STATISTICAL INFERENCE BASED ON CHARACTERISTIC FUNCTIONS FOR INTRACTABLE LIKELIHOOD PROBLEMS

BY

FAN YANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Professor Yuguo Chen, Advisor
Associate Professor Liming Feng, Co-advisor
Assistant Professor Alexandra Chronopoulou
Professor Xiaofeng Shao

Abstract

This dissertation is devoted to statistical inference based on characteristic functions. For some popular stochastic processes (e.g., Lévy processes, Lévy driven Ornstein−Uhlenbeck processes), the transition density may not be available. However, the (conditional) characteristic function is sometimes known. We study various statistical inference methods for fitting those processes with implicit characteristic functions.

In the first part, an efficient sampling method based on Bayesian empirical likelihood is developed. The method involves pseudo-marginal Markov chain Monte Carlo with temperature and is shown to be effective for Lévy processes. In the second part and third part, we study maximum likelihood methods and empirical characteristic function estimation based on characteristic functions. We find the analyticity of the characteristic function can make efficient implementations of both methods possible, guaranteeing asymptotic properties as well. We also find, for certain models, very large samples might be needed to accurately identify the true parameters. Numerical results show the appealingness of some infinite activity models. In the last part, this dissertation includes my another project, which is about *truth discovery* in data mining. A dynamic model is developed to discover the truth between information sources across time. Experiments on real-world applications demonstrate its advantages over previous approaches.

Acknowledgments

I would like to thank Professor Yuguo Chen and Professor Liming Feng for their guidance and support throughout my graduate study at University of Illinois at Urbana-Champaign. This work would not have been possible without their help.

I also want to thank my committee members, Xiaofeng Shao and Alexandra Chronopoulou, who provide valuable suggestions and advices.

And finally, thanks to my parents, my girlfriend and numerous friends who always be with me for all joyful moments and difficult time.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this dissertation, we study statistical inference based on characteristic functions for intractable likelihood problem. For some popular stochastic processes (e.g., Lévy processes, Lévy driven Ornstein−Uhlenbeck processes), the transition density may not be available. However, the (conditional) characteristic function is sometimes known. We study various statistical inference methods for fitting those processes with implicit characteristic functions.

In Chapter 2, we mainly focus on Bayesian empirical likelihood inference based on characteristic functions. We construct a maximum empirical posterior estimator with asymptotic properties. We utilize pseudo-marginal Markov chain Monte Carlo with temperature method to make efficient implementation of Bayesian empirical likelihood inference possible. The numerical study confirms the effectiveness of our maximum empirical posterior estimator.

In Chapter 3 and Chapter 4, we study maximum likelihood methods and empirical characteristic function estimation methods based on characteristic functions for Lévy processes that are commonly used in finance. In both chapters, we utilize the analyticity of the characteristic functions of such Lévy processes.

For maximum likelihood methods, the transition probability densities can be computed very fast and accurately. This makes efficient implementations of maximum likelihood inference possible. We provide regularity conditions that guarantee consistency, asymptotic normality and efficiency of maximum likelihood estimation and validity of likelihood ratio tests. Our approach extends to Ornstein-Uhlenbeck or Lévy based processes with explicit and analytic conditional characteristic functions. Simulation studies show the effectiveness of our method. While parameter estimation is relatively easier for some Lévy processes, very large samples might be needed for others. This is illustrated by some popular Lévy models, including the CGMY model, where the log-likelihood surface is saddle-shaped with relatively flat areas. Finally, fitting of equity returns shows the appealingness of some infinite activity Lévy models.

For empirical characteristic function (ECF) estimation methods, the analyticity of the characteristic functions is also helpful to provide an efficient approximation of the estimation target function. Specifically, empirical characteristic function estimation is a generalized moment match method to match the empirical

characteristic function with the model's characteristic function. To match them, the usual method is to minimize an integration of the distance between empirical characteristic function and the model's characteristic function, which is computationally intensive. We utilize the analyticity of the characteristic functions (for certain stochastic processes) to show that the integration involved in ECF estimation can be computed very fast and accurately. We also provide regularity conditions to show the estimates based on our implementation have consistency and asymptotic normality properties. The simulation study shows the effectiveness and efficiency of our methods.

This dissertation also includes my another project in Chapter 5, which is a joint project with Shi Zhi, Zheyi Zhu, Qi Li, Zhaoran Wang and Jiawei Han. I am the equally contributed first author with Shi Zhi. We build a state space model for evolving truth discovery. *Truth discovery* is an important topic in data mining. Untrustworthy information is ubiquitous in big data, which gives rise to the challenge of *truth discovery*. The goal of truth discovery is to distill the most credible information from noisy but redundant sources. Despite recent progress, the time-varying structure of truth discovery problems remains less explored, e.g., the latent truth may evolve over time and have a temporal correlation. In this project, we propose a general framework named *EvolvT* to explicitly model the dynamics of truth evolution. At the core of such a framework is an adaptation of hidden Markov model to characterize the trustworthiness of different sources. Based on Kalman filtering and smoothing techniques, we establish an expectation-maximization (EM) algorithm with strong theoretical guarantees. In comparison with existing approaches, our framework captures source dependency between information sources across time, and furthermore, automatically allows for missing data. Experiments on real-world applications demonstrate its advantages over the state-of-the-art truth discovery approaches.

In Chapter 6, we discuss the potential extension of our proposed estimation methods to more general Markov processes.

Chapter 2

An MCMC approach for empirical likelihood inference based on characteristic functions

2.1 Introduction

Common statistical procedures such as maximum likelihood can provide efficient estimates under mild regularity conditions if the likelihood function is available. However, a number of difficult problems might arise for certain models of which the likelihood is not tractable. In this chapter, we mainly focus on the case that characteristic functions have analytic forms rather than density functions. One potential application is Lévy processes because their characteristic functions are more likely to be available than densities due to Lévy-Khintchine representation.

Statistical inference based on characteristic functions includes several methods. The first method is empirical characteristic function (ECF) estimation investigated by Paulson et al. (1975); Feuerverger and Mureika (1977); Feuerverger and McDunnough (1981*a*). The basic idea of this method is to match the characteristic function implied by the model and the empirical characteristic function obtained from data. A review can be found from Yu (2004). A Markov chain Monte Carlo approach developed by Chernozhukov and Hong (2003) is utilized for ECF estimation with application to fit stock prices by Lévy processes in Grynkiv (2010). We propose an approximation method to conduct ECF estimation efficiently in Chapter 4. Another choice to do statistical inference based on characteristic function is to obtain densities by Fourier inversion transform. We will discuss it in Chapter 3.

In this work, we focus on the statistical inference method based on empirical likelihood. Kunitomo and Owada (2006) and Chan et al. (2009) construct maximum empirical likelihood estimator (MELE) based on characteristic functions. The basic idea is to find optimal parameters which maximize the empirical likelihood constructed based on characteristic functions. Empirical likelihood, proposed by Owen (1988), sometimes can be regarded as an alternative to parametric likelihood especially when model's density function is not available (DiCiccio et al. (1989)). Moreover, Qin and Lawless (1994) show that maximum empirical likelihood estimates (MELE) are asymptotically normally distributed under certain conditions.

When it comes to the implementations of empirical likelihood inference, generally, there are two common

ways to incorporate characteristic functions into empirical likelihood. Kunitomo and Owada (2006) chooses a finite number of grid points $u_1, u_2, \ldots, u_m$, and use estimation equations based on characteristic functions $\phi(u_i) = E(\exp(iu_iX))$ with $i = 1, 2, \ldots, m$, to construct empirical likelihood, so as to obtain MELE. However, efficiency cannot be reached based on finite points $\{u_i\}$. To determine how dense and how many grids points $\{u_i\}_{i=1}^m$ are enough to get valid estimates is still an unanswered question.

The alternative approach is to construct an empirical likelihood $L(\theta|u)$ based on characteristic function $\phi(u) = E(\exp(iuX))$ conditionally on a fixed u. Then, construct integrated log-empirical likelihood $T(\theta) = \int \log(L(\theta|u))dG(u)$, where $G(u)$ represents the distribution of u. This method can be regarded as the continuous version of empirical likelihood based on characteristic function, and the MELE with a judiciously chosen $G(u)$ can reach full maximum likelihood efficiency (See Chan et al. (2009)). But this method comes with a big computing burden. To be specific, we have three computing issues. Firstly, the function inside the integral $T(\theta)$ includes empirical likelihood $L(\theta|u)$, which requires one optimization procedure. Secondly, $T(\theta) = \int \log(L(\theta|u))dG(u)$ is an integral of $\log(L(\theta|u))$, which requires a rigorous integration procedure. Thirdly, to obtain MELE, we need to find the values of our parameters to maximize $T(\theta)$, avoiding several local maximums. All of them are computationally extensive. Especially, the second issue and third issue are more challenging because there are already many works (for example, Owen (1990) and Wu (2004)) about the first issue to improve the efficiency of calculating empirical likelihood.

In this chapter, we mainly focus on the sampling based optimization technique and propose a pseudo-marginal Markov chain Monte Carlo (MCMC) based simulated annealing algorithm (which is called integrated empirical likelihood sampler) to deal with the second and the third issue. We resolve the second issue by setting a different target function as integrated empirical likelihood $T(\theta) = \int (L(\theta|u))dG(u)$ and sample the parameter θ from it via pseudo-marginal MCMC. During this procedure, we can sample from $T(\theta)$ without evaluating the integral form $T(\theta)$. What's more, estimates based on new integrated empirical likelihood here still keep common asymptotic properties (i.e. asymptotic normality). To resolve the third issue, we use the idea in the simulated annealing which sample from $T(\theta)^{1/t_n}$ with a sequence of decreasing temperatures t_n. Under certain regularity condition, it can be shown that samples from $T(\theta)^{1/t_n}$ will converge to the global maximum of our target $T(\theta)$ when $t_n \to 0$.

All in all, in this chapter, we construct a new continuous version of empirical likelihood inference and prove the asymptotic properties of our estimates based on this new version. Noticing that, one of the regularity conditions for asymptotic properties is not easy to verify. We provide a simple equivalent condition to it, which is easy to understand and test. Based on our integrated empirical likelihood, pseudo-marginal MCMC based samplers are proposed to obtain the parameter estimates.

We organize this chapter as follows. In Chapter 2.2, we construct our Bayesian empirical likelihood and maximum empirical posterior estimates (MEPE). In Chapter 2.3, we present several asymptotic properties of our MEPE and Bayesian empirical likelihood. We also discuss regularity conditions and provide two equivalent regularity conditions. In Chapter 2.4, we propose integrated empirical likelihood sampler to estimate MEPE. In Chapter 2.5 and Chapter 2.6, simulation and case studies are performed to demonstrate its effectiveness. Chapter 2.7 is the conclusion of this study. All proofs are in the Appendices.

2.2 Bayesian empirical likelihood based on characteristic functions

In this section, we construct Bayesian empirical likelihood based on characteristic functions. Suppose we have n i.i.d. random variables $X = \{X_1, \ldots, X_n\}$ from d variate distribution F with characteristic function $\Phi(u;\theta_0)$ for $u \in R^d$ where θ_0 is the p dimensional true parameter. The real part and imaginary part of the characteristic function are $\phi^R(u;\theta_0) = \Re(E(e^{iu^T X_i}))$ and $\phi^{Im}(u;\theta_0) = \Im(E(e^{iu^T X_i}))$. Assume θ is a p dimensional parameter associated with F within compact parameter space $\Theta \subset R^p$. Suppose that the information of parameter θ can be summarized in the form of infinite moment conditions $g(u, X_i;\theta)$ based on characteristic functions $\phi(u;\theta)$, such that $E\{g(u, X_i, \theta_0)\} = 0$. $g(u, X_i, \theta)$ is defined as

$$g(u, X_i;\theta) := (\cos(u^T X_i) - \phi^R(u;\theta),\ \sin(u^T X_i) - \phi^I(u,\theta))^T,$$

and θ_0 is uniquely determined above.

Similar to Chan et al. (2009); Chen et al. (2013), the profile empirical likelihood proposed in Qin and Lawless (1994) with given u is:

$$L_n(u;\theta) = \max\{\prod_{j=1}^{n} w_j |\ w_j \geq 0, \sum_{j=1}^{n} w_j = 1, \sum_{j=1}^{n} w_j \cos(u^T X_j) = \phi^R(u;\theta),\ \sum_{j=1}^{n} w_j \sin(u^T X_j) = \phi^I(u;\theta)\}. \tag{2.1}$$

By solving it using Lagrange multiplier , we have the profile likelihood $L_n(u;\theta) = \prod_{i=1}^{n} w_i(u;\theta)$ and $w_i(u;\theta) = \frac{1}{n(1+\lambda_n(u;\theta)^T g(u,X_i;\theta))}$. $\lambda_n(u;\theta)$ is the Lagrange multiplier satisfies

$$\sum_{i=1}^{n} \frac{g(u, X_i;\theta)}{1 + \lambda_n(u;\theta)^T g(u, X_i;\theta)} = 0, \tag{2.2}$$

given $u \in R^d$. Then, the integrated empirical likelihood $T_n(\theta)$ is defined as

$$T_n(\theta) = \int_S L_n(u;\theta) dG(u), \tag{2.3}$$

and $G(u)$ is either a given discrete distribution or a smooth distribution function of u with support on compact set $S \subset R^d$ satisfying regularity conditions in Chapter 2.3.

With a prior specification $p_0(\theta)$ on the parameter θ, we have the posterior density

$$p(\theta|X) \propto p_0(\theta) T_n(\theta). \tag{2.4}$$

We call $p(\theta|X)$ the posterior distribution based on integrated empirical likelihood approach. Then, we define maximum empirical posterior estimator (MEPE) $\hat{\theta}$ as

$$\hat{\theta} = \arg\max_{\theta} p(\theta|X). \tag{2.5}$$

In this chapter, we mainly investigate both inference and computing of our proposed posterior distribution in Equation (2.4). We construct asymptotic properties of posterior density and MEPE. We also propose a pseudo-marginal MCMC algorithm to estimate the MEPE.

Remark 1. We add a prior distribution in our setting of Equation 2.4. Prior information is important to identify parameters, which is also indicated in Johannes and Polson (2003). In our model setting, Bayesian setting will lead to Bayesian empirical likelihood. Lazar (2003) discusses the validity of this procedure first. Grendár and Judge (2009) investigates the asymptotic equivalence of Bayesian maximum posteriori estimator and empirical likelihood. Also, Bayesian empirical likelihood with Monte Carlo has been applied to several areas such as population genetics (Mengersen et al. (2013)), quantile regression (Yang and He (2012)). Thus, it might be reasonable to consider priors in our work. When priors are flat, it is obvious to see that Bayesian point estimation will be consistent to the point estimation under Frequentest framework.

2.3 Asymptotic properties of the posterior

We write down several regularity conditions below.

A.1 $E\{g(u, X_i; \theta_0) g^T(u, X_i; \theta_0)\}$ is positive definite for $u \in S$ with arbitrary fixed i within $\{1, 2, 3, \ldots, n\}$.

A.2 $\frac{\partial}{\partial \theta} g(u, x; \theta)$ is continuous with θ in Ω_0, the neighborhood of θ_0 for $u \in S$ and $x \in R^d$.

A.3 $\sup_{\theta\in\Omega_0} ||\frac{\partial}{\partial\theta}g(u,x;\theta)|| \leq H(u,x)$, where $H(u,x)$ is a function satisfying $\int_S \int_{R^d} H(u,x)dF(x)dG(u) < \infty$.

A.4 $\int_S (E(\frac{\partial}{\partial\theta}g(u,x;\theta_0))^T (E\{g(u,X_i;\theta_0)g^T(u,X_i;\theta_0)\})^{-1}(E(\frac{\partial}{\partial\theta}g(u,x;\theta_0))))dG(u)$ is positive definite.

A.5 $\frac{\partial^2}{\partial\theta\partial\theta^T}g(u,x;\theta)$ is continuous in θ for $\theta\in\Omega_0, t\in S, x\in R^d$.

A.6 $\sup_\theta \in \Omega_0 ||\frac{\partial^2}{\partial\theta\partial\theta^T}g(u,x;\theta)|| \leq H(u,x)$, where $H(u,x)$ is given in A.3.

A.7 Log-prior $\log\{p_0(\theta)\}$ has bounded first derivative in Ω_0, which is the neighborhood of θ_0.

A.8 Let $J(\hat{\theta}_n) = -\frac{\partial^2}{\partial\theta\theta^T}\log T(\theta)|_{\theta=\hat{\theta}_n}$ and assume it is positive definite.

2.3.1 Asymptotic properties of our defined MEPE

The following theorems in Chapter 2.3.1 shows the asymptotic normality of MEPE estimates.

Theorem 1 (Consistency). *Under regularity conditions A.1-A.4 and A.7, our posterior $p(\theta|X)$ attains its maximum $\hat{\theta}_n$ in a $n^{-\frac{1}{3}}$ neighborhood of θ_0, $||\hat{\theta}-\theta_0|| \leq n^{-\frac{1}{3}}$, almost surely satisfying*

$$\begin{cases} Q_{1n}(u;\hat{\theta}_n,\lambda_n(u;\hat{\theta}_n)) = 0 \\ \int_S \{p_0(\hat{\theta}_n)L_n(u;\hat{\theta}_n)Q_{2n}(u;\hat{\theta}_n,\lambda_n(u;\hat{\theta}_n)) - \frac{1}{n}\frac{\partial p_0(\hat{\theta}_n)}{\partial\theta}L_n(u;\hat{\theta}_n)\}dG(u) = 0, \end{cases} \tag{2.6}$$

where

$$\begin{cases} Q_{1n}(u;\theta,\lambda) = \frac{1}{n}\sum_{i=1}^n \frac{1}{1+\lambda^T g(u,X_i;\theta)}g(u,X_i;\theta) \\ Q_{2n}(u;\theta,\lambda) = \frac{1}{n}\sum_{i=1}^n \frac{1}{1+\lambda^T g(u,X_i;\theta)}\frac{\partial g(u,X_i;\theta)}{\partial\theta}\lambda. \end{cases} \tag{2.7}$$

As for the asymptotic normality property, we will obtain the exactly same asymptotic covariance matrix as the one in Chan et al. (2009) with the similarly proof.

Theorem 2 (Asymptotic normality). *Under regularity conditions A.1-A.7, for $\hat{\theta}_n$ given in Theorem 1, when $n\to\infty$*

$$\begin{aligned} \sqrt{n}\{\hat{\theta}_n-\theta_0\} &= -\{\int_S s_{21}(u)s_{11}^{-1}(u)s_{12}(u)dG(u)\}^{-1} \\ &\quad\times\{\int_S s_{21}(u)s_{11}^{-1}(u)\sqrt{n}Q_{1n}(u;\theta_0 0)dG(u)\} + o_p(1) \\ &\xrightarrow{d} N(0,\Sigma), \end{aligned} \tag{2.8}$$

where

$$
\begin{aligned}
s_{11}(u) =& - E\{g(u, X_1; \theta_0) g^T(u, X_1; \theta_0)\},\\
s_{12}(u) =& s_{21}^T = E\{\frac{\partial}{\partial \theta} g(u, X_1; \theta_0)\},\\
\Sigma =& \{\int_S s_{21}(u) s_{11}^{-1}(u) s_{12}(u) dG(u)\}^{-1} \{\int_S \int_S s_{21}(u_1) s_{11}^{-1}(u_1) E\{g(u_1, X_1; \theta_0) g^T(u_2, X_1; \theta_0)\} s_{11}^{-1}(u_2)\\
& \times s_{12}(u_2) dG_(u_1) dG(u_2)\} \times \{\int_S s_{21}(u) s_{11}^{-1}(u) s_{12}(u) dG(u)\}^{-1}.
\end{aligned}
$$

To reach asymptotic efficiency of MEPE, we define the $G(u)$ in (2.3) as same as the one mentioned in Section 2.1 of Chan et al. (2009) due to the same form of asymptotic covariance matrix in Theorem 2. For the testing, since our estimator has the same rate of convergence as the MELE proposed in Chan et al. (2009), our estimates MEPE can be employed in the empirical likelihood test in Chan et al. (2009). Specifically, it is stated in Corollary 2.1.

Corollary 2.1. *Suppose we want to test a model with a characteristic function $\phi(u; \theta)$ and we have hypotheses:*

$$H_0 : \phi(u) \in \{\phi(u; \theta) : \exists \theta \in \Omega\},$$

against

$$H_a : \phi(u) \notin \{\phi(u; \theta) : \forall \theta \in \Omega\},$$

where $\Omega \subset R^p$ is a given set. Then, we have test statistics:

$$\mathcal{T}(\hat{\theta}) = \int_S \log(L_n(u; \hat{\theta})) dG(u),$$

where $\hat{\theta}$ is the MEPE defined in Equation (2.5). Under regularity conditions A.1-A.7, as $n \to \infty$, we have

$$|\mathcal{T}(\hat{\theta} - \{-W_2 + W_1^T \{\int_{-a}^{a} s_{21}(u) s_{11}^{-1}(u) s_{12}(u) dG(u)\}^{-1} W_1\})| = o_p(1),$$

where

$$W_1 = \int_{-\infty}^{\infty} \{\int_{-a}^{a} s_{21}(u) s_{11}^{-1}(u) (\cos(ux), \sin(ux))^T dG(u)\} dB_n(F(x)),$$

$$W_2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{\int_{-a}^{a} (\cos(ux), \sin(ux)) s_{11}^{-1}(u) (\cos(uy), \sin(uy))^T dG(u)\} dB_n(F(x)) dB_n(F(y)),$$

and $\{B_n(y) : 0 \le y \le 1\}$ is a sequence of Brownian bridges.

Considering the limit distribution of test statistics $\mathcal{T}(\hat{\theta})$ in Corollary 2.1 is complicated, following the

suggestions in Chan et al. (2009), we also recommend performing parametric bootstrap to approximate the limit distribution of this test statistics.

2.3.2 Bayesian properties of posterior based on empirical likelihood

In this section, we discuss the asymptotic normality of Bayesian posterior based on empirical likelihood. Lazar (2003) discusses the possibility to use empirical likelihood in Bayesian inference. Inspired by the Theorem 1 in Lazar (2003), we have Bayesian asymptotically normality theorem below.

Theorem 3 (Posterior local normality)**.** *$\hat{\theta}_n$ is MEPE given in Theorem 1, under regularity conditions A.1-A.8, for $\{\theta : ||\theta - \theta_0|| = O(n^{-\frac{1}{2}})\}$, the posterior distribution of θ (in (2.4)) has the density*

$$p(\theta|X) \propto \exp\{-\frac{1}{2}(\theta - \hat{\theta}_n)^T J(\hat{\theta}_n)(\theta - \hat{\theta}_n) + o_p(1),\}$$

and if $J(\hat{\theta}_n)$ is positive definite,

$$J(\hat{\theta}_n)^{\frac{1}{2}}(\theta - \hat{\theta}_n) \xrightarrow{d} N(0, I),$$

where $\hat{\theta}_n$ is MEPE and

$$J(\hat{\theta}_n) = -\frac{\partial^2}{\partial\theta\theta^T} \log p(\theta|X)|_{\theta=\hat{\theta}_n}.$$

The approximated normality stated in Theorem 3 makes MCMC techniques possible to draw samples around MEPE from our posterior $p(\theta|X)$. However, this theorem (Theorem 3) only holds in a neighborhood of true parameter θ_0. That is, the posterior may not be approximated normally distributed over the whole parameter space Θ. One possibility is that the distribution implied by the sample you draw from posterior might be skewed. Then, the sample mean might not be a good estimator of our MEPE. In addition, more than one local maxima of $p(\theta|X)$ might appear in the place not around the true parameters. This problem might be even serious when the sample size is not big enough. Thus, we decide to borrow tempering idea from simulated annealing algorithm to obtain the estimates of MEPE accurately rather than sampling from our posterior $p(\theta|X)$ directly. This algorithm will be introduced in Chapter 2.4.

2.3.3 Discussion of regularity conditions

The regularity conditions in section 2.3 are standard compared with Qin and Lawless (1994) and Chan et al. (2009) except A.4. To be more specific, we find that the rank of $\frac{\partial}{\partial\theta}g(u, x; \theta_0)$ is p (the dimension of parameter space). The rank of $(E\{g(u, X_i; \theta_0)g^T(u, X_i; \theta_0)\})$ is 2 (dimension of $g(u, X_i, \theta_0)$ including one

real part and one imaginary part). Then, from linear algebra, the rank of

$$E(\frac{\partial}{\partial\theta}g(u,x;\theta_0))^T(E\{g(u,X_i;\theta_0)g^T(u,X_i;\theta_0)\})^{-1}(E(\frac{\partial}{\partial\theta}g(u,x;\theta_0))) \tag{2.9}$$

is 2 which could be less than p. Thus, (2.9) is not positive definite when $p > 2$. But interestingly, after integration,

$$\int_S(E(\frac{\partial}{\partial\theta}g(u,x;\theta_0))^T(E\{g(u,X_i;\theta_0)g^T(u,X_i;\theta_0)\})^{-1}(E(\frac{\partial}{\partial\theta}g(u,x;\theta_0))))dG(u)$$

could be positive definite (this is our regularity condition A.4). In this section, we discuss this phenomenon and investigate the equivalent condition to A.4. Hopefully, our provided regularity conditions alternative to A.4 in Theorem 4 and Theorem 5 can help us easily to check regularity condition A.4 considering the integral form in A.4 is usually complicated.

Here, we discuss $G(u)$ in the smooth distribution case or discrete distribution case separately.

- When $G(u)$ is a smooth distribution:

When $G(u)$ is a smooth distribution, obviously we can conclude that the support of $G(u)$ contains open sets. Then, we have the following theorem:

Theorem 4. *If the support of distribution $G(u)$, S, contains open sets, there exists an open set $I \subset S$ so that*

$$\int_I(E(\frac{\partial}{\partial\theta}g(u,x;\theta_0))^T(E\{g(u,X_i;\theta_0)g^T(u,X_i;\theta_0)\})^{-1}(E(\frac{\partial}{\partial\theta}g(u,x;\theta_0))))dG(u) \tag{2.10}$$

is positive definite if and only if there is no non-zero constant β (not a function of u) satisfying $E(\frac{\partial}{\partial\theta}g(u,x;\theta_0))^T(E\{g(u,X_i;\theta_0)g^T(u,X_i;\theta_0)\})^{-1}(E(\frac{\partial}{\partial\theta}g(u,x;\theta_0)))\beta = 0$ for $\forall u \in I$.

Because the open set I is the subset of S, whenever (2.10) is positive definite, A.4 will hold. Then, we have the following corollary:

Corollary 4.1. *If there exists an open set I so that there is no non-zero constant β(not a function of u) satisfying*

$$E(\frac{\partial}{\partial\theta}g(u,x;\theta_0))^T(E\{g(u,X_i;\theta_0)g^T(u,X_i;\theta_0)\})^{-1}(E(\frac{\partial}{\partial\theta}g(u,x;\theta_0)))\beta = 0$$

for $\forall u \in I$, then A.4 holds.

- When $G(u)$ is a discrete distribution:

When $G(u)$ is a discrete distribution, obviously we can conclude that the support of $G(u)$ contains countable points. Then, we have the following theorem:

Theorem 5. *If the support of distribution $G(u)$, S, is a set of countable points, then, there exists $G(u)$ to make regularity condition A.4 hold, which is equivalent to the condition that there is no non-zero constant β (not a function of u) so that*

$$E(\frac{\partial}{\partial\theta}g(u,x;\theta_0))^T(E\{g(u,X_i;\theta_0)g^T(u,X_i;\theta_0)\})^{-1}(E(\frac{\partial}{\partial\theta}g(u,x;\theta_0)))G(u)\beta = 0$$

for $\forall u \in S$. Moreover, based on equivalent regularity condition, we have a fixed way to construct or check the support of $G(u)$ with $n+1-m$ point masses (See lemma 24.2 for the way of the support construction by setting $A(t)$ to be (2.9)). $G(u)$ with point masses less than $n+1-m$ might be possible depending on the structure of matrix (2.9).

Remark 2. Theorem 4 and Theorem 5 provide us a way to check regularity condition A.4 without integral. To check that there is no non-zero constant β(not a function of u) so that $E(\frac{\partial}{\partial\theta}g(u,x;\theta_0))^T(E\{g(u,X_i;\theta_0)g^T(u,X_i;\theta_0)\})^{-1}(E(\frac{\partial}{\partial\theta}g(u,x;\theta_0)))\beta = 0$ for $\forall u \in S$, we can assume β exists and check the coefficients of all different terms in it including u are zero or not. This is relatively easy compared with checking A.4.

2.4 Sampling from posterior distribution

In this section, we mainly introduce the algorithm to obtain MEPE in (2.5). Specifically, we have empirical posterior

$$p(\theta|X) = \int_S L_n(u;\theta)p_0(\theta)dG(u), \tag{2.11}$$

which is defined in (2.3) and (2.4). MEPE is the $\hat{\theta}$ maximizing empirical posterior $p(\theta|X)$. This indicates a nonlinear optimization is required. In addition, $p(\theta|X)$ might be multi-modal. An optimization algorithm avoiding samples getting trapped in local maxima is required. Moreover, $p(\theta|X)$ is an integration without analytic form. Maximizing it directly will introduce a huge computational burden. Computation of standard error of MEPE requiring numerical derivatives of $p(\theta|X)$ will further aggravate computation difficulties. Instead of performing integral approximation technique before optimization, we use Markov chain Monte Carlo based sampling idea to estimate MEPE, which alleviates the issues mentioned above.

In this section, we propose a so called integrated empirical likelihood sampler based on pseudo-marginal MCMC and simulated annealing. Simulated annealing is a sampling based optimization algorithm. Its key

feature is to generate samples from the targets with different temperature so as to allow samples escape local optima in hopes of finding a global optimum. Specially, our posterior with temperature t is

$$p_t(\theta|X) = p(\theta|X)^{\frac{1}{t}} = (\int_S L_n(u;\theta)p_0(\theta)dG(u))^{\frac{1}{t}}. \tag{2.12}$$

Under the regularity conditions from Hwang (1980), it can be showed that the sequence of distributions $p_t(\theta|X)$ concentrates upon the MEPE when $t \to 0$. Furthermore, simulated annealing generates Markov Chain through MCMC technique following the distribution $p_t(\theta|X)$ with different temperature t. With detailed control of this generation procedure, the generated Markov chain will converge to the MEPE. In our case, $p_t(\theta|X)$ is a continuous function of θ. There are several papers presenting convergence properties and correspondent regularity conditions of simulated annealing global optimization for continuous functions including Locatelli (2000) and Yang (2000). Algorithm 1 is the simulated annealing algorithm to obtain MEPE. With fixed temperature, the acceptance rate $h(\theta, \vartheta)$ is from Metropolis hastings algorithm.

Initialize $\theta^0 = (\theta_1^0, \theta_2^0, \ldots, \theta_k^0)$ and temperature t_0;
Set iteration $j = 0$.;
while *Stopping criterion not satisfied* **do**
  Sample a random parameter vector ϑ^{j+1} from transition kernel $q(\theta \to \vartheta)$;
  Sample v from uniform distribution U(0,1);
  Calculate acceptance rate $h(\vartheta^{j+1}, \theta^j) = \min\{1, p_{t_j}(\vartheta^{j+1}|X)/p_{t_j}(\theta^j|X)\}$, where t_j is the temperature at iteration j. $p_t(\theta|X)$ is defined in (2.12);
  if $v \leq h(\vartheta^{j+1}, \theta^j)$ **then**
    Set $\theta^{j+1} = \vartheta^{j+1}$
  else
    Set $\theta^{j+1} = \theta^j$;
  end
  set j=j+1;
end

Algorithm 1: Simulated annealing

Simulated annealing algorithm listed above might help us to obtain the global maximum of $p(\theta|X)$. However, it leaves us a computation problem. That is, the acceptance rate $h(\theta, \vartheta)$ in algorithm 1 has an integration form which is hard to calculate.

One way to solve this problem is to change our target from $p(\theta|X)$ to

$$p_t(\theta, u) = p_0(\theta)\prod_{i=1}^{1/t} L(\theta|u_i)g(u_i), \tag{2.13}$$

where t is the temperature and $1/t$ is assumed to be an integer. $g(u)$ is the density function of u corresponding to its CDF $G(u)$. This idea is called State Augmentation for Marginal Estimation (SAME) from Doucet

et al. (2002) and this idea also appears in Jacquier et al. (2007).

We can find that the marginal distribution of (2.13) is $p_t(\theta|X)$ in (2.12). Then, simulating θ following $p_t(\theta|X)$ is equivalent to simulating θ and u separately from $p_t(\theta, u)$ by Gibbs sampling and getting rid of generated u. Similar to simulated annealing idea, by carefully designing sampling procedure and temperature t updating, simulated θ is supposed to coverage to the MEPE. Extension of $1/t$ to any real number in SAME is from Johansen et al. (2008).

SAME idea can avoid us calculating integral required by simulated annealing. However, we need to sample extra u from $p_t(u|\theta)$ via Gibbs sampling. Furthermore, with different u, we expect empirical likelihood $L(\theta|u)$ is different because different u implies different moment conditions inside (2.1). Therefore, a strong correlation between u and θ with respect to SAME idea target $p_t(\theta, u)$ might exist and sampling efficiency from $p_t(\theta, u)$ via Gibbs sampling might be a problem. Considering that we actually don't need to estimate u in our case, we adopt another sampling idea which is called pseudo-marginal MCMC to resolve the computation problem left by simulated annealing.

Pseudo marginal MCMC is first used by Beaumont (2003) and established formally by Andrieu and Roberts (2009). It can sample θ from intractable target function whenever unbiased estimates of the target function are available. One MCMC without likelihood algorithm proposed by Marjoram et al. (2003) can be regarded as a typical example of this algorithm. In our case, we replace acceptance rate from Metropolis hastings by the one based on pseudo-marginal MCMC to avoid the integral form of acceptance rate $h(\theta, \vartheta)$ in simulated annealing algorithm 1. Specifically, we list our pseudo-marginal MCMC with simulated annealing sampler in Algorithm 2 to show its advantage over standard simulated annealing algorithm 1. We name our pseudo-marginal MCMC with simulated annealing sampler in Algorithm 2 as integrated empirical likelihood sampler.

Initialize $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_k^{(0)})$;

`// Initialize unbiased estimates of` $p_{t_0}(\theta^{(0)}|X)$ `in the following steps`

for *integer r in set* $\{1, \ldots, 1/t_0\}$ **do**

Sample $u_1^{(r)}, \ldots, u_{n_0}^{(r)}$ following i.i.d. G(u) Obtain rth unbiased estimates of $p(\theta^{(0)}|X)$ in (2.11): $\hat{p}^{(r)}(\theta^{(0)}|X) = \frac{1}{n_0}\sum_{w=1}^{n_0} L_n(\theta^{(0)}; u_w^{(r)})$

end

Calculate unbiased estimates of $p_{t_0}(\theta|X)$ defined in (2.12): $\hat{p}_{t_0}(\theta^{(0)}|X) = \prod_{r=1}^{1/t_0} \hat{p}^{(r)}(\theta^{(0)}|X)$;

Initialize temperature pattern $t_0, t_1, \ldots, t_i, \ldots$, and for each i, $\frac{1}{t_i}$ is a integer.;

Set iteration $j = 0$;

while *Stopping criterion not satisfied* **do**

Sample a random parameter vector $\vartheta^{(j+1)}$ from transition kernel $q(\theta \to \vartheta)$;

`// Obtain unbiased estimates of` $p_{t_j}(\vartheta^{(j+1)}|X)$ `in the following steps`

for *integer r in set* $\{1, 2, \ldots, 1/t_j\}$ **do**

Sample $u_1^{(r)}, \ldots, u_{n_j}^{(r)}$ following i.i.d. G(u) Obtain rth unbiased estimates of $p(\vartheta^{(j+1)}|X)$ in (2.11): $\hat{p}^{(r)}(\vartheta^{(j+1)}|X) = \frac{1}{n_j}\sum_{w=1}^{n_j} L_n(\vartheta^{(j+1)}; u_w^{(r)})$;

end

Calculate unbiased estimates of $p_{t_j}(\vartheta^{(j+1)}|X)$: $\hat{p}_{t_j}(\vartheta^{(j+1)}|X) = \prod_{r=1}^{1/t_j} \hat{p}^{(r)}(\vartheta^{(j+1)}|X)$;

Sample v from uniform distribution U(0,1);

Calculate acceptance rate `// based on pseudo-marginal MCMC`

$h_p(\vartheta^{(j+1)}, \theta^{(j)}) = \min\{1, \hat{p}_{t_j}(\vartheta^{(j+1)}|X)/\hat{p}_{t_j}(\theta^{(j)}|X)\}$, where t_j is the temperature at iteration j.;

if $v \leq h(\vartheta^{(j+1)}, \theta^{(j)})$ **then**

Set $\theta^{(j+1)} = \vartheta^{(j+1)}$;

Set $\hat{p}_{t_{j+1}}(\theta^{(j+1)}|X) = \hat{p}_{t_j}(\vartheta^{(j+1)}|X)$;

else

Set $\theta^{j+1} = \theta^j$;

Set $\hat{p}_{t_{j+1}}(\theta^{(j+1)}|X) = \hat{p}_{t_j}(\theta^j|X)$;

end

set j=j+1;

end

Algorithm 2: Pseudo marginal MCMC with Simulated annealing

The key difference between pseudo-marginal MCMC with simulated annealing (Algorithm 2) and simulated annealing (Algorithm 1) is the acceptance rate. The acceptance rate in Algorithm 2 is $h_p(\vartheta^{(j+1)}, \theta^{(j)})$

of which the numerator and denominator are unbiased estimates of their counterparts in Algorithm 1. If we let temperature pattern $\{t_i\}$ be a fixed value t. Algorithm 1 is the Metropolis hastings, while, algorithm 2 is the pseudo-marginal MCMC. Both of them can generate Markov chains in such a way that, as more and more sample values are produced, the distribution of values more closely approximates the desired distribution $p_t(\theta|X)$. But, the pseudo-marginal MCMC provides us easy-to-calculate acceptance rate $h_p(\vartheta^{(j+1)}, \theta^{(j)})$ when the desired distribution $p_t(\theta|X)$ is not tractable.

Generally, it is normal to design a pattern of temperature $\{t_i\}_{i=1}^{\infty}$ decreasing to zero when performing simulated annealing. In our case, we keep temperature decreasing gently and gradually like the downward stairs. Then, use the normality check in introduced Johansen et al. (2008) to check the convergence of the Markov chain with this temperature. When the Markov chain generated from our Algorithm 2 converges, we stop decreasing the temperature and use the average of the sample with respect to the last temperature to estimate the parameter.

2.5 Simulation study

In this section, we apply integrated empirical likelihood sampler to some typical Lévy processes in Finance. The main motivation of using Lévy processes in finance is from the goodness of fit of asset return. The return distributions generally have skewness and heavy tails than the normal distribution, which is noted by Fama (1965). Based on flexible designs, Lévy processes can capture the skewness and heavy tails property of financial asset returns. Suppose S_t represents the price of financial securities at time t, it is common to assume that logarithm of price, $Y_t = \log S_t$, follows Lévy processes.

2.5.1 Lévy processes

An R^d-valued stochastic process $\{Y_t\}$ is said to be a Lévy process if it is continuous in probability and has independent and stationary increments. Independent increments mean that increments $Y_s - Y_t$ and $Y_u - Y_v$ are independent where $[t, s]$ and $[v, u]$ are intervals without overlap. Stationary increments mean that the distribution of $Y_t - Y_s$ is the same as Y_{t-s}. Thus, if we observe data based on several discrete points with same time interval δ, say $\{\delta, 2\delta, \ldots, (n+1)\delta\}$ in time span $[0, T]$, our data will be $\{Y_\delta,\ Y_{2\delta}, \ldots, Y_{n\delta}\}$. Then, the increments $\{X_t = Y_{(t+1)\delta} - Y_{t\delta}\}_{t=1}^{n}$ will be i.i.d.. Moreover, the characteristic function of $\{X_t\}_{t=1}^{n}$ is available.

The existence of expression of characteristic function for Lévy increments $\{X_t\}_{t=1}^{n}$ is due to Lévy-Khintchine formula. Based on this formula, simple analytic expressions of characteristic functions are avail-

able for a wide range of Lévy processes. It is stated that the characteristic function of a Lévy increment X_t can be expressed as

$$\phi(u) \equiv E\{\exp(iu'X_t)\} = \exp\{\delta[i\mu u - \frac{1}{2}u'\Sigma u + \int_{R^d}(\exp(iu'x) - 1 - iu'xI_{\{|x|\leq 1\}}J(dx))]\}, \tag{2.14}$$

where (μ, Σ, J) is the triplet of the Lévy processes including the drift $\mu \in R^d$, the volatility matrix of its diffusion component Σ which is a symmetric non-negative definite $d \times d$ matrix and the Lévy measure $J(dx)$ on R^d with $J(\{0\}) = 0$ and $\int_{R^d} \min(|x|^2, 1)J(dx) < \infty$.

Lévy triplet (μ, Σ, J) can determine the Lévy process X_t due to Lévy-Khintchine formula. The Lévy measure $J(dx)$ describes the expected number of jumps with jump size x in a time interval of length 1. If $J = 0$, this Lévy process is just Brownian motion with drift μ and volatility matrix Σ without jumps. There are two types of Lévy processes, finite activity processes and infinite activity Lévy process. If $\int_{R^d} J(dx) < \infty$, it means that there are only finitely many jumps in any given finite interval. We call them finite activity Lévy processes. It also can be proved that every jump process with finite activity is a compound Poisson type process. Thus, Lévy processes with finite activities are also called jump-diffusion process of which the jump component is of compound Poisson type with Poisson arrival intensity $\lambda = \int_{R^d} J(dx)$ and jump size distribution $\lambda^{-1}J$. Typical jump-diffusion models in finance include Merton's jump-diffusion model (Merton (1976)) in which the jump size follows normal distribution and Kou's double exponential jump-diffusion model(Kou (2002)), which allows asymmetric jump size following double exponential distributions.

For infinite activity Lévy processes, $\int_{R^d} J(dx)$ is infinite. It means that there are infinite many jumps in any given finite time intervals. Compared with diffusion jump models, pure jump Lévy processes with infinite activity might have better representations of stock price dynamics (Geman (2002)). Typical financial models with infinite activities include the generalized hyperbolic model (Barndorff-Nielsen (1977)) and its subclasses, CGMY model (Carr et al. (2002)) and its subclasses.

In this section, we will implement our integrated empirical likelihood sampler to do parameter estimation for several financial models based on Lévy processes with explicit expression of characteristic functions including multivariate version of Black-Shores model, Kou's jump-diffusion model(Kou (2002)), Variance Gamma model(Madan and Seneta (1990)).

2.5.2 A multivariate version of Black-Scholes-Merton Model

We start from a simple example, a multivariate version of Black-Scholes-Merton Model (Black and Scholes (1973); Merton (1973)). Under the setting of this model, the log-equity return is

$$X_t \equiv \log \frac{S_{(t+1)\delta}}{S_{t\delta}} = \mu\delta + \sigma\sqrt{\delta}\epsilon_t, \tag{2.15}$$

where δ is the fixed time interval for observed prices $S_t \in R^d$; $\mu \in R^d$ is the drift term; $\Sigma = \sigma\sigma^T$ is the $d \times d$ diffusion matrix in annualized scale. The Lévy triplet of this process is $(\mu, \Sigma, J(dx) = 0)$ and its characteristic function is

$$\phi(u) = \exp\{\delta(iu^T\mu - \frac{1}{2}u^T\Sigma u)\}. \tag{2.16}$$

Algorithm setting

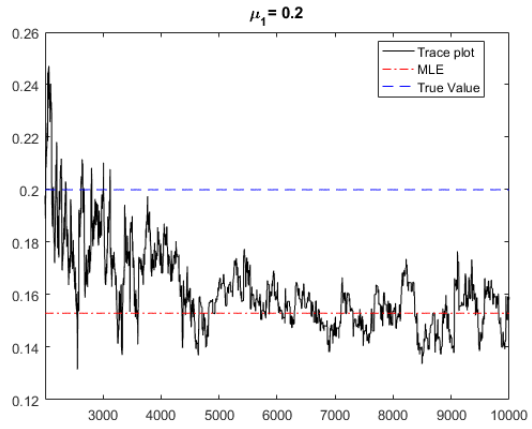
For this case, we simply set the support of $G(u)$ is the region where the real part of empirical characteristic function (ECF) value based on our simulated sample between 0.1 and 1 (See Fig 2.1f). The reason for choosing 0.1 to be the lower bound is that ECF real value is noisy in the area far away from original zero point. This effects are even more dramatic when sample size is small. This noisy area gives us barely useful information. Thus, we set a lower-bound to obtain useful information of ECF. The distribution of $G(u)$ is set to be uniform distribution. For the integrated empirical likelihood sampler, we run 10000 draws with 2000 burn in draws. We set the temperature pattern $T_0 = 1,\ T_{9999} = 0.01$. When $i < 5000$, $T_i = T_{9999}^{(\lfloor i/100 \rfloor)/50}$. When $i \geq 5000$, $T_i = T_{9999}$. The number of particles for unbiased estimates on ith iteration is $N_i = N_0\lceil\frac{1}{T_i}\rceil$. N_0 is 2 here. The proposal we use for μ and Σ is normal inverse Wishart distribution. Supposing we don't have any useful information of parameters, we set the prior $p(\theta)$ to be flat for all parameters in a certain large range.

Performance

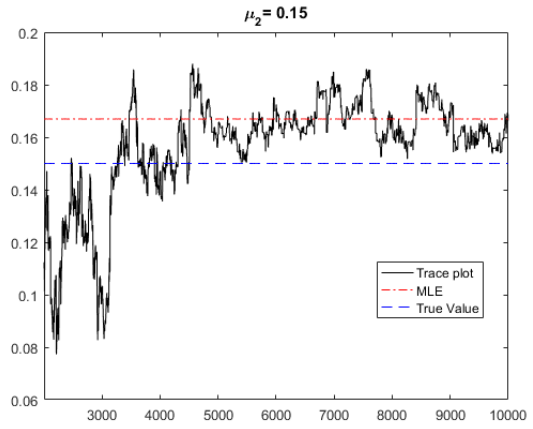
We demonstrate the validity of our algorithm for the two-dimensional version of Black-Shores-Merton model. We simulate 1000 weekly returns based on the annualized drift, $\mu = (\mu_1,\ \mu_2) = (0.2, 0.15)$, annualized diagonal elements of diffusive matrix Σ, $\sigma_1^2 = 0.04,\ \sigma_2^2 = 0.0225$ and annualized off diagonal elements of diffusive matrix to make correlation between two assets ρ is 50%. Our around 20% annualized return with around 0.2 volatility is quite typical in stock market. Also, Chan et al. (2009) use similar parameters. For the 50% correlations of assets, this is also quite standard (See Jacquier et al. (2007)).

Fig 2.1 shows the trace plot of our sampler to identify our MEPE. We also plot the true value and MLE

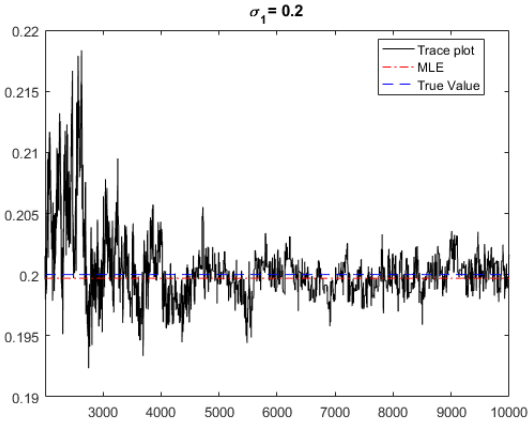
in the graph. We can see that if we use last several samples mean in our trace plot to estimate MEPE, our estimated MEPE will have quite comparable accuracy with MLE. Also, we can find that the trace plot for the elements in diffusive matrix Σ has better performance compared with drift term and it converges much quicker. The possible reason for this is that the standard error of our estimated μ is larger than Σ based on MEPE. This is also consistent to the result in Chan et al. (2009).



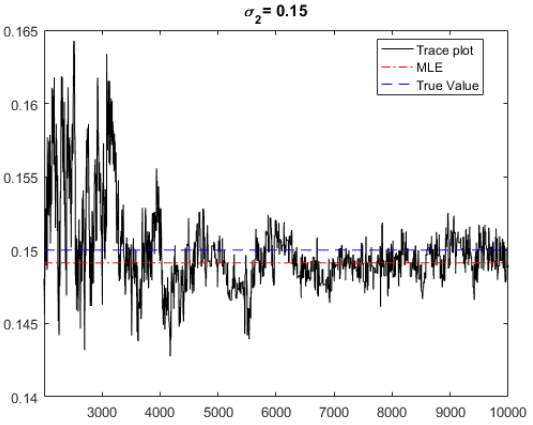
(a) μ_1 trace plot



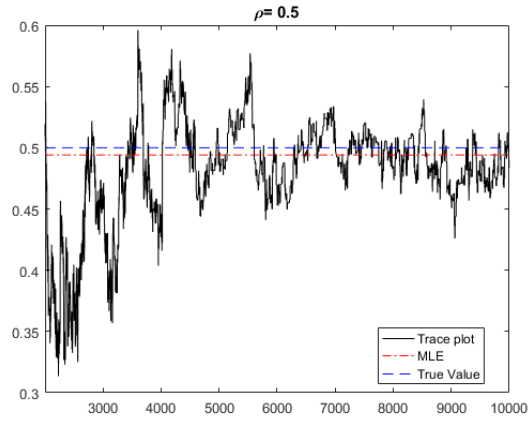
(b) μ_2 trace plot



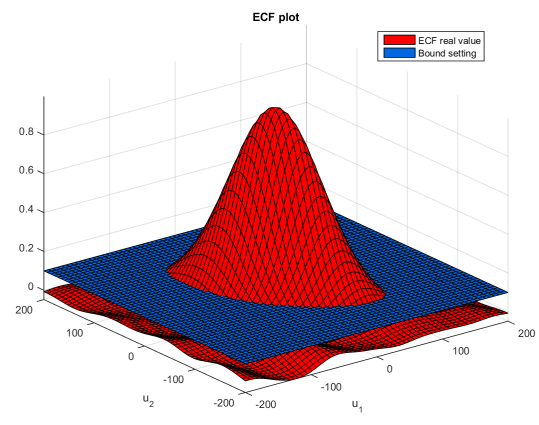
(c) σ_1 trace plot



(d) σ_2 trace plot



(e) ρ trace plot



(f) Real part of ECF value

Figure 2.1: Empirical characteristic function information and the trace plot of multivariate version of Black-Scholes-Merton model where $\mu_1 = 0.2,\ \mu_2 = 0.15,\ \sigma_1^2 = 0.04,\ \sigma_2^2 = 0.0225,\ \rho = 0.5$.

2.5.3 Kou's jump-diffusion model

Let's consider a finite activity Lévy process, Kou's jump-diffusion model (Kou (2002)). In this model, we assume the logarithm of the equity return

$$X_t \equiv \log \frac{S_{(t+1)\delta}}{S_{t\delta}} = \mu\delta + \sigma\sqrt{\delta}\epsilon_t + \sum_{i=1}^{N_t} Z_i, \tag{2.17}$$

where δ is the fixed time interval for observed prices $S_t \in R$. $\mu \in R$ is the drift term and σ is the volatility in annualized scale. These setting is similar to the Black-Shores-Merton model. In addition to drift term and diffusion term, we also assume jumps arrive according to a Poisson process N_t with intensity λ. The jump sizes $\{Z_i\}$ follow i.i.d. asymmetric double exponential distribution with density

$$f_Z(z) = p\eta_1 \exp(-\eta_1 z)1_{\{z\geq 0\}} + (1-p)\eta_2 \exp(\eta_2 z)1_{\{z<0\}}. \tag{2.18}$$

That is, the positive jumps probability is p, mean positive jump size is $\frac{1}{\eta_1}$ and mean netative jump size is $\frac{1}{\eta_2}$. The Lévy triplet of this process is $(\mu, \sigma^2, J(dz) = \lambda f_Z(z)dz)$ and its characteristic function is

$$\phi(u) = \exp\{\delta[i\mu u - \frac{1}{2}\sigma^2 u^2 - \lambda(1 - \frac{p\eta_1}{\eta_1 - iu} - \frac{(1-p)\eta_2}{\eta_2 + iu})]\}. \tag{2.19}$$

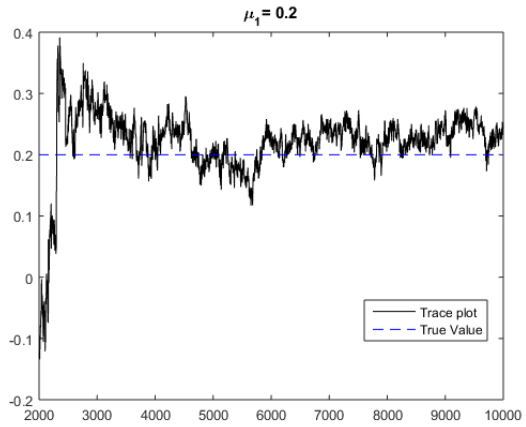
Algorithm setting

In this case, we set the support of $G(u)$ to be positive ECF real part area like the setting in Black-Shores-Merton model. It is easy to show that the real jump part of the characteristic function $\exp\{\delta\lambda(1 - \frac{p\eta_1}{\eta_1 - iu} - \frac{(1-p)\eta_2}{\eta_2+iu})\}$ normally decays to zero much more quickly than the drift and diffusion part $\exp[\delta(i\mu u - \frac{1}{2}\sigma^2 u^2)]$ when $u \to \infty$. To capture more information about jump parameters, we put more weights around zero by setting $G(u)$ to be a triangular distribution. For other settings, they are same as Black-Shores-Merton model's.

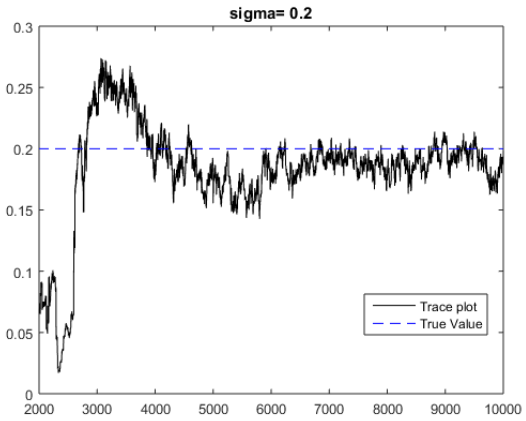
Performance

The annualized parameter set we use is that $\delta = \frac{1}{52},\ \mu = 0.2,\ \sigma = 0.2,\ \lambda = 10,\ p = 0.4,\ \eta_1 = 5,\ \eta_2 = 2.5$. That is, we assume that there are about 10% weekly returns jump with average negative jump size 40% and average positive jump size 20% per year. This is reasonable for U.S. stocks with high volatility. To simulate the sample path of Kou's jump-diffusion model, we use the simulation method of jump-diffusion models indicated in Glasserman (2003). The simulated data are 1000 weekly returns.

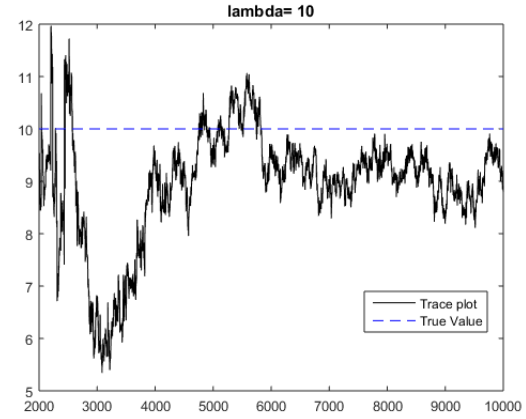
Fig 2.2 shows the trace plot of our sampler to identify our MEPE. We also plot the true value in the graph. At first, when temperature is low, trace plot search some areas which even very far away from the initial value. Then, with the temperature rising, trace plot converges to the value around true value. The trace plot demonstrates the validity of algorithm for the Kou's jump-diffusion model.



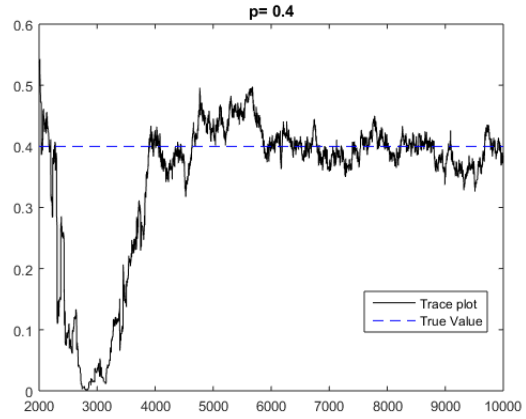
(a) μ trace plot



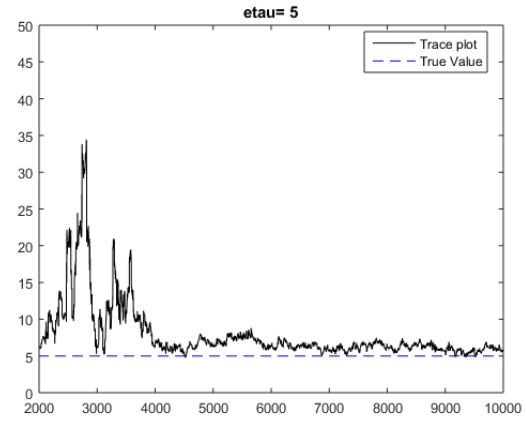
(b) σ trace plot



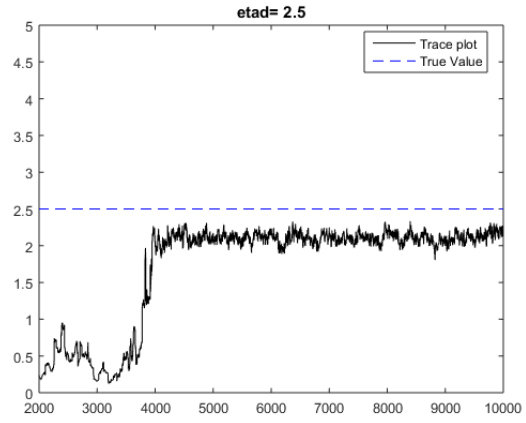
(c) λ trace plot



(d) p trace plot



(e) η_u trace plot



(f) η_d trace plot

Figure 2.2: Trace plot of Kou's jump-diffusion model where $\mu = 0.2,\ \sigma^2 = 0.04,\ \lambda = 10,\ p = 0.4,\ \eta_u = 5,\ \eta_d = 2.5$.

2.5.4 Variance Gamma model

Now, we move to an asset pricing model with infinite activity process, variance Gamma (VG) model. It is proposed by Madan and Seneta (1987) for stock market data. It is also a degenerate case of CGMY process (Carr et al. (2002)) with the Lévy triplet $(\gamma, 0, J(dx))$, where

$$\gamma = C(MG)^{-1}(G(\exp(-M) - 1) - M(\exp(-G) - 1))$$

$$J(dx) = (C\exp(-Mx)|x|^{-1}1_{\{x>0\}} + C\exp Gx|x|^{-1}1_{\{x<0\}})dx.$$

if we define the drift (μ) adjusted log-equity return as $X_t - \mu\delta \equiv \log\frac{S_{(t+1)\delta}}{S_{t\delta}} - \mu\delta$,its characteristic function of VG process is given by

$$\phi(u) = (\frac{GM}{GM + (M-G)iu + u^2})^{C\delta}, \tag{2.20}$$

where δ is the fix time interval for observed prices $S_t \in R$. Parameter G and M control the sign of skewness. If $G = M$, the density of log-equity return is symmetric with mean 0. The skewness will be negative when $G < M$ indicating roughly larger negative jumps of returns happen more frequently compared with positive jumps. Similarly, $G > M$ indicates more frequently larger positive jumps. Parameter C mainly controls the kurtosis of log-equity return distribution.

It is not difficult to show that Lévy measure $J(dx)$ of VG process has infinite mass. Thus, it is an infinite activity process with infinitely many jumps in any finite intervals. There are several alternative representations of the VG processes. To better understand VG process, we recommend several references: Madan and Seneta (1990); Madan and Milne (1991); Madan et al. (1998).

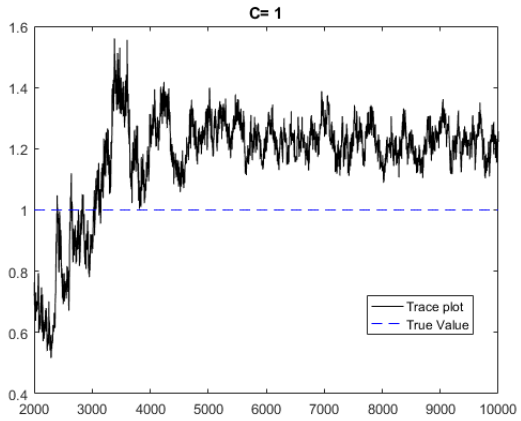
Algorithm setting

Analyzing the characteristic function of VG process, one thing we need to mention is that absolute value of $\phi(u)$, $|\phi(u)|$, decays to zero with polynomial rate which is too slow for us to define the support region of $G(u)$ to include most of the positive area of $|\phi(u)|$. To make it simple, we set the support of $G(u)$ to be the region including all largely and relatively different slop of $|\phi(u)|$ to catch pattern of characteristic function. The distribution of $G(u)$ is set to be uniform distribution over the support. For other settings, they are same as the previous two's.

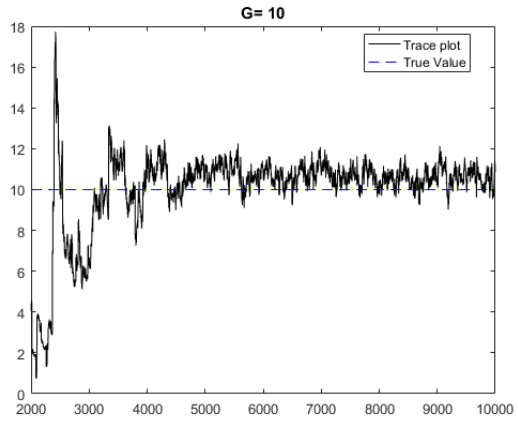
Performance

We demonstrate the validity of our algorithm for VG model. We simulate 1000 weekly returns based on the annualized parameter, $C = 1,\ G = 10$ and $M = 5$. This is reasonable for relatively highly volatile equities in the United States.

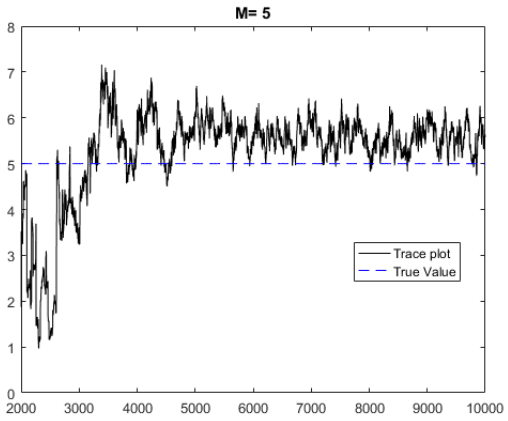
Fig 2.3 shows the trace plot of our sampler to identify MEPE. We also plot the true value of the parameters. We can find that after about 4000 draws, our trace plots of C, G and M are stable around the true values. In the very first draws, it can search a large area which is far away from true values and initial values. All in all, with temperature increasing, our sampler can give us more and more accurate estimates of parameters.



(a) C trace plot



(b) G trace plot



(c) M trace plot

Figure 2.3: Trace plot of VG model where $C = 1,\ G = 10,\ M = 5$.

	Black-Shores-Merton	Kou	VG
Mean	$\mu\delta$	$(\mu + \frac{\lambda p}{\eta_u} - \frac{\lambda(1-p)}{\eta_d})\delta$	$(\mu + C(\frac{1}{M} - \frac{1}{G}))\delta$
Variance	$\sigma^2\delta$	$(\sigma^2 + 2(\frac{\lambda p}{\eta_u^2} + \frac{\lambda(1-p)}{\eta_d^2}))\delta$	$C(\frac{1}{M^2} + \frac{1}{G^2})\delta$
Skewness	0	$\frac{6(\frac{\lambda p}{\eta_u^3} - \frac{\lambda(1-p)}{\eta_d^3})}{(\sigma^2 + 2\frac{\lambda p}{\eta_u^2} + \frac{\lambda(1-p)}{\eta_d^2})^{1.5}\sqrt{\delta}}$	$\frac{2(\frac{1}{M^3} - \frac{1}{G^3})}{(\frac{1}{M^2} + \frac{1}{G^2})^{1.5}\sqrt{C\delta}}$

Table 2.1: Moments for selected models

2.6 Case Study

In this section, we use real data to fit Black-Scholes-Merton model, Kou's jump-diffusion model and Variance Gamma model to check the validity of our algorithm. The data we use is weekly log-returns of the S&P500 index from January 2, 1987 to December 28, 2007. It includes 1058 returns in total. The data is plotted in Figure 2.4. The mean of weekly log-returns is 0.0017. The volatility of it is 0.0219. The weekly log-returns have negative skewness -0.5647, which implies negative jumps are generally bigger than positive jumps.
The setting of algorithm for each model is same as the one in the simulation study. We choose the mean value of last 1500 generated sample to be our parameter estimates which are reported in Table 2.2.
For Black-Shores-Merton model, the estimates of the drift μ is 0.0938. This indicates an estimated mean of log returns $\hat{\mu}\delta = 0.0938/52 = 0.0018$ which is similar to the sample mean of log-returns, 0.0017. The estimates of the σ is 0.1568. Then, the estimated volatility is $\hat{\sigma}\sqrt{\delta} = 0.1568/\sqrt{52} = 0.0217$, which perfectly match the sample volatility 0.0219. Unfortunately, Black-Shores-Merton model cannot capture higher moments. For example, the kurtosis implied by this model is always zero, while, the log-return data has a negative kurtosis. Fortunately, the following two models, Kou's jump-diffusion model and Variance Gamma model, have capabilities to match higher moments.

Compared with Black-Shores-Merton model, Kou's jump-diffusion models introduce positive jumps and negative jumps in log-return processes. Our log-return processes have estimated drift $\hat{\mu} = 0.15$ and volatility $\hat{\sigma} = 0.0930$. The estimated jump intensity parameter in the process $\hat{\lambda} = 9.0485$, which indicates that there are more or less 10 jumps occurring per year. The probability of negative jumps when a jump appears is $1 - \hat{p} = 1 - 0.4318 = 0.5682$, which is bigger than half. The jump size of positive jumps follows exponential distribution with mean $1/\eta_u = 0.0269$; the negative jumps,$1/\eta_d = 0.0331$. Jumps in the process can capture the volatility induced by large movements so that our estimated volatility in the Kou's jump-diffusion model is smaller than the one in Black-Shores-Merton model. Moreover, indicated by estimated p, η_u and η_d, negative jumps are more frequent and bigger in general compared with positive jumps. Thus, our estimated

drift is bigger than the correspondent one in Black-Shores-Merton model to compensate the impact of negative jumps. According to Table 2.1, the estimated mean and volatility in Kou's jump-diffusion model are 0.0016 and 0.0222, which match the sample mean (0.0017) and volatility (0.0219) of log-returns very well. The estimated skewness via Table 2.1 is negative (-1.1592), which is consistent to the negative sample skewness (-0.5647).

For Variance Gamma model, because we don't include drift term μ in VG characteristic function (2.5.4) in Chapter 2.5.4, we estimate C, G, M through mean-adjusted log-return data. Then, the corresponding characteristic function will be $\phi(u)/\phi(-i)$ where $\phi(u)$ is in Equation (2.20) and i is the imaginary unit. The use of mean adjusted data to estimate pure jump processes parameters also appears in Carr et al. (2002). The estimated volatility of log-returns is 0.0233 (through Table 2.1), which can match the sample volatility of log-returns. Moreover, estimated skewness is -0.4011, which indicates a negative skewness. This is also consistent to our sample skewness of log-returns.

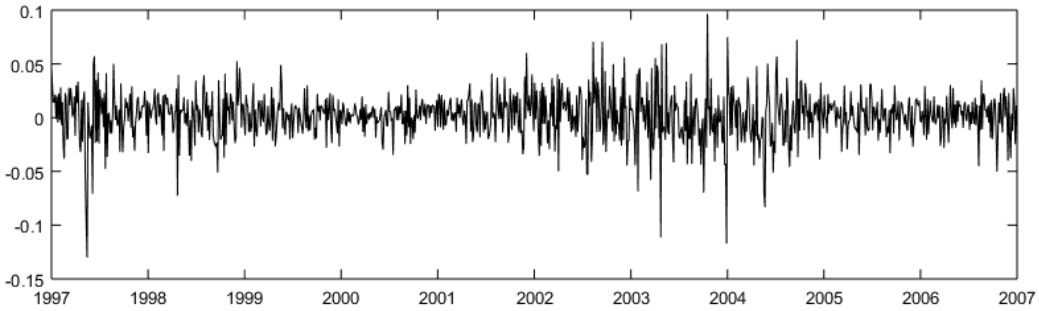


Figure 2.4: The weekly log-return of S&P500 index from January 2, 1987 to December 28, 2007.

2.7 Concluding Remarks

In this chapter, we propose MEPE estimator based on characteristic functions. Our MEPE inherits asymptotic properties of MELE from Chan et al. (2009). Moreover, by using our integrated empirical likelihood sampler, we can obtain MEPE accurately in a relatively efficient way without approximation, which adopted to obtain MELE in Chan et al. (2009). We also propose an easy-to-verify equivalent regularity conditions for A.4 because verifying the positive-definiteness for a integral form directly is generally difficult. There are some futures works could be done. For example, in fact, our integrated empirical likelihood sampler is designed for our integral target (posterior in (2.4)), which is based on pseudo-marginal MCMC with simulated annealing. It should work for other statistical inferences with the target of integral form such as continuous version of ECF estimation (See the review of it in Yu (2004)). Also, with very low temperature,

Table 2.2: Estimates of parameters for the S&P500 index

(a) Black-Scholes-Merton model

	μ	σ
EL	0.0938	0.1568

(b) Kou's jump-diffusion model

	μ	σ	λ	p	η_u	η_d
EL	0.1500	0.0930	9.0485	0.4318	37.1441	30.2522

(c) Variance Gamma model

	C	G	M
EL	4.7770	17.8791	18.9350

the computational burden will increase dramatically. This is also a common problem for simulated annealing based algorithm. So, to further increase the efficiency of this algorithm could be interesting as well especially for some models needing very low temperatures.

Chapter 3

Maximum likelihood inference for Lévy process based models in finance

3.1 Introduction

Lévy processes are popular in quantitative finance since they fit financial data with skewness and fat tails better while staying computationally tractable. See Tankov (2003) for discussions of Lévy process models in finance.

Suppose $Y = \{Y_t, t = 0, \delta, 2\delta, \ldots\}$ is a sequence of an evenly sampled financial variable, e.g., the log price of a certain asset. If we assume that Y is a Lévy process, then $\{X_k = Y_{k\delta} - Y_{(k-1)\delta}, k \geq 1\}$ is an i.i.d. sequence due to the fact that Lévy processes have independent and stationary increments. The distribution of X_i is usually complex or not available explicitly. But its characteristic function $\phi(u) = E(\exp(iuX_k))$ often has an explicit form due to the Lévy-Khintchine formula (refer to Tankov (2003)). In some other cases, we don't have i.i.d. data, and assume that $\{Y_t\}$ is sampled from a Lévy driven Markov process:

$$dY_t = \mu(t, Y_t)dt + \sigma(t, Y_t)dL_t$$

for some Lévy process $\{L_t, t \geq 0\}$. Again, the transition density of such a process may not be available. However, the conditional characteristic function of $X_k = Y_{k\delta}$, $\phi_k(x, u) = E(\exp(iuX_{k+1})|X_k = x)$, is sometimes known.

Given the model and data, our primary missions are parameter estimation and empirical assessment of the model. Testing of trading strategies as in Avellaneda and Lee (2010) and volatility forecasting as in Kim et al. (2008) can be done only when one has a appropriately fitted model.

Various statistical inference methods are available for fitting processes with explicit characteristic functions. One class of moment based methods are the empirical characteristic function (ECF) methods, first proposed by A.Feuerverger and R.A.Mureika (1977). The case of independent data was studied by Feuerverger and McDunnough (1981*a*). K.J.Singleton (2001) studied the case of dependent data. Under some mild regularity conditions, ECF estimators are consistent. However, they are usually not efficient. Refer to Yu (2004)

for a review of the ECF method. Another class of non-parametric methods are the empirical likelihood methods, first proposed by Owen (1988). Kunitomo and Owada (2006),Chan et al. (2009) and Chen et al. (2013) have studied the empirical likelihood methods for fitting Lévy processes with explicit characteristic functions and more general Lévy driven processes with explicit conditional characteristic functions.

Compared to the above methods, the maximum likelihood estimation (MLE) has several advantages. First, under certain regularity conditions, maximum likelihood estimators are asymptotically efficient. Second, various hypothesis testing and model selection techniques are based on the maximum likelihood method. Last but not the least, the MLE can be accomplished efficiently using Fourier methods when the characteristic functions are available and analytic in the complex plane. The maximum likelihood method has been used in various quantitative finance works. For example, when fitting Kou's jump-diffusion model, Ramezani and Zeng (2007) truncates the infinite series representation of the transition density to obtain approximated maximum likelihood estimates. However, in many Lévy process models, the densities do not admit explicit expressions. K.J.Singleton (2001) implements the maximum likelihood method using the Gauss-Legendre quadrature for inverse Fourier transform. However, when the characteristic functions are analytic, the simplest trapezoidal rule is highly accurate with exponentially decaying errors.

In this Chapter, we utilize the analyticity of the characteristic function and obtain highly accurate values of the density using the trapezoidal rule for the inverse Fourier transform. We focus on likelihood inference for Lévy process models used in quantitative finance as well as one-dimensional Lévy driven processes with explicit conditional characteristic functions. Our theoretical framework establishes asymptotic properties of the proposed approximated maximum likelihood estimation (AMLE) and AMLE based hypothesis testing. We present regularity conditions based on characteristic functions. We present efficient implementation of the AMLE. To examine the effectiveness of our estimation procedure, we perform simulation studies and empirical studies. Simulation studies show that very large samples may be needed to accurately identify the true parameters for some popular models, such as the CGMY model. Further analysis shows that the log-likelihood surface is saddle-shaped with relatively flat areas, which is causing the difficulties. On the other hand, the parameters of commonly used jump-diffusion models, including Merton and Kou's models, are relatively easier to identify. We also fit empirical equity return data with some popular Lévy models. Numerical results show the appealingness of some infinite activity models, such as the normal inverse Gaussian model.

This Chapter is organized as follows. In Chapter 3.2, we present our AMLE with asymptotic properties and the corresponding regularity conditions. Likelihood based hypothesis testing and model selection are also presented here. In Chapter 3.3, we present some commonly used Lévy models. In Chapter 3.4, we

provide details of the numerical implementation, perform simulation studies as well as empirical studies using real world financial data. Chapter 3.4.4 summarizes the conclusion. All proofs are in Appendices.

3.2 Maximum likelihood inference

In this section, we mainly focus on approximated likelihood based on inverting characteristic function technique. We assume that $X = \{X_1, \dots X_N\}$ follows i.i.d distribution F with implicit characteristic function $\phi(u;\theta_0)$ for $u \in R$. θ_0 is the true parameter vector. $\theta = (\theta_1, \cdots, \theta_p)$ is the vector of unknown parameters associated with distribution F within parameter space $\Theta \subset R^p$. Considering that a Lévy process has independent and stationary increments and the form of its characteristic function is available due to Lévy-Khintchine formula, the assumption of data can be applied to Lévy processes directly. Later, we will extent i.i.d case to non-i.i.d case in Chapter 3.2.3.

3.2.1 Parameter estimation

Suppose we have the characteristic function associated with distribution function $F(x;\theta)$:

$$\phi(u;\theta) = \int_{-\infty}^{\infty} \exp(iux) F(dx;\theta).$$

If $\phi(u;\theta)$ is absolutely integrable,i.e., $\int_{-\infty}^{\infty} |\phi(u;\theta)| du \leq \infty$, through Fourier inversion theorem, there exists a bounded and uniformly continuous probability density function $f(x;\theta)$ corresponding to $F(x;\theta)$:

$$f(x;\theta) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-iux)\phi(u;\theta) du. \tag{3.1}$$

The approximated density with fixed $M \in \mathbf{Z}^+$,$h \in \mathbf{R}^+$ and $a \in \mathbf{R}$ is defined below

$$f^{M,h,a}(x;\theta) = \frac{1}{2\pi} \sum_{m=-M}^{M} e^{-ix(mh+ia)} \phi(mh + ia;\theta) h. \tag{3.2}$$

This form is the inverse Fourier transform integral approximated by trapezoidal summation. h mainly controls the error between the integral and summation approximation (discretization error) and Mh controls the error due to the truncation in the summation approximation. If $a = 0$, the integral in (3.1) is along the real line. By the Cauchy integral theorem, this integration can be shifted to a horizontal line $x + ia, x \in R$ for a in a certain region to perform trapezoidal rule on this horizontal line. A good review about trapezoidal rule and Fourier transform is Abate and Whitt (1992).

If we denote the likelihood function as $L(\theta; x) = f(x; \theta)$, the log-likelihood function is

$$l(\theta; X) := \frac{1}{N}\sum_{j=1}^{N}\log(f(X_j, \theta)) = \frac{1}{N}\sum_{j=1}^{N}\log(\frac{1}{2\pi}\int_R e^{-iX_j y}\phi(y)dy)$$

and likelihood function will be $L(\theta; X) = \exp(l(\theta; X))$.

Correspondingly, the approximated-log-likelihood is defined as

$$l^{M,h,a}(\theta; X) := \Re(\frac{1}{N}\sum_{j=1}^{N}\log(f^{M,h,a}(X_j, \theta))) = \Re(\frac{1}{N}\sum_{j=1}^{N}\log(\frac{1}{2\pi}\sum_{m=-M}^{M} e^{-iX_j(mh+ia)}\phi(mh+ia)h)), \quad (3.3)$$

where $\Re(\cdot)$ represents the real part of any given complex number.

The approximated maximum likelihood estimator (AMLE) $\hat{\theta}_N^{M,h,a}$ is defined as

$$\hat{\theta}_N^{M,h,a} = \arg\max_{\theta\in\Theta} l^{M,h,a}(\theta; X).$$

Before describing the asymptotic properties about AMLE, we need the following definition and regularity conditions.

A.1 For any $\theta \in \Theta$, the characteristic function $\phi(u; \theta)$ is analytic in $\mathcal{D}_{(d_-,d_+)}$ where $\mathcal{D}_{(d_-,d_+)} = \{z \in C : \Im(z) \in (d_-, d_+)\}$, $-\infty < d_- < 0 < d_+ < \infty$ and d_-, d_+ do not depend on θ. $\Re(z)$ is the real part and $\Im(z)$ is the imaginary part of z.

A.2 For any given $\theta \in \Theta$, $\int_{d-}^{d+} |\phi(x + iy; \theta)|dy \to 0$ when $x \to \pm\infty$.
$||\phi||^{\pm} := \lim_{\epsilon\to 0+}\int_R |\phi(x + i(d_\pm) \mp \epsilon; \theta)|dx < +\infty$ uniformly on $\theta \in \Theta$.

A.3 The parameter in Equation (3.2), a, is a given real value and $a \in (d_-, d_+)$.

A.4 $|\phi(x + ia; \theta) \le k|x|^n \exp(-c|x|^\nu)$, $x \in \mathbf{R}$ with $a \in (d_-, d_+)$ for some $\kappa > 0, \nu > 0, c > 0, n \in \mathbf{R}$ or $\kappa > 0, \nu > 0, c = 0, n < -1$. Here,$\kappa, \nu, c$ and n are not related to the parameter $\theta \in \Theta$ and they may or may not depend on a.

A.5 $Mh \ge (n/c\nu)^{1/\nu} 1_{c>0,n>0}$ for n, c, ν defined above.

Remark 3. Regularity condition A.1 requires the analyticity of the characteristic function $\phi(u; \theta)$ in $\mathcal{D}$ given θ. For regularity condition A.2, noticing that the analyticity band $\mathcal{D}$ is an open set, we can arbitrarily choose d_+ and d_- inside analyticity band $\mathcal{D}$. Then, the characteristic function $\phi(u; \theta)$ is analytic in $\{z \in C : \Im(z) \in [d_-, d_+]\}$ and $||\phi||^{\pm} := \int_R |\phi(x + i(d_\pm); \theta)|dx$. Regularity condition A.3 implies that our manipulation are restricted to the analyticity band of $\phi(u; \theta)$. Regularity A.4 guarantee that the characteristic function has

exponential or binomial tails which grantees its absolute integrability. Then, distribution density function defined in (3.1) exists. Moreover, the smoothness of density can be determined by the tail behavior of the characteristic functions. Specifically, if $\nu > 0$, the characteristic function has exponentially-decaying tails and its distribution density function has derivatives of all orders. If $c = 0,\ n < -1$, the characteristic function has binomially-decaying tails and its distribution density function has derivatives of orders up to $\lfloor -n \rfloor$. Last but not least, if we can choose proper d_- and d_+ to make A.4 hold without dependency on $a \in (d_-, d_+)$, it is easy to prove that A.2 will hold. In fact, this is the case for a lot of Levy processes.

Class A regularity conditions are used to control the difference between approximated-log-likelihood $l^{M,h,a}(\theta; x)$ and real likelihood function $l(\theta; x)$ uniformly on $\theta \in \Theta$. Specifically, we have the following theorem.

Theorem 6. *Under regularity conditions of class A, we define the error of approximation of probability density function as*

$$E^F_{h,M}(\phi, a)(x) = f(x;\theta) - f^{M,h,a}(x;\theta) = \frac{1}{2\pi}\int_R e^{-ixy}\phi(y;\theta)dy - \frac{1}{2\pi}\sum_{m=-M}^{M} e^{-ix(mh+ia)}\phi(mh+ia;\theta)h.$$

Then, we have the bound of error

$$|E^F_{h,M}(\phi, a)(x)| \leq \frac{e^{-2\pi(a-d_-)/h}}{2\pi(1-e^{-2\pi(a-d_)/h})}e^{xd_-}||\phi||^- + \frac{e^{-2\pi(d_+-a)/h}}{2\pi(1-e^{-2\pi(d_+-a)/h})}e^{xd_+}||\phi||^+ + T_{Mh} \tag{3.4}$$

where $T_{Mh} = \frac{ke^{ax}}{|n+1|\pi}(Mh)^{n+1}$ *if* $c = 0, n < -1$, *and* $T_{Mh} = \frac{ke^{ax}}{\pi\nu c^{(n+1)/\nu}}\Gamma(\frac{n+1}{\nu}, c(Mh)^\nu)$ *if* $c > 0$. *Incomplete Gamma function* $\Gamma(s,b) = \int_b^\infty e^{-t}t^{s-1}dt$. *Moreover, let* $Mh \to \infty$ *and* $h \to 0$, *then, the bound of error* $E^F_{h,M}(\phi, a)(x)$ *will decay to zero uniformly on* $\theta \in \Theta$. *That is, with any given* x, $f^{M,h,a}(x;\theta)$ *converges to* $f(x;\theta)$ *uniformly for* $\theta \in \Theta$ *when* $Mh \to \infty$ *and* $h \to 0$.

Based on the Theorem 6, we have following asymptotic property between MLE and AMLE.

Theorem 7. *Suppose* $\hat{\theta}_N$ *is the unique MLE defined as* $\arg\max_\theta l(\theta, X)$ *with large enough* N. *We assume parameter space* Θ *is compact and likelihood function is continuous at* θ. *Fixing large enough sample size* N, *under A class regularity conditions*

$$\hat{\theta}^{M,h,a}_N \xrightarrow{p} \hat{\theta}_N, \tag{3.5}$$

when $h \to 0$ *and* $Mh \to \infty$ *with fixed* a.

Remark 4. In fact, the continuity can be generalized to be upper-continuous. The uniqueness of MLE $\hat{\theta}_N$ combined with compact parameter space and continuity can be replaced by a more general condition: $\hat{\theta}_N$ is

a well-separated point of the maximum (See the remark of Lemma 24.3 in Appendices).

Theorem 8. *Suppose* $\hat{\theta}_N$ *is the unique MLE defined as* $\arg\max_\theta l(\theta, X)$ *with large enough* N*. We introduce B class of regularity conditions in appendix B.1.1, to guarantee the asymptotic property of MLE. Under A class of regularity condition and B class of regularity conditions , there exists* $M(N)$ *and* $h(N)$ *with respect to* N *and for AMLE* $\hat{\theta}_N^{M,h,a}$*,* $\hat{\theta}_N^{M(N),h(N),a} \xrightarrow{p} \theta_0$ *with fixed* a *when* $N \to \infty$*. Furthermore,*

$$\sqrt{N}(\hat{\theta}_N^{M(N),h(N),a} - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0)),$$

where $I(\theta_0)$ *is the fisher information matrix.*

Remark 5. The first 4 regularity conditions in B class is responsible for consistency of MLE , while, the last 6 regularity conditions in B class is specific for the asymptotic normality and asymptotic efficiency of MLE. Combined with A class regularity conditions, AMLE $\hat{\theta}_N^{M,h,a}$ will be consistent and asymptotic efficient.

Remark 6. In theorem 8, we assume that the uniqueness of MLE $\hat{\theta}_N$ with large enough N. In fact, we only need to guarantee the unique maximum of the likelihood function in the neighborhood of true parameter θ_0 considering the consistency of MLE $\hat{\theta}_N$ (Lemme 24.7). To show this uniqueness property, we can use Corollary 1 provided by Little et al. (2010). That is, if $rank[(\frac{\partial^2 L(\theta_0;X)}{\partial\theta_i\partial\theta_j})_{i,j=1}^p] = p$, there will be at most one maximum in the neighborhood of true parameter.

All in all, under certain regularity conditions, consistency, asymptotic normality and asymptotic efficiency can be reached by the AMLE by maximizing approximated-log-likelihood $l^{M,h,a}(\theta; X)$.

3.2.2 Likelihood ratio test and model selection

Other than parameter estimation, another important part within MLE framework is likelihood ratio test (LRT). It is a standard method for hypothesis testing. Under mild regularity conditions, test statistics has an asymptotic χ^2 distribution (Wilks (1938)).

In this subsection, we give a nested likelihood ratio test based on our approximated likelihood by inverting characteristic functions.

Suppose the hypotheses are

$$H_0 : g_1(\theta) = 0, \ldots, g_q(\theta) = 0, \qquad 1 \le q < p.$$

against

$$H_a : \exists k, \quad g_k(\theta) \neq 0, \qquad 1 \le k \le q.$$

Theorem 9. *Let $\{X\}$ be i.i.d data. Assume that we can treat H_0 as $\theta_1 = \theta_1^0, \cdot, \theta_r = \theta_r^0$ via reparametrization. When H_0 holds, under A and B classes of regularity conditions applied to both parameter space Θ and its subspace Θ_0 of parameter vector $(\theta_{r+1}, \ldots, \theta_p)$, there exists $M_0(N)$, $h_0(N), M_1(N)$ and $h_1(N)$ with respect to N so that*

$$-2\log\Lambda \xrightarrow{d} \chi_k^2$$

where

$$\Lambda = \frac{L_0^{M_0(N),h_0(N),a}(\hat{\theta}_N^{M_0(N),h_0(N),a}; X)}{L_1^{M_1(N),h_1(N),a}(\hat{\theta}_N^{M_1(N),h_1(N),a}; X)}$$

with fixed a. $\hat{\theta}_N^{M_0(N),h_0(N),a} \in \Theta_0$ is the AMLE of the approximated likelihood $L_0^{M_0(N),h_0(N),a}(\theta; X)$ under H_0. $\hat{\theta}_N^{M_1(N),h_1(N),a} \in \Theta_1$ is the AMLE of the approximated likelihood $L_1^{M_1(N),h_1(N),a}(\theta; X)$ in parameter space Θ. The dimension of parameter space is p for Θ and $(p-q)$ for Θ_0. χ_k^2 is a Chi-squared distribution with degree freedom k.

When we perform likelihood ratio test, we need to be careful with the regularity condition B.5, which θ_0 should lies in the interior of compact parameter space Θ. In several cases, this regularity condition is not satisfied. For example, if we want to compare Black-Shores-Merton model with Jump diffusion model, we might need to test jump intensity $\lambda = 0$. However, 0 is the boundary of the parameter space of λ. In this case, test statistics will not asymptotically χ^2 distributed. For another example, if we want to check diffusion term is necessary or not in an infinite activity Lévy process, we need to test volatility term $\sigma = 0$. However, 0 is also the boundary of the parameter space of σ. In this case, the asymptotic distribution of our test statistics is also not χ^2 distributed. Theoretical result of likelihood ratio tests on non-regular condition can be found in Self and Liang (1987) following previous work Chernoff (1954); Feder (1968); Moran (1971); Chant (1974). Sinha et al. (2007); Kopylev and Sinha (2011) derive the asymptotic distributions based on some special cases. A explicit review of this topic can refer toKopylev (2012). Generally speaking, asymptotic distribution of test statistics could be mixture of χ^2 distribution or even more complicated distributions. Instead of deriving explicit asymptotic distribution of test statistics, bootstrapping approaches are also suggested by many authors. For example, Chan et al. (2009) compare different financial models by performing empirical likelihood test utilizing parametric bootstrapping.

For more non-nest likelihood ratio based test for model comparison like Vuong's test (Vuong (1989)), information based model selection technique, Akaike information criterion (AIC), Bayesian information criterion (Schwarz (1978)), approximated likelihood could also be performed. For example, the approximated

AIC and approximated BIC are in the form below.

$$AIC^{M,h} = -2\log L^{M,h,a}(\hat{\theta}_N^{M,h,a}; X) + 2p$$
$$BIC^{M,h} = -2\log L^{M,h,a}(\hat{\theta}_N^{M,h,a}; X) + p\log(N)$$

where N is the sample size and p is the dimension of parameter space Θ. Generally with fine setting of M and h, $AIC^{M,h}$ and $BIC^{M,h}$ can be as close as AIC and BIC as we desire. As penalty based model selection methods discouraging over-fitting, a smaller value of them will indicate a preference of the model.

3.2.3 Extension to Markov processes

Now, we consider the case that data $X = \{X_0, X_1, X_2, \ldots, X_N\}$ are not i.i.d but rather discrete trajectory observations of Markov processes with stationary transition measures

$$F(\xi, A; \theta_0) = P(x_{j+1} \in A | x_j = \xi; \theta_0), \tag{3.6}$$

where θ_0 is the true parameter vector. Suppose the characteristic function associated with $F(\xi, A; \theta)$ is available for any given $\xi \in \mathcal{X} \subset \mathbf{R}$:

$$\phi(u, \xi; \theta) = \int_{-\infty}^{\infty} \exp(iux) F(\xi, dx; \theta),$$

where $\mathcal{X}$ is the space of the value of $\{X_i\}$. Similarly to the I.I.D case, when $\phi(u, \xi; \theta)$ is absolutely integral with every $\xi \in \mathcal{X}$, we have bounded and uniformly continuous transition density given ξ:

$$f(x|\xi; \theta) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-iux)\phi(u, \xi; \theta) du. \tag{3.7}$$

The approximated transition density with fixed $M \in \mathbf{Z}^+$,$h \in \mathbf{R}^+$ and $a \in \mathbf{R}$ is

$$f^{M,h,a}(x|\xi; \theta) = \frac{1}{2\pi} \sum_{m=-M}^{M} e^{-ix(mh+ia)} \phi(mh + ia, \xi; \theta) h. \tag{3.8}$$

If we define the likelihood function $L(\theta; X_j, X_{j-1}) = f(X_j | X_{j-1}; \theta)$ and the log-likelihood function of $X = \{X_0, \ldots, X_N\}$ is

$$l(\theta; X) := \frac{1}{N}(\sum_{j=1}^{N} \log(f(X_j | X_{j-1}; \theta))) = \frac{1}{N}(\sum_{j=1}^{N} \log(\frac{1}{2\pi} \int_R e^{-iX_j y} \phi(y, X_{j-1}; \theta) dy)).$$

Correspondingly, the approximated-log-likelihood is defined as

$$\begin{aligned} l^{M,h,a}(\theta;X) :=& \Re(\frac{1}{N}(\sum_{j=1}^{N}\log(f^{M,h,a}(X_j|X_{j-1},\theta)))) \\ =& \Re(\frac{1}{N}(\sum_{j=1}^{N}\log(\frac{1}{2\pi}\sum_{m=-M}^{M}e^{-iX_j(mh+ia)}\phi(mh+ia,X_{j-1};\theta)h))). \end{aligned} \tag{3.9}$$

Definitely, the AMLE $\hat{\theta}_N^{M,h,a}$ in this case is

$$\hat{\theta}_N^{M,h,a} = \arg\max_{\theta\in\Theta} l^{M,h,a}(\theta;X)$$

Remark 7. We don't add the initial distribution of X_0 because the asymptotic property can be still obtained without it. In practice, initial distribution might have a undeniable effect especially in small sample size. In that case, we can incorporate the initial distribution in our likelihood.

Similar to the case of Lévy processes, the following regularity conditions and Theorem 10 which can control the difference between approximated-log-livelihood and real likelihood uniformly on $\theta\in\Theta$.

C.1 For any $\theta\in\Theta$ and $\xi\in\mathcal{X}$, the characteristic function$\phi(u,\xi;\theta)$ is analytic in $\mathcal{D}_{(d_-,d_+)}$ where $\mathcal{D}_{(d_-,d_+)} = \{z\in C:\Im(z)\in(d_-,d_+)\}$, $-\infty<d_-<0<d_+<\infty$ and d_-, d_+ might depend on ξ; but not θ. $\Re(z)$ is the real part and $\Im(z)$ is the imaginary part of z.

C.2 Given $\xi\in\mathcal{X}$, for any $\theta\in\Theta$, $\int_{d-}^{d+}|\phi(x+iy,\xi;\theta)|dy\to 0$ when $x\to\pm\infty$
$||\phi||^{\pm} := \lim_{\epsilon\to 0+}\int_R|\phi(x+i(d_\pm\mp\epsilon),\xi;\theta)|dx<+\infty$ uniformly on $\theta\in\Theta$.

C.3 The parameter in Equation (3.8), a, is a real value and $a\in(d_-,d_+)$

C.4 For any $\xi\in\mathcal{X}$, $|\phi(x+ia,\xi;\theta)\le k|x|^n\exp(-c|x|^\nu)$, $x\in\mathbf{R}$ with $a\in(d_-,d_+)$ for some $\kappa>0,\nu>0,c>0,n\in\mathbf{R}$ or $\kappa>0,\nu>0,c=0,n<-1$. κ,ν,c and n might depends on ξ and a, but not related to the parameter $\theta\in\Theta$.

C.5 $Mh\ge(n/c\nu)^{1/\nu}1_{c>0,n>0}$ for n,c,ν defined above.

Theorem 10. *Given $\xi\in\mathcal{X}$, if the error of approximation of transition density function*

$$E_{h,M}^F(\phi,\xi,a)(x) = \frac{1}{2\pi}\int_R e^{-ixy}\phi(y,\xi;\theta)dy - \frac{1}{2\pi}\sum_{m=-M}^{M}e^{-ix(mh+ia)}\phi(mh+ia,\xi;\theta)h,$$

under class C of regularity condition, we have the bound of error

$$|E^F_{h,M}(\phi,a)(x)| \leq \frac{e^{-2\pi(a-d_-)/h}}{2\pi(1-e^{-2\pi(a-d_)/h})}e^{xd_-}||\phi||^- + \frac{e^{-2\pi(d_+-a)/h}}{2\pi(1-e^{-2\pi(d_+-a)/h})}e^{xd_+}||\phi||^+ + T_{Mh}, \tag{3.10}$$

where $T_{Mh} = \frac{ke^{ax}}{|n+1|\pi}(Mh)^{n+1}$ *if* $c = 0, n < -1$, *and* $T_{Mh} = \frac{ke^{ax}}{\pi\nu c^{(n+1)/\nu}}\Gamma(\frac{n+1}{\nu}, c(Mh)^\nu)$ *if* $c > 0$. *Incomplete Gamma function* $\Gamma(s,b) = \int_b^\infty e^{-t}t^{s-1}dt$. *Moreover, let* $Mh \to \infty$ *and* $h \to 0$, *then, bound of error* $E^F_{h,M}(\phi,a)(x)$ *will decay to zero uniformly for* $\theta \in \Theta$.

We also list regularity conditions for the weak law of large number (WLLN) theorem and central limit (CLT) theorem for Markov processes which are proposed by Billingsley (1961). The Markov version of WLLN and CLT will be key ingredients to show the consistency and asymptotic normality of AMLE. It is worth to mention that Billingsley (1961) developed consistency and asymptotic normality for Markov processes (Theorem 2.1) in a 'local' version (the parameter space is around the true value) with different definition of MLE compared with our work.

D.1 There exists an unique stationary distribution $p(\cdot;\theta_0)$. That is, for true parameter $\theta_0 \in \Theta$, there exists an unique probability measure $p(\cdot;\theta_0)$ such that $p(A;\theta_0) = \int_{\mathcal{X}} p(d\xi;\theta_0)F(\xi,A;\theta_0)$ where $A \in \mathcal{F}_{\mathcal{X}}$.

D.2 Given $x \in \mathcal{X}$ and for true parameter $\theta_0 \in \Theta$, the transition probability measure $F(\xi,\cdot;\theta_0)$ is absolutely continuous with respect to stationary probability measure $p(\cdot;\theta_0)$: $F(\xi,\cdot;\theta_0) \ll p(\cdot;\theta_0)$.

D.3 Suppose true parameter $\theta_0 = \{\theta_1^0,\ldots,\theta_p^0\}$. $E_{\theta_0}[|\frac{\partial l(\theta_0;x_0,x_1)}{\partial\theta_j^0}|^2] \leq \infty$ for $j = 1,2,\ldots,p$. $E_{\theta_0}[\cdot]$ denotes an expected value computed under the assumption that initial distribution is the stationary distribution (We just use this assumption when we calculate the expected value here. In fact, the initial distribution even does not appear in our likelihood function).

Under Markov processes framework, $E_{\theta_0}[\cdot]$ below denote an expected value computed under the assumption that the initial distribution is the stationary distribution. This is just a computational device. We don't assume the initial distribution is in fact the stationary distribution. Let's list the WLLN and CLT for Markov processes proposed by Billingsley (1961).

Lemma 10.1 (WLLN and CLT for Markov processes)**.** *Under the regularity conditions of D.1 and D.2, for any* $\theta_0 \in \Theta$, *no matter what the initial distribution is, if* $E_{\theta_0}|(k(X_0,X_1))| < \infty$, *then,*

$$\lim_{N\to\infty}\frac{1}{N}\sum_{j=1}^{N}k(X_{j-1},X_j) = E_{\theta_0}[k(X_0,X_1)]$$

with probability one. Moreover, under all the regularity conditions of D class, we additionally have

$$\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\varphi(X_j, X_{j-1};\theta_0) \xrightarrow{d} N(0, I(\theta_0)),$$

where $\varphi(X_j, X_{j-1};\theta) = \frac{\partial}{\partial\theta} f(X_j|X_{j-1};\theta)$.

Similar to the i.i.d case, we have correspondent theorems about asymptotic properties of AMLE based on WLLN and CLT for Markov processes. E class regularity conditions are listed in the Appendix B.1.2 to guarantee the asymptotic property of MLE for Markov processes.

Theorem 11. *Suppose* $\hat{\theta}_N$ *is the MLE defined as* $\arg\max_\theta l(\theta, X)$ *which is unique. Fix sample size* N*, under* C *class regularity conditions,* $E.2$ *and* $E.3$*,*

$$\hat{\theta}_N^{M,h,a} \xrightarrow{p} \hat{\theta}_N \tag{3.11}$$

when $h \to 0$ *and* $Mh \to \infty$ *with fixed* a*.*

Theorem 12. *Suppose MLE* θ_N *is unique given likelihood function for every large enough* N*. Under C,D,E class of regularity condition, there exists M(n) and h(n) with respect to n and for AMLE* $\hat{\theta}_N^{M,h,a}$*,* $\hat{\theta}_N^{M(n),h(n),a} \xrightarrow{a.s} \theta_0$ *with fixed a when* $N \to \infty$*.Furthermore,*

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

The first 4 regularity conditions in B class is responsible for consistency, while, the last 6 regularity conditions in C class is specific for the asymptotic normality and asymptotic efficiency of AMLE $\hat{\theta}_N^{M,h,a}$*.*

Similarly to the Theorem 9, we also have likelihood ratio tests for Markov process with similar proof. Based on it, AIC and BIC might be available as well. We do not list all of them here.

3.3 Lévy process based models in finance

We consider several Lévy processes and one Lévy driven Ornastein-Uhlenbeck process. For Lévy processes, There are two types of Lévy processes, finite activity Lévy processes and infinite activity Lévy processes. Finite activity Lévy processes allow only finitely many jumps in any given time interval, while, infinite activity Lévy processes include infinitely many jumps in any given time interval. Also, it can be shown that finite activity Lévy processes are compound Poisson type processes typically including Merton's jump-

diffusion model (Merton (1976)) and Kou's jump-diffusion model (Kou (2002)). Compared with compound Poisson type processes, pure jump Lévy processes with infinite activity might have better representations of stock price dynamics (Geman (2002)). Thus, we also consider several commonly used financial models with infinite activities including normal inverse Gaussian model (Barndorff-Nielsen (1997)) and CGMY model (Carr et al. (2002)). In addition to Lévy processes, we applied non-i.i.d data to one Lévy driven Ornastein-Uhlenbeck process, IG-OU process to show validity of the extension described in Chapter 3.2.3.

3.3.1 Lévy processes

Lévy processes are commonly used in Finance due to its flexibility to model heavy tails and skewness of financial time series. In our case, we assume the equity value series are $S_t = \exp(Y_t)$ and $\{Y_t\}$ is a Lévy process. Suppose data is observed over equally spaced timestamps $(0, \delta, 2\delta, \ldots, n\delta)$. Then, total time interval will be $T = n\delta$ and sample size will be $n + 1$. In practice, we model logarithm of equity returns $\{X_t = Y_{t\delta} - Y_{(t-1)\delta}\}_1^n$. They will follow the same distribution F due to the fact that Lévy processes have independent stationary increments.

One important property we need to utilize later is Lévy-Khintchine formula. Suppose X_t is a Lévy process described above, the characteristic function $\phi(u;\theta)$ of X_t has the form

$$\phi(u) = \exp(\delta(iua - \frac{bu^2}{2} + \int_R (\exp(iux) - 1 - iux1_{|x|\leq 1})J(dx))), \tag{3.12}$$

where (a, b, J) is called Lévy triplet which fully determines a Lévy process. Lévy triplet includes the drift parameter $a \in R$, the diffusion component $\sigma \geq 0$ and Lévy measure $J(dx)$ satisfying $J(\{0\}) = 0$ and $\int \min(1, x^2)J(dx) < \infty$. Through Lévy-Khintchine formula, implicit characteristic function is more likely available compared with density functions for Lévy processes. Thus, if the logarithm of equity value series $\{Y_t\}_1^n$ follow Lévy processes, then the Levy increments $\{X_t\}_1^N$ will be i.i.d data with characteristic function described in (3.12).

In this chapter, we mainly focus on several typical Lévy processes based models due to their availability of the characteristic function. Introduction of those models are below and moments information about those models are in Table 3.1. We will use them later in the simulation study and case study.

Table 3.1: Moments of returns

Merton	Kou	NIG	CGMY
Mean			
$(\mu+\lambda\mu_j)\delta$	$(\mu+\frac{\lambda p}{\eta_u}-\frac{\lambda(1-p)}{\eta_d})\delta$	$(\mu+\frac{\beta\lambda}{(\alpha^2-\beta^2)^{0.5}})\delta$	$(\mu+C\Gamma(1-Y)(\mathcal{M}^{Y-1}-G^{Y-1}))\delta$
Variance			
$(\sigma^2+\lambda(\mu_j^2+\sigma_j^2))\delta$	$(\sigma^2+2\frac{\lambda p}{\eta_u^2}+2\frac{\lambda(1-p)}{\eta_d^2})\delta$	$\frac{\lambda\alpha^2\delta}{(\alpha^2-\beta^2)^{1.5}}$	$C\Gamma(2-Y)(\mathcal{M}^{Y-2}+G^{Y-2})\delta$
Skewness			
$\frac{\lambda\mu_j(\mu_j^2+3\sigma_j^2)}{(\sigma^2+\lambda(\sigma_j^2+\mu_j^2))^{1.5}\delta^{0.5}}$	$\frac{6(\frac{\lambda p}{\eta_u^3}-\frac{\lambda(1-p)}{\eta_d^3})}{(\sigma^2+2\frac{\lambda p}{\eta_u^2}+2\frac{\lambda(1-p)}{\eta_d^2})^{1.5}\delta^{0.5}}$	$\frac{3\beta}{\alpha\lambda^{0.5}(\alpha^2-\beta^2)^{0.25}\delta^{0.5}}$	$\frac{\Gamma(3-Y)(\mathcal{M}^{Y-3}-G^{Y-3})}{C^{0.5}(\Gamma(2-Y)(\mathcal{M}^{Y-2}+G^{Y-2}))^{1.5}\delta^{0.5}}$
Kurtosis			
$\frac{\lambda(\mu_j^4+6\mu_j^2\sigma_j^2+3\sigma_j^4)}{(\sigma^2+\lambda(\mu_j^2+\sigma_j^2))^2\delta}$	$\frac{24(\frac{\lambda p}{\eta_u^4}+\frac{\lambda(1-p)}{\eta_d^4})}{(\sigma^2+2\frac{\lambda p}{\eta_u^2}+2\frac{\lambda(1-p)}{\eta_d^2})^2\delta}$	$\frac{3(1+4\frac{\beta^2}{\alpha^2})}{\lambda(\alpha^2-\beta^2)^{0.5}\delta}$	$\frac{\Gamma(4-Y)(\mathcal{M}^{Y-4}+G^{Y-4})}{C(\Gamma(2-Y)(\mathcal{M}^{Y-2}+G^{Y-2}))^2\delta}$

Merton's jump-diffusion model

In Merton's jump-diffusion model, the observed stock price, S_t, satisfies the following equation:

$$X_t \equiv \log(S_{(t+1)\delta}/S_{t\delta}) = \mu\delta + \sigma\sqrt{\delta}Z + \sum_{i=N_t+1}^{N_{t+1}} Z_i, \tag{3.13}$$

where δ is the evenly spaced time interval of observed data; μ is drift; σ is volatility; Z is standard $N(0,1)$; N_t is a Poisson process with intensity λ; $\{Z_i\}$ are jump sizes following i.i.d $f_z(x) \sim N(\mu_j, \sigma_j^2)$. That is, in Merton's jump-diffusion model, jumps occur according to Poisson process N_t with jump sizes $\{Z_i\}$.
The Lévy triplet of this model is $(\mu, \sigma^2, J(dx) = \lambda f_z(x)dx)$. Although, the density of X_t has a complex form (Tankov (2003)), its characteristic function, $\phi(u;\theta)$, has a simple form through Lévy-Khintchine formula:

$$\phi(u;\theta) = \exp(\delta(iu\mu - \sigma^2u^2/2 + \lambda(\exp(i\mu_j u - \sigma_j^2u^2/2) - 1))).$$

For regularity conditions, we can show that d_+ and d_- satisfying A.1 could be arbitrary positive number and negative number. A.3 can be shown directly. A.4 can be satisfied when $c = \sigma^2/2$ and $\nu = 2$ if we allow c and ν can depend on parameter θ. To make sure c is not related parameter θ, we assume the parameter space Θ of Merton's jump model is $-\infty < L_\mu \le \mu \le U_\mu < \infty$, $0 < L_\sigma \le \sigma \le U_\sigma < \infty$, $0 < L_\lambda \le \lambda \le U_\lambda < \infty$, $-\infty < L_{\mu_j} \le \mu_j \le U_{\mu_j} < \infty$, $0 < L_{\sigma_j} \le \sigma_j \le U_{\sigma_j} < \infty$, and true parameter $\theta_0 \in \Theta$. Then, Θ is a compact parameter space and we can let c to be $L_\sigma^2/2$ which does not depend on parameter σ. In this case we do not consider the case when $\sigma = 0$. The likelihood of Merton's jump model can be regarded as a mixture distribution and its value can reach infinity on some points of which $\sigma = 0$ (See Kiefer (1978)). Furthermore, when $\sigma = 0$, the model degenerate to a compound Poisson process which is also not our primary interest.

To simulate Merton's jump-diffusion model in the simulation study, we can utilize the method described

in Glasserman (2003).

Kou's jump-diffusion model

Compared with Merton's jump-diffusion model, Kou's jump-diffusion model allows the distribution of jump sizes $\{Z_i\}$ to be asymmetric. To be specific, the observed stock price, S_t, satisfying the following equation:

$$X_t \equiv \log(S_{(t+1)\delta}/S_{t\delta}) = \mu\delta + \sigma\sqrt{\delta}Z + \sum_{i=N_t+1}^{N_{t+1}} Z_i, \tag{3.14}$$

where the jump size here, $\{Z_i\}$, follows i.i.d. double exponential distribution, which has the density

$$f_z(x) = p\eta_u \exp(-\eta_u x 1_{\{x>0\}}) + (1-p)\eta_d \exp(\eta_d x 1_{\{x<0\}}).$$

p is the positive jump probability; $1/\eta_u$ is the mean positive jump size; $1/\eta_d$ is the mean negative jump size. The Lévy triplet of this model is $(\mu, \sigma^2, J(dx) = \lambda f_z(x)dx)$. The density of X_t has a complex form (Ramezani and Zeng (2007)), while, its characteristic function, $\phi(u;\theta)$, has a simple form through Lévy-Khintchine formula:

$$\phi(u;\theta) = \exp(\delta(iu\mu - \sigma^2u^2/2 - \lambda(1 - \frac{p\eta_u}{\eta_u - iu} - \frac{(1-p)\eta_d}{\eta_d + iu}))).$$

For regularity conditions, $d_+ \in (0, \eta_d)$ and $d_- \in (-\eta_u, 0)$ satisfy A.1 and A.3, which can be shown directly. A.4 can be satisfied when $c = \sigma^2/2$ and $\nu = 2$, if we allow c and ν can depend on parameter θ. We assume the similar parameter space to the Merton's jump model: $-\infty < L_\mu \le \mu \le U_\mu < \infty$, $0 < L_\sigma \le \sigma \le U_\sigma < \infty$, $0 < L_\lambda \le \lambda \le U_\lambda < \infty$, $0 < L_p \le p \le U_p < 1$, $0 < L_{\eta_u} \le \eta_u \le U_{\eta_u} < \infty$, $0 < L_{\eta_d} \le \eta_d \le U_{\eta_d} < \infty$, and the true parameter $\theta_0 \in \Theta$. Then, Θ is a compact parameter space. We can let $d_+ = L_{\eta_d}/2$ and $d_- = -L_{\eta_u}/2$. This setting is generally practical because η_u and η_d are positive which cannot be zero exactly. We can let c to be $L_\sigma^2/2$ which does not depend on parameter σ. In this case we do not consider the case when $\sigma = 0$. The reason for it is same as Merton's jump-diffusion model.

To simulate it in simulation study, we also utilize the method described in Glasserman (2003).

Normal inverse Gaussian model

Normal inverse Gaussian (NIG) model belongs to a more general class of Lévy processes, generalized hyperbolic model (Eberlein et al. (1998)). It is a infinite activity Lévy process. It can be characterized by

$$X_t \equiv \log(S_{(t+1)\delta}/S_{t\delta}) = \mu\delta + \beta z_\delta + \lambda W_{z_\delta}, \tag{3.15}$$

where z_δ is the first time when a Brownian motion with drift γ reaches the positive level δ. The density of z_δ is inverse Gaussian (IG) distribution. W_{z_δ} is a Brownian motion of which the calendar time is a random time z_δ.

Set $\alpha = \sqrt{\beta^2 + \gamma^2}$, we have the Lévy triplet $(\mu, 0, J(dx) = f(x)dx)$, where

$$f(x) = \frac{\lambda\alpha}{\pi|x|} \exp(\beta x) K_1(\alpha|x|).$$

$K_n(x)$ is the modified Bessel function of the second kind with order n. Then, the probability density function of X_t contains Bessel function, but the characteristic function is in a simple form

$$\phi(u;\theta) = \exp(\delta(iu\mu - \lambda(\sqrt{(\alpha^2 - (\beta + iu)^2)} - \sqrt{\alpha^2 - \beta^2}))).$$

For regularity conditions, $d_+ \in (0, \beta + \alpha]$ and $d_- \in [\beta - \alpha, 0)$ can make the characteristic function analytic, also satisfying A.3. A.4 can be satisfied when $c = \lambda$ and $\nu = 1$ if we allow c and ν can depend on parameter θ. We assume the parameter space of the NIG model: $-\infty < L_\mu \le \mu \le U_\mu < \infty$, $\infty < L_\alpha \le \alpha \le U_\alpha < \infty$, $\infty < L_\beta \le \beta \le U_\beta < \infty$, $0 < L_{\alpha-|\beta|} \le \alpha - |\beta|$, $0 < L_\lambda \le \lambda \le U_\lambda \le \infty$, and true parameter $\theta_0 \in \Theta$. Then, Θ is a compact parameter space. We can let $d_+ = L_{L(\alpha-\beta)}$ and $d_- = -L_{\alpha-\beta}$. This setting is generally practical because $\alpha - |\beta|$ is required to be positive by NIG model. We can let c to be L_λ, which does not depend on parameter λ.

To simulate NIG process X_t in simulation study, we refer the method in Rydberg (1997).

CGMY model

Unlike finite-activity jump processes such as Merton's jump-diffusion process, CGMY could be either finite activity or infinite activity processes. That is, it could have infinitely many jumps in any finite time interval. It came out as the generalization of Variance Gamma (Madan and Seneta (1987)) Lévy density with parameters C, G, $\mathcal{M}$ and Y. Its Lévy triplet is $(\mu, 0, J(dx) = f(x)dx)$ and $f(x)$ is:

$$f(x) = \begin{cases} C\frac{\exp(-\mathcal{M}x)}{x^{1+Y}} & x > 0 \\ C\frac{\exp(-Gx)}{|x|^{1+Y}} & x < 0. \end{cases} \tag{3.16}$$

where $C > 0, G \ge 0, \mathcal{M} \ge 0, Y < 2$.

In addition, when $Y < 0$, CGMY is a finite activity process. It could be regarded as a compound Poisson process. When $0 < Y < 2$, the process has infinite activities. The case $Y = 0$ degenerates to a variance

gamma process. Here, we only consider infinite activity CGMY process with $0 < Y < 2$.

The density function of CGMY process $\{X_t\}$ is unknown for us. However, the characteristic function is available for us.

For $Y \in (0,1) \cup (1,2)$:

$$\phi(u;\theta) = \exp(\delta(i\mu u + C\Gamma(-Y)[(\mathcal{M} - iu)^Y - \mathcal{M}^Y + (G + iu)^Y - G^Y])), \tag{3.17}$$

and for $Y = 1$:

$$\phi(u;\theta) = \exp(\delta(i\mu u + C((\mathcal{M} - iu)\log(1 - iu/\mathcal{M}) + (G + iu)\log(1 + iu/G) - iu(\log(\mathcal{M}) - \log(G))))).$$

It is not hard to show that the characteristic function of CGMY model is smooth with respect to parameters C, G, $\mathcal{M}$ and Y, where $C > 0, G \geq 0, \mathcal{M} \geq 0, 0 < Y < 2$.

For regularity conditions, $d_+ \in (0, G)$ and $d_- \in (-\mathcal{M}, 0)$ make the characteristic function analytic. A.3 can be shown directly. A.4 can be satisfied when $c = f_c(C,\ Y)$ and $\nu = Y$, where $f_c(C,\ Y) = \delta C|\Gamma(-Y)\cos(\pi Y/2)|$ when $Y \in (0,1) \cup (1,2))$ and $f_c(C,\ Y) = \delta\frac{\pi}{2}C$ when $Y = 1$, if we allow c and ν can depend on parameter θ. We assume the parameter space of the CGMY model: $-\infty < L_\mu \leq \mu \leq U_\mu < \infty$, $0 < L_C \leq C \leq U_C < \infty$, $0 < L_G \leq G \leq U_G < \infty$, $0 < L_\mathcal{M} \leq \mathcal{M} \leq U_\mathcal{M} < \infty$, $0 < L_Y \leq Y \leq U_Y < 2$, and true parameter $\theta_0 \in \Theta$. Then, Θ is a compact parameter space. To keep A.1 hold, we can let $d_+ = L_G/2$ and $d_- = -L_\mathcal{M}/2$. This setting is generally practical because $\mathcal{M}$ and G are required to be positive by CGMY model. To keep A.4 hold, $c = \inf_{C,\ Y} f_c(C,\ Y)$ (considering $f_c(C,\ Y)$ is continuous and C,Y are in compact parameter space) and $\nu = L_Y$ which are not related to parameters any more.

We can simulate CGMY process by inverting characteristic function introduced in Chen et al. (2012).

3.3.2 Lévy driven Ornstein-Uhlenbeck processes

Lévy driven Ornstein-Uhlenbeck processes (Lévy driven OU processes) is a sub-class of continuous-time Markov processes. Such processes are popular to model stochastic volatility and interest rate. (see Barndorff-Nielsen and Shephard (2001)).

The process $X = \{X_t\}_{t\geq 0}$ is called a Lévy driven OU process when

$$dX_t = -\lambda X_t dt + dZ_{\lambda t}, \qquad X_0 > 0,$$

where λ is the drift parameter and $\lambda > 0$. $\{Z_t\}_{t\geq 0}$ is called background driving Lévy process (BDLP) with

Lévy triplet $(a, 0, J)$. BDLP is required to have no Brownian part, a non-negative drift and only positive increments.

As mentioned in section 2 of Shiga (1990), the process is time-homogeneous satisfying the equation (3.6) in Chapter 3.2.3.

What's more, there exists an unique stationary distribution π such that $X_t \xrightarrow{d} \pi$ (See Sato (1999) Theorem 17.5 and Corollary 17.9) when

$$\int_{|x|>2} \log(|x|) J(dx) < \infty, \tag{3.18}$$

and π is the distribution implied by Lévy triplet (a^*, b^*, J^*) where

$$a^* = \frac{a_0}{2} \quad b^* = \int_{|x|>1} \frac{x}{|x|} J(dx),$$

and

$$g(x) = \begin{cases} J((x, \infty)) & x > 0 \\ J((-\infty, x)) & x < 0. \end{cases}$$

π is not related to λ due to the deliberate setting of subscript λt of BDLP. Then, we find the Equation (3.18) is equivalent to the Regularity condition D.1.

Moreover, we suppose the equity value series are $\{P_t\}_{t\geq 0}$. If our discrete-time observations $\{X_t = P_{t\delta}\}$ where δ is fixed time interval, the conditional characteristic function of the transition density $f(X_t|X_{t-1}, \theta)$ is available (See Sato (1999) lemma 17.1):

$$\phi(u, X_{t-1}; \theta) = \exp(iu \exp(-\lambda\delta) X_{t-1} + \lambda \int_0^\delta g(\exp(\lambda(z-\delta))u) dz), \tag{3.19}$$

where $g(x)$ is the cumulation of $Z(1)$ which is $\log E(ixz_1)$. Thus, we have shown that with Equation (3.18) holding, Lévy driven OU process is time-homogeneous with explicit formula of condition characteristic function (3.19), and has unique stationary distribution. These prosperities satisfy the assumptions for Markov processes in Chapter 3.2.3.

One way to construct Lévy driven OU processes is to specify the stationary distribution for the process. Specifically, we define the marginal law D, when y_t follows distribution D for arbitrary $t > 0$, if the initial distribution (the distribution of y_0) follows distribution D. Then, the OU process with marginal law D is called D-OU process. Barndorff-Nielsen and Shephard (2001) showed that D-OU processes exists if and only

if D is self-decomposable. Moreover, the relations between law D and BDLP is below:

$$\begin{aligned} g(x) &= u\frac{dg_D(x)}{dx}, \\ f_z(x) &= -u(x) - xu'(x), \end{aligned} \tag{3.20}$$

where $g(x)$ is the cumulant function of BDLP Z(1) (needed in (3.19)); $g_d(x)$ is the cumulant function of the self-decomposable law D; $f_z(x)$ is the levy density, if exists, of Z(1); $u(x)$ is the levy density, if exists, of law D.

Remark 8. For IG-OU process, we assume the OU process follows inverse Gaussian (IG) law (IG distribution is self-decomposable). We have the Lévy density of IG(a,b):

$$u(x) = (2\pi)^{-1/2} a x^{-3/2} \exp(-\frac{1}{2}b^2 x), \qquad x > 0.$$

Then, through Equation (3.20), we have levy density of correspondent BDLP:

$$f_z(x) = \frac{1}{2}(2\pi)^{-1/2} a x^{-1/2}(x^{-1} + b^2) \exp(-\frac{1}{2}b^2 x), \qquad x > 0.$$

And the cumulant function of BDLP Z(1):

$$g(x) = -iab^{-1}x(1 + 2xb^{-2})^{-1/2}. \tag{3.21}$$

We will use IG-OU process in our simulation study in Chapter 3.3.2.

Remark 9. Class D is the requirements for the Markov processes. Specifically, for Lévy driven OU processes, Equation (3.18) is equivalent to D.1.

To check D.2, we might need to prove absolutely continuity. One idea to prove it is that, for typical D-OU processes, we know the stationary distribution is D distribution. In most cases, it is not hard to show that density of D is positive at every feasible point. Then, D and Lebesgue measure are equivalent. To show the transition probability measure $F(\xi, \cdot; \theta_0)$ is absolutely continuous with respect to D, we just need to show transition probability measure is absolutely continuous with respect to Lebesgue measure. By showing the conditional characteristic function is absolutely integral (or its binomial or exponential tails required by C.4), absolute continuity with respect to Lebesgue measure is guaranteed by Fourier inversion theorem.

To check D.3, we generally can prove that $|\frac{\partial l(\theta_0; x_0, x_1)}{\partial \theta_j^0}|$ is bounded on the support $\mathcal{X} \times \mathcal{X}$ in most cases via Fourier inversion theorem.

Remark 10. Another models might satisfy our setting of Markov processes is affine processes. A sub-class of affine processes is called canonical affine processes. They are still time-homogeneous Markov processes and stochastically continuous. Its characteristic function has fixed form called affine property. See, for example, Duffie et al. (2003) for more details. Also, the unique existence of stationary distribution of canonical affine processes can be found in Glasserman and Kim (2010). For more general cases, if we add jumps in affine processes, which is called affine jump-diffusion models, Jin et al. (2016) shows the exponential ergodicity of jump CIR process. All in all, some models based on affine processes also satisfy our requirement for Markov process (Class D regularity conditions), which can be potentially applied to use likelihood inference.

IG-OU process

IG-OU process is Lévy driven Ornstein-Uhlenbeck process defined by

$$dY_t = -\lambda Y_t dt + dL_{\lambda t} \tag{3.22}$$

where Y_t follows inverse Gaussian law and $\lambda > 0$. We assume Y_0 is generated from $IG(a, b)$. Though the transition density is not available, we have conditional characteristic function by using Equation (3.19) and (3.21):

$$\phi_t(u:\theta) \equiv E(\exp(iuY_{(t+1)\delta})|Y_{t\delta}) = \exp(-a(\sqrt{-2iu+b^2} - \sqrt{-2iu\exp(-\lambda\delta)+b^2}) + iu\exp(-\lambda\delta)Y_{t\delta}). \tag{3.23}$$

3.3.3 Verifications of regularity conditions

Verification of A class of regularity conditions

In this section, we provide several ideas to prove the A class of maximum likelihood estimator (MLE) for common Lévy processes: Merton's jump-diffusion model, Kou's jump-diffusion model, normal inverse Gaussian model and CGMY model. Other typical Levy processes might also use similar ways to prove. Once regularity conditions are fulfilled, our approximated likelihood function will converge to its corresponding likelihood function with exponential decay rate, which is fast and efficient. Thus, our proposed implementation should be efficient, in theory.

Theorem 13. *The characteristic function of Merton's jump-diffusion model, Kou's jump-diffusion model, NIG model and CGMY model satisfy the regularity condition of class A if the analytic strip $\mathcal{D}_{(d_-,d_+)}$ and compact parameter space Θ are selected regarding to the Chapter 3.3.*

Positive definiteness

To conduct maximum likelihood inference for Merton's jump-diffusion model, Kou's jump-diffusion model, NIG model and CGMY model, we might also be interested in the positive-definiteness of their fisher information matrix, so that AMLE is asymptotic efficient.

To prove the fisher information is positive definite, we provide a sufficient condition implied by the following lemma:

Lemma 13.1. *Suppose $g(X)$ is vector of statistics with positive definite covariance matrix $V(\theta)$. Let $h(\theta) = E_\theta g(X)$. Then, the matrix:*

$$I(\theta) - (\frac{\partial h(\theta)}{\partial \theta})^T V(\theta)^{-1} (\frac{\partial h(\theta)}{\partial \theta})$$

is nonnegative definite, where $I(\theta)$ is the fisher information matrix. That is, if the matrix $\partial h(\theta)/\partial\theta$ has full column rank, the fisher information matrix is positive definite.

We can apply lemma 13.1 to prove the fisher information matrix is positive definite given the set of parameters. Notice that Lemma 13.1 is the sufficient condition. If $\partial h(\theta)/\partial\theta$ is not full rank given the parameter θ, it doesn't mean fisher information is not positive definite. In that case, we can try different statistics vector $g(X)$ and the new $\partial h(\theta)/\partial\theta$ might be full rank which indicates the positive definiteness of the fisher information matrix.

To utilize the Lemma 13.1, we first construct statistics vector g(X). To guarantee the positive definiteness of $V(\theta)$, we choose $g(X) = (X, X^2, ...X^p)$. Then, we have the following lemma:

Lemma 13.2. *if X has continuous probability density over real line $\boldsymbol{R}$ given parameter θ, and X has finite moments with orders up to $2p$, $g(X) = (X, X^2, ...X^p)$ has positive definite covariance matrix $V(\theta)$.*

Now, we choose $g(X) = (X, X^2, ...X^p)$. Then,we provide several propositions with respect to the full rank of $h(\theta)/\partial\theta$ for multiple Levy processes.

Proposition 14. *For Merton's jump-diffusion model, if $g(X) = (X, X^2, \ldots, X^5)$, the determinant $|h(\theta)/\partial\theta| = 4\delta^5\lambda^2\sigma\mu_j\sigma_j(\mu_j^4 + (6\sqrt{6} - 9)\sigma_j^4)(\mu_j^4 - (6\sqrt{6} - 9)\sigma_j^4)$*

Notice that, if we choose the first 5 moments to construct $g(x)$ for Merton's jump-diffusion model, $h(\theta)/\partial\theta$ is a square matrix. Then for $h(\theta)/\partial\theta$, the full column rank is equivalent to the non-zero determinant. From the proposition 14, if true parameters satisfy: $\lambda \neq 0$, $\sigma \neq 0$, $\mu_j \neq 0$, $\sigma_j \neq 0$ and $\mu_j \neq \pm\sqrt[4]{6\sqrt{6} - 9}$, the $h(\theta)/\partial\theta$ has full column rank. That is, combined with Lemma 13.2, the fisher information matrix will

be positive definite. The true parameters λ, σ, μ_j and σ_j are unlike zero. However, $\mu_j = \pm\sqrt[4]{6\sqrt{6}-9}$ is possible. If $\mu_j = \pm\sqrt[4]{6\sqrt{6}-9}$, we can choose another $g(X)$. For example, we also try g(X) = $g(X) = (X, X^2, \ldots, X^4, X^6)$ and we find $\mu_j = \pm\sqrt[4]{6\sqrt{6}-9}$ is not the factor of $h(\theta)/\partial\theta$ anymore.

Proposition 15. *For Kou's jump-diffusion model, if* $g(X) = (X, X^2, \ldots, X^6)$*, the determinant* $|h(\theta)/\partial\theta| = 24883200\delta^6\sigma\eta_u^{-12}\eta_d^{-12}(\eta_u + \eta_d)^4\lambda^3(p-1)p$

Similarly, we choose the first 6 moments to construct $g(x)$ for Kou's jump-diffusion model to guarantee the $h(\theta)/\partial\theta$ is a square matrix. Then for $h(\theta)/\partial\theta$, the full column rank is equivalent to the non-zero determinant. From the proposition 15, if true parameters satisfy: $\lambda \neq 0$, $\sigma \neq 0$, $p \neq 0$, $p \neq 1$, $\eta_u \neq 0$ and $\eta_d \neq 0$, the $h(\theta)/\partial\theta$ has full column rank. That is, combined with Lemma 13.2, the fisher information matrix will be positive definite. Those conditions are fulfilled based on our restricted parameter space stated in Chapter 3.3.1.

Proposition 16. *For NIG model, if* $g(X) = (X, X^2, \ldots, X^4)$*, the determinant* $|h(\theta)/\partial\theta| = 18\delta^4\lambda^2\alpha^7(\alpha^2 - \beta^2)^{-15/2}$

Similarly, we choose the first 4 moments to construct $g(x)$ for NIG model to guarantee the $h(\theta)/\partial\theta$ is a square matrix. Then for $h(\theta)/\partial\theta$, the full column rank is equivalent to the non-zero determinant. From the proposition 16, if true parameters satisfy: $\lambda \neq 0$, $\alpha \neq 0$, $\alpha^2 \neq \beta^2$, the $h(\theta)/\partial\theta$ has full column rank. That is, combined with Lemma 13.2, the fisher information matrix will be positive definite. Those conditions are fulfilled based on our restricted parameter space stated in Chapter 3.3.1.

Proposition 17. *For CGMY model, if* Y *is fixed and known, we choose* $g(X) = (X, X^2, \ldots, X^4)$*, the determinant* $|h(\theta)/\partial\theta| = \delta^4C^2G^{Y-8}\mathcal{M}^{Y-8}(G+\mathcal{M})Gamma(-Y)^3Y^3(Y-1)^3(Y-2)^2(Y-3)((4-Y)(G^{1+Y}\mathcal{M}^3 + G^3\mathcal{M}^{1+Y}) + (2-Y)(G^Y\mathcal{M}^4 + G^4\mathcal{M}^Y))$*. If* Y *is also an unknown parameter, the determinant* $|h(\theta)/\partial\theta|$ *is far more complex which is* $C^3\delta^5G^{(-13+Y)}\mathcal{M}^{(-13+Y)}(G+\mathcal{M})(-3+Y)(-2+Y)^2Gamma[2-Y]^4(G^7\mathcal{M}^{2Y}(\mathcal{M}^2(-5+Y)(-4+Y) + G^2(-2+Y)^2 + G\mathcal{M}(-4+Y)(-5+2Y)) - G^{2Y}\mathcal{M}^7(G^2(-5+Y)(-4+Y) + \mathcal{M}^2(-2+Y)^2 + G\mathcal{M}(-4+Y)(-5+2Y)) + G^{3+Y}\mathcal{M}^{3+Y}((G-\mathcal{M})(4(5G^2+11G\mathcal{M}+5\mathcal{M}^2) - 13(G+\mathcal{M})^2Y + 2(G+\mathcal{M})^2Y^2) - (G+\mathcal{M})^3(-4+Y)(-3+Y)(-2+Y)(\log G - \log\mathcal{M}))$.

If we assume Y is fixed and known, we choose the first 4 moments to construct $g(x)$ for CGMY model to guarantee the $h(\theta)/\partial\theta$ is a square matrix. Then for $h(\theta)/\partial\theta$, the full column rank is equivalent to the non-zero determinant. From the proposition 16, if true parameters satisfy: $C \neq 0$, $G \neq 0$, $\mathcal{M} \neq 0$, the $h(\theta)/\partial\theta$ has full column rank. That is, combined with Lemma 13.2, the fisher information matrix will be positive definite. Those conditions are fulfilled based on our restricted parameter space stated in Chapter

3.3.1. However, if Y is included in the model, the determinant is far more complex. Though, $|h(\theta)/\partial\theta|$ is not zero in a large parameter set, if $G \to \mathcal{M}$, $|h(\theta)/\partial\theta| \to 0$. Thus, if there is no significant skewness ($G \approx \mathcal{M}$), we might want to use other choice of statistics vector $g(X)$ to prove the positive definiteness of fisher information matrix. For example, we try $g(X) = (X, X^2, X^3, X^4, X^6)$. Then, $G = \mathcal{M}$ will not imply $|h(\theta)/\partial\theta| = 0$.

3.4 Implementation and numerical studies

3.4.1 Implementation

In Chapter 3.2, we provide several regularity conditions to make sure that AMLE is consistent, asymptotically normal and asymptotically efficient. Also, these regularity conditions can be satisfied on specific compact parameter space for several typical Lévy processes. In this section, we illustrate how to obtain AMLE effectively to do parameter estimation and model selection.

In each model, we define a compact subset of parameter space and assume the true value is located inside (See Merton's jump-diffusion model, Kou's jump-diffusion model, normal inverse Gaussian model, CGMY model and IG-OU model in Chapter 3.3).

To perform simulation study, we need to simulate samples from different models we introduced above. The simulation methods introduced by each model are the exact simulation method except of CGMY model. To simulate the sample from CGMY model, we use very rigorous error controls, based on the inverting CDF methods Chen et al. (2012). Specific setting for the simulation study is in Appendix B.

To obtain reasonable AMLE, for each group of parameters and data, we properly select M, h to bound the distance between approximated density (3.2) and real density less than 10^{-8}. That is, we don't select a common M, h for all parameter values in our parameter space. Instead, we choose optimal $M(\theta), h(\theta)$ for every different parameter θ, so that the distance between approximated density and real density will be small enough uniformly on the parameter space. It will reduce a lot of computational burden because for some infeasible parameter values, M is required to be a huge number. Thus, we maximize (3.2) with optimal M, h to control the error bound. We choose 'NOMAD' optimization in 'OPTI' toolbox (Abramson et al. (n.d.)) of MATLAB to be our numerical optimization procedure here for three reasons. First, it is fast and stable. Second, it is a nonlinear and non-smooth optimization algorithm. In our case, we choose different M and h for different parameter to control the distance between approximated likelihood and real livelihood uniformly for parameter. Thus, our approximated likelihood is non-linear and non-smooth. Last, it is a constrained global optimization procedure which can deal with our compact parameter space which

is not sensitive to the initial parameter setting. Implementation details are put in the Appendix B.

Whenever we have AMLE, we can perform likelihood ratio test and model selection described in Chapter 3.2.2. To perform likelihood ratio test correctly, we need to check the regularity conditions. Unfortunately, regularity conditions cannot be satisfied for several typical cases. More details are in Chapter 3.2.2. As for model comparison, by rigorously controlling the approximation error of approximated likelihood, our $AIC^{M,h}$ and $BIC^{M,h}$ defined in Chapter 3.2.2 will be reasonably closed to AIC and BIC. We can use AIC and BIC to select appropriate model by penalizing over-fitting. Again, the asymptotic property of AIC and BIC might not be applied for our models. But, hopefully, they can at least get some useful hints. It is still an open question which model selection technique is the most appropriate one for financial models with good asymptotic properties.

3.4.2 Simulation study

In this section, we conduct the simulation study to examine the performance of AMLE. We mainly have two targets. First, we want to check if our AMLE matches the true parameter based on simulated data. In addition, we compare our AMLE estimates with the estimates based on other methods (approximated ECF estimation) to illustrate the smaller asymptotic variance of AMLE. Second, we want to discover more asymptotic properties of the estimated parameter from our simulation study. To the best of knowledge, there is few rigorous study of the asymptotic property of MLE based on Lévy processes. The main reason for it is that the density function implied by Lévy processes usually don't have a implicit form. Then, it might be too difficult to verify the regularity conditions of MLE asymptotic properties. Even, those regularity conditions can be verified, it is still hard to know the property of asymptotic variance of MLE due to the complicated form of fisher information matrix. Thus, considering AMLE inherits the asymptotic property of MLE (Theorem 8),we are able to find some interesting implications of MLE's asymptotic properties for several popular Lévy processes.

We consider the Lévy process based models described in Chapter 3.3 including Merton's jump-diffusion model, Kou's jump-diffusion model, NIG model, CGMY model and IG-OU processes.

For each Lévy model in Chapter 3.3, we fit the model based on simulated 500 sample paths with different sample size $N = \{200, 100, 500\}$, daily frequency ($\delta = 1/252$) and weekly frequency ($\delta = 1/52$). We also provide the approximated ECF estimates (proposed in Chapter 4.4) based on the same data dataset and same model. Global optimization 'NOMAD' is used to search the AMLE. The large enough compact parameter space required by Theorem 7 and Theorem 8 is set following the way mentioned in Chapter 3.3 for each model. Specific setting of the parameter space is reported in Table 3.2. For the Lévy driven Markov model,

Table 3.2: Parameter spaces

Merton	Kou	NIG	CGMY
True values			
$\mu = 0.1,\ \sigma = 0.3,\ \lambda = 20$	$\mu = 0.2,\ \sigma = 0.2,\ \lambda = 30$	$\mu = 0, \alpha = 50,\ \beta = -5,\ \lambda = 5$	$C = 3,\ G = 79,\ \mathcal{M} = 83$
$\mu_j = -0.5,\ \sigma_j = 0.25$	$p = 0.4,\ \eta_u = 5,\ \eta_d = 2.5$		$Y = 0.9$
Parameter spaces			
$-1 \leq \mu \leq 1$	$-1 \leq \mu \leq 1$	$1 \leq \alpha - \lvert\beta\rvert$	$1 \leq C \leq 100$
$0.01 \leq \sigma \leq 1$	$0.01 \leq \sigma \leq 1$	$-10 \leq \beta \leq 10$	$1 \leq G \leq 100$
$1 \leq \lambda \leq 50$	$1 \leq \lambda \leq 50$	$0.5 \leq \delta \leq 10$	$1 \leq \mathcal{M} \leq 100$
$-1 \leq \mu_j \leq 1$	$0.1 \leq p \leq 0.9$	$1 \leq \alpha \leq 150$	$0.1 \leq Y \leq 1.9$
$0.05 \leq \sigma_j \leq 1$	$1 \leq \eta_u \leq 20$		
	$1 \leq \eta_d \leq 20$		

IG-OU process, we use monthly frequency ($\delta = 1/12$) which is consistent to the setting in Chen et al. (2012).

In Table 3.3, we report the empirical means, corresponding standard errors of parameter estimates and true parameter values based on 500 simulated sample paths. For jump-diffusion models [Table3.3a,3.3b,3.3c and 3.3d], empirical means are reasonably closed to true parameters. With sample size N increasing, empirical means tend to converge to true value. This finding supports the consistency of AMLE (Theorem 7). In addition, when the sample size gets 5 times bigger (sample size from 200 to 1000 or from 1000 to 5000), the standard error becomes roughly $1/\sqrt{5}$ smaller. This is consistent to the convergence rate of asymptotic normality, $\sqrt{N}$, indicated by Theorem 8. No matter for Merton's jump-diffusion model or Kou's jump-diffusion model, we find 200 sample size is enough to recover the true parameter value by AMLE for weekly data, while 1000 sample size is needed for daily data. That is, roughly five year's data can match the true parameter with AMLE very well. If we have less data, for example, one year's data, most parameter can be estimated very well except of the drift parameter μ and certain jump's parameter. Take jump size parameter η_u in Kou's jump-diffusion model for an instance. The standard error is 2.9812 when the daily sample size is 200. This makes the estimate value 5.9695 less than three standard error away from zero. If we check further the histogram of η_u's estimates based on 500 simulated sample path (not shown in the work), we find a few estimates hits the parameter space boundary which makes the histogram deviate from normal distribution density. Thus, 200 daily sample might not be enough to estimate the parameters of jump-diffusion's model very well for our parameter setting.

What's more, for jump-diffusion models, if we set the sample size to be equal, we can find generally weekly data estimates have less or similar bias (difference between empirical mean and true value) and standard error, while, daily data estimates have larger or similar bias and standard error. This might be an indication that relative high frequency sampled data might imply relative slow or similar convergence speed compared with relative low frequency data. In general, it seems that when the total sample size is fixed, dense data might lead to more difficulties to estimate parameters. This finding is also intuitive. For finite activity Lévy processes (jump-diffusion models), there are finite big jumps in time interval. Weekly data

with fixed sample size have a longer time span, indicating more big jumps compared with daily data. Thus, denser data with less big jumps information lead to less accuracy of the estimates of jump parameters.

It is worth noticing that this is not the case for infinite activity Lévy processes [Table 3.3e,3.3f,3.3g and 3.3h]. For example, for parameters in NIG model and CGMY model, denser data (e.g. daily data) implies a less bias and standard error of the AMLE. Intuitively, the reason is that for these models, jumps will happen on each data point no matter it is daily data or weekly data. The data with a longer time span doesn't contain more jump information because in all fixed time interval, jumps will happen infinitely.

For infinite activity Lévy processes, the estimation performance varies. The estimated parameters in NIG model imply the consistency and convergence rate information. Based on empirical mean with true initial value in Table [Table 3.3e,3.3f], larger sample size N implies a smaller estimated bias and smaller standard error. For CGMY model, performance of estimation is not stable. Table 3.3g and Table 3.3h report the estimation result of CGMY model. We find parameter C and Y are hard to be identified with a big estimation standard error especially for the data. For the daily data, estimation performance gets even worse. It still seems that a very large sample size (larger than 5000) is necessary to identify C and Y.

We find there are two reasons for the issue above. First, the characteristic function decay slowly for CGMY model, especially when Y parameter is small ($Y < 0.4$). In our theory, this is because $\nu = Y$ for CGMY model (Check section 3.3.1 and regularity condition A.4) and smaller ν leads to slower decay of the characteristic function's tail. And this slower decay will generally lead to large Mh to make truncation approximation error bound T_{Mh} the same. Now, we let the discretization approximation error bound the same. That is, h is unchanged. We will have a larger M to control overall bound of error to be same. This usually leads to a very huge value of M to control our approximation error. For daily data, δ in the characteristic function (Equation 3.17) is smaller, making the characteristic function decay even slower (because smaller δ implies smaller c in regularity condition A.4 (check Section 3.3.1)). Thus, it implies that daily data even need bigger M compared with weekly data. In practice, when optimization algorithm searches the parameters in those unfeasible parameter space (e.g. Y is very small), a very large M ($M > 5000$) is needed to control the approximation error to be small (less than 10^{-8}). We set an upper-bound for M during each optimization iteration to reduce the computational burden in practice. This leads to large errors of the approximation for certain parameter values. This is more serious for daily data because daily data in general need larger M than weekly data.

Moreover, we find there is a relative flat trace in the likelihood surface of C and Y, so that smaller Y will lead to dramatically big Y and bigger Y will make C closed to zero. This makes the optimization algorithm tend to search those unfeasible area ($Y < 0.4$) easily and frequently, amplifying the effects of approximation

error caused by M. Figure 3.1a shows us a saddle-shaped log-likelihood surface of C and Y, when fixing G, $\mathcal{M}$ to be the true parameter based on 1000 generated daily data. Figure 3.1b exhibits a relative flat trace of C and Y (Yellow trace) including the maximum point on this saddle-shaped log-likelihood surface. Log-likelihood is estimated from Equation (3.3) which has an accuracy of 6 significant digits and a precision of 2 decimal places by controlling truncation error, discretization error and set $a = 0$ (We only draw the log-likelihood surface when $Y > 0.5$ because controlling the error to be smaller needs a huge M when $Y < 0.5$). It is reasonable to believe there exists even flatter space when relaxing G and $\mathcal{M}$. This supports our opinions about the poor performance because optimization algorithm might search the area where C is large and Y is small in that flat trace, causing big approximation errors and wrong estimates. We also find log-likelihood of parameter G and $\mathcal{M}$ when fixing C and G is hump-shaped and is relatively fat which also indicating not small standard errors of parameter estimates [Figure 3.1c and 3.1d].

To express the issue of CGMY model estimation further, we redo the simulation study with larger sample size and a little smaller parameter space (Reduce the computational burden). Then, we plot the histogram of estimated parameters based on 500 simulated sample paths. All models show the normal distributed shapes except of CGMY model. To address this difficulty, we list the histogram of estimated parameters for CGMY model below. Figure 3.2 and Figure 3.3 show the histogram of estimated parameters C, G, $\mathcal{M}$, Y with different sample size n and different frequencies (daily and weekly). Parameters C, G, $\mathcal{M}$, Y are listed from first row to forth row in each figure. Sample size increases with column number increasing. For weekly data (Figure 3.2), 5000 sample size might be able to make parameter G and $\mathcal{M}$ in normal distribution shape. But, histogram of C and Y are still not in a good shape. It is skewed for the histogram of parameter C and multi-peaked for the histogram of parameter Y. Then, when it comes to daily data, it seems the histogram has better shape. When sample size is as big as 25000, the mode of histogram is around the true parmeters. However, parameter C's histogram is still a little skewed. Overall, parameters in CGMY model are relatively hard to estimate. Parameter C and Y might have some identification problems of practical simulated data, making the likelihood surface relative flat in certain part. This amplifies the issues caused by the big error of the approximation. The CGMY model might need a very large sample size to make the flat part of log-likelihood surface steeper, making AMLE reasonably closed to true parameter. Or, simply a smaller parameter space is suggested.

We also compare our AMLE with approximated ECF estimates for Merton's jump-diffusion model, Kou's jump-diffusion model and NIG model. ECF estimation is implemented based on Chapter 4 (See the Chapter 4.3 and Chapter 4.4 for implementation details). Empirical characteristic function (ECF) is another efficient method to do parameter estimation for Lévy processes. It can be regarded as a generalized moment match

method to match empirical characteristic function and the model's characteristic function. The approximated ECF estimation introduced in Chapter 4 also has asymptotic normality property. However, the asymptotic variance cannot reach the MLE efficiency. That is, the asymptotic variance of MLE should be smaller than the one of approximated ECF estimate in theory. We set the tuning parameter in AECF estates to be a big number ($M = 200$) and conduct both MLE and ECF methods based on same 5000 simulated data in Table 3.4. We find that the empirical means of both estimates (MLE and ECF) are similar and the standard error of the estimates is smaller for AMLE. This is consistent to the asymptotic efficiency of AMLE.

Another interesting finding is about the running time. The running time in the table is the median of running time based on 500 simulated paths. Running time depends on the optimization algorithm, likelihood surface and our computing device. All the simulation study was conducted by the same laptop with Intel Core i5-5300U CPU (2.30GHz). Table 3.3 shows the increasing ruining time (per simulated path in 'second' scale) when sample size is increasing. This is reasonable due to large sample size indicating large computational burden of approximated likelihood (Equation (3.2)). Moreover, denser (daily) simulated data indicates a larger running time than more scattered data (weekly data) when controlling the same sample size N. This can be explained by our Equation (3.4) in Theorem 6. That is, denser data can make the model's characteristic function decay more slowly with bigger parameter k in truncation error bound T_{Mh}. This indicates we need larger Mh to make the truncation error bound unchanged if we have larger k. Suppose we also control the discretization error bound in Equation (3.4) which is only related to parameter h, we find we need larger M to control the total error bound with larger k. And, larger M leads to higher computation cost due to our approximated likelihood equation (3.2). Thus, to control the same likelihood approximation error bound, denser data from the same model need larger M which indicating larger computation cost (running time).

For a Lévy driven Markov process, IG-OU process, the proposed estimates as reported in Table 3.3i are reasonably closed to the true values and standard errors decrease with increasing sample size. In summary, this simulation results suggest that AMLE can accurately estimates most models introduced here with reasonable sample size. CGMY model has relatively flat log-likelihood and more samples might be necessary to make AMLE closed to true parameters. But, this doesn't mean extreme large data set is needed to fit CGMY model in practice. It is believed that practical data can't be exactly from a model. In this way, true parameter might only live in our simulation study rather than real world. Comparing estimated parameter with true parameter is even more impossible. If we just utilize information of log-likelihood or to match several moments, relatively smaller sample size might be possible.

Table 3.3: Empirical averages and their standard errors (in parentheses) of the approximated maximum likelihood estimates (AMLE) with sample size $N = 200, 1000, 5000$. Running time is also reported in the 'second' scale

(a) Merton's jump-diffusion model (weekhly data)

N	$\mu = 0.1$	$\sigma = 0.3$	$\lambda = 10$	$\mu_j = -0.5$	$\sigma_j = 0.25$	time
200	0.0860(0.1745)	0.2989(0.0178)	10.0263(1.8771)	−0.5048(0.0585)	0.2397(0.0482)	6.4756
1000	0.0956(0.0771)	0.3001(0.0081)	9.9184(0.8146)	−0.5039(0.0261)	0.2482(0.0209)	11.7263
5000	0.1012(0.0337)	0.3002(0.0037)	9.9912(0.3634)	−0.5004(0.0109)	0.2490(0.0093)	35.4301

(b) Merton's jump-diffusion model (daily data)

N	$\mu = 0.1$	$\sigma = 0.3$	$\lambda = 10$	$\mu_j = -0.5$	$\sigma_j = 0.25$	time
200	0.1030(0.3609)	0.2989(0.0152)	9.8928(3.5901)	−0.5017(0.1107)	0.2152(0.0792)	10.8874
1000	0.1025(0.1567)	0.3001(0.0068)	9.9730(1.7330)	−0.5007(0.0464)	0.2411(0.0335)	16.4473
5000	0.0991(0.0682)	0.3002(0.0031)	9.9354(0.7451)	−0.4997(0.0204)	0.2480(0.0158)	41.0280

(c) Kou's jump-diffusion model (weekly data)

N	$\mu = 0.2$	$\sigma = 0.2$	$\lambda = 30$	$p = 0.4$	$\eta_u = 5$	$\eta_d = 2.5$	time
200	0.1868(0.1667)	0.1993(0.0188)	30.0854(4.1172)	0.4041(0.0679)	5.1941(1.1422)	2.5463(0.3884)	8.9960
1000	0.1979(0.0708)	0.1997(0.0083)	30.0221(1.7681)	0.3999(0.0301)	5.0303(0.5109)	2.5180(0.1717)	19.3745
5000	0.1970(0.0327)	0.2000(0.0038)	29.9995(0.8350)	0.3997(0.0134)	4.9838(0.2307)	2.5061(0.0797)	64.5251

(d) Kou's jump-diffusion model (daily data)

N	$\mu = 0.2$	$\sigma = 0.2$	$\lambda = 30$	$p = 0.4$	$\eta_u = 5$	$\eta_d = 2.5$	time
200	0.2078(0.2483)	0.1993(0.0117)	30.2998(7.2142)	0.4066(0.1231)	5.9695(2.9812)	2.7645(0.9177)	29.5504
1000	0.2034(0.1159)	0.2002(0.0053)	30.0147(3.2294)	0.3994(0.0530)	5.1325(0.8929)	2.5381(0.3481)	48.1453
5000	0.2009(0.0468)	0.2001(0.0023)	29.9410(1.3322)	0.3999(0.0234)	5.0509(0.3810)	2.4981(0.1526)	146.1137

Table 3.3 (cont.)

(e) Normal inverse Gamma model (NIG)(weekly data)

n	μ	$\alpha = 50$	$\beta = -5$	$\lambda = 5$	time
200	−0.1738(0.5910)	58.8797(26.0311)	−3.2901(6.0864	5.7439(2.3579)	4.4612
1000	−0.0133(0.3862)	53.3473(14.0727)	−4.9311(3.9311)	5.3110(1.3506)	5.8038
5000	0.0041(0.1906)	50.8198(5.7839)	−5.0761(1.9559)	5.0781(0.5307)	13.2580

(f) Normal inverse Gamma model (NIG)(daily data)

N	μ	$\alpha = 50$	$\beta = -5$	$\lambda = 5$	time
200	−0.0970(0.5134)	55.5646(19.7875)	−3.9036(5.5808)	5.3684(1.4257)	33.3807
1000	0.0087(0.2759)	51.0985(6.7145)	−5.2008(3.0870)	5.0792(0.5097)	51.8738
5000	0.0014(0.1253)	50.2575(3.0809)	−5.0808(1.4329)	5.0235(0.2149)	124.7928

(g) CGMY model (weekly data)

N	$C = 3$	$G = 79$	$\mathcal{M} = 83$	$Y = 0.9$	time
200	11.5818(20.7850)	80.2868(15.9604)	84.7726(15.8860)	0.8538(0.3248)	37.3605
1000	5.7926(8.4604)	80.4154(8.3584)	84.4865(8.4944)	0.8987(0.1760)	52.9537
5000	3.4610(2.5674)	79.0453(3.7624)	83.1361(3.8104)	0.8851(0.0955)	75.6349

(h) CGMY model (daily data)

N	$C = 3$	$G = 79$	$\mathcal{M} = 83$	$Y = 0.9$	time
200	8.6386(20.3133)	87.7222(18.6472)	91.1323(16.2169)	0.3980(0.3919)	113.7943
1000	4.7185(8.1174)	88.2859(12.4441)	91.4449(10.9292)	0.5302(0.3882)	176.2361
5000	3.4651(3.0954)	84.9388(9.9740)	88.1790(8.6123)	0.6962(0.3344)	244.2984

(i) Inverse Gaussian-OU model (monthly data)

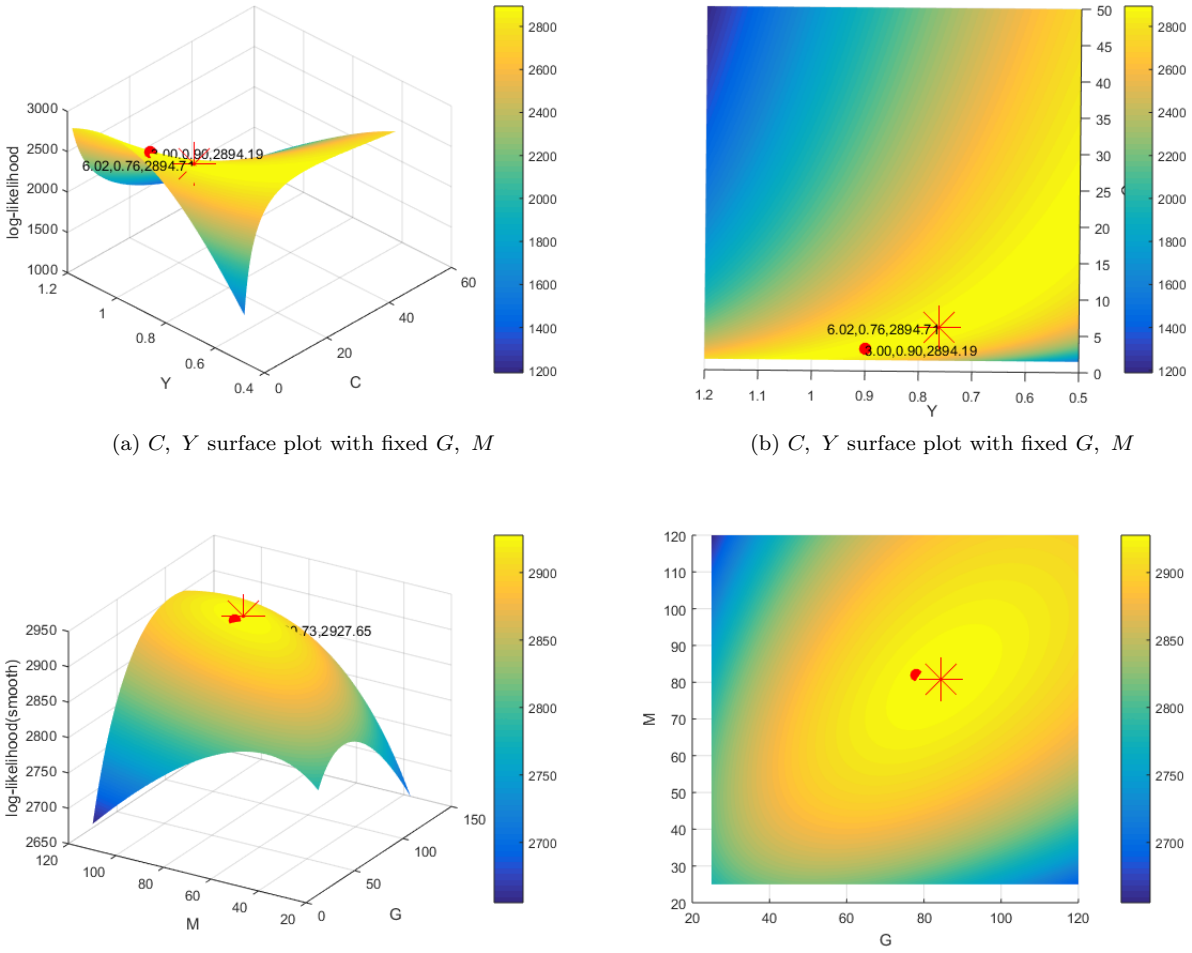
	$\lambda = 10$	$a = 1$	$b = 20$
$N = 125$	10.3244(1.8016)	1.0146(0.0799)	20.2936(1.6439)
$N = 250$	10.1797(1.1867)	1.0098(0.0533)	20.1805(1.0898)
$N = 500$	10.0187(0.8242)	1.0036(0.0400)	20.0888(0.8298)

Table 3.4: Comparison of empirical averages and their standard errors (in parentheses) between the approximated maximum likelihood estimates (AMLE) and approximated empirical characteristic function estimates (AECF). 500 sample paths are generated and each sample path has 5000 data.

Merton							
Method	Frequency	$\mu = 0.1$	$\sigma = 0.3$	$\lambda = 10$	$\mu_j = -0.5$	$\sigma_j = 0.25$	time
MLE	Weekly	0.1012(0.0337)	0.3002(0.0037)	9.9912(0.3634)	−0.5004(0.0109)	0.2490(0.0093)	35.4301
ECF	Weekly	0.1015(0.0341)	0.3001(0.0041)	10.0037(0.3936)	−0.4996(0.0140)	0.2498(0.0129)	25.2845
MLE	Daily	0.0991(0.0682)	0.3002(0.0031)	9.9354(0.7451)	−0.4997(0.0204)	0.2480(0.0158)	41.0280
ECF	Daily	0.0999(0.0702)	0.3003(0.0033)	9.9268(0.7451)	−0.4995(0.0250)	0.2463(0.0219)	41.2290

Kou								
Method	Frequency	$\mu = 0.2$	$\sigma = 0.2$	$\lambda = 30$	$p = 0.4$	$\eta_u = 5$	$\eta_d = 2.5$	time
MLE	Weekly	0.1970(0.0327)	0.2000(0.0038)	29.9995(0.8350)	0.3997(0.0134)	4.9838(0.2307)	2.5061(0.0797)	64.5251
ECF	Weekly	0.1963(0.0342)	0.1999(0.0050)	30.0004(0.9112)	0.3999(0.0137)	4.9919(0.3216)	2.5011(0.1014)	78.1217
MLE	Daily	0.2009(0.0468)	0.2001(0.0023)	29.9410(1.3322)	0.3999(0.0234)	5.0509(0.3810)	2.4981(0.1526)	146.1137
ECF	Daily	0.2007(0.0478)	0.2001(0.0024)	29.9443(1.3620)	0.4193(0.0239)	5.0413(0.6093)	2.5081(0.2253)	90.0734

NIG						
Method	Frequency	$\mu = 0$	$\alpha = 50$	$\beta = -5$	$\lambda = 5$	time
MLE	weekly	0.0041(0.1906)	50.8198(5.7839)	−5.0761(1.9559)	5.0781(0.5307)	13.2580
ECF	weekly	0.0041(0.1995)	50.7725(5.9230)	−5.0778(2.0616)	5.0732(0.5433)	38.7957
MLE	daily	0.0014(0.1253)	50.2575(3.0809)	−5.0808(1.4329)	5.0235(0.2149)	124.7928
ECF	daily	0.0002(0.1354)	50.1876(3.2926)	−5.0785(1.6281)	5.0185(0.2287)	27.2182



(a) $C,\ Y$ surface plot with fixed $G,\ M$

(b) $C,\ Y$ surface plot with fixed $G,\ M$

(c) $G,\ M$ surface plot with fixed $C,\ Y$

(d) $G,\ M$ surface plot with fixed $C,\ Y$

Figure 3.1: Log-likelihood surface of CGMY model with fixed $G,\ \mathcal{M}$ and fixed $C,\ Y$ where true parameters are $C = 3,\ G = 78,\ \mathcal{M} = 82,\ Y = 0.9$ (denoted by red $'\bullet'$) based on 1000 simulated data. Red $'*'$ is the maximum likelihood point when fixing $G,\ \mathcal{M}$ or $C,\ Y$ to true parameter. Log-likelihood has an accuracy of 6 significant digits and a precision of 2 decimal places

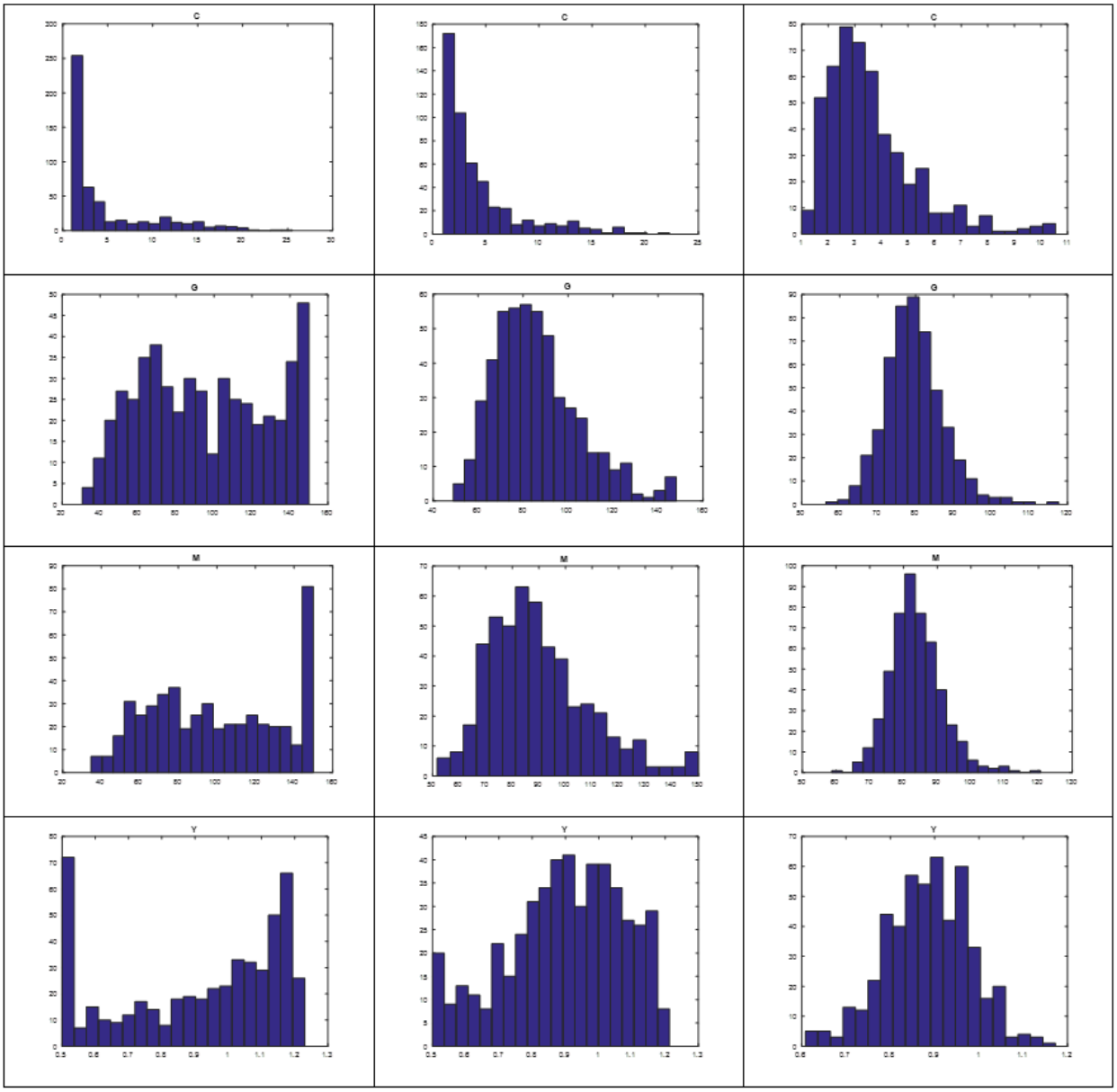


Figure 3.2: Histogram of parameters of CGMY model based on simulated weekly data: sample size n are $200, 1000, 5000$ from first column to third column. Parameters C, G, $\mathcal{M}$, Y are from first row to forth row

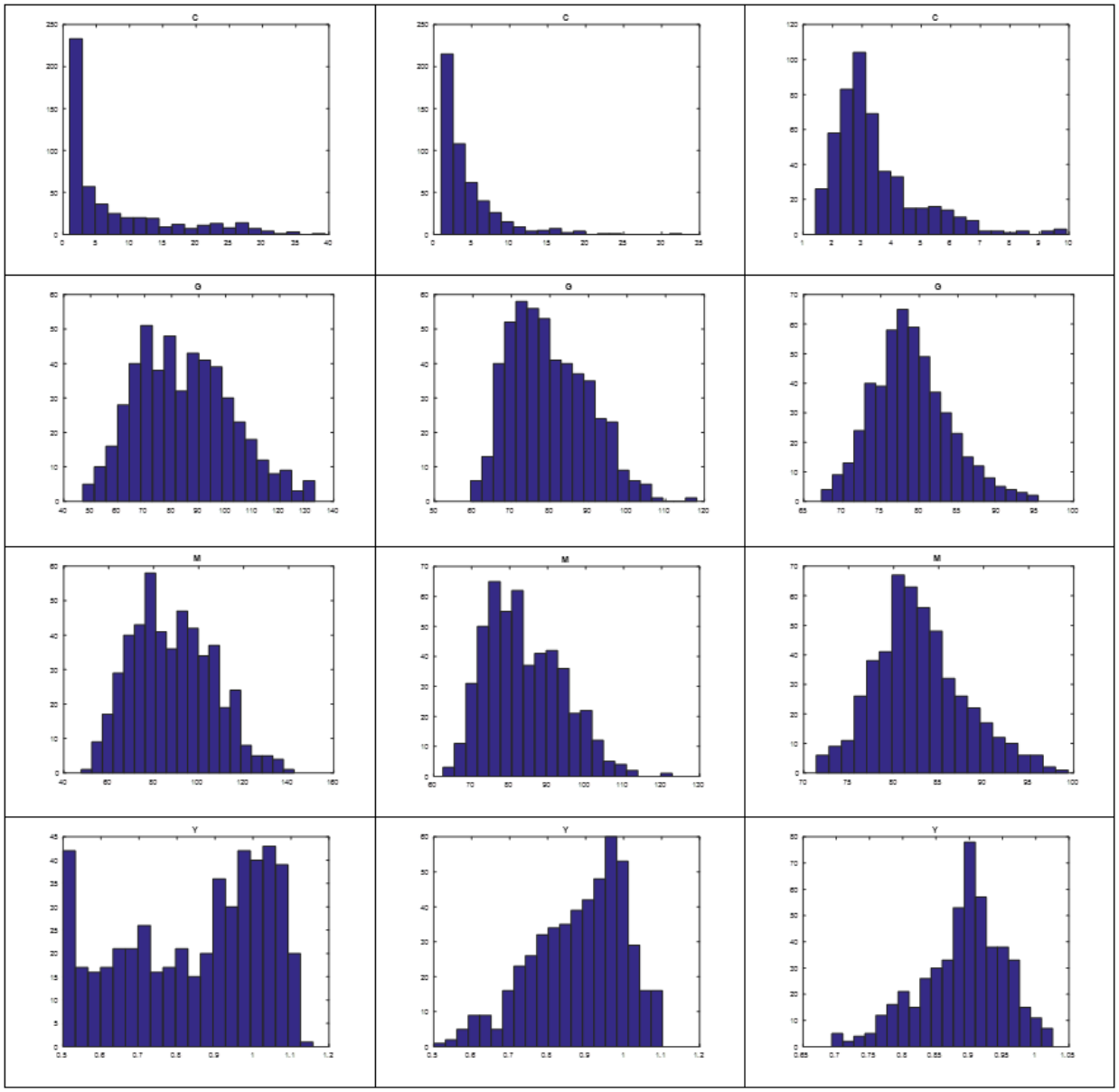


Figure 3.3: Histogram of parameters of CGMY model based on simulated daily data: sample size n are $1000, 5000, 25000$ from first column to third column. Parameters C, G, $\mathcal{M}$, Y are from first row to forth row

Table 3.5: Maximum likelihood estimates, AIC and BIC

Model	Parameters						AIC	BIC
Ticker (Date): CRSP (7/2/1962-12/31/1987)								
Merton	μ	σ	λ	μ_j	σ_j			
	0.1318	0.0949	33.1195	−0.0003	0.0138		−22241.75	−22207.93
Kou	μ	σ	λ	p	η_u	η_d		
	0.1128	0.08892	76.2045	0.3986	129.8429	169.2817	−22247.83	−22207.23
NIG	μ	α	β	λ				
	0.3136	117.6695	−14.9778	1.8168			**-22301.12**	**-22274.05**
CGMY	μ	C	G	$\mathcal{M}$	Y			
	0.2820	7.9363	122.2715	149.0953	0.6068		−22298.90	−22265.08
Ticker (Date): SPD (1/3/1995-12/30/2005)								
Merton	μ	σ	λ	μ_j	σ_j			
	0.2057	0.097	175.5738	−0.0005	0.0110		−8677.01	−8647.38
Kou	μ	σ	λ	p	η_u	η_d		
	0.4646	0.0917	319.4898	0.3560	147.8200	182.5037	−8679.49	−8643.92
NIG	μ	α	β	λ				
	0.2053	85.3495	−3.1783	2.6518			**-8683.59**	**-8659.88**
CGMY	μ	C	G	$\mathcal{M}$	Y			
	0.2100	6.6229	86.9645	93.6477	0.6792		−8681.91	−8652.28
Ticker (Date): LCBM (1/3/1995-12/30/2005)								
Merton	μ	σ	λ	μ_j	σ_j			
	−0.7005	0.3596	79.6837	0.01541	0.0628		−5183.74	−5154.10
Kou	μ	σ	λ	p	η_u	η_d		
	−0.2784	0.2851	197.3305	0.4409	25.7382	42.4000	−5205.31	−5169.75
NIG	μ	α	β	λ				
	−0.9066	13.5180	2.7874	6.1501			**-5223.93**	**-5200.22**
CGMY	μ	C	G	$\mathcal{M}$	Y			
	−0.8350	7.1241	16.7069	11.3917	0.7334		−5223.80	−5194.16

Table 3.5 (cont.)

Ticker (Date): INTC (1/3/1995-12/30/2005)								
Merton	μ	σ	λ	μ_j	σ_j			
	0.2129	0.3375	55.1745	0.0022	0.0418		−6011.52	−5981.89
Kou	μ	σ	λ	p	η_u	η_d		
	−0.2867	0.3206	120.1293	0.6686	51.7683	40.3296	−6013.49	−5977.93
NIG	μ	α	β	λ				
	−0.0698	37.3429	1.6204	7.7685			−6012.79	**-5989.08**
CGMY	μ	C	G	$\mathcal{M}$	Y			
	0.0160	0.7001	20.1865	17.7136	1.2777		**-6015.26**	−5985.63
Ticker (Date): DOW (1/3/1995-12/30/2005)								
Merton	μ	σ	λ	μ_j	σ_j			
	−0.0258	0.2045	74.3908	0.0026	0.0255		−7200.55	−7170.91
Kou	μ	σ	λ	p	η_u	η_d		
	−0.2746	0.1822	182.2514	0.5976	81.1307	78.7373	−7202.45	−7166.88
NIG	μ	α	β	λ				
	−0.2798	50.7061	4.7008	4.5448			**-7206.89**	**-7183.18**
CGMY	μ	C	G	$\mathcal{M}$	Y			
	−0.2493	0.9548	37.2887	28.6712	1.1076		−7205.89	−7176.25
Ticker (Date): TBL(1/3/1995-12/30/2005)								
Merton	μ	σ	λ	μ_j	σ_j			
	−0.1301	0.2786	73.8747	0.0067	0.0401		−6191.56	−6161.93
Kou	μ	σ	λ	p	η_u	η_d		
	−0.2466	0.2457	169.454	0.5418	45.8613	54.3658	−6197.51	−6161.95
NIG	μ	α	β	λ				
	−0.4690	27.7655	4.0255	5.4340			**-6208.18**	**-6184.48**
CGMY	μ	C	G	$\mathcal{M}$	Y			
	−0.4630	2.2258	25.0830	17.1563	0.9812		−6206.86	−6177.22
Ticker (Date): TIF(1/3/1995-12/30/2005)								

Table 3.5 (cont.)

Merton	μ	σ	λ	μ_j	σ_j			
	-0.2880	0.2668	86.7467	0.0070	0.03509		-6277.41	-6247.78
Kou	μ	σ	λ	p	η_u	η_d		
	-0.5308	0.2476	167.8447	0.6036	50.4098	55.8244	-6282.77	-6247.20
NIG	μ	α	β	λ				
	-0.4690	27.7655	4.0255	5.4340			**-6290.54**	**-6266.83**
CGMY	μ	C	G	$\mathcal{M}$	Y			
	-0.6357	1.3821	26.5877	16.4158	1.0882		-6289.98	-6260.34

3.4.3 Fitting equity returns

In this section, we study daily financial series. The weekly series can be analyzed in the same way. The daily data we use include value weighted CRSP, S&P-500 and 5 individual stocks. CRSP is the daily series with 6410 value weighted return from 1926 to 1987. The S&P-500 and individual stock series are from 1995 to 2005 containing 2771 daily returns. Individual series are adjusted by dividends. Individual stocks are chosen based on large range of kurtosis [DOW (6.6754), INTC (7.5635), TBL (9.5721), TIF (10.0038) and LCBM (26.4485)].

We estimate Merton's jump-diffusion model, Kou's jump-diffusion model, NIG model and CGMY model for each series and utilize AIC and BIC to do model selection. To confirm the accuracy of estimation, we rigorously control discretization and truncation error described in 6 so that Log-likelihood has a precision of 2 decimal places.

Table 3.5 summarizes the AMLE of four models with corresponding AIC and BIC. First, we find different models contain consistent information. For the drift term μ, all models implies the same sign except INTC of which drift term is around zero. Because NIG and CGMY model doesn't contain diffusion term, we compare diffusion term estimates of Merton's jump-diffusion and Kou's jump-diffusion model and find they are similar in most cases. For the skewness, negative μ_j estimates in Merton's jump-diffusion model, negative $\frac{p}{\eta_u} - \frac{(1-p)}{\eta_d}$ in Kou's jump-diffusion model, negative β and negative $G - \mathcal{M}$ roughly indicate a density skewing to the left. Positive values imply a density skew to the right. It is obviously to find for all financial series we use, all models imply consistent skewness. What's more, if we only focus on jumps, Kou's jump-diffusion model, NIG model and CGMY model indicate that symmetric jumps (p=0.5, $G = \mathcal{M}$ and $\beta = 0$) rarely happen. Thus, Merton's jump-diffusion model might not be the best choice in a lot of cases.

For the kurtosis, we find CRSP and CPD both have relatively small μ_j and σ_j for Merton's jump-diffusion model compared with individual stocks which indicates lighter tails. For Kou's jump-diffusion model, larger η_u and η_d in CRSP and SPD than individual stocks also have the same implications which is also consistent to the result in Ramezani and Zeng (2007). NIG model with large α and CGMY model with large C also roughly indicate high kurtosis of CRSP and SPD compared with most individual stocks.

If we roughly compare our estimation results with Ramezani and Zeng (2007), we can find that most parameters are in reasonable range except extreme large λ in SPD. Even if we don't consider that λ, both our estimates and estimates in Ramezani and Zeng (2007) still imply a large jump intensity λ. That is, jump frequency is very high. For jump sizes η_u and η_d, we find that they range from 25.74 to 182. Similar to the analysis Ramezani and Zeng (2007), considering density of up jumps following Pareto distribution with parameter $(1, \eta_u)$, over 95% of the up jumps will be less than3% when $\eta_u = 80$. This might be appropriate to be regarded as diffusion part for some individual stocks. Thus, jump-diffusion model might come cross a problem for several real world data to identify jumps from diffusion part due to the fact that jumps might include both small high frequent jumps and large low frequent jumps. In this way, infinite frequency jump structure like NIG model and CGMY model might be a good alternative to jump-diffusion models. Also, our parameter estimates of CGMY model are comparable to those reported in Kim et al. (2008). That estimates of CGMY model is roughtly $C = 3, G = 78, \mathcal{M} = 82, Y = 0.9$ for the S&P-500 series from 2000 to 2005. Ours are $C = 6.6, G = 86.9, \mathcal{M} = 93.62, Y = 0.67$ for the S&P-500 series from 1995 to 2005.

Turning to the model selection, it has been mentioned that smallest BIC (AIC) provides the best fits of data. We find that NIG model provides the best fit with respect to BIC (AIC) in all (6 out of 7) time series. CGMY model has comparable performance with NIG model and provides the lowest AIC for INTC. Compared with jump-diffusion model, infinitely activity Lévy processes including NIG model and CGMY model are preferred. The reason for it might be what we mentioned in the last paragraph that small jumps are too frequent in some return series. If we only compare jump-diffusion models, AIC prefers to select Kou's jump-diffusion model for all time series, while BIC tends to select Kou's jump-diffusion model for 2 out of 7 return series in case study.

In conclusion, we fit both jump-diffusion models and infinite activity Lévy models to the daily return series. Most AMLE we get is reasonable and comparable with other studies. From our AMLE of parameters, we confirm the result in Ramezani and Zeng (2007) that high frequency of jumps are identified by jump-diffusion models for return series. Our model selection result also prefers infinite activity Lévy processes to finite activity Lévy processes (Jump diffusion processes). This is also partly consistent to the result mentioned in Geman (2002) that infinitely activity and finite variation pure jump Lévy processes are preferred.

Moreover, all of our selected optimal models have infinite variation here even for our selected CGMY model with $Y = 1.2777$ based on INTC ($Y > 1$ indicates infinite variation , while $Y < 1$ implies finite variation). Because testing finite variation is not our main focus here, we don't discuss it here. But, it is an interesting topic to discriminate finite variation from infinite variation based on discrete and equally spaced data since variation is a limit concept while discrete data only contain limited jump information.

3.4.4 Concluding remarks

In this chapter, we construct approximated likelihood based on inverse Fourier transform. We propose the approximated maximum likelihood estimation (AMLE) and model selection. On the theoretical side, we propose asymptotic properties of AMLE. On the application side, AMLE is applied to Lévy and its based processes. Simulation study indicates the asymptotic property of AMLE. Also, it raised a problem of statistical estimation for CGMY model. The likelihood surface, especially of parameter C and Y is flat and large sample size might be necessary to match the estimated parameter back to the true parameter based on simulated paths. Numerical results show the apprealingness of some infinite activity models, such as NIG model. Both simulation study and numerical results show the effectiveness and efficiency of our implementation of AMLE.

Chapter 4

Empirical characteristic function estimation for Levy processes in finance

4.1 Introduction

Lévy processes are widely favored in economic and financial applications due to their capability of capturing skewness and fat tails. Suppose $Y = \{Y_t, t = 0, \delta, 2\delta, \ldots\}$ is a sequence of an evenly sampled financial variable, e.g., the log price of a certain asset. If we assume that Y is a Lévy process, then $\{X_k = Y_{k\delta} - Y_{(k-1)\delta}, k \geq 1\}$ is an i.i.d. sequence due to the fact that Lévy processes have independent and stationary increments.

To estimate the parameters in the Lévy processes, we can conduct traditional inference methods on this i.i.d. sequence $X = \{X_1, \ldots, X_n\}$. For example, we can use maximum likelihood methods, which have asymptotic efficiency property under mild regularity conditions. However, maximum likelihood methods usually require the explicit form of the distribution of X_i, which is usually complex or not available explicitly for Lévy processes. A discussion of maximum likelihood inference for Lévy processes is in Chapter 3.

Although the distribution form of Lévy processes is usually not available, their characteristic function $\phi(u;\theta) = E(\exp(iuX_k))$ often has an explicit form due to the Lévy-Khintchine formula (refer to Tankov (2003)). Given $u \in \mathbf{R}$, the characteristic function can be regarded as a generalized moment of Lévy processes (in $\mathbf{R}$) with close form. This makes generalized moment match (GMM) inference possible. We do not introduce GMM in details here. More discussion of GMM can be found in Hansen (1982).

We focus on an ECF estimation method belonging to the generalized moment match inference. This method exploits the empirical characteristic function $\phi_n(u)$, to estimate the model's parameters based on the data $X = \{X_1, \ldots, X_n\}$ (Definition of empirical characteristic function is in Chapter 4.2). Specifically, this method searches parameters θ by minimizing the distance between empirical characteristic function $\phi_n(u)$ and the model's characteristic function $\phi(u;\theta)$ continuously over $u \in \mathbf{R}$:

$$\arg\min_{\theta} \int |\phi(u,\theta) - \phi_n(u)|^2 dG(u),$$

where $G(u)$ is a weight function. Considering an integration is involved in the ECF target function, this

might lead to a huge computational cost for ECF estimation procedure.

In this chapter, we provide a way to implement ECF estimation efficiently and effectively by using the analyticity of the characteristic function. We show that our method works for popular Lévy processes. We also provide the theoretical framework to establish the asymptotic property of the ECF estimation obtained by our implementation. Simulation study shows that our ECF estimates match the true parameter very well. In addition, few points (about 20 points) can approximate the integration in the ECF estimation very well in most cases.

The chapter is organized as follows. In Chapter 4.2, we present our ECF estimates with asymptotic properties and the corresponding regularity conditions. In Chapter 4.3, we present the detailed implementation and some commonly used Lévy models. In Chapter 4.4, we perform simulation studies. Chapter 4.5 is the conclusion. All proofs are in the Appendix C.

4.2 Methods

In this section, we study the approximated empirical characteristic estimation based on trapezoidal rule. We assume $X = \{X_1, \ldots, X_n\}$ is identically and independently distributed (i.i.d) and its characteristic function $\phi(u; \theta_0)$ is available, where θ_0 is the true parameter vector and $u \in \mathbf{R}$. However, its probability density function (PDF) $f(x; \theta)$ might not have close form. We denote $\theta = (\theta_1, \ldots, \theta_p)$ as unknown parameters of distribution f within parameter space $\Theta \subset \mathbf{R}^p$. $F(x; \theta)$ is the cumulative distribution function (CDF) and $F_n(x)$ is the empirical CDF.

The increments of Leýy processes are independent and stationary. In addition, their density function are usually not tractable, but the form of their characteristic function is known due to Leýy-Khintchine formula. Thus, our i.i.d assumptions above can be applied to Leýy processes directly.

4.2.1 Empirical characteristic function estimation

Empirical characteristic function estimation is an estimation method to match the empirical characteristic function implied by the data and the characteristic function derived from the model. It can be viewed as a generalized method of moment (GMM) estimation (Hansen (1982)).

Suppose we have the characteristic function:

$$\phi(u; \theta) = E(\exp(iuX)) = \int \exp(iux) dF(x; \theta) = \int \exp(iux) f(x; \theta) dx, \tag{4.1}$$

and the empirical characteristic function is:

$$\phi_n(u) = \int \exp(iux) dF_n(x) = \frac{1}{n} \sum_{j=1}^{n} \exp(iuX_j), \tag{4.2}$$

where $i = \sqrt{-1}$.

Then, we have following infinite moment conditions:

$$m(u, X_i; \theta) = \exp(iuX_i) - \phi(u; \theta),$$

satisfying $Em(u, X_i; \theta_0) = 0$ for infinite and arbitrary choices of $u \in R$.

If we select q discrete $u = \{u_1, \ldots, u_q\}$ and denote $m(X_i, \theta) = (m(u_1, X_i; \theta), \ldots, m(u_q, X_i; \theta))$ as the vector of moment conditions, we also have $Em(X_i, \theta) = 0$ ($'0'$ is a vector of 0 here). Then, we can apply GMM estimation method to estimate the parameter θ. Specifically, we minimize the GMM target function:

$$\hat{\theta}_{GMM} = \arg\min_{\theta} \frac{1}{n} \sum_{j=1}^{n} m(X_j; \theta)' W_n \frac{1}{n} \sum_{j=1}^{n} m(X_j; \theta),$$

where W_n is a positive semidefinite weighting matrix. Under some regularity conditions, the GMM estiamte $\hat{\theta}_{GMM}$ is consistent and asymptotic normally distributed. A good choice of W_n can further make the asymptotic variance of the GMM estiamte reach a lower bound. We refer Hansen (1982) for detailed properties of GMM estimates.

However, the asymptotic variance lower bound of GMM estimates highly depends on the choice of u. An interesting question will be how to select the optimal discrete points u to minimize the lower bound of GMM estimates' asymptotic variance, finally reaching the Cramér-Rao lower bound (maximum likelihood efficiency). Feuerverger and McDunnough (1981*b*) shows that Cramér-Rao lower bound can be reached by sufficiently large number of discrete points and the grid also should be sufficiently dense and extended. This might suggest that a 'continuous' weight function with support on the sufficient large space might be more appropriate than a weight matrix with finite discrete points u.

To use this 'continuous' weight function, Heathcote (1977) and Knight and Yu (2002) propose the ECF method, which minimizes the integral:

$$\hat{\theta}_{ECF} = \arg\min_{\theta} \int |\phi(u, \theta) - \phi_n(u)|^2 dG(u),$$

where $G(u)$ is the weight function.

They also assume $G(u)$ is bounded and non-increasing to guarantee the consistency and normality of ECF estimate $\hat{\theta}_{ECF}$. This assumption implies that $G(u)$ is differentiable almost everywhere. We assume $g(u) = G'(u)$. Then, we have:

$$\hat{\theta}_{ECF} = \arg\min_{\theta} \int |\phi(u,\theta) - \phi_n(u)|^2 g(u) du. \tag{4.3}$$

To minimize the Equation (4.3), integration calculation is involved in, which might be computationally intensive. In this chapter, we show that trapezoidal rule can approximate this integration efficiently under mild regularity conditions on characteristic functions $\phi(u;\theta)$ and weight functions $g(u)$. Typical financial models such as Merton's jump-diffusion model, Kou's jump-diffusion model, normal inverse Gaussian (NIG) model and CGMY model can be applied, with a broad choice of weight function $g(u)$.

4.2.2 The Sinc expansion and trapezoidal rule approximation

Suppose we have the ECF target function:

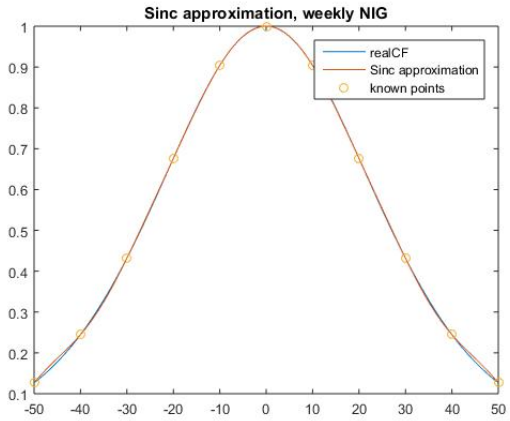
$$e(\theta; X) \equiv \int_{\mathbf{R}} |\phi(u;\theta) - \phi_n(u)|^2 g(u) du. \tag{4.4}$$

To do the integration in Equation (4.4), we need to know the value of $\phi(u;\theta)$, based on every single point of u. If we only know the value of characteristic function on an evenly spaced grids set $u \in \{-\mathcal{M}h, \ldots, \mathcal{M}h\}$, we can reconstruct the characteristic function based on its Sinc expansion $C_{h,\mathcal{M}}(\phi)(u;\theta)$:

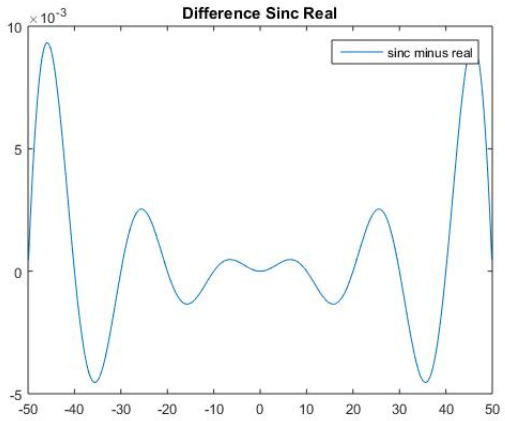
$$C_{h,\mathcal{M}}(\phi)(u;\theta) = \sum_{j=-\mathcal{M}}^{\mathcal{M}} \phi(jh;\theta) \frac{\sin(\pi(u-jh)/h)}{\pi(u-jh)/h},$$

where $\mathcal{M}$ and h are two tuning parameters of the evenly spaced grids $\{-\mathcal{M}h, \ldots, \mathcal{M}h\}$, controlling the error of the approximation.

Feng and Lin (2013) shows only a few points can recover $\phi(u;\theta)$ very well based on the Sinc expansion when $\phi(u;\theta)$ has good properties in analytic strip. Figure 4.1 shows the minor difference between the characteristic function of NIG model (will be introduced in Chapter 4.3) and its Sinc expansion approximation with $\mathcal{M} = 5$ and $h = 10$. The difference between the Sinc expansion approximation and corresponding characteristic function is in the 10^{-3} level and even smaller when u is around 0. Thus, we can find only 11 points in the evenly spaced grids $\{-\mathcal{M}h, \ldots, \mathcal{M}h\}$ can approximate the characteristic function very well (only real part of the characteristic function is shown here). This is due to the analyticity of the NIG model's characteristic function. This inspires us that good analyticity property can also lead to the approximation



(a) Sinc expansion approximation



(b) Difference between Sinc and Real

Figure 4.1: (Real part of) characteristic function of normal inverse Gaussian process (NIG) weekly increments and its corresponding Sinc expansion approximation. Parameters of NIG processes are $\alpha = 50,\ \beta = -5,\ \lambda = 5$ and $\mu = 0$. The First figure shows both NIG's characteristic function and its Sinc approximation. The second figure shows the difference between them.

of the ECF target in Equation (4.4), which is a function of the characteristic function.

To reconstruct the integral in Equation (4.4), we try to use the trapezoidal rule based on evenly spaced grids. Hopefully, when $\phi(u;\theta)$ and $g(u)$ have good analytic properties, we can also find that only a few grid points can approximate the Equation (4.4) very well.

Specifically, we have the following trapezoidal rule approximation of Equation (4.4):

$$e^{\mathcal{M},\hbar,a}(\theta;X) = \sum_{j=-\mathcal{M}}^{\mathcal{M}} |\phi(j\hbar + ia;\theta) - \phi_n(j\hbar + ia)|^2 g(j\hbar + ia)\hbar, \tag{4.5}$$

where $\mathcal{M}$ and $\hbar$ are two tuning parameters of the evenly spaced grids $\{-\mathcal{M}\hbar, \ldots, \mathcal{M}\hbar\}$, controlling the error of the approximation. a controls the shifted horizontal line $\{x + ia; x \in \mathbf{R}\}$ to conduct trapezoidal rule, by the Cauchy integral theorem.

The approximated ECF estimate is defined as:

$$\hat{\theta}_n^{\mathcal{M},\hbar,a} = \arg\min_{\theta\in\Theta} e^{\mathcal{M},\hbar,a}(\theta;X).$$

Before describing trapezoidal rule approximation performance and asymptotic properties of this approximated ECF estimate $\hat{\theta}_n^{\mathcal{M},\hbar,a}$, we need the following regularity conditions of A for characteristic functions $\phi(u;\theta)$ and weight functions $g(u)$:

A.1 For any $\theta \in \Theta$, $|\phi(u;\theta) - \phi_n(u)|^2 g(u)$ is analytic in strip $\mathcal{D}_{[d_-,d_+]}$ where $\mathcal{D}_{[d_-,d_+]} = \{z \in \mathbf{C} : \Im(z) \in [d_-, d_+]\}$, $-\infty < d_- < 0 < d_+ < \infty$ and d_-, d_+ do not depend on θ. $\Re(z)$ is the real part and $\Im(z)$ is the imaginary part of z.

A.2 For any given $x \in \mathbf{R}$, $|\phi(x + id_\pm;\theta)|$ is uniformly bounded by C with respect to $\theta \in \Theta$.

A.3 $\int_{d-}^{d+} |g(x+iy)|dy \to 0$ when $x \to \pm\infty$.
$||g||^\pm := \int_R |g(x + i(d_\pm))|dx < +\infty$.

A.4 $|\phi(x+i\mathit{a};\theta) - \phi_n(x+i\mathit{a})|^2 |g(x+i\mathit{a})| \le \mathit{k}|x|^{\mathit{n}} \exp(-\mathit{c}|x|^\nu)$, $x \in \mathbf{R}$ with $\mathit{a} \in [d_-, d_+]$ for some $\mathit{k} > 0, \nu > 0, \mathit{c} > 0, \mathit{n} \in \mathbf{R}$ or $\mathit{k} > 0, \nu > 0, \mathit{c} = 0, \mathit{n} < -1$. Here, k, ν, c, n may or may not depend on a and not related to the parameter $\theta \in \Theta$. a is trapezoidal approximation parameter in Equation (4.5), which a given real value, $\mathit{a} \in [d_-, d_+]$.

A.5 $\mathcal{M}\mathit{h} \ge (\mathit{n}/\mathit{c}\nu)^{1/\nu} 1_{\mathit{c}>0,\mathit{n}>0}$ for $\mathit{n}, \mathit{c}, \nu$ defined above.

Remark 11. Regularity condition A.1 and A.2 are the conditions for characteristic function. Considering $\phi_n(u)$ is a simple function, we only need $\phi(u : \theta)$ and $g(u)$ is analytic in the strip $\mathcal{D}_{[d_-,d_+]}$. For regularity condition A.2, because our characteristic function has closed form, it is easy to verify when a compact parameter space Θ is given. Actually, it is easy prove that this is true for common Lev́y processes such as Merton's jump-diffusion model, Kou's jump-diffusion model, normal inverse Gaussian model and CGMY model. Regularity condition A.3 and A.4 are two regularity conditions for the continuous weight function $g(u)$. For the most weight functions with exponential tail, A.3 is easily verified. For the regularity condition A.4, in the usual case, we do the trapezoidal rule over the real line $\mathbf{R}$. That is, $\mathit{a} = 0$. Then, $|\phi(x + i\mathit{a};\theta) - \phi_n(x+i\mathit{a})|$ are obviously bounded by 2. A.4 generally requires that $g(u)$ at least has a polynomial tail. $g(u)$ can be one of most smooth probability densities with exponential tail on the whole real line such as normal distribution, Gumbel distribution, normal inverse Gaussian distribution, which are easily verified to satisfy Regularity conditions. A.5 is the condition for tuning parameters $\mathcal{M}$ and h which are trivial to verify.

Now, we have following theorem to control the bound of trapezoidal rule approximation error based on the regularity conditions of class A:

Theorem 18. *Under regularity conditions of class A, we define the error of approximation by trapezoidal rule as*

$$\begin{aligned} E^F_{\mathit{h},\mathcal{M}}(e, \mathit{a})(\theta; X) &= e(\theta; X) - e^{\mathcal{M},\mathit{h},\mathit{a}}(\theta; X) \\ &= \int_{\boldsymbol{R}} |\phi(u;\theta) - \phi_n(u)|^2 g(u)du - \sum_{j=-\mathcal{M}}^{\mathcal{M}} |\phi(j\mathit{h} + i\mathit{a};\theta) - \phi_n(j\mathit{h} + i\mathit{a})|^2 g(j\mathit{h} + i\mathit{a})\mathit{h}. \end{aligned}$$

Then, we have the bound of error

$$|E^F_{\hbar,\mathcal{M}}(e,\boldsymbol{a})(\theta;X)| \leq \frac{e^{-2\pi(a-d_-)/\hbar}}{1-e^{-2\pi(a-d_-)/\hbar}}C_1||g||^- + \frac{e^{-2\pi(d_+-a)/\hbar}}{1-e^{-2\pi(d_+-a)/\hbar}}C_2||g||^+ + T_{\mathcal{M}\hbar} \tag{4.6}$$

where $T_{\mathcal{M}\hbar} = \frac{2\mathfrak{K}}{|n+1|}(\mathcal{M}\hbar)^{n+1}$ *if* $\mathfrak{c} = 0, \mathfrak{n} < -1$, *and* $T_{\mathcal{M}\hbar} = \frac{2\mathfrak{K}}{\nu\mathfrak{c}^{(n+1)/\nu}}\Gamma(\frac{n+1}{\nu}, \mathfrak{c}(\mathcal{M}\hbar)^\nu)$ *if* $\mathfrak{c} > 0$. *Incomplete Gamma function* $\Gamma(s,b) = \int_b^\infty e^{-t}t^{s-1}dt$. $C_1 = C + 1 + \sum_{j=1}^n \exp(-d_-X_j)$ *and* $C_2 = C + 1 + \sum_{j=1}^n \exp(-d_+X_j)$ *which are not related to parameter* θ *(*C *is the uniform bound in regularity condition A.2.). Moreover, let* $\mathcal{M}\hbar \to \infty$ *and* $\hbar \to 0$, *then, the bound of error* $E^F_{\hbar,\mathcal{M}}(\phi,\boldsymbol{a})(x)$ *will decay to zero uniformly on* $\theta \in \Theta$. *That is, with any given data* X, $e^{\mathcal{M},\hbar,\boldsymbol{a}}(\theta;X)$ *converges to* $e(\theta;X)$ *uniformly for* $\theta \in \Theta$ *when* $\mathcal{M}\hbar \to \infty$ *and* $\hbar \to 0$.

Thus, under regularity conditions of class A, we have the uniform convergence with respect to $\theta \in \Theta$ of the trapezoidal approximation. Moreover, the first two terms of Equation (4.6) are only related to $\hbar$ and have an exponential decay rate. The last term of Equation (4.6) is related to $\mathcal{M}\hbar$ and also has an exponential decay rate. We denote the first two terms as discretization error bound and the last term as truncation error bound. More explanations are in the proof.

Thus, from the error bound listed in Equation (4.6), we find the bound of error decays very quickly (exponential rate) with respect to bigger $\mathcal{M}\hbar$ and smaller $\hbar$. Considering the evenly spaced grids $\{-\mathcal{M}\hbar, \ldots, \mathcal{M}\hbar\}$, to control the error bound to be small, we even don't need very small interval length $\hbar$ or very big extension $\mathcal{M}\hbar$ due to the exponentially decay of error bound. In fact, implied by a lot of cases in our simulation study, $\hbar = 10$ can give us a very good approximation.

4.2.3 Asymptotic properties of the approximated empirical characteristic function estimation

In this section, we provide asymptotic properties of our approximated ECF estimate $\hat{\theta}_n^{\mathcal{M},\hbar,a}$.

Based on Theorem 18, we have the following asymptotic properties between ECF estimate and approximated ECF estimate.

Theorem 19. *Suppose* $\hat{\theta}_n$ *is the unique ECF estimate defined as* $\arg\min_\theta e(\theta,X)$ *with large enough sample size* n. *We assume parameter space* Θ *is compact and* $e(\theta,X)$ *is continuous with respect to* θ. *Fixing large enough sample size* n, *under A class regularity conditions*

$$\hat{\theta}_n^{\mathcal{M},\hbar,a} \xrightarrow{p} \hat{\theta}_n, \tag{4.7}$$

when $\hbar \to 0$ *and* $\mathcal{M}\hbar \to \infty$ *with fixed* a.

Theorem 20. *We introduce B class of regularity conditions in appendix C.1.1 to guarantee the asymptotic properties of ECF. Under A class of regularity condition and B class of regularity conditions , there exists* $\mathcal{M}(n)$ *and* $\hbar(n)$ *with respect to* n *and for Approximated ECF estimate* $\hat{\theta}_n^{\mathcal{M},\hbar,a}$, $\hat{\theta}_n^{\mathcal{M}(n),\hbar(n),a} \xrightarrow{p} \theta_0$ *with fixed* a *when* $n \to \infty$. *Furthermore,*

$$\sqrt{n}(\hat{\theta}_n^{\mathcal{M}(n),\hbar(n),a} - \theta_0) \xrightarrow{d} N(0, B^{-1}(\theta_0)A(\theta_0)B^{-1}(\theta_0)),$$

where $A(\theta_0) = var(K(X_1;\theta_0))$.

Remark 12. The regularity conditions in B class are borrowed from Knight and Yu (2002). The first 4 regularity conditions in B class are responsible for consistency of ECF estimate , while, the last 3 regularity conditions in B class are specifically designed for the asymptotic normality of ECF estimates. Combine B class regularity conditions with A class regularity conditions, the approximated ECF estimate $\hat{\theta}_n^{\mathcal{M},\hbar,a}$ is consistent and asymptotic normally distributed.

All in all, under certain regularity conditions, consistency, asymptotic normality can be reached by the approximated ECF estimate by utilizing trapezoidal rule approximation in Equation (4.5).

4.3 Implementation

In this section, we study the detailed implementation of our approximated ECF estimation based on trapezoidal rule. Specifically, we list some typical Levy processes in quantitative finance which satisfying our regularity conditions of class A and B. Then, we propose some suggestions on the selection of the weight function $g(u)$ and tuning parameters $\mathcal{M}$, $\hbar$, d_+ and d_- in our approximated ECF estimation.

4.3.1 Selected Lévy processes

We first introduce some typical Lévy processes. Lévy processes are commonly used in Finance due to its flexibility to model heavy tails and skewness of financial time series. One of its property is that they have independent stationary increments. That is, if we assume equity value series are $S_t = \exp(Y_t)$ and $\{Y_t\}$ is a Lévy process, the logarithm of the evenly spaced equity returns $\{X_t = Y_{t\delta} - Y_{(t-1)\delta}\}_1^n$ will follow the same distribution F, where evenly spaced timestamps are defined as $(0, \delta, 2\delta, \ldots, n\delta)$. Then, we can use the i.i.d. data to fit the model. Another important property of Levy processes is Lévy-Khintchine formula. This

formula provides implicit form of the characteristic function of Y_t:

$$\phi(u) = \exp(\delta(iua - \frac{bu^2}{2} + \int_R (\exp(iuy) - 1 - iux1_{|y|\leq 1})J(dy))), \tag{4.8}$$

where (a, b, J) is called Lévy triplet which fully determines a Lévy process Y_t. $a \in R$ is drift parameter, $\sigma \geq 0$ is the diffusion component and Lévy measure $J(dx)$ satisfying $J(\{0\}) = 0$ and $\int \min(1, x^2)J(dx) < \infty$. We can derive the characteristic function via Lévy-Khintchine formula no matter the density function of Lévy processes is available or not.

In this chapter, we will implement our parameter estimation algorithm for several typical Lévy processes including: Merton's jump-diffusion model (Merton (1976)), Kou's jump-diffusion model (Kou (2002)), NIG model (Barndorff-Nielsen (1997)) and CGMY model (Carr et al. (2002)). Merton's jump-diffusion model and Kou's jump-diffusion model belong to the finite activity Lévy processes class, which allow only finitely many jumps in any given time interval, while, infinite activity Lévy processes include infinitely many jumps in any given time interval including NIG model and CGMY model.

Merton's jump-diffusion model

In Merton's jump-diffusion model, the observed stock price, S_t, satisfying the following equation:

$$X_t \equiv \log(S_{(t+1)\delta}/S_{t\delta}) = \mu\delta + \sigma\sqrt{\delta}Z + \sum_{i=N_t+1}^{N_{t+1}} Z_i, \tag{4.9}$$

where δ is the evenly spaced time interval of observed data; μ is drift; σ is volatility; Z is standard $N(0, 1)$; N_t is Poisson process with intensity λ; $\{Z_i\}$ are jump sizes following i.i.d $f_z(x) \sim N(\mu_j, \sigma_j^2)$. That is, in Merton's jump-diffusion model, jumps occur according to a Poisson process N_t with jump sizes $\{Z_i\}$.

The Lévy triplet of this model is $(\mu, \sigma^2, J(dx) = \lambda f_z(x)dx)$. Although, the density of X_t has a complex form (Tankov (2003)), its characteristic function, $\phi(u; \theta)$, has a simple form through Lévy-Khintchine formula:

$$\phi(u; \theta) = \exp(\delta(iu\mu - \sigma^2u^2/2 + \lambda(\exp(i\mu_j u - \sigma_j^2 u^2/2) - 1))). \tag{4.10}$$

We set the following compact parameter space: $-\infty < L_\mu \leq \mu \leq U_\mu < \infty$, $0 < L_\sigma \leq \sigma \leq U_\sigma < \infty$, $0 < L_\lambda \leq \lambda \leq U_\lambda < \infty$, $-\infty < L_{\mu_j} \leq \mu_j \leq U_{\mu_j} < \infty$, $0 < L_{\sigma_j} \leq \sigma_j \leq U_{\sigma_j} < \infty$, and true parameter $\theta_0 \in \Theta$. For regularity conditions, we can show that d_+ and d_-, satisfying A.1 and A.2, could be arbitrary positive number and negative number due to this compact parameter space setting and the holomorphic property of the characteristic function of Merton's jump-diffusion model. Regularity condition A.4 also

holds, if $|g(x+i\boldsymbol{a})|$ has an exponential or binomial decay tail due to uniform boundedness of $\phi(x+i\boldsymbol{a};\theta)$ and $\phi_n(x+i\boldsymbol{a})$ with respect to the parameter $\theta \in \Theta$. Thus, Merton's jump-diffusion model can easily fulfill the A class of regularity conditions.

To simulate Merton's jump-diffusion model in the simulation study, we can utilize the method described in Glasserman (2003).

Kou's jump-diffusion model

Merton's jump-diffusion model assumes the jump size is symmetric following normal distribution. Kou's jump-diffusion model allows the distribution of jump sizes $\{Z_i\}$ to be asymmetric following double exponential distribution. To be specific, the observed stock price, S_t, satisfying the following equation:

$$X_t \equiv \log(S_{(t+1)\delta}/S_{t\delta}) = \mu\delta + \sigma\sqrt{\delta}Z + \sum_{i=N_t+1}^{N_{t+1}} Z_i, \tag{4.11}$$

where jump size here $\{Z_i\}$ follows i.i.d. double exponential distribution and its distribution density is

$$f_z(x) = p\eta_u \exp(-\eta_u x 1_{\{x>0\}}) + (1-p)\eta_d \exp(\eta_d x 1_{\{x<0\}}),$$

where p is the positive jump probability; $1/\eta_u$ is mean positive jump size; $1/\eta_d$ is mean negative jump size.

The Lévy triplet of this model is $(\mu, \sigma^2, J(dx) = \lambda f_z(x)dx)$. The density of X_t has a complex form (Ramezani and Zeng (2007)). Its characteristic function, $\phi(u;\theta)$, has a simple form through Lévy-Khintchine formula

$$\phi(u;\theta) = \exp(\delta(iu\mu - \sigma^2u^2/2 - \lambda(1 - \frac{p\eta_u}{\eta_u - iu} - \frac{(1-p)\eta_d}{\eta_d + iu}))).$$

For regularity conditions, $d_+ \in (0, \eta_d)$ and $d_- \in (-\eta_u, 0)$ can make the characteristic function analytic and also satisfy A.2 with following parameter space: $-\infty < L_\mu \le \mu \le U_\mu < \infty$, $0 < L_\sigma \le \sigma \le U_\sigma < \infty$, $0 < L_\lambda \le \lambda \le U_\lambda < \infty$, $0 < L_p \le p \le U_p < 1$, $0 < L_{\eta_u} \le \eta_u \le U_{\eta_u} < \infty$, $0 < L_{\eta_d} \le \eta_d \le U_{\eta_d} < \infty$, and true parameter $\theta_0 \in \Theta$. Then, Θ is a compact parameter space. To make this model satisfy regularity condition A.1, d_+ and d_- should not be related to parameters. We can let $d_+ = L_{\eta_d}/2$ and $d_- = -L_{\eta_u}/2$ due to our compact parameter setting. This setting is generally practical because true parameter of η_u and η_d are positive which cannot be zero exactly. Regularity condition A.4 also holds, if $|g(x+i\boldsymbol{a})|$ has an exponential or binomial decay tail due to uniform boundedness of $\phi(x+i\boldsymbol{a};\theta)$ and $\phi_n(x+i\boldsymbol{a})$ with respect to the parameter $\theta \in \Theta$. Thus, Kou's jump-diffusion model can easily fulfill the regularity conditions of class A.

To simulate it in simulation study, we also utilize the method described in Glasserman (2003).

Normal inverse Gaussian model

Normal inverse Gaussian (NIG) model belongs to a more general class of Lévy processes, generalized hyperbolic model (Eberlein et al. (1998)). It can be characterized by

$$X_t \equiv \log(S_{(t+1)\delta}/S_{t\delta}) = \mu\delta + \beta z_\delta + \lambda W_{z_\delta}, \tag{4.12}$$

where z_δ is the first time when a Brownian motion with drift γ reaches the positive level δ. The density of z_δ is inverse Gaussian (IG) distribution. W_{z_δ} is a Brownian motion of which the calendar time is a random time z_δ.

Set $\alpha = \sqrt{\beta^2 + \gamma^2}$, we have the Lévy triplet $(\mu, 0, J(dx) = f(x)dx)$ where

$$f(x) = \frac{\lambda\alpha}{\pi|x|} \exp(\beta x) K_1(\alpha|x|).$$

$K_n(x)$ is the modified Bessel function of the second kind with order n. Then, the probability density function of X_t contains Bessel function but the characteristic function is in a simple form

$$\phi(u;\theta) = \exp(\delta(iu\mu - \lambda(\sqrt{(\alpha^2 - (\beta + iu)^2)} - \sqrt{\alpha^2 - \beta^2}))).$$

For regularity conditions of class A, $d_+ \in (0, \beta + \alpha)$ and $d_- \in (\beta - \alpha, 0)$ can make the characteristic function analytic and fulfill A.2 as well, with following parameter space: $-\infty < L_\mu \leq \mu \leq U_\mu < \infty$, $\infty < L_\alpha \leq \alpha \leq U_\alpha < \infty$, $\infty < L_\beta \leq \beta \leq U_\beta < \infty$, $0 < L_{\alpha-|\beta|} \leq \alpha - |\beta|$, $0 < L_\lambda \leq \lambda \leq U_\lambda \leq \infty$, and true parameter $\theta_0 \in \Theta$. Then, Θ is a compact parameter space. We can simply let $d_+ = L_{(\alpha-\beta)}/2$ and $d_- = -L_{\alpha-\beta}/2$ to make sure the characteristic function is analytic in $\mathcal{D}_{[d_-,d_+]}$ which is not related parameters. Then, regularity condition A.1 holds. Regularity condition A.4 also holds, if $|g(x + i\boldsymbol{a})|$ has an exponential or binomial decay tail due to uniform boundedness of $\phi(x + i\boldsymbol{a};\theta)$ and $\phi_n(x + i\boldsymbol{a})$ with respect to the parameter $\theta \in \Theta$. Thus, NIG model can easily fulfill the regularity conditions of class A.

To simulate NIG process X_t in simulation study, we refer the method in Rydberg (1997).

CGMY model

CGMY model can be regarded as the generalization of Variance Gamma (Madan and Seneta (1990)). The Lévy triplet of it is $(\mu, 0, J(dx) = f(x)dx)$ and $f(x)$ is:

$$f(x) = \begin{cases} C\frac{\exp(-Mx)}{x^{1+Y}} & x > 0 \\ C\frac{\exp(-Gx)}{|x|^{1+Y}} & x < 0, \end{cases} \tag{4.13}$$

where $C > 0, G \geq 0, M \geq 0, Y < 2$.

In addition, when $Y < 0$, CGMY is a finite activity process. It could be regarded as a compound Poisson process. When $0 < Y < 2$, the process is an infinite activity process. The case, when $Y = 0$, is variance gamma process. Here, we only consider infinite activity CGMY process with $0 < Y < 2$.
The density function of CGMY process $\{X_t\}$ is unknown for us. However, the characteristic function is available for us.

For $Y \in (0, 1) \cup (1, 2)$:

$$\phi(u; \theta) = \exp(\delta(i\mu u + C\Gamma(-Y)[(M - iu)^Y - M^Y + (G + iu)^Y - G^Y])), \tag{4.14}$$

and for $Y = 1$:

$$\phi(u; \theta) = \exp(\delta(i\mu u + C((M - iu)\log(1 - iu/M) + (G + iu)\log(1 + iu/G) - iu(\log(M) - \log(G))))). \tag{4.15}$$

It is easily verified that the characteristic function of CGMY model is smooth with respect to its parameters. For regularity conditions, $d_+ \in (0, G)$ and $d_- \in (-M, 0)$ can make the characteristic function analytic in $\mathcal{D}_{[d_-,d_+]}$. To further make CGMY model satisfy regularity condition A.1, we can let $d_+ = L_G/2$ and $d_- = -L_M/2$, so that $\mathcal{D}_{[d_-,d_+]}$ not related to parameters. A.2 and A.4 can be easily verified when $|g(x+i\boldsymbol{a})|$ has an exponential or binomial decay tail and parameter space are set to be: $-\infty < L_\mu \leq \mu \leq U_\mu < \infty$, $0 < L_C \leq C \leq U_C < \infty$, $0 < L_G \leq G \leq U_G < \infty$, $0 < L_M \leq M \leq U_M < \infty$, $0 < L_Y \leq Y \leq U_Y < 2$. and true parameter $\theta_0 \in \Theta$. Thus, CGMY also satisfy regularity conditions of class A.

We can simulate CGMY process by inverting characteristic function introduced in Chen et al. (2012).

4.3.2 Selection of weights function

To do ECF estimation, we need to select a weight function $g(u)$. It can be shown that the ECF estimation can reach maximum likelihood efficiency by choosing an optimal weight function (Feuerverger and McDunnough

(1981a)). However, this weight function is based on Fourier inversion transform of a function of log likelihood which is unknown when the likelihood function has no closed form. If we only need the estimator to be consistent and asymptotic normally distributed, any bounded and non-decreasing $G(u)$ can guarantee it, where $g(u) = G'(u)$ (Knight and Yu (2002)).

Our regularity condition A.1 implies that analytic functions in strip $\mathcal{D}_{[d_-,d_+]}$ could be a good choice for $g(u)$. A.3 and A.4 should hold when this analytic function has an exponential decay tails. Actually, we can choose $g(u)$ to be a probability density function of common distributions on the whole real line such as normal distribution, Gumbel distribution, normal inverse Gaussian distribution. They all satisfy the regularity condition A.1, A.3 and A.4. We select probability density function of the normal distribution to be our weight distribution in this work and prove that it satisfying our regulatory condition A.1, A.3 and A.4 as following. In the simulation study, we find that normal distribution weight function have a good and efficient estimation performance.

Proposition 21. *if $g(u)$ follows normal distribution $N(\mu, \sigma^2)$, it satisfies our regularity condition of class A.*

The normal distribution also suggests putting more weights on the points around the origin, which is also indicated by our Proposition 22.

Proposition 22. *If we fix u, following the central limit theorem, we have:*

$$\sqrt{n}(\Re(\phi_n(u)) - \Re(\phi(u,\theta))) \xrightarrow{d} N(0, \frac{1}{2} + \frac{1}{2}\Re(\phi(2u,\theta)) - (\Re(\phi(u,\theta)))^2),$$

$$\sqrt{n}(\Im(\phi_n(u)) - \Im(\phi(u,\theta))) \xrightarrow{d} N(0, \frac{1}{2} - \frac{1}{2}\Re(\phi(2u,\theta)) - (\Im(\phi(u,\theta)))^2).$$

Proposition 22 provides the central limit theorem of the empirical characteristic function with fixed u. It shows that moment matching is generally easier for u around the origin than the large enough u. That is, the asymptotic variance is smaller around the origin and bigger when u is sufficient large. Because $\phi(u,\theta)$ is available for Lévy processes, we can even write down an implicit expression of the asymptotic variance. For the typical Lévy listed in our work, $|\phi(u,\theta)| \to 0$ when $u \to 0$. Thus, it is easily verified that asymptotic variance of both real and imaginary part of empirical characteristic function is around $\frac{1}{2}$ with sufficient large u compared with the 0 asymptotic variance around the origin. That is, it might be more appropriate to put more weights around the origin of u.

Suppose we have a weight distribution $g(u)$, we also need to pin down the parameters within that

distribution. We describe a way to do it, which gives us a stable and good estimation performance. The general idea is to capture all useful information of the characteristic function where its value deviated from 0. However, when u is large, empirical ECF could be very noisy and volatile. This is also supported by the large asymptotic variance of large u in Proposition 22. Thus, we can set a threshold value and mainly use the information when empirical likelihood function value larger than this value (which is not noisy and volatile). Thus, we define the ECF information L and its threshold $L_{threshold}$ as following:

$$L = \{u : L_{threshold} \leq \Re(\phi_n(u)) \leq 1\}, \tag{4.16}$$

where $L_{threshold}$ can be selected to be $5 * \sqrt{(1/2n)}$ and n is the sample size. The reason for this selection is $\sqrt{(1/2n)}$ is the standard deviation of $\phi(u, \theta)$ with sufficient large u and sufficient large sample size n (See Proposition 22). $5 * \sqrt{(1/2n)}$ can exclude noises of $\phi_n(u)$ with large u very well in practice because of the asymptotic normal distribution of $\phi_n(u)$ indicated in Proposition 22. We use the real part of the empirical characteristic function to find L because imaginary part is relative more volatile in practice. To find L efficiently given $L_{threshold}$, binary search algorithm can be utilized.

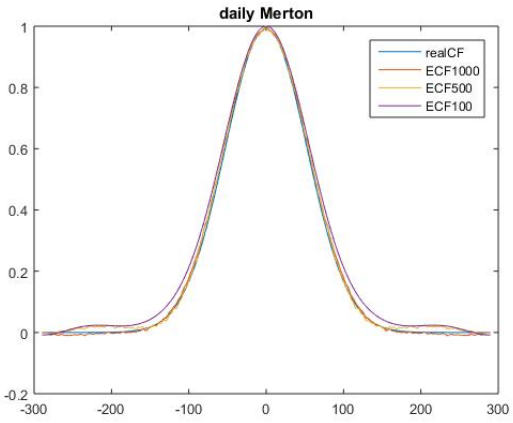
Figure 4.2 describe the real part and imaginary part of the empirical characteristic function and the corresponding characteristic function of Merton's jump-diffusion model. The empirical characteristic function is defined in Equation (4.2) based on 100, 500 and 1000 simulated data and the characteristic function is from Equation (4.10). The model's parameter is selected to be same as the one in our simulation study. We can find that empirical characteristic function is more closed to real characteristic function when u is more closed to origin and sample size is bigger. When u is bigger, empirical characteristic function is more volatile and can deviate more from the morel's characteristic function.

Figure 4.3 shows an example to choose L and $L_{threshold}$ for NIG model with simulated sample size 100. $L_{threshold} = 0.35$ from the definition in (4.16). L value can be found by binary searching the value $L_{threshold}$ on empirical characteristic function of sample size 100. We can find the interval $u \in [-L, L]$ covers the most area that empirical characteristic function closed to the model's characteristic function.

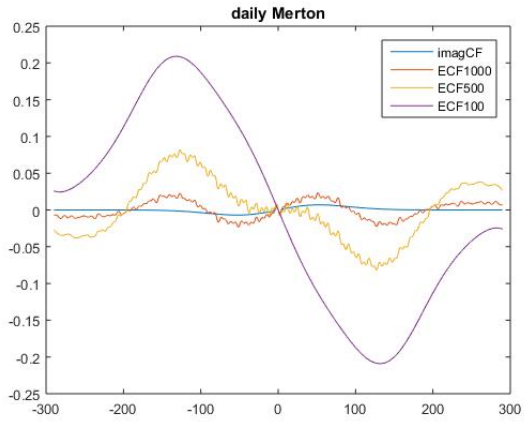
When we have ECF information value L, we can design the parameter in $g(u)$ to cover this interval $[-L, L]$. For example, if $g(u) \sim N(\mu, \sigma^2)$. We can let $\mu = 0$ and $3\sigma = L$ due to 3σ covers the interval $[-L, L]$ in probability more than 99.5%.

In summary, our weight function can be selected in three steps when we have data:

- Select $g(u)$ satisfying regularity conditions A.1, A.3 and A.4 (for example, normal distribution density:$N(\mu, \sigma^2)$).
- Binary search ECF information value L on empirical characteristic function regarding (4.16) with



(a) Real part of Merton's CF



(b) Imaginary part of Merton's CF

Figure 4.2: Characteristic function of Merton's model and its corresponding empirical characteristic function with sample size 100, 500 and 1000. Parameters are $\mu = 0.1,\ \sigma = 0.3,\ \lambda = 10,\ \mu_j = -0.5$ and $\sigma_j = 0.25$. Blue curve is the characteristic function. Red, yellow and purple curves represent empirical characteristic function with 1000, 500, 100 simulated samples.

$L_{threshold} = 5 * \sqrt{(1/2n)}$.

- Design parameters in $g(u)$ to cover interval $[-L, L]$ well (for normal distribution $N(\mu, \sigma)$, $\mu = 0$ and $\sigma = L/3$).

4.3.3 Selection of tuning parameters in trapezoidal approximation

In this section, we discuss the way to select tuning parameters a, $\mathcal{M}$ and h in trapezoidal approximation (4.5). From Theorem 18, we find that approximation error will decay to zero when $\mathcal{M}h \to \infty$ and $h \to 0$. d_a is defined as $2\min(d_+ - a, a - d_-)$. Furthermore, the first two terms of Equation (4.6) implies that the decay rate is exponential, which is $\exp(-\pi d_a/h)$. The last term of Equation (4.6) implies that the decay rate is also exponential which is $\exp(-c(\mathcal{M}h)^\nu)$ if $c > 0$. This suggests that we can select h to make three terms in Equation (4.6) converge to zero at the same exponential rate. That is, we have:

$$\exp(-\pi d_a/h) = \exp(-c(\mathcal{M}h)^\nu).$$

Then,

$$h(\mathcal{M}) = (\pi d_a/c)^{\frac{1}{\nu+1}} \mathcal{M}^{-\frac{\nu}{\nu+1}}, \tag{4.17}$$

which is also suggested by Feng and Lin (2013). For a, we simply select $a = 0$ in this work and it has efficient and good performance.

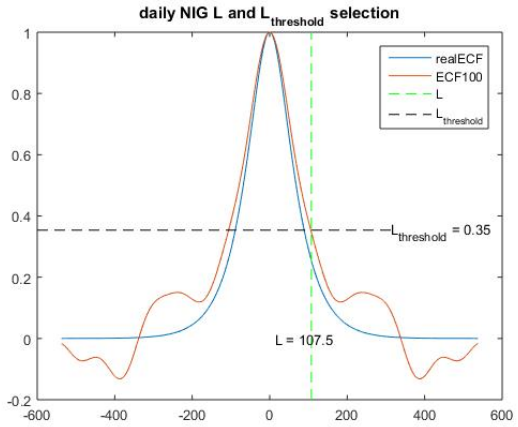


Figure 4.3: Characteristic function of NIG model and its corresponding empirical characteristic function with sample size 100. Parameters are $\mu = 0,\ \alpha = 50,\ \beta = -5,\ \lambda = 5$. Blue curve is the characteristic function. Red curve represents empirical characteristic function with 100 simulated samples. Green dashed line indicates that ECF information value $L = 107.5$ based on the threshold $L_{threshold} = 0.35$ (calculated from the definition in (4.16))

.

If $\mathcal{M}$ is given, we still need to know $d_a, \mathfrak{c}$ and ν. $\mathfrak{c}$ and ν parameters can be simply determined from regularity condition A.4. For example, if we define continuous weight function $g(u)$ to be normal distribution density function $N(\mu, \sigma^2)$ and $\mathfrak{a} = 0$, the corresponding $\nu = 2$ and $\mathfrak{c} = \frac{1}{2\sigma^2}$ which is $\frac{9}{2L^2}$ given that $\sigma = \frac{L}{3}$ suggested in Chapter 4.3.2.

For d_a, we notice that we generally like bigger d_a to have bigger $\hbar(\mathcal{M})$. Then, if we the fix grid extension $\mathcal{M}\hbar$, $\mathcal{M}$ could be smaller which reduce our computation burden ($\mathcal{M}$ is the number of grid points in trapezoidal rule.). We select different d_a based on different given parameter in each iteration of the optimization program. To explain it in details, recall that we need search the estimated parameter by minimize Equation (4.5). The optimization program will evaluate Equation (4.5) in each iteration with given parameter value θ. Suppose the parameter value is given in one optimization iteration, we can determine the corresponding biggest d_a, which is the value to make sure characteristic function is analytic in a strip (regularity condition A.1). This provides a connection between d_a and θ. Specifically, Chapter 4.3.1 provides a general guide to choose $d_\pm$ based on parameters. For example, Merton's jump-diffusion model can allow arbitrary $d_\pm$ because its characteristic function is holomorphic. In practice, we can choose $d_a = 10$ with a good performance. Kou's jump-diffusion model need $d_+ \in (0, \eta_d)$ and $d_- \in (-\eta_u, 0)$. We can simply make d_a to be a value slightly smaller than $2\min(\eta_u, \eta_d)$. Similarly, we can make d_a to be a value slightly smaller than $2\min(\alpha - \beta, \alpha + \beta)$ for NIG model and d_a to be a value slightly smaller than $2\min(G,\ M)$ for CGMY model given parameter in each iteration of the optimization program.

Once we can determine $\hbar$ given $\mathcal{M}$ and parameter θ, we only need to determine $\mathcal{M}$. We can do it in two

ways.

In the first way, we can select $\mathcal{M}$ in each iteration of the optimization program. That is, when we minimize Equation (4.5), optimization program will evaluate it in each iteration with given parameter value θ. Thus, for a fix parameter θ, we can increase $\mathcal{M}$ gradually to make $e^{\mathcal{M},\hbar,a}(\theta;X)$ stable with changes less than a error threshold (i.e. 10^{-8}). From our practical experience, if we set initial $\mathcal{M} = 50$, for most reasonable parameters in the parameter space, there is no need to increase $\mathcal{M}$ multiple times.

In the second way, we can simply try different fixed $\mathcal{M}$ and minimize $e^{\mathcal{M},\hbar,a}(\theta;X)$ based on fixed $\mathcal{M}$. Then, use the $\mathcal{M}$ when we get stable approximated ECF estimate $\hat{\theta}_n^{\mathcal{M},\hbar,a}$ (ECF estimate converges to a stable value by increasing $\mathcal{M}$). We conduct the second way in our simulation study.

We recommend using global optimization program which can also deal with non-smooth function. We choose 'NOMAD' optimization in 'OPTI' toolbox (Abramson et al. (n.d.)) of MATLAB to be our numerical optimization procedure which is stable and fast implied by our simulation study.

4.3.4 Verifications of regularity conditions

In this section, we want to show that the selected popular Lévy processes satisfy the regularity condition of class A and class B. If class A regularity conditions holds for those processes, our ECF target approximation (Equation (4.5)) converges to the ECF target (4.4) with the exponential rate. That is, our approximation will be accurate due to fast convergence. Also, if class B regularity conditions holds, our approximated ECF estimates will have good asymptotic properties.

Theorem 23. *If $g(u)$ satisfies the regularity conditions of class A (e.g. normal distribution shown in Proposition 21), the characteristic function of Merton's jump-diffusion model, Kou's jump-diffusion model, NIG model and CGMY model satisfy the regularity condition of class A if the analytic strip $\mathcal{D}_{[d_-,d_+]}$ and compact parameter space Θ are selected regarding to the Chapter 4.3.1.*

Theorem 24. *If $g(u)$ satisfies the regularity conditions of class A (e.g. normal distribution shown in Proposition 21), satisfies the regularity conditions of class A (e.g. normal distribution shown in Proposition 21), the characteristic function of Merton's jump-diffusion model, Kou's jump-diffusion model, NIG model and CGMY model satisfy the regularity condition of class B.*

In our simulation study, we just select $g(u)$ to be the probability density function of normal distribution (suggested by selection 4.3). This setting satisfies the requirements of Theorem 23 and Theorem 24.

Table 4.1: Parameter spaces

Merton	Kou	NIG	CGMY
True values			
$\mu = 0.1,\ \sigma = 0.3,\ \lambda = 20$	$\mu = 0.2,\ \sigma = 0.2,\ \lambda = 30$	$\mu = 0, \alpha = 50,\ \beta = -5,\ \lambda = 5$	$C = 3,\ G = 79,\ M = 83$
$\mu_j = -0.5,\ \sigma_j = 0.25$	$p = 0.4,\ \eta_u = 5,\ \eta_d = 2.5$		$Y = 0.9$
Parameter spaces			
$-1 \leq \mu \leq 1$	$-1 \leq \mu \leq 1$	$1 \leq \alpha - \lvert\beta\rvert$	$1 \leq C \leq 100$
$0.01 \leq \sigma \leq 1$	$0.01 \leq \sigma \leq 1$	$-10 \leq \beta \leq 10$	$1 \leq G \leq 100$
$1 \leq \lambda \leq 50$	$1 \leq \lambda \leq 50$	$0.5 \leq \delta \leq 10$	$1 \leq M \leq 100$
$-1 \leq \mu_j \leq 1$	$0.1 \leq p \leq 0.9$	$1 \leq \alpha \leq 150$	$0.1 \leq Y \leq 1.9$
$0.05 \leq \sigma_j \leq 1$	$1 \leq \eta_u \leq 20$		
	$1 \leq \eta_d \leq 20$		

4.4 Simulation Study

In this section, we conduct a simulation study to examine the performance of approximated ECF estimate. We mainly have two targets. First, we want to show that only a small $\mathcal{M}$ in Equation (4.5) is needed to approximate the ECF target (4.2) very well, when the model satisfies our regularity conditions. That is, we can nearly use few points to approximate ECF target without computational intensive integration. This is shown theoretically by Theorem 18 due to the very fast decay rate of approximation error (exponential decay). Second, we want to check if our approximated ECF estimate will match the true parameter based on simulated data. This is already implied by the asymptotic properties of approximated ECF estimates (Theorem 20).

We conduct the simulation study based on the Lévy processes described in Chapter 4.3.1 including Merton's jump-diffusion model, Kou's jump-diffusion model, NIG model and CGMY model.

For each Lévy model in Chapter 4.3.1, we fit the model based on simulated 500 sample paths with different sample size n, daily frequency ($\delta = 1/252$) and weekly frequency ($\delta = 1/52$). In each sample path, we simulate 5000 samples for each model and fit the model based on first 200, 1000 and 5000 samples separately. Based on each same sample, we also set fixed tuning parameter $\mathcal{M} = \{10, 30, 50, 100, 200\}$ to illustrate that in most cases, a small $\mathcal{M}$ ($\mathcal{M} = 10$ or $\mathcal{M} = 30$) can give us a very stable estimation result.

The large enough compact parameter space required by regularity conditions A and B is set following the way mentioned in Chapter 4.3.1 for each model. Specific setting of the parameter space is reported in Table 4.1. In Table 4.2, we report the empirical means, corresponding standard errors of parameter estimates and true parameter values used for simulating 500 sample paths. Global optimization 'NOMAD' is used to search the approximate ECF estimate. The implementation follows Chapter 4.3 with normal distributed weight function $g(u)$.

4.4.1 Tuning parameters analysis

We discuss the result based on our tuning parameter $\mathcal{M}$ in the approximated ECF estimation (Equation (4.5). $\mathcal{M}$ is the key parameter in our ECF approximation. By selecting the tuning parameter $\hbar$ based on equation (4.17), we can show the approximation error bound decay exponentially with respect to $\mathcal{M}$. $\mathcal{M}$ is also the parameter to control the computational cost of our estimation procedure. Larger $\mathcal{M}$ indicates more terms involved in our ECF approximation which implies more computational cost. The exponential decay of error bound in our theoretical result indicates that we don't need a large enough $\mathcal{M}$ usually, so that our approximated ECF estimates are stable. Table 4.2 provides the empirical means, corresponding standard errors of parameter estimates and true parameter values based on same 500 simulated sample path, different sample size and different tuning parameter $\mathcal{M}$. For Merton's jump-diffusion model in Table 4.2a and Table 4.2b, we find $\mathcal{M} = 50$ is generally large enough to guarantee that empirical means and their standard errors converge to a number with a precision of 4 decimal places for weekly data. While, for daily data, $\mathcal{M}$ might be a little larger, which is 100. This can be explained by our Equation (4.6). It can be proved that denser data (daily data) have a smaller parameter δ in the characteristic functions (Check expressions of characteristic functions in Chapter 4.3.1). And a smaller δ will lead to a larger $\mathcal{K}$ and larger truncation error bound $T_{\mathcal{M}\hbar}$ (check Equation(4.6) and regularity condition A.4), making the bound of error $T_{\mathcal{M}\hbar}$ decay to zero more slowly. This implies that we need a larger $\mathcal{M}\hbar$ to make the truncation error bound unchanged if we have a larger $\mathcal{K}$. Then, $\mathcal{M}$ will be larger due to our tuning parameter $\hbar$ selection equation (4.17). Overall, daily data (denser data) usually need larger $\mathcal{M}$ (more computational cost) to guarantee the same ECF approximation error bound from Equation (4.6). If we don't need this high precision of 4 decimal places, $\mathcal{M} = 10$ can give us a reasonably good ECF estimate which is closed to the true parameter with stable standard error. That is, around 20 points in total can estimate the integration form in Equation (4.4) very well, which gives us effective estimates.

For Kou's jump-diffusion model in Table 4.2c and Table 4.2d, we have a little higher requirement for $\mathcal{M}$. That is, $\mathcal{M} = 10$ in a lot of cases cannot give us a reasonable result. For example, empirical mean of η_d is 6.3708, which is too far away from 2.5 compared with the case of $\mathcal{M} \geq 30$. Similar issues happen for daily simulated data with $n = 1000$ and $n = 5000$. This is mainly due to the small true parameter value of η_d. That is, the characteristic function of Kou's jump-diffusion model is analytic in the strip $[-\eta_d, \eta_u]$. Thus, based on the selection rule described in the Chapter 4.3.3, d_a is a value slightly smaller than $2\min(\eta_u, \eta_d)$. That is, if η_u and η_d are small, d_a is small. Then, we need a larger M to make sure we have the same tuning parameter $\hbar$ with the same exponential decay rate of approximation error bound (4.6) for the true parameter. Actually, this issue is not obvious for weekly data because we have shown that daily data usually

need larger $\mathcal{M}$ compared with weekly data to have the same bound of error. This is also an indication that daily data (denser data) require a relative large $\mathcal{M}$ to control the ECF approximation error bound (explained in the last paragraph for Merton's jump-diffusion model).

For NIG and CGMY model in Table [4.2e, 4.2f, 4.2g and 4.2h], we find the requirement for tuning parameter is low. $\mathcal{M} = 10$ is enough to give us a stable estimates. Again, this is due to the large d_a around the true parameter. Compared with Kou's jump-diffusion model of which $d_a \approx 5$, $d_a \approx 90$ for NIG model and $d_a \approx 158$ for CGMY model. Thus, with the same exponential error bound decay rate in Equation (4.6), the CGMY and NIG model need much smaller $\mathcal{M}$ compared with Kou's jump-diffusion model around the true parameter. Thus, $\mathcal{M} = 10$ is generally big enough to give us reasonable approximate ECF estiamte for NIG model and CGMY model in this parameter setting. More detailed explanation about tuning parameters $\mathcal{M}$, $\hbar$ and d_a can also be found in the Chapter 4.3.3.

We also plot the tuning parameter $\hbar$ value for each model in Figure 4.4, based on different choices of $\mathcal{M}$. This can give us a general idea of the $\hbar$ values corresponding to the different values of $\mathcal{M}$. Figure 4.4 shows the $\hbar$ values in each NOMAD optimization iteration, based on one simulated path. It is calculated via the Equation (4.17). Recall that $\mathcal{M}$ indicates the number of grid points in the evenly spaced grids set $\{-\mathcal{M}\hbar, -\hbar(\mathcal{M}-1), \ldots, \hbar(\mathcal{M}-1), \mathcal{M}\hbar\}$. Thus, $\hbar$ represents its dense degree in this grids set. We can find that if $\mathcal{M}$ is as small as 10, $\hbar$ could be around 10 for the daily simulated data of Merton's jump-diffusion model, Kou's jump-diffusion model, or even larger than 20 for NIG model and CGMY model. Actually, $\hbar = 10$ can give us reasonably good estimates from our analysis above for Merton's jump-diffusion model, NIG model and CGMY model. Thus, we can conclude that scattered grids (large $\hbar$) might be able to provide good approximation of the ECF target (Equation 4.4), which gives us stable approximated ECF estimates. For Kou's jump-diffusion model, the requirement of $\mathcal{M}$ and $\hbar$ is relative high, $\hbar$ is around 1 for daily data when $\mathcal{M} = 200$ and $\hbar$ is around 0.5 for weekly data when $\mathcal{M} = 200$. That is due to our parameter setting of small d_a.

Overall, we find that a relative small $\mathcal{M}$ and relative large $\hbar$ can approximate the ECF target (4.2) very well when the model satisfying our regularity conditions. That is, we can nearly use few scattered points with large spans to approximate ECF target without computational intensive integration. This will save us a lot of computation time and burden. We report the median of running time based on 500 simulated paths. All the simulation study was conducted by the same laptop with Intel Core i5-5300U CPU (2.30GHz). We find the running time of $\mathcal{M} = 10$ is 1/5 even 1/10 of the running time of $\mathcal{M} = 200$.

4.4.2 Asymptotic properties evidences

In this section, we want to check if our approximated ECF estimate matches the true parameter based on simulated data. This is implied by the asymptotic properties of approximated ECF estimate (Theorem 20). For jump-diffusion models [Table4.2a,4.2b,4.2c and 4.2d], empirical means are reasonably closed to true parameters. With sample size n increasing, empirical means tend to converge to true value and the standard error tends to converge. The convergence rate is roughly $\sqrt{n}$ because when the sample size gets 5 times bigger (sample size from 200 to 1000 or from 1000 to 5000), the standard error becomes roughly $1/\sqrt{5}$ smaller. These findings are consistent to the consistency and asymptotic normality of approximated ECF estimates (Theorem 19 and Theorem 20). For weekly data, small sample size $n = 200$ might be enough to identify volatility parameter σ and jump parameters λ, μ_j, σ_j (Merton's jump-diffusion model) or λ, p, η_u and η_d (Kou's jump-diffusion model) about three standard errors far away from zero. For daily data, this corresponding sample size might be as large as $n = 1000$. In addition, it seems that when the total sample size is fixed, dense data might lead to more difficulties to estimate parameters. This is consistent to our intuition. For finite activity Lévy processes, there are finite big jumps in time interval. Weekly data with fixed sample size have longer time spans, indicating more big jumps compared with daily data. Thus, more big jumps information leads to more accuracy of the estimate. However, this is not the case for infinite activity Lévy processes [Table 4.2e,4.2f,4.2g and 4.2h]. For example, for parameters in NIG model and CGMY model, denser data implies a less bias and standard error of the approximated ECF estimates. Intuitively, the reason is that for these models, jumps will happen on each data point no matter it is daily data or weekly data. The data with longer time spans doesn't contain more jump information because in all time intervals, jump will happen infinitely.

For infinite activity Lévy processes, the estimation performance is a little worse than finite activity Lévy processes above due to the model itself. The estimated parameters in NIG and CGMY model also imply the consistency and convergence rate information. Based on empirical mean with true initial value in Table [Table 4.2e,4.2f,4.2g and 4.2h], larger sample size n implies a smaller estimated bias and smaller standard error. For CGMY model, we find the parameter C identification problem. Table 4.2g and Table 4.2h report the estimation result of CGMY model. We find parameter C is hard to be identified with a big estimation standard error. For example, 1000 sample size might not be enough to identify C from 0 for weekly data. 5000 weekly data might be needed to match the estimates with true value with reasonable small standard error. However, 5000 weeks are rough 100 years.

In summary, this simulation results show that approximated ECF estimates are effective which can accurately match the true parameter values with reasonable small sample size. CGMY model might need

more data. More importantly, our approximated ECF estimates doesn't need a large tunning parameter $\mathcal{M}$, which is computationally efficient to implement.

Table 4.2: Empirical averages and their standard errors (in parentheses) of the approximated empirical characteristic function(ECF) estimates with sample size $n = (200, 1000, 5000)$ and different choices of $\mathcal{M} = (10, 30, 50, 100, 200)$.

(a) Merton's jump-diffusion model (Weekly data)

n	$\mathcal{M}$	$\mu = 0.1$	$\sigma = 0.3$	$\lambda = 10$	$\mu_j = -0.5$	$\sigma_j = 0.25$	time
$n = 200$	$\mathcal{M} = 10$	0.0855(0.1774)	0.2978(0.0197)	10.2269(2.1213)	−0.0043(0.6148)	0.2443(0.0660)	1.2707
$n = 200$	$\mathcal{M} = 30$	0.0862(0.1776)	0.2987(0.0188)	10.1107(1.9900)	−0.4924(0.1193)	0.2448(0.0763)	1.2413
$n = 200$	$\mathcal{M} = 50$	0.0862(0.1777)	0.2987(0.0187)	10.1035(1.9862)	−0.4985(0.0724)	0.2413(0.0598)	1.3396
$n = 200$	$\mathcal{M} = 100$	0.0862(0.1777)	0.2987(0.0187)	10.1034(1.9862)	−0.4985(0.0724)	0.2413(0.0598)	1.5193
$n = 200$	$\mathcal{M} = 200$	0.0862(0.1777)	0.2987(0.0187)	10.1034(1.9862)	−0.4985(0.0724)	0.2413(0.0598)	1.9045
$n = 1000$	$\mathcal{M} = 10$	0.0957(0.0782)	0.2998(0.0088)	9.9613(0.9061)	−0.1383(0.5771)	0.2496(0.0292)	1.3785
$n = 1000$	$\mathcal{M} = 30$	0.0956(0.0782)	0.2999(0.0086)	9.9503(0.8735)	−0.4983(0.0750)	0.2509(0.0436)	1.6762
$n = 1000$	$\mathcal{M} = 50$	0.0957(0.0781)	0.2999(0.0086)	9.9498(0.8737)	−0.5013(0.0335)	0.2494(0.0280)	2.0314
$n = 1000$	$\mathcal{M} = 100$	0.0957(0.0781)	0.2999(0.0086)	9.9498(0.8736)	−0.5013(0.0335)	0.2494(0.0280)	3.3563
$n = 1000$	$\mathcal{M} = 200$	0.0957(0.0781)	0.2999(0.0086)	9.9499(0.8737)	−0.5013(0.0335)	0.2494(0.0280)	5.7590
$n = 5000$	$\mathcal{M} = 10$	0.1012(0.0343)	0.3000(0.0042)	10.0044(0.4107)	−0.2053(0.5370)	0.2497(0.0132)	2.0658
$n = 5000$	$\mathcal{M} = 30$	0.1015(0.0341)	0.3001(0.0041)	10.0033(0.3938)	−0.4996(0.0140)	0.2498(0.0129)	4.5251
$n = 5000$	$\mathcal{M} = 50$	0.1015(0.0341)	0.3001(0.0041)	10.0036(0.3936)	−0.4996(0.0140)	0.2498(0.0129)	6.8596
$n = 5000$	$\mathcal{M} = 100$	0.1015(0.0341)	0.3001(0.0041)	10.0036(0.3936)	−0.4996(0.0140)	0.2498(0.0129)	12.8914
$n = 5000$	$\mathcal{M} = 200$	0.1015(0.0341)	0.3001(0.0041)	10.0037(0.3936)	−0.4996(0.0140)	0.2498(0.0129)	25.2845

(b) Merton's jump-diffusion model (daily data)

n	$\mathcal{M}$	$\mu = 0.1$	$\sigma = 0.3$	$\lambda = 10$	$\mu_j = -0.5$	$\sigma_j = 0.25$	time
$n = 200$	$\mathcal{M} = 10$	0.1053(0.3656)	0.2981(0.0168)	10.3672(4.5100)	0.1964(0.5946)	0.2057(0.2118)	2.2306
$n = 200$	$\mathcal{M} = 30$	0.1061(0.3655)	0.2989(0.0163)	9.8865(3.6263)	0.0757(0.7118)	0.2062(0.1011)	2.3533
$n = 200$	$\mathcal{M} = 50$	0.1066(0.3653)	0.2987(0.0164)	10.0009(3.6660)	−0.2945(0.5091)	0.2795(0.2582)	2.3917
$n = 200$	$\mathcal{M} = 100$	0.1065(0.3656)	0.2989(0.0165)	9.9237(3.6510)	−0.4881(0.1419)	0.1921(0.0965)	2.9422
$n = 200$	$\mathcal{M} = 200$	0.1064(0.3655)	0.2988(0.0165)	9.9629(3.6589)	−0.4867(0.1457)	0.1915(0.0962)	3.6945
$n = 1000$	$\mathcal{M} = 10$	0.1035(0.1648)	0.2999(0.0076)	10.0810(2.0117)	0.1716(0.5945)	0.2739(0.2063)	2.5013
$n = 1000$	$\mathcal{M} = 30$	0.1038(0.1645)	0.3001(0.0074)	9.9627(1.7658)	−0.1368(0.6344)	0.2445(0.0724)	3.0720
$n = 1000$	$\mathcal{M} = 50$	0.1039(0.1644)	0.3000(0.0074)	9.9879(1.7594)	−0.4098(0.3604)	0.2791(0.1788)	3.7781
$n = 1000$	$\mathcal{M} = 100$	0.1040(0.1645)	0.3001(0.0074)	9.9652(1.7588)	−0.4995(0.0566)	0.2362(0.0460)	6.0785
$n = 1000$	$\mathcal{M} = 200$	0.1040(0.1645)	0.3001(0.0074)	9.9652(1.7590)	−0.4995(0.0566)	0.2363(0.0460)	10.1221
$n = 5000$	$\mathcal{M} = 10$	0.1002(0.0705)	0.3001(0.0034)	10.0000(0.8495)	0.3414(0.6189)	0.2874(0.1782)	4.0002
$n = 5000$	$\mathcal{M} = 30$	0.0998(0.0702)	0.3003(0.0033)	9.9121(0.7500)	−0.0823(0.6623)	0.2495(0.0410)	7.5911
$n = 5000$	$\mathcal{M} = 50$	0.0998(0.0702)	0.3002(0.0033)	9.9318(0.7455)	−0.4726(0.2009)	0.2588(0.0980)	11.5485
$n = 5000$	$\mathcal{M} = 100$	0.0999(0.0702)	0.3003(0.0033)	9.9265(0.7452)	−0.4995(0.0250)	0.2463(0.0219)	21.1049
$n = 5000$	$\mathcal{M} = 200$	0.0999(0.0702)	0.3003(0.0033)	9.9268(0.7451)	−0.4995(0.0250)	0.2463(0.0219)	41.2290

Table 4.2 (cont.)

(c) Kou's jump-diffusion model (weekly data)

n	$\mathcal{M}$	$\mu = 0.2$	$\sigma = 0.2$	$\lambda = 30$	$p = 0.4$	$\eta_u = 5$	$\eta_d = 2.5$	time
$n = 200$	$\mathcal{M} = 10$	0.1786(0.2076)	0.1931(0.0353)	30.9883(5.3240)	0.4194(0.0780)	5.7473(3.0542)	2.7305(1.1330)	5.5777
$n = 200$	$\mathcal{M} = 30$	0.1839(0.1778)	0.1960(0.0247)	30.4946(4.6563)	0.4072(0.0695)	5.4008(2.0607)	2.5580(0.6144)	4.9446
$n = 200$	$\mathcal{M} = 50$	0.1834(0.1771)	0.1965(0.0240)	30.4218(4.5907)	0.4068(0.0690)	5.3832(2.0102)	2.5336(0.5551)	4.9779
$n = 200$	$\mathcal{M} = 100$	0.1836(0.1771)	0.1965(0.0239)	30.4223(4.5856)	0.4068(0.0690)	5.3806(2.0052)	2.5361(0.5528)	5.6622
$n = 200$	$\mathcal{M} = 200$	0.1836(0.1771)	0.1965(0.0239)	30.4231(4.5856)	0.4068(0.0690)	5.3807(2.0057)	2.5364(0.5526)	7.0234
$n = 1000$	$\mathcal{M} = 10$	0.1948(0.0851)	0.1978(0.0146)	30.2325(2.1517)	0.4035(0.0377)	5.1789(1.1742)	2.5397(0.4867)	5.7777
$n = 1000$	$\mathcal{M} = 30$	0.1981(0.0766)	0.1990(0.0108)	30.0752(1.8969)	0.3999(0.0316)	5.0673(0.7259)	2.5022(0.2375)	6.2845
$n = 1000$	$\mathcal{M} = 50$	0.1983(0.0762)	0.1990(0.0107)	30.0678(1.8948)	0.3999(0.0315)	5.0602(0.7166)	2.5027(0.2295)	7.3662
$n = 1000$	$\mathcal{M} = 100$	0.1983(0.0762)	0.1990(0.0107)	30.0677(1.8934)	0.3999(0.0315)	5.0599(0.7163)	2.5031(0.2290)	11.7326
$n = 1000$	$\mathcal{M} = 200$	0.1984(0.0762)	0.1990(0.0107)	30.0675(1.8932)	0.3999(0.0315)	5.0596(0.7161)	2.5031(0.2290)	19.2467
$n = 5000$	$\mathcal{M} = 10$	0.1958(0.0379)	0.1995(0.0066)	30.0455(1.0070)	0.4017(0.0166)	4.9983(0.5568)	2.5194(0.2086)	8.1813
$n = 5000$	$\mathcal{M} = 30$	0.1963(0.0343)	0.1998(0.0050)	30.0032(0.9159)	0.3999(0.0137)	4.9953(0.3267)	2.5011(0.1046)	15.6615
$n = 5000$	$\mathcal{M} = 50$	0.1963(0.0342)	0.1999(0.0050)	30.0008(0.9116)	0.3999(0.0137)	4.9926(0.3220)	2.5009(0.1014)	22.4953
$n = 5000$	$\mathcal{M} = 100$	0.1963(0.0342)	0.1999(0.0050)	30.0009(0.9115)	0.3999(0.0137)	4.9921(0.3219)	2.5011(0.1012)	40.4674
$n = 5000$	$\mathcal{M} = 200$	0.1963(0.0342)	0.1999(0.0050)	30.0004(0.9112)	0.3999(0.0137)	4.9919(0.3216)	2.5011(0.1014)	78.1217

(d) Kou's jump-diffusion model (daily data)

n	$\mathcal{M}$	$\mu = 0.2$	$\sigma = 0.2$	$\lambda = 30$	$p = 0.4$	$\eta_u = 5$	$\eta_d = 2.5$	time
$n = 200$	$\mathcal{M} = 10$	0.2174(0.2698)	0.1966(0.0159)	32.0937(8.0065)	0.5054(0.2102)	6.8176(6.2990)	6.3708(6.0965)	5.4246
$n = 200$	$\mathcal{M} = 30$	0.2080(0.2586)	0.1986(0.0125)	31.1331(7.4277)	0.4295(0.1412)	7.0238(5.8910)	3.7597(3.8401)	5.8465
$n = 200$	$\mathcal{M} = 50$	0.2079(0.2595)	0.1988(0.0118)	31.0456(7.3824)	0.4154(0.1307)	6.8661(5.6867)	3.5158(3.8262)	6.5503
$n = 200$	$\mathcal{M} = 100$	0.2090(0.2588)	0.1987(0.0118)	31.1609(7.3101)	0.4164(0.1234)	6.8023(5.6340)	3.5981(3.8678)	7.2795
$n = 200$	$\mathcal{M} = 200$	0.2096(0.2589)	0.1986(0.0118)	31.2379(7.2819)	0.4193(0.1236)	6.8508(5.4292)	3.6818(3.6920)	8.5291
$n = 1000$	$\mathcal{M} = 10$	0.2090(0.1244)	0.1996(0.0076)	30.7498(3.8770)	0.4810(0.1665)	5.5683(4.3218)	4.5213(3.7909)	7.5394
$n = 1000$	$\mathcal{M} = 30$	0.2020(0.1197)	0.2001(0.0058)	30.1173(3.3195)	0.4043(0.0704)	5.4637(2.4314)	2.6179(0.9473)	9.1346
$n = 1000$	$\mathcal{M} = 50$	0.2021(0.1206)	0.2001(0.0056)	30.1009(3.2952)	0.3991(0.0567)	5.3977(2.0309)	2.5391(0.6569)	10.8261
$n = 1000$	$\mathcal{M} = 100$	0.2024(0.1206)	0.2001(0.0056)	30.0933(3.2994)	0.3993(0.0544)	5.3290(1.8427)	2.5472(0.5657)	14.5648
$n = 1000$	$\mathcal{M} = 200$	0.2027(0.1209)	0.2001(0.0056)	30.0961(3.2975)	0.4004(0.0544)	5.2813(1.7398)	2.5733(0.5510)	21.9689
$n = 5000$	$\mathcal{M} = 10$	0.2036(0.0487)	0.2004(0.0032)	29.9217(1.6269)	0.4678(0.1512)	4.8981(2.8994)	3.3779(1.9191)	13.9644
$n = 5000$	$\mathcal{M} = 30$	0.1999(0.0476)	0.2001(0.0024)	29.9313(1.3701)	0.3983(0.0311)	5.1329(0.8225)	2.4587(0.3844)	22.6049
$n = 5000$	$\mathcal{M} = 50$	0.2006(0.0478)	0.2001(0.0024)	29.9470(1.3632)	0.3996(0.0247)	5.0697(0.6424)	2.4953(0.2702)	30.7117
$n = 5000$	$\mathcal{M} = 100$	0.2006(0.0479)	0.2001(0.0024)	29.9473(1.3700)	0.4000(0.0244)	5.0623(0.7379)	2.5036(0.2308)	47.4964
$n = 5000$	$\mathcal{M} = 200$	0.2007(0.0478)	0.2001(0.0024)	29.9443(1.3620)	0.4193(0.0239)	5.0413(0.6093)	2.5081(0.2253)	90.0734

Table 4.2 (cont.)

(e) Normal inverse Gamma model (NIG)(weekly data)

n	$\mathcal{M}$	$\mu = 0$	$\alpha = 50$	$\beta = -5$	$\lambda = 5$	time
$n = 200$	$\mathcal{M} = 10$	$-0.1766(0.5791)$	$54.2762(24.5082)$	$-3.2620(5.8970)$	$5.3479(2.2210)$	3.4409
$n = 200$	$\mathcal{M} = 30$	$-0.1771(0.5808)$	$53.5725(23.5995)$	$-3.2511(5.9281)$	$5.2978(2.1778)$	3.5649
$n = 200$	$\mathcal{M} = 50$	$-0.1765(0.5804)$	$53.6826(23.6676)$	$-3.2583(5.9247)$	$5.3095(2.1892)$	3.8705
$n = 200$	$\mathcal{M} = 100$	$-0.1759(0.5804)$	$53.6546(23.7462)$	$-3.2657(5.9275)$	$5.3043(2.1847)$	4.3231
$n = 200$	$\mathcal{M} = 200$	$-0.1774(0.5803)$	$53.6072(23.6188)$	$-3.2496(5.9259)$	$5.3015(2.1805)$	5.5611
$n = 1000$	$\mathcal{M} = 10$	$-0.0273(0.3862)$	$52.3664(13.5416)$	$-4.7776(3.9328)$	$5.2237(1.3066)$	3.9003
$n = 1000$	$\mathcal{M} = 30$	$-0.0292(0.3850)$	$52.2119(13.4305)$	$-4.7549(3.9196)$	$5.2108(1.2944)$	4.7854
$n = 1000$	$\mathcal{M} = 50$	$-0.0290(0.3839)$	$52.2234(13.4435)$	$-4.7575(3.9123)$	$5.2117(1.2942)$	5.6850
$n = 1000$	$\mathcal{M} = 100$	$-0.0295(0.3845)$	$52.2338(13.4900)$	$-4.7515(3.9148)$	$5.2129(1.2988)$	8.3360
$n = 1000$	$\mathcal{M} = 200$	$-0.0300(0.3837)$	$52.1743(13.3832)$	$-4.7466(3.9079)$	$5.2075(1.2906)$	13.0213
$n = 5000$	$\mathcal{M} = 10$	$0.0056(0.1980)$	$50.8440(6.0785)$	$-5.0980(2.0446)$	$5.0784(0.5536)$	4.8432
$n = 5000$	$\mathcal{M} = 30$	$0.0033(0.1999)$	$50.7352(5.9069)$	$-5.0695(2.0654)$	$5.0698(0.5414)$	8.7562
$n = 5000$	$\mathcal{M} = 50$	$0.0045(0.2002)$	$50.7944(5.9501)$	$-5.0695(2.0686)$	$5.0751(0.5454)$	12.0437
$n = 5000$	$\mathcal{M} = 100$	$0.0039(0.1996)$	$50.7684(5.9226)$	$-5.0764(2.0636)$	$5.0728(0.5431)$	20.7758
$n = 5000$	$\mathcal{M} = 200$	$0.0041(0.1995)$	$50.7725(5.9230)$	$-5.0778(2.0616)$	$5.0732(0.5433)$	38.7957

(f) Normal inverse Gamma model (NIG)(daily data)

n	$\mathcal{M}$	$\mu = 0$	$\alpha = 50$	$\beta = -5$	$\lambda = 5$	time
$n = 200$	$\mathcal{M} = 10$	$-0.1023(0.5289)$	$52.9748(19.5433)$	$-3.8273(5.7551)$	$5.1952(1.4367)$	2.0956
$n = 200$	$\mathcal{M} = 30$	$-0.0967(0.5253)$	$53.3950(18.7111)$	$-3.9115(5.7534)$	$5.2323(1.3583)$	2.3162
$n = 200$	$\mathcal{M} = 50$	$-0.0966(0.5253)$	$53.3845(18.6927)$	$-3.9123(5.7534)$	$5.2316(1.3571)$	2.3873
$n = 200$	$\mathcal{M} = 100$	$-0.0966(0.5253)$	$53.3914(18.7059)$	$-3.9125(5.7534)$	$5.2321(1.3580)$	2.7628
$n = 200$	$\mathcal{M} = 200$	$-0.0966(0.5253)$	$53.3961(18.7360)$	$-3.9127(5.7533)$	$5.2323(1.3581)$	3.5458
$n = 1000$	$\mathcal{M} = 10$	$-0.0002(0.2849)$	$50.8573(7.3647)$	$-5.0857(3.3062)$	$5.0646(0.5456)$	2.0279
$n = 1000$	$\mathcal{M} = 30$	$0.0003(0.2835)$	$50.8586(7.3807)$	$-5.0979(3.2650)$	$5.0652(0.5465)$	2.6538
$n = 1000$	$\mathcal{M} = 50$	$0.0003(0.2835)$	$50.8583(7.3807)$	$-5.0981(3.2650)$	$5.0652(0.5465)$	3.3830
$n = 1000$	$\mathcal{M} = 100$	$0.0003(0.2835)$	$50.8580(7.3810)$	$-5.0981(3.2641)$	$5.0652(0.5466)$	5.0605
$n = 1000$	$\mathcal{M} = 200$	$0.0003(0.2835)$	$50.8575(7.3809)$	$-5.0985(3.2642)$	$5.0651(0.5466)$	8.1698
$n = 5000$	$\mathcal{M} = 10$	$-0.0005(0.1365)$	$50.1891(3.2879)$	$-5.0647(1.6587)$	$5.0186(0.2285)$	2.9783
$n = 5000$	$\mathcal{M} = 30$	$0.0002(0.1354)$	$50.1865(3.2925)$	$-5.0782(1.6279)$	$5.0185(0.2287)$	5.5391
$n = 5000$	$\mathcal{M} = 50$	$0.0002(0.1354)$	$50.1880(3.2925)$	$-5.0782(1.6282)$	$5.0186(0.2288)$	7.9247
$n = 5000$	$\mathcal{M} = 100$	$0.0002(0.1354)$	$50.1871(3.2930)$	$-5.0764(1.6301)$	$5.0188(0.2287)$	14.4405
$n = 5000$	$\mathcal{M} = 200$	$0.0002(0.1354)$	$50.1876(3.2926)$	$-5.0785(1.6281)$	$5.0185(0.2287)$	27.2182

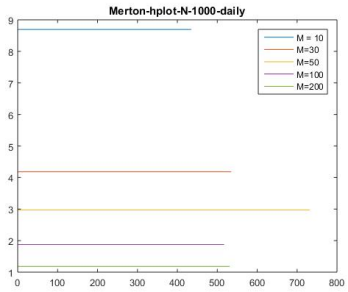
Table 4.2 (cont.)

(g) CGMY (weekly data)

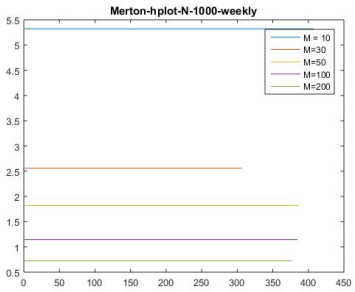
n	$\mathcal{M}$	$C = 3$	$G = 79$	$M = 83$	$Y = 0.9$	time
$n = 200$	$\mathcal{M} = 10$	4.9237(5.5964)	79.8082(10.1230)	84.3237(9.5742)	0.8597(0.1761)	1.6727
$n = 200$	$\mathcal{M} = 30$	5.1597(6.2451)	79.5460(10.1620)	84.0606(9.7654)	0.8541(0.1842)	1.7228
$n = 200$	$\mathcal{M} = 50$	5.1942(6.1846)	79.5789(10.1352)	84.0939(9.6720)	0.8530(0.1847)	1.9034
$n = 200$	$\mathcal{M} = 100$	5.0816(6.2178)	79.7728(9.9587)	84.2880(9.4997)	0.8571(0.1796)	2.3687
$n = 200$	$\mathcal{M} = 200$	5.1459(6.0211)	79.5272(10.1491)	84.0432(9.6020)	0.8547(0.1837)	3.1707
$n = 1000$	$\mathcal{M} = 10$	3.7650(2.0202)	79.8530(4.5741)	83.8771(4.6435)	0.8717(0.0810)	1.4976
$n = 1000$	$\mathcal{M} = 30$	3.7557(2.0022)	79.9348(4.4372)	83.9586(4.5874)	0.8722(0.0820)	2.0228
$n = 1000$	$\mathcal{M} = 50$	3.7480(1.9923)	79.9167(4.4308)	83.9406(4.5271)	0.8724(0.0813)	2.5595
$n = 1000$	$\mathcal{M} = 100$	3.7668(2.0794)	79.8664(4.3671)	83.8905(4.5587)	0.8720(0.0826)	4.5197
$n = 1000$	$\mathcal{M} = 200$	3.7592(2.0995)	79.9457(4.5301)	83.9696(4.6172)	0.8727(0.0824)	8.0488
$n = 5000$	$\mathcal{M} = 10$	3.3548(0.8212)	79.0259(1.9861)	83.0857(2.1687)	0.8837(0.0383)	2.1797
$n = 5000$	$\mathcal{M} = 30$	3.3481(0.7239)	79.0482(1.8807)	83.1079(2.0624)	0.8837(0.0365)	5.0230
$n = 5000$	$\mathcal{M} = 50$	3.3414(0.6963)	79.0641(1.8569)	83.1238(2.0343)	0.8840(0.0353)	7.7427
$n = 5000$	$\mathcal{M} = 100$	3.3460(0.7089)	79.0709(1.9131)	83.1307(2.0942)	0.8838(0.0356)	14.7654
$n = 5000$	$\mathcal{M} = 200$	3.3474(0.7228)	79.0499(1.8839)	83.1096(2.0661)	0.8838(0.0364)	28.6536

(h) CGMY (daily data)

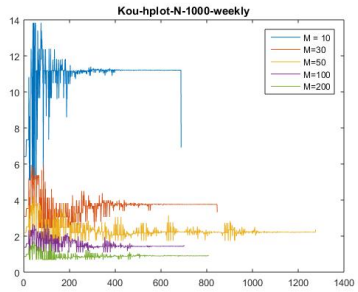
n	$\mathcal{M}$	$C = 3$	$G = 79$	$M = 83$	$Y = 0.9$	time
$n = 200$	$\mathcal{M} = 10$	6.7472(11.0038)	82.3170(10.6091)	87.4827(9.8201)	0.8541(0.2002)	3.0932
$n = 200$	$\mathcal{M} = 30$	6.8214(10.7326)	82.2827(10.9506)	87.4494(9.9995)	0.8518(0.2010)	3.3347
$n = 200$	$\mathcal{M} = 50$	6.7942(10.8521)	82.3795(11.1892)	87.5430(10.1011)	0.8536(0.2002)	3.7647
$n = 200$	$\mathcal{M} = 100$	6.6717(10.7255)	82.0844(10.9620)	87.2545(10.2268)	0.8560(0.2010)	4.6166
$n = 200$	$\mathcal{M} = 200$	6.8321(11.6814)	82.1539(10.8451)	87.3220(9.6834)	0.8540(0.2006)	6.3195
$n = 1000$	$\mathcal{M} = 10$	4.0526(2.5416)	81.8516(5.5046)	85.8938(5.7504)	0.8687(0.0920)	2.1551
$n = 1000$	$\mathcal{M} = 30$	4.1071(2.7390)	81.9503(5.4208)	85.9927(5.7259)	0.8673(0.0937)	3.0954
$n = 1000$	$\mathcal{M} = 50$	4.0733(2.8436)	81.8911(5.6086)	85.9365(5.7532)	0.8678(0.0909)	3.9532
$n = 1000$	$\mathcal{M} = 100$	4.0862(2.5802)	81.9416(5.5369)	85.9842(5.7541)	0.8675(0.0929)	7.2494
$n = 1000$	$\mathcal{M} = 200$	4.0202(2.3862)	81.8797(5.5703)	85.9194(5.8295)	0.8691(0.0921)	12.8699
$n = 5000$	$\mathcal{M} = 10$	3.3465(0.8003)	79.7171(2.2706)	83.8696(2.5123)	0.8858(0.0382)	2.7697
$n = 5000$	$\mathcal{M} = 30$	3.3501(0.6946)	79.7185(2.3513)	83.8703(2.4740)	0.8853(0.0387)	6.7488
$n = 5000$	$\mathcal{M} = 50$	3.3289(0.6620	79.7172(2.3483)	83.8690(2.5180)	0.8864(0.0378)	10.5584
$n = 5000$	$\mathcal{M} = 100$	3.3403(0.6977)	79.7283(2.3870)	83.8783(2.4906)	0.8859(0.0387)	19.5148
$n = 5000$	$\mathcal{M} = 200$	3.3274(0.6761)	79.7588(2.3416)	83.9093(2.5314)	0.8866(0.0379)	37.5094



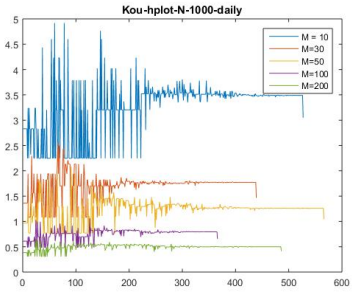
(a) Merton-h-N-1000-daily



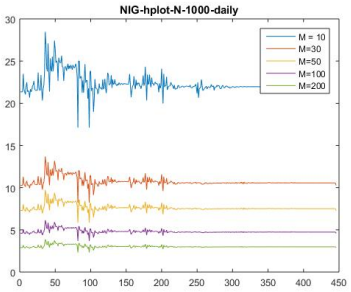
(b) Merton-h-N-1000-weekly



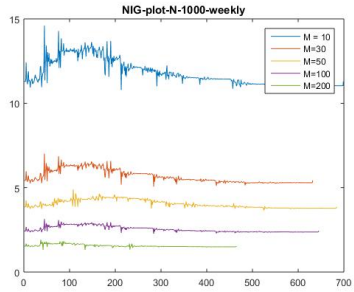
(c) Kou-h-N-1000-daily



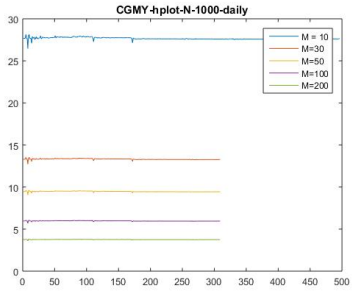
(d) Kou-h-N-1000-weekly



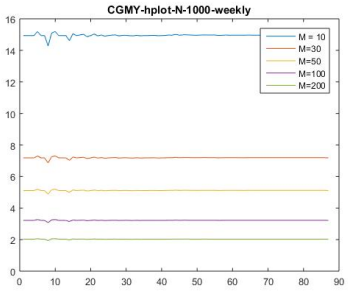
(e) NIG-h-N-1000-daily



(f) NIG-h-N-1000-weekly



(g) CGMY-h-N-1000-daily



(h) CGMY-h-N-1000-weekly

Figure 4.4: For one simulated path with sample size 1000, the tuning parameter $\hbar$'s value in NOMAD optimization procedure. $\mathcal{M} = \{10, 30, 50, 100, 200\}$. $\hbar$ is calculated based on Equation (4.17), which is suggested in Chapter 4.3.3.

4.5 Concluding remarks

In this chapter, we construct approximated ECF estimates based on the trapezoidal approximation. On the theoretical side, we show the exponential decay of the error bound of our ECF estimation target function's

approximation. This fast decay rate reduces the computational burden of approximation procedure. In addition, we also propose the asymptotic property of our approximated ECF estimates and prove the selected popular Lévy processes satisfy the regularity conditions. On the application side, our estimation procedure is applied to popular Lévy processes. Simulation study indicates that our approximated ECF estimates are efficient and effective.

Chapter 5

A dynamic model for evolving truth discovery

This Chapter includes my joint project, collaborated with Shi Zhi[1], Zheyi Zhu[2], Qi Li[3], Zhaoran Wang[4] and Jiawei Han[5]. I am the equally contributed first author with Shi Zhi.

5.1 Introduction

Nowadays, people can access vast amount of information from all kinds of sources every day. Suppose a traveler would like to check the current departure time of his flight, he may go to some frequently used websites for it. Also, when the business owners investigate the daily pedestrian counts to make an investment, they may get this information from sensors of traffic lights, the number of cell phones connected to a Wi-Fi hotspot, or from some hired workers. Multiple information sources provide us chances to validate the truth, but at the same time bring possible misinformation due to lack of expertise, malicious purposes or broadcasting failures. Take the stock market as an example. Market capitalization is one of the key information that investors are interested in. Based on the statistics on the collected market capitalization data from 55 sources during 2011[6], we find that within 1000 stocks, the sources provide market capitalization information with 95.6% on average, and 19.7 days on average out of 21 trading days . Meanwhile, the only source that provides information for all stocks during July 2011, 'pc-quote', ranks at the bottom in terms of the precision. This drives us to develop an efficient algorithm to discover the reliable information with complete coverage along time. Thus, if we collect the information about the same object from different information sources, it is common that data have both consensus and conflict at the same time. This brings in the *truth discovery* problem, which aims to find the most trustworthy information collected from multiple sources.

One naive approach to resolve the conflicts of the data is median/mean. However, this is not always the

[1]shizhi2@illinois.edu
[2]zzhu27@illinois.edu
[3]qili5@illinois.edu
[4]zhaoran@princeton.edu
[5]hanj@illinois.edu
[6]http://lunadong.com/fusionDataSets.htm

case. Thus, the general principle is introduced here: if the piece of information is from a reliable source, then it is more trustworthy, and the source that provides trustworthy information is more reliable. This intuition is emerging as a promising methodology in resolving conflicts in various scenarios Yin et al. (2008) and various domains such as natural language processing tasks Yu et al. (2014); Liu et al. (2017) and online health community Mukherjee et al. (2014). These applications reveal the necessity to develop truth discovery methods.

Most of the existing truth discovery algorithms are proposed to work on static data. However, batch algorithms would not properly solve the dynamic truth discovery problem mainly due to three reasons. First, since the data arrives sequentially, it is costly to re-run the batch algorithms all over from the first timestamp on large-scale data.

Second, we observe that in real-world scenarios, the truth of the same entity usually will fluctuate as time changes and would not stay as a constant value, and the truths of consecutive timestamps are correlated in most cases. For example, we examine the auto-correlation of market capitalization of 100 stock symbols, and with a significant portion of time the auto-correlation is larger than 0.2. This auto-correlation provides us another evidence that using estimated true value from the history can benefit the estimation of current true value. Moreover, a common case is that the sources are unlikely to provide an observation to every objects at every timestamp, or even no sources provide an observation at some date, we name it as 'missing data problem'. If we can correctly estimate the correlation along time, we may alleviate the missing data problem by using the latent truths from last timestamps as a smoother for the current estimation and making the best guess on it. Li et al. Li et al. (2015) also re-weigh the current estimated true value with the estimated truth from last timestamp using a smoothing factor as a fixed parameter. However, in the real world data, the balance of value from last timestamp and the integration from current observations always fluctuate along time, and differ from entity to entity, and both parameters in Li et al. (2015) are given by users and fixed for all entities along time. This raises a huge challenge to dynamically estimate the smoothing strength and the current estimation.

Thirdly, inspired by Li et al. (2015), we also observe that the source quality would also evolves over time and the source quality consistency assumption of existing methods does not hold any more. However, Li et al. (2015) assumes the source quality stays the same along time, and compensates the sudden change by a decaying factor assigned on source quality, which is pre-defined by users. Our proposed model can naturally incorporate this decaying effect by dynamically estimating it. Also, as illustrated in Dong et al. (2009*b*), sources may copy from each other, or get information from similar sources. Similar situations appear when multiple stock information websites are actually operated by the same head company. This

may be harmful to validate the truthfulness when happening among bad sources. Thus, understanding the source dependency can help us to better estimate the truth.

In this work, a new truth discovery method for evolving numerical data based on hidden Markov model is developed for dynamic scenarios. We take into account evolving truth, source quality, source correlation, objects correlation in our model. The case study shows its effectiveness compared with previous methods.

In summary, our contributions in this work are as follows.

1. Model the evolving truth discovery in probability and statistics methods with theoretical guarantees.
2. Balance truth smoothing and current estimation dynamically.
3. The model is robust to missing observations.
4. Estimate source dependency in a unified hidden Markov model.
5. Develop both batch-mode algorithm and efficient O(T) online version.
6. Case study shows its effectiveness over several real datasets.

5.2 Related works

Truth discovery problem has been studied to resolve the conflict among sources. The essential idea is by incorporating the source quality, information from high-quality sources are more trustworthy, and should weigh more in truth estimation. It is first formally introduced by Yin et. al. Yin et al. (2008), which models source quality as a single score and iteratively updates source quality and truth value in an unsupervised way. The idea is shared in some early works Pasternack and Roth (2010); Vydiswaran et al. (2011); Liu et al. (2011). These algorithms focus on categorical truth.

Then more papers propose new truth discover algorithms in various scenarios. CRH Li, Li, Gao, Zhao, Fan and Han (2014) is an integrated framework for both numerical and categorical truths, by defining different loss function and combines them into an optimization object function, and Li, Li, Gao, Su, Zhao, Demirbas, Fan and Han (2014) proposes a new framework for long-tail phenomenon. Probabilistic graphical model is also widely adopted in truth discovery domain, where the latent truth and source quality can be modeled as latent variables. Expectation-Maximization (EM) is naturally used to infer the truth and source quality Zhao et al. (2012); Wang et al. (2012); Qi et al. (2013); Ma et al. (2015); Zhi et al. (2015).

There have been some recent works that solve the sub-problems discussed in our model. As for source dependency analysis, Dong et al. (2009a) is the first to consider copy-cats among sources and integrate it

into the inference of truth values. Zhao and Han (2012) also directly models the real-value truth by putting normal distribution assumption on the observation given the latent truth. Since both of them are focused on static data, it does not make use of the correlation of truths between timestamps.

In works that study the temporal change of truth, Pal et. al. Pal et al. (2012) models the history of the objects using hidden semi-Markovian process. However, it assumes source independency, which is not true in the real cases. We will show the dependency analysis in Chapter 5.4.2. Li et. al. Li et al. (2015) provides an incremental framework that updates truths and source weights as new data come, but the temporal correlation is captured by manually prefixed parameters. We can achieve the same function as those prefixed parameters when estimating the parameters in our model. A recent work Garcia-Ulloa et al. (2017) proposes a truth discovery model based on recursive Bayesian estimation in spatio-temporal tasks, but specifically for categorical truth value. Our method models the temporal correlation of numerical truth and incorporates source dependency, and we further propose an efficient and effective online estimation algorithm.

5.3 The model

In this section, we first formulate the problem of evolving truth discovery problem using the hidden Markov model, where the truths are the hidden vairables. Then, we provide Kalman filter and smoother with the efficient blocked parameter updating under expectation maximization (EM) schema. We provide an effective data pre-processing method and an online algorithm with pre-train step for practical use.

5.3.1 Problem formulation

Notation

Input. Let $\mathcal{O} = \{o_1, o_2, \ldots, o_O\}$ be the objects that we are interested in. Let $\mathcal{S} = \{s_1, s_2, \ldots, s_S\}$ be the set of sources. Numerical observations of O objects can be collected from S sources at each timestamp $t \in \{1, 2, \ldots, T\}$ ($t \in 1:T$ [7]). Let $v^i_{j,t}$ represent the observation provided by the source s_i of the object o_j at the t-th timestamp. For convenience, we denote all the observations from source s_i at time t as $\mathcal{X}^i_t$, that is, $\mathcal{X}^i_t = \{v^i_{j,t}\}_{o_j \in \mathcal{O}}$. Further, the size of this set is denoted as $c^i_t = |\mathcal{X}^i_t|$.

Output. Let $\mu_{j,t}$ be the truth for object o_j at time t, and the output is the whole set of truths at time t, denoted by $\mathcal{T}_t = \{\mu_{1,t},\ \mu_{2,t}, ...,\ \mu_{O,t}\}$.

Besides inferring truths, truth discovery methods can also estimate source reliability degrees. Let Σ

[7] $a:b$, where a, b are arbitrary integers and $b \geq a$, represents the set $\{a, a+1, \ldots, b\}$ in this work.

denote the source covariance matrix. Its diagonal element σ_i^2 can be interpreted as the source quality of source s_i. Its off-diagonal elements $\sigma_{i,i'}$ can be used to measure the source dependency between source s_i and $s_{i'}$.

Task definition

We formally define the task in this paper as follows.

Inferring truth. Until timestamp T, we collect observations $\{v_{j,t}^{1:S}\}$ of O objects from S sources. Our goal is to infer the true values $\{\mu_{j,1:T}\}$ for each object o_j by aggregating observations $\{v_{1:O,1:T}^{1:S}\}$ of O objects from S sources.

Inferring source quality. Besides inferring truths, we also would like to infer source covariance matrix during timestamp $t \in 1:T$ given observations of all objects from all sources.

Inferring other parameters. We can also infer sources dependency $\sigma_{i,i'}$ in our work. It is useful to capture the effects of copying among sources.

5.3.2 Batch solution: hidden Markov model

We first build a hidden Markov model for evolving truth discovery when we can observe all data until time T. Figure 5.1 shows the dynamics of truths and observations. We use Markov processes to model the dynamics of truths with the underlying temporal correlations. We assume first-order Markov property on latent truths, where the current latent truth depends on the latent truth from the last timestamp. The current observations of objects only depends on the latent truth of current timestamp, and observations are conditionally independent along the timeline. Assume that we observe the same set of objects along the timeline. $\boldsymbol{\mu}_t = (\mu_{1,t}, \ldots, \mu_{O,t})^T$ denotes the vector of latent truths, where the superscript T represents transpose of a vector or a matrix. Then, the dynamics of truths can be written in the following form

$$\boldsymbol{\mu}_{t+1} = A\boldsymbol{\mu}_t + \boldsymbol{\omega}_t \tag{5.1}$$

, where $t \in \{1, 2, \ldots, T\}$, $\boldsymbol{\mu}_t$ is the latent vector we aim to estimate, and $A \in \mathbb{R}^{O \times O}$ is the transition matrix of latent truth for all timestamps. Eq. 5.1 means that the current truth is the linear combination of truths from last timestamp, plus an error term. If A is a diagonal matrix, truth of each object will only depend on the previous true value of the same object, whereas if A is non-diagonal, truth of an object will also depend on the true values of other objects at last timestamp.

Let the initial distribution of $\boldsymbol{\mu}_1$ follow a multivariate normal distribution

$$\boldsymbol{\mu}_1 \sim Normal(\boldsymbol{\pi}_1, V_1) \tag{5.2}$$

, where $\boldsymbol{\pi}_1 \in \mathbb{R}^{O\times 1}$ and $V_1 \in \mathbb{R}^{O\times O}$ are the mean vector and covariance matrix of the initial state, respectively.

Then, let the error vector $\boldsymbol{\omega}_t \in \mathbb{R}^O$ follow multivariate normal distribution

$$\boldsymbol{\omega}_t \sim Normal(\mathbf{0}, \Gamma) \tag{5.3}$$

, where $\Gamma \in \mathbb{R}^{O\times O}$ is the covariance matrix of the truths of all objects. It, together with the transition matrix A, reflects the dependency among objects.

We assume the observation for object o_j from source s_i at time t, i.e., $v^i_{j,t}$, fully depends on the truths at time t, i.e. $\boldsymbol{\mu}_t$, following the multivariate normal distribution

$$\mathbf{v_t} = C\boldsymbol{\mu}_t + \boldsymbol{\epsilon}_t, \quad C = I_O \otimes \mathbf{1_S}. \tag{5.4}$$

. $\mathbf{v_t} \in \mathbb{R}^{OS\times 1}$ is the stacked observation vector including all objects from their sources at timestamp t. It is firstly ordered by sources, then by objects. Specifically, $\mathbf{v_t} = (v^1_{1,t}, \ldots, v^S_{1,t}, v^1_{2,t} \ldots, v^S_{O,t})^T$. $\otimes$ is Kronecker product. $\mathbf{1_S} \in \mathbb{R}^{S\times 1}$is a vector in which all elements are ones. $I_O \in \mathbb{R}^{O\times O}$ is the identity matrix thus $C \in \mathbb{R}^{OS\times O}$. Implied by Eq. 5.4, we assume the mean of the observations are the centered at the truths. If $v^i_{j,t}$ is not provided by source s_i, we regard it as *missing data*. Note that the missing data is prevalent in the real truth discovery cases, where not all sources will provide observations for every object.

We assume $\boldsymbol{\epsilon}_t$ follows multivariate normal distribution independently as follows

$$\boldsymbol{\epsilon}_t \sim Normal(0, \Pi) \tag{5.5}$$

, where the diagonal blocks of Π are denoted by Σ, and the off-diagonal blocks all 0. The diagonal elements of Σ $(\sigma^2_1, \sigma^2_2 \ldots, \sigma^2_S)$ are interpreted as source quality, because large variability of the observations could most likely be from unreliable sources. The off-diagonal elements of Σ represent the correlation between each pair of sources. The observations from the same source or each pair of sources will share the same diagonal or non-diagonal parameters from Σ. Thus the source quality σ^2_i and source dependency $\sigma_{i,i'}$ are actually estimated from observations on all objects. If we assume sources are not correlated to each other,

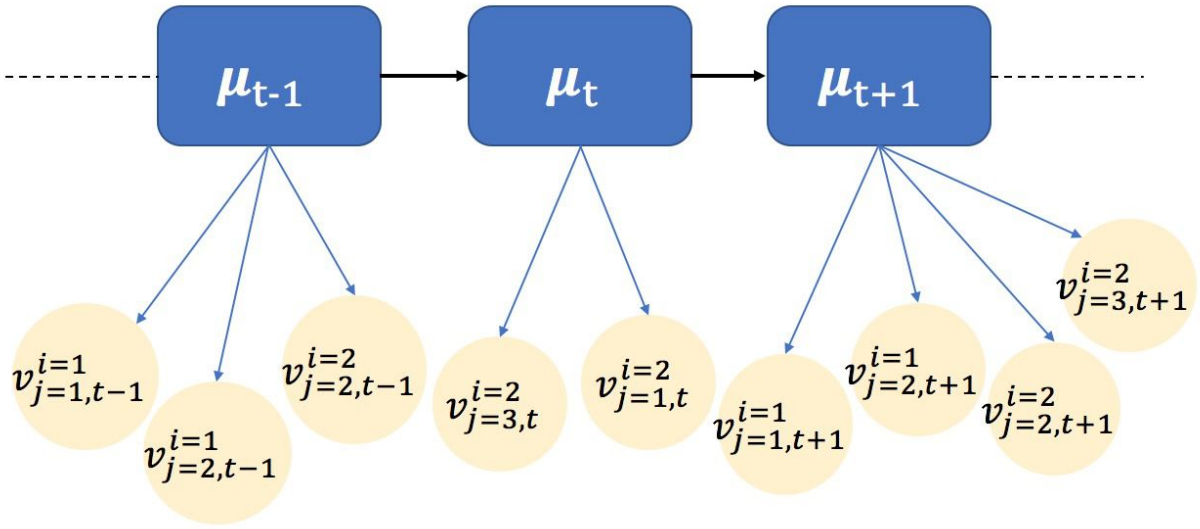


Figure 5.1: Hidden Markov model with observations from multiple sources

$\sigma_{i,i'}$ can also set to 0. Otherwise, they can be estimated from the data. We will discuss both the diagonal and non-diagonal cases in Chapter 5.3.

The hidden Markov model is composed of Eq. (5.1)-(5.5). The parameters to estimate are transition matrix A, objects covariance matrix Γ, initial truth parameters $\boldsymbol{\pi}_1$, V_1 and source quality covariance matrix Σ. Given parameter values, Kalman filter and smoother are typical methods Harvey (1990) to infer the latent truth at timestamp t by estimating $E(\mu_t|v_{1:t})$ (filtering) or $E(\mu_t|v_{1:T})$ (smoothing). The essential difference is that when estimating the expected value of current latent truth, filtering only uses previous observations, while smoothing uses observations from the past, the present and the future.

Model inference

To estimate the parameters in the model, both maximum likelihood and Bayesian methods are available. We refer Durbin and Koopman (2012) for background discussions on hidden Markov models. In this work, we adopt EM algorithm to estimate the Kalman filter and smoother, the time-variant truths and the parameters iteratively. In most cases, not all sources will provide observations for all objects at any timestamp t. It is prevalent that one source does not provide any observation about some objects at timestamp t. Here, we treat unavailable data as missing data. We adopt the EM algorithm Shumway and Stoffer (1982) to infer the truths, source quality and dependency information based on our model with missing data.

The joint log likelihood of the complete data $\boldsymbol{\mu}_{1:T}, \mathbf{v_{1:T}}$ can be written in the following form

$$\begin{aligned}\log P(\boldsymbol{\mu}_{1:T}, \mathbf{v_{1:T}}) &= -\frac{1}{2}\log|V_1| - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\pi}_1)V_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\pi}_1) \\ &- \frac{T}{2}\log|\Gamma| - \frac{1}{2}\sum_{t=2}^{T}(\boldsymbol{\mu}_t - A\boldsymbol{\mu}_{t-1})^T\Gamma^{-1}(\boldsymbol{\mu}_t - A\boldsymbol{\mu}_{t-1}) \\ &- \frac{TO}{2}\log|\Sigma| - \frac{1}{2}\sum_{t=1}^{T}(\mathbf{v_t} - C\boldsymbol{\mu}_t)^T(I_O \otimes \Sigma^{-1})(\mathbf{v_t} - C\boldsymbol{\mu}_t)\end{aligned} \tag{5.6}$$

.

E-step:

We use $\mu_{t|\tau}$ to denote the conditional expectation $E(\mu_t|v_{1:\tau})$, $V_{t|\tau}$ to denote the conditional covariance matrix $Var(\mu_t|v_{1:\tau})$ and $V_{t,t-1|\tau}$ to denote the conditional cross-covariance matrix $Cov(\mu_t, \mu_{t-1}|v_{1:\tau})$ for $t, \tau \in \{1, 2, \ldots, T\}$. Starting from $t = 1$, we have the following Kalman filter forward recursions at r-th round

$$\begin{aligned}\mu_{t|t-1} &= A_{\langle r\rangle}\mu_{t-1|t-1} \\ V_{t|t-1} &= A_{\langle r\rangle}V_{t-1|t-1}A_{\langle r\rangle}^T + \Gamma_{\langle r\rangle} \\ \mu_{t|t} &= \mu_{t|t-1} + K_t(\mathbf{v}_t^* - C^*\mu_{t|t-1}) \\ V_{t|t} &= V_{t|t-1} - K_tC^*V_{t|t-1} \\ K_t &= V_{t|t-1}C^{*T}(C^*V_{t|t-1}C^{*T} + \Pi_{\langle r\rangle})^{-1}\end{aligned} \tag{5.7}$$

, where $\mathbf{v}_t^*$ is the vector by entering zeros in the $\mathbf{v}_t$ if the object is not observed and C^* is the matrix by zeroing out the corresponding row of the matrix C in Eq. (5.4). $A_{\langle r\rangle}$, $\Gamma_{\langle r\rangle}$ and $\Pi_{\langle r\rangle}$ are the parameters estimated from r-th round of M-step. The Kalman gain K_t is deducted by minimizing the trace of the covariance matrix $V_{t|t}$. The estimation of current timestamp $\mu_{t|t}$ is the combination of the prediction from previous timestamp and the current observations from all sources, and Kalman gain automatically balances these two parts.

Eq. (5.7) also reflects the advantages of the use of source quality. If we assume latent truths from different objects are conditional independent at time t, and sources are independent, the terms $C^*V_{t|t-1}(C^*)^T$ and Σ in K_t will be diagonal. When the j-th source is more reliable, i.e. the corresponding variance σ_j^2 in Σ is small, the entries related to s_j in K_t will be large. It would put more weight of the observations from s_j on the estimation of latent truth, and also put more weight in reducing the uncertainty, i.e. $V_{t|t}$.

The initial state prediction is $\mu_{1|0} = \boldsymbol{\pi}_{1,\langle r\rangle}$ and $V_{1|0} = V_{1,\langle r\rangle}$. Starting from $t = T$, we have Kalman

smoother backward recursions at r-th round using observations from 1 to T

$$\begin{aligned}
\mu_{t-1|T} &= \mu_{t-1|t-1} + J_{t-1}(\mu_{t|T} - A_{\langle r\rangle}\mu_{t-1|t-1}) \\
V_{t-1|T} &= V_{t-1|t-1} + J_{t-1}(V_{t|T} - V_{t|t-1})J_{t-1}^T \\
J_{t-1} &= V_{t-1|t-1}A_{\langle r\rangle}^T(V_{t|t-1})^{-1} \\
V_{t-1,t-2|T} &= V_{t-1|t-1}J_{t-2}^T + J_{t-1}(V_{t,t-1|T} - A_{\langle r\rangle}V_{t-1|t-1})J_{t-2}^T
\end{aligned} \tag{5.8}$$

with initial value $V_{T,T-1|T} = A_{\langle r\rangle}V_{T-1|T-1} - K_T C^* A_{\langle r\rangle}V_{T-1|T-1}$.

Blocked Kalman filter and smoother in E-step

In the E-step, we use Kalman filter and Kalman smoother to estimate the covariance matrix and the cross-covariance matrix which will be used in M-step. If large numbers of objects are considered together, the dimension of $\boldsymbol{\mu}_t$ will be high. The matrix calculation in E-step will take huge memory and turn out to be computationally expensive. If objects are not correlated, or only correlated in their own group, the computational cost may be greatly reduced. Here, we provide a blocked Kalman filter and smoother when the following conditions are satisfied. (1) Given truths at time $t-1$, μ_{t-1}, truths in different blocks are independent at timestamp t. (2) Observations in different blocks are independent given truths at any timestamp t. Then, conditional latent truth of block b, i.e. $\mu_t^b|v_{1:t}$ and $\mu_t^b|v_{1:T}$, are independent among different blocks. That is, we can implement Kalman filter and Kalman smoother in E-step for decomposed blocked objects independently. Then we can update the blocks of $E(\mu_t|v_{1:t})$ and $E(\mu_t|v_{1:T})$ independently by conducting the same Kalman filtering and smoothing Eq. (5.7), (5.8) based on block related observations $v_{block,t}^{1:S}$ and parameters.

M-step:

In the M-step, we maximize the conditional expectation of log likelihood in Eq. (5.6) with respect to the latent truths. We add prior distribution of source quality matrix Σ in log likelihood. We can update parameters at r-th iteration as follows.

If we assume sources are independent, Σ will be a diagonal matrix with elements $\{\sigma_{1:S}^2\}$. We set the prior distribution of σ_i^2 as independent inverse gamma distribution, $Inv-Gamma(\alpha_i, \beta_i)$, and σ_i^2 is in the form

$$\begin{aligned}
\sigma_{i,\langle r+1\rangle}^2 &= \frac{2\beta_i + \sum_{j=1}^{O}\sum_{t=1}^{T} E((v_{j,t}^i - \mu_{j,t})^2|v_{1:T})}{2(\alpha_i+1) + \sum_{t=1}^{T} c_t^i} \\
&= \frac{2\beta_i + \sum_{j=1}^{O}\sum_{t=1}^{T} ({v_{j,t}^i}^2 - 2v_{j,t}^i\mu_{j,t|T} + \mu_{j,t|T}^2 + V_{j,t|T})}{2(\alpha_i+1) + \sum_{t=1}^{T} c_t^i}
\end{aligned} \tag{5.9}$$

, where $\mu_{j,t|T}$ and $V_{j,t|T}$ are the element of object o_j in $\mu_{t|T}$ and $V_{t|T}$ obtained in E-step, respectively.

If sources are dependent, we set the prior distribution of Σ inverse Wishart distribution, $\mathcal{W}(\Phi, \nu)$, and $\Sigma_{\langle r+1\rangle}$ is in the form

$$\Sigma_{\langle r+1\rangle} = \frac{\sum_{t=1}^{T}\sum_{j=1}^{O} D_{jt}\Sigma_{jt}^{i}D_{jt} + \Phi}{T \times O(T \times O + S + \nu + 1)} \tag{5.10}$$

, where $D_{jt} \in \mathbb{R}^{S\times S}$ is the permutation matrix that switches missing values in the observations of o_j, i.e. $\mathbf{v}_{j,t} = (v_{j,t}^1, v_{j,t}^2, \ldots, v_{j,t}^S)^T$, to the end in order. That is, $D_{jt}\mathbf{v}_{j,t} = (\mathbf{v}_{j,t}^{(1)}, \mathbf{v}_{j,t}^{(2)})$ where $\mathbf{v}_{j,t}^{(1)}$ is the observed portion and $\mathbf{v}_{j,t}^{(2)}$ is the unobserved portion, where (1) and (2) are the lengths of the two vectors. The corresponding entries of o_j matrix C in Eq. (5.4) is permuted to $(C_{jt}^{(1)}, C_{jt}^{(2)})^T = D_{jt}$. The expression of covariance matrix of $(\mathbf{v}_{j,t}^{(1)}, \mathbf{v}_{j,t}^{(2)})$ is as follows

$$\Sigma_{jt}^{i} = \begin{Bmatrix} \Sigma_{jt}^{11} & \Sigma_{jt}^{12} \\ \Sigma_{jt}^{21} & \Sigma_{jt}^{22} \end{Bmatrix} \tag{5.11}$$

, where

$$\begin{aligned}
\Sigma_{jt}^{11} &= (v_{j,t}^{(1)} - C_{jt}^{(1)}\mu_{j,t|T})(v_{j,t}^{(1)} - C_{jt}^{(1)}\mu_{j,t|T})^T + C_{jt}^{(1)}(V_{j,t|T})(C_{jt}^{(1)})^T \\
\Sigma_{jt}^{12} &= (\Sigma_{jt}^{21})^T = \Sigma_{jt}^{11}(\Sigma_{\langle r\rangle}^{11})^{-1}\Sigma_{\langle r\rangle}^{12} \\
\Sigma_{jt}^{22} &= \Sigma_{\langle r\rangle}^{22} - \Sigma_{\langle r\rangle}^{21}(\Sigma_{\langle r\rangle}^{11})^{-1}\Sigma_{\langle r\rangle}^{12} + \Sigma_{\langle r\rangle}^{21}(\Sigma_{\langle r\rangle}^{11})^{-1}\Sigma_{jt}^{11}(\Sigma_{\langle r\rangle}^{11})^{-1}\Sigma_{\langle r\rangle}^{12}
\end{aligned} \tag{5.12}$$

, and the r-th iteration of source covariance matrix Σ with respect to $(\mathbf{v}_{j,t}^{(1)}, \mathbf{v}_{j,t}^{(2)})$ is denoted as $\begin{Bmatrix} \Sigma_{\langle r\rangle}^{11} & \Sigma_{\langle r\rangle}^{12} \\ \Sigma_{\langle r\rangle}^{21} & \Sigma_{\langle r\rangle}^{22} \end{Bmatrix}$.

The initial state parameters π_1 and V_1 are estimated as follows

$$\pi_{1,\langle r+1\rangle} = \mu_{1|T}, \quad V_{1,\langle r+1\rangle} = V_{1|T} \tag{5.13}$$

.

The transition matrix A is estimated in the following way

$$A_{\langle r+1\rangle} = (\sum_{t=2}^{T} V_{t,t-1|T} + \mu_{t|T}\mu_{t-1|T}^T)(\sum_{t=2}^{T} V_{t-1|T} + \mu_{t-1|T}\mu_{t-1|T}^T)^{-1} \tag{5.14}$$

. If A is blocked matrix following the rules in Chapter 5.3.2, the blocks in A can be calculated in the same way of Eq. (5.14) using corresponding blocked objects information in $V_{t,t-1|T}, \mu_{t|T}, \mu_{t-1|T}$ and $V_{t-1|T}$.

The covariance matrix of the truths Γ is updated as follows

$$\Gamma_{\langle r+1\rangle} = \frac{1}{T-1}(\sum_{t=2}^{T} V_{t|T} + \mu_{t|T}\mu_{t|T}^T - A_{\langle r+1\rangle} \sum_{t=2}^{T} V_{t,t-1|T} + \mu_{t|T}\mu_{t-1|T}^T) \tag{5.15}$$

. If Γ is blocked matrix following the rules in Chapter 5.3.2, the blocks in Γ will be calculated in the same way of Eq. (5.15) using corresponding block information in $V_{t,t-1|T}$, $\mu_{t|T}$, $\mu_{t-1|T}$, $V_{t|T}$ and A.

Comparison with previous methods

To explicitly illustrate the power of our model, we compare the deduction of our model with the evolving truth model Li et al. (2015). In Li et al. (2015), source quality is modeled as source weight in the following way

$$w_i = \frac{2(\alpha_i - 1) + \sum_{t=1}^{T} \gamma^{T-t} c_t^i}{2\beta_i + \theta \sum_{j=1}^{O} \sum_{t=1}^{T} (\gamma^{T-t}(v_{j,t}^i - \mu_{j,t})^2)} \tag{5.16}$$

, where (α_i, β_i) is the parameter of Gamma prior distribution, $\mu_{j,t}$ is estimated using source weighted sum of observations, γ is the decay factor used to adjust the source weight, and θ is the regularization parameter. Compared with the update of our source quality σ_j^2 in Eq. (5.9), we can see that our source quality parameter σ_j^2 is similar to the inverse of source weight w_j in Li et al. (2015), i.e. $\sigma_j^2 \approx 1/w_j$. Therefore, if we assume truths are independent of each other, sources are independent of each other, and observations given the truths are conditionally independent, our model will reduce to a similar solution of the model in Li et al. (2015). The key advantage of our model, in addition to the power to model dependencies, is to use the population variance adjusted by Kalman smoother rather than the population variance adjusted by a fixed decay factor.

The updated $\mu_{j,t}$ in Li et al. (2015) is

$$\mu_{j,t} = \frac{\sum_{i=1}^{n_i} w_i v_{j,t}^i + \lambda \hat{v}_{j,(t-1)}^*}{\sum_{i=1}^{n_i} w_i + \lambda} \tag{5.17}$$

. Comparing with the update of $E(\mu_{j,t}|v_{1:t})$ in Eq. (5.7) and $E(\mu_{j,t}|v_{1:T})$ in Eq. (5.8) where K_t is Kalman Gain matrix related to source quality matrix Σ, we can see both Eq. (5.7) and (5.8) balance the new observation and previous state estimation. However, our balance is dynamic based on the estimated Kalman Gain. In Li et al. (2015), the balancing parameter λ is predefined and fixed.

Moreover, by relaxing the independence assumptions, our model is sufficiently general and flexible to provide other estimations such as source correlation and object correlation.

5.3.3 Data preprocessing

The detection of outliers is important in making an accurate estimation of truth and source quality. Implied by Eq. (5.2), (5.3) and (5.5), we utilize the normal distribution to model the truths and observations. In most cases, the outliers of observations are unlikely to be the truth in most practical cases. The normalization of observations is also necessary especially we take into account a large number of objects. The expressions in Eq. (5.9)-(5.12) imply that if one object has an extreme large scale compared to others, this object would dominate the source quality, making the source quality estimation biased.

We implement data preprocessing step before EM algorithm.To detect outliers of the observations of each object at time t, we use the median absolute deviation to find outliers Iglewicz and Hoaglin (1993). After removing all outliers, we normalize the observations to its z-scores as the input of the EM algorithm. Specifically, for each object o_j at time t, normalized $v_{j,t}^i$ is $(v_{j,t}^i\text{-mean}(v_{j,t}^{1:S}))/\text{std}(v_{j,t}^{1:S})$. When there are no sufficient data at time t for object o_j, we aggregate observations from consecutive timestamps in a fixed length sliding window to normalize the data. The detailed data preprocessing algorithm is described in Algorithm 3. The observation whose modified z-score is larger than the threshold parameter δ will be classified as outliers and removed from the estimation. Following Iglewicz and Hoaglin (1993), we use 0.6745 as the constant multiplier in the modified z-score and set the threshold $\delta = 3.5$.

Data: The observation vector $v_{j,t}$ for object o_j from S sources at time t.
{Outlier Detection:}
med=nanmedian($v_{j,t}$) ; `// calculate the median without missing value.`
$diff$=abs($v_{j,t} - med$) ; `// calculate the absolute deviations between` $v_{j,t}^i$ `and its median.`
med_abs_dev=nanmedian($diff$) ; `// calculate the median of absolute deviations.`
if $med\_abs\_dev == 0$ **then**
  for *i in 0:(S-1)* **do**
    if $diff[i] == 0.0$ *or isnan(*$diff[i]$*)==True* **then**
      $out\_lier[v_{j,t}[i]] \leftarrow False$
    else
      $out\_lier[v_{j,t}[i]] \leftarrow True$
    end
  end
else
  for *i in 0:(S-1)* **do**
    $modified\_zscore[i] = 0.6745 * diff[i]/med\_abs\_dev$ **if** $modified\_zscore[i] > \delta$ **then**
      $out\_lier[v_{j,t}[i]] \leftarrow True$
    else
      $out\_lier[v_{j,t}[i]] \leftarrow False$
    end
  end
end
{Normalization:calculating z-scores:}
The normalized z-score of $v_{j,t}^i$ from source s_i is $(v_{j,t}^i - mean(v_{j,t}))/std(v_{j,t})$

Algorithm 3: Data Preprocessing

5.3.4 Online solution: $EvolveT(T^*)$

The key challenge of integrating streaming data is that (1) we are not able to use the future data to estimate the present, and (2) using all the data at each timestamp is time-consuming. Here, we propose an online solution. When a new data point arrives, we update the estimation of current truth with the parameters at previous timestamp using Kalman filtering one step further without running Kalman smoother backwards in E-step, and update the parameters in M-step in an incremental way with O(1) complexity. The reason we can update the estimates incrementally is that Kalman forward recursion is defined in an incremental way, and parameters in Eq. (5.9), (5.10), (5.13), (5.14) and (5.15) contain only accumulated term $\sum_{t=1}^{T}(\cdot)$. Thus, this online version can incrementally update the truths and parameters sequentially with time complexity $O(T)$.

Though future data is not accessible during the estimation, the historical records can help to better initialize the parameters. We first run the batch solution with both Kalman filtering and smoothing until EM step converges for the first few timestamps, followed by the online version. We call it the pre-train step.

We summarize our entire algorithm in Algorithm 4. We call it *Evolving Truth* algorithm, denoted by $EvolvT(T^*)$. T^* denotes the history length to run the batch-mode version. We will compare the performance of different T^* in the experiments section. For historical T^* timestamps data, we iteratively update Kalman recursions $\mu_{t|T^*}$, $V_{t|T^*}$ based on Eq. (5.7), (5.8) and parameters based on Eq. (5.9), (5.10), (5.13), (5.14) and (5.15) until all of them converge. At the pre-train step, we assume the source quality matrix Σ to be consistent for stable initialization.After first T^* pre-train timestamps, we update truths and parameters in an incremental way and only scan the remaining data once from time T^* to T. Truths are updated based on Kalman forward recursions $\mu_{t|t}$, $V_{t|t}$ in Eq. (5.7). Parameters are updated following Eq. (5.9), (5.10), (5.13), (5.14) and (5.15), but replacing smoothing estimates $(\mu_{t|T}, V_{t|T}, V_{t-1,t|T})$ by filtering estimates $(\mu_{t|t}, V_{t|t}, V_{t-1,t|t})$. Here, the source quality matrix Σ is actually changing over time.

Theoretical analysis

According to Wu et al. Wu (1983), when the complete likelihood function satisfies the continuity condition, then all the limit points of any parameters of an EM algorithm are local maxima of the likelihood, and converge monotonically for some local maximum. Since Eq. (5.6) is continuous in terms of all the parameters, our EM algorithm will achieve the local maxima of the log likelihood.

Data: The observation vector $v_{j,t}$ for object o_j from S sources at time t.
{Data preprocessing step:};
Follow Algorithm 3;
{Training step:(if $T^* > 0$)}
while $||para\langle r\rangle - para\langle r-1\rangle||_2 > \delta_{em}$; `// para⟨r⟩ is the vector of all parameters to be estimated at r-th iteration.`
do
  E-Step:
  for *block b in 1:B tcp*B: number of independent blocks* **do**
    for h *in* $1 : T^*$ **do**
      update filtering estimates $\mu^b_{h|h}$ and $V^b_{h|h}$ based on Eq. (5.7) in blocks
    end
    for h *in* $(T^*, T^* - 1, \ldots, 1)$ **do**
      update smoothing estimates $\mu^b_{h|T^*}$, $V^b_{h|T^*}$ and $V^b_{T^*-1,T^*|T^*}$ based on Eq. (5.8) in blocks
    end
  end
  M-Step:
  Update source quality Σ based on Eq. (5.9) or (5.10)
  Update initial state parameters π_1 and V_1 based on Eq. (5.13)
  Update transition matrix A based on Eq. (5.14)
  Update truth covariance matrix Γ based on Eq. (5.15)
end
{Incremental updating step:};
for *t in T*:T*+T* **do**
  for *block b in 1:B* **do**
    update filtering estimates $\mu^b_{t|t}$ and $V^b_{t|t}$ based on Eq. (5.7) in blocks
  end
  Repeat M-Step above and replace smoothing estimates by Kalman estimates if they are used to update parameters
end

Algorithm 4: Evolving Truth algorithm, $EvolvT(T^*)$

5.4 Experiments

In this section we show the effectiveness, efficiency and case study of our new model on the real-world datasets. All the experiments are conducted on a laptop with 4 GB RAM, 1.4 GHz Intel Core i5 CPU, and OS X 10.11.6, with Python 3.6.

5.4.1 Experiment setup

Datasets

We adopt the market capitalization data, fight arrival data, weather forecast data and pedestrian counts data to evaluate the algorithms.

Market capitalization data (Stock). The market capitalization data consists of 1000 stock symbols from 55 sources on trading days in July 2011 Li et al. (2012). The ground truth for evaluation is built on NASDAQ100 stocks collected by taking the majority values provided by five stock data providers: nasdaq.com, yahoo finance, google finance, bloomberg and MSN finance.

Flight arrival data (Flight). The flight arrival time contains 3000 flights from 38 sources over 1-month period (31 days) (December 2011) from Li et al. (2012). We normalize the arrival time into minutes. Ground truth for evaluation is provided by corresponding airline websites.

Weather forecast data (Weather). We collect the highest temperature weather forecast data for 88 U.S. cities from 6 websites: wunderground.com, worldweatheronline.com, openweathermap.org, DarkSky.net, APIXU.com and yahoo.com. The data last for 2 months, from June 8th, 2017 to August 8th, 2017 (61 days). We also collect the actual highest temperature (°F) observations as the ground truth. **Pedestrian data (Pedestrian)**. Given by Dublin City Council [8] , the data consist of daily pedestrian counts of four streets in 2015. There are many sources that may provide the pedestrian counts, such as sensors from traffic light, surveillance cameras, infrared beam counters, etc. Since it is not easy to collect the real data from the aforementioned sources, we simulate six different sources by varying the Gaussian noise level with different variances. We use data from November 1st to December 31st and set the variances as 1, 1.44, 1.96, 2.56, 3.24, 4, respectively.

Evaluation metrics

We use mean absolute error and root mean square error to measure the correctness of all the truth discovery algorithms.

[8] https://data.gov.ie/dataset/pedestrian_footfall

Mean Absolute Error (MAE) is the average of the sum of absolute distance between truth and the estimated value.

Root Mean Square Error (RMSE) is the root of the average on the sum of square loss between the truth and the estimated value.

Compared methods

A large portion of the existing methods are only working on stable truths. To demonstrate the advantages of modeling dynamic truths, we treat the streaming data in a batch way to run these methods.

The following algorithms are designed for categorical truth, where distance between answers are not measurable in Euclidean distance. Thus, the true answer is selected from one of the candidate answers. Since our paper is focused on numerical truth, we treat numerical truth as one of the candidate to fit in the models. **TruthFinder** Yin et al. (2008) and **AverageLog** Pasternack and Roth (2010) iteratively estimate the truths and source quality using additive or multiplicative ways. In **Investment** and **Pooled-Investment** Pasternack and Roth (2010), each source uniformly invests its quality among the answers they provide, and its quality is a weighted sum of the credibility of those answers. **3Estimates** Galland et al. (2010) extends the framework further by introducing an additional factor, i.e. difficulty of the question when evaluating the truths and source quality.

For these methods modeling continuous data, we list them all as follows. **Median** and **Mean** are two naive methods that do not consider the source quality and take median or mean at each timestamp independently. **GTM** Zhao and Han (2012) is a probabilistic graphical model designed for continuous data in static truth discovery. **DynaTD+ALL** is the incremental version with both decay and smoothing factors and the best performed algorithm reported in Li et al. (2015). We also use our data preprocessing algorithm for the data.

For our proposed method, *EvolvT*, we develop a set of different versions by varying the pre-train step T^*. We take three sets of (T^*) to illustrate their difference. $EvolveT(0)$ is a fully online version starting with no historical data and random initialized parameters. Parameters such as source quality, transition matrix, and object covariance matrix will also be updated along time. $EvolvT(t)(t \neq 0)$ is to use historical t timestamps to evaluate truths and parameters, then conduct the $O(T)$ algorithm to infer the truths left. *EvolvT* is a fully batch-mode version when we observe all the streaming data and estimate the truths at all timestamps and fix source quality along time. parameter from last timestamp, this method For all baseline methods, we use the suggested parameters, initializations and convergence conditions in the original papers. For our model, the parameters are set $\nu = 2$, $\Phi = (S + \nu + 1 * I)$, $\alpha_i = \beta_i = 10$ for each source s_j, $\boldsymbol{\mu}_1 = \mathbf{0}$,

Table 5.1: Performance comparison

	Stock			Flight			Weather			Pedestrian		
Methods	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)
TruthFinder	3.49	9.44	98.12	1.19	5.85	50.04	2.81	3.33	2.11	1.79	2.28	1.11
3Estimates	6.05	27.70	208.99	28.90	119.90	146.75	3.34	4.42	1.76	1.79	2.28	1.11
AverageLog	3.60	9.70	31.53	0.98	4.85	22.20	2.53	3.65	0.78	1.79	2.28	0.05
Investment	6.05	27.50	44.48	28.91	119.90	30.76	3.34	4.42	0.93	1.79	2.28	0.05
PooledInv	3.23	11.19	50.69	**0.39**	**3.13**	44.68	2.55	3.36	1.41	1.79	2.28	0.08
Median	2.97	13.29	39.93	2.55	6.40	45.44	2.49	3.20	3.08	0.79	0.95	0.18
Mean	4.60	19.26	57.87	5.78	10.75	62.19	2.46	3.25	4.42	0.87	1.01	0.36
GTM	2.24	11.83	103.28	2.70	3.31	125.82	2.55	3.35	0.70	0.98	1.19	0.10
GTM+ours	1.63	9.27	131.32	3.37	6.73	150.35	2.46	3.53	2.75	0.80	0.96	0.11
DynaTD+All	2.05	8.01	8.82	2.89	8.48	0.23	2.97	4.45	10.23	0.78	1.01	0.16
EvolvT	2.05	7.97	3.63	2.54	4.31	140.1	2.52	3.78	54.78	0.72	0.95	0.12
EvolvT(0)	1.96	8.35	4.01	2.80	9.18	0.49	2.64	3.49	2.33	0.72	0.93	0.03
EvolvT(5)	1.86	7.54	7.93	3.25	8.53	3.85	2.48	3.28	11.38	**0.69**	**0.89**	0.04
EvolvT(10)	**1.54**	**6.91**	3.85	3.31	8.61	22.89	**2.42**	**3.21**	22.89	0.70	0.90	0.06

V_1 and A to identity matrix. We evaluate *Evolve* in the batch mode without any historical data. As for T^*, we set it to [0, 5, 10], and evaluate all methods from 11th timestamps for fair comparison.

5.4.2 Experimental results

In this section, we empirically demonstrate the effectiveness and efficiency of our algorithm, $EvolvT(T^*)$, and illustrate the impact of different factors to the estimation performance.

Dynamic Truth Inference

Table 5.1 shows the performance comparison of all the models. In general, the methods for numerical data perform better that those for categorical data on stock, weather and pedestrian datasets, and dynamic models performs better than static numerical models. Our method are significantly better than all other algorithms on stock, weather and pedestrian datasets. With pre-train step of historical records, our method gets better performance than random initialization $EvolvT(0)$. The batch-mode version performs slightly better than online version, but would cost more time due to the convergence requirements.

Discussion on numerical truth discovery

If we assume truth is from one candidate, choose the candidate which is most closed to our estimated truth. One interesting finding is that though flight arrival time is real-valued parameter, whose difference can be measured using Euclidean distance, the actual distribution of flight arrival time observations do not follow a normal distribution. Due to the delays or accidental incidents, some websites may not immediately update the information timely, or not updated until the flight actually arrives, leading to scattered distribution

centered at some discrete value. Thus, the optimal solution, with high probability, will be one of the time provided by certain sources. Table 5.1 shows that TruthFinder, AverageLog, and PooledInvestment have better performance than the Gaussian-based method, GTM.

Efficiency

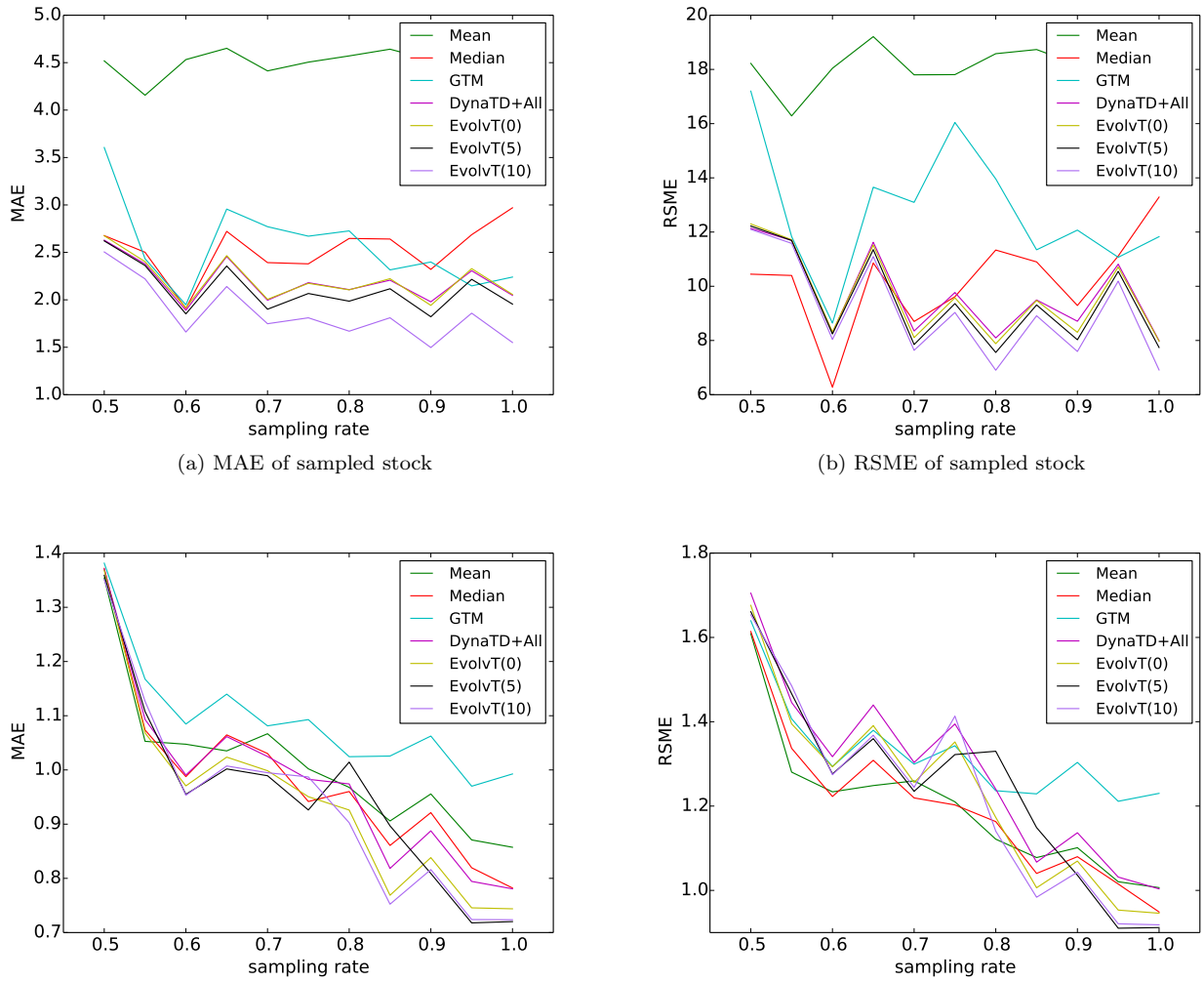
We also report the running time of all the algorithms. For efficiency, the running time (s) of our proposed model is close to Median and Mean. Iterative methods usually takes 10 times more times to converge at each timestamp, while our single-pass O(T) version gains better performance with even shorter running time. Our algorithm can largely reduce the running time by single-pass O(T) algorithm without loss in performance. The reason behind it is that algorithm can keep track of the parameter information from the history such that source quality, object dependency and prediction power are properly inherited, while the batch-mode algorithms do not use the history information, making the running time relatively long.

Missing observations.

To illustrate the robustness of our algorithm to missing observations, we randomly remove some of the observations of sources. Since we have shown in Table 5.1 that the methods specifically working on numerical truth work consistently better than those on categorical truth, we only run part of the baseline algorithms for further comparison on missing observations. The sampling rate demonstrates the proportion that we keep out of all the data. We range the sampling rate in $[0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.]$. Figure 5.2 shows the MAE and RSME changes along with the sampling rate. We can observe that our method $Evolve(10)$ performs the best in terms of MAE in most cases on both stock and pedestrian datasets. One interesting observations is that Median performs well in terms of RSME when we only keep half of the data on both datasets. one our model can still effectively give an estimation on the missing data. One possible reason is that as the number of data points significantly drops, there are no sufficient data to estimate the parameters of the truth discovery model.

Effectiveness of data preprocessing.

To demonstrate the effectiveness of our data preprocessing step, we replace the outlier detection and normalization step of GTM, denoted by GTM+ours. We can see that using our preprocessing can reduce the MAE of GTM on three datasets.



(a) MAE of sampled stock

(b) RSME of sampled stock

(c) MAE of sampled pedestrian

(d) RSME of sampled pedestrian

Figure 5.2: MAE, RSME of sampled stock, pedestrian

Source dependency

We also compare the performance with and without the source dependency assumption. Table 5.2 list the performance of dependent sources model. We find that assuming all sources are correlated would not have a good estimation on the truths. The major reason is that the number of parameters is square of the number of sources in the dependency case.

For source dependency study, we also rank all the sources by their quality at the last timestamp, i.e. $1/\sigma_i^2$ from top to low, and separate them into three different groups. The sources in highest quality group is listed in Figure 5.3a, and the corresponding source name is listed on our website [9]. 2:bloomberg, 21:tmx-quotemedia and 10:investoguide,24:yahoo-finance have highest correlation in this group. Figure 5.3b and

[9]http://shizhi2.web.engr.illinois.edu/

Table 5.2: Performance of the dependent sources model

	Stock		Flight		Weather	
Methods	MAE	RMSE	MAE	RMSE	MAE	RMSE
dep-EvolvT(0)	2.59	9.24	3.18	15.78	3.60	2.44
dep-EvolvT(5)	2.58	8.74	3.17	16.23	3.78	11.38
dep-EvolvT(10)	2.56	8.65	3.50	16.3	3.40	22.94

5.3c shows the second and last quality group with decreasing source quality.(4:cnn-money, 15:optimum) are highly correlated, and (barrons[10], and marketwatch[11], screamingmedia) are highly correlated. We further check the the origins of these three websites, and find barrons and marketwatch are all owned by Dow Jones & Company, which is an American publishing and financial information firm, and screamingmedia is just pre-owned by Dow Jones & Company[12]. With high possibility, these websites get information from an identical source.

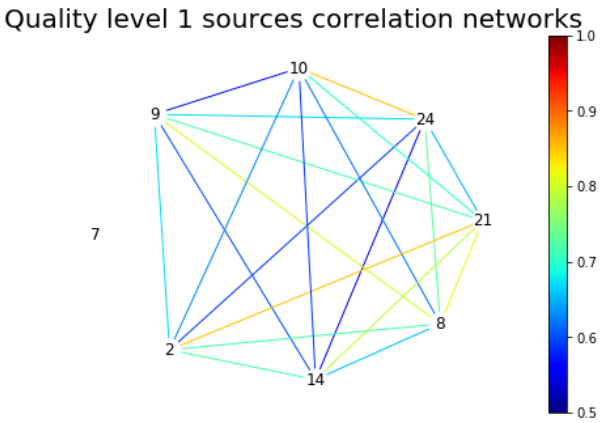
As for the flight data, we plot the graph with all sources as shown in Figure 5.3d. (6:flights, 7:businesstravellogue, 8:flylouisville) are highly correlated, and (9:flightview, 10:panynj, 11:gofox, 12:foxbusiness, 13:allegiantair, 14:boston) are highly correlated. One possible reason is they may copy the flight information from each other, or they achieve information from similar sources.

For the weather data, we find that within the 6 sources, only APIXU and worldweatheronline are highly correlated, with correlation score=0.9. We further check on the web and find that World Weather Online acquires Apixu platform [13]. Thus, the source correlation is validated, showing that our method can effectively detect the source dependencies.

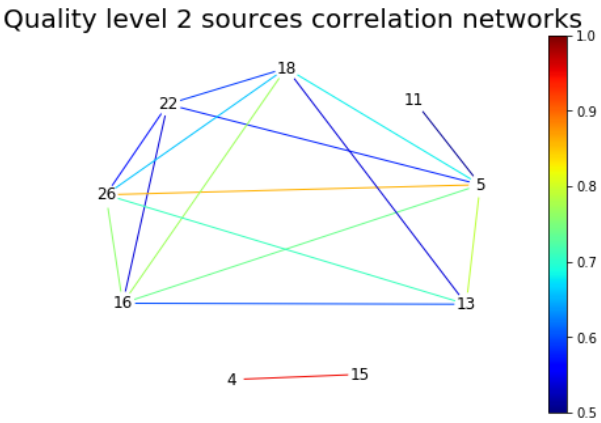
5.5 Conclusions

In this chapter, we aim to discovery the evolving truths from the streaming data. We propose a general hidden Markov model with analytical solution to effectively model the temporal dynamics of the truths with the noisy observations from multiple sources. We propose both the batch version, blocked version and an efficient online version algorithms with effective data preprocessing step. These three methods reduce the computational cost and boost the performance, allowing the real-time truth discovery applications. To address the efficiency, we propose the blocked version and one-pass version to reduce the computational cost and to allow online inference. Experiments on the real-world datasets demonstrate the great effectiveness of our algorithm. In the future, it would be interesting to examine the object dependency using our model.

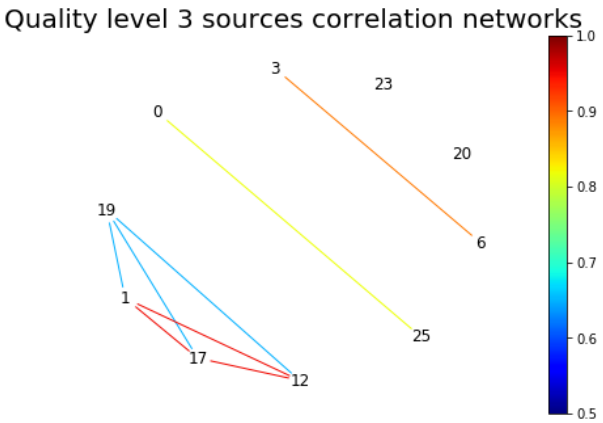
[10]https://en.wikipedia.org/wiki/Barron%27s_(newspaper)
[11]https://en.wikipedia.org/wiki/MarketWatch
[12]http://adage.com/article/btob/dow-jones-sells-screaming-media-yellowbrix/276811/
[13]https://www.facebook.com/worldweatheronline/posts/1160440014032686



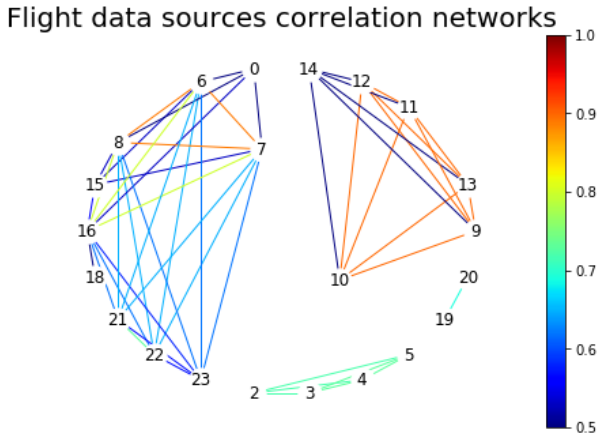
(a) Stock: quality level 1 group



(b) Stock: quality level 2 group



(c) Stock: quality level 3 group



(d) Flight: all sources

Figure 5.3: Source dependency.

Chapter 6

Future directions

In Chapters 2 and 3, we study the statistical inference methods based on characteristic functions. In the future work, we might be able to extend the pseudo-marginal MCMC algorithm proposed in Chapter 2 and approximated ECF methods in Chapter 4 to Markov processes with implicit conditional characteristic functions.

6.1 An MCMC approach for Lévy process based models in finance

In Chapter 2, we propose an MCMC approach to do inference based on characteristic functions. It mainly can be applied to Lévy processes. Actually, we might be able to extend it to certain Markov processes, Lévy process based models, because their conditional characteristic functions are usually available (See Chapter 3.2.3). Specifically, we can follow the logic in Chen et al. (2013) to build our integrated empirical likelihood below.

Let $\{X_{t\delta}\}_{t=1}^{n}$ be n discretely sampled observations of Markov process. For notation simplification, we denote $X_{t\delta}$ as X_t. Suppose that we can get conditional characteristic function:

$$\Phi_t(u;\theta) = E_\theta(e^{iu^T X_{t+1}}|X_t).$$

Thus, for all $u \in R^d$:

$$E_\theta(\Phi_t(u;\theta) - e^{iu^T X_{t+1}}|X_t) = 0.$$

Then, for any weight factor, $A(X_t, u, s)$, and all u,

$$E_\theta[(\Phi_t(u;\theta) - e^{iu^T X_{t+1}})A(X_t, u, s)] = 0.$$

For any $\tau = (u, s) \in R^{2d}$, define:

$$\epsilon_t(\tau, \theta) = (\Phi_t(u; \theta) - e^{iu^T X_{t+1}})A(X_t, u, s).$$

Thus, we have the moment condition:

$$E(\epsilon_t(\tau, \theta)) = 0.$$

Due to the Markov property, also for $t1 \neq t2$, we have

$$Cov(\epsilon_{t1}(\tau; \theta), \epsilon_{t2}(\tau; \theta)) = 0.$$

Because the observations are equally spaced, $\epsilon_{t1}(\tau; \theta)$ has the same distribution as $\epsilon_{t2}(\tau; \theta)$ when $t_1 \neq t_2$. Thus, we use the same EL strategy in Chapter 2.2 to deal with Markov process under the moment condition above.

$$\begin{aligned} L_n(\tau; \theta) = \max\{&\prod_{j=1}^{n}(nw_j) : w_j \geq 0, \sum_{j=1}^{n} w_j = 1, \\ &\sum_{j=1}^{n} p_j \tilde{\epsilon}_j(\tau, \theta) = 0, \} \end{aligned}$$

where $\tilde{\epsilon}_j = (\epsilon_j^R, \epsilon_j^I)^T$. Then, we can follow the exact same procedure in Chapter 2.2 to build the integrated empirical likelihood $T_n(\theta)$:

$$T_n(\theta) = \int_S L_n(\tau; \theta) dG(\tau), \tag{6.1}$$

and $G(\tau)$ is either a given discrete distribution or a smooth distribution function of τ with support on compact set $S \subset R^{2d}$.

With a prior specification $p_0(\theta)$ on the parameter θ, we have the posterior density

$$p(\theta|X) \propto p_0(\theta) T_n(\theta). \tag{6.2}$$

Then, the maximum empirical posterior estimator (MEPE) $\hat{\theta}$ as

$$\hat{\theta} = \arg\max_{\theta} p(\theta|X). \tag{6.3}$$

Thus, we can follow Chapter 2.3 to study this MEPE's asymptotic properties and utilize the exact same sampling algorithm in Chapter 2.4 to estimate the parameter.

In this framework, our computing dimension is doubled to $2d$. Thus, the efficiency of sampling algorithm is worthy to study. Moreover, it could be very interesting to study how to design the weight factor $A(X_t, u, s)$ to generally provide a lower asymptotic variance of the MEPE.

6.2 Empirical characteristic function estimation for Lévy process based models in finance

In Chapter 4, we propose an approximated ECF estimation based on characteristic function. It mainly can be applied to Lévy processes. Actually, we might be able to apply it to certain Markov processes, Lévy process based models, if the joint characteristic function is available.

The approach through joint characteristic function with asymptotic property is discussed in Knight and Yu (2002). The idea is to define the moving blocks of data. Specifically, $Z_j = (X_j, \ldots, X_{j+p})^T \quad (j = \{1, \ldots, T-p\})$, could be regarded as $T-p$ moving blocks for $\{X_1, \ldots, X_T\}$. Each block size is p and the joint characteristic of each moving blocks is:

$$\phi(u;\theta) = E(\exp(iu^T Z_j)), \tag{6.4}$$

and the empirical characteristic function is:

$$\phi_n(u) = \int \exp(iu^T z) dF_n(z) = \frac{1}{n}\sum_{j=1}^{n} \exp(iu^T Z_j), \tag{6.5}$$

where $i = \sqrt{-1}$ and $u = (u_1, \ldots, u_{p+1})$.

Then, similar to Chapter 4, we can estimate the parameter by minimizing the distance between joint characteristic function and empirical joint characteristic function:

$$\hat{\theta}_{ECF} = \arg\min_{\theta} \int |\phi(u,\theta) - \phi_n(u)|^2 dG(u)$$

where $G(u)$ is the continuous weight function.

Then, we need define the trapezoidal rule approximation for multiple integration $\int |\phi(u,\theta) - \phi_n(u)|^2 dG(u)$ and build a similar theorem to Theorem 18 to control the error bound of trapezoidal rule approximation. Hopefully, the approximation will again related to the analyticity property of the joint characteristic function.

.

Appendix A

Appendix of Chapter 2

A.1 Lemmas of the positive-definiteness of integrated matrices

Lemma 24.1. *$\{A(t)\}_{t\in R^d}$ are positive-semidefinite $n\times n$ matrices of which entries are continuous functions of $t \in R^d$. $G(t)$ is a probability measure with support S containing an open set $I \subset S$. Then, there will be no open set $\mathcal{I} \subseteq I$ so that $\int_{\mathcal{I}} A(t)dG(t)$ is positive definite.*
$\Leftrightarrow$ there exists non-zero constant β so that $\forall t \in I$, $A(t)\beta = 0$.

Proof. $\Leftarrow$ Note that there exists non-zero constant vector β(not related to t) so that for $\forall t$ $A(t)\beta = 0$ with $\forall t \in I$. Thus, for any open set $\mathcal{I} \subset I$, we have $\int_{\mathcal{I}}(A(t))dG(t)\beta = \int_{\mathcal{I}}(A(t))\beta dG(t) = 0$. That is, $\int_{\mathcal{I}} A(t)dG(t)$ is not full rank.

$\Rightarrow$ We prove it by contradiction. If $\forall \mathcal{I} \subseteq I$, $\int_{\mathcal{I}} A(t)dG(t)$ is not full rank. First we select an arbitrary open set $\mathcal{I}_1 \subseteq I$. Because $\int_{\mathcal{I}_1} A(t)dG(t)$ is not full rank, there exists non-zero constant vector β_1 satisfying $\int_{\mathcal{I}_1} A(t)dG(t)\beta_1 = 0$. Then, $\int_{\mathcal{I}_1} \beta_1^T A(t)\beta_1 dG(t) = 0$. Considering $A(t)$ is positive-semidefinite matrix and $A(t)$ is continuous with respect to t, we have $\beta_1^T A(t)\beta_1^T = 0$ for $\forall t \in \mathcal{I}_1$. Also, $A(t)$ can be written as $P(t)^T P(t)$ by Cholesky decomposition. Then, we have $(P(t)\beta_1^T)^T(P(t)\beta_1) = 0$ which is $A(t)\beta_1 = 0$ for $\forall t \in \mathcal{I}_1$. Because there is no non-zero constant vector β so that $A(t)\beta = 0$ for $\forall t \in I$, there exists $t_2 \notin \mathcal{I}_1$ so that $A(t_2)\beta_1 \neq 0$. Thus, we select another open set $\mathcal{I}_2$ satisfying $\mathcal{I}_2 \supset (\mathcal{I}_1 \cup t_2)$ and $\mathcal{I}_2 \subseteq I$. Because $\int_{\mathcal{I}_2} A(t)dG(t)$ is not full rank, there exists non-zero constant vector β_2 satisfying $\int_{\mathcal{I}_2} A(t)dG(t)\beta_2 = 0$. Definitely we also have $A(t_2)\beta_2 = 0$. Repeating the procedure above, we will have:

1. Open set $\mathcal{I}_1 \subset \mathcal{I}_2 \cdots \subseteq I$;

2. Non-zero constant vector β_i satisfying $A(t)\beta_i = 0$ for $\forall t \in \mathcal{I}_i$ $(i = 1, 2, \ldots)$;

3. For $j = 2, 3, \ldots$, there exists $t_j \in \mathcal{I}_j \setminus \mathcal{I}_{i-1}$ satisfying $A(t_j)\beta_{j-1} \neq 0$. But, $A(t_j)\beta_k = 0$ for $\forall k \geq j$.

Suppose we select $\beta_1, \beta_2, \ldots, \beta_{n+1}$. Then, they must be linearly dependent (β is $n \times 1$ vector). That is, there exists at least one non-zero real numbers $c_1, \ldots, c_{n+1}$ so that $c_1\beta_1 + c_2\beta_2 + \cdots + c_{n+1}\beta_{n+1} = 0$. Then, $A(t_2)(c_1\beta_1 + c_2\beta_2 + \cdots + c_{n+1}\beta_{n+1}) = 0$. Then, we have $c_1 A(t_2)\beta_1 = 0$ which lead to $c_1 = 0$. Similarly, we multiply $A(t_3)$ by $c_2\beta_2 + \cdots + c_{n+1}\beta_{n+1}$ to get $c_2 = 0$. In the end, we will get $c_1 = c_2 = \cdots = c_{n+1} = 0$.

This is a contradiction. ■

Lemma 24.2. *$\{A(t)\}_{t\in R^d}$ is positive-semidefinite $n\times n$ matrices with entries are functions of $t\in S$. Then, we can at least find $t_1, t_2, \ldots, t_{n+1-m}$ so that $A(t_1)+A(t_2)+\cdots+A(t_{n+1-m})$ is full rank where $m=\max_{t\in S}\{rank(A(t))\}$.*

$\Leftrightarrow$ There will be no non-zero constant vector β(not related to t), so that, for any $t\in S$, $A(t)\beta=0$.

What is more, if we suppose that there will be no non-zero constant vector β(not related to t), so that, for any $t\in S$, $A(t)\beta=0$, we can either construct or check $t_1, t_2, \ldots$ in this way: sequentially selecting $\{\beta_i\}_{i=1}^k$ satisfying (A.1). Whenever β_k is non-exist, $A(t_1)+A(t_2)+\cdots+A(t_k)$ is positive definite. You will never select $\{\beta_i\}_{i=1}^k$ endlessly. The maximum of k is $n+1-m$.

Proof. $'\Rightarrow'$ If there exists non-zero constant β so that $A(t)\beta=0$ for $\forall t\in R^d$, then, $(A(t_1)+A(t_2)+\cdots+A(t_{n+1-m}))\beta=0$ for $\forall t_1, t_2, \ldots t_{n+1-m}$. Then, $A(t_1)+\cdots+A(t_{n+1-m})$is not full rank.

$'\Leftarrow'$ First, let's select t_1 so that $rank(A(t_1))=m\leq n$. if $A(t_1)$ is full rank, then $r=1$. If $A(t_1)$ is not full rank, there exists non-zero constant vector β_1 so that $A(t_1)\beta=0$. But, from our condition, there is no non-zero constant vector β to make $A(t)t=0$ for any $t\in R^d$. That is, we can find t_2 so that $A(t_2)\beta_1\neq 0$. If $A(t_1)+A(t_2)$ is full rank, then, $r=2$. Otherwise, there exists non-zero constant vector β_2 so that $(A(t_1)+A(t_2))\beta_2=0$. Because $A(t_1)$ and $A(t_2)$ are non-negative definite matrices, we have $A(t_1)=L_1L_1^T$ and $A(t_2)=L_2L_2^T$ due to Cholesky decomposition where L_1 and L_2 are lower triangular matrices. Then, we have $\beta_2^T(A(t_1)+A(t_2))\beta=(L_1^T\beta_1)^T(L_1^T\beta_1)+(L_2^T\beta_2)^T(L_2^T\beta_2)=0\Leftrightarrow A(t_1)\beta_2=A(t_2)\beta_2=0$. Similarly, we can select t_3 so that $A(t_3)\beta\neq 0$ by using our condition. We repeat the procedure above until $r=k\geq 1$. That is, if $A(t_1)+A(t_2)+\cdots+A(t_k)$ is full rank, $r=k$. Otherwise, $A(t_1)+A(t_2)+\cdots+A(t_k)$ is not full rank, we will use the same procedure above to have $\beta_1, \beta_2, \ldots, \beta_k$ satisfying

$$\begin{aligned} &A(t_1)\beta_1=0, && A(t_2)\beta_1\neq 0,\\ &A(t_1)\beta_2=A(t_2)\beta_2=0, && A(t_3)\beta_2\neq 0,\\ &\quad\vdots && \quad\vdots\\ &A(t_1)\beta_k=\ldots=A(t_k)\beta_k=0, && A(t_k)\beta_{k-1}\neq 0. \end{aligned} \tag{A.1}$$

Note that $\beta_1, \ldots, \beta_k\subseteq\{\beta : A(t_1)\beta\}$ *i.o.* the solution space S of $A(t_1)\beta=0$. It is easy to know $Dim(S)=n-m$ due to linear algebra where $rank(A(t_1))=m$. Now, we prove $\beta_1, \ldots, \beta_k$ are linearly independent. if $c_1\beta_1+\cdots+c_k\beta_k=0$ where $c_1, \ldots, c_k$ are real constant. Then, $A(t_2)(c_1\beta_1+\cdots+c_k\beta_k)=c_1(A(t_2)\beta_1)=0$. Because $A(t_2)\beta_2\neq 0$, we have $c_1=0$. Then, we have $A(t_3)(c_2\beta_2+\cdots+c_k\beta_k)=c_2A(t_3)\beta_2=0$. That is, $c_2=0$ because $A(t_3)\beta_2\neq 0$. Similarly, we have $c_1=c_2=\cdots=c_k=0$. In other words, $\beta_1, \ldots, \beta_k$ are linearly independent. Thus, $k\leq n-m$. That is, if we repeat our procedure $n+1-m$ times, we will have

full rank $A(t_1) + \cdots + A(t_{n+1-m})$.

The selection of $\{t_i\}_{i=1}^k$ will just follow the procedure of proof above based on the selection of $\{\beta_i\}_{i=1}^k$. ■

Remark 13. $n + 1 - m$ is optimal for Corollary 24.2 holding for all positive-semidefinite $n \times n$ matrices $\{A(t)\}_{t\in R}$. That is, Here, we give an example of $A(t)$ so that $A(t_1) + \cdots + A(t_r)$ is not full rank no matter what $t_1, \ldots, t_r$ we use, where $r \leq n - m$ and $m = \max_{t\in R}\{rank(A(t))\}$.

$$A(t) = \begin{pmatrix} f_1(t) & & & \\ & f_2(t) & & \\ & & \ddots & \\ & & & f_n(t) \end{pmatrix},$$

where

$$f_j(t) = \begin{cases} 1 & t = j \text{ or } j < m \\ 0 & t \neq j \text{ and } j \geq m \end{cases}.$$

The proof of it is trivial because it is easy for us to find a row in $A(t_1) + \cdots + A(t_{n-m})$ is all zeros.

Remark 14. For some certain positive-semidefinite $n \times n$matrices $\{A(t)\}_{t\in R}$, to make $A(t_1) + \cdots + A(t_r)$ full rank, r could be smaller than $n + 1 - m$. But, r still has a lower bound for it which is $\lceil \frac{n}{m} \rceil$. The reason for it is that $n = rank(A(t_1) + \cdots + A(t_r)) \leq rank(A(t_1)) + \cdots + rank(A(t_r)) \leq r \times m$. The example to attain equality is trivial and we don't discuss more here.

A.2 Proofs

A.2.1 Proof of Theorem 1

Proof. Similar to the proof of Lemma 1 in Qin and Lawless (1994), proof of (**detail**) in Owen (1990) and proof of lemma 1 (ii) in Yuan et al. (2014), we can get

$$\lambda_n(u, \theta) = \{\frac{1}{n}\sum_{i=1}^{n} g(u, X_i; \theta)g^T(u, X_i; \theta)\}^{-1}\{\frac{1}{n}\sum_{i=1}^{n} g(u, X_i; \theta)\} + o(n^{-\frac{1}{3}}) \tag{A.2}$$

almost surely and uniformly in $||\theta - \theta_0|| \leq n^{-\frac{1}{3}}$ with $u \in S$ due to the fact that each component in $g(u, X_i; \theta)$ is between 0 and 1(satisfying the condition $E||g(u, X_i; \theta)||^3 \leq 1$ in Qin and Lawless (1994),Owen (1990) and Yuan et al. (2014)).

Denote $\theta = \theta_0 + vn^{-\frac{1}{3}}$. Let's give a lower bound for $T_n(\theta)$ on the surface of the ball $||\theta - \theta_0|| = n^{-\frac{1}{3}}$ when

$||v|| = 1$. Scaling our posterior(2.4) by multiplying n^n and performing Taylor expansion, we have

$$\begin{aligned}
n^n p(\theta|X) &= \int_S n^n p_0(\theta) L_n(u;\theta) dG(u) \\
&= \int_S \exp\{-\sum_{i=1}^{n} \log\{1 + \lambda_n^T(u;\theta) g(u, X_i;\theta)\} + \log(p(\theta_0))\} dG(u) \\
&= \int_S \exp\{-\sum_{i=1}^{n} \lambda_n^T(u;\theta) g(u, X_i;\theta) + \frac{1}{2}\sum_{i=1}^{n} \{\lambda_n^T(u;\theta) g(u, X_i;\theta)\}^2 \\
&+ o(n^{\frac{1}{3}})\} dG(u) \\
&= \int_S \exp\{-\frac{n}{2}\{\frac{1}{n}\sum_{i=1}^{n} g(u, X_i;\theta)\}^T \{\frac{1}{n}\sum_{i=1}^{n} g(u, X_i;\theta) g^T(u, X_i;\theta)\}^{-1} \\
&\times \{\frac{1}{n}\sum_{i=1}^{n} g(u, X_i;\theta)\} + o(n^{\frac{1}{3}})\} dG(u) \\
&= \int_S \exp\{-\frac{n}{2}\{\frac{1}{n}\sum_{i=1}^{n} g(u, X_i;\theta_0) + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial g(u, X_i;\theta_0)}{\partial\theta} v n^{-\frac{1}{3}}\}^T \{\frac{1}{n}\sum_{i=1}^{n} g(u, X_i;\theta) g^T(u, X_i;\theta)\}^{-1} \\
&\times \{\frac{1}{n}\sum_{i=1}^{n} g(u, X_i;\theta_0) + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial g(u, X_i;\theta_0)}{\partial\theta} v n^{-\frac{1}{3}}\} + o(n^{\frac{1}{3}})\} dG(u) \\
&= \int_S \exp\{-\frac{n}{2}\{O(n^{-1/2}(\log\log n)^{1/2}) + E(\frac{\partial g(u, X_i;\theta)}{\partial\theta}) v n^{-\frac{1}{3}}\}^T \{E g(u, X_i;\theta_0) g^T(u, X_i;\theta_0)\}^{-1} \\
&\times \{O(n^{-1/2}(\log\log n)^{1/2}) + E(\frac{\partial g(u, X_i;\theta)}{\partial\theta}) v n^{-\frac{1}{3}}\} + o(n^{\frac{1}{3}})\} dG(u).
\end{aligned} \tag{A.3}$$

We rewrite equation (A.3) as $\int_S \exp(-n^{\frac{1}{3}}(f(u) + o(1)))dG(u)$, where $f(u)$ denotes $v^T E(\frac{\partial}{\partial\theta} g(u, x;\theta_0))^T (E g(u, X_i;\theta_0) g^T(u, X_i;\theta_0))^{-1} E(\frac{\partial}{\partial\theta} g(u, x;\theta_0)) v$ and $f(u) \geq 0$.
Now, we prove that there exists large enough $K > 0$ so that $\int_S \exp(-n^{\frac{1}{3}}(f(u)+o(1)))dG(u) \leq \exp(-n^{\frac{1}{4}}(\int_S (f(u) dG(u))))$ holds with $n > K$. To prove it, equivalently, we can show for large enough n,

$$n^{-\frac{1}{4}}(\log \int_S \exp(-n^{\frac{1}{3}} f(u)) dG(u)) \leq -\int_S f(u) dG(u). \tag{A.4}$$

From regularity condition A.4, we know that for any non-zero vector v, $E(f(u))$ is positive. Because $Ef(u) = \int_0^\infty P(f(u) > x)dx$, there exists $m > 0$ so that $P(f(u) > m) > 0$. Then,

$$n^{-\frac{1}{4}}(\log \int_S \exp(-n^{\frac{1}{3}} f(u)) dG(u)) \leq P(f(u) > m) m(-n^{\frac{1}{12}}) + P(f(u) \leq m) n^{-\frac{1}{4}}.$$

Then, it is obvious to see that left side of equation $(A.4)$ will go to $-\infty$ as $n \to \infty$. Thus, we prove the inequality above. Then, we have

$$\begin{aligned}
&\int_S \exp\{-\frac{n}{2}\{O(n^{-1/2}(\log\log n)^{1/2}) + E(\frac{\partial g(u,X_i;\theta)}{\partial\theta})vn^{-\frac{1}{3}}\}^T\{Eg(u,X_i;\theta_0)g^T(u,X_i;\theta_0)\}^{-1} \\
&\times \{O(n^{-1/2}(\log\log n)^{1/2}) + E(\frac{\partial g(u,X_i;\theta)}{\partial\theta})vn^{-\frac{1}{3}}\} + o(n^{\frac{1}{3}})\}dG(u) \\
\leq &\exp(-n^{\frac{1}{4}}(\int_S (f(u)dG(u)))) \\
\leq &\exp(-(c-\epsilon)n^{\frac{1}{4}})
\end{aligned} \tag{A.5}$$

almost surely and uniformly in $||v|| = 1$, where $c > 0$ and c is the smallest eigenvalue of

$$\int_S E(\frac{\partial}{\partial\theta}g(u,x;\theta_0))^T(Eg(u,X_i;\theta_0)g^T(u,X_i;\theta_0))^{-1}E(\frac{\partial}{\partial\theta}g(u,x;\theta_0))dG(u),$$

Similarly,

$$\begin{aligned}
n^nT_n(\theta_0) &= \int_S \exp\{-\frac{n}{2}\{\frac{1}{n}\sum_{i=1}^n g(u,X_i;\theta_0)\{\frac{1}{n}\sum_{i=1}^n g(u,X_i;\theta_0)g^T(u,X_i;\theta_0)\}^{-1} \\
&\quad \times \{\frac{1}{n}\sum_{i=1}^n g(u,X_i;\theta_0)\} + o(1)\}dG(u) \\
&= O((\log n)^{-1}).
\end{aligned} \tag{A.6}$$

Combining (A.5) and (A.6), we find that $T_n(\theta)$ attains its maximum value in the interior of the ball $||\theta-\theta_0|| \leq n^{-\frac{1}{3}}$.

Also, $\hat{\theta}_n$ satisfies $\frac{\partial}{\partial\theta}p(\theta|X) = 0$ which is the second equation in (2.6). The first equation is from (2.2) ■

A.2.2 Proof of Theorem 2

Proof. Similar to the proof of Theorem 2 in Chan et al. (2009), it can be shown by uniform law of large number that $\frac{\partial}{\partial\theta}Q_{1n}(u;\theta_0,0) \xrightarrow{p} s_{12}(u)$, $\frac{\partial}{\partial\lambda^T}Q_{1n}(u;\theta_0,0) \xrightarrow{p} s_{11}(u)$, $\frac{\partial}{\partial\theta}Q_{2n}(u;\theta_0,0) = 0$ and $\frac{\partial}{\partial\lambda^T}Q_{1n}(u;\theta_0,0) \xrightarrow{p} s_{21}(u)$ uniformly in $u \in S$.

Denote $Q_{3n}(u;\theta,\lambda) = \exp(-\sum_{i=1}^n \log(1+\lambda^T g(u,X_i;\theta)))$. Similarly, according to (2.6) in Theorem 1,we have

$$\begin{aligned}
0 =&Q_{1n}(u;\hat{\theta}_n,\lambda_n(u;\hat{\theta}_n)) \\
=&Q_{1n}(u;\theta_0,0) + \frac{\partial Q_{1n}(u;\theta_0,0)}{\partial\theta}(\hat{\theta}_n-\theta_0) + \frac{\partial Q_{1n}(u;\theta_0,0)}{\partial\lambda^T}\lambda_n(u;\hat{\theta}_n) + o_p(\delta_n)
\end{aligned} \tag{A.7}$$

uniformly in $u \in S$, and

$$\begin{aligned}
0 =& n^n \int_S p_0(\hat{\theta}_n) L_n(u;\hat{\theta}_n) Q_{2n}(u;\hat{\theta}_n, \lambda_n(u;\hat{\theta}_n)) - \frac{1}{n}\frac{\partial p_0(\hat{\theta}_n)}{\partial \theta} L_n(u;\hat{\theta}_n) dG(u) \\
=& \int_S \{p_0(\theta_0) Q_{3n}(u;\theta_0,0) Q_{2n}(u;\theta_0,0) + p_0(\theta_0) Q_{3n}(u;\theta_0,0)\frac{\partial Q_{2n}(u;\theta_0,0)}{\partial \theta}(\hat{\theta}_n - \theta_0) + \\
& Q_{2n}(u;\theta_0,0) Q_{3n}(u;\theta_0,0)\frac{\partial p_0(\theta_0)}{\partial \theta}(\hat{\theta}_n - \theta_0) + Q_{2n}(u;\theta_0,0) p_0(\theta_0)\frac{\partial Q_{3n}(u;\theta_0,0)}{\partial \theta}(\hat{\theta}_n - \theta_0) + \\
& p_0(\theta_0) Q_{3n}(u;\theta_0,0)\frac{\partial Q_{2n}(u;\theta_0,0)}{\partial \lambda^T}\lambda_n(u;\hat{\theta}_n) + p_0(\theta_0) Q_{2n}(u;\theta_0,0)\frac{\partial Q_{3n}(u;\theta_0,0)}{\partial \lambda^T}\lambda_n(u;\hat{\theta}_n)\} dG(u) + o_p(\delta_n)
\end{aligned} \tag{A.8}$$

Then, with $Q_{3n}(u;\theta_0,0) = 1$, $Q_{2n}(u;\theta_0,0) = 0$ and $\frac{\partial}{\partial\theta} Q_{2n}(u;\theta_0,0) = 0$,

$$0 = \int_S p_0(\theta_0)\frac{\partial Q_{2n}(u;\theta_0,0)}{\partial \lambda^T}\lambda_n(u;\hat{\theta}_n) dG(u) + o_p(\delta_n).$$

Then, we have

$$0 = \int_S \frac{\partial Q_{2n}(u;\theta_0,0)}{\partial \lambda^T}\lambda_n(u;\hat{\theta}_n) dG(u) + o_p(\delta_n). \tag{A.9}$$

because $p_0(\theta)$ is positive in the neighbor of θ_0 implied by regularity condition A.7.

Following (A.7), we get

$$\lambda_n(u;\hat{\theta}_n) = -s_{11}^{-1}(u) Q_{1n}(u;\theta_0,0) - s_{11}^{-1}(u) s_{12}(u)(\hat{\theta}_n - \theta_0) + o_p(\delta_n). \tag{A.10}$$

Plugging (A.10) into (A.9), we find

$$\hat{\theta}_n - \theta_0 = -\{\int_S s_{21}(u) s_{11}^{-1}(u) s_{12}(u) dG(u)\}^{-1}\{\int_S s_{21}(u) s_{11}^{-1}(u) Q_{1n}(u;\theta_0,0) dG(u)\} + o_p(\delta_n), \tag{A.11}$$

Then, conclusion (2.8) in Theorem 2 is directly from (A.10) and (A.11). ■

A.2.3 Proof of Theorem 3

Proof. For $\theta \in \{||\theta - \theta_0|| = O(n^{-\frac{1}{2}})\}$, by Taylor expansion,

$$\log p(\theta|X) = \log p(\hat{\theta}_n|X) - \frac{1}{2}(\theta - \hat{\theta}_n)^T J(\hat{\theta}_n)(\theta - \hat{\theta}_n) + o_p(||\theta - \hat{\theta}_n||^2).$$

Because $o_p||\theta - \hat{\theta}_n||^2 \le o_p||\theta - \theta_0||^2 + o_p||\theta_0 - \hat{\theta}_n||^2$, we have $o_p||\theta - \hat{\theta}_n||^2 = o_p(n^{-1})$. Then,

$$p(\theta|X) \propto \exp\{-\frac{1}{2}(\theta - \hat{\theta}_n)^T J(\hat{\theta}_n)(\theta - \hat{\theta}_n) + o_p(1)\}. \tag{A.12}$$

Now, suppose $J(\hat{\theta}_n)$ is positive definite and let $D = J(\hat{\theta}_n)^{\frac{1}{2}}(\theta - \hat{\theta}_n)$. Then, from (A.12), we have

$$p(D|X) \propto \exp\{-\frac{1}{2}D^T D + o_p(1)\}.$$

To show the convergence property of cumulated distribution function

$$J(\hat{\theta}_n)^{\frac{1}{2}}(\theta - \hat{\theta}_n) \xrightarrow{d} \mathrm{N}(0, I),$$

we need to prove tails probability goes to zero based on the convergence property of posterior density in (A.12). That is, we need to prove

$$P(||J(\hat{\theta}_n)^{\frac{1}{2}}(\theta - \hat{\theta}_n)|| > \delta) \to 0,$$

when $\delta \to \infty$ and $n \to \infty$.

From (A.12), for any $\theta = \hat{\theta} + J(\hat{\theta}_n)^{-\frac{1}{2}}D$, we have

$$p(\theta|X) \xrightarrow{p} p_0(\theta_0)\exp\{-||D||^2/2\}$$

By the dominate convergence theorem due to the fact that $p(\theta|X) \leq p_0(\theta)$,we have

$$\int_{||D||>\delta} p(\hat{\theta}_n + J(\hat{\theta}_n)^{-\frac{1}{2}}D|X)dD \xrightarrow{p} p_0(\theta_0)\int_{||D||>\delta}\exp\{-||D||^2/2\}dD$$

for any $\delta \geq 0$. Then, we have

$$\begin{aligned} P(||J(\hat{\theta}_n)^{\frac{1}{2}}(\theta - \hat{\theta}_n)|| > \delta|X) &= \frac{\int_{||D||>\delta} p(\hat{\theta}_n + J(\hat{\theta}_n)^{-\frac{1}{2}}D|X)dD}{\int_{||D||>0} p(\hat{\theta}_n + J(\hat{\theta}_n)^{-\frac{1}{2}}D|X)dD} \\ &\to \frac{p_0(\theta_0)\int_{||D||>\delta}\exp\{-||D||^2/2\}dD}{p_0(\theta_0)\int_{||D||>0}\exp\{-||D||^2/2\}dD} \\ &= (2\pi)^{-\frac{p}{2}}\int_{||D||>\delta}\exp\{-||D||^2/2\}dD < \epsilon, \end{aligned}$$

where δ is a number larger than $1-\epsilon$ quantile of standard normal distribution. ■

A.2.4 Proof of Theorem 4

Proof. **Proof of Theorem 4:**

We notice that (2.9) is continuous with respect to θ and it is positive-semidefinite. We can apply Lemma 24.1 by setting $A(t)$ to be (2.9). Then, Theorem 4 is proved. ■

A.2.5 Proof of Theorem 5

Proof. We notice that (2.9) is continuous with respect to θ and it is positive-semidefinite. Due to the discussion in Section 2.3.3, we know that $m = 2$ in our case. We can apply Lemma 24.2 by setting $A(t)$ to be (2.9). Then, Theorem 5 is proved. ■

Appendix B

Appendix of Chapter 3

B.1 Asymptotic properties of maximum likelihood

We list several typical asymptotic properties of maximum likelihood for Lévy processes and Markov processes with proofs under B, E class regularity conditions. Those properties will be used to prove the theorems of previous sections in this paper.

Lemma 24.3. *$X^n \in \mathcal{X}$ is a $n-$dimensional random vector and $Q(\theta; X^n)$ is a real-valued function of θ given X^n. $Q_0(\theta)$ is a real-valued function of θ. Suppose $\theta \in \Theta \subset R^p$ and $\hat{\theta}_n$ is defined as the value of $\theta \in \Theta$ maximizing $Q(\theta; X^n)$. Under such regularity conditions below:*

L24.3.1 Parameter space Θ is compact.

L24.3.2 $Q_0(\theta)$ is uniquely maximized at θ_0.

L24.3.3 $Q_0(\theta)$ is continuous in $\theta \in \Theta$.

L24.3.4 $Q(\theta; X^n)$ converges uniformly in probability to $Q_0(\theta)$. That is, $\sup_{\theta\in\Theta} |Q(\theta; X^n) - Q_0(\theta)| \xrightarrow{p} 0$ as $n \to \infty$.

then, $\hat{\theta}_n \xrightarrow{p} \theta_0$ as $n \to \infty$.

Remark 15. This lemma is the Theorem 2.1 in Newey and McFadden (1994). In fact, the condition L24.3.3 can be generalized to be upper-continuous. The combination of conditions L24.3.1, L24.3.2 and L24.3.3 can be replaced by a more general condition: θ_0 is a well-separated point of the maximum (See Corollary 3.2.3 in Van der Vaart (2000)).

Proof. let $B_\epsilon(\theta_0) = \{\theta : ||\theta - \theta_0|| < \epsilon\}$. Because $\Theta \cap B_\epsilon^C(\theta_0)$ is compact (L24.3.1) and $Q_0(\theta)$ is a continuous function (L24.3.3), there exists $\theta^* \in \Theta \cap B_\epsilon^C(\theta_0)$ to achieve $\sup_{\theta\in\Theta\cap B_\epsilon^C(\theta_0)}\{Q_0(\theta)\}$. Because θ_0 is the unique to maximize $Q_0(\theta)$ (L24.3.2), we denote $Q_0(\theta_0) - Q_0(\theta^*)$ as $\delta > 0$.

Notice that:

$$\sup_{\theta\in(\Theta\cap B_\epsilon^C(\theta_0))}|Q(\theta;X^n)-Q_0(\theta)|<\frac{\delta}{2}\Rightarrow Q(\theta;X)<Q_0(\theta)+\frac{\delta}{2}\leq Q_0(\theta^*)+\frac{\delta}{2}=Q_0(\theta_0)-\frac{\delta}{2}.$$

$$\sup_{\theta\in\Theta\cap B_\epsilon(\theta_0)}|Q(\theta;X^n)-Q_0(\theta)|<\frac{\delta}{2}\Rightarrow Q(\theta_0;X^n)>Q_0(\theta_0)-\frac{\delta}{2}.$$

Then, we have:

$$\sup_{\theta\in\Theta}|Q(\theta;X^n)-Q_0(\theta)|<\frac{\delta}{2}\Rightarrow Q(\theta_0;X^n)>Q(\theta;X^n)\text{ for }\theta\in\Theta\cap B_\epsilon^C(\theta_0)\Rightarrow\theta_n\in\Theta\cap B_\epsilon(\theta_0).$$

Due to the L24.3.4, as $n\to\infty$, we have

$$P(\sup_{\theta\in\Theta}|Q(\theta;X^n)-Q_0(\theta)|<\frac{\delta}{2})\xrightarrow{p}1.$$

Then, we have $P(\theta_n\in\Theta\cap B_\epsilon(\theta_0))\xrightarrow{p}1$ as $n\to\infty$. Equivalently, we have $\hat{\theta}_n\xrightarrow{p}\theta_0$ as $n\to\infty$. ■

Lemma 24.4 (Uniform convergence)**.** *Let $a(X,\theta)$ be a matrix of functions of random vector X and the parameter θ. $||a||$ is the Frobenius norm (also called as $L2,2$ norm) of the matrix a. Under such regularity conditions below:*

L24.4.1 Parameter space Θ is compact.

L24.4.2 $a(x,\theta)$ is continuous at $\theta\in\Theta$ given observation $x\in\mathcal{X}$.

L24.4.3 There exists $d(x)$ with $||a(x,\theta)||\leq d(x)$ for all $\theta\in\Theta$ and $E_{\theta_0}[d(X)]<\infty$.

L24.4.4 Suppose we have random vectors $X_1,\ldots,X_n$ from a specific stochastic process satisfying above conditions separately. They also have ergodic properties: $\frac{1}{n}\sum_{i=1}^n a(X_i,\theta)\xrightarrow{p}E_{\theta_0}[a(X_j,\theta)]$. Moreover, $\Delta(X_i;\delta)=\sup_{||\theta_1-\theta_2||<\delta}||a(X_i,\theta_1)-a(X_i,\theta_2)||$ also has ergodic property that $\frac{1}{n}\sum_{i=1}^n\Delta(X_i,\theta)\xrightarrow{p}E_{\theta_0}[\Delta(X_j,\theta)]$ (This indicates $E_{\theta_0}[a(X_j,\theta)]$ and $E_{\theta_0}[\Delta(X_j,\theta)]$ does not depend on subscript j where $1\leq j\leq n$. Thus, we might write X denoting X_j under E_{θ_0} later in these situations) .

then, $E[a(X,\theta)]$ is continuous and we have uniform convergence law: $\sup_{\theta\in\Theta}||\frac{1}{n}\sum_{i=1}^n a(X_i,\theta)-E_{\theta_0}[a(X,\theta)]||\xrightarrow{p}0$.

Remark 16. This lemma basically is the Lemma 2.4 in Newey and McFadden (1994) to deal with i.i.d data. We extend it to deal with non-i.i.d case which is also suggested in page 2129 in Newey and McFadden (1994).

Proof. **Continuity of** $E[a(X,\theta)]$:

Suppose for every $\theta \in \Theta$, we construct $\theta_k \to \theta$. By L24.4.2, we have $a(x,\theta_k) \to a(x,\theta)$ given x. Since L24.4.3, $E_{\theta_0}[a(X,\theta_k)] \to E_{\theta_0}[a(X,\theta)]$ due to dominated convergence theorem. Thus, $E_{\theta_0}[a(X,\theta)]$ is continuous in $\theta \in \Theta$.

Uniform convergence law:

We need to show $\forall \epsilon, \eta$, there exists $H(\epsilon,\eta)$ so that $P(\sup_{\theta\in\Theta} ||\frac{1}{n}\sum_{i=1}^n a(X_i,\theta) - E_{\theta_0}[a(X,\theta)]|| > \epsilon) < \eta$.

First, let us to decompose $\sup_{\theta\in\Theta} ||\frac{1}{n}\sum_{i=1}^n a(X_i,\theta) - E_{\theta_0}[a(X,\theta)]||$ into three parts. Suppose $\cup_{\theta\in\Theta} B_\delta(\theta)$ is the open cover of parameter space Θ. Since Θ is compact (L24.4.1), there exists finite sub-cover so that $\Theta \subset \cup_{j=1}^J B_\delta(\theta_j)$. Thus, we have this decomposition

$$\sup_{\theta\in\Theta} ||\frac{1}{n}\sum_{i=1}^n a(X_i,\theta) - E_{\theta_0}[a(X,\theta)]|| = \sup_{1\le j\le J}\sup_{\theta\in B_\delta(\theta_j)\cap\Theta} ||\frac{1}{n}\sum_{i=1}^n a(X_i,\theta) - E_{\theta_0}[a(X,\theta)]||$$

$$\le \sup_{1\le j\le J}\sup_{\theta\in B_\delta(\theta_j)\cap\Theta} \Big\{ \underbrace{||\frac{1}{n}\sum_{i=1}^n a(X_i,\theta) - \frac{1}{n}\sum_{i=1}^n a(X_i,\theta_j)||}_{① \ (including \ \sup\sup \ part)} + \underbrace{||\frac{1}{n}\sum_{i=1}^n a(X_i,\theta_j) - E_{\theta_0}[a(X,\theta_j)]||}_{② \ (including \ \sup\sup \ part)}$$

$$+ \underbrace{||E_{\theta_0}[a(X,\theta_j)] - E_{\theta_0}[a(X,\theta)]||}_{③ \ (including \ \sup\sup \ part)} \Big\}.$$

Thus, we have

$$P(\sup_{\theta\in\Theta} ||\frac{1}{n}\sum_{i=1}^n a(X_i,\theta) - E_{\theta_0}[a(X,\theta)]|| > \epsilon) \le P(① > \frac{\epsilon}{3}) + P(② > \frac{\epsilon}{3}) + P(③ > \frac{\epsilon}{3}).$$

Second, let us study ①, ② and ③ separately.

For ①, we utilize $\Delta(X;\delta) = \sup_{||\theta_1-\theta_2||<\delta} ||a(X_i,\theta_1) - a(X,\theta_2)||$. Thus, $① \le \frac{1}{n}\sum_{i=1}^n \Delta(X_i;\delta)$. Thus we have

$$\begin{aligned} P(① > \frac{\epsilon}{3}) &\le P(\sum_{i=1}^n \Delta(X_i;\delta) > \frac{\epsilon}{3}) \\ &= P((\sum_{i=1}^n \Delta(X_i;\delta) - E_{\theta_0}[\Delta(X_1,\delta)]) + E_{\theta_0}[\Delta(X_1,\delta)] > \frac{\epsilon}{3}). \end{aligned}$$

Since $a(x,\delta)$ is continuous in θ (L24.4.2) and $\theta \in \Theta$ which is compact (L24.4.1), $a(x,\delta)$ is uniformly continuous in θ. In addition, $||\Delta(x;\delta)|| \le 2d(x)$ where $E_{\theta_0}[d(X_1)] \le \infty$ (L24.4.3), by dominated convergence theorem, we have $E_{\theta_0}[\Delta(X_1,\delta)] \to 0$ as $\delta \to 0$. Thus, we can let $\delta < \delta_1$ so that $E_{\theta_0}[\Delta(X_1,\delta)] < \frac{\epsilon}{6}$. Then, we have

$$P(① > \frac{\epsilon}{3}) \le P(\sum_{i=1}^n \Delta(X_i;\delta) - E_{\theta_0}[\Delta(X_1,\delta)] > \frac{\epsilon}{6}).$$

Through ergodic property in L24.4.4, we can find $H_1(\epsilon,\eta)$, when $n > H_1(\epsilon,\eta)$, we have

$$P(① > \frac{\epsilon}{3}) \leq P(\sum_{i=1}^{n} \Delta(X_i;\delta) - E_{\theta_0}[\Delta(X_1,\delta)] > \frac{\epsilon}{6}) < \frac{\eta}{2}.$$

For ②, it is obvious that

$$P(② > \frac{\epsilon}{3}) \leq \sum_{j=1}^{J} P(||\frac{1}{n}\sum_{i=1}^{n} a(X_i,\theta_j) - E_{\theta_0}[a(X,\theta_j)]|| > \frac{\epsilon}{3}).$$

Through ergodic property L24.4.4, there exists $H_{2j}(\epsilon,\eta)$ so that ,when $n > H_{2j}(\epsilon,\eta)$, we have

$$P(||\frac{1}{n}\sum_{i=1}^{n} a(X_i,\theta_j) - E_{\theta_0}[a(X,\theta_j)]|| > \frac{\epsilon}{3}) < \frac{\eta}{2J}.$$

Let $H_2(\epsilon,\eta) = \max_{1\leq j\leq J} H_{2j}(\epsilon,\eta)$. Then, when $n > H_2(\epsilon,\eta)$, we have

$$P(② > \frac{\epsilon}{3}) \leq \sum_{j=1}^{J} P(||\frac{1}{n}\sum_{i=1}^{n} a(X_i,\theta_j) - E_{\theta_0}[a(X,\theta_j)]|| > \frac{\epsilon}{3}) < \frac{\eta}{2}.$$

For ③, we have shown that $E[a(X,\theta)]$ is continuous (See the first paragraph of this proof). We also know that Θ is compact (①), we have $E[a(X,\theta)]$ is uniformly continuous. Thus, we can find δ_2 so that, when$\delta < \delta_2$, we have $P(③ > \frac{\epsilon}{3}) = 0$.

Last, by setting $\delta = \min(\delta_1,\delta_2)$ to form the open cover of Θ at the very first step, we can select $H(\eta,\epsilon) = \max(H_1(\epsilon,\eta), H_2(\epsilon,\eta))$ and have

$$P(\sup_{\theta\in\Theta} ||\frac{1}{n}\sum_{i=1}^{n} a(X_i,\theta) - E_{\theta_0}[a(X,\theta)]|| > \epsilon) < \frac{\eta}{2} + \frac{\eta}{2} + 0 = \eta$$

when $n > H(\eta,\epsilon)$. That is, we have the uniform convergence law, which is

$$\sup_{\theta\in\Theta} ||\frac{1}{n}\sum_{i=1}^{n} a(X_i,\theta) - E_{\theta_0}[a(X,\theta)]|| \xrightarrow{p} 0.$$

■

Lemma 24.5 (Interchange of differentiation and integration)**.** *$a(x,\theta)$ is a real-valued function of an observation $x \in \mathcal{X}$ and the parameter θ. If $a(x,\theta)$ is continuously differentiable at $\theta \in \mathcal{N}$ for all $x \in \mathcal{X}$ where $\mathcal{N}$ is an open set and $\int_{\mathcal{X}} (\sup_{\theta\in\mathcal{N}} ||\frac{\partial}{\partial\theta} a(x,\theta)||) dx < \infty$, then, $\int_{\mathcal{X}} a(x,\theta)dx$ is continuously differentiable and $\frac{\partial}{\partial\theta}\int_{\mathcal{X}} a(x,\theta)dx = \int_{\mathcal{X}} (\frac{\partial}{\partial\theta} a(x,\theta))dx$.*

Proof. For any $\theta \in \mathcal{N}$, we can choose a sequence $\theta_n \in \mathcal{N}$ so that $\theta_n \to \theta$ because $\mathcal{N}$ is an open set. Since $a(x,\theta)$ is continuously differentiable at $\theta \in \mathcal{N}$ for all $x \in \mathcal{X}$, we have $\frac{\partial}{\partial \theta_n} a(x,\theta) \to \frac{\partial}{\partial \theta_n} a(x,\theta)$. In addition, we know that $\int_{\mathcal{X}} (\sup_{\theta \in \mathcal{N}} ||\frac{\partial}{\partial \theta} a(x,\theta)||) dx < \infty$, Thus, $\int_{\mathcal{X}} (\frac{\partial}{\partial \theta} a(x,\theta_n)) dx \to \int_{\mathcal{X}} (\frac{\partial}{\partial \theta} a(x,\theta)) dx$ due to dominated convergence theorem. Thus, $\int_{\mathcal{X}} (\frac{\partial}{\partial \theta} a(x,\theta)) dx$ is continuous. By mean-value expansion theorem, we have $a(x,\theta_n) = a(x,\theta) + \frac{\partial}{\partial \theta} a(x,\theta)'(\theta_n - \theta) + r(x,\theta_n)$ where $r(x,\theta_n) = [\frac{\partial}{\partial \theta} a(x,\theta_n^*) - \frac{\partial}{\partial \theta} a(x,\theta)]'(\theta_n - \theta)$ and θ_n^* is the mean value located in the line between θ_n and θ. Because $\mathcal{N}$ is an open set, for all θ_n closed enough to θ, we can assume that $\theta_n^* \in \mathcal{N}$ without losing generality. Then, we have $||r(x,\theta_n)||/||\theta_n - \theta|| \leq ||\frac{\partial}{\partial \theta} a(x,\theta_n^*) - \frac{\partial}{\partial \theta} a(x,\theta)|| \to 0$ by the continuity of $\frac{\partial}{\partial \theta} a(x,\theta)$. In addition, $||r(x,\theta_n)||/||\theta_n - \theta|| \leq 2 \sup_{\theta \in \mathcal{N}} ||\frac{\partial}{\partial \theta} a(x,\theta)||$, we have $||r(x,\theta_n)||/||\theta_n - \theta|| \to 0$ by dominated convergence theorem. Therefore, $|\int_{\mathcal{X}} a(x,\theta_n) - \int_{\mathcal{X}} a(x,\theta) - [\int_{\mathcal{X}} \frac{\partial}{\partial \theta} a(x,\theta) dx]'(\theta_n - \theta) = |\int_{\mathcal{X}} r(x,\theta_n) dx| \leq \int_{\mathcal{X}} |r(x,\theta_n)| dx = o(||\theta_n - \theta||)$. Equivalently, we have $\frac{\partial}{\partial \theta} \int_{\mathcal{X}} a(x,\theta) dx = \int_{\mathcal{X}} (\frac{\partial}{\partial \theta} a(x,\theta)) dx$. ■

Lemma 24.6 (Unique maximum for I.I.D)**.** *If $X_1, \ldots, X_n$ are I.I.D. The true parameter θ_0 is identified. That is, $\theta \neq \theta_0 \Leftrightarrow f(x;\theta) \neq f(x;\theta_0)$ where $f(x|\theta)$ is the density function of i.i.d data. In addition, if we have $E_{\theta_0} |\log f(X;\theta)| < \infty$ for all $\theta \in \Theta$ where Θ is the parameter space. Then, $E_{\theta_0}[\log f(X;\theta)]$ is uniquely maximized at θ_0.*

Proof. Take $g(y) = -\log(y)$ and we have $E[g(Y)] \geq g(EY)$ due to Jensen's inequality. Let $Y = \frac{f(X;\theta)}{f(X;\theta_0)}$. Because the true parameter θ_0 is identified, Y is not a constant when $\theta \neq \theta_0$. Thus, we have $E[g(Y)] > g(EY)$ which is

$$E_{\theta_0}[-\log(\frac{f(X;\theta)}{f(X;\theta_0)})] > -\log(E_{\theta_0}[\frac{f(X;\theta)}{f(X;\theta_0)}]).$$

Notice that

$$E_{\theta_0}[\frac{f(X;\theta)}{f(X;\theta_0)}] = \int_{\mathcal{X}} \frac{f(x;\theta)}{f(x;\theta_0)} f(x;\theta_0) dx = \int_{\mathcal{X}} f(x;\theta) = 1$$

Therefore, we have $E_{\theta_0}[\log f(X;\theta_0)] > E_{\theta_0}[\log f(X;\theta)]$ for any $\theta \neq \theta_0$. ■

B.1.1 B regularity conditions in Chapter 3

We list the following B class of regularity conditions which will be used for consistency and asymptotic normality of MLE for I.I.D case.

B.1 $\theta \neq \theta_0 \Leftrightarrow L(\theta;\cdot) \neq L(\theta_0;\cdot)$

B.2 The parameter space Θ is compact.

B.3 $L(\theta;x)$ is a continuous function of θ for almost all $x \in \mathcal{X} \subset R$. $\mathcal{X}$ is the support of $f(x;\theta)$.

B.4 $||l(\theta;x)|| \leq g_1(x)$ for all $\theta \in \Theta$ and $E_{\theta_0}[g_1(X)] < \infty$.

B.5 θ_0 lies in the interior of compact parameter space Θ.

B.6 $L(x;\theta)$ is twice continuously differentiable and $L(x;\theta) > 0$ in a neighborhood of θ_0, Ω_0.

B.7 $||\frac{\partial L(\theta;x)}{\partial \theta}|| \leq g_2(x)$ for all $\theta \in \Omega_0$ and $\int_{\mathcal{X}} g_2(x)dx < \infty$.

B.8 $I(\theta_0) = E_{\theta_0}[\frac{\partial \log L(\theta;x)}{\partial \theta} \frac{\partial \log L(\theta;x)}{\partial \theta}^T]$ is positive definite.

B.9 $||\frac{\partial^2 l(\theta;x)}{\partial \theta \partial \theta^T}|| \leq g_3(x)$ for all $\theta \in \Omega_0$ and $E_{\theta_0}[g_3(X)] < \infty$.

B.10 $||\frac{\partial^2 L(\theta;x)}{\partial \theta \partial \theta^T}|| \leq g_4(x)$ for all $\theta \in \Omega_0$ and $\int_{\mathcal{X}} g_4(x)dx < \infty$.

Lemma 24.7 (Consistency and normality for I.I.D)**.** *Under regularity conditions B.1 $\sim$ B.4, the MLE $\hat{\theta}_N$ is a consistent estimator. That is,*

$$\hat{\theta}_N \xrightarrow{p} \theta_0 \tag{B.1}$$

when $N \to \infty$. When all B class of regularity conditions hold, then

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

Proof. Consistency:

To prove the consistency of MLE, we mainly use Lemma 24.3 by setting $Q(\theta; X^N) = \frac{1}{N}\sum_{j=1}^{N} l(\theta; X_j)$. Now we prove that regularity conditions B.1,B.2,B.3 and B.4 are equivalent to regularity conditions L24.3.1, L24.3.2, L24.3.3 and L24.3.4 provided by Lemma 24.3.

First, L24.3.1 is the same as B.2.

Second, to show L24.3.2, we use Lemma 24.6. In fact, regularity B.2 and B.4 guarantee the availability of Lemma 24.6 so that $Q_0(\theta)$ is uniquely maximized at true parameter θ_0.

Last, to show L24.3.3 and L24.3.4 together, we use Lemma 24.4 by setting $a(X_j, \theta) = l(\theta; X_j)$ for $j = 1, \ldots, N$. That is, we need to verify L24.4.1, L24.4.2, L24.4.3 and L24.4.4. In fact, L24.4.1 is guaranteed by B.2. L24.4.2 is guaranteed by B.3. L24.4.3 is guaranteed by B.4. L24.4.4 is guaranteed by B.3 and B.4 together via the Weak law of large numbers for I.I.D data. Thus, we have the consistency property, that is, $\hat{\theta}_N \xrightarrow{p} \theta_0$.

Asymptotic normality:

Let $\mathcal{N} \subset \Theta$ is an open set containing true parameter θ_0. Due to the consistency of MLE $\hat{\theta}_N$ and regularity condition B.5, we have $1_{\{\hat{\theta}_N \in \mathcal{N}\}} \xrightarrow{p} 1$. We denote $1_{\{\hat{\theta}_N \in \mathcal{N}\}}$ as $\hat{1}$. When MLE $\hat{\theta}_N$ is inside $\mathcal{N}$, we

have $\hat{1} \cdot \frac{\partial}{\partial \theta} Q(\hat{\theta}_N; X^N) = 0$ due to the first order condition. By the mean-value expansion theorem (regularity condition B.6):

$$\hat{1}_N \cdot [\frac{1}{\sqrt{N}} \sum_{j=1}^{N} \varphi(X_j; \theta_0) + \{-J_N^*\} \sqrt{N}(\hat{\theta}_N - \theta_0)] = 0 \tag{B.2}$$

where $\varphi(X_j; \theta)$ is defined as $\frac{\partial}{\partial \theta} l(\theta; X_j)$. J_N^* is a $p \times p$ random matrix where jth row of the matrix is the jth row of J_N evaluated at the mean value θ_{jN}^* between $\hat{\theta}_N$ and θ_0 and

$$J_N(\theta) = [-\frac{1}{N} \sum_{j=1}^{N} \frac{\partial^2 \log L(\theta; X_j)}{\partial \theta \partial \theta'}]$$

Suppose we have: (we will prove them later)

①: $\frac{1}{\sqrt{N}} \sum_{j=1}^{N} \varphi(X_j; \theta_0) \xrightarrow{D} N(0, I(\theta_0))$ as $N \to \infty$.

②: $J_N^* \xrightarrow{p} I(\theta_0)$ as $N \to \infty$.

Then, we denote $1_{\{\hat{\theta}_N \in \mathcal{N} \cap J_N^* \text{is non-singular}\}}$ as $\tilde{1}$. Then, by Equation (B.2), we have

$$\tilde{1} \cdot \sqrt{N}(\hat{\theta}_N - \theta_0) = \tilde{1} \cdot (J_N^*)^{-1} \sum_{j=1}^{N} \varphi(X_j; \theta_0)$$

Due to regularity condition B.8, we know $\tilde{1} \xrightarrow{p} 1$. Combined with ①, ② and Slutsky's theorem, we have

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

Proof of ①: We only need to show $E_{\theta_0} \varphi(X_j; \theta_0) = 0$, then, through central limit theorem for I.I.D data, we have $\frac{1}{\sqrt{N}} \sum_{j=1}^{N} \varphi(X_j; \theta_0) \xrightarrow{D} N(0, I(\theta_0))$ as $N \to \infty$. To show $E_{\theta_0} \varphi(X_j; \theta_0) = 0$, we utilize Lemma 24.5. Set $a(x, \theta)$ in Lemma 24.5 to be $L(\theta; x)$. The regularity conditions in Lemma 24.5 can be guaranteed by B.6 and B.7. Then, we have

$$E[\varphi(x; \theta)] = \int_{\mathcal{X}} \frac{\partial f(x; \theta)/\partial \theta}{f(x; \theta)} f(x; \theta) d\mu(x) = \frac{\partial}{\partial \theta} \int_{\mathcal{X}} f(x; \theta) dx = 0$$

Proof of ②: Similar to the proof of ①, we have $E_{\theta_0}[J_n(\theta_0)] = I(\theta_0)$ by utilizing Lemma 24.5 based on regularity conditions B.6, B.7 and B.10. If we denote $E_{\theta_0}[-\frac{\partial^2}{\partial \theta \partial \theta'} \log f(X; \theta)]$ as $J(\theta)$, we have $J_n(\theta) \xrightarrow{p} J(\theta)$ due to the weak law of large number and B.10. We utilize Lemma 24.4 by setting $a(X, \theta)$ to be the random matrix $J_n(\theta)$. Based on the regularity condition B2, B.6, B.9, B.10 and week law of large number for I.I.D data, we have that $J(\theta)$ is continuous at $\theta \in \mathcal{N}$ and uniform continuity $\sup_{\theta \in \mathcal{N}} ||J_n(\theta) - J(\theta)|| \xrightarrow{p} 0$. To

prove ②, we use triangle inequality

$$||J_n^* - I(\theta_0)|| \leq ||J_n^* - J(\theta^*)|| + ||J(\theta^*) - I(\theta_0)||$$

$||J_n^* - J(\theta^*)|| \xrightarrow{p} 0$ is due to the uniform convergence proved above and $||J(\theta^*) - I(\theta_0)|| \xrightarrow{p} 0$ is since θ^* is consistent defined in (B.2). ■

Lemma 24.8 (Unique maximum for Markovian data)**.** *If $X_1, \ldots, X_n$ are following Markov processes. $f(\cdot|\cdot;\theta)$ is the stationary transition density. The true parameter θ_0 is identified. That is, $\theta \neq \theta_0 \Leftrightarrow f(\cdot|\cdot;\theta) \neq f(\cdot|\cdot;\theta_0)$. In addition, if we have $E_{\theta_0}|\log f(X_1|X_0;\theta)| < \infty$ for all $\theta \in \Theta$ where Θ is the parameter space. Then, $E_{\theta_0}[\log f(X_1|X_0;\theta)]$ is uniquely maximized at θ_0.*

Remark 17. Through our definition of $E_{\theta_0}[\cdot]$, $E_{\theta_0}[\log f(X_j|X_{j-1};\theta)]$ will be not related to j. Without losing generality, we normally write it as $E_{\theta_0}[\log f(X_1|X_0;\theta)]$.

Proof. Take $g(y) = -\log(y)$ and we have $E[g(Y)] \geq g(EY)$ due to Jensen's inequality. Let $Y = \frac{f(X_1|X_0;\theta)}{f(X_1|X_0;\theta_0)}$. Because the true parameter θ_0 is identified, Y is not a constant when $\theta \neq \theta_0$. Thus, we have $E[g(Y)] > g(EY)$ which is

$$E_{\theta_0}[-\log(\frac{f(X_1|X_0;\theta)}{f(X_1|X_0;\theta_0)})] > -\log(E_{\theta_0}[\frac{f(X_1|X_0;\theta)}{f(X_1|X_0;\theta_0)}]).$$

Notice that

$$E_{\theta_0}[\frac{f(X_1|X_0;\theta)}{f(X_1|X_0;\theta_0)}] = \int_{\mathcal{X}\times\mathcal{X}} \frac{f(x_1|x_0;\theta)}{f(x_1|x_2;\theta_0)} f(x_1|x_0;\theta_0)p(x_0)dx_0dx_1 = \int_{\mathcal{X}\times\mathcal{X}} f(x_1|x_0;\theta)p(x_0)dx_0dx_1 = 1$$

where $p(x_0)$ is the stationary distribution of the Markov process. Therefore, we have $E_{\theta_0}[\log f(X_1|X_0;\theta_0)] > E_{\theta_0}[\log f(X_1|X_0;\theta)]$ for any $\theta \neq \theta_0$. ■

B.1.2 E regularity conditions in Chapter 3

We list the following E class of regularity conditions which will be used for consistency and asymptotic normality of MLE for Markovian case.

E.1 $\theta \neq \theta_0 \Leftrightarrow f(\cdot|\cdot;\theta) \neq f(\cdot|\cdot;\theta_0)$. $f(\cdot|\cdot;\theta)$ is transition density.

E.2 The parameter space Θ is compact.

E.3 $l(\theta; x_0, x_1)$ is a continuous function of θ for all $x_0 \in \mathcal{X} \subset R$ and $x_1 \in \mathcal{X} \subset R$.

E.4 $||l(\theta; x_0, x_1)|| \leq g_1(x_0, x_1)$ for all $\theta \in \Theta$ and $E_{\theta_0}[g_1(X_0, X_1)] < \infty$. $E_{\theta_0}[\cdot]$ is defined in D.3.

E.5 θ_0 lies in the interior of compact parameter space Θ.

E.6 $L(\theta; x_0, x_1)$ is twice continuously differentiable and $L(\theta; x_0, x_1) > 0$ in a neighborhood of θ_0, Ω_0 for all $x_0 \in \mathcal{X} \subset R$ and $x_1 \in \mathcal{X} \subset R$.

E.7 $||\frac{\partial L(\theta;x_0,x_1)}{\partial \theta}|| \leq g_2(x_0, x_1)$ for all $\theta \in \Omega_0$ and $\int_{\mathcal{X}\times\mathcal{X}} g_2(x_0, x_1)\mu(dx_0 \times dx_1) < \infty$.

E.8 $I(\theta_0) = E_{\theta_0}[\frac{\partial \log L(\theta;x_0,x_1)}{\partial \theta} \frac{\partial \log L(\theta;x_0,x_1)}{\partial \theta}^T]$ is positive definite. $E_{\theta_0}[\cdot]$ is defined in D.3.

E.9 $||\frac{\partial^2 \log L(\theta;x_0,x_1)}{\partial \theta \partial \theta^T}|| \leq g_3(x_0, x_1)$ for all $\theta \in \Omega_0$ and $E_{\theta_0}[g_3(X_0, X_1)] < \infty$. $E_{\theta_0}[\cdot]$ is defined in D.3.

E.10 $||\frac{\partial^2 L(\theta;x_0,x_1)}{\partial \theta \partial \theta^T}|| \leq g_4(x_0, x_1)$ for all $\theta \in \Omega_0$ and $\int_{\mathcal{X}\times\mathcal{X}} g_4(x_0, x_1)\mu(dx \times dx) < \infty$.

Lemma 24.9 (Consistency and normality for I.I.D)**.** *Under regularity conditions D.1, D.2 and E.1 $\sim$ E.4, the MLE $\hat{\theta}_N$ is a consistent estimator. That is,*

$$\hat{\theta}_N \xrightarrow{p} \theta_0 \tag{B.3}$$

when $N \to \infty$. When all D and E class of regularity conditions hold, then

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

Proof. Consistency:

To prove the consistency of MLE, we mainly use Lemma 24.3 by setting $Q(\theta; X^N) = \frac{1}{N}\sum_{j=1}^{N} l(\theta; X_j, X_{j-1})$. Now we prove that regularity conditions E.1,E.2,E.3 and E.4 are equivalent to regularity conditions L24.3.1, L24.3.2, L24.3.3 and L24.3.4 provided by Lemma 24.3.

First, L24.3.1 is the same as E.2.

Second, to show L24.3.2, we use Lemma 24.8. In fact, regularity E.2 and E.4 guarantee the availability of Lemma 24.8 so that $Q_0(\theta)$ is uniquely maximized at true parameter θ_0.

Last, to show L24.3.3 and L24.3.4 together, we use Lemma 24.4 by setting $a(Y_j, \theta) = l(\theta; X_j, X_{j-1})$ for $j = 1, \ldots, N$ (We use Y_j to represent the X_j in Lemma 24.4). That is, we need to verify L24.4.1, L24.4.2, L24.4.3 and L24.4.4. In fact, L24.4.1 is guaranteed by E.2. L24.4.2 is guaranteed by E.3. L24.4.3 is guaranteed by E.4. L24.4.4 is guaranteed by E.3 and E.4 and the Weak law of large numbers for Markovian data (D.1 and D.2). Thus, we have the consistency property, that is, $\hat{\theta}_N \xrightarrow{p} \theta_0$.

Asymptotic normality:

Let $\mathcal{N} \subset \Theta$ is an open set containing true parameter θ_0. Due to the consistency of MLE $\hat{\theta}_N$ and regularity condition E.5, we have $1_{\{\hat{\theta}_N \in \mathcal{N}\}} \xrightarrow{p} 1$. We denote $1_{\{\hat{\theta}_N \in \mathcal{N}\}}$ as $\hat{1}$. When MLE $\hat{\theta}_N$ is inside $\mathcal{N}$, we have $\hat{1} \cdot \frac{\partial}{\partial \theta} Q(\hat{\theta}_N; X^N) = 0$ due to the first order condition. By the mean-value expansion theorem (regularity condition E.6):

$$\hat{1}_N \cdot [\frac{1}{\sqrt{N}} \sum_{j=1}^{N} \varphi(X_j, X_{j-1}; \theta_0) + \{-J_N^*\}\sqrt{N}(\hat{\theta}_N - \theta_0)] = 0 \tag{B.4}$$

where $\varphi(X_j, X_{j-1}; \theta)$ is defined as $\frac{\partial}{\partial \theta} l(\theta; X_j, X_{j-1})$. J_N^* is a $p \times p$ random matrix where jth row of the matrix is the jth row of J_N evaluated at the mean value θ_{jN}^* between $\hat{\theta}_N$ and θ_0 and

$$J_N(\theta) = [-\frac{1}{N} \sum_{j=1}^{N} \frac{\partial^2 \log L(\theta; X_j, X_{j-1})}{\partial \theta \partial \theta'}]$$

Suppose we have: (we will prove them later)

①: $\frac{1}{\sqrt{N}} \sum_{j=1}^{N} \varphi(X_j, X_{j-1}; \theta_0) \xrightarrow{D} N(0, I(\theta_0))$ as $N \to \infty$.

②: $J_N^* \xrightarrow{p} I(\theta_0)$ as $N \to \infty$.

Then, we denote $1_{\{\hat{\theta}_N \in \mathcal{N} \cap J_N^* \text{is non-singular}\}}$ as $\tilde{1}$. Then, by Equation (B.4), we have

$$\tilde{1} \cdot \sqrt{N}(\hat{\theta}_N - \theta_0) = \tilde{1} \cdot (J_N^*)^{-1} \sum_{j=1}^{N} \varphi(X_j, X_{j-1}; \theta_0)$$

Due to regularity condition E.8, we know $\tilde{1} \xrightarrow{p} 1$. Combined with ①, ② and Slutsky's theorem, we have

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

Proof of ①: This is guaranteed by Lemma 10.1 under regularity conditions of D class.

Proof of ②: We have $E_{\theta_0} \varphi(X_j, X_{j-1}; \theta_0) = 0$ and then $E_{\theta_0}[J_n(\theta_0)] = I(\theta_0)$ by utilizing Lemma 24.5 based on regularity conditions E.6, E.7 and E.10. If we denote $E_{\theta_0}[-\frac{\partial^2}{\partial \theta \partial \theta'} \log f(X_1 | X_0; \theta)]$ as $J(\theta)$, we have $J_n(\theta) \xrightarrow{p} J(\theta)$ due to the Lemma 10.1 (D.1 and D.2) and E.10. We utilize Lemma 24.4 by setting $a(X, \theta)$ to be the random matrix $J_n(\theta)$. Based on the regularity condition E.2, E.6, E.9, E.10 and Lemma 10.1 (D.1 and D.2), we have that $J(\theta)$ is continuous at $\theta \in \mathcal{N}$ and uniform continuity $\sup_{\theta \in \mathcal{N}} ||J_n(\theta) - J(\theta)|| \xrightarrow{p} 0$. To prove ②, we use triangle inequality

$$||J_n^* - I(\theta_0)|| \leq ||J_n^* - J(\theta^*)|| + ||J(\theta^*) - I(\theta_0)||$$

where θ^* is mean value defined in (B.4). $||J_n^* - J(\theta^*)|| \xrightarrow{p} 0$ is due to the uniform convergence proved above

and $||J(\theta^*) - I(\theta_0)|| \xrightarrow{p} 0$ is since θ^* is consistent defined in (B.4). ■

Under the framework of Section 3.2.2, we have the hypothesis:

$$H_0 : g_1(\theta) = 0, \ldots, g_q(\theta) = 0, \qquad 1 \leq q < p.$$

against

$$H_a : \exists k, \quad g_k(\theta) \neq 0, \qquad 1 \leq k \leq q.$$

Lemma 24.10. *Let $\{X\}$ be i.i.d data. Assume that we can treat H_0 as $\theta_1 = \theta_1^0, \cdot, \theta_r = \theta_r^0$ via reparametrization where θ_j^0 is the jth true parameter. When H_0 holds, B classes of regularity conditions applied to both parameter space Θ and its subspace Θ_0 of parameter vector $(\theta_{r+1}, \ldots, \theta_p)$, we have*

$$-2 \log \Lambda \xrightarrow{d} \chi_k^2$$

where

$$\Lambda = \frac{L_0(\hat{\hat{\theta}}_N; X)}{L_1(\hat{\theta}_N; X)},$$

$\hat{\hat{\theta}}_N \in \Theta_0$ is the MLE of the likelihood $L_0(\theta; X)$ under H_0. $\hat{\theta}_N \in \Theta_1$ is the MLE of the likelihood $L_1(\theta; X)$ in parameter space Θ. The dimension of parameter space is p for Θ and $(p-q)$ for Θ_0. χ_{k-p}^2 is a Chi-squared distribution with degree freedom k.

Proof. If we denote $\{\theta_1^0, \ldots, \theta_r^0, \hat{\hat{\theta}}_N\}$ as $\tilde{\theta}$, we have $L_1(\tilde{\theta}_N; X) = L_0(\hat{\hat{\theta}}_N; X)$. Suppose log-likelihood is $l_1(\theta; X) = \log L_1(\theta; X)$ ($l_0(\theta; X) = \log L_0(\theta; X)$). Noticing that:

$$l_1(\theta; X) = \sum_{j=1}^{N} l_1(\theta; X_j),$$

we first perform mean-value expansion theorem (regularity condition B.6) to both $l_0(\hat{\hat{\theta}}_N; X)$ and $l_1(\hat{\theta}_N; X)$:

$$l_0(\hat{\hat{\theta}}_N; X) = \sum_{j=1}^{N} l_1(\tilde{\theta}_N; X_j) = \sum_{j=1}^{N} l_1(\theta_0; X_j) + \sum_{j=1}^{N} \varphi(\theta_0; X_j)'(\tilde{\theta}_N - \theta_0) + \frac{N}{2}(\tilde{\theta}_N - \theta_0)'(-J_{1N}^*)(\tilde{\theta}_N - \theta_0)$$

$$l_1(\hat{\theta}_N; X) = \sum_{j=1}^{N} l_1(\hat{\theta}_N; X_j) = \sum_{j=1}^{N} l_1(\theta_0; X_j) + \sum_{j=1}^{N} \varphi(\theta_0; X_j)'(\hat{\theta}_N - \theta_0) + \frac{N}{2}(\hat{\theta}_N - \theta_0)'(-J_{2N}^*)(\hat{\theta}_N - \theta_0)$$

where $\varphi(\theta; X_j) = \frac{\partial}{\partial \theta} l_1(\theta; X_j)$ and J_{1N}^*, J_{2N}^* are $p \times p$ random matrix where jth row of the matrix is the jth

row of J_N evaluated at the mean value between $\hat{\theta}_N$ and θ_0 and

$$J_N(\theta) = [-\frac{1}{N}\sum_{j=1}^{N}\frac{\partial^2 \log L_1(\theta; X_j)}{\partial\theta\partial\theta'}].$$

Then, we apply Equation (B.2) and ② in Lemma 24.7 to φ, J_{1N}^* and J_{2N}^* based on B class of regularity conditions combined with Slutsky's theorem, we have

$$l_0(\hat{\tilde{\theta}}_N; X) = \sum_{j=1}^{N} l_1(\tilde{\theta}_N; X_j) = \sum_{j=1}^{N} l_1(\theta_0; X_j) + \frac{1}{2}\sqrt{N}(\tilde{\theta}_N - \theta_0)'(I(\theta_0))\sqrt{N}(\tilde{\theta}_N - \theta_0) + o_p(1)$$

$$l_1(\hat{\theta}_N; X) = \sum_{j=1}^{N} l_1(\hat{\theta}_N; X_j) = \sum_{j=1}^{N} l_1(\theta_0; X_j) + \frac{1}{2}\sqrt{N}(\hat{\theta}_N - \theta_0)'(I(\theta_0))\sqrt{N}(\hat{\theta}_N - \theta_0) + o_p(1)$$

Then, we have

$$-2\log\Lambda = \frac{1}{2}\sqrt{N}(\hat{\theta}_N - \tilde{\theta})'(I(\theta_0))\sqrt{N}(\hat{\theta}_N - \tilde{\theta}) + o_p(1)$$

We decompose $\hat{\theta}_N = \{\hat{\theta}_N^1, \hat{\theta}_N^2\}$ and $\tilde{\theta}_N = \{\tilde{\theta}_N^1, \tilde{\theta}_N^2\}$ in the following way:

$$\hat{\theta}_N^1 = \{\hat{\theta}_1, \ldots, \hat{\theta}_k\} \quad \hat{\theta}_N^2 = \{\hat{\theta}_{k+1}, \ldots, \hat{\theta}_p\} \quad \tilde{\theta}_N^1 = \{\theta_1^0, \ldots, \theta_k^0\} \quad \tilde{\theta}_N^2 = \{\hat{\tilde{\theta}}_{k+1}, \ldots, \hat{\tilde{\theta}}_p\}$$

Thus, we know that

$$\sqrt{N}(\hat{\theta}_N^1 - \tilde{\theta}_N^1) \xrightarrow{d} N(0, I_{1:k,1:k}^{-1}(\theta_0))$$

$$\sqrt{N}(\hat{\theta}_N^2 - \tilde{\theta}_N^2) \xrightarrow{d} 0$$

via asymptotic normality indicated by Lemma 24.7. Then, obviously we have $-2\log\Lambda \xrightarrow{d} \chi_k^2$ which is a Chi-squared distribution with degree freedom k. ■

B.2 Proofs

B.2.1 Proof of Theorem 6

For each $\theta \in \Theta$, $E_{h,M}^F(\phi, a)(x)$ will decay to zero (see Therorem 2.3 and Corollary 2.4 in Feng and Lin (2013)). Because the bound of $|E_{h,M}^F(\phi, a)(x)|$ in Equation (3.4) doesn't depend on parameter θ based on the A class regularity conditions, $|E_{h,M}^F(\phi, a)(x)|$ decay to zero uniformly on $\theta \in \Theta$ as $Mh \to \infty$ and $h \to 0$.

B.2.2 Proof of Theorem 7

Proof. We utilize Lemma 24.3 by setting $Q(\theta, X, M, h, a) = l^{M,h,a}(\theta; X)$ and $Q_0(\theta, X) = l(\theta; X)$. We know that $Q(\theta, X, M, h, a)$ converges to $Q_0(\theta, X)$ uniformly for $\theta \in \Theta$ when $Mh \to \infty$ and $h \to 0$ via Theorem 6. Then, Lemma 24.3 holds and $\hat{\theta}_N^{M,h,a} \xrightarrow{p} \hat{\theta}_N$. ■

B.2.3 Proof of Theorem 8

Proof. Under B class regularity conditions, we have

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

via Lemma 24.7. Combined with Theorem 7, we have $\hat{\theta}_N^{M,h,a}$, $\hat{\theta}_N^{M(N),h(N),a} \xrightarrow{p} \theta_0$ with fixed a when $N \to \infty$, and

$$\sqrt{N}(\hat{\theta}_N^{M(N),h(N),a} - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0))$$

via Slutsky's theorem. ■

B.2.4 Proof of Theorem 9

Proof. Under B class regularity conditions, we have

$$-2 \log \Lambda^* \xrightarrow{d} \chi_k^2$$

where

$$\Lambda^* = \frac{L_0(\hat{\hat{\theta}}_N; X)}{L_1(\hat{\theta}_N; X)},$$

via Lemma 24.10. Combined with Theorem 7, we have $\Lambda \xrightarrow{p} \Lambda^*$ as $Mh \to \infty$ and $h \to 0$. Then,

$$-2 \log \Lambda \xrightarrow{d} \chi_k^2$$

via Slutsky's theorem. ■

B.2.5 Proof of Theorem 10

For each $\theta \in \Theta$, $E_{h,M}^F(\phi, a)(x)$ will decay to zero (see Therorem 2.3 and Corollary 2.4 in Feng and Lin (2013)). Because the bound of $|E_{h,M}^F(\phi, a)(x)|$ in (3.4) doesn't depend on parameter θ based on the C class

regularity conditions, then, $|E^F_{h,M}(\phi, a)(x)|$ decay to zero uniformly on $\theta \in \Theta$ as $Mh \to \infty$ and $h \to 0$.

B.2.6 Proof of Theorem 11

Proof. We utilize Lemma 24.3 by setting $Q(\theta, X, M, h, a) = l^{M,h,a}(\theta; X)$ and $Q_0(\theta, X) = l(\theta; X)$. We know that $Q(\theta, X, M, h, a)$ converges to $Q_0(\theta, X)$ uniformly for $\theta \in \Theta$ when $Mh \to \infty$ and $h \to 0$ via Theorem 10. Then, Lemma 24.3 holds combined with regularity condition E.2 and E.3, and $\hat{\theta}^{M,h,a}_N \xrightarrow{p} \hat{\theta}_N$. ■

B.2.7 Proof of Theorem 12

Proof. Under E class regularity conditions, we have

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

via Lemma 24.9. Combined with Theorem 11, we have $\hat{\theta}^{M,h,a}_N$, $\hat{\theta}^{M(N),h(N),a}_N \xrightarrow{p} \theta_0$ with fixed a when $N \to \infty$, and

$$\sqrt{N}(\hat{\theta}^{M(N),h(N),a}_N - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0))$$

via Slutsky's theorem. ■

B.2.8 Proof of Theorem 13

Proof. If we follow the selection of analytic strip $\mathcal{D}_{[d_-,d_+]}$ for each model in Section 3.3, A.1 is satisfied because their characteristic function $\phi(u; \theta)$ is analytic in the analytic strip. The proof of analyticity for each model is trivial.

For A.2 and A.4, the model's characteristic function forms are provided in the Section 3.3. And all of them $|(\phi(x + iy))|$ are bounded, analytic and having exponential tails with respect to $x \in \mathbf{R}$ within the analytic strip. Given the compact parameter space (actually it is a closed interval for each parameter), A.2 and A.4 are satisfied. The specific computation step is omitted.

■

B.2.9 Proof of Lemma 13.1

Proof. Please see the proof of Lemma 2.3 in Mukerjee and Sutradhar (2002) or 5a.3 in Rao et al. (1973). ■

B.2.10 Proof of Lemma 13.2

Proof. if $V(\theta)$ is not positive definite, there exists non-zero $\beta \in \mathbf{R}^p$, so that $cov(\beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p) = 0$. Then, $\beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p = C$, where C is a constant. Because X has a continuous probability density over real line $\mathbf{R}$, thus X could be any value at least within a small open interval. Notice that $\{X, X^2, \ldots, X^p\}$ are linearly independent. Thus, β is a zero vector. This is a contradiction to non-zero $\beta \in \mathbf{R}^p$. Thus, $V(\theta)$ is positive definite. ■

B.3 The simulation of CGMY processes

We simulate CGMY processes with parameter $C = 3$, $G = 78$, $M = 82$ and $Y = 0.9$. We mainly use the method proposed by Chen et al. (2012). For the daily data ($\delta = 1/252$), we choose $M = 150$ and $h = 3$ to approximate CDF $\tilde{F}(x)$ of Levy processes $X(\delta)$ at time δ by using Hilbert transform method (Section 2 in Chen et al. (2012)). The choice of M and h can guarantee the approximation error less than 10^{-12} for every possible x in this parameter set. Suppose the expected value of $X(\delta)$ is m_{cgmy} and variance is var_{cgmy}. We choose n_1 and n_2 so that $\tilde{F}(x_l) < 10^{-8}$ and $1 - \tilde{F}(x_u) < 10^{-8}$ based on $x_u = m_{cgmy} + n_1(var_{cgmy})^{0.5}$ and $x_l = m_{cgmy} - n_2(var_{cgmy})^{0.5}$. Then, we construct 2^{22} points $\{x\}$ between x_u and x_l uniformly and calculate their corresponding approximated CDF $\tilde{F}(x)$. Then we simulate u following uniform distribution and calculate $\tilde{F}^{-1}(u)$ based on the linear interpolation or a solver. $\tilde{F}^{-1}(u)$ will follow the distribution of $X(\delta)$ approximately. We find interpolation can give us very accurate $\tilde{F}^{-1}(u)$ and the error of $\tilde{F}^{-1}(u)$ less than 10^{-12}. if simulated $u < 10^{-8}$ or $u > 1 - 10^{-8}$, we set $\tilde{F}^{-1}(u)$ to be x_u and x_l. This averagely happens every 10^8 samples which is very rare and the error of $\tilde{F}^{-1}(u)$ is also less than 10^{-12}.All in all, our implementation of the simulation method can control the error less than 10^{-12}.

B.4 Approximated MLE implementation details

To approximate the likelihood $l(\theta; x)$, for our simulated data $x = \{x_j\}_j^N$ and parameter θ, we find M and h so that the approximation error is small. To be simple, we set $a = 0$ to approximate log-likelihood. Thus, our approximated likelihood is $l^{M,h,0}(\theta; x)$. First, through Theorem 6, we find Mh is only related to the parameter and h is related to both parameter and data x for the bound of error. Moreover, for one data x_j, larger $|x_j|$ generally requires smaller h to control the approximation error. Following these two phenomenon, we separate data x into two groups x^1 and x^2. $\max|x^1| < 0.1$, $\min|x^2| > 0.1$. If we want to model equity values, most data will be has the absolute value less than 0.1. If we fix Mh,

we can find that group 1 generally can have larger h with smaller M and notice that large M is mainly the large computation requirement from based on Equation 3.2. Then, we choose common M and h for either group 1 data or group 2 data to guarantee both $|l^{M,h,0}(\theta; x_{max}) - l^{2M,h,0}(\theta; x_{max})| < 10^{-8}$ and $|l^{M,h,0}(\theta; x_{max}) - l^{2M,h/2,0}(\theta; x_{max})| < 10^{-8}$ where x_{max} is either $\max|x^1|$ or $\max|x^2|$. In this way separating data into two groups can effectively reduce the computation burden. In the end, I will check if selected M and h can satisfying $|l^{M,h,0}(\theta; x_{max}) - l^{2M,h,0}(\theta; x_j)| < 10^{-8}$ and $|l^{M,h,0}(\theta; x_{max}) - l^{2M,h/2,0}(\theta; x_j)| < 10^{-8}$ for all $1 \leq j \leq N$. In our experience, this is general the case. If error can be controlled for data x_{max},(in our case, error is controlled by 10^{-8}), similar error can also be controlled for every data point. In this way, we don't need to search M and h for each data point of x and we can save a lot of time. To search M and h, we set the initial search point is $M = 50$ and $h = 5$. This initial point is set by our practical experience. The search pattern for M, h is from the expression of error of bound in Lemma 6. It can be shown that the first two terms converge to zero at the exponential rate $\exp(-\pi d/h)$ and T_{Mh} converges to zero at the exponential rate $\exp(-c(Mh)^\nu)$. Thus, we choose the search new M_1 and h_1 to be $\exp(-(M_1 h)^\nu) \approx \exp(-(Mh)^\nu) * \delta_1$ and $\exp(-(\pi)/h_1) = \exp(-(\pi)/h) * \delta_2$. We set $\delta_1 = 0.1$ and $\delta_2 = 0.5$ in our algorithm applied to the simulation study and case study and it shows high efficiency to search M and h.

Appendix C

Appendix of Chapter 4

C.1 Useful lemmas

We list several typical asymptotic properties of ECF estimates for Lévy processes with proofs under B class regularity conditions. Those properties will be used to prove the theorems of previous sections in this chapter.

We first list the lemma to help us to prove the Theorem 18. This is basically the Theorem 2.3 in Feng and Lin (2013).

A function f is in $H(\mathcal{D}_{(d_-,d_+)})$ if it is analytic in $\mathcal{D}_{(d_-,d_+)}$ and satisfies

$$\int_{d_-}^{d_+} |f(x+iy)|dy \to 0, \qquad x \to \pm\infty.$$

$$||f||^{\pm} \equiv \lim_{\epsilon \to 0+} \int_{\mathbf{R}} |f(x+i(d_{\pm} \mp \epsilon))|dx < +\infty.$$

Lemma 24.11. *Suppose $f \in H(\mathcal{D}_{d_-,d_+})$. The trapezoidal sum approximation:*

$$T_{h,M}(f,a) = \sum_{m=-M}^{M} f(mh+ia)h$$

. The approximation error is:

$$E_{h,M}^{T}(f,a) = \int_{-\infty}^{\infty} f(x)dx - T_{h,M}(f,a).$$

Then for any $a \in (d_-, d_+)$,

$$|E_{h,M}^{T}(f;a)| \leq \frac{e^{-2\pi(a-d_-)/h}}{1-e^{-2\pi(a-d_-)/h}}||f||^{-} + \frac{e^{-2\pi(d_+-a)/h}}{1-e^{-2\pi(d_+-a)/h}}||f||^{+} + T_{Mh}.$$

When f satisfies that:

$$|f(x+ia)| \leq k|x|^n \exp(-c|x|^\nu), \qquad x \in \boldsymbol{R}$$

for some $k > 0, \nu \geq 0$ and either $c > 0, n \in \boldsymbol{R}$ or $c = 0, n < -1$, we have $T_{Mh} = \frac{2k}{|n+1|}(Mh)^{n+1}$ if $c = 0, n < -1$, and $T_{Mh} = \frac{2k}{\nu c^{(n+1)/\nu}}\Gamma(\frac{n+1}{\nu}, c(Mh)^\nu)$ if $c > 0$. Incomplete Gamma function $\Gamma(s,b) = \int_b^\infty e^{-t}t^{s-1}dt$.

Proof. see Theorem 2.3 in Feng and Lin (2013). ■

We list the following lemma which helps us to prove the Theorem 19.

Lemma 24.12. *$X^n \in \mathcal{X}$ is a $n-$dimensional random vector and $Q(\theta; X^n)$ is a real-valued function of θ given X^n. $Q_0(\theta)$ is a real-valued function of θ. Suppose $\theta \in \Theta \subset R^p$ and $\hat{\theta}_n$ is defined as the value of $\theta \in \Theta$ maximizing $Q(\theta; X^n)$. Under such regularity conditions below:*

L24.12.1 Parameter space Θ is compact.

L24.12.2 $Q_0(\theta)$ is uniquely maximized at θ_0.

L24.12.3 $Q_0(\theta)$ is continuous in $\theta \in \Theta$.

L24.12.4 $Q(\theta; X^n)$ converges uniformly in probability to $Q_0(\theta)$. That is, $\sup_{\theta\in\Theta} |Q(\theta; X^n) - Q_0(\theta)| \xrightarrow{p} 0$ as $n \to \infty$.

then, $\hat{\theta}_n \xrightarrow{p} \theta_0$ as $n \to \infty$.

Remark 18. This lemma is the Theorem 2.1 in Newey and McFadden (1994). In fact, the condition L24.12.3 can be generalized to be upper-continuous. The combination of conditions L24.12.1, L24.12.2 and L24.12.3 can be replaced by a more general condition: θ_0 is a well-separated point of the maximum (See Corollary 3.2.3 in Van der Vaart (2000)).

Proof. let $B_\epsilon(\theta_0) = \{\theta : ||\theta - \theta_0|| < \epsilon\}$. Because $\Theta \cap B_\epsilon^C(\theta_0)$ is compact (L24.12.1) and $Q_0(\theta)$ is a continuous function (L24.12.3), there exists $\theta^* \in \Theta \cap B_\epsilon^C(\theta_0)$ to achieve $\sup_{\theta\in\Theta\cap B_\epsilon^C(\theta_0)}\{Q_0(\theta)\}$. Because θ_0 is the unique to maximize $Q_0(\theta)$ (L24.12.2), we denote $Q_0(\theta_0) - Q_0(\theta^*)$ as $\delta > 0$.

Notice that:

$$\sup_{\theta\in(\Theta\cap B_\epsilon^C(\theta_0))} |Q(\theta; X^n) - Q_0(\theta)| < \frac{\delta}{2} \Rightarrow Q(\theta; X) < Q_0(\theta) + \frac{\delta}{2} \leq Q_0(\theta^*) + \frac{\delta}{2} = Q_0(\theta_0) - \frac{\delta}{2}.$$

$$\sup_{\theta\in\Theta\cap B_\epsilon(\theta_0)} |Q(\theta; X^n) - Q_0(\theta)| < \frac{\delta}{2} \Rightarrow Q(\theta_0; X^n) > Q_0(\theta_0) - \frac{\delta}{2}.$$

Then, we have:

$$\sup_{\theta\in\Theta}|Q(\theta;X^n)-Q_0(\theta)|<\frac{\delta}{2}\Rightarrow Q(\theta_0;X^n)>Q(\theta;X^n) \text{ for } \theta\in\Theta\cap B_\epsilon^C(\theta_0)\Rightarrow\theta_n\in\Theta\cap B_\epsilon(\theta_0).$$

Due to the L24.12.4, as $n\to\infty$, we have

$$P(\sup_{\theta\in\Theta}|Q(\theta;X^n)-Q_0(\theta)|<\frac{\delta}{2})\xrightarrow{p}1.$$

Then, we have $P(\theta_n\in\Theta\cap B_\epsilon(\theta_0))\xrightarrow{p}1$ as $n\to\infty$. Equivalently, we have $\hat{\theta}_n\xrightarrow{p}\theta_0$ as $n\to\infty$. ■

C.1.1 B class regularity conditions in Chapter 4

We list the following B class of regularity conditions for the asymptotic property of ECF estimation, which is proposed by Knight and Yu (2002). This class of regularity conditions work for both i.i.d and dependent data. We discard the regularity conditions which are for non-i.i.d data (always hold for i.i.d data) and only list the ones for i.i.d data.

B.1 θ_0 lies in the interior of compact parameter space Θ.

B.2 With probability one, $e(\theta;X)$ is twice continuously differentiable under the integral sign with respect to θ over Θ.

B.3 Let $e_0(\theta)=\int|\phi(u;\theta)-\phi(u;\theta_0)|^2g(u)du$ and $e_0(\theta)=0$ only if $\theta=\theta_0$.

B.4 $K(x;\theta)$ is a measurable function of x for all θ and bounded, where

$$K(x;\theta)=\int\{(\cos(ux)-\Re\phi(u;\theta))\frac{\partial\Re\phi(u;\theta)}{\partial\theta}+(\sin(ux)-\Im\phi(u;\theta))\frac{\partial\Im\phi(u;\theta)}{\partial\theta}\}g(u)du$$

B.5 $B(\theta_0)=\int(\partial\phi(u;\theta_0)/\partial\theta)(\partial\phi(u;\theta_0)/\partial\theta^T)g(u)du$ is nonsingular and $\partial^2\phi(u;\theta)/\partial\theta\partial\theta^T$ is uniformly bounded by a g-integrable function over Θ.

Now, we list the lemma to prove the Theorem 20.

Lemma 24.13 (Consistency and normality of ECF estimates for I.I.D data)**.** *Under regularity conditions B.1 ∼ B.3, the ECF estimate $\hat{\theta}_{ECF}$ is a consistent estimator. That is,*

$$\hat{\theta}_{ECF}\xrightarrow{p}\theta_0 \tag{C.1}$$

when sample size $n \to \infty$. When all B class of regularity conditions hold, then

$$\sqrt{n}(\hat{\theta}_{ECF} - \theta_0) \xrightarrow{d} N(0, B^{-1}(\theta_0)A(\theta_0)B^{-1}(\theta_0)),$$

where $A(\theta_0) = var(K(x1; \theta_0))$.

Proof. See the Theorem 2.1 in Knight and Yu (2002) ■

This interchange of differentiation and integration lemma will be used to prove Theorem 24.

Lemma 24.14 (Interchange of differentiation and integration)**.** *$a(x, \theta)$ is a real-valued function of an observation $x \in \mathcal{X}$ and the parameter θ. If $a(x, \theta)$ is continuously differentiable at $\theta \in \mathcal{N}$ for all $x \in \mathcal{X}$ where $\mathcal{N}$ is an open set and $\int_{\mathcal{X}}(\sup_{\theta\in\mathcal{N}} ||\frac{\partial}{\partial\theta}a(x, \theta)||)dx < \infty$, then, $\int_{\mathcal{X}} a(x, \theta)dx$ is continuously differentiable and $\frac{\partial}{\partial\theta}\int_{\mathcal{X}} a(x, \theta)dx = \int_{\mathcal{X}}(\frac{\partial}{\partial\theta}a(x, \theta))dx$.*

Proof. For any $\theta \in \mathcal{N}$, we can choose a sequence $\theta_n \in \mathcal{N}$ so that $\theta_n \to \theta$ because $\mathcal{N}$ is an open set. Since $a(x, \theta)$ is continuously differentiable at $\theta \in \mathcal{N}$ for all $x \in \mathcal{X}$, we have $\frac{\partial}{\partial\theta_n}a(x, \theta) \to \frac{\partial}{\partial\theta_n}a(x, \theta)$. In addition, we know that $\int_{\mathcal{X}}(\sup_{\theta\in\mathcal{N}} ||\frac{\partial}{\partial\theta}a(x, \theta)||)dx < \infty$, Thus, $\int_{\mathcal{X}}(\frac{\partial}{\partial\theta}a(x, \theta_n))dx \to \int_{\mathcal{X}}(\frac{\partial}{\partial\theta}a(x, \theta))dx$ due to dominated convergence theorem. Thus, $\int_{\mathcal{X}}(\frac{\partial}{\partial\theta}a(x, \theta))dx$ is continuous. By mean-value expansion theorem, we have $a(x, \theta_n) = a(x, \theta) + \frac{\partial}{\partial\theta}a(x, \theta)'(\theta_n - \theta) + r(x, \theta_n)$ where $r(x, \theta_n) = [\frac{\partial}{\partial\theta}a(x, \theta_n^*) - \frac{\partial}{\partial\theta}a(x, \theta)]'(\theta_n - \theta)$ and θ_n^* is the mean value located in the line between θ_n and θ. Because $\mathcal{N}$ is an open set, for all θ_n closed enough to θ, we can assume that $\theta_n^* \in \mathcal{N}$ without losing generality. Then, we have $||r(x, \theta_n)||/||\theta_n - \theta|| \leq ||\frac{\partial}{\partial\theta}a(x, \theta_n^*) - \frac{\partial}{\partial\theta}a(x, \theta)|| \to 0$ by the continuity of $\frac{\partial}{\partial\theta}a(x, \theta)$. In addition, $||r(x, \theta_n)||/||\theta_n - \theta|| \leq 2\sup_{\theta\in\mathcal{N}} ||\frac{\partial}{\partial\theta}a(x, \theta)||$, we have $||r(x, \theta_n)||/||\theta_n - \theta|| \to 0$ by dominated convergence theorem. Therefore, $|\int_{\mathcal{X}} a(x, \theta_n) - \int_{\mathcal{X}} a(x, \theta) - [\int_{\mathcal{X}} \frac{\partial}{\partial\theta}a(x, \theta)dx]'(\theta_n - \theta) = |\int_{\mathcal{X}} r(x, \theta_n)dx| \leq \int_{\mathcal{X}} |r(x, \theta_n)|dx = o(||\theta_n - \theta||)$. Equivalently, we have $\frac{\partial}{\partial\theta}\int_{\mathcal{X}} a(x, \theta)dx = \int_{\mathcal{X}}(\frac{\partial}{\partial\theta}a(x, \theta))dx$. ■

We also need to use the following lemma proposed in Chapter 2.3.3 (Lemma 24.1) to prove the regularity condition B.5.

Lemma 24.15. *$\{A(t)\}_{t\in R^d}$ are positive-semidefinite $n\times n$ matrices of which entries are continuous functions of $t \in R^d$. $G(t)$ is a probability measure with support S containing an open set $I \subset S$. Then, there will be no open set $\mathcal{I} \subseteq I$ so that $\int_{\mathcal{I}} A(t)dG(t)$ is positive definite.*
$\Leftrightarrow$ there exists non-zero constant β so that $\forall t \in I$, $A(t)\beta = 0$.

Proof. $\Leftarrow$ Note that there exists non-zero constant vector β(not related to t) so that for $\forall t$ $A(t)\beta = 0$ for $\forall t \in I$. Thus, for any open set $\mathcal{I} \subset I$, we have $\int_{\mathcal{I}}(A(t))dG(t)\beta = \int_{\mathcal{I}}(A(t))\beta dG(t) = 0$. That is, $\int_{\mathcal{I}} A(t)dG(t)$

is not full rank.

$\Rightarrow$ We prove it by contradiction. If for $\forall \mathcal{I} \subseteq I$, $\int_{\mathcal{I}} A(t)dG(t)$ is not full rank. First we select an arbitrary open set $\mathcal{I}_1 \subseteq I$. Because $\int_{\mathcal{I}_1} A(t)dG(t)$ is not full rank, there exists non-zero constant vector β_1 satisfying $\int_{\mathcal{I}_1} A(t)dG(t)\beta_1 = 0$. Then, $\int_{\mathcal{I}_1} \beta_1^T A(t)\beta_1 dG(t) = 0$. Considering $A(t)$ is positive-semidefinite matrix and $A(t)$ is continuous with respect to t, we have $\beta_1^T A(t)\beta_1^T = 0$ for $\forall t \in \mathcal{I}_1$. Also, $A(t)$ can be written as $P(t)^T P(t)$ by Cholesky decomposition. Then, we have $(P(t)\beta_1^T)^T(P(t)\beta_1) = 0$ which is $A(t)\beta_1 = 0$ for $\forall t \in \mathcal{I}_1$. Because there is no non-zero constant vector β so that $A(t)\beta = 0$ for $\forall t \in I$, there exists $t_2 \notin \mathcal{I}_1$ so that $A(t_2)\beta_1 \neq 0$, Thus, we select another open set $\mathcal{I}_2$ satisfying $\mathcal{I}_2 \supset (\mathcal{I}_1 \cup t_2)$ and $\mathcal{I}_2 \subseteq I$. Because $\int_{\mathcal{I}_2} A(t)dG(t)$ is not full rank, there exists non-zero constant vector β_2 satisfying $\int_{\mathcal{I}_2} A(t)dG(t)\beta_2 = 0$. Definitely we also have $A(t_2)\beta_2 = 0$. Repeating the procedure above, we will have:

1. Open set $\mathcal{I}_1 \subset \mathcal{I}_2 \cdots \subseteq I$;

2. Non-zero constant vector β_i satisfying $A(t)\beta_i = 0$ for $\forall t \in \mathcal{I}_i$ $(i = 1, 2, \ldots)$;

3. For $j = 2, 3, \ldots$, there exists $t_j \in \mathcal{I}_j \setminus \mathcal{I}_{i-1}$ satisfying $A(t_j)\beta_{j-1} \neq 0$. But, $A(t_j)\beta_k = 0$ for $\forall k \geq j$.

Suppose we select $\beta_1, \beta_2, \ldots, \beta_{n+1}$. Then, they must be linearly dependent (β is $n \times 1$ vector). That is, there exists not all zero real numbers $c_1, \ldots, c_{n+1}$ so that $c_1\beta_1 + c_2\beta_2 + \cdots + c_{n+1}\beta_{n+1} = 0$. Then, $A(t_2)(c_1\beta_1 + c_2\beta_2 + \cdots + c_{n+1}\beta_{n+1}) = 0$. Then, we have $c_1 A(t_2)\beta_1 = 0$ which lead to $c_1 = 0$. Similarly, we multiply $A(t_3)$ by $c_2\beta_2 + \cdots + c_{n+1}\beta_{n+1}$ to get $c_2 = 0$. In the end, we will get $c_1 = c_2 = \cdots = c_{n+1} = 0$. This is a contradiction. ■

C.2 Proofs

C.2.1 Proof of Theorem 18

We basically follow the proof of Theorem 2.2 in Feng and Lin (2013) (Lemma 24.11). For simplicity, we use a close analytic strip $\mathcal{D}_{[d_-,d_+]}$ in our work instead of the open one in Feng and Lin (2013). This will not make a difference.

Let $f = |\phi(u;\theta) - \theta_n(u)|^2 g(u)$. Then, for each fixed $\theta \in \Theta$, regularity condition A.1, A.2 and A.3 can guarantee $f \in H(\mathcal{D}_{[d_-,d_+]})$, $||f||^- \leq C_1||g||^-$ and $||f||^+ \leq C_2||g||^+$. Applying Lemma 24.11 with regularity condition A.4 and A.5, we have:

$$|E^F_{\hbar,\mathcal{M}}(e, \boldsymbol{a})(\theta; X)| \leq \frac{e^{-2\pi(a-d_-)/\hbar}}{1 - e^{-2\pi(a-d_-)/\hbar}} C_1||g||^- + \frac{e^{-2\pi(d_+-a)/\hbar}}{1 - e^{-2\pi(d_+-a)/\hbar}} C_2||g||^+ + T_{\mathcal{M}\hbar}.$$

Moreover, Because bound of $|E^F_{\hbar,\mathcal{M}}(\phi, \boldsymbol{a})(x)|$ in (4.6) doesn't depend on parameter θ based on the A class

regularity conditions, then, $|E^F_{\hbar,\mathcal{M}}(\phi, a)(x)|$ decay to zero uniformly on $\theta \in \Theta$ as $\mathcal{M}\hbar \to \infty$ and $h \to 0$.

C.2.2 Proof of Theorem 19

Proof. We utilize Lemma 24.12 by setting $Q(\theta, X, \mathcal{M}, \hbar, a) = -e^{\mathcal{M},\hbar,a}(\theta; X)$ and $Q_0(\theta, X) = -e(\theta; X)$. We know that $Q(\theta, X, \mathcal{M}, \hbar, a)$ converges to $Q_0(\theta, X)$ uniformly for $\theta \in \Theta$ when $\mathcal{M}\hbar \to \infty$ and $\hbar \to 0$ via Theorem 18. Then, Lemma 24.12 holds and $\hat{\theta}_n^{\mathcal{M},\hbar,a} \xrightarrow{p} \hat{\theta}_{ECF}$ when $n \to \infty$. ■

C.2.3 Proof of Theorem 20

Proof. Under B class regularity conditions, we have

$$\sqrt{n}(\hat{\theta}_{ECF} - \theta_0) \xrightarrow{d} N(0, B^{-1}(\theta_0)A(\theta_0)B^{-1}(\theta_0))$$

via Lemma 24.13. Combined with Theorem 19, we have $\hat{\theta}_n^{\mathcal{M},\hbar,a}$, $\hat{\theta}_n^{\mathcal{M}(n),\hbar(n),a} \xrightarrow{p} \theta_0$ with fixed a when sample size $n \to \infty$, and

$$\sqrt{n}(\hat{\theta}_{ECF}^{\mathcal{M}(n),\hbar(n),a} - \theta_0) \xrightarrow{d} N(0, B^{-1}(\theta_0)A(\theta_0)B^{-1}(\theta_0))$$

via Slutsky's theorem. ■

C.2.4 Proof of Proposition 21

Proof. $g(u) = \frac{1}{\sqrt{2\pi}} \exp -\frac{u^2}{2\sigma^2}$, which is analytic function. This satisfies the A.1.
$g(x + iy) = \frac{1}{\sqrt{2\pi}} \exp -\frac{x^2 - y^2 + 2ixy}{2\sigma^2}$ converge to zero when $x \to \infty$ given any $y \in [-d, d]$. Also, it is simply to show $||g||^+$ is bounded due to its exponential tail $\exp(-x^2/(2\sigma^2))$. Thus, A.3 is fulfilled.
For the common lèvy processes, $|\phi(x + ia; \theta) - \phi_n(x + ia)|$ is uniformly bounded with respect to θ into a compact parameter space. Also, it is easy to show that $g(x + ai) \leq \mathcal{K}|x|^{\imath} \exp(-\mathcal{c}|x|^{\nu}), \quad x \in \mathbf{R}$ with $\imath = 0$, $\nu = 2$, $\mathcal{c} = 1/(2\sigma^2)$ and $\mathcal{K} = \frac{1}{\sqrt{2\pi}} \exp(\max^2(-d_-, d_+)/(2\sigma^2))$. Thus, A.4 also hold. Thus, normal distribution density satisfies the regularity condition. ■

C.2.5 Proof of Proposition 22

Proof. Utilize cental limit theorem for $\cos(uX)$ and $\sin(uX)$. Also, notice that $\cos(2uX) = 1 - 2\cos^2(uX)$. Then, we have the result in Proposition 22. ■

C.2.6 Proof of Theorem 23

Proof. If we follow the selection of analytic strip $\mathcal{D}_{[d_-,d_+]}$ for each model in Section 4.3.1, A.1 is satisfied because their characteristic function $\phi(u;\theta)$ is analytic in the analytic strip. The proof of analyticity for each model is trivial.

For A.2, the model's characteristic function forms are provided in the Section 4.3.1. And all of them, $|(\phi(x+iy))|$, are bounded, analytic and having exponential tails with respect to $x \in \mathbf{R}$ within the analytic strip. Given the compact parameter space (actually it is a closed interval for each parameter), A.2 is satisfied. The specific computation step is omitted.

Similarly to A.2, we can show $|(\phi(x+i\boldsymbol{a}))|$ is uniformly bounded with respect to parameter θ given $\boldsymbol{a}$. Also, empirical characteristic function $\phi_n(x+i\boldsymbol{a})$ is also obviously bounded given the data set X and $\boldsymbol{a}$ due to its definition. Thus, $||\phi(x+i\boldsymbol{a};\theta) - \phi_n(x+i\boldsymbol{a})||$ is uniformly bounded with respect to θ into a compact parameter space. Also, we know that $g(x+\boldsymbol{a}i) \leq \mathcal{k}|x|^n \exp(-\mathcal{c}|x|^\nu), \quad x \in \mathbf{R}$ based on our selection of $g(u)$ which satisfies regularity condition A.4. Thus, A.4 is satisfied. ■

C.2.7 Proof of Theorem 24

Proof. B.1 is satisfied based on our choice of compact parameter space Θ in Section 4.3.1.

To prove B.2, we can use the Lemma 24.14 and show $\int_{\mathcal{X}}(\sup_{\theta\in\mathcal{N}} ||\frac{\partial}{\partial\theta}|\phi(x;\theta) - \phi_n(x)|g(x)||)dx < \infty$. The computation is tough. We use Mathematica and find $\frac{\partial}{\partial\theta}|\phi(x;\theta) - \phi_n(x)|$ is uniformly bounded for each model with respect to the parameter in the compact parameter space. Also, we know $g(x)$ is selected to satisfy class A regularity conditions. That is, $g(x) \leq \mathcal{k}|x|^n \exp(-\mathcal{c}|x|^\nu), \quad x \in \mathbf{R}$. Then, $\int_{\mathcal{X}}(\sup_{\theta\in\mathcal{N}} ||\frac{\partial}{\partial\theta}|\phi(x;\theta) - \phi_n(x)|g(x)||)dx < \infty$. Also, it is easy to show that $|\phi(x;\theta) - \phi_n(x)|g(x)$ is continuously differentiable at θ inside the parameter space, Thus, B.2 is satisfied.

To prove B.3, we only need to show $\phi(u;\theta)$ and $\phi(u;\theta_0)$ have different values for u within an interval. This is also easy to prove by using the linear independent property in the linear algebra. We take NIG model as an instance. $\log\phi(u;\theta) = i\mu u - \lambda(\sqrt{(\alpha^2-(\beta+iu)^2)} - \sqrt{\alpha^2-\beta^2})$. Also, $\log\phi(u;\theta_0) = i\mu_0 u - \lambda_0(\sqrt{(\alpha_0^2-(\beta_0+iu)^2)} - \sqrt{\alpha_0^2-\beta_0^2})$. We can find that $i\mu u$, $\lambda\sqrt{(\alpha^2-(\beta+iu)^2)}$ and $\lambda\sqrt{\alpha^2-\beta^2}$ are linear independent in terms of u. Then, $i\mu u = i\mu_0 u$, $\lambda(\sqrt{(\alpha^2-(\beta+iu)^2)}) = \lambda_0(\sqrt{(\alpha_0^2-(\beta_0+iu)^2)}$ and $\lambda\sqrt{\alpha^2-\beta^2} = \lambda_0\sqrt{\alpha_0^2-\beta_0^2}$ for all u in the interval. Then, we use the similar logic for $\lambda(\sqrt{(\alpha^2-(\beta+iu)^2)}) = \lambda_0(\sqrt{(\alpha_0^2-(\beta_0+iu)^2)}$ by square both sides, we can find $\lambda = \lambda_0$ and $\beta = \beta_0$. Then, we can find $\alpha = \alpha_0$ from the equation $\lambda\sqrt{\alpha^2-\beta^2} = \lambda_0\sqrt{\alpha_0^2-\beta_0^2}$ and $\mu = \mu_0$ from the equation $i\mu u = i\mu_0 u$. Thus, B.3 is satisfied for NIG model. Similarly, Merton's jump-diffusion model, Kou's jump-diffusion model also can be proved simply. For CGMY model, to prove the identification, we can use the (7.14) and (7.15) in Miyahara (2002).

To prove B.4, we can show stronger result that $K(x, \theta)$ is continuously differentiable under the integral sign based on the Lemma 24.14. The proof is similar to the proof of B.2 and we use Mathematica to conduct the calculation, finding the part inside the integral is smooth, bounded and exponential tail with respect to u uniformly on θ in the compact parameter space (defined in Section 4.3.1). Computational details are skipped here.

To prove B.5, we first prove $B(\theta_0)$ is nonsingular. This is a little tricky because we do not have a close form of $B(\theta_0)$ because it is an integration with respect to u. We use Lemma 24.15 which proposed in Chapter 2.3.3. Notice that in our framework, $G'(u) = g(u)$ and $G(u)$ is bounded and increasing function. Thus, if we define $A(t)$ in the Lemma 24.15 to be $A(t) = (\partial\phi(t;\theta_0)/\partial\theta)(\partial\phi(t;\theta_0)/\partial\theta^T)$, Lemma 24.15 can be applied. Then, we only need to show that there is no β so that $A(t)\beta = 0$ for $t \in R$, where β is not a function of t. This is relative easy to prove because we know the close form of $A(t)$ and $A(t)$ does not have the integration part anymore. We use Mathematica to calculate $A(t)$ for each model first, we find for the most cases, all the elements in the first row of $A(t)$ is even linearly independent with respect to t. Thus, $\beta = 0$. Then, there is no β to guarantee $A(t)\beta = 0$ for $t \in R$. Then, $B(\theta_0)$ is nonsingular.

To prove that $\partial^2\phi(u;\theta)/\partial\theta\partial\theta^T$ is uniformly bounded by a g-integrable function over Θ. We use Mathematica to derive the close form of $\partial^2\phi(u;\theta)/\partial\theta\partial\theta^T$ for each model and find each element in $\partial^2\phi(u;\theta)/\partial\theta\partial\theta^T$ can be uniformly bounded by a constant, given compact parameter space in Section 4.3.1. A constant is certainly a g-integrable function, because $\int_{\mathbf{R}} g(u) = G(u)|_0^\infty$, which is bounded in ECF framework. ■

References

Abate, J., and Whitt, W. (1992), "The Fourier-series method for inverting transforms of probability distributions," *Queueing systems*, 10(1-2), 5–87.

Abramson, M., Audet, C., Couture, G., Dennis, Jr., J., LeD̃igabel, S., and Tribes, C. (n.d.), "The NOMAD project,", Software available at `https://www.gerad.ca/nomad/`.

A.Feuerverger, and R.A.Mureika (1977), "The empirical characteristic function and its applications.," *The annals of statistics*, 5, 88–97.

Andrieu, C., and Roberts, G. O. (2009), "The pseudo-marginal approach for efficient Monte Carlo computations," *The Annals of Statistics*, pp. 697–725.

Avellaneda, M., and Lee, J.-H. (2010), "Statistical arbitrage in the US equities market," *Quantitative Finance*, 10(7), 761–782.

Barndorff-Nielsen, O. E. (1977), "Exponentially decreasing distributions for the logarithm of particle size," *Royal Society of London. Proceedings. Mathematical, Physical and Engineering Sciences*, .

Barndorff-Nielsen, O. E. (1997), "Processes of normal inverse Gaussian type," *Finance and stochastics*, 2(1), 41–68.

Barndorff-Nielsen, O. E., and Shephard, N. (2001), "Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 167–241.

Beaumont, M. A. (2003), "Estimation of population growth or decline in genetically monitored populations," *Genetics*, 164(3), 1139–1160.

Billingsley, P. (1961), *Statistical Inference for Markov Processes*, Statistical Research Monographs University of Chicago Press.

Black, F., and Scholes, M. (1973), "The pricing of options and corporate liabilities," *The journal of political economy*, pp. 637–654.

Carr, P., Geman, H., Madan, D. B., and Yor, M. (2002), "The fine structure of asset returns: An empirical investigation," *The Journal of Business*, 75(2), 305–333.

Chan, N. H., Chen, S. X., Peng, L., and Yu, C. L. (2009), "Empirical likelihood methods based on characteristic functions with applications to Lévy processes," *Journal of the American Statistical Association*, 104(488), 1621–1630.

Chant, D. (1974), "On asymptotic tests of composite hypotheses in nonstandard conditions," *Biometrika*, 61(2), 291–298.

Chen, S. X., Peng, L., and Cindy, L. Y. (2013), "Parameter estimation and model testing for Markov processes via conditional characteristic functions," *Bernoulli*, 19(1), 228–251.

Chen, Z., Feng, L., and Lin, X. (2012), "Simulating Lévy processes from their characteristic functions and financial applications," *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 22(3), 14.

Chernoff, H. (1954), "On the distribution of the likelihood ratio," *The Annals of Mathematical Statistics*, pp. 573–578.

Chernozhukov, V., and Hong, H. (2003), "An MCMC approach to classical estimation," *Journal of Econometrics*, 115(2), 293–346.

DiCiccio, T. J., Hall, P., and Romano, J. P. (1989), "Comparison of parametric and empirical likelihood functions," *Biometrika*, 76(3), 465–476.

Dong, X., Berti-Equille, L., and Srivastava, D. (2009*a*), "Integrating conflicting data: the role of source dependence," *PVLDB*, 2(1), 550–561.

Dong, X., Berti-Equille, L., and Srivastava, D. (2009*b*), "Truth discovery and copying detection in a dynamic world," *PVLDB*, 2(1), 562–573.

Doucet, A., Godsill, S. J., and Robert, C. P. (2002), "Marginal maximum a posteriori estimation using Markov chain Monte Carlo," *Statistics and Computing*, 12(1), 77–84.

Duffie, D., Filipović, D., and Schachermayer, W. (2003), "Affine processes and applications in finance," *Annals of applied probability*, pp. 984–1053.

Durbin, J., and Koopman, S. J. (2012), *Time series analysis by state space methods*, Vol. 38 OUP Oxford.

Eberlein, E., Keller, U., and Prause, K. (1998), "New insights into smile, mispricing, and value at risk: The hyperbolic model*," *The Journal of Business*, 71(3), 371–405.

Fama, E. F. (1965), "The behavior of stock-market prices," *The journal of Business*, 38(1), 34–105.

Feder, P. I. (1968), "On the distribution of the log likelihood ratio test statistic when the true parameter is" near" the boundaries of the hypothesis regions," *The Annals of Mathematical Statistics*, 39(6), 2044–2055.

Feng, L., and Lin, X. (2013), "Inverting analytic characteristic functions and financial applications," *SIAM Journal on Financial Mathematics*, 4(1), 372–398.

Feuerverger, A., and McDunnough, P. (1981*a*), "On some Fourier methods for inference," *Journal of the American Statistical Association*, 76(374), 379–387.

Feuerverger, A., and McDunnough, P. (1981*b*), "On the efficiency of empirical characteristic function procedures," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 20–27.

Feuerverger, A., and Mureika, R. A. (1977), "The empirical characteristic function and its applications," *The annals of statistics*, pp. 88–97.

Galland, A., Abiteboul, S., Marian, A., and Senellart, P. (2010), Corroborating information from disagreeing views,, in *Proc. of WSDM*.

Garcia-Ulloa, D. A., Xiong, L., and Sunderam, V. (2017), "Truth discovery for spatio-temporal events from crowdsourced data," *Proceedings of the VLDB Endowment*, 10(11), 1562–1573.

Geman, H. (2002), "Pure jump Lévy processes for asset price modelling," *Journal of Banking & Finance*, 26(7), 1297–1316.

Glasserman, P. (2003), *Monte Carlo methods in financial engineering*, Vol. 53 Springer Science & Business Media.

Glasserman, P., and Kim, K.-K. (2010), "Moment explosions and stationary distributions in affine diffusion models," *Mathematical Finance*, 20(1), 1–33.

Grendár, M., and Judge, G. (2009), "Asymptotic equivalence of empirical likelihood and Bayesian MAP," *The Annals of Statistics*, pp. 2445–2457.

Grynkiv, I. (2010), "Estimation of Jump-Diffusion and pure jump models of stock prices," *Working paper*, .

Hansen, L. P. (1982), "Large sample properties of generalized method of moments estimators," *Econometrica: Journal of the Econometric Society*, pp. 1029–1054.

Harvey, A. C. (1990), *Forecasting, structural time series models and the Kalman filter* Cambridge university press.

Heathcote, C. (1977), "The integrated squared error estimation of parameters," *Biometrika*, 64(2), 255–264.

Hwang, C.-R. (1980), "Laplace's method revisited: weak convergence of probability measures," *The Annals of Probability*, pp. 1177–1182.

Iglewicz, B., and Hoaglin, D. C. (1993), *How to detect and handle outliers*, Vol. 16 Asq Press.

Jacquier, E., Johannes, M., and Polson, N. (2007), "MCMC maximum likelihood for latent state models," *Journal of Econometrics*, 137(2), 615–640.

Jin, P., Rüdiger, B., and Trabelsi, C. (2016), "Exponential ergodicity of the jump-diffusion CIR process," in *Stochastics of Environmental and Financial Economics* Springer, pp. 285–300.

Johannes, M. S., and Polson, N. (2003), "MCMC methods for continuous-time financial econometrics," *Available at SSRN 480461*, .

Johansen, A. M., Doucet, A., and Davy, M. (2008), "Particle methods for maximum likelihood estimation in latent variable models," *Statistics and Computing*, 18(1), 47–57.

Kiefer, N. M. (1978), "Discrete parameter variation: efficient estimation of a switching regression model," *Econometrica: Journal of the Econometric Society*, pp. 427–434.

Kim, Y. S., Rachev, S. T., Bianchi, M. L., and Fabozzi, F. J. (2008), "Financial market models with Lévy processes and time-varying volatility," *Journal of Banking & Finance*, 32(7), 1363–1378.

K.J.Singleton (2001), "Estimation of affine asset pricing models using the empirical characteristic function," *Journal of Econometrics*, 102, 111–141.

Knight, J. L., and Yu, J. (2002), "Empirical characteristic function in time series estimation," *Econometric Theory*, 18(3), 691–721.

Kopylev, L. (2012), "Constrained parameters in applications: Review of issues and approaches," *ISRN Biomathematics*, 2012.

Kopylev, L., and Sinha, B. (2011), "On the asymptotic distribution of likelihood ratio test when parameters lie on the boundary," *Sankhya B*, 73(1), 20–41.

Kou, S. G. (2002), "A jump-diffusion model for option pricing," *Management science*, 48(8), 1086–1101.

Kunitomo, N., and Owada, T. (2006), "Empirical likelihood estimation of Lévy processes," *CIRJE Discussion Papers*, .

Lazar, N. A. (2003), "Bayesian empirical likelihood," *Biometrika*, 90(2), 319–326.

Li, Q., Li, Y., Gao, J., Su, L., Zhao, B., Demirbas, M., Fan, W., and Han, J. (2014), "A confidence-aware approach for truth discovery on long-tail data," *PVLDB*, .

Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W., and Han, J. (2014), Resolving Conflicts in Heterogeneous Data by Truth Discovery and Source Reliability Estimation,, in *SIGMOD*.

Li, X., Dong, X. L., Lyons, K., Meng, W., and Srivastava, D. (2012), Truth finding on the deep web: Is the problem solved?,, in *Proceedings of the VLDB Endowment*, Vol. 6, VLDB Endowment, pp. 97–108.

Li, Y., Li, Q., Gao, J., Su, L., Zhao, B., Fan, W., and Han, J. (2015), On the discovery of evolving truth,, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 675–684.

Little, M. P., Heidenreich, W. F., and Li, G. (2010), "Parameter identifiability and redundancy: theoretical considerations," *PloS one*, 5(1), e8915.

Liu, L., Ren, X., Zhu, Q., Zhi, S., Gui, H., Ji, H., and Han, J. (2017), "Heterogeneous Supervision for Relation Extraction: A Representation Learning Approach," *arXiv preprint arXiv:1707.00166*, .

Liu, X., Dong, X. L., Ooi, B. C., and Srivastava, D. (2011), "Online data fusion," *Proceedings of the VLDB Endowment*, 4(11), 932–943.

Locatelli, M. (2000), "Simulated annealing algorithms for continuous global optimization: convergence conditions," *Journal of Optimization Theory and applications*, 104(1), 121–133.

Ma, F., Li, Y., Li, Q., Qiu, M., Gao, J., Zhi, S., Su, L., Zhao, B., Ji, H., and Han, J. (2015), Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation,, in *KDD*.

Madan, D. B., Carr, P. P., and Chang, E. C. (1998), "The variance gamma process and option pricing," *European finance review*, 2(1), 79–105.

Madan, D. B., and Milne, F. (1991), "Option Pricing With VG Martingale Components1," *Mathematical finance*, 1(4), 39–55.

Madan, D. B., and Seneta, E. (1987), "Chebyshev polynomial approximations and characteristic function estimation," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 163–169.

Madan, D. B., and Seneta, E. (1990), "The variance gamma (VG) model for share market returns," *Journal of business*, pp. 511–524.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003), "Markov chain Monte Carlo without likelihoods," *Proceedings of the National Academy of Sciences*, 100(26), 15324–15328.

Mengersen, K. L., Pudlo, P., and Robert, C. P. (2013), "Bayesian computation via empirical likelihood," *Proceedings of the National Academy of Sciences*, 110(4), 1321–1326.

Merton, R. C. (1973), "Theory of rational option pricing," *The Bell Journal of economics and management science*, pp. 141–183.

Merton, R. C. (1976), "Option pricing when underlying stock returns are discontinuous," *Journal of financial economics*, 3(1-2), 125–144.

Miyahara, Y. (2002), "Estimation of Lévy processes," , .

Moran, P. A. (1971), Maximum-likelihood estimation in non-standard conditions,, in *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 70, Cambridge Univ Press, pp. 441–450.

Mukerjee, R., and Sutradhar, B. C. (2002), "On the positive definiteness of the information matrix under the binary and Poisson mixed models," *Annals of the Institute of Statistical Mathematics*, 54(2), 355–366.

Mukherjee, S., Weikum, G., and Danescu-Mizil, C. (2014), People on drugs: credibility of user statements in health communities,, in *KDD*.

Newey, W. K., and McFadden, D. (1994), "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

Owen, A. (1988), "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, 75(2), 237–249.

Owen, A. (1990), "Empirical likelihood ratio confidence regions," *The Annals of Statistics*, pp. 90–120.

Pal, A., Rastogi, V., Machanavajjhala, A., and Bohannon, P. (2012), Information integration over time in unreliable and uncertain environments,, in *Proceedings of the 21st international conference on World Wide Web*, ACM, pp. 789–798.

Pasternack, J., and Roth, D. (2010), Knowing what to believe (when you already know something),, in *COLING*.

Paulson, A. S., Holcomb, E. W., and Leitch, R. A. (1975), "The estimation of the parameters of the stable laws," *Biometrika*, 62(1), 163–170.

Qi, G.-J., Aggarwal, C. C., Han, J., and Huang, T. (2013), Mining collective intelligence in diverse groups,, in *WWW*.

Qin, J., and Lawless, J. (1994), "Empirical likelihood and general estimating equations," *The Annals of Statistics*, pp. 300–325.

Ramezani, C. A., and Zeng, Y. (2007), "Maximum likelihood estimation of the double exponential jump-diffusion process," *Annals of Finance*, 3(4), 487–507.

Rao, C. R., Rao, C. R., Statistiker, M., Rao, C. R., and Rao, C. R. (1973), *Linear statistical inference and its applications*, Vol. 2 Wiley New York.

Rydberg, T. H. (1997), "The normal inverse Gaussian Lévy process: simulation and approximation," *Communications in statistics. Stochastic models*, 13(4), 887–910.

Sato, K. (1999), *Lévy processes and infinitely divisible distributions* Cambridge university press.

Schwarz, G. (1978), "Estimating the dimension of a model," *The annals of statistics*, 6(2), 461–464.

Self, S. G., and Liang, K.-Y. (1987), "Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions," *Journal of the American Statistical Association*, 82(398), 605–610.

Seneta, E. (2004), "Fitting the variance-gamma model to financial data," *Journal of Applied Probability*, 41, 177–187.

Shiga, T. (1990), "A recurrence criterion for Markov processes of Ornstein-Uhlenbeck type," *Probability Theory and Related Fields*, 85(4), 425–447.

Shumway, R. H., and Stoffer, D. S. (1982), "An approach to time series smoothing and forecasting using the EM algorithm," *Journal of time series analysis*, 3(4), 253–264.

Sinha, B., Kopylev, L., and Fox, J. (2007), Some new aspects of dose-response multistage models with applications,, in *Environmental and Ecological Statistics (accepted for publication). Platinum Jubilee Conference of ISI, World Scientific Publishing, Singapore. Earlier version is available as UMBC technical report: http://www. math. umbc. edu/˜ kogan/technical_papers/2007/Sinha_Kopylev_Fox. pdf*, Citeseer.

Tankov, P. (2003), *Financial modelling with jump processes*, Vol. 2 CRC press.

Van der Vaart, A. W. (2000), *Asymptotic statistics*, Vol. 3 Cambridge university press.

Vuong, Q. H. (1989), "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica: Journal of the Econometric Society*, pp. 307–333.

Vydiswaran, V., Zhai, C., and Roth, D. (2011), Content-driven trust propagation framework,, in *Proc. of SIGKDD*.

Wang, D., Kaplan, L., Le, H., and Abdelzaher, T. (2012), "On truth discovery in social sensing: A maximum likelihood estimation approach," *IPSN*, .

Wilks, S. S. (1938), "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, 9(1), 60–62.

Wu, C. (2004), "Some algorithmic aspects of the empirical likelihood method in survey sampling," *Statistica Sinica*, pp. 1057–1067.

Wu, C. J. (1983), "On the convergence properties of the EM algorithm," *The Annals of statistics*, pp. 95–103.

Yang, R. (2000), "Convergence of the simulated annealing algorithm for continuous global optimization," *Journal of optimization theory and applications*, 104(3), 691–716.

Yang, Y., and He, X. (2012), "Bayesian empirical likelihood for quantile regression," *The Annals of Statistics*, 40(2), 1102–1131.

Yin, X., Han, J., and Yu, P. (2008), "Truth discovery with multiple conflicting information providers on the web," *TKDE*, 20(6), 796–808.

Yu, D., Huang, H., Cassidy, T., Ji, H., Wang, C., Zhi, S., Han, J., Voss, C., and Magdon-Ismail, M. (2014), The Wisdom of Minority: Unsupervised Slot Filling Validation based on Multi-dimensional Truth-Finding,, in *COLING*, ACM.

Yu, J. (2004), "Empirical characteristic function estimation and its applications," *Econometric reviews*, 23(2), 93–123.

Yuan, A., Xu, J., and Zheng, G. (2014), "On empirical likelihood statistical functions," *Journal of Econometrics*, 178, 613–623.

Zhao, B., and Han, J. (2012), "A probabilistic model for estimating real-valued truth from conflicting sources," , .

Zhao, B., Rubinstein, B., Gemmell, J., and Han, J. (2012), "A Bayesian approach to discovering truth from conflicting sources for data integration," *PVLDB*, 5(6), 550–561.

Zhi, S., Zhao, B., Tong, W., Gao, J., Yu, D., Ji, H., and Han, J. (2015), Modeling truth existence in truth discovery,, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 1543–1552.