558, 2016.

E. Parner. A composite likelihood approach to multivariate survival data. *Scand. J. Statist.*, 28(2):295–302, 2001.

U. Picchini and R. Anderson. Approximate maximum likelihood estimation using data-cloning abc. *Comput. Statist. Data Anal.*, 105:166–183, 2017.

J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Mol. Biol. Evol.*, 16(12):1791–1798, 1999.

G. D. Rayner and H. L. MacGillivray. Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Stat. Comput.*, 12(1):57–75, 2002.

F. Rubio and A. M. Johansen. A simple approach to maximum intractable likelihood estimation. *Electron. J. Stat.*, 7:1632–1654, 2013.

A. Rényi. On the theory of order statistics. *Acta Mathematica Academiae Scientiarum Hungarica*, 4:191–231, 1953.

R. D. Schafer. *An Introduction to Nonassociative Algebras*. Elsevier, New York, 1966.

S. A. Sisson, Y. Fan, and M. A. Beaumont. Overview of abc. In *Handbook of approximate Bayesian computation*, pages 3–54. Chapman and Hall/CRC, 2018.

G. Szegö. *Orthogonal Polynomials*. American Mathematical Society, New York, revised edition, 1959.

T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*, 6(31):187–202, 2009.

M. A. Van Duijn, K. J. Gile, and M. S. Handcock. A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Soc. Netw.*, 31(1):52–62, 2009.

G. G. Walter. Properties of hermite series estimation of probability density. *Ann. Statist.*, 5(6):1258–1264, 1977.

Y. Wang, W. Yang, and S. Hu. The bahadur representation of sample quantiles for weakly dependent sequences. *Stochatic: An International Journal of Probability and Stochastic Processes*, pages 428–436, 2016.

S. N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104, 2010.

W. Zhu, J. M. Marin, and F. Leisen. A bootstrap likelihood approach to bayesian computation. *Aust. N. Z. J. Stat.*, 58(2):227–244, 2016.