Estimation and Testing of the Jump Component in Lévy Processes

by

Zhaoxia Ren

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2013

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this thesis, a new method based on characteristic functions is proposed to estimate the jump component in a finite-activity Lévy process, which includes the jump frequency and the jump size distribution. Properties of the estimators are investigated, which show that this method does not require high frequency data. The implementation of the method is discussed, and examples are provided. We also perform a comparison which shows that our method has advantages over an existing threshold method. Finally, two applications are included: one is the classification of the increments of the model, and the other is the testing for a change of jump frequency.

Acknowledgement

First of all, I would like to express my sincere gratitude to my supervisor, Dr. Adam W. Kolkiewicz, for his inspiration, his guidance, his patience, and his great efforts in supervising me. Throughout my Ph.D. studies, he has provided constant support and encouragement. He is also a good friend who gave me suggestions in teaching, interview preparation and life itself.

I would like to thank the members of my thesis committee, Dr. Don L. McLeish, Dr. Tony S. Wirjanto, Dr. Ken Vetzal, and Dr. Hao Yu (from the University of Western Ontario) for their insightful comments and valuable advice.

I am also grateful to Dr. Phelim P. Boyle, Dr. Ken Seng Tan, Dr. Gord Willmot and Dr. Mu Zhu for the enlightening discussions during my PhD studies. My special thanks goes to Dr. Steve Drekic for his help when we were teaching the same undergraduate course, Mathematical Statistics.

It is a pleasure to acknowledge the financial support from the Department of Statistics and Actuarial Science, the Waterloo Research institute in Insurance, Securities and Quantitative finance (WatRISQ) and the University of Waterloo.

There are many colleagues within the department whom I wish to thank: Hua Shen, Michelle Zhou, Zhijian Chen, Zhenyu Cui, Pengfei Li, Chris Men, Reza Ramezan, Chengguo Weng, Longyang Wu, Yanqiao Zhang and many others. We accompanied each other on the comprehensive exams, courses, research discussions; we had a great time together and watched each other growing in those years. I also want to thank Ms. Mary Lou Dufton for her help throughout my Ph.D. studies.

Finally, I would like to thank my family for their endless love and support. To my parents, my brother and my husband, I dedicate this thesis.

Contents

List of Tables

List of Figures

List of Symbols

$L = \{L_t, t \geq 0\}$	Lévy process with time index t
$W = \{W_t, t \geq 0\}$	standard Brownian motion with time index t
$N = \{N_t, t \geq 0\}$	Poisson process with time index t
μ_0	drift term (of a diffusion process)
σ_0	volatility term (of a diffusion process)
Δ	length of time interval between two consecutive observations
λ	jump frequency
$\alpha \equiv \lambda\Delta$	jump ratio
φ_X	characteristic function of random variable X
n	number of observations
$\Re(\cdot)$	real part of a complex number or function
$\Im(\cdot)$	imaginary part of a complex number or function
$R(t) \equiv \Re(\varphi_X(t))$	real part of characteristic function of random variable X
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
t_{M1}	selection of t using Method 1
t_{M2}	selection of t using Method 2
t_{M3}	selection of t using Method 3

Chapter 1

Introduction

Lévy processes have been well-studied since long time ago, see Lévy (1954) and Itô (1961). In recent years there is a renaissance of interest in Lévy processes. Applications of Lévy processes exist in various fields, such as finance, economics, insurance, biology, queueing, telecommunications, quantum theory and many others, see Barndorff-Nielsen et al. (2001) for an overview.

A Lévy process consists of two parts: the diffusion component and the jump component. In finance, Lévy processes are popular models since they accommodate both continuous evolution and abrupt changes in the underlying assets. Compared with the famous Black-Scholes models which consist of geometric Brownian motions, Lévy processes allow heavy tails, excess kurtosis and asymmetry.

Not only the Lévy processes include a wide range of non-continuous model settings, they are also sufficiently mathematically tractable to permit estimators of parameters and distributions to be constructed relatively simply. This is a major attraction of the Lévy processes approach.

1.1 Lévy Processes

Loosely speaking, a Lévy process is a continuous-time stochastic process that has independent and stationary increments. It can be thought of as an analogue of a random walk in continuous time. In the following we first introduce the definition of Lévy processes, and then present an important characteristic representation of Lévy processes. Finally we discuss some well-known examples of Lévy processes.

Definition 1.1.1. Let $\mathbb{P}$ be a probability measure on $(\Omega, \mathcal{F})$. We say that L is a *Lévy process* for $(\Omega, \mathcal{F}, \mathbb{P})$ if it satisfies:

1. **Independent increments**: for every $s, t \geq 0$, $L_{t+s} - L_t$ is independent of the process $(L_v, 0 \leq v \leq t)$;
2. **Stationary increments**: $L_{t+s} - L_t$ has the same law as L_s;
3. **Stochastic continuity**: For each $\epsilon > 0, \lim_{h \to 0} P(|L_{t+h} - L_t| \geq \epsilon) = 0$.

Typically we assume $\mathbb{P}(L_0 = 0) = 1$. Every Lévy process can be characterized by its characteristic function. The following famous result, Lévy-Khintchine representation, implies many important properties of the Lévy processes.

Theorem 1.1.2 (Lévy-Khintchine)**.** *Given $a \in \mathbb{R}$, $\sigma \geq 0$ and a measure v on $\mathbb{R}\backslash\{0\}$ such that $\int (1 \wedge |x|^2) v(dx) < \infty$, for every $u \in \mathbb{R}$ let*

$$\psi(u) := iau - \frac{1}{2}\sigma^2 u^2 + \int_{\mathbb{R}\backslash\{0\}} \left(e^{iux} - 1 - iux\mathbf{1}_{|x|<1}\right) v(dx). \tag{1.1}$$

Then there exists a unique probability measure $\mathbb{P}$ on Ω under which L is a Lévy process with characteristic function $\varphi_{L_t}(u) := \mathbb{E}\left[e^{iuL_t}\right] = \exp\left\{t\psi(u)\right\}$. Moreover, the jump process of L, namely $\Delta L = (\Delta L_t, t \geq 0)$, is a Poisson point process with measure v.

The function ψ is called the *characteristic exponent* of the Lévy process L. The triplet $(a, \sigma^2, v(dx))$ is called the *Lévy triplet*. Specifically, the measure v is called the *Lévy measure*. It dictates how the jumps occur: jumps of size in the set A occur according to a Poisson process with intensity $\int_A v(dx)$. If the Lévy measure is of the form $v(dx) = \xi(x)dx$, we call $\xi(x)$ the *Lévy density*.

The formula (1.1) has another representation form, which was frequently used in early books that address the generalized central limit problem. As we know, the generalized central limit problem is related to the *infinitely divisible distributions*, which is in turn related to the Lévy processes. For example, in Loève (1963, page 298) or Breiman (1968, page 194), the function ψ was written as

$$\psi(u) = ibu + \int_{\mathbb{R}} \left(e^{iux} - 1 - \frac{iux}{1+x^2} \right) \frac{1+x^2}{x^2} G(dx), \tag{1.2}$$

where $b \in \mathbb{R}$, and G is a distribution function up to a multiplicative constant, with $G(0+) - G(0-) = \sigma^2$. The function G was called the *jump function*. Since the integrand in (1.2) has a limit of $-u^2/2$ as $x \to 0$, we may also write

$$\psi(u) = ibu - \frac{1}{2}\sigma^2 u^2 + \int_{\mathbb{R}\setminus\{0\}} \left(e^{iux} - 1 - \frac{iux}{1+x^2} \right) \frac{1+x^2}{x^2} G(dx). \tag{1.3}$$

Let us compare the two formulas in (1.1) and (1.3). The term $\frac{1+x^2}{x^2}G(x)$ in (1.3) has the same property as the measure $v(x)$ in (1.1) does that $\int (1 \wedge |x|^2) v(dx) < \infty$. The term $\frac{iux}{1+x^2}$ in (1.3) plays the role of avoiding the integrand in (1.3) exploding when $x \to 0$, and the term $iux\mathbf{1}_{|x|<1}$ in (1.1) plays the same role. The difference between a and b reflects the difference between the two integrals in (1.1) and (1.3).

Examples of Lévy Processes

The most well-known examples of Lévy processes are the *Brownian motion* and the *Poisson processes*, which are the building blocks of Lévy processes. In the following we present them and some other important examples.

1. Brownian motion with drift

A *standard Brownian motion or Wiener process* W_t is a Lévy process whose increments follow normal distributions. More specifically, $W_t - W_s \sim \mathcal{N}(0, t-s)$.

A *Brownian motion with drift* can be written as

$$L_t = \mu t + \sigma W_t$$

where $\mu \in \mathbb{R}$, $\sigma > 0$ and W_t is a standard Brownian motion. The characteristic exponent of N_t is $\psi(u) = i\mu u - \frac{1}{2}\sigma^2 u^2$. Then, by Theorem 1.1.2, we can see that the corresponding Lévy triplet is $(\mu, \sigma^2, 0)$. Thus, this process is the only continuous Lévy process.

2. Poisson processes

A *Poisson process* N_t with *intensity* λ is a Lévy process whose increments follow Poisson distributions. More specifically, $N_t - N_s \sim Poisson(\lambda(t-s))$.

The characteristic exponent of N_t is $\psi(u) = \lambda(e^{iu} - 1)$, from which we can see that the Lévy triplet is $(0, 0, \lambda\delta(1))$, where $\delta(1)$ denotes the Dirac measure at point 1. This means that the jump size always equals one. A Poisson process is an increasing pure jump Lévy process.

3. Compound Poisson processes

A *compound Poisson process* can be written as

$$L_t = \sum_{i=1}^{N_t} X_i \tag{1.4}$$

where N_t is a Poisson process with intensity λ, $\{X_i, i = 1, 2, ...\}$ is a sequence of independently and identically distributed random variables representing jump sizes with distribution $\mathbb{P}$, and N_t and $\{X_i\}_{i \geq 1}$ are independent.

Define $\nu(A) = \lambda \cdot \mathbb{P}(X \in A)$, then it can be shown that the characteristic function of L_t is

$$E[e^{iuL_t}] = \exp\left(t \int_{-\infty}^{+\infty} (e^{iux} - 1)\nu(dx)\right).$$

Therefore, the Lévy triplet is $\left(\int_{-1}^{1} x\nu(dx), 0, \nu(dx)\right)$. If X has a density function $f_X(x)$, i.e. $\mathbb{P}(dx) = f_X(x)dx$, then the Lévy density of L_t is $\xi(x) = \lambda f_X(x)$.

4. Jump-diffusion processes

Combining the Brownian motion (with drift) and the compound Poisson process, we obtain the *jump-diffusion models* as follows:

$$L_t = \mu t + \sigma W_t + \sum_{i=1}^{N_t} X_i \tag{1.5}$$

where $\mu t + \sigma W_t$ is the diffusion component, $\sum_{i=1}^{N_t} X_i$ is the jump component, and $(W_t, N_t, X_i's)$ are mutually independent.

Two of the most widely used jump-diffusion models are the Merton model and the Kou model:

- *Merton model*: The jump size follows a normal distribution, i.e.

  $$X_i \sim N(\mu_X, \sigma_X^2).$$

- *Kou model*: The jump size follows a double exponential distribution, i.e. the probability density function of X_i is of the form

  $$f(x) = p\eta_1 e^{-\eta_1 x} 1_{\{x>0\}} + (1-p)\eta_2 e^{-\eta_2|x|} 1_{\{x<0\}} \tag{1.6}$$

  where $\eta_1 > 1, \eta_2 > 0$ and $p > 0$. Note that the requirement that $\eta_1 > 1$ is to ensure that $\mathbb{E}\left[e^{L_t}\right] < \infty$ since e^{L_t} is typically used to model the stock price. It essentially means that the average upward jump cannot exceed 100%, which is a reasonable requirement.

5. Infinite-activity Lévy processes

In a jump-diffusion process, jumps are considered as rare events, in the sense that within any given finite interval there are only finite number of jumps. For *infinite activity* Lévy processes, there are infinitely many jumps in any finite time interval. Many of these model can be constructed via Brownian subordination, i.e. time-changed Brownian motion. The following are two rich classes of this type of models.

1. Generalized hyperbolic model: The increments of this process follows a generalised hyperbolic (GH) distribution. The GH distribution has five parameters $(\alpha, \beta, \lambda, \delta, \mu)$ with the probability density given by

$$f_{GH}(x) = C(\delta^2 + (x-u)^2)^{\frac{\lambda}{2}-\frac{1}{4}} K_{\lambda-\frac{1}{2}}(\alpha\sqrt{\delta^2+(x-\mu)^2}) e^{\beta(x-\mu)}$$

   where C is a known function of $(\alpha, \beta, \lambda, \delta)$ and K is the modified Bessel function of the second kind. It includes the following well-known distributions:

   - Normal distribution: $\delta \to \infty$ and $\delta/\alpha \to \sigma^2$.
   - Normal inverse Gaussian: $\lambda = -1/2$.
   - Variance gamma: $\delta = 0$ and $\mu = 0$.

2. Tempered stable processes: It is obtained by modifying the Lévy measure for the α-stable process.

In the financial area, the testing results in a number of articles indicate that jumps are necessary to be included for modelling the data, see Barndorff-Nielsen and Shephard (2006), Aït-Sahalia and Jacod (2009), among others. Aït-Sahalia and Jacod (2011) further proposed tests to discriminate between the finite-activity and infinite-activity jumps, while allowing the presence of a continuous component in

both cases. The testing results on the stock returns indicate the presence of infinite-activity jumps. Aït-Sahalia and Jacod (2010) also discussed the problem that, while allowing the presence of the infinite-activity jumps, whether the Brownian motion is necessary to be included for modelling the high frequency data. When applied to individual stock data, the test results point toward the necessity to include the Brownian motion.

In this thesis, we mainly consider the estimation of the finite-activity Lévy process, equivalently, the jump-diffusion model (1.5). The estimation of the infinite-activity Lévy processes will be mentioned in Chapter 8. For the jump-diffusion processes, the diffusion parameters μ, σ, the jump intensity λ and the distribution of the jump size are all assumed unknown. The only available information is a discrete record of observations of the process. We will use these observations to estimate all the components of the jump-diffusion model.

1.2 Nonparametric Estimation of Lévy Processes

The work on the estimation of Lévy process can go back to Rubin and Tucker (1959), where the parameter b and the jump function G in (1.2) were consistently estimated based on data virtually observed in the continuum. Basawa and Brockwell (1982) proposed three estimators of the jump function G of non-decreasing pure jump Lévy processes, in the case that the jump sizes are assumed directly observable.

In the following we consider the cases when processes are observed at discrete time points only. Suppose we have a discrete record of equidistant observations of a Lévy process L_t, denoted by $\{L_0, L_\Delta, L_{2\Delta}, ..., L_{n\Delta}\}$, where Δ denotes the *data frequency*. Let the *time horizon* be $T = n\Delta$. The increments are given by

$$Y_j \equiv L_{j\Delta} - L_{(j-1)\Delta}, \quad j = 1, 2, ..., n.$$

We first review the methods proposed in a **high frequency** framework, and then look at the methods proposed when the data frequency is fixed.

Threshold method

The threshold method was discussed in Mancini (2004) and Gegler and Stadtmüller (2010). The idea is that we define a threshold and whenever a movement is larger than the threshold we classify it as a jump.

In Mancini (2004), a generalized jump-diffusion model was considered of the form

$$dL_t = a_t dt + \sigma_t dW_t + \gamma_t dN_t, \quad t > 0 \tag{1.7}$$

where $|a_t| \leq M, |\sigma_t| \leq M$ for some $M > 0$, and $\gamma_t \geq \Gamma$ for some $\Gamma > 0$. That is, the diffusion coefficients are bounded from above, and the jump coefficient is bounded away from zero.

The contribution of the diffusion part to the increments $\{Y_j, j = 1, 2..., n\}$ tends to zero quickly as Δ decreases, because the stochastic integral $\int \sigma_t dW_t$, after a change of time, behaves as a Brownian motion. The Lévy's modulus of continuity theorem for the path of a Brownian motion asserts that the rate $\sqrt{2\Delta \log(1/\Delta)}$ measures the speed at which the increment of a Brownian motion over a time step Δ goes to zero. Based on this result, the author proposed to let $r(\Delta) = \sqrt{8M^2\beta} \cdot \sqrt{\Delta \log(1/\Delta)}$ for some $\beta \in (1, 2]$ (or let $r(\Delta)$ be any other function of Δ which goes to zero more slowly), and conclude that a jump occurs if $|Y_j| > r(\Delta)$.

Then the author constructed consistent estimators of $\{N_{j\Delta}, j = 1, ..., n\}$, of the jump intensity λ and of the size of jump γ_{τ_j}, where τ_j are the instants of jump within the time interval $[0, T]$. The estimator of λ is asymptotically Gaussian. These results hold when the time horizon $T \to \infty$, the data frequency $\Delta \to 0$, such that $n\Delta^\beta \to 0$ for $\beta \in (1, 2]$.

Since the method requires the condition that the jump size is bounded away from zero, many widely-used financial models cannot be applied, such as the Merton

model and the Kou model.

The threshold method was also discussed in Gegler and Stadtmüller (2010), where a general Lévy process was considered. The authors constructed consistent estimators of the diffusion parameters and of the Lévy measure for both the finite-activity case and the infinite-activity case. In the finite activity case, the estimators are shown to be asymptotically normally distributed. These results hold when $T \to \infty$ and $\Delta \to 0$ such that $T\Delta \to 0$. The implementation of this method was also included in the paper. For more details, see Chapter 6, where we compare our method with the threshold method. The threshold method was later extended to multivariate Lévy process in Gegler (2011).

Sieve method

A *sieve* method, or a *penalized projection* method, was used to propose estimators of the Lévy density; see Figueroa-López and Houdrél (2006, 2008), Figueroa-Lopéz (2009). An orthogonal projection of Lévy density ξ onto some space $\mathcal{S} := \{\beta_1\phi_1 + ... + \beta_d\phi_d : \beta_1, ..., \beta_d \in \mathbb{R}\}$ is given by

$$\xi^{\perp} := \sum_{j=1}^{d} a(\phi_j)\phi_j(x)$$

where the coefficients are $a(\phi_j) := \int \phi_j(x)\xi(x)dx$. Using the fact that, if ϕ is ν-continuous, bounded and vanishing in a neighborhood of the origin, then $\int \phi(x)\xi(x)dx = \lim_{\Delta\to 0} \frac{1}{\Delta}\mathbb{E}[\phi(L_\Delta)]$, the coefficients are estimated by

$$\hat{a}(\phi_j) = \frac{1}{T}\sum_{k=1}^{n} \phi_j(L_{k\Delta} - L_{(k-1)\Delta}).$$

Based on this property, estimators of the Lévy density were constructed. This method works when $\Delta \to 0$ and $T \to \infty$.

Estimation of a function of the Lévy density

In Comte and Genon-Catalot (2009), a sub-class of the Lévy processes is considered, which is virtually a compound Poisson process given by (1.4) with the

requirements that $\int_{\mathbb{R}} |x| v(dx) < \infty$ and the Levy density ξ exists. The goal is to estimate the function

$$g(x) := x \cdot \xi(x).$$

The idea is to use the Fourier transform of the function g which is given by

$$g^{\text{ft}}(u) := \int e^{iux} g(x) dx = -i \frac{\varphi'_{L_\Delta}(u)}{\Delta \varphi_{L_\Delta}(u)} \approx \frac{1}{\Delta} \mathbb{E}(Y_j e^{iuY_j}),$$

where Y_j is the increment of a Levy process over time step Δ. An empirical version of $g^{\text{ft}}(u)$ is defined as

$$\hat{g}^{\text{ft}}(u) = \frac{1}{n\Delta} \sum_{j=1}^{n} Y_j e^{iuY_j}.$$

Then two methods were proposed to recover the function g. One is to use the Fourier inversion, where a cutoff parameter is adaptively selected. This approach is similar to our method for recovery of the jump size distribution, which will be discussed in Chapter 5 of this thesis. The other one is to apply the penalized projection method. The asymptotic results are derived when the data frequency $\Delta \to 0$ and the time horizon $T \to \infty$.

In this method, it is a function (g) of the Lévy density that was estimated. After an estimate of the function g is obtained, we have to transform it back to obtain an estimate of the Lévy density.

Kernel method

In Shimizu (2006), a kernel based estimator of the Lévy density was proposed for the case of finite activity, where the threshold idea was also applied. This estimator is rate optimal under certain smoothness assumptions. The method requires that the time horizon $T \to \infty$ and the data frequency $\Delta \to 0$.

Now we look at some existing methods in the **low frequency** framework. In this case, the problem of estimating the Lévy measure is closely related to a deconvolution problem.

Minimum-distance estimator based on ECF

In Neumann and Reiß (2009), the volatility parameter σ and the Lévy measure ν in the Lévy-Khintchine formula (1.1) for a general Lévy process are merged into a single quantity ν_σ:

$$\nu_\sigma := \sigma^2 \delta_0(dx) + \frac{x^2}{1+x^2}\nu(dx),$$

where δ_0 denotes the point measure in zero. Note that this ν_σ is exactly the same measure as $G(x)$ in formula (1.2). Estimators of the drift parameter b and the measure ν_σ are proposed based on the minimum-distance criterion:

$$(\hat{b}, \hat{\nu_\sigma}) := \arg\inf_{b,\nu_\sigma} d(\hat{\varphi}_n, \varphi(\cdot\,; b, \nu_\sigma)),$$

where φ is the characteristic function of the increments, $\hat{\varphi}_n$ is the empirical characteristic function of the increments, and d is an appropriately selected metric.

It was shown that for equidistant observations, when time horizon $T \to \infty$ and the data frequency Δ being fixed, the estimators of b and ν_σ are strongly consistent under some conditions on the choice of the metric d. Thus, this method works for low frequency data.

Later, Kappus and Reiß (2010) extended this approach to an arbitrary data frequency, where an asymptotic upper bound for the estimator of the jump measure was obtained with the assumption that $\int_{\mathbb{R}\setminus\{0\}} x^2 \nu(dx) < \infty$.

However, the combined form of ν_σ indicates that the volatility σ and the Lévy measure ν cannot be estimated separately using this minimum-distance method.

Regression based method

A regression based method was proposed by Chen et al. (2010) to estimate the Lévy characteristics of a sub-class of Lévy processes, which is modelled as a sum of a drift, a symmetric stable process and a compound Poisson process. When the index parameter of the stable law was assumed known, a regression based method was used to estimate the drift, the scale of the stable process and the jump intensity.

A nonparametric deconvolution approach was used to estimate the distribution of the jump size. Later the index parameter of the stable law was estimated by some existing method called the Hill estimator.

A lease square criterion was performed for the regression method. For this, an interval of t over which the integrals of the squared residuals are taken has to be selected. This selection represents a trade-off between the error from using empirical characteristic function to approximate the characteristic function and the error from assuming the characteristic function of the jump size being zero when t is outside of the selected region. The consistency of the estimators was shown under the asymptotics that the time horizon T increases while the data frequency Δ is fixed.

This method has similarities to our method. But our method in selecting t is more objective and the entire estimation procedure has less computational burden.

Fourier inversion and Kernel smoothing

In Gugushvili (2009), the method based on Fourier inversion and Kernel smoothing was applied to estimate the components of a jump-diffusion model, which includes the diffusion parameters, the jump frequency and the Lévy density. Different kernels with different restrictions were used in the construction of the estimators of different parameters or distribution. For the estimation of the Lévy density, the Fourier inversion technique is applied. This method works when the data frequency is fixed. The asymptotic result is based on the sample size $n \to \infty$, which makes the window size go to zero. The implementation of this method was not discussed in the paper.

In this thesis, we propose a new method based on the characteristic function approach to estimate the jump component of a finite-activity Lévy process, when the data frequency is fixed. It includes the estimation of the jump frequency and the estimation of the jump size distribution, and thus we face a semi-parametric

estimation problem. Since the sizes of the jumps are typically much larger on average than the sizes of the diffusion increments, the characteristic function of an increment with jump converges to zero much faster than the characteristic function of the increment without jump does. This difference in the convergence rates of the characteristic functions between different components of the model is the main driver of our method. The simulation results show that our method has many advantages over the existing threshold method discussed above.

1.3 Testing for Jumps

After obtaining estimates for all the components of the finite-activity Lévy process, we consider the classification problem that, given an increment of the process, we classify it as a jump or non-jump. This problem is related to the testing of jumps. Different methods have been proposed in the literature to test the jumps.

A non-parametric statistic which tests whether a jump has occurred or not was proposed by Aït-Sahalia and Jacod (2007):

$$S_n = \frac{\sum_{i=1}^{[n/K]} |Y_{iK\Delta} - Y_{(i-1)K\Delta}|^p}{\sum_{i=1}^{n} |Y_{i\Delta} - Y_{(i-1)\Delta}|^p}.$$

It converges to two different deterministic constants depending on whether the process has jumps or not. The test is valid for all Itô semi-martingales. In this paper the authors discuss the testing of jumps only. Later the detection of jumps based on the same idea was extended by Fan and Fan (2011).

Jiang and Oomen (2005) propose a test statistic that measures the impact of jumps on the third and higher order return moments. Barndorff-Nielsen and Shephard (2006) introduce a test statistic based on the bipower variation of the asset price, which is consistent and asymptotically normal with mean zero under the null hypothesis of no jumps. Lee and Mykland (2007) propose a non-parametric test

not only to detect the presence of jump but also to give estimates of the realized jump sizes.

In this thesis, we use a Bayes procedure to classify the observations with or without jump. The theoretical results show that the misclassification probability decreases to zero when the data frequency increases.

1.4 Outline of the Thesis

In Chapter 2, we discuss the estimation of the jump frequency of a jump-diffusion model, assuming that the parameters in the diffusion component are known and the data frequency is fixed. A new estimator of jump frequency based on characteristic functions is proposed. The bias and variance analysis of this estimator is performed. The consistency and asymptotic normality are proved, which shows that our method works for low-frequency data. The selection of the tuning parameter t for the finite sample case is also discussed.

In Chapter 3, we investigate the properties of the proposed estimator of jump frequency when the data frequency increase, assuming that the parameters in the diffusion component are known. We also compare our method with the maximum likelihood method when the data frequency increases.

In Chapter 4, we discuss the estimation of the jump frequency when the diffusion parameters are unknown. Firstly, we use robust procedures to obtain robust estimates of the diffusion parameters. Then we use the same method proposed in previous chapters to estimate the jump frequency.

In Chapter 5, we propose an estimator of the density function of jump size. This is a de-convolution problem, for which we need to use truncation as a regularization step to ensure that the Fourier inversion exists. The truncation point is the same as the selected value of the tuning parameter t used for the estimation of the jump

frequency. The consistency of the density estimator is investigated. Simulation studies and real data examples are also provided.

In Chapter 6, we compare our method with an existing threshold method. When the variance of jump size distribution is large (relative to the volatility in the diffusion), the threshold method works well. However, when the variance of jump size distribution is not too large, our method performs better. Moreover, when the sample size increases, our method outperforms the threshold method even when the variance of jump size distribution is large.

In Chapter 7, two applications are discussed. One is the classification of the increments, based on a Bayes procedure. The other is the testing for a change of jump frequency, based on the cusum method. Simulation studies for both topics are provided.

In Chapter 8, we discuss some future work, including the extension of our method to the infinite-activity Lévy processes.

Chapter 2

Estimation of the Jump Frequency when the Diffusion Parameters are Known

The jump-diffusion models are widely used in finance and insurance. We propose a new method to estimate the jump components of these models, which includes the jump frequency and the jump size distribution. In this chapter, we consider the estimation of the jump frequency. To separate the sources of the error, we assume that the two parameters in the diffusion component, i.e. the drift and the volatility, are known in this chapter. This assumption will be removed in Chapter 4. In this chapter the frequency with which we observe the data, which we call *data frequency*, is fixed. The high frequency case will be discussed in Chapter 3.

The proposed method is based on a characteristic function approach. Consider a discretely observed realization of a jump-diffusion model with the length of the time interval fixed. Then each increment may or may not involve a jump (or jumps). If there is no jump, then it is simply an increment of the diffusion component. Since

on average the jump size is typically much larger than the increment of the diffusion component, the increment of the model with a jump (or jumps) would correspond to a distribution with much heavier tails than the increment of the model without a jump does. Using characteristic function techniques, we transform the distribution with heavier tails to a characteristic function converging to zero at a faster rate. By capturing this difference in the convergence rates of the characteristic functions, we construct a new estimator.

We show that, when the intensity of jumps is low, considering the jump-diffusion model is equivalent to considering a two-component mixture model. Then the estimation of the jump-diffusion model is equivalent to the estimation of the different components of the mixture model. In this chapter, the data frequency is fixed, so all the results in this chapter are also applicable in the context of mixture models. Most of the literature on mixture models considers a parametric setting, i.e. the distributions of the components are known up to a finite number of parameters, and then the goal is to estimate the number of components, the mixing proportions and the component parameters. Even in the case when the nonparametric setting is considered for mixture models, all components of the model usually belong to the same family of distributions, i.e. $F(x) = \sum_{j=1}^{k} \alpha_j G(x - \mu_j)$ with G unknown. However, in our problem, the distribution of jump sizes is completely unknown, and of course it could be different from the distribution in the diffusion component, which is a normal distribution under our assumptions.

There are two different types of asymptotics that we may consider in our estimation problem. The one that we discuss in this chapter is a "typical" one, since it corresponds to adding more observations coming from the same distribution. This approach has the following features:

- The sample size increases by increasing the time horizon (i.e. the total number of years) of the data set while keeping the data frequency fixed.

- The distribution of the sample remains the same.
- Since the data frequency is fixed, the jump frequency can be explained by the *jump ratio*, i.e. the expected percentage of jump observations among all observations.

Another type of asymptotics is discussed in Chapter 3 and has the following features:

- The sample size increases by increasing the data frequency while keeping the the time horizon fixed.
- The distribution of the sample changes when the sample size increases. In particular, the jump ratio decreases when the sample size increases. This is because the jump intensity λ is fixed, which implies that the expected total number of jumps in a given time horizon is fixed, but meanwhile the total number of observations increases.
- In this case, we consider the estimation of the jump intensity λ, which is fixed. We investigate some properties of this estimator and compare it with the maximum likelihood estimator.

It is possible to consider another type of asymptotics, where we let both the data frequency and the time horizon increase but control the speed of the increase for each. However, it will not be discussed in this thesis.

In the title of this chapter, we used the term "jump frequency". In the first type of asymptotics, "jump frequency" is simply equivalent to "jump ratio"(i.e. the expected proportion of observations with jumps among all observations). In the second type of asymptotics, by "jump frequency" we mean the "jump intensity" which is the intensity parameter λ of the Poisson process.

This chapter is organized as follows. In Section 2.1, the proposed method is introduced, where the characteristic function of the observable variable is estimated by the empirical characteristic function. Properties of the proposed estimator are discussed in Section 2.2, including the bias and variance analysis, consistency and asymptotic normality of the estimator. The proposed estimator involves a tuning parameter t, whose proper selection is essential to guarantee the good performance of the estimator. By considering the asymptotic behavior of the proposed estimator, we derive the rate at which t should increase to infinity as the sample size n increases. However, in practice for finite samples, we have to propose methods of selecting t. This will be discussed in Section 2.3 where three methods are proposed. In Section 2.4 we consider alternative estimators (other than the empirical characteristic function) of the characteristic function which are uniformly consistent. Our analysis shows that, however, the method based on the empirical characteristic function leads to the most efficient estimation procedure.

2.1 Proposed Method

In this section we introduce the proposed method. In Section 2.1.1 we specify a jump-diffusion model and in Section 2.1.2 we present the main idea that is behind our new estimator. A two-component mixture model is introduced in Section 2.1.3, which provides a good approximation to the jump-diffusion model. We specify the form of the proposed estimator in Section 2.1.4. Finally the identifiability problem for the mixture model is discussed in Section 2.1.5.

2.1.1 Model Settings

Suppose we have a discretely observed realization $\{L_\Delta, L_{2\Delta}, ..., L_{n\Delta}\}$, where Δ is the length of the time interval between two consecutive observations, and $T = n\Delta$

is the time horizon, from the following jump-diffusion model (representing the log-price, $\ln S_t$), which we call **Model 1**:

$$L_t \equiv \ln S_t = \mu_0 t + \sigma_0 W_t + \sum_{i=1}^{N_t} X_i, \quad t \geq 0^* \tag{2.1}$$

where $L_0 = 0$, $\{W_t\}$ is a standard Brownian motion, $\{N_t\}$ is a Poisson process with an intensity parameter λ, $X_1, X_2, ...$ are independently and identically distributed (i.i.d.) random variables representing jump sizes, and (W_t, N_t, X_i) are mutually independent. This process L_t consists of two components: the first one is the diffusion component $\mu_0 t + \sigma_0 W_t$, and the second one is the jump component $\sum_{i=1}^{N_t} X_i$.

In the remainder of the thesis, we will impose the following assumptions concerning Model 1:

(A1-1) The distribution of the jump size X is absolutely continuous with respect to the Lebesgue measure.

(A1-2) The product $\lambda\Delta$ is small so that the term of order $o(\lambda\Delta)$ in the expansion $e^{-\lambda\Delta} = 1 - \lambda\Delta + o(\lambda\Delta)$ is negligible.

The assumption (A1-1) is equivalent to the one that the distribution of jump size X has a density function with respect to the Lebesgue measure. This is the case for almost all financial applications where the jump-diffusion model is used. Particular examples for jump size distribution include the normal distribution and the double exponential distribution.

With the rapidly increasing computing power, the data set we work with in the financial area can be daily or higher frequency. To justify the assumption (A1-2), let us consider the following example: when $\Delta = 1/250$ (daily data) and $\lambda = 12.5^\dagger$,

*This is equivalent to $dL_t = d\ln S_t = \mu_0 dt + \sigma_0 dW_t + X dN_t$. By Ito's formula for seminartingales, this is also equivalent to $\frac{dS_t}{S_{t-}} = \left(\mu_0 + \frac{\sigma_0^2}{2}\right) dt + \sigma_0 dW_t + (e^X - 1) dN_t$.

†In Kou (2002), $\lambda = 10$ was used. Here we use $\lambda = 12.5$ for convenience.

then $\lambda\Delta = 5\%$, which gives us $e^{-\lambda\Delta} = 0.9512$ and $1 - \lambda\Delta = 0.9500$. The difference between them is 0.0012, which is small. Thus the assumption (A1-2) seems to be reasonable.

We are interested in the estimation of the jump component, including the jump frequency and the jump size distribution. However, in practice, the parameters in the diffusion component are also unknown. To separate the error due to the estimation of the diffusion parameters from the error due to the estimation of the jump component, in Chapters 2 and 3 we adopt the following assumption:

(A1-3) μ_0 and σ_0 are known.

This assumption will be removed in Chapter 4 and the subsequent chapters.

2.1.2 Methodology

The estimators of jump frequency and jump size distribution that we propose are based on characteristic functions. The *characteristic function* (c.f.) of a random variable X is defined by

$$\varphi_X(t) := \mathbb{E}[e^{itX}], \quad t \in \mathbb{R}.$$

Its real and imaginary parts are denoted by

$$\Re(\varphi_X(t)) = \mathbb{E}[\cos(tX)] = \int_{\mathbb{R}} \cos(tx) \cdot f_X(x)dx,$$

and

$$\Im(\varphi_X(t)) = \mathbb{E}[\sin(tX)] = \int_{\mathbb{R}} \sin(tx) \cdot f_X(x)dx,$$

respectively, where f_X is the pdf of the random variable X. Note that the notation t in this thesis may have two meanings: one is the argument of a characteristic function, and the other is the time index of a stochastic process. It should be easy to distinguish them by the context.

The following lemma describes the tail behavior of the characteristic function of an absolutely continuous random variable. This result can be found in Lukacs (1970, page 19). It can also be proved directly by the Riemann-Lebesgue Lemma.

Lemma 2.1.1. *When a random variable is absolutely continuous, its characteristic function satisfies* $\lim_{|t|\to\infty} \varphi(t) \to 0$.

In the following we will find the characteristic function of increments of the process L_t in (2.1). Let us denote the increments of L_t by

$$\begin{aligned} Y_j &:= L_{j\Delta} - L_{(j-1)\Delta} \\ &= \mu_0\Delta + \sigma_0\left(W_{j\Delta} - W_{(j-1)\Delta}\right) + \sum_{k=N_{(j-1)\Delta}+1}^{N_{j\Delta}} X_k, \quad j = 1, 2, ..., n. \end{aligned}$$

Then the variables $\{Y_j, j = 1, 2, ..., n\}$ are i.i.d. Denote the increments of the diffusion component by

$$Z_j \equiv \mu_0\Delta + \sigma_0\left(W_{j\Delta} - W_{(j-1)\Delta}\right), \quad j = 1, 2, ..., n. \tag{2.2}$$

Then $\{Z_j, j = 1, 2, ..., n\}$ are i.i.d. and $Z_j \sim \mathcal{N}(\mu_0\Delta, \sigma_0^2\Delta)$. Let Y be a random variable that has the same distribution as the increments Y_j's, and Z be a random variable with the same distribution as Z_j's. Since $Y_j = Z_j + \sum_{k=N_{(j-1)\Delta}+1}^{N_{j\Delta}} X_k$, the characteristic function of Y is given by

$$\begin{aligned} \varphi_Y(t) &= \mathbb{E}[\exp\{itY_j\}] \\ &= \mathbb{E}\left[\exp\left\{it\left(Z_j + \sum_{k=N_{(j-1)\Delta}+1}^{N_{j\Delta}} X_k\right)\right\}\right] \\ &= \mathbb{E}\left[\exp\left\{it\left(Z_j + \sum_{k=1}^{N_\Delta} X_k\right)\right\}\right] \\ &= \varphi_Z(t) \cdot \mathbb{E}\left[e^{it\sum_{k=1}^{N_\Delta} X_k}\right] \end{aligned} \tag{2.3}$$

where $\varphi_Z(t) = e^{it\mu_0\Delta - t^2\sigma_0^2\Delta/2}$ is the characteristic function of Z, the third line is due to the fact that $\sum_{k=N_{(j-1)\Delta}+1}^{N_{j\Delta}} X_k$ and $\sum_{k=1}^{N_\Delta} X_k$ have the same distribution, and in the last line the independence between Z_j and (N_t, X) is used.

By conditioning on the number of jumps occurring in the time interval $[0, \Delta]$, we have

$$\begin{aligned}
\mathbb{E}\left[e^{it\sum_{k=1}^{N_\Delta} X_k}\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{it\sum_{k=1}^{N_\Delta} X_k} | N_\Delta\right]\right] \\
&= \sum_{l=0}^{\infty} \mathbb{E}\left[e^{it\sum_{k=1}^{l} X_k}\right] \cdot \mathbb{P}(N_\Delta = l) \\
&= \sum_{l=0}^{\infty} \left(\varphi_X(t)\right)^l \cdot \frac{(\lambda\Delta)^l}{l!} e^{-\lambda\Delta} \\
&= e^{-\lambda\Delta(1-\varphi_X(t))}.
\end{aligned}$$

Therefore, by (2.3), we obtain the characteristic function of Y as

$$\varphi_Y(t) = \varphi_Z(t) \cdot e^{-\lambda\Delta(1-\varphi_X(t))}. \tag{2.4}$$

Due to the assumption (A1-2) and the fact that $\|1 - \varphi_X(t)\| \leq 2$, we have

$$e^{-\lambda\Delta(1-\varphi_X(t))} = 1 - \lambda\Delta(1 - \varphi_X(t)) + o(\lambda\Delta),$$

and the error term $o(\lambda\Delta)$ is negligible.

Let us denote

$$\alpha \equiv \alpha(\Delta) := \lambda\Delta. \tag{2.5}$$

Note that

$$\mathbb{P}(N_\Delta \geq 1) = 1 - \mathbb{P}(N_\Delta = 0) = 1 - e^{-\lambda\Delta} = 1 - (1 - \lambda\Delta + o(\lambda\Delta)) = \lambda\Delta + o(\lambda\Delta),$$

that is, under the assumption (A1-2), α provides a good approximation of the probability of at least one jump occurring during a time step of length Δ. Thus, we may call α the *jump ratio*, since it approximates the expected proportion of increments with jumps.

Now, if we ignore the term $o(\lambda\Delta)$, we can write (2.4) as

$$\varphi_Y(t) = \varphi_Z(t) \cdot (1 - \alpha + \alpha\varphi_X(t)). \tag{2.6}$$

The consequences of replacing (2.4) by (2.6) will be discussed in Section 2.1.3. Representation (2.6) is an important formula, which we will use in the following to derive estimators of the jump frequency and the jump size distribution.

Dividing both sides of (2.6) by $\varphi_Z(t)$, we get

$$\frac{\varphi_Y(t)}{\varphi_Z(t)} = 1 - \alpha + \alpha\varphi_X(t). \tag{2.7}$$

By the assumption (A1-1) and Lemma 2.1.1, we have

$$\lim_{t\to\infty} \varphi_X(t) = 0.$$

Then, by taking the limits of both sides of (2.7), we get

$$\alpha = 1 - \lim_{t\to\infty} \frac{\varphi_Y(t)}{\varphi_Z(t)}. \tag{2.8}$$

Since α is a real number, we may take the real part of the right-hand side to obtain

$$\alpha = 1 - \lim_{t\to\infty} \Re\left(\frac{\varphi_Y(t)}{\varphi_Z(t)}\right). \tag{2.9}$$

To characterize the distribution of X, we use equation (2.6) again to obtain

$$\varphi_X(t) = \frac{\frac{\varphi_Y(t)}{\varphi_Z(t)} - (1 - \alpha)}{\alpha}. \tag{2.10}$$

The two equations (2.9) and (2.10) form the basis of the method we propose in the thesis to estimate the jump frequency and the jump size distribution.

2.1.3 An Equivalent Model

In this section we explain that equation (2.6) corresponds to a two-component mixture model, and when we replace (2.4) by (2.6) we in fact approximate the jump-diffusion model, i.e. Model 1 described in (2.1), by this mixture model.

Suppose we have a random sample, $\{Y_1, Y_2, ..., Y_n\}$, from the following model, which we call **Model 2**:

$$Y = (1 - I)Z + I(Z + X) \tag{2.11}$$

where Z follows a normal distribution, $Z \sim \mathcal{N}(\mu_0\Delta, \sigma_0^2\Delta)$; I is a Bernoulli random variable with $\mathbb{P}(I = 1) = \lambda\Delta = \alpha$; X represents the jump size with an unknown distribution function; and (Z, X, I) are mutually independent. This model corresponds to a two-component mixture model, where one component follows a normal distribution and the other component follows the convolution of the normal distribution and the jump size distribution. We assume that Model 2 satisfies:

(A2-1) The distribution function of X is absolutely continuous.

(A2-2) μ_0 and σ_0 are known.

In the following we show that Model 2 with assumption (A2-1) approximates Model 1 with assumptions (A1-1) and (A1-2).

Let us observe that the mixture model (2.11) can be rewritten as

$$Y = Z + V,$$

where $V := IX$. Since Z is independent of (I, X), Z is independent of V. Thus

$$\varphi_Y(t) = \mathbb{E}[e^{itY}] = \mathbb{E}[e^{it(Z+V)}] = \mathbb{E}[e^{itZ}] \cdot \mathbb{E}[e^{itV}] = \varphi_Z(t) \cdot \varphi_V(t), \tag{2.12}$$

where

$$\begin{aligned}\varphi_V(t) = \mathbb{E}[e^{itV}] &= \mathbb{E}[e^{itIX}] \\ &= 1 \cdot \mathbb{P}(I = 0) + \mathbb{E}[e^{itX}] \cdot \mathbb{P}(I = 1) \\ &= (1 - \alpha) + \alpha\varphi_X(t).\end{aligned}$$

Therefore,

$$\varphi_Y(t) = \varphi_Z(t) \cdot (1 - \alpha + \alpha\varphi_X(t)).$$

This is the same equation as (2.6) for Model 1, which was obtained by ignoring the error term, $o(\lambda\Delta)$.

Therefore, the distribution of Y in the mixture model (i.e. Model 2) approximates the distribution of the increments Y in the jump-diffusion model (i.e. Model 1). More specifically, Model 2 with assumption (A2-1) approximates Model 1 with assumptions (A1-1) and (A1-2).

In the remainder of the thesis, we consider a sequence of independent random variables $Y_1, Y_2, ..., Y_n, ...$ following the distribution with the c.f. given by (2.6). Note that this distribution describes Model 2 exactly, but Model 1 only approximately.

Now we derive the cumulative distribution function (cdf) and probability density function (pdf) of Y in Model 2. Equation (2.11) can be rewritten as

$$Y = (1 - I)Z + IG, \tag{2.13}$$

where $G := Z + X$. By the independence between I and (Z, G), the cdf of Y is

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}((1 - I)Z + IG \leq y) \\ &= \mathbb{P}\left(Z \leq y | I = 0\right) \cdot \mathbb{P}(I = 0) + \mathbb{P}(G \leq y | I = 1) \cdot \mathbb{P}(I = 1) \\ &= (1 - \alpha)\Phi_Z(y) + \alpha F_G(y). \end{aligned} \tag{2.14}$$

Since $G = Z + X$ and $Y = Z + IX$, by Theorem C.0.4 in Appendix C, we have that G and Y are both absolutely continuous, i.e. their density functions exist. Then, from (2.14), the pdf of Y is given by

$$f_Y = (1 - \alpha)\phi_Z + \alpha f_G. \tag{2.15}$$

Furthermore, since $G = Z + X$, we have $f_G = \phi_Z * f_X$, where $*$ denotes the convolution of two functions. Then equation (2.15) becomes

$$f_Y = (1 - \alpha)\phi_Z + \alpha(\phi_Z * f_X). \tag{2.16}$$

2.1.4 Proposed Estimator

From Section 2.1.2 we see that Z follows a normal distribution $Z \sim \mathcal{N}(\mu_0\Delta, \sigma_0^2\Delta)$. To ease the notation, we may define, for a given data frequency Δ,

$$\mu := \mu_0\Delta \quad \text{and} \quad \sigma := \sigma_0\sqrt{\Delta}. \tag{2.17}$$

Then the characteristic function of Z is given by

$$\varphi_Z(t) = e^{i\mu t - \frac{1}{2}\sigma^2 t^2}. \tag{2.18}$$

By assumption (A1-3), μ and σ are known. Thus, using (2.9) we may define an estimator of the jump ratio α by

$$\hat{\alpha} = 1 - \lim_{t\to\infty} \Re\left(\frac{\hat{\varphi}_Y(t)}{e^{i\mu t - \frac{1}{2}\sigma^2 t^2}}\right), \tag{2.19}$$

where $\hat{\varphi}_Y$ is an estimator of the characteristic function of Y. After obtaining $\hat{\alpha}$, we can use (2.10) to introduce an estimator of the distribution of the jump size X in terms of its characteristic function by

$$\hat{\varphi}_X(t) = \frac{\frac{\hat{\varphi}_Y(t)}{e^{i\mu t - \frac{1}{2}\sigma^2 t^2}} - (1-\hat{\alpha})}{\hat{\alpha}}. \tag{2.20}$$

To define an estimator completely, we have to address the following two issues:

- how to obtain an estimator of $\varphi_Y(t)$, where Y is observable.
- how to deal with the limit of t going to infinity in (2.19). An appropriate selection of t might be necessary.

We first look at methods of estimating the c.f. of Y. One straightforward approach is to use the *empirical characteristic function*. Other estimators of the characteristic function will be discussed in Section 2.4.

Given a random sample $Y_1, Y_2, ..., Y_n$ from the distribution of Y, the *empirical characteristic function* (e.c.f.) of Y is defined by

$$\hat{\varphi}_Y(t) := \frac{1}{n}\sum_{j=1}^{n} e^{itY_j}, \quad t \in \mathbb{R}. \tag{2.21}$$

By the strong law of large numbers, the e.c.f. is a strongly consistent estimator of the underlying characteristic function, i.e. $\hat{\varphi}_Y(t) \to \varphi_Y(t)$ a.s. for any fixed $t \in \mathbb{R}$.

Notice that in the estimator of the jump frequency and the estimator of the jump size distribution, the ratio of characteristic functions $\varphi_Y(t)/\varphi_Z(t)$ plays an important role. By (2.12), this ratio is equal to $\varphi_V(t)$. Using the e.c.f. (2.21) and the c.f. of Z in (2.18), we can estimate the c.f. of V by

$$\begin{aligned}
\hat{\varphi}_V(t) &= \hat{\varphi}_Y(t)/\varphi_Z(t) \\
&= \frac{1}{n}\sum_{j=1}^{n} e^{itY_j} \cdot e^{-i\mu t + \frac{1}{2}\sigma^2 t^2} \\
&= e^{\frac{1}{2}\sigma^2 t^2} \frac{1}{n}\sum_{j=1}^{n} e^{i(Y_j-\mu)t} \\
&= e^{\frac{1}{2}\sigma^2 t^2} \left\{ \frac{1}{n}\sum_{j=1}^{n} \cos((Y_j-\mu)t) + i \cdot \frac{1}{n}\sum_{j=1}^{n} \sin((Y_j-\mu)t) \right\}.
\end{aligned} \tag{2.22}$$

Therefore, the estimator (2.19) becomes

$$\begin{aligned}
\hat{\alpha} &= 1 - \lim_{t\to\infty} \Re(\hat{\varphi}_V(t)) \\
&= 1 - \lim_{t\to\infty} \left\{ e^{\frac{1}{2}\sigma^2 t^2} \cdot \frac{1}{n}\sum_{j=1}^{n} \cos((Y_j-\mu)t) \right\}.
\end{aligned} \tag{2.23}$$

However, this estimator cannot be used directly because it involves the limit of t going to infinity and the limit does not exist. In practice we have to use a finite t, and thus in the following we consider the estimators of the form

$$\begin{aligned}
\hat{\alpha}(t) &:= 1 - \Re(\hat{\varphi}_V(t)) \\
&= 1 - e^{\frac{1}{2}\sigma^2 t^2} \cdot \frac{1}{n}\sum_{j=1}^{n} \cos((Y_j-\mu)t)
\end{aligned} \tag{2.24}$$

with $t > 0$, and then we select an appropriate value of t. Since α is allowed to take values only in the interval [0,1], we may trim $\hat{\alpha}(t)$ by one from above and by zero from below. However, it might be difficult to establish any theoretical property of such an estimator. In the following we still consider $\hat{\alpha}(t)$ in (2.24), but in the simulation studies we will truncate values of our estimators to the interval $[0, 1]$.

The formula (2.23) suggests that it is better to choose a large value of t. However, as we will see later, when t becomes large, the variance of $\hat{\alpha}(t)$ increases quickly, which suggests that t cannot be chosen too large. A detailed discussion on the selection of t will be provided in Section 2.2 and Section 2.3.

To gain some insight into the basic properties of the estimator $\hat{\alpha}(t)$, in the following we present two examples based on simulated data. We simulate the data using the mixture model (i.e. Model 2), instead of the jump-diffusion model (i.e. Model 1). This assures that there will be at most one jump occurring in one time interval and α is exactly the expected proportion of jump observations. We have checked that, comparing with generating observations based on the jump-diffusion model, the difference in the estimation results of the jump frequency is small.

Example 2.1.2 (Merton Model)**.** In the Merton model, the jump size X follows a normal distribution. Denote $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, where typically $\sigma_X > \sigma$. The parameters are set as follows:

- $\Delta = 1/250$, i.e. daily data. Other values of Δ will be discussed in Chapter 3.
- $\mu_0 = 0.1$, $\sigma_0 = 0.2$.[‡] This implies that the expected daily return is $\mu = 0.1 \times \Delta = 4 \times 10^{-4}$, and the daily volatility is $\sigma = 0.2 \times \sqrt{\Delta} \approx 1.26\%$.

[‡]Throughout the thesis, the parameters $\mu_0 = 0.1$, $\sigma_0 = 0.2$ are used in all examples, unless otherwise stated. Note that it is the difference between the diffusion distribution and the jump size distribution, rather than the diffusion distribution itself, that decides the performance of the method.

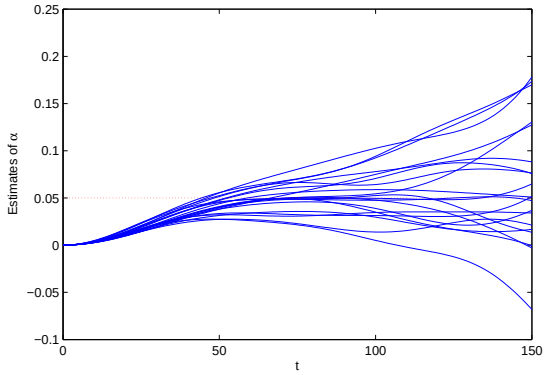
- $\alpha = 0.05$. Equivalently, the jump intensity is $\lambda = 12.5$, i.e. there are 12.5 jumps per year on average.

- $\mu_X = 0$, $\sigma_X = 3 \times 0.2 \times \sqrt{1/250} \approx 3.79\%$.[§]

Note that the parameters μ_0, σ_0 in the diffusion component and the jump intensity λ are close to the ones in Kou (2002). The expected value μ_X is assumed zero for convenience, and later other values of μ_X will also be discussed. The standard deviation σ_X of the jump size distribution is not far from the one in Kou (2002).

For the time horizon, we use data set over a 10-year period, i.e. $T = 10$. Then the sample size is $n = T/\Delta = 2500$.

Figure 2.1: Twenty curves of $\hat{\alpha}(t)$ [Merton]



After generating a sample of $n = 2500$ observations, we use equation (2.24) to obtain a curve of $\hat{\alpha}(t)$ with respect to t. Repeating this 20 times we obtain Figure 2.1. From this figure we can see that, as t increases from zero, $\hat{\alpha}(t)$ starts at zero and then increases but later it could increase to large positive values or decrease to

[§]Here $\sigma_X = 3 * \sigma$, i.e. the standard deviation of jump size is equal to triple of daily volatility. In the following we show that this is large enough to obtain good estimates of jump frequency, although a larger value of σ_X would definitely lead to a better estimate.

large negative values. This suggests that the variance of $\hat{\alpha}(t)$ increases significantly as t increases.

Example 2.1.3 (Kou Model)**.** In the Kou model, the jump size X follows a double exponential distribution with the pdf given by (1.6):

$$f(x) = p\eta_1 e^{-\eta_1 x} 1_{\{x>0\}} + (1-p)\eta_2 e^{-\eta_2|x|} 1_{\{x<0\}}.$$

As in the Merton model, we set

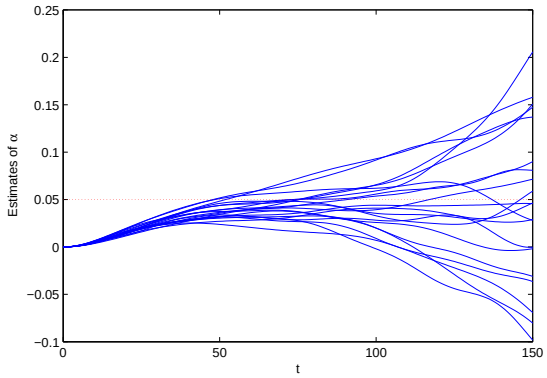
- $\Delta = 1/250$; $\mu_0 = 0.1$; $\sigma_0 = 0.2$; $\alpha = 0.05$ (i.e. $\lambda = 12.5$).

The double exponential parameters are the same as the ones in Kou (2002), i.e.

- $p = 0.3; \eta_1 = 1/0.02; \eta_2 = 1/0.04$.

Then we have $E(X) = -2.2\%$ and $SD(X) = 4.47\%$. That is, the average jump size is -2.2%, the standard deviation of jump size is 4.47%, and on average 70% of jumps are downward since p is the probability of a jump being positive.

Figure 2.2: Twenty curves of $\hat{\alpha}(t)$ [Kou]



As in the Merton model, we consider daily data over a 10-year period, i.e. $n = 2,500$. Figure 2.2 shows 20 curves of $\hat{\alpha}(t)$ for $t \in [0, 150]$.

Similarly to the Merton model, the variance of $\hat{\alpha}(t)$ increases quickly as t increases. As we will explain in Section 2.2, in both cases the main reason for this behavior is the term $e^{\frac{1}{2}\sigma^2 t^2}$ in equation (2.24). From the plot we may infer that, for a fixed sample size n, a judicious selection of t is necessary to obtain an accurate estimate of α.

Remark 2.1.4. Although invisible in Figure 2.1 or Figure 2.2, each curve $\hat{\alpha}(t)$, $t \geq 0$, in these two figures is oscillating between large positive numbers and large negative numbers. This is because the term $\frac{1}{n}\sum_{j=1}^{n}\cos((Y_j - \mu)t)$ in equation (2.24), which corresponds to a sample average, is oscillating around its theoretical mean, but the term $e^{\frac{1}{2}\sigma^2 t^2}$ inflates the pattern significantly, especially when t is large.

2.1.5 Identifiability Problem

Now we discuss an important issue for mixture models: the identifiability problem. In general, a parametric distribution family is said to be *identifiable* if different parametric values generate different members of the family. The identifiability for our problem can be defined similarly. The estimation problem would become meaningless if the model is not identifiable.

For the mixture model in (2.11), if we only look at the pdf given by (2.15):

$$f_Y = (1 - \alpha)\phi_Z + \alpha f_G, \tag{2.25}$$

we might infer that α and f_G are not identifiable, even in the case when the parameters μ and σ in the normal density ϕ_Z are known. This is because for any α belonging to the set $A = \{\alpha \in (0, 1) : f_G = (f_Y - (1 - \alpha)\phi_Z)/\alpha \geq 0\}$, there exists a well-defined f_G for which (2.25) holds. There is no guarantee that the set A contains only one value: for example, if $\alpha_0 \in A$, then $\alpha_0 + \epsilon$ for some $\epsilon > 0$ might also belong to A. Therefore, α and f_G in (2.25) might not be identifiable. A

specific counter-example is given by the following equality:

$$(1-\alpha)\cdot\phi(x)+\alpha\cdot\frac{\phi(x-2)+\phi(x)}{2}=\left(1-\frac{\alpha}{2}\right)\cdot\phi(x)+\frac{\alpha}{2}\cdot\phi(x-2),$$

where ϕ is an arbitrary even pdf. Comparing with the right side of (2.25), we may see that α and f_G in (2.25) are not identifiable.

However, based on the procedure we used to derive the proposed estimators, as summarized by equations (2.9) and (2.10), we see that, when μ and σ are known, α is indeed uniquely determined by φ_Y and φ_Z, and so is the distribution of X. In the following we provide some further explanations as to why the mixture model (2.11) is indeed identifiable.

Let us look at the pdf of Y given by (2.16). To simplify the notation, we drop the subscripts and, with some abuse of the notation, rewrite (2.16) as

$$f=(1-\alpha)\phi+\alpha(\phi*g), \tag{2.26}$$

where g is the pdf of the jump size X. We would like to investigate whether α and g are identifiable. For any fixed $\beta\in(0,1)$, (2.26) can be represented as

$$f=(1-\beta)\phi+\beta\left[\left(\frac{\alpha}{\beta}\right)\phi*g+\left(1-\frac{\alpha}{\beta}\right)\phi\right].$$

Now we ask this question: can we write $\left(\frac{\alpha}{\beta}\right)\phi*g+\left(1-\frac{\alpha}{\beta}\right)\phi=\phi*g^{new}$ for some g^{new}? The answer is yes.

$$\begin{aligned}
&\left[\left(\frac{\alpha}{\beta}\right)\phi*g+\left(1-\frac{\alpha}{\beta}\right)\phi\right](x)\\
&=\frac{\alpha}{\beta}\int_{-\infty}^{\infty}\phi(y)g(x-y)dy+\left(1-\frac{\alpha}{\beta}\right)\phi(x)\\
&=\int_{-\infty}^{\infty}\phi(y)\left(\frac{\alpha}{\beta}\right)g(x-y)dy+\int_{-\infty}^{\infty}\phi(y)\left(1-\frac{\alpha}{\beta}\right)\delta(x-y)dy\\
&=\int_{-\infty}^{\infty}\phi(y)\left[\left(\frac{\alpha}{\beta}\right)g(x-y)+\left(1-\frac{\alpha}{\beta}\right)\delta(x-y)\right]dy\\
&=\left[\phi*g^{new}\right](x)
\end{aligned}$$

where

$$g^{new} = \left(\frac{\alpha}{\beta}\right) g + \left(1 - \frac{\alpha}{\beta}\right) \delta,$$

and δ will be defined shortly. However, this g^{new} is not absolutely continuous. The reason is given as follows. Firstly, let us explain the notation $\delta(\cdot)$. It can be viewed as an *identity element for convolution*, i.e. $h * \delta = h$ for any function h. As we know, there is no identity element for convolution in L^1 (otherwise, we may use the Fourier inversion combining with the Riemann-Lebesgue Lemma to reach a contradiction). In other words, there is no absolute continuous distribution such that its density function $\delta(\cdot)$ satisfies $\phi * \delta = \phi$. Instead, there exist other "functions" which work as the identity element for convolution: the Dirac delta function and the Kronecker delta function. They correspond to a singular distribution and a discrete distribution whose cumulative distribution functions are the same and given by

$$F(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

Therefore, $\delta(\cdot)$ is not absolutely continuous. This implies that g^{new} is not absolutely continuous. Thus, we conclude that, when the diffusion parameters are known, the condition that "X is absolutely continuous" ensures the identifiability of the two-component mixture model (2.11).

When the diffusion parameters μ and σ are unknown, the identifiability problem may occur. However, robust estimation can provide good estimates of μ and σ when the jump ratio is not too large. Besides, the jump ratio decreases by the availability of higher frequency data, and as a result we can have more efficient estimation of μ and σ. Therefore, we do not need to worry too much about the identifiability problem for this case. More details about this case will be discussed in Chapter 3.

2.2 Properties of the Proposed Estimator

As an estimator of α we can use (2.24), which is of the form

$$\hat{\alpha}(t) = 1 - e^{\frac{1}{2}\sigma^2 t^2} \cdot \frac{1}{n} \sum_{j=1}^{n} \cos((Y_j - \mu)t)$$

for an appropriate value of t. To select t, it is helpful to analyze the bias and variance of $\hat{\alpha}(t)$ as functions of t. We discuss this in Section 2.2.1. The consistency and asymptotic distribution of $\hat{\alpha}(t)$ are discussed in Section 2.2.2 and Section 2.2.3. Some further explanation of our method are provided in Section 2.2.4. The selection of t for finite samples will be investigated in Section 2.3.

2.2.1 Mean and Variance

The following result provides the formulas for the bias, variance and mean squared error (MSE) of the estimator $\hat{\alpha}(t), t > 0$.

Proposition 2.2.1. *Consider the estimator $\hat{\alpha}(t)$ defined in (2.24), and denote $R(t) := \Re(\varphi_X(t)), t > 0$. Then we have the following:*

(i) The expectation of $\hat{\alpha}(t)$ is given by

$$\mathbb{E}(\hat{\alpha}(t)) = \alpha - \alpha R(t), \tag{2.27}$$

and hence the bias of $\hat{\alpha}(t)$ is

$$bias(\hat{\alpha}(t)) = \mathbb{E}[\hat{\alpha}(t)] - \alpha = -\alpha R(t). \tag{2.28}$$

Moreover, $bias(\hat{\alpha}(t)) \to 0$ as $t \to \infty$.

(ii) The variance of $\hat{\alpha}(t)$ is given by

$$Var(\hat{\alpha}(t)) = \frac{1}{2n}\left(e^{\sigma^2 t^2} + e^{-\sigma^2 t^2}\left[1 - \alpha + \alpha R(2t)\right] - 2\left[1 - \alpha + \alpha R(t)\right]^2\right) \tag{2.29}$$

and the mean squared error is given by

$$\begin{aligned} MSE(\hat{\alpha}(t)) &= (bias(\hat{\alpha}(t)))^2 + Var(\hat{\alpha}(t)) \\ &= \frac{1}{n}\left\{\frac{1}{2}\left(e^{\sigma^2t^2} + e^{-\sigma^2t^2}\left[1-\alpha+\alpha R(2t)\right]\right)\right. \\ &\quad \left. -(1-\alpha)^2 - 2\alpha(1-\alpha)R(t) + (n-1)\alpha^2R^2(t)\right\}. \end{aligned} \tag{2.30}$$

(iii) The optimal value of t which minimizes the MSE is the solution to the equation

$$\begin{aligned} &\sigma^2 t\left(e^{\sigma^2t^2} - e^{-\sigma^2t^2}\left[1-\alpha+\alpha R(2t)\right]\right) + \alpha e^{-\sigma^2t^2}R'(2t) \\ &- 2\alpha(1-\alpha)R'(t) + 2(n-1)\alpha^2R(t)R'(t) = 0 \end{aligned} \tag{2.31}$$

provided that it exists and at this value of t the second derivative of $MSE(\hat{\alpha}(t))$ is non-negative.

Proof. (i) The expectation of $\hat{\alpha}(t)$ is given by

$$\begin{aligned} \mathbb{E}[\hat{\alpha}(t)] &= \mathbb{E}\left[1 - e^{\frac{1}{2}\sigma^2t^2}\cdot\frac{1}{n}\sum_{j=1}^{n}\cos((Y_j-\mu)t)\right] \\ &= 1 - e^{\frac{1}{2}\sigma^2t^2}\mathbb{E}\left[\cos((Y-\mu)t)\right]. \end{aligned}$$

Let $\varphi_{Y-\mu}(t)$ denote the characteristic function of the random variable $Y-\mu$. Then

$$\mathbb{E}\left[\cos((Y-\mu)t)\right] = \Re\left(\varphi_{Y-\mu}(t)\right).$$

Using the fact that

$$\varphi_{Y-\mu}(t) = \mathbb{E}\left[e^{it(Y-\mu)}\right] = e^{-i\mu t}\varphi_Y(t)$$

and the equation (2.6), we obtain

$$\begin{aligned} \mathbb{E}\left[\cos((Y-\mu)t)\right] &= \Re\left(e^{-i\mu t}\cdot e^{i\mu t - \frac{1}{2}\sigma^2t^2}\cdot\left[(1-\alpha)+\alpha\varphi_X(t)\right]\right) \\ &= e^{-\frac{1}{2}\sigma^2t^2}\cdot\Re\left(\left[1-\alpha+\alpha\varphi_X(t)\right]\right) \\ &= e^{-\frac{1}{2}\sigma^2t^2}\left[1-\alpha+\alpha R(t)\right]. \end{aligned} \tag{2.32}$$

Therefore,

$$\begin{aligned}\mathbb{E}[\hat{\alpha}(t)] &= 1 - e^{\frac{1}{2}\sigma^2 t^2} \cdot e^{-\frac{1}{2}\sigma^2 t^2} \cdot [1 - \alpha + \alpha R(t)] \\ &= 1 - [1 - \alpha + \alpha R(t)] \\ &= \alpha - \alpha R(t).\end{aligned}$$

It follows that the bias of $\hat{\alpha}(t)$ is given by

$$\text{bias}(\hat{\alpha}(t)) = \mathbb{E}[\hat{\alpha}(t)] - \alpha = -\alpha R(t).$$

With the assumption (A2-1), we have that $\lim_{t\to\infty} \varphi_X(t) = 0$, which implies that $\text{bias}(\hat{\alpha}(t)) \to 0$ as $t \to \infty$.

(ii) The variance of $\hat{\alpha}(t)$ is given by

$$\begin{aligned}\text{Var}(\hat{\alpha}(t)) &= \text{Var}\left(1 - e^{\frac{1}{2}\sigma^2 t^2} \cdot \frac{1}{n}\sum_{j=1}^{n} \cos((Y_j - \mu)t)\right) \\ &= e^{\sigma^2 t^2} \cdot \frac{1}{n}\text{Var}\left(\cos((Y - \mu)t)\right) \\ &= \frac{1}{n} e^{\sigma^2 t^2} \cdot \left\{\mathbb{E}\left[\cos^2((Y - \mu)t)\right] - \left(\mathbb{E}\left[\cos((Y - \mu)t)\right]\right)^2\right\}.\end{aligned}$$

Then using the formula $\cos^2 x = (1 + \cos 2x)/2$ we obtain

$$\begin{aligned}\mathbb{E}\left[\cos^2((Y - \mu)t)\right] &= \frac{1}{2} \cdot \mathbb{E}\left[1 + \cos((Y - \mu) \cdot 2t)\right] \\ &= \frac{1}{2}\left(1 + e^{-2\sigma^2 t^2}\left[1 - \alpha + \alpha R(2t)\right]\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\text{Var}(\hat{\alpha}(t)) &= \frac{1}{n} e^{\sigma^2 t^2} \cdot \left\{\frac{1}{2}\left(1 + e^{-2\sigma^2 t^2}\left[1 - \alpha + \alpha R(2t)\right]\right) - \left(e^{-\frac{1}{2}\sigma^2 t^2}\left[1 - \alpha + \alpha R(t)\right]\right)^2\right\} \\ &= \frac{1}{n} e^{\sigma^2 t^2} \cdot \frac{1}{2}\left\{1 + e^{-2\sigma^2 t^2}\left[1 - \alpha + \alpha R(2t)\right] - 2e^{-\sigma^2 t^2}\left[1 - \alpha + \alpha R(t)\right]^2\right\} \\ &= \frac{1}{2n}\left(e^{\sigma^2 t^2} + e^{-\sigma^2 t^2}\left[1 - \alpha + \alpha R(2t)\right] - 2\left[1 - \alpha + \alpha R(t)\right]^2\right),\end{aligned}$$

where in the first line we used (2.32).

(iii) Now we can calculate the MSE:

$$\begin{aligned}
\mathrm{MSE}(\hat{\alpha}(t)) &= [\mathrm{bias}(\hat{\alpha}(t))]^2 + \mathrm{Var}(\hat{\alpha}(t)) \\
&= [-\alpha R(t)]^2 + \frac{1}{2n}\left(e^{\sigma^2 t^2} + e^{-\sigma^2 t^2}\left[1-\alpha+\alpha R(2t)\right]\right) - \frac{1}{n}\left[1-\alpha+\alpha R(t)\right]^2 \\
&= [-\alpha R(t)]^2 + \frac{1}{2n}\left(e^{\sigma^2 t^2} + e^{-\sigma^2 t^2}\left[1-\alpha+\alpha R(2t)\right] - 2\left[1-\alpha+\alpha R(t)\right]^2\right) \\
&= \frac{1}{2n}\left(e^{\sigma^2 t^2} + e^{-\sigma^2 t^2}\left[1-\alpha+\alpha R(2t)\right] - 2\left[1-\alpha+\alpha R(t)\right]^2 + 2n\alpha^2 R^2(t)\right) \\
&= \frac{1}{n}\left\{\frac{1}{2}\left(e^{\sigma^2 t^2} + e^{-\sigma^2 t^2}\left[1-\alpha+\alpha R(2t)\right]\right)\right. \\
&\quad \left. -(1-\alpha)^2 - 2\alpha(1-\alpha)R(t) + (n-1)\alpha^2 R^2(t)\right\}.
\end{aligned}$$

To find the optimal value of t which minimizes the MSE, we differentiate $\mathrm{MSE}(\hat{\alpha}(t))$ with respect to t to obtain

$$\begin{aligned}
\frac{\partial(\mathrm{MSE}(\hat{\alpha}(t)))}{\partial t} = \frac{1}{n}\Big\{&\sigma^2 t\left(e^{\sigma^2 t^2} - e^{-\sigma^2 t^2}[(1-\alpha)+\alpha R(2t)]\right) + \alpha e^{-\sigma^2 t^2} R'(2t) \\
&-2\alpha(1-\alpha)R'(t) + 2(n-1)\alpha^2 R(t)R'(t)\Big\}.
\end{aligned}$$

Thus the optimal value of t which minimizes the MSE is the solution to the equation

$$\begin{aligned}
&\sigma^2 t\left(e^{\sigma^2 t^2} - e^{-\sigma^2 t^2}\left[1-\alpha+\alpha R(2t)\right]\right) + \alpha e^{-\sigma^2 t^2} R'(2t) \\
&- 2\alpha(1-\alpha)R'(t) + 2(n-1)\alpha^2 R(t)R'(t) = 0,
\end{aligned}$$

provided that it exists and at this value of t the second derivative of $\mathrm{MSE}(\hat{\alpha}(t))$ is non-negative. ⊠

The special cases in which X follows either a normal distribution (in the Merton model) or a double exponential distribution (in the Kou model) are discussed in the following two corollaries.

Corollary 2.2.2 (Merton model)**.** *When the jump size $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, we have $R(t) = \Re(\varphi_X(t)) = \cos(\mu_X t)\cdot e^{-\frac{1}{2}\sigma_X^2 t^2}$. Thus the expectation of $\hat{\alpha}(t)$ is*

$$\mathbb{E}(\hat{\alpha}(t)) = \alpha - \alpha\cos(\mu_X t)\cdot e^{-\frac{1}{2}\sigma_X^2 t^2},$$

and the variance of $\hat{\alpha}(t)$ *is*

$$Var(\hat{\alpha}(t)) = \frac{1}{2n}\left\{e^{\sigma^2 t^2} + e^{-\sigma^2 t^2}\left[1-\alpha+\alpha\cos(2\mu_X t)\cdot e^{-2\sigma_X^2 t^2}\right]\right.$$
$$\left.-2\left(1-\alpha+\alpha\cos(\mu_X t)\cdot e^{-\frac{1}{2}\sigma_X^2 t^2}\right)^2\right\}.$$

The mean squared error can be found by (2.30), and the optimal value of t *which minimizes the MSE is the solution to equation (2.31).*

Corollary 2.2.3 (Kou model)**.** *When the jump size* X *follows a double exponential distribution with the pdf (1.6), we have*

$$\varphi_X(t) = \frac{p\eta_1}{\eta_1 - it} + \frac{q\eta_2}{\eta_2 + it} \tag{2.33}$$

and thus

$$R(t) \equiv \Re(\varphi_X(t)) = \frac{p\eta_1^2}{\eta_1^2 + t^2} + \frac{q\eta_2^2}{\eta_2^2 + t^2}.$$

Then the expectation of $\hat{\alpha}(t)$ *is*

$$\mathbb{E}(\hat{\alpha}(t)) = \alpha - \alpha\left(\frac{p\eta_1^2}{\eta_1^2 + t^2} + \frac{q\eta_2^2}{\eta_2^2 + t^2}\right),$$

and the variance of $\hat{\alpha}(t)$ *is*

$$Var(\hat{\alpha}(t)) = \frac{1}{2n}\left\{e^{\sigma^2 t^2} + e^{-\sigma^2 t^2}\left[1-\alpha+\alpha\left(\frac{p\eta_1^2}{\eta_1^2+4t^2} + \frac{q\eta_2^2}{\eta_2^2+4t^2}\right)\right]\right.$$
$$\left.-2\left[1-\alpha+\alpha\left(\frac{p\eta_1^2}{\eta_1^2+t^2} + \frac{q\eta_2^2}{\eta_2^2+t^2}\right)\right]^2\right\}.$$

The mean squared error can be found by (2.30), and the optimal value of t *which minimizes the MSE is the solution to equation (2.31).*

In the above two corollaries, the optimal values of t which minimize the corresponding MSE's cannot be solved explicitly. But we can use numerical procedure to evaluate them.

Remark 2.2.4. We observe that the parameter μ does not enter the bias or variance formula for $\hat{\alpha}(t)$, and hence it has no effect on the accuracy of estimation of α. However, when we assume μ unknown, as is the case in Chapter 4, the accuracy of its estimate does play a role in the accuracy of the estimation of α (see Proposition 4.2.2).

In the following we discuss the asymptotic unbiasedness of the estimator $\hat{\alpha}(t)$. By the expectation formula in (2.27), $\hat{\alpha}(t)$ is biased for any fixed value of t. Since the expectation does not depend on the sample size n, we might think that it is even not asymptotically unbiased as n goes to infinity. However, there is a tuning parameter t which, as we will explain later, may be allowed to depend on the sample size n in a way similar to how the window width depends on the sample size in the context of kernel density estimation. When the sample size increases, the selected value of t can increase. We also have $\lim_{t\to\infty} R(t) = 0$ by Lemma 2.1.1. Thus, we obtain the following corollary.

Corollary 2.2.5. *The estimator $\hat{\alpha}(t)$ defined by (2.24) is asymptotically unbiased if the selected value of t satisfies $\lim_{n\to\infty} t(n) \to \infty$.*

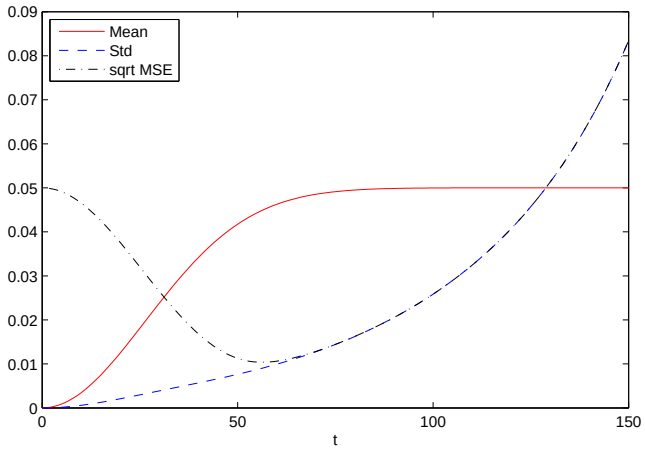
Note that this corollary holds regardless of the speed of t going to infinity as a function of n.

In the following examples we apply the above bias and variance analysis to the Merton model and the Kou model.

Example 2.2.6 (Merton Model)**.** With the same parameter setting as in Example 2.1.2, Figure 2.3 shows the expectation, variance and MSE of $\hat{\alpha}(t)$ using the formulas presented in Proposition 2.2.1.

From the figure we can see that, as t increases, the expected value of the estimator approaches the true value of $\alpha \equiv 0.05$. But the variance increases rapidly, which confirms the first impression we had from Examples 2.1.2 and 2.1.3. The

Figure 2.3: Mean, std and MSE of $\hat{\alpha}(t)$ [Merton]

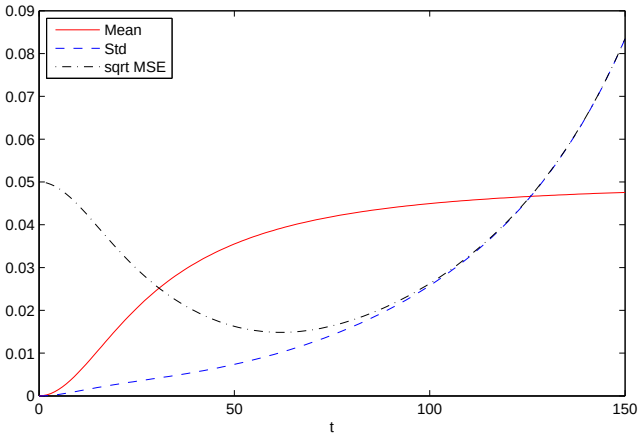


MSE first decreases and then increases to infinity, which suggests an "optimal" selection of t. Using a numerical procedure, we find that the smallest value of the MSE corresponds to $t_{opt} = 56.2789$. At this point, we have $\text{MSE}_{min} = 1.0767 \times 10^{-4}$, Mean = 0.0449 (i.e. bias=0.0041) and SD = 0.0090. Since the mean value is close to the true value of α, and the standard deviation is small relative to the true value of α, we can conclude that the estimate $\hat{\alpha}(t)$ at this "optimal" selection of t performs well.

Example 2.2.7 (Kou Model)**.** With the same parameter setting as in Example 2.1.3, Figure 2.4 shows the expectation, variance and MSE of $\hat{\alpha}(t)$ using the formulas in Proposition 2.2.1.

Comparing this figure with Figure 2.3 for the Merton model, the plot of the standard deviation is similar, but the expectation approaches the true level of α at a lower speed. This would suggest that our estimator should perform better for the Merton model than for the Kou model. In this figure, the lowest point on the MSE curve corresponds to $t_{opt} = 61.9214$ at which $\text{MSE}_{min} = 2.2055 \times 10^{-4}$, Mean = 0.0392 (i.e. bias=0.0108) and SD = 0.0102. The bias is significantly larger

Figure 2.4: Mean, std and MSE of $\hat{\alpha}(t)$ [Kou]



than the one for the Merton model.

2.2.2 Consistency

In this section we investigate the conditions under which our estimator is consistent. We assume that the time horizon T of the available data increases but the data frequency Δ remains constant. Another type of asymptotics for which the data frequency increases will be discussed in Chapter 3.

Let T_n denote the selection of t for a given sample size n. Define a *triangular array* of random variables by

$$B_{nj} \equiv B_j(T_n) := 1 - e^{\frac{1}{2}\sigma^2 T_n^2} \cos((Y_j - \mu)T_n), \quad 1 \leq j \leq n, \quad n \geq 1. \tag{2.34}$$

Then our estimator becomes

$$\hat{\alpha}_n \equiv \hat{\alpha}(T_n) = \frac{1}{n}\sum_{j=1}^{n} B_j(T_n) = \frac{1}{n}\sum_{j=1}^{n} B_{nj}, \quad n \geq 1. \tag{2.35}$$

Note that, for a fixed n, $\hat{\alpha}_n$ can be viewed as a sample average of the i.i.d. random

variables $B_{n1}, B_{n2}, ...B_{nn}$. But each of these random variables depend on the sample size n through T_n.

In the following we first discuss the weak consistency of $\hat{\alpha}_n$, and then discuss its strong consistency.

Theorem 2.2.8 (Weak consistency)**.** *Consider Model 2 in (2.11) with assumptions (A2-1) and (A2-2), and the estimator $\hat{\alpha}(t)$ defined by (2.24). Let C_n be a sequence of positive real numbers such that*

$$C_n \to \infty \quad \textit{and} \quad C_n/n \to 0$$

as $n \to \infty$. Then for

$$T_n = \sqrt{\ln C_n}/\sigma$$

we have that, as $n \to \infty$,

$$\hat{\alpha}_n \equiv \hat{\alpha}(T_n) \xrightarrow{p} \alpha.$$

Proof. It suffices to show that for any $\epsilon > 0$, we have

$$\mathbb{P}(|\hat{\alpha}_n - \alpha| > \epsilon) \to 0 \quad \text{as} \quad n \to \infty.$$

Using Proposition 2.2.1, we have

$$\begin{aligned}
\mathbb{P}(|\hat{\alpha}_n - \alpha| > \epsilon) &\leq \mathbb{P}(|\hat{\alpha}_n - \mathbb{E}\hat{\alpha}_n| + |\mathbb{E}\hat{\alpha}_n - \alpha| > \epsilon) \\
&= \mathbb{P}(|\hat{\alpha}_n - \mathbb{E}\hat{\alpha}_n| > \epsilon - |\mathbb{E}\hat{\alpha}_n - \alpha|) \\
&= \mathbb{P}\left(|\hat{\alpha}(T_n) - \mathbb{E}(\hat{\alpha}(T_n))| > \epsilon - \alpha|R(T_n)|\right).
\end{aligned}$$

By Chebyshev's inequality, the expression in the last line is bounded from above by

$$\begin{aligned}
&\frac{\mathrm{Var}(\hat{\alpha}(T_n))}{(\epsilon - \alpha|R(T_n)|)^2} \\
&= \frac{\frac{1}{2n}\left(e^{\sigma^2 T_n^2} + e^{-\sigma^2 T_n^2}\left[1 - \alpha + \alpha R(2T_n)\right] - 2\left[1 - \alpha + \alpha R(T_n)\right]^2\right)}{(\epsilon - \alpha|R(T_n)|)^2},
\end{aligned}$$

where the variance formula (2.29) is used. Note that "$C_n \to \infty$ and $C_n/n \to 0$" is equivalent to "$T_n \to \infty$ and $e^{\sigma^2 T_n^2}/n \to 0$". Moreover, $R(T_n) \to 0$ by assumption (A2-1). Hence, the numerator in the above ratio goes to zero and the denominator goes to ϵ^2. Therefore,

$$\mathbb{P}(|\hat{\alpha}_n - \alpha| > \epsilon) \to 0 \quad \text{as} \quad n \to \infty.$$

The proof is complete. $\boxtimes$

We actually have a stronger result under the same conditions.

Theorem 2.2.9. *With the same conditions as in Theorem 2.2.8, the estimator $\hat{\alpha}_n \equiv \hat{\alpha}(T_n)$ is consistent in the mean-square sense, meaning that $\mathbb{E}|\hat{\alpha}_n - \alpha|^2 \to 0$ as $n \to \infty$.*

Proof. To see this, we can write $\mathbb{E}|\hat{\alpha}_n - \alpha|^2 = \mathrm{Var}(\hat{\alpha}_n) + \mathrm{bias}^2(\hat{\alpha}_n)$. From the proof of Theorem 2.2.8 we know that $\mathrm{Var}(\hat{\alpha}_n) \to 0$ and $\mathrm{bias}(\hat{\alpha}_n) \to 0$ as $n \to \infty$. Thus, $\mathbb{E}|\hat{\alpha}_n - \alpha|^2 \to 0$ as $n \to \infty$. $\boxtimes$

In the following we prove the strong consistency of $\hat{\alpha}_n$. Firstly, we state a probability inequality for sums of bounded random variables, which was proved by Hoeffding (1963).

Lemma 2.2.10 (Hoeffding's inequality)**.** *Suppose that $X_1, X_2, ..., X_n$ are independent, $\mu_i = \mathbb{E}[X_i]$, and $a_i \le X_i \le b_i$ almost surely for $i = 1, 2, ..., n$. Then, for $\epsilon > 0$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mu_i \ge \epsilon\right) \le e^{-2n^2\epsilon^2/\sum_{i=1}^{n}(b_i - a_i)^2}.$$

Theorem 2.2.11 (Strong consistency)**.** *Consider Model 2 in (2.11) with assumptions (A2-1) and (A2-2), and the estimator $\hat{\alpha}(t)$ defined by (2.24). Let C_n be a sequence of positive real numbers such that*

$$C_n \to \infty$$

as $n \to \infty$, and

$$\sum_{n=1}^{\infty} e^{-\frac{\gamma \cdot n}{C_n}} < \infty$$

for any $\gamma > 0$. Then for

$$T_n = \sqrt{\ln C_n}/\sigma$$

we have that, as $n \to \infty$,

$$\hat{\alpha}_n \equiv \hat{\alpha}(T_n) \xrightarrow{a.s.} \alpha.$$

Proof. We want to show that $\hat{\alpha}_n - \alpha \xrightarrow{a.s.} 0$, i.e. $\hat{\alpha}_n - \mathbb{E}\hat{\alpha}_n + \mathbb{E}\hat{\alpha}_n - \alpha \xrightarrow{a.s.} 0$. Since $\mathbb{E}\hat{\alpha}_n - \alpha = \alpha R(T_n) \to 0$, we only need to show $\hat{\alpha}_n - \mathbb{E}\hat{\alpha}_n \xrightarrow{a.s.} 0$. By the Borel-Cantelli lemma, it suffices to prove that for any $\epsilon > 0$, we have

$$\sum_{n=1}^{\infty} \mathbb{P}(|\hat{\alpha}_n - \mathbb{E}\hat{\alpha}_n| > \epsilon) < \infty.$$

From equation (2.35) we see that, for a fixed n, $\hat{\alpha}_n$ can be viewed as a sample average of the i.i.d. random variables B_{nj}'s, $j = 1, 2, ..., n$. Since

$$1 - e^{\frac{1}{2}\sigma^2 T_n^2} \leq B_{nj} \leq 1 + e^{\frac{1}{2}\sigma^2 T_n^2},$$

we apply Hoeffding's inequality given in Lemma 2.2.10 to obtain

$$\begin{aligned}
\mathbb{P}(\hat{\alpha}_n - \mathbb{E}\hat{\alpha}_n \geq \epsilon) &= \mathbb{P}\left(\frac{1}{n}\sum_{j=1}^{n} B_{nj} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}B_{nj} \geq \epsilon\right) \\
&\leq \exp\left\{-\frac{2n^2\epsilon^2}{\sum_{j=1}^{n}\left[1 + e^{\frac{1}{2}\sigma^2 T_n^2} - (1 - e^{\frac{1}{2}\sigma^2 T_n^2})\right]^2}\right\} \\
&= \exp\left\{-\frac{2n^2\epsilon^2}{\sum_{j=1}^{n}\left[2e^{\frac{1}{2}\sigma^2 T_n^2}\right]^2}\right\} \\
&= \exp\left\{-\frac{n\epsilon^2}{2e^{\sigma^2 T_n^2}}\right\}.
\end{aligned} \tag{2.36}$$

It is easy to check that, if we reverse the sign of B_{nj}'s, the inequality in (2.36) still holds. That is, we also have

$$\mathbb{P}(\hat{\alpha}_n - \mathbb{E}\hat{\alpha}_n \leq -\epsilon) \leq \exp\left\{-\frac{n\epsilon^2}{2e^{\sigma^2 T_n^2}}\right\}.$$

Therefore,

$$\begin{aligned}
\sum_{n=1}^{\infty}\mathbb{P}(|\hat{\alpha}_n - \mathbb{E}\hat{\alpha}_n| \geq \epsilon) &= \sum_{n=1}^{\infty}\mathbb{P}(\hat{\alpha}_n - \mathbb{E}\hat{\alpha}_n \geq \epsilon) + \sum_{n=1}^{\infty}\mathbb{P}(\hat{\alpha}_n - \mathbb{E}\hat{\alpha}_n \leq -\epsilon) \\
&\leq 2\sum_{n=1}^{\infty}\exp\left\{-\frac{n\epsilon^2}{2e^{\sigma^2 T_n^2}}\right\}.
\end{aligned}$$

Let $\gamma = \frac{\epsilon^2}{2} > 0$, then

$$\sum_{n=1}^{\infty}\mathbb{P}\left(|\hat{\alpha}_n - \mathbb{E}\hat{\alpha}_n| \geq \epsilon\right) \leq 2\sum_{n=1}^{\infty}\exp\left\{-\frac{\gamma \cdot n}{e^{\sigma^2 T_n^2}}\right\}.$$

Therefore, if $\sum_{n=1}^{\infty}\exp\left\{-\frac{\gamma\cdot n}{e^{\sigma^2 T_n^2}}\right\} < \infty$ for any $\gamma > 0$, then the strong consistency of $\hat{\alpha}_n$ holds. Let $C_n = e^{\sigma^2 T_n^2}$, we reach the statement in the theorem. ☒

Remark 2.2.12. As we know, the series $\sum_{n=1}^{\infty} a_n < \infty$ implies $a_n \to 0$. Therefore, the sequence C_n which satisfies $\sum_{n=1}^{\infty} e^{-\frac{\gamma\cdot n}{C_n}} < \infty$ for any $\gamma > 0$ also satisfies $\frac{C_n}{n} \to 0$.

That is, the condition for strong consistency of $\hat{\alpha}_n$ is stronger than the condition for weak consistency of $\hat{\alpha}_n$.

Remark 2.2.13. One example of sequences satisfying the conditions "$C_n \to \infty$ as $n \to \infty$" and "$\sum_{n=1}^{\infty} e^{-\frac{\gamma \cdot n}{C_n}} < \infty$ for any $\gamma > 0$" is $C_n = \frac{n}{(\ln n)^2}$. To see this, we first apply the L'Hospital's rule to prove $C_n = \frac{n}{(\ln n)^2} \to \infty$. Then we can show "$\sum_{n=1}^{\infty} e^{-\frac{\gamma \cdot n}{C_n}} < \infty$ for any $\gamma > 0$" as follows: For any $\gamma > 0$, we have

$$\lim_{x \to \infty} \gamma \cdot \frac{(\ln x)^2}{\ln x^2} = \lim_{x \to \infty} \gamma \cdot \frac{2 \ln x \cdot \frac{1}{x}}{\frac{1}{x^2} \cdot 2x} = \lim_{x \to \infty} \gamma \cdot \ln x = \infty.$$

Then there exists (for any γ) an integer N such that for all $n > N$

$$\gamma (\ln n)^2 > \ln n^2,$$

which implies that

$$e^{-\gamma (\ln n)^2} < e^{-\ln n^2} = \frac{1}{n^2}.$$

Then

$$\sum_{n>N} e^{-\gamma \frac{n}{C_n}} < \sum_{n>N} \frac{1}{n^2}.$$

Since $\sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$, we have $\sum_{n=1}^{\infty} e^{-\frac{\gamma \cdot n}{C_n}} < \infty$.

In the following we use the Merton model as an example to illustrate the impact of the sample size n on the estimation accuracy, where the selection of t is based on the criterion of minimizing the MSE.

Example 2.2.14 (Merton model)**.** With all the other parameters the same as the ones in Example 2.1.2, we increase the sample size to $n = 4 * 2,500 = 10,000$, which corresponds to a 40-year daily data set. In Figure 2.5, we plot the expectation, standard deviation and MSE of $\hat{\alpha}(t)$ using the formulas in Proposition 2.2.1. By comparing this graph with Figure 2.3, we can see that the mean value stays the same (as expected, because the expectation of $\hat{\alpha}(t)$ does not depend n), but the variance decreases for any fixed value of t. From Proposition 2.2.1 we can see

that when the sample size increases to four times of the original one, the standard deviation reduces to half of the original standard deviation. Now the lowest point on the MSE curve corresponds to $t_{opt} = 63.1587$; at which $\text{MSE}_{min} = 3.6958 \times 10^{-5}$, Mean $= 0.0472$ and SD $= 0.0054$. Thus, the optimal value of t (denoted by t_{opt}) increases, and the performance of $\hat{\alpha}(t)$ at the point t_{opt} improves in both bias and variance.

Figure 2.5: Mean, std and MSE of $\hat{\alpha}(t)$ [Merton] ($n = 10,000$)

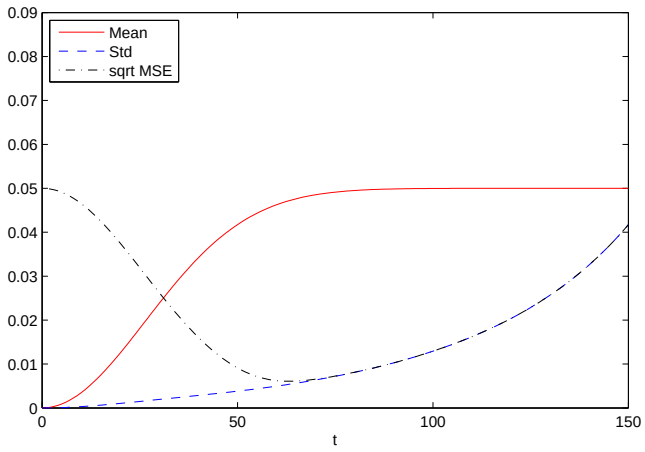


Table 2.1 further shows how t_{opt} depends on the sample size, which is implied by the time horizon T since the data frequency is fixed (as being daily). We can see that t_{opt} increases when the sample size increases, but the rate of the increase is very slow (and the rate is difficult to determine). Both the bias and the variance of $\hat{\alpha}(t)$ at t_{opt} decrease when the sample size increases. Note that in reality it is impossible to have 10^6 as the number of years of data, and this table only provides theoretical findings.

Table 2.1: Relation between the selection of t and the time horizon T

T	t_{opt}	MSE_{min}	Mean	St. Dev.
10	56.2789	1.0767e-004	0.0449	0.0090
40	63.1587	3.6958e-005	0.0472	0.0054
10^2	67.3448	1.7914e-005	0.0481	0.0038
10^3	76.9574	2.7742e-006	0.0493	0.0015
10^4	85.5996	4.0888e-007	0.0497	5.8614e-004
10^5	93.5120	5.8100e-008	0.0499	2.2274e-004
10^6	100.8452	8.0314e-009	0.0500	8.3318e-005

2.2.3 Asymptotic Normality

In this section we discuss the asymptotic distribution of the proposed estimator. The type of asymptotics is the same as the one for consistency discussed in the beginning of the previous section. There are two results for the asymptotic distribution: the first one is applicable in the case when t is fixed, and the second one is for the case when the selection of t depends on the sample size.

Theorem 2.2.15. *For a fixed t, as $n \to \infty$ we have*

$$\sqrt{n}\left(\hat{\alpha}(t) - \mu^*\right) \xrightarrow{d} \mathcal{N}(0, \sigma^{*2})$$

where $\mu^* = \alpha - \alpha R(t)$ and

$$\sigma^{*2} = \frac{1}{2}\left(e^{\sigma^2 t^2} + e^{-\sigma^2 t^2}\left[1 - \alpha + \alpha R(2t)\right] - 2\left[1 - \alpha + \alpha R(t)\right]^2\right).$$

Proof. The definition of $\hat{\alpha}(t)$ in equation (2.24) can be rewritten as

$$\hat{\alpha}(t) = \frac{1}{n}\sum_{j=1}^{n}\left[1 - e^{\frac{1}{2}\sigma^2 t^2}\cos((Y_j - \mu)t)\right].$$

Thus, for a fixed t, $\hat{\alpha}(t)$ can be viewed as a sample average of the i.i.d. random variables $D_j \equiv 1 - e^{\frac{1}{2}\sigma^2 t^2}\cos((Y_j - \mu)t), j = 1, 2, ..., n$. Applying the central limit theorem, we obtain

$$\sqrt{n}\left(\hat{\alpha}(t) - \mathbb{E}[\hat{\alpha}(t)]\right) \xrightarrow{d} \mathcal{N}(0, \sigma^{*2})$$

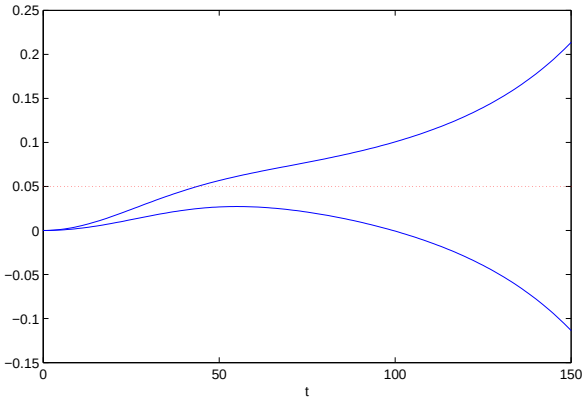
as $n \to \infty$, where $\sigma^{*2} = n\text{Var}[\hat{\alpha}(t)]$, and the expectation and variance of $\hat{\alpha}(t)$ are given in Proposition 2.2.1. ⊠

Using the above result, we can find the confidence interval (CI) of $\hat{\alpha}(t)$ for any fixed value of t.

Corollary 2.2.16. *For any fixed value of t, the β-CI of $\hat{\alpha}(t)$ is given by $(\mu^* - k\sigma^*/\sqrt{n}, \mu^* + k\sigma^*/\sqrt{n})$, where $k = \Phi^{-1}\left(1 - \frac{1-\beta}{2}\right)$, and μ^* and σ^* are defined in Theorem 2.2.15 .*

Example 2.2.17 (Merton Model)**.** With the same parameters as in Example 2.1.2, the 95%-CI of $\hat{\alpha}(t)$ are plotted in Figure 2.6. Since the variance of $\hat{\alpha}(t)$ gets larger when t increases, the confidence interval becomes wider.

Figure 2.6: Confidence interval (95%) of $\hat{\alpha}(t)$ [Merton]



In the following we derive another result on the asymptotic distribution, where the selection of t depends on the sample size.

From equation (2.35), we know that $\{\hat{\alpha}_n, n \geq 1\}$ involves a triangular array of random variables $\{B_{nj}, 1 \leq j \leq n, n \geq 1\}$. To prove the asymptotic normality, we first recall the central limit theorem for an array of random variables. McLeish (1974) proved the central limit theorems for martingales and near-martingales without the existence of moments or the full Lindeberg condition. For our problem, the following more basic result, which can be found in Loève (1963, page 316), will be sufficient.

Theorem 2.2.18 (Loève 1963, page 316)**.** *Suppose that for each n the variables $X_{n,1}, X_{n,2}, ..., X_{n,r_n}$ are independent and such that $\sum_{j=1}^{r_n} \mathbb{E}(X_{n,j}) \to \mu$ and $\sum_{j=1}^{r_n} Var(X_{n,j}) \to \sigma^2$ as $n \to \infty$. Then*

$$\sum_{j=1}^{r_n} X_{n,j} \xrightarrow{d} \mathcal{N}(\mu, \sigma^2) \text{ and } \max_{1 \leq j \leq r_n} \mathbb{P}[|X_{n,j}| \geq \epsilon] \longrightarrow 0$$

as $n \to \infty$ if and only if for every $\epsilon > 0$,

$$\sum_{j=1}^{r_n} \mathbb{P}[|X_{n,j}| \geq \epsilon] \longrightarrow 0 \quad as \quad n \to \infty.$$

Let $r_n = n$ and $X_{n,j} = \frac{B_{nj} - \mathbb{E}[B_{nj}]}{\sqrt{n} \cdot SD(B_{nj})}, j = 1, 2, ..., n$. Then we obtain a simplified version of the above theorem which can be used directly in our problem.

Theorem 2.2.19 (Loève 1963, page 316, special case)**.** *Let $B_{n1}, B_{n2}, ..., B_{nn}$ be a sequence of i.i.d. random variables for each n. Then*

$$\frac{\sum_{j=1}^{n} (B_{nj} - \mathbb{E}B_{nj})}{\sqrt{n \cdot Var(B_{nj})}} \xrightarrow{d} \mathcal{N}(0, 1)$$

as $n \to \infty$ if for every $\epsilon > 0$,

$$n\mathbb{P}\left[\left|\frac{B_{nj} - \mathbb{E}[B_{nj}]}{SD(B_{nj})}\right| \geq \epsilon\sqrt{n}\right] \longrightarrow 0 \quad as \quad n \to \infty.$$

Using Theorem 2.2.19 we obtain the following result on the asymptotic normality of our estimator.

Theorem 2.2.20 (Asymptotic Normality)**.** *Consider Model 2 in (2.11) with the assumptions (A2-1) and (A2-2), and the estimator $\hat{\alpha}(t)$ defined by (2.24). Denote $T_n = \sqrt{\ln C_n}/\sigma$, where C_n is a sequence of positive real numbers satisfying*

$$C_n \to \infty, \qquad C_n/n \to 0, \quad \text{and} \quad R(T_n)/\sqrt{C_n/n} \to 0$$

as $n \to \infty$. Then the estimator $\hat{\alpha}_n \equiv \hat{\alpha}(T_n)$ satisfies

$$\frac{\hat{\alpha}_n - \alpha}{\sqrt{C_n/(2n)}} \xrightarrow{d} \mathcal{N}(0,1).$$

Proof. By the Markov's inequality, $\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}(|X|^r)}{a^r}$ for any $r > 0$ and $a > 0$, we obtain

$$\mathbb{P}\left[\left|\frac{B_{nj} - \mathbb{E}B_{nj}}{SD(B_{nj})}\right| \geq \epsilon\sqrt{n}\right] \leq \frac{\mathbb{E}(|B_{nj} - \mathbb{E}B_{nj}|^{2+\delta})}{[\epsilon\sqrt{n} \cdot SD(B_{nj})]^{2+\delta}}$$

for some $\delta > 0$. Thus,

$$\begin{aligned} n\mathbb{P}\left[\left|\frac{B_{nj} - \mathbb{E}B_{nj}}{SD(B_{nj})}\right| \geq \epsilon\sqrt{n}\right] &\leq n \cdot \frac{\mathbb{E}(|B_{nj} - \mathbb{E}B_{nj}|^{2+\delta})}{(\epsilon\sqrt{n} \cdot SD(B_{nj}))^{2+\delta}} \\ &= \frac{\mathbb{E}(|B_{nj} - \mathbb{E}B_{nj}|^{2+\delta})}{\epsilon^{2+\delta} n^{\frac{\delta}{2}} \cdot (\mathrm{Var}(B_{nj}))^{1+\frac{\delta}{2}}}. \end{aligned}$$

From the proof of Proposition 2.2.1, the mean and variance of $B_{nj} \equiv B_j(T_n)$ for $j = 1, 2, ..., n$ are given by

$$\mathbb{E}(B_{nj}) = \alpha - \alpha R(T_n),$$

$$\mathrm{Var}(B_{nj}) = \frac{1}{2}\left(e^{\sigma^2 T_n^2} + e^{-\sigma^2 T_n^2}\left[1 - \alpha + \alpha R(2T_n)\right] - 2\left[1 - \alpha + \alpha R(T_n)\right]^2\right).$$

Then the numerator of the above ratio is

$$\begin{aligned}
\mathbb{E}(|B_{nj} - \mathbb{E}B_{nj}|^{2+\delta}) &= \mathbb{E}(|B_{nj} - \alpha(1 - R(T_n))|^{2+\delta}) \\
&\leq \mathbb{E}\left[(|B_{nj}| + 2)^{2+\delta}\right] \\
&= \mathbb{E}\left[\left(\left|1 - e^{\frac{1}{2}\sigma^2 T_n^2}\cos((Y_j - \mu)T_n)\right| + 2\right)^{2+\delta}\right] \\
&\leq \mathbb{E}\left[\left(e^{\frac{1}{2}\sigma^2 T_n^2} + 3\right)^{2+\delta}\right] \\
&= e^{\left(1+\frac{\delta}{2}\right)\sigma^2 T_n^2} \cdot (1 + o(1)),
\end{aligned}$$

where by $o(1)$ we mean this term has a limit of zero as $n \to \infty$. The last equality can be seen from the fact that $\lim_{n\to\infty}\left(e^{\frac{1}{2}\sigma^2 T_n^2} + 3\right)^{2+\delta} / e^{\left(1+\frac{\delta}{2}\right)\sigma^2 T_n^2} = 1$. Similarly, the denominator is

$$\begin{aligned}
&\epsilon^{2+\delta} n^{\frac{\delta}{2}} \cdot (\mathrm{Var}(B_{nj}))^{1+\frac{\delta}{2}} \\
&= \epsilon^{2+\delta} n^{\frac{\delta}{2}} \cdot \left(\frac{1}{2}\left(e^{\sigma^2 T_n^2} + e^{-\sigma^2 T_n^2}\left[1 - \alpha + \alpha R(2T_n)\right] - 2\left[1 - \alpha + \alpha R(T_n)\right]^2\right)\right)^{1+\frac{\delta}{2}} \\
&= \left(\frac{\epsilon}{\sqrt{2}}\right)^{2+\delta} \cdot n^{\frac{\delta}{2}} \cdot e^{\left(1+\frac{\delta}{2}\right)\sigma^2 T_n^2} \cdot (1 + o(1)).
\end{aligned}$$

Therefore,

$$\begin{aligned}
n\mathbb{P}\left[\left|\frac{B_{nj} - \mathbb{E}B_{nj}}{SD(B_{nj})}\right| \geq \epsilon\sqrt{n}\right] &\leq \frac{e^{\left(1+\frac{\delta}{2}\right)\sigma^2 T_n^2} \cdot (1 + o(1))}{\left(\frac{\epsilon}{\sqrt{2}}\right)^{2+\delta} \cdot n^{\frac{\delta}{2}} \cdot e^{\left(1+\frac{\delta}{2}\right)\sigma^2 T_n^2} \cdot (1 + o(1))} \\
&= \left(\frac{\sqrt{2}}{\epsilon}\right)^{2+\delta} \cdot n^{-\frac{\delta}{2}} \cdot (1 + o(1))
\end{aligned}$$

which goes to zero as $n \to \infty$. By Theorem 2.2.19, we have

$$\frac{\sum_{j=1}^{n}(B_{nj} - \mathbb{E}B_{nj})}{\sqrt{n \cdot \mathrm{Var}(B_{nj})}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Equivalently,

$$\frac{\hat{\alpha}_n - \mathbb{E}(\hat{\alpha}_n)}{\sqrt{\mathrm{Var}(\hat{\alpha}_n)}} \xrightarrow{d} \mathcal{N}(0, 1)$$

i.e.

$$\frac{\hat{\alpha}_n - (\alpha - \alpha R(T_n))}{\sqrt{\frac{1}{2n}\left(e^{\sigma^2 T_n^2} + e^{-\sigma^2 T_n^2}\left[1 - \alpha + \alpha R(2T_n)\right] - 2\left[1 - \alpha + \alpha R(T_n)\right]^2\right)}} \xrightarrow{d} \mathcal{N}(0,1).$$

Note that on the left side, only $\hat{\alpha}_n$ is random; the other terms are all deterministic. Since $T_n \to \infty$ as $n \to \infty$ and $R(t) \to 0$ as $t \to \infty$, we have

$$\lim_{n\to\infty} \frac{\sqrt{\frac{1}{2n}\left(e^{\sigma^2 T_n^2} + e^{-\sigma^2 T_n^2}\left[1 - \alpha + \alpha R(2T_n)\right] - 2\left[1 - \alpha + \alpha R(T_n)\right]^2\right)}}{\sqrt{\frac{1}{2n}e^{\sigma^2 T_n^2}}} = 1.$$

Then, using Slutsky's theorem, we have

$$\frac{\hat{\alpha}_n - \alpha + \alpha R(T_n)}{\sqrt{\frac{1}{2n}e^{\sigma^2 T_n^2}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Using the condition that $R(T_n)/\sqrt{e^{\sigma^2 T_n^2}/n} \to 0$, and applying Slutsky's theorem again, we obtain

$$\frac{\hat{\alpha}_n - \alpha}{\sqrt{\frac{1}{2n}e^{\sigma^2 T_n^2}}} \xrightarrow{d} \mathcal{N}(0,1).$$

This implies that $\frac{1}{2n}e^{\sigma^2 T_n^2} \to 0$. Let $C_n = e^{\sigma^2 T_n^2}$, we obtain the statement in the theorem. ⊠

Remark 2.2.21. We use the Merton model as an example to illustrate when the conditions in Theorem 2.2.20 are satisfied. In this case, we require both $e^{\sigma^2 T_n^2}/n \to 0$ and $e^{(\sigma^2+\sigma_X^2)T_n^2}/n \to \infty$. One example of a sequence satisfying these conditions is to let T_n be such that $e^{\sigma^2 T_n^2}/n = cn^q$, where $-\sigma_X^2/(\sigma^2 + \sigma_X^2) < q < 0$ and c is a constant.

Remark 2.2.22. In Theorem 2.2.15 and Theorem 2.2.20, we derived two different results on the asymptotic normality of $\hat{\alpha}_n$. In the following we list some of the basic facts about the two approaches:

- In Theorem 2.2.15:
  - t is fixed;
  - the bias stays at the same level;
  - asymptotic normality holds, and the convergence rate is $\sqrt{1/n}$.
- In Theorem 2.2.20:
  - the selected value of t increases as the sample size increases, although at a low speed;
  - the bias decreases;
  - asymptotic normality holds, and the convergence rate is $\sqrt{C_n/n}$, which is slower than $\sqrt{1/n}$.
  - For finite samples, this result does not tell how to select t.

In practice, when we always have a finite number of observations, we have to select t according to a certain procedure (e.g. method 3 which will be introduced in Section 2.3).

2.2.4 Further Explanations of our Method

In this section we use another approach, which is of "geometric" nature, to recover the estimator $\hat{\alpha}(t)$ in (2.24) of the jump frequency. Meanwhile, this approach also provides some insight into the selection of t, from a theoretical perspective, which will be used in both the estimation of jump frequency and the subsequent estimation of jump size distribution (in Chapter 5). The selection of t for finite samples will be discussed in Section 2.3.

The mixture model was given by (2.11):

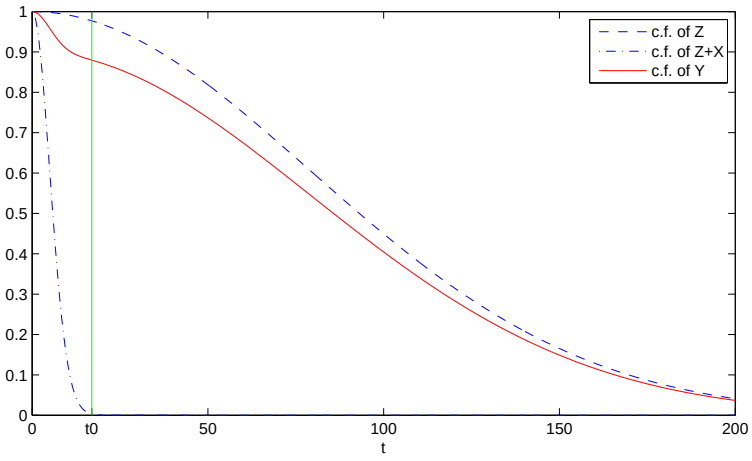
$$Y = (1 - I)Z + I(Z + X),$$

whose characteristic function was given by (2.6):

$$\varphi_Y(t) = (1-\alpha)\varphi_Z(t) + \alpha\varphi_Z(t)\varphi_X(t), \quad t \in \mathbb{R}. \tag{2.37}$$

In the following we take the Merton model with $\mu_Z = \mu_X = 0$ as an example to illustrate the idea. Let $\sigma_0 = 0.2$, $\sigma_X = 0.2$ and $\alpha = 0.1$. Figure 2.7 shows the c.f. of different components of the mixture model: the dashed curve corresponds to the c.f. of the diffusion component Z; the dash-dotted curve corresponds to the c.f. of the jump component $Z + X$; and the solid curve corresponds to the c.f. of the mixture Y. Typically the distribution of jump size (X) corresponds to a much larger proportion of large observations than the distribution of diffusion (Z) does, and moreover, $Z+X$ is a convolution of Z and X, so the convergence of $\varphi_{Z+X}(t)$ to zero would be much faster than the convergence of $\varphi_Z(t)$ to zero. In the following we show that this difference between the convergence rates of the c.f.'s of different components is the main driver of our method.

Figure 2.7: Characteristic functions of different components of the mixture model



Let t_0 be the smallest point from which the c.f. of X can be considered to be close to 0, i.e.

$$t_0 = \inf\{t \geq 0 : \varphi_X(s) \approx 0 \text{ for all } s \geq t\}.$$

We may understand the notation "≈ 0" by the following more rigorous way of defining t_0: for a small $\epsilon > 0$, let

$$t_0 = \inf\{t \geq 0 : |\varphi_X(s)| < \epsilon \text{ for all } s \geq t\}. \tag{2.38}$$

Of course, $\varphi_{Z+X}(t)$ will also be close to zero for $t \geq t_0$, since $|\varphi_{Z+X}(t)| = |\varphi_X(t)\varphi_Z(t)|$ $\leq |\varphi_X(t)|$. Ideally, we want to use this t_0 as the selection of t so that the bias is close to zero while the variance is still small, which follows from our bias and variance analysis presented in Proposition 2.2.1.

As we explained earlier, the selection of t is important, since the accuracy of the estimator of the jump frequency (and later the estimator of the jump size distribution) relies heavily on this selection of t. Below we provide an alternative explanation of this fact, where we consider two different cases.

- Case I: $t \in [t_0, \infty)$:
  On this interval, the c.f. of $Z + X$ is approximately equal to zero, i.e. $\varphi_Z(t)\varphi_X(t) \approx 0$. Then by (2.37) we have $\varphi_Y(t) \approx (1-\alpha)\varphi_Z(t)$ and it implies that
  $$\alpha \approx 1 - \frac{\varphi_Y(t)}{\varphi_Z(t)}.$$
  This "recovers" the formula (2.24) which was used to estimate the jump frequency. The bias of the estimator for $t \in [t_0, \infty)$ can be derived as follows. From equations (2.37) and (2.38) we have $|\varphi_Y(t) - (1-\alpha)\varphi_Z(t)| = |\alpha\varphi_Z(t)\varphi_X(t)| < \alpha\epsilon|\varphi_Z(t)|$. Dividing both sides by $|\varphi_Z(t)|$, we obtain
  $$\left|\alpha - \left(1 - \frac{\varphi_Y(t)}{\varphi_Z(t)}\right)\right| < \alpha\epsilon.$$
  This result matches the bias formula in (2.28). To avoid a large variance, we would select the smallest value of this interval, which is t_0.

- Case II: $t \in [0, t_0]$:
  On this interval, the c.f. $\varphi_X(t)$ starts from one and goes to a small value, i.e.

a value close to zero. After t_0, $\varphi_X(t)$ stays close to zero. See the following Remark 2.2.23 for the discussion on the estimation of jump size distribution.

Remark 2.2.23. To estimate the jump size distribution (discussed in details in Chapter 5), we would use mostly information about $\varphi_X(t)$ for $t \leq t_0$. This can be explained as follows:

- When $t > t_0$, $\varphi_X(t)$ stays close to zero. This suggests that there is not much information on the distribution of X remanning in $\varphi_X(t)$ for $t > t_0$, since we have that $\lim_{t\to\infty} \varphi_X(t) = 0$.
- We do not have the access to the true $\varphi_X(t)$, but only to its empirical version $\hat{\varphi}_X(t)$. However, $\hat{\varphi}_X(t)$ for $t > t_0$ would mask any remaining information on the distribution of X by bringing in estimation errors. This follows from the observation that $\hat{\varphi}_X(t)$ does not stay close to zero when $t > t_0$ but oscillates with an increasing amplitude as t increases.

The above facts explain why in Chapter 5 we use the method of truncating the empirical version of $\varphi_X(t)$ at some point to recover the distribution of X (see equation (5.16)), and the truncating point is the same as the selected value of t used for the estimation of jump frequency.

2.3 Selection of t for Finite Samples

For the selection of t, we obtained some theoretical results in Theorem 2.2.8 for the weak consistency of the estimator, Theorem 2.2.11 for the strong consistency and Theorem 2.2.20 for the asymptotic normality. However, these results do not suggest how to select t in practice when the sample size is finite. The bias and variance analysis in Proposition 2.2.1 provides some insight into the selection of t:

- Larger values of t will decrease the bias of the estimator but increase the variance. That is, t is a tuning parameter in our estimator, and its selection rests on a trade-off between the bias and the variance.
- When the sample size n increases, we may select a larger value of t, which leads to a smaller bias and possibly a smaller variance.
- The "optimal" value of t which minimizes the MSE solves some equation, but in practice we cannot use this equation since it involves the unknown jump frequency and jump size distribution.

In this section we propose three different methods of selecting t for finite samples. The first method is based on the variance formula of $\hat{\alpha}(t)$ presented in Proposition 2.2.1. The second method involves using a bootstrap procedure to obtain an estimate of the variance of $\hat{\alpha}(t)$. In the third method, we select t by using the shape of the curve $\hat{\alpha}(t)$ with respect to t. From the form of the estimator $\hat{\alpha}(t)$ in (2.24), we see that it is an even function of t. Thus, in the following we consider $t \geq 0$ only.

2.3.1 Method 1

In Proposition 2.2.1, we obtained the variance formula of $\hat{\alpha}(t)$ in (2.29):

$$\mathrm{Var}(\hat{\alpha}(t)) = \frac{1}{2n}\left(e^{\sigma^2 t^2} + e^{-\sigma^2 t^2}\left[1-\alpha+\alpha R(2t)\right] - 2\left[1-\alpha+\alpha R(t)\right]^2\right) \tag{2.39}$$

From this result we can see that the variance of $\hat{\alpha}(t)$ eventually increases exponentially as t increases. Thus, we need to bound t to avoid a large variance.

Since $R(2t) \in [-1,1]$ and $\alpha \in [0,1]$, it is easy to verify that

$$-1 \leq 1-\alpha+\alpha R(t) \leq 1, \quad \forall\, t \geq 0.$$

Then, the second term in (2.39) satisfies

$$\left|e^{-\sigma^2 t^2}\left[1-\alpha+\alpha R(2t)\right]\right| \leq 1$$

and the third term satisfies

$$\left|2\left[1-\alpha+\alpha R(t)\right]^2\right| \leq 2.$$

Also, as $t \to \infty$, we have $R(t) \to 0$ due to Lemma 2.1.1, and $e^{-\sigma^2 t^2} \to 0$. Thus, as t increases, the second term decreases to a small value; the third term will be around level 2 given that α is typically small; but the first term $e^{\sigma^2 t^2}$ increases quickly and could be very large. Therefore, we propose to choose t so that $e^{\sigma^2 t^2} = 2$, or

$$t_{M1} = \sqrt{\ln 2}/\sigma.$$

For such a selection of t, we hope the first term in (2.39) would be comparable in size with the remaining terms. Note that the number 2 in t_{M1} is not essential, and can be changed to any other number between 1.5 and 5.

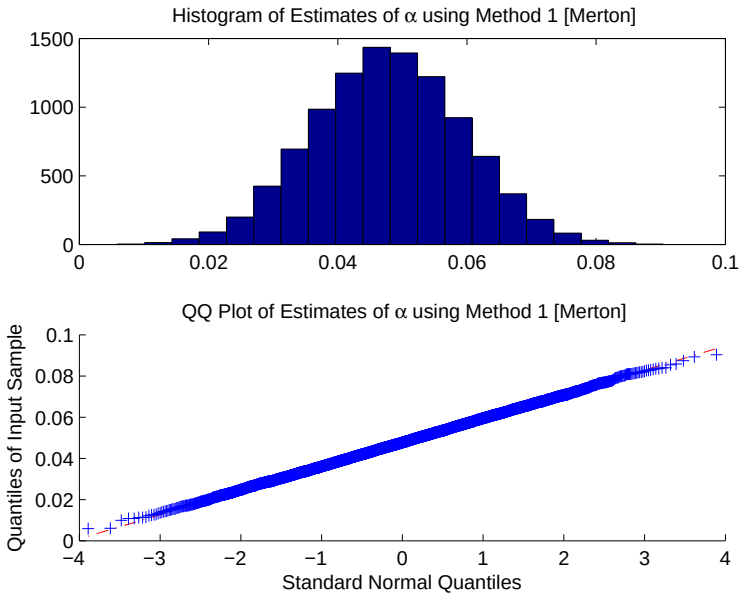
The estimate of α is then given by $\hat{\alpha}(t_{M1})$. If $\hat{\alpha}(t_{M1}) < 0$, we let $\hat{\alpha}(t_{M1}) = 0$. If $\hat{\alpha}(t_{M1}) > 1$, we let $\hat{\alpha}(t_{M1}) = 1$.

In our numerical studies we have found that this method works well when the data frequency is daily. However, when the data frequency is high, i.e. Δ is small, we have $\sigma \equiv \sigma_0\sqrt{\Delta}$ is small. Then the resulted t_{M1} tends to be too large, and thus Method 1 does not work well. This can also be explained by the fact that the second term in the variance formula (2.39) cannot be ignored when the data frequency is high. For example, when $\Delta = 1/2500$, we have $t_{M1} = 208.1387$ for Merton model, and in this case the variance is too large (comparing with the true value of α). See Figure 3.1 in Chapter 3, where the intersection of the curve "Approx2" and the t-axis is the selected value of t based on Method 1.

In the following we apply this method to the Merton model and the Kou model when the data frequency is daily.

Example 2.3.1 (Merton and Kou Models)**.** With the same parameter settings as in Example 2.1.2 for the Merton Model, the value of t is chosen to be $t_{M1} =$

Figure 2.8: Histogram and QQ plot of $\hat{\alpha}(t_{M1})$ using Method 1 [Merton]



$\frac{\sqrt{\ln 2}}{0.2*\sqrt{1/250}} = 65.8192$. This value is not far from the "theoretical" optimal value, 56.2789, which minimizes the MSE as shown in Example 2.2.6.

Figure 2.8 shows the histogram and the QQ plot of the estimates $\hat{\alpha}(t_{M1})$ based on 10,000 different samples. From the histogram we see that the estimates are bell-shaped, which suggests that $\hat{\alpha}(t_{M1})$ is normally distributed. This is confirmed by the QQ plot, since it shows a straight line (except for the two tails).

Table 2.2 provides a summary of the statistics for $\hat{\alpha}(t_{M1})$. The results are encouraging. Note that the true value of the jump ratio is $\alpha = 0.05$, and the theoretical expectation and standard deviation of $\hat{\alpha}(t_{M1})$ are 0.0478 and 0.0115, respectively. The term "MAD" denotes the median absolute deviation, i.e. the median of the absolute differences between observations and the median.

Now we turn to the Kou model. With the same parameter settings as in Example 2.1.3, the value of t_{M1} is the same as above: $t_{M1} = 65.8192$, because the same σ is

Table 2.2: Estimation results for $\hat{\alpha}(t_{M1})$ using Method 1 [Merton]

	Mean	Std	25% quantile	Median	75% quantile	MAD
$\hat{\alpha}(t_{M1})$	0.0479	0.0115	0.0400	0.0478	0.0556	0.0077

used for both models. Figure 2.9 shows the histogram and the QQ plot for $\hat{\alpha}(t_{M1})$ based on 10,000 different samples. The QQ plot also suggests that the estimates are normally distributed.

Figure 2.9: Histogram and QQ plot of $\hat{\alpha}(t_{M1})$ using Method 1 [Kou]

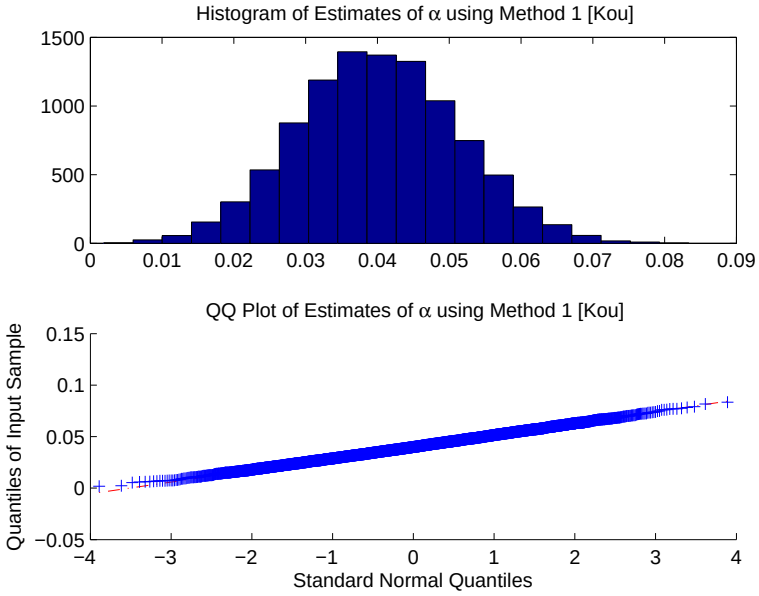


Table 2.3 provides a summary of the statistics for the Kou model. Note that the theoretical expectation and standard deviation of $\hat{\alpha}(t_{M1})$ are 0.0401 and 0.0113, respectively. When compared with the Merton model, the downward bias is significantly larger. This has been explained in Example 2.2.7. However, the standard deviation is similar, and the overall result is acceptable.

To summarize, Method 1 is easy to implement, since the selected value of t

Table 2.3: Estimation results for $\hat{\alpha}(t_{M1})$ using Method 1 [Kou]

	Mean	Std	25% quantile	Median	75% quantile	MAD
$\hat{\alpha}(t_{M1})$	0.0403	0.0113	0.0328	0.0401	0.0478	0.0075

depends on σ only. However, the drawback is that this method does not incorporate other information about the model, such as the distribution of the jump size or the sample size n. Also, it is not applicable for the high frequency data.

2.3.2 Method 2: Bootstrap

The bootstrap, introduced by Efron (1979), is a method typically used to estimate the sampling distribution of a statistic (or, an estimator) of interest. Suppose $Y_1, Y_2, ..., Y_n$ is a sequence of i.i.d. random variables with a common distribution, and θ is a parameter. Let $\hat{\theta}$ be an estimator of θ, which is a function of $Y_1, Y_2, ..., Y_n$. The basic bootstrap method can be described as follows:

Step 1 Construct a *re-sample* of equal size by *random sampling with replacement* from the original data set, and use this re-sample to compute $\hat{\theta}_1^*$.

Step 2 Repeat Step 1 for M times to obtain $\hat{\theta}_1^*, \hat{\theta}_2^*, ..., \hat{\theta}_M^*$.

Step 3 Use $\hat{\theta}_1^*, \hat{\theta}_2^*, ..., \hat{\theta}_M^*$ to obtain an empirical distribution of the estimator $\hat{\theta}$. Then further inference about $\hat{\theta}$ can be made. For example, the variance of $\hat{\theta}$ can be estimated by the sample variance of $\hat{\theta}_1^*, \hat{\theta}_2^*, ..., \hat{\theta}_M^*$.

In our problem, the observations $Y_1, Y_2, ..., Y_n$ from the distribution given by (2.6) are i.i.d., so we can use the basic bootstrap method to obtain an estimate of the variance of $\hat{\alpha}(t)$ for any value of t. Then we propose to select a value of t for which this estimated variance does not exceed a given level. This is the basic idea of Method 2. To describe it in steps, we have the following:

Step 1 For a given sample $Y_1, Y_2, ..., Y_n$, use the basic bootstrap method to obtain an estimate of the variance of $\hat{\alpha}(t)$ for a set of values of t that we are interested in. Denote this estimated variance by $\mathrm{Var}_B(\hat{\alpha}(t))$, and the corresponding standard deviation by $\mathrm{Std}_B(\hat{\alpha}(t))$.

Step 2 Set a threshold value ϵ (details on how to set this value is given in the following Example 2.3.2), and select t by

$$t_{M2} := \max\{t \geq 0 : \mathrm{Std}_B(\hat{\alpha}(t)) \leq \epsilon\}.$$

Step 3 The estimate of α is given by $\hat{\alpha}(t_{M2})$. If $\hat{\alpha}(t_{M2}) < 0$, let $\hat{\alpha}(t_{M2}) = 0$. If $\hat{\alpha}(t_{M2}) > 1$, let $\hat{\alpha}(t_{M2}) = 1$.

Example 2.3.2 (Merton and Kou Models)**.** The same parameter settings as in Example 2.1.2 for the Merton Model are used. We found that using $M = 1,000$ re-samples already provides good estimates of the variance of $\hat{\alpha}(t)$. The resample size is equal to the sample size, i.e. 2500. The set of values of t we consider is $t = 0, 1, 2, ..., 200$.

Figure 2.10 shows one curve of $\mathrm{Std}_B(\hat{\alpha}(t))$, where one sample is simulated and used for the bootstrap procedure. We can see that the bootstrapped estimate of standard deviation almost coincides with the true standard deviation. And it has been checked that, for different simulated samples, the bootstrapped estimate of standard deviation is stable.

In the following we propose two different methods of selecting t, both based on the bootstrapped estimate of the variance of $\hat{\alpha}(t)$.

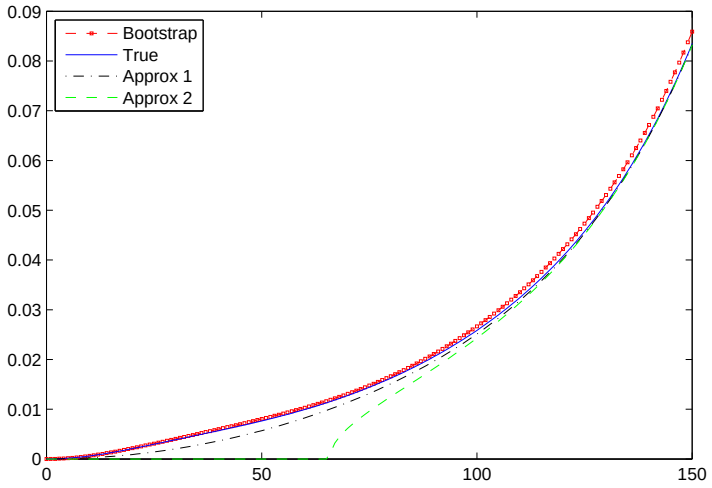
For the first method, we introduce an "approximation" of the variance of $\hat{\alpha}(t)$ by

$$\mathrm{Std}_{\mathrm{Approx1}}(\hat{\alpha}(t)) = \sqrt{\frac{1}{2n}\,(e^{\sigma^2 t^2} + e^{-\sigma^2 t^2} - 2)}, \tag{2.40}$$

which is obtained by assuming $R(\cdot) = 0$ and $\alpha = 0$ in the formula of $\mathrm{Std}(\hat{\alpha}(t))$ given in (2.29). Then we may select t by comparing the two curves: the approximation

$\text{Std}_{\text{Approx1}}(\hat{\alpha}(t))$ and the bootstrapped estimate $\text{Std}_B(\hat{\alpha}(t))$, both of which are available. More specifically, as α is typically small, the value of t for which $R(\cdot)$ gets close to zero should also be the value of t starting from which the two curves become close. From Figure 2.10, we can choose $t = 75$, or equivalently $\text{Std}_B(\hat{\alpha}(t)) = 0.015$. This actually defines the threshold ϵ in the above Step 2 to be $\epsilon = 0.015$.

Figure 2.10: One bootstrapped estimate and two approximations of $\text{Std}(\hat{\alpha}(t))$



Note that in Figure 2.10 we also plot another approximation, which is based on

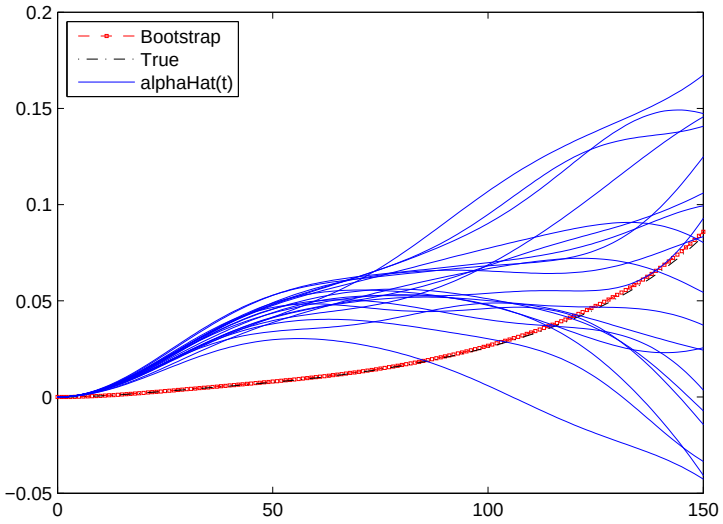
$$\text{Std}_{\text{Approx2}}(\hat{\alpha}(t)) = \sqrt{\frac{1}{2n}\,(e^{\sigma^2 t^2} - 2)}. \tag{2.41}$$

This one is obtained by further ignoring the second term, $e^{-\sigma^2 t^2}$, in (2.29). We can see that the intersection of t-axis and the curve of this approximation defines t_{M1}, i.e. the selected value of t based on Method 1. If we compare the two approximations, the first one is more accurate than the second one, and moreover, the first one is always positive for $t > 0$.

For the second method of selecting t, we use the curve of $\hat{\alpha}(t)$ itself. To illustrate the idea, in Figure 2.11 we add 20 curves of $\hat{\alpha}(t)$, which corresponds to 20 different

samples. We can see that a value between 60 and 80 would be appropriate for the selection of t, which corresponds to $\text{Std}(\hat{\alpha}(t))$ between 0.01 and 0.02. At those values of t, the standard deviation is more or less one third of the level of $\hat{\alpha}(t)$, which is considered appropriate.

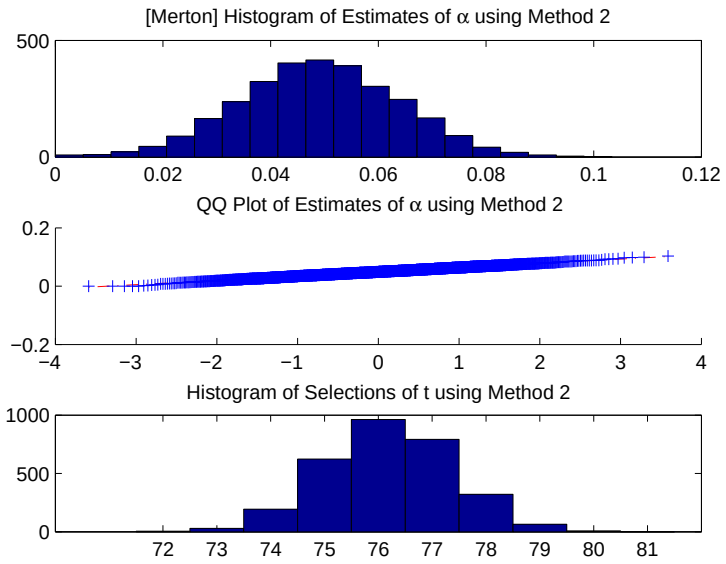
Figure 2.11: One bootstrapped estimate of $\text{Std}(\hat{\alpha}(t))$ and twenty curves of $\hat{\alpha}(t)$



In the simulation we cannot select t manually for every sample. Since both of the above methods lead to $\epsilon = 0.015$ as the threshold value for $\text{Std}_B(\hat{\alpha}(t))$, in the following we apply the above 3-step procedure with setting $\epsilon = 0.015$. Repeat the procedure $3,000$ times. Figure 2.12 shows the histogram and the Q-Q plot of $\hat{\alpha}(t_{M2})$, and the histogram of the selected values t_{M2}. From the plots we see that the estimates $\hat{\alpha}(t_{M2})$ are close to being normally distributed, and most of the selected values of t are between 74 and 78.

Table 2.4 provides a summary of the statistics for $\hat{\alpha}(t_{M2})$. The mean is 0.0488, which is quite close to the true value $\alpha = 0.05$; and the standard deviation is 0.0147, which is close to the pre-assigned level $\epsilon = 0.015$. Comparing these results with

Figure 2.12: Histogram of $\hat{\alpha}(t_{M2})$ and t_{M2} using Method 2 [Merton]



those obtained from Method 1, we can see that the bias decreases and the variance increases.

Table 2.4: Estimation results for $\hat{\alpha}(t_{M2})$ using Method 2 [Merton]

	Mean	Std	25% quantile	Median	75% quantile	MAD
t_{M2}	76.1953	1.2513	75.0000	76.0000	77.0000	1.0000
$\hat{\alpha}(t_{M2})$	0.0488	0.0147	0.0390	0.0488	0.0586	0.0098

For the Kou model, the same parameter settings as in Example 2.1.3 are used. As in Merton model, we use $M = 1,000$ resamples with the resample size equal to the sample size, the threshold value for $\mathrm{Std}_B(\hat{\alpha}(t))$ is set as $\epsilon = 0.015$, and repeat 3,000 times. Figure 2.13 shows the histogram and the Q-Q plot of the estimates $\hat{\alpha}(t_{M2})$, and the histogram of the selected values t_{M2}. The estimates $\hat{\alpha}(t_{M2})$ are close to being normally distributed, and most of the selected values of t are between

75 and 78. Table 2.5 provides a summary of the statistics for $\hat{\alpha}(t_{M2})$. The mean value is 0.0420 and the standard deviation is 0.0146 (close to the pre-assigned level $\epsilon = 0.015$). The estimates of α are still biased downward, and the reason has been provided in Example 2.2.7.

Figure 2.13: Histogram of $\hat{\alpha}(t_{M2})$ and t_{M2} using Method 2 [Kou]

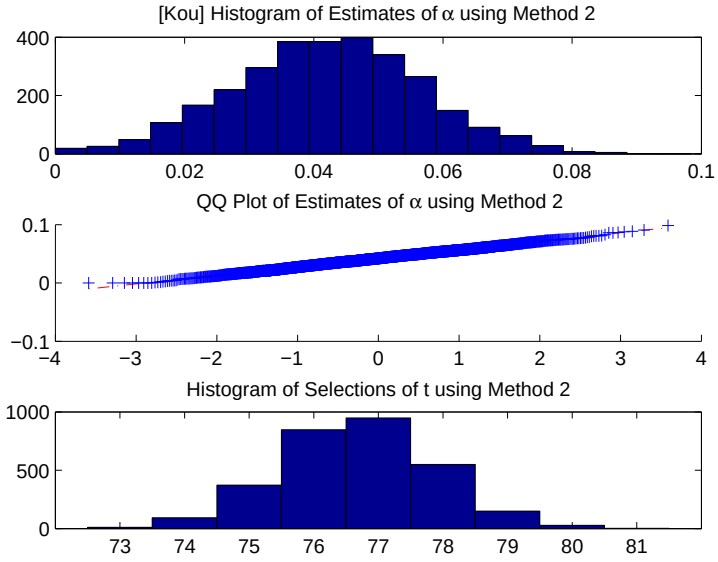


Table 2.5: Estimation results for $\hat{\alpha}(t_{M2})$ using Method 2 [Kou]

	Mean	Std.	25% quantile	Median	75% quantile	MAD
t_{M2}	76.6750	1.2173	76.0000	77.0000	77.0000	1.0000
$\hat{\alpha}(t_{M2})$	0.0420	0.0146	0.0324	0.0422	0.0521	0.0099

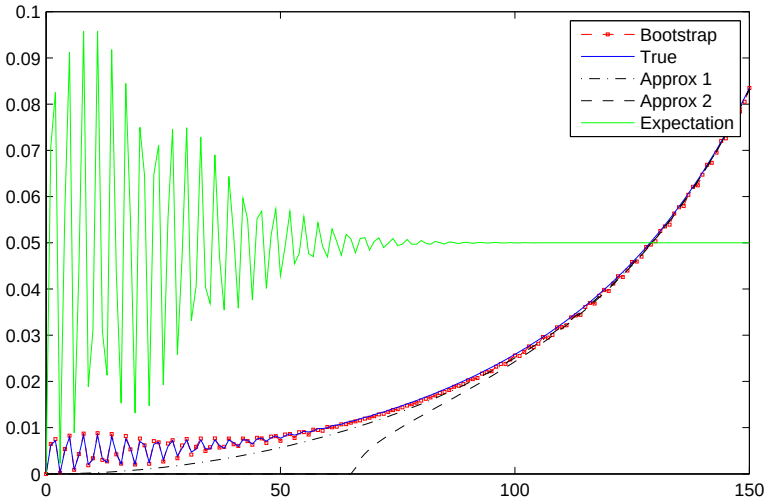
The following example shows that Method 2 also works for some "unusual" cases.

Example 2.3.3 (Merton Model with some extreme values of the parameter)**.** We

use the same parameter settings as in Example 2.1.2 for the Merton Model, except that μ_X is changed from 0 to 2, noting that 2 is very big and "unusual" for the financial data.

Similarly to Figure 2.10, Figure 2.14 shows the bootstrapped standard deviation $\text{Std}_B(\hat{\alpha}(t))$, the true standard deviation $\text{Std}(\hat{\alpha}(t))$ and two approximations of the standard deviation: $\text{Std}_{\text{approx},1}(\hat{\alpha}(t))$ and $\text{Std}_{\text{approx},2}(\hat{\alpha}(t))$. Moreover, we add one more curve, which is the expectation of $\hat{\alpha}(t)$. As we can see, the curves of the expectation and the true standard deviation are oscillating, since $R(t) = \cos(\mu_X t)e^{-\frac{1}{2}\sigma_X^2 t^2}$ is oscillating when $\mu_X \neq 0$. This might add difficulty to the selection of t. However, the bootstrapped standard deviation still estimates the true standard deviation well, and we can apply the same rule as before: to compare the first approximation, $\text{Std}_{\text{approx},1}(\hat{\alpha}(t))$, with the bootstrapped $\text{Std}_B(\hat{\alpha}(t))$ to select the value of t using the point where the two curves gets close. From the figure, we may choose $t_{M2} = 80$.

Figure 2.14: Oscillating curves of standard deviation and expectation of $\hat{\alpha}(t)$



For the cases where the true standard deviation oscillates, like in this example, there is a more direct method to select t: find the point where the bootstrapped

$\text{Std}_B(\hat{\alpha}(t))$ "stops" oscillating. This is because it is $R(t)$ that makes the standard deviation (and the expectation) oscillate; and thus when $R(t)$ gets close to zero, the oscillation should seem to disappear (but not exactly, since it is just the magnitude of the oscillation becomes too small so that it will not be seen). We can check that, in Figure 2.14, the bootstrapped $\text{Std}_B(\hat{\alpha}(t))$ "stops" oscillating at $t_{M2} = 80$.

Now we summarize the basic features of Method 2 as follows:

- Unlike Method 1, the selected value of t using Method 2 is sample-dependent, and thus it may incorporate more information, such as the sample size n, the jump size distribution and the jump frequency. Unlike Method 3 introduced below, there is no restriction on the underlying jump size distribution for Method 2.

- However, the disadvantage is that this method depends heavily on the threshold ϵ for $\text{Std}_B(\hat{\alpha}(t))$, and the selection of ϵ described in this section is not completely "automatic". To make it worse, when later in Chapter 4 we consider the case that the diffusion parameters are unknown and estimated using some robust procedure, the curve of the bootstrapped variance $\text{Std}_B(\hat{\alpha}(t))$ is no longer stable across different samples, and $\text{Std}_B(\hat{\alpha}(t))$ may become far from the true $\text{Std}(\hat{\alpha}(t))$ as t increases.

Therefore, we suggest to use this method as a follow-up step to check whether $\text{Std}_B(\hat{\alpha}(t))$ is acceptable at some value of t selected by using another method, for example, using Method 3 introduced below.

2.3.3 Method 3: First Stationary Point

The *stationary points* of a function are defined as the points at which the first derivative of the function is zero. In this section, we select t by using the first stationary point of $\hat{\alpha}(t), t > 0$. This method works under the following assumption:

Assumption (B1) : $R(t), t > 0$ is strictly decreasing.

Among the financial models, the Merton model with $\mu_X = 0$ and the Kou model satisfy the assumption (B1). For the models where assumption (B1) is not satisfied, see Remark 2.3.10 below.

Now we investigate some properties of the curve of $\hat{\alpha}(t), t \geq 0$:

(P1) As we know, $R(0) = 1$ and $R(t) \to 0$ as $t \to \infty$. Then, under the assumption (B1), $R(t)$ is strictly decreasing from one and goes to zero. Since $\mathbb{E}(\hat{\alpha}(t)) = \alpha(1 - R(t))$, we introduce a notation

$$\alpha(t) \equiv \alpha(1 - R(t)),$$

which is equal to the expectation of $\hat{\alpha}(t)$. Then, under the assumption (B1), $\alpha(t)$ is strictly increasing from zero to α.

(P2) The $\mathrm{Var}(\hat{\alpha}(t))$ in (2.29) is small when t is small, and $\mathrm{Var}(\hat{\alpha}(0)) = 0$. This implies that $\hat{\alpha}(t)$ estimates $\alpha(t)$ well when t is small.

Due to property (P1), $\alpha(t), t \geq 0$ is strictly increasing from zero to α, and thus has no stationary point. However, its estimator $\hat{\alpha}(t)$ will be "oscillating" because of the nature of the trigonometric functions, and thus $\hat{\alpha}(t)$ will likely have one or more stationary points. Due to property (P2), $\hat{\alpha}(t)$ approximates $\alpha(t)$ well when t is small, so the first stationary point is approximately the smallest value of t such that $\alpha(t)$ gets close to α, especially when $\alpha(t)$ converges to α quickly (or equivalently, $R(t)$ converges to zero quickly). Thus we propose to select t by

$$t_{M3} = \text{first stationary point of } \hat{\alpha}(t), \quad t \in (0, t_{\max}], \tag{2.42}$$

where $t_{\max}$ is an appropriately selected upper bound of t. This is the basic idea of Method 3.

In the following we discuss some issues in the implementation of this method. In our empirical studies we have found that the shape of $\hat{\alpha}(t)$ varies and sometimes the stationary point of $\hat{\alpha}(t)$ may not exist (for t in an appropriate range that we consider). For this case, we could use $t_{\max}$ as the selected t. However, to improve the method, in such cases we would rather use the point with the smallest first derivative. The reason of doing this can be seen from Example 2.3.7 below.

The following Lemma 2.3.4 shows that $\alpha'(t)|_{t=0} = 0$, so we first need to exclude an interval starting at zero, say $[0, t_{\min})$ for some $t_{\min} > 0$. One way to select $t_{\min}$ is to use the "inflection point". Note that the *inflection points* of a function are the points at which the second derivative of the function, assuming it exits, is zero.

Lemma 2.3.4. *For any random variable X with $\mathbb{E}|X| < \infty$, we have $R'(t)|_{t=0} = 0$, where $R(t) \equiv \Re(\varphi_X(t))$.*

Proof. By Theorem C.0.5 (b) in the Appendix and letting $k = 1$, we have that if $\mathbb{E}|X| < \infty$, then $\varphi'(t)|_{t=0} = i\mathbb{E}(X)$. Taking the real part of both sides, we obtain the result. (Note that in the last step we assumed the ability to interchange the order of taking the real part and the differentiation of a complex-valued function. The reason we can do so is given as follows. The definition of the differentiation of a complex-valued function is analogous to the one for a real-valued function, i.e. $f'(z_0) = \lim_{z \to z_0} \frac{f(z)-f(z_0)}{z-z_0}$. Since in the current problem, the argument t is a real number, so we can interchange the order freely.) $\boxtimes$

The above Lemma 2.3.4 leads to the following result.

Lemma 2.3.5. *For any absolutely continuous random variable X with $\mathbb{E}|X| < \infty$, there exists at least one inflection point of $R(t)$, where $R(t) \equiv \Re(\varphi_X(t))$.*

Proof. Since $R(t) \to 0$ as $t \to \infty$ by Lemma 2.1.1 and using the fact that $R(0) = 1$, there exists some interval (a, b) for $0 < a < b$ on which $R'(t) < 0$. Furthermore,

$R'(0) = 0$ by Lemma 2.3.4. Then there exists an interval (c, d) for $0 < c < d$ on which $R''(t) < 0$. Assume that there exists no inflection point of $R(t)$. Then $R''(t) \leq 0$ for all $t > 0$. That is, $R'(t)$ is non-increasing for all $t > 0$. Therefore, there exist $\delta < 0$ such that $R'(t) < \delta$ for all $t > a$. This implies that $R(t) \to -\infty$ as $t \to \infty$, which contradicts the fact that $R(t) \to 0$ as $t \to \infty$. $\boxtimes$

Based on Lemma 2.3.5, there exists at least one inflection point of $\alpha(t)$. Using this result and property P2, we may let

$$t_{\min} := \text{ first inflection point of } \hat{\alpha}(t),\ t \in [0, t_{\max}). \tag{2.43}$$

Since $\hat{\alpha}(t)$ is strictly increasing and concave upward on the interval $t \in [0, t_{\min})$, we have $t_{\min} < t_{M3}$.

In our empirical studies we have found that the above-defined $t_{\min}$ exists for almost all samples. Very rarely (less than 0.01% for the model in Example 2.3.7 below) when $t_{\min}$ does not exist, we could let $t_{M3} = t_{\max}$. However, in this situation, $\hat{\alpha}(t)$ is strictly increasing and always concave upward for $t \in [0, t_{\max})$, so $\hat{\alpha}(t_{\max})$ is typically bias up high. Therefore, we would rather not apply this method to such cases, but refer to other methods, like Method 2.

Based on the above ideas we propose the following procedure for Method 3:

Step 1 Determine an appropriately selected upper bound of t, which is denoted by $t_{\max}$.

Step 2 Determine the lower bound of t by (2.43):

$$t_{\min} := \text{ first inflection point of } \hat{\alpha}(t),\ t \in [0, t_{\max}).$$

If $t_{\min}$ does not exist, then set flag=0, and stop.

Step 3 Select t based on the "first stationary point":

3-1 If there exists a stationary point of $\hat{\alpha}(t)$ for $t \in [t_{\min}, t_{\max}]$, then we select t by (2.42):

$$t_{M3} = \text{ first stationary point of } \hat{\alpha}(t),\ t \in [t_{\min}, t_{\max}].$$

Then set flag=1, and go to Step 4.

3-2 If there exists no stationary point of $\hat{\alpha}(t)$ for $t \in [t_{\min}, t_{\max}]$, then we let t_{M3} be the point belonging to $[t_{\min}, t_{\max}]$ with the smallest first derivative of $\hat{\alpha}(t)$. Then set flag=2, and go to Step 4.

Step 4 The estimate of α is given by $\hat{\alpha}(t_{M3})$. If $\hat{\alpha}(t_{M3}) < 0$, then we let $\hat{\alpha}(t_{M3}) = 0$, and set flag=31; if $\hat{\alpha}(t_{M3}) > 1$, then we let $\hat{\alpha}(t_{M3}) = 1$, and set flag=32.

Remark 2.3.6. For the selection of $t_{\max}$, it would be enough to use $t_{\max} = 100$ for the financial data with yearly volatility $\sigma_0 \geq 20\%$, since R(t) is already close to zero at $t = 100$ in these cases. However, for markets with $\sigma_0 < 20\%$, we might need to increase $t_{\max}$. This is because those markets are typically less risky and the jump size are typically smaller, and then $R(t)$ becomes close to zero at a larger value of t.

In the remaining of the thesis, we always set the upper bound to be $t_{\max} = 100$ (since we always set $\sigma_0 = 20\%$), unless stated otherwise.

Example 2.3.7 (Merton Model)**.** We use the same parameter settings as in Example 2.1.2 for the Merton Model, and repeat the procedure 100,000 times. For the lower bound $t_{\min}$, the result shows that the mean is 26.6 (while the "theoretical" value is 26.2), and the st.dev. is 3.1.

Table 2.6 shows the number of samples for each "flag". We see that, for more than a half of samples, the first stationary point exists. There are 6 (i.e. 0.006%) samples where the first inflection point does not exist, and we exclude them from consideration. There are 41,134 samples with flag=2, among which only 8,350 (i.e.

around 8% of all observations) samples having $t_{M3} = t_{\max}$. This explains why we do not use $t_{\max}$ as the selected t in the above Step 3-2, for which the first stationary point of $\hat{\alpha}(t)$ does not exist and $\hat{\alpha}(t)$ is strictly increasing for $t \in [t_{\min}, t_{\max}]$.

Table 2.6: Number of flags [Merton]

flag	1	2	31	32	0	sum
No. of samples	58,860	41,134	0	0	6	100,000

Figure 2.15 shows the histograms of t_{M3} and $\hat{\alpha}(t_{M3})$. For t_{M3}, there is a high "peak" in the last bin, which occurs because, as indicated above, there are around 8% of samples having $t_{M3} = t_{\max}$. The histogram of $\hat{\alpha}(t_{M3})$ shows a slightly heavy right tail. This is because for the samples where $t_{M3} = t_{\max}$ the obtained estimates $\hat{\alpha}(t_{M3})$ tend to be larger than the true value.

Figure 2.15: Histograms of $\hat{\alpha}(t_{M3})$ and t_{M3} using Method 3 [Merton]

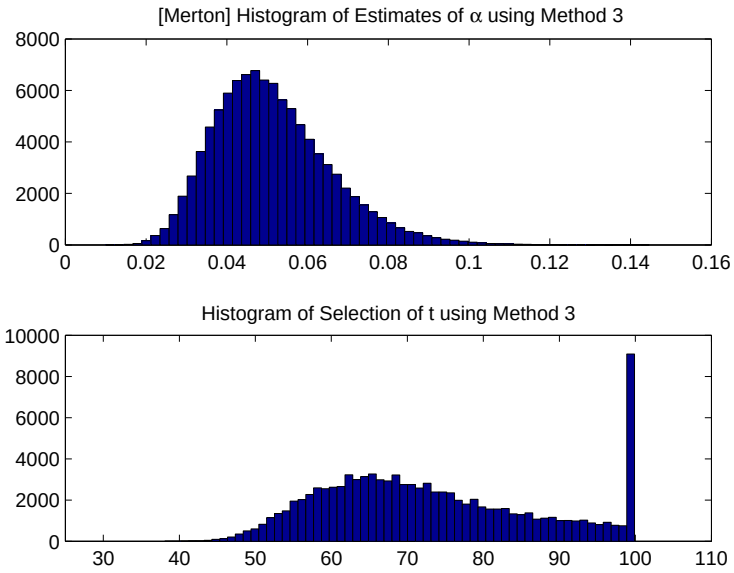


Table 2.7 provides the statistics of the estimates. The mean value of $\hat{\alpha}(t_{M3})$

is 0.0512 and the standard deviation is 0.0145. It is biased up, due to the right heavy tail. The result is better than Method 1, and is comparable with Method 2. However, this Method 3 is more "automatic" than Method 2 for implementation.

Table 2.7: Estimation results for $\hat{\alpha}(t_{M3})$ using Method 3 [Merton]

	Mean	Std	25% quantile	Median	75% quantile	MAD
t_{M3}	73.3072	14.2709	62.3000	70.8000	83.2000	9.9000
$\hat{\alpha}(t_{M3})$	0.0512	0.0145	0.0409	0.0495	0.0597	0.0093

To see how Method 3 performs for different scenarios of parameter settings, Table 2.8 shows estimation results for different jump size distribution (implied by σ_X), jump frequency λ and sample size n. Note that here we record $\hat{\lambda}(t_{M3})$ instead of $\hat{\alpha}(t_{M3})$, where $\hat{\lambda}(\cdot)$ is defined in Chapter 3 by equation (3.3): $\hat{\lambda}(t) = \hat{\alpha}(t)/\Delta$ (in that section it is explained that $\hat{\lambda}(\cdot)$ estimates the intensity parameter λ). The reason for doing so is that we will use this table in later chapters where λ is the parameter we focus on. For every scenario in this table, $t_{\max} = 100$ is used. Note that the true value is $\lambda = 12.5$. The values inside the parenthesis are the corresponding standard deviations.

Table 2.8: Estimation results using Method 3 for different scenarios [Merton]

		$\sigma_X \approx 3.79\%$		$\sigma_X = 10\%$		$\sigma_X = 60\%$	
		$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$
n=1000	t_{M3}	68.9(15.9)	74.5(14.0)	35.8(13.5)	36.7(12.1)	5.7 (1.7)	6.0 (1.7)
	$\hat{\lambda}_{M3}$	13.0 (5.3)	25.6 (6.5)	13.6 (3.2)	26.3 (3.9)	13.1 (2.2)	25.9 (3.0)
n=2500	t_{M3}	73.3(14.3)	78.2(12.5)	37.6(13.2)	38.5(12.0)	6.0 (1.7)	6.2 (1.7)
	$\hat{\lambda}_{M3}$	12.8 (3.6)	25.3 (4.3)	13.2 (2.1)	25.9 (2.5)	12.9 (1.4)	25.6 (1.9)
n=10000	t_{M3}	79.1(12.1)	83.4(10.5)	40.3(13.2)	40.8(11.8)	6.5 (1.7)	6.6 (1.6)
	$\hat{\lambda}_{M3}$	12.7 (2.0)	25.1 (2.4)	12.9 (1.1)	25.4 (1.3)	12.7 (0.7)	25.3 (0.9)

From this table, we can see the following:

1. As n increases, the bias and variance of $\hat{\lambda}(t_{M3})$ decrease.

2. As λ increases, the bias and variance of $\hat{\lambda}(t_{M3})/\lambda$ decrease.

3. As σ_X increases, the bias and variance of $\hat{\lambda}(t_{M3})$ should decrease. This is intuitive since a larger σ_X implies easier classification between the jump observations and diffusion observations. However, the presented results are not consistent with this relation when σ_X changes from 3.79% to 10%. This is just because of different degrees of influence of $t_{\max}$ on $\hat{\lambda}(t_{M3})$ for different σ_X (see item 6 below for more details).

4. As n increases, the mean value of t_{M3} increases and the variance decreases. This matches our expectation, since a larger n implies a smaller variance of $\hat{\alpha}(t)$ for any t, i.e. $\hat{\alpha}(t)$ becomes closer to $\alpha(t)$ on average for any t, and then the first stationary point of $\alpha(t)$ would occur at a larger value of t. This fact suggests that the estimator $\hat{\lambda}(t_{M3})$ may possess some asymptotic properties, such as the consistency and the asymptotic normality; however, the speed of the increase of t is difficult to determine.

5. As λ increases, the mean value of t_{M3} increases and the variance decreases.

6. As σ_X increases, the mean value of t_{M3} decreases and the variance decreases. (There are some exceptions from $\sigma_X \approx 3.79\%$ to $\sigma_X = 10\%$ due to the same reason as given in the above item 3.) This is because of the following property of $R(t)$: when σ_X increases, $R(t)$ gets close to zero at smaller values of t. This also explains that, as σ_X increases, $t_{\max} = 100$ has less effect on the selection of t. For example, Figure 2.16 shows the histograms for the case when $\sigma_X = 0.1$ (and $\lambda = 12.5$, $n = 2500$), from which we see that $t_{\max} = 100$ plays almost no role in the selection of t_{M3}. Figure 2.17 shows the histograms for the case when $\sigma_X = 0.6$ (and $\lambda = 12.5$, $n = 2500$), from which we see that $t_{\max} = 100$ plays no role at all in the selection of t_{M3}.

Figure 2.16: Histograms of $\hat{\alpha}(t_{M3})$ and t_{M3} when $\sigma_X = 0.1$ [Merton]

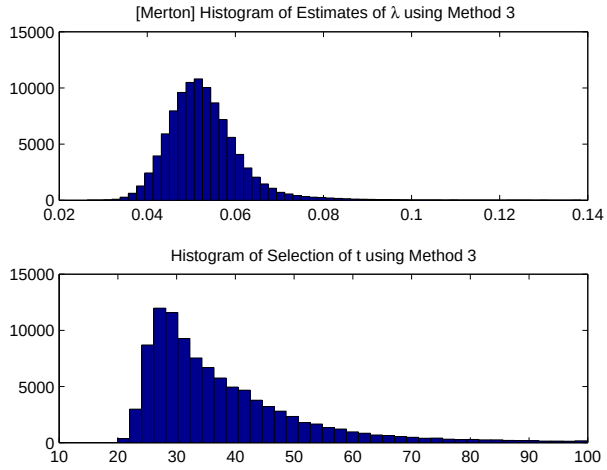
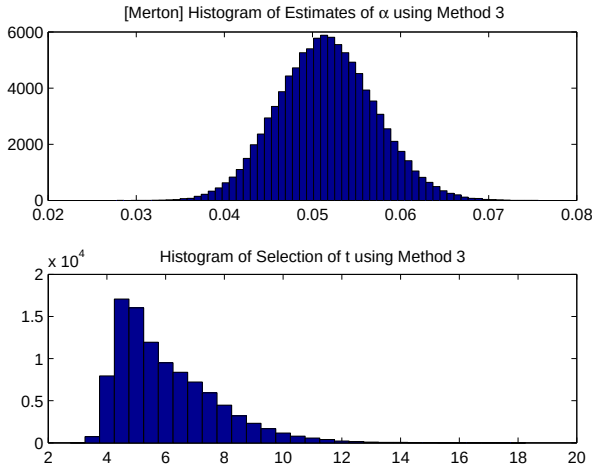


Figure 2.17: Histograms of $\hat{\alpha}(t_{M3})$ and t_{M3} when $\sigma_X = 0.6$ [Merton]



The following remark emphasizes some aspect of the role that $t_{\max}$ plays, as discovered from the above example.

Remark 2.3.8. As the only tuning parameter for the implementation of Method 3, $t_{\max} = 100$ may have different degrees of influence on the estimation results

for different scenarios of parameter settings. This can be seen from the above Merton model example: when $\sigma_X = 3.79\%$, $t_{\max} = 100$ plays an important role to "truncate" for the selection of t; but when $\sigma_X = 60\%$, $t_{\max} = 100$ plays no role at all because all the selected values of t are much smaller than 100. This may have an impact on other relations, and in some cases it may obscure them. Examples of such situations are provided in items 3 and 6 above. For these situations, we may also say that there is a "confounding effect" between $t_{\max}$ and other factors.

Example 2.3.9 (Kou Model)**.** For the Kou model, we use the same parameter settings as in Example 2.1.3. As before, set $t_{\max} = 100$, and repeat the procedure 100,000 times. From the simulation results we have that the mean of $t_{\min}$ is 16.0 (the "theoretical" value is 15.4), and the standard deviation is 3.9.

Table 2.9 shows the number of samples for each flag. As we can see, there is no sample leading to flag=0, i.e. the first inflection point exists for all samples. There are 34,961 samples with flag=2, among which 3,727 (i.e. about 4% of all observations) samples having $t_{M3} = t_{\max}$.

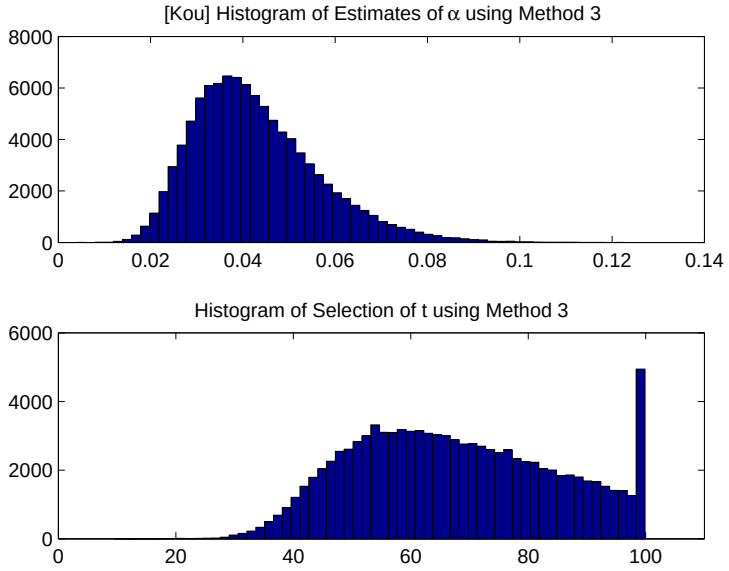
Table 2.9: Number of flags [Kou]

flag	1	2	31	32	0	sum
No. of samples	65,039	34,961	0	0	0	100,000

Figure 2.18 shows the histograms of t_{M3} and $\hat{\alpha}(t_{M3})$. Similarly, we can see that for t_{M3} there is a high "peak" in the last bin, and $\hat{\alpha}(t_{M3})$ has a heavy right tail.

Table 2.10 provides the statistics of the estimates. The mean value of $\hat{\alpha}(t_{M3})$ is 0.0429 and the standard deviation is 0.0137. It is biased down (although there is a heavy right tail), unlike in the Merton model. This is because $R(t)$ in the Kou model converges to zero very slowly, which can be seen from the fact that $\mathbb{E}[\hat{\alpha}(t_{\max})] = 0.0449$ only when $t_{\max} = 100$. (However, $t_{\max}$ can not be chosen too large because otherwise the variance would be very large). Similarly to the Merton

Figure 2.18: Histogram of $\hat{\alpha}(t_{M3})$ and t_{M3} using Method 3 [Kou]



model, for the Kou model the results based on Method 3 are better than for Method 1 and comparable with Method 2. However, Method 3 is more "automatic" than Method 2.

Table 2.10: Estimation results for $\hat{\alpha}(t_{M3})$ using Method 3 [Kou]

	Mean	Std	25% quantile	Median	75% quantile	MAD
t_{M3}	67.7136	17.1301	54.2000	66.3000	80.8000	13.1000
$\hat{\alpha}(t_{M3})$	0.0429	0.0137	0.0329	0.0407	0.0507	0.0087

Table 2.11 shows the estimation results for different parameters of the jump size distribution, jump frequency λ and sample size n. Note that in the two different jump size distributions, $p = 0.3$ is the same. The six findings we presented for the Merton model still hold for the Kou model.

Table 2.11: Estimation results using Method 3 for different parameters [Kou]

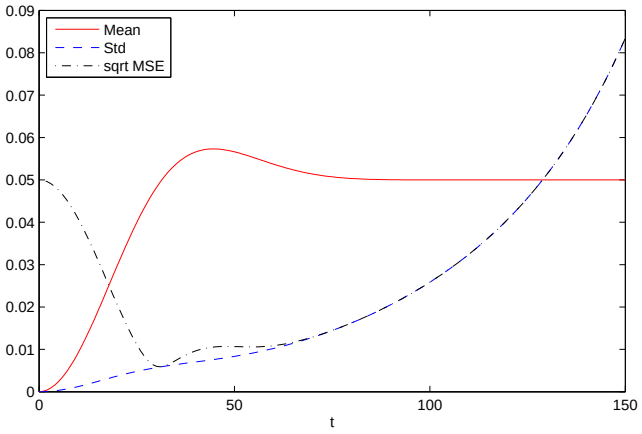
		$\eta_+ = 1/0.02$, $\eta_- = 1/0.04$		$\eta_+ = 1/0.1$, $\eta_- = 1/0.2$	
		$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$
n=1000	t_{M3}	60.3(19.0)	68.6(16.8)	16.3 (5.8)	18.2 (5.4)
	$\hat{\lambda}_{M3}$	10.6 (4.8)	21.5 (6.3)	11.2 (2.4)	22.8 (3.4)
n=2500	t_{M3}	67.5(17.0)	75.3(14.6)	18.9 (5.4)	21.0 (5.3)
	$\hat{\lambda}_{M3}$	10.7 (3.4)	21.6 (4.4)	11.4 (1.6)	23.1 (2.3)
n=10000	t_{M3}	77.6(14.0)	84.4(11.4)	23.6 (5.4)	26.3 (5.5)
	$\hat{\lambda}_{M3}$	10.8 (2.1)	21.8 (2.6)	11.7 (0.9)	23.7 (1.2)

Remark 2.3.10. When assumption (B1) is not satisfied, i.e when $R(t), t > 0$ is not strictly decreasing, it is still possible to apply Method 3. For example, for the Merton model with all the other parameters the same as in Example 2.1.2, but with μ_X changed from 0 to a number belonging to $[-0.05, 0.05]\backslash\{0\}$, the assumption (B1) is not satisfied. For this case, we may still apply Method 3, but the only problem is that the estimate may be biased up high. See Figure 2.19 for the case when $\mu_X = -0.05$, from which we can see that $\mathbb{E}[\hat{\alpha}(t)]$ is greater than $\alpha = 0.05$ at the first stationary point of $\mathbb{E}[\hat{\alpha}(t)]$. The simulation results show that the mean value of $\hat{\lambda}(t_{M3})$ is $14.7 > 12.5$ and the standard deviation is 2.2. Below we present another example for this case, which we call "Sepp model".

Example 2.3.11 (Sepp model)**.** In the paper by Sepp (2012), the author discussed a jump-diffusion model with the jump size following a mixture of four different normal distributions. In the following, we call this model the "Sepp model". The following are the parameters for this model:

- $\Delta = 1/250$, i.e. daily data.
- The diffusion parameters are $\sigma_0 = 0.1348$, $\mu_0 = -0.0091$.

Figure 2.19: Mean, std and MSE of $\hat{\alpha}(t)$ (Merton model with $\mu_X = -0.05$)



- The jump size follows a mixture of four normal distributions with mean values $\mu_{X,1}$ = -0.0733, $\mu_{X,2}$ = -0.0122, $\mu_{X,3}$ = 0.0203, $\mu_{X,4}$ = 0.1001 and the same standard deviation σ_X = 0.0127; the weights of the four components are p_1 = 0.0208, p_2 = 0.5800, p_3 = 0.3954, p_4 = 0.0038.

- The jump intensity is $\lambda = 46.4444$, which implies $\alpha = \lambda \cdot \Delta = 0.1858$.

As before, we simulate a 10-year data set, i.e. the sample size $n = 2,500$. As we demonstrate below, the assumption (B1) is not satisfied under this model, but we may still consider applying Method 3. Since $\sigma_0 = 0.1348$, which is less than the value 0.2 used in all previous examples, by Remark 2.3.6 we reset the upper bound of t to be $t_{\max}$=200. From the simulation results we have that the mean of the lower bound $t_{\min}$ is 46.3 and the st.dev. is 9.9.

Figure 2.20 shows the mean, st.dev. and MSE of $\hat{\alpha}(t)$. We see that the mean curve is not strictly increasing, which verifies that the assumption (B1) is not satisfied.

Table 2.12 shows the number of samples for each flag. Among samples with flag=2, there are 1,440 samples (1.4% of all observations) having $t_{M3} = t_{\max}$.

Figure 2.20: Mean, std and MSE of $\hat{\alpha}(t)$ [Sepp]

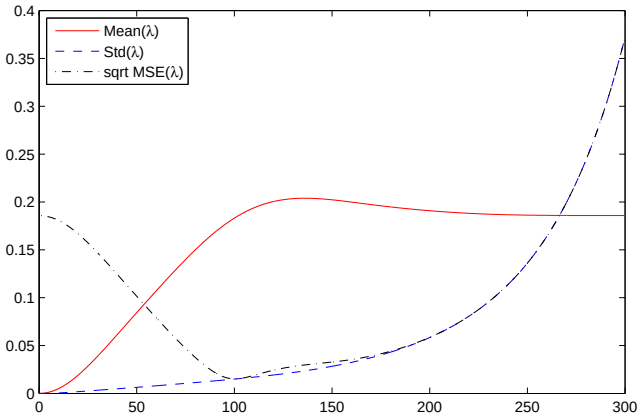


Table 2.12: Number of flags [Sepp]

flag	1	2	31	32	0	sum
No. of samples	81,680	18,320	0	0	0	100,000

Figure 2.21 shows the histograms of t_{M3} and $\hat{\alpha}(t_{M3})$. The last bin in the histogram of t_{M3} is less "peaked" than in the Merton and Kou models, suggesting that the selection of t has less dependence on the upper bound $t_{\max}$. The estimates $\hat{\alpha}(t_{M3})$ still show a heavy right tail somehow.

Table 2.13 summarizes the statistics of the results. Here we add the result for $\hat{\lambda}(t_{M3}) \equiv \hat{\alpha}(t_{M3})/\Delta$. Comparing with the true value $\lambda = 46.4444$, we have that $\hat{\lambda}(t_{M3})$ is biased up by about 12%. This can be explained by two reasons: one is the right heavy tail, and the other one, which is more important, is similar to the argument presented in Remark 2.3.10.

Remark 2.3.12. In Method 3, we select t by using the first stationary point of $\hat{\alpha}(t), t > 0$. However, for a model for which the c.f. of the jump size has only one inflection point, we may also consider selecting t by using the second inflection

Figure 2.21: Histogram of $\hat{\alpha}(t_{M3})$ and t_{M3} using Method 3 [Sepp]

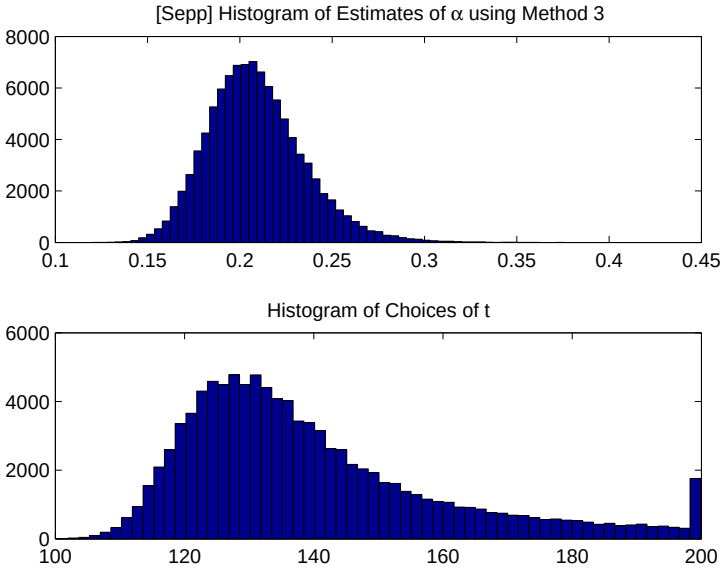


Table 2.13: Estimation results for $\hat{\alpha}(t_{M3})$ using Method 3 [Sepp]

	Mean	Std	25% quantile	Median	75% quantile	MAD
t_{M3}	139.7477	20.0610	125.5000	134.6000	148.6000	10.7000
$\hat{\alpha}(t_{M3})$	0.2087	0.0262	0.1907	0.2065	0.2241	0.0166
$\hat{\lambda}(t_{M3})$	52.1847	6.5584	47.6703	51.6204	56.0142	4.1500

point of $\hat{\alpha}(t), t > 0$. Examples are the Kou model with parameters suitable for the financial data and the Merton model. Unfortunately, using the second inflection point is not as stable as using the first stationary point. It works well for the Merton model. But for models with the c.f. of jump size converging to zero slowly, like the Kou model, it may be biased down significantly. This is because for such models the $\text{Var}[\hat{\alpha}(t)]$ is already large before $\mathbb{E}[\hat{\alpha}(t)]$ becomes close to α, and thus $\hat{\alpha}(t)$ is very oscillating and then the "second inflection point" tends to come earlier than

it should be.

Previously we presented examples where either the normal (or mixed normal) distribution or the double exponential distribution was assumed as the jump size distribution. Now we look at the case where the t distribution, which has heavier tails, is used as the jump size distribution.

Example 2.3.13 (t distribution)**.** The probability density function of a t distribution with $\nu > 0$ degrees of freedom is given by

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

where Γ is the *gamma function*. The expectation is zero for $\nu > 1$, and the variance is $\nu/(\nu-2)$ for $\nu > 2$. The characteristic function is given by

$$\varphi(t) = \frac{K_{\nu/2}\left(\sqrt{\nu}|t|\right)\cdot\left(\sqrt{\nu}|t|\right)^{\nu/2}}{\Gamma(\nu/2)2^{\nu/2-1}} \tag{2.44}$$

for $v > 0$, where $K_\nu(x)$ is the *modified Bessel function of the second kind*.

Suppose W is a random variable which follows a t distribution with ν degrees of freedom, where $\nu > 2$. To make a t distribution comparable to the normal distribution (with mean zero and variance σ_X^2) in the Merton model, we scale this t distribution to make its variance equal to the variance of the normal distribution, while keeping its expectation equal to zero. Following this idea, we let the jump size be

$$X = \sigma_X\sqrt{(\nu-2)/\nu}\cdot W$$

for $\sigma_X > 0$. Thus, X follows a scaled t distribution with mean zero, variance σ_X^2 and the number of degrees of freedom ν. The characteristic function of X is $\varphi_X(t) = \varphi(\sigma_X\sqrt{(\nu-2)/\nu}\cdot t)$, where φ is given by (2.44).

In our numerical example we choose $\nu = 3$. The other parameters are set the same as in Example 2.1.2 for the Merton Model. Figure 2.22 shows twenty curves

of $\hat{\alpha}(t)$. Figure 2.23 shows the expectation, standard deviation, and the square root of the MSE of $\hat{\alpha}(t)$. Comparing with Figure 2.3 for Merton model and Figure 2.4 for Kou model, we can see that the mean curve in Figure 2.23 approaches the true level 0.05 more slowly. This is a consequence of the fact that the t distribution has heavier tails than the normal distribution and the double exponential distribution.

Figure 2.22: Twenty curves of $\hat{\alpha}(t)$ [t distribution]

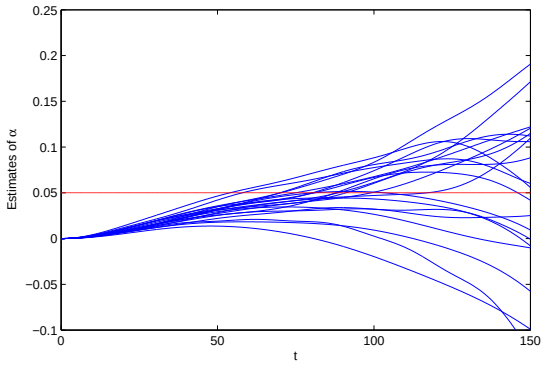
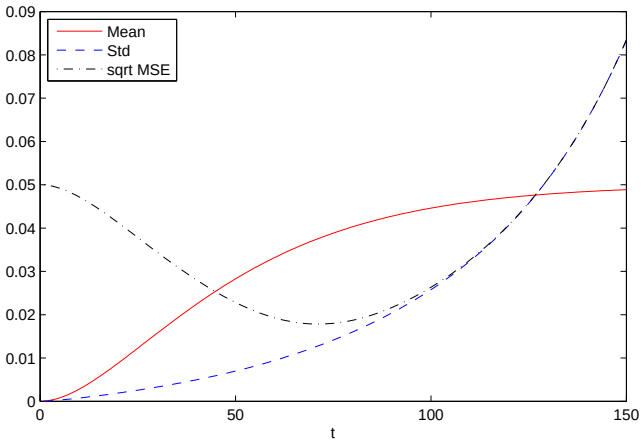


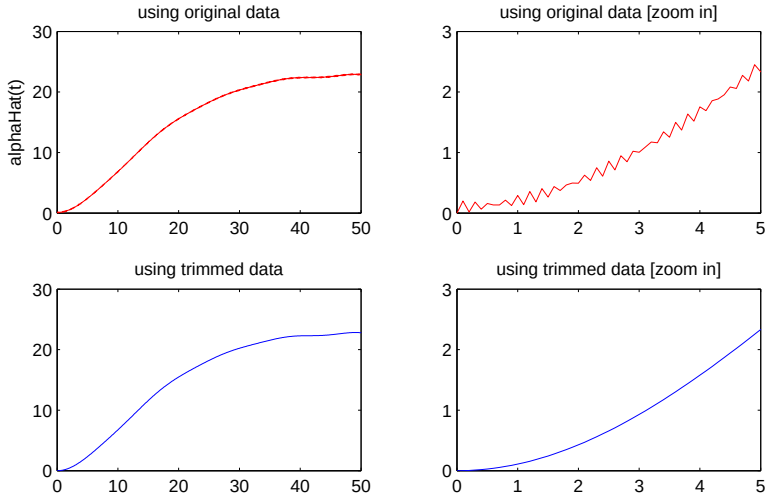
Figure 2.23: Mean, std and MSE of $\hat{\alpha}(t)$ [t distribution]



For the selection of a proper value of t, we consider using Method 3. In our simulation study we have found that for some samples the curve of $\hat{\alpha}(t)$ can show

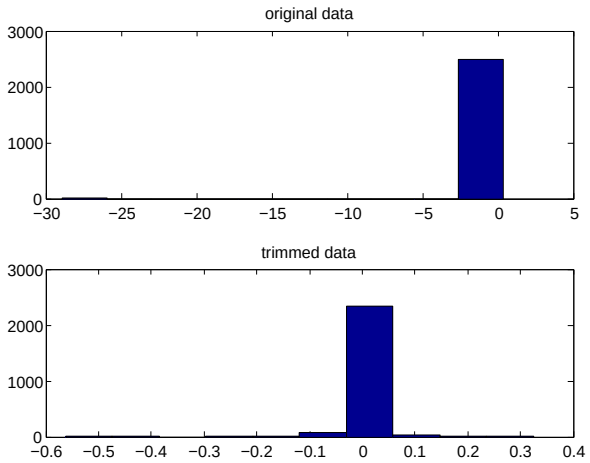
significant local oscillation so that Method 3 returns t_{M3} equal to zero. Such situations occur with probability less that 0.1%. In Figure 2.24 we show one such "extreme" case (with $\sigma_X = 10\%, \lambda = 25, n = 2,500$). The reason for this behavior is the presence of an extremely large jump size simulated from the t distribution. For example, for Figure 2.24, the observation with the largest size is -28.8827, while all the other observations are of sizes less than 0.6. See Figure 2.25 for the corresponding histogram. Then, by using formula (2.24), the cosine curve corresponding to this large observation has such a small period so that there is no other cosine curve that could "smooth" out this oscillation.

Figure 2.24: $\hat{\alpha}(t)$ based on one extreme sample [t distribution]



To solve this issue, we may consider either applying some smoothing techniques to the estimated curve of $\hat{\alpha}(t)$ or using some methods of removing outliers. In our study we have decided to remove the observation with the largest size from every extreme sample, if it exits, and use the "trimmed" data to estimate the jump frequency, and at the end we correct the resulting estimate by increasing its value

Figure 2.25: Histograms based on one extreme sample [t distribution]



by one[¶].

Based on the above revision for Method 3, Table 2.14 shows the estimation results for different values of σ_X, jump frequency λ and sample size n. From the results, we can observe the same pattern as for the Merton model example, although the estimation accuracy is not as good as for the Merton model.

Table 2.14: Estimation results using Method 3 for different scenarios [t distribution]

		$\sigma_X \approx 3.79\%$		$\sigma_X = 10\%$		$\sigma_X = 60\%$	
		$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$
n=1000	t_{M3}	66.4(22.0)	76.2(18.5)	44.8(16.9)	48.0(14.7)	7.5 (2.5)	8.1 (2.2)
	$\hat{\lambda}_{M3}$	9.4 (5.6)	19.7 (7.5)	12.6 (4.4)	24.8(5.5)	12.1(2.7)	24.4(3.8)
n=2500	t_{M3}	74.1(19.5)	83.8(15.0)	49.2(15.1)	52.3(13.2)	8.3 (2.1)	8.9 (1.9)
	$\hat{\lambda}_{M3}$	9.7 (4.1)	20.6 (5.5)	12.6 (3.0)	24.9 (3.7)	12.2(1.7)	24.6(2.2)
n=10000	t_{M3}	85.7(14.1)	92.1(10.2)	55.8(13.3)	58.8(11.6)	9.5 (1.8)	10.1(1.7)
	$\hat{\lambda}_{M3}$	10.4 (2.6)	21.5 (3.3)	12.5 (1.7)	24.9 (2.1)	12.3(0.8)	24.8(1.0)

[¶]However, when we discuss the estimation of jump size distribution in Chapter 5, we still use the original data.

2.3.4 Method 4: Iteration of Minimizing MSE [failed]

In this subsection, we investigate whether a method based on direct minimization of MSE would work. Let us look at the formula (2.30) for MSE($\hat{\alpha}(t)$): only α and the function $R(\cdot)$ are unknown. If we have an initial value of t_0, then we could obtain an estimate of α, which is $\hat{\alpha}(t_0)$. Then, by substituting this value into equation (2.20):

$$\hat{\varphi}_X(t) = \frac{\frac{\hat{\varphi}_Y(t)}{\varphi_Z(t)} - (1 - \hat{\alpha})}{\hat{\alpha}} \tag{2.45}$$

we would obtain an estimate of the function $R(t)$ by

$$\hat{R}(t) = \Re(\hat{\varphi}_X(t)). \tag{2.46}$$

Thus, we can have an estimate of MSE($\hat{\alpha}(t)), t \geq 0$. Then we minimize this estimated MSE($\hat{\alpha}(t)$) to obtain a new value of t. This procedure may be repeated till it converges. Fortunately the inverse Fourier transform is not involved in this procedure since we only need $\hat{R}(t)$.

To make it precise, we write this procedure in the following steps:

Step 1: Given an initial value of t_0, use equation (2.24) to obtain an estimate of α, denoted by $\hat{\alpha}_0 \equiv \hat{\alpha}(t_0)$. (Or, an initial value of $\hat{\alpha}_0$ is given directly.)

Step 2: Substitute $\hat{\alpha}_0$ into (2.45) to obtain $\hat{\varphi}_X(t)$ and then $\hat{R}(t) = \Re(\hat{\varphi}_X(t))$.

Step 3: Substitute $\hat{\alpha}_0$ and $\hat{R}(t)$ into the MSE formula (2.30) to obtain the empirical version of MSE($\hat{\alpha}(t)$), and denote it by $\widehat{\text{MSE}(\hat{\alpha}(t))}$. Then minimize $\widehat{\text{MSE}(\hat{\alpha}(t))}$ with respect to t to obtain a new value of t, denoted by t_1. Repeat the above till it converges.

At a first glance, this method should work. However, the values of t_0 and $\hat{\alpha}_0$ may decrease in every iteration, and finally α could be underestimated significantly, or even possibly $\hat{\alpha}_0$ can go to zero. Some explanations are provided in the following.

In Step 3, an alternative way, which leads to the same result, to obtain the empirical version of $\text{MSE}(\hat{\alpha}(t))$ is to obtain the empirical version of the mean and variance of $\hat{\alpha}(t)$ first by using formulas (2.27) and (2.29). This becomes

$$\widehat{\mathbb{E}[\hat{\alpha}(t)]} = \hat{\alpha}_0 - \hat{\alpha}_0 \Re\left(\frac{\hat{\varphi}_V(t) - (1-\hat{\alpha}_0)}{\hat{\alpha}_0}\right) = 1 - \Re\left(\hat{\varphi}_V(t)\right) \equiv \hat{\alpha}(t)$$

and

$$\widehat{\text{Var}(\hat{\alpha}(t))} = \frac{1}{2n}\left(e^{\sigma^2 t^2} + e^{-\sigma^2 t^2}\left(1 - \hat{\alpha}(2t)\right) - 2\left(1 - \hat{\alpha}(t)\right)^2\right) \equiv h(\hat{\alpha}(t)),$$

where $h(\cdot)$ is a function defined above. Note that neither of the above two equations depend on the initial value $\hat{\alpha}_0$. In other words, no matter what the initial value $\hat{\alpha}_0$ is, these two functions $\widehat{\mathbb{E}[\hat{\alpha}(t)]}$ and $\widehat{\text{Var}(\hat{\alpha}(t))}$ would remain the same. Then the above three steps can be rewritten in a more compact way as follows:

Step 1: Given t_0, we obtain an estimate $\hat{\alpha}_0 \equiv \hat{\alpha}(t_0)$.

Step 2: The new t is given by

$$t_1 = \min_t \widehat{\text{MSE}(\hat{\alpha}(t))} = \min_t \left\{(\hat{\alpha}(t) - \hat{\alpha}_0)^2 + h(\hat{\alpha}(t))\right\}$$

and the new estimate is $\hat{\alpha}_1 = \hat{\alpha}(t_1)$. Then repeat till it converges.

Within each iteration and at $t = t_0$, we have $\widehat{\text{bias}(t_0)} = \widehat{\mathbb{E}[\hat{\alpha}(t_0)]} - \hat{\alpha}_0 = \hat{\alpha}(t_0) - \hat{\alpha}_0 = \hat{\alpha}_0 - \hat{\alpha}_0 = 0$, and thus $\widehat{\text{MSE}(\hat{\alpha}(t_0))} = \widehat{\text{Var}(\hat{\alpha}(t_0))}$. Suppose that the curve of the empirical variance $\widehat{\text{Var}(\hat{\alpha}(t))}$ is strictly increasing, i.e. $\widehat{\text{Var}(\hat{\alpha}(t))} > \widehat{\text{Var}(\hat{\alpha}(s))}$ if $t > s$. Then for $t \in (t_0, \infty)$,

$$\widehat{\text{MSE}(\hat{\alpha}(t))} \geq \widehat{\text{Var}(\hat{\alpha}(t))} > \widehat{\text{Var}(\hat{\alpha}(t_0))} = \widehat{\text{MSE}(\hat{\alpha}(t_0))}.$$

Hence the minimum of $\widehat{\text{MSE}(\hat{\alpha}(t))}$ can only occur at a point smaller than t_0, i.e.

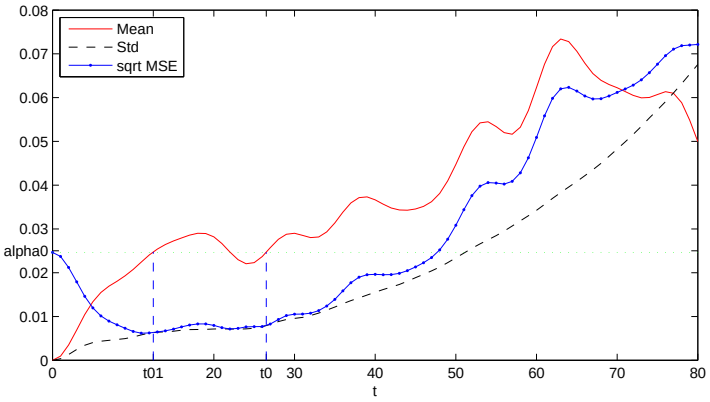
$$t_1 \leq t_0.$$

Therefore, if $\widehat{\mathrm{Var}(\hat{\alpha}(t))}$ is strictly increasing, then $t_1 \leq t_0$ in every iteration.

Actually if we let $t_{01} := \inf\{t \geq 0 : \hat{\alpha}(t) = \hat{\alpha}_0\}$, then $t_1 \leq t_{01}$ for the same reason as above. It is obvious that $t_{01} \leq t_0$. Therefore, $t_1 \leq t_{01} \leq t_0$.

Figure 2.26 shows an example of the empirical mean, standard deviation and square root of MSE of $\hat{\alpha}(t)$ in one iteration.

Figure 2.26: Empirical mean, std and MSE of $\hat{\alpha}(t)$ [Merton]



For the parameter settings suitable to financial applications, it happens frequently that the plot of $\widehat{\mathrm{Var}(\hat{\alpha}(t))}$ shows an increasing curve, and $t_1 < t_0$ in every iteration; and the estimate of α can even possibly decreases to zero through the iterations. Therefore, the method of minimizing MSE does not work properly.

2.3.5 Summary

So far we have proposed several different methods of selecting t. Since the method in Section 2.3.4 fails, in the following we only discuss the first three methods.

- Method 1 is based on analyzing the variance formula of $\hat{\alpha}(t)$. The idea is straightforward and the method is easy to implement. It depends only on

$\sigma \equiv \sigma_0\sqrt{\Delta}$. However, it does not incorporate other information about the model, and it does not work well for high frequency data.

- Method 2 is based on the bootstrapped estimate of the variance of $\hat{\alpha}(t)$. By setting a threshold value for this estimated variance, we obtain a critical value of t. This method incorporates more information about the model than Method 1 does, and there is no constraint on the underlying distribution of jump size. However, this method depends strongly on the selection of the threshold value, and thus it may not be completely objective or "automatic". However, this method can be used as a check-up step after we use another method to select a value of t.

- Method 3 is based on the shape of the curve $\hat{\alpha}(t), t \geq 0$. Under the assumption that $R(t), t \geq 0$ is strictly decreasing, we select t by using the first stationary point of $\hat{\alpha}(t), t \geq 0$. This method is objective. The simulation results also show that it is either better or comparable with the results obtained by using Method 1 or Method 2. For the models where $R(t), t \geq 0$ is not strictly decreasing, we may still apply this method but possibly with larger errors.

Based on our numerical studies, among the above three methods, we recommend Method 3.

Bias Correction [failed]

In the following, we consider the possible bias correction, after selecting t and obtaining an estimate of α. Suppose t_M is our selection of t and $\hat{\alpha} \equiv \hat{\alpha}(t_M)$ is the estimate of jump ratio. Note that we should be able to distinguish the two notations: $\hat{\alpha}$ is for the estimate, and $\hat{\alpha}(\cdot)$ is for the function. From equation (2.27):

$$\mathbb{E}(\hat{\alpha}(t)) = \alpha - \alpha R(t)$$

we might consider correcting the bias by defining a new estimate α^* as

$$\alpha^* = \hat{\alpha} - \hat{\alpha}\hat{R}(t_M), \tag{2.47}$$

where $\hat{R}(\cdot)$ is given by (2.46) and (2.45):

$$\hat{R}(t) = \Re(\hat{\varphi}_X(t)) = \Re\left(\frac{\frac{\hat{\varphi}_Y(t)}{\varphi_Z(t)} - (1-\hat{\alpha})}{\hat{\alpha}}\right).$$

Unfortunately, this method for bias correction does not work. The reason is given as follows. At the selected point t_M, we have

$$\begin{aligned}
\hat{R}(t_M) &= \Re\left(\frac{\frac{\hat{\varphi}_Y(t_M)}{\varphi_Z(t_M)} - (1-\hat{\alpha}(t_M))}{\hat{\alpha}(t_M)}\right) \\
&= \Re\left(\frac{\frac{\hat{\varphi}_Y(t_M)}{\varphi_Z(t_M)} - (1-\hat{\alpha}(t_M))}{\hat{\alpha}(t_M)}\right) \\
&= \frac{\hat{\alpha}(t_M) - \left(1 - \Re\left(\frac{\hat{\varphi}_Y(t_M)}{\varphi_Z(t_M)}\right)\right)}{\hat{\alpha}(t_M)} \\
&= 0
\end{aligned}$$

where in the last line we used $\hat{\alpha}(t_M) = 1 - \Re\left(\frac{\hat{\varphi}_Y(t_M)}{\varphi_Z(t_M)}\right)$ obtained by (2.24). Thus, the new estimate is

$$\alpha^* = \hat{\alpha} - \hat{\alpha}\hat{R}(t_M) = \hat{\alpha}.$$

That is, the new estimate is the same as the estimate before correction. Therefore, this procedure for bias correction does not work.

2.4 Application of other Estimators of the Characteristic Function

In the previous sections the characteristic function of the observable variable Y was estimated by the empirical characteristic function (e.c.f.). In this section we

consider alternative estimators of the characteristic function. The objective is to see whether the estimation of jump frequency can be improved.

Firstly, let us review some properties of the e.c.f.. The definition of the e.c.f. was given by Parzen (1962) (althoug in that paper the terminology "e.c.f." was not used yet). Later Feuerverger and Mureika (1977) initiated a systematic study of the e.c.f.. The convergence properties of the e.c.f. were investigated by Csörgő (1981a, 1981b, 1985) and Csörgő and Totik (1983). Properties in the context of statistical inference were discussed by Feuerverger and McDunnough (1981a, 1981b, 1984).

By the strong law of large numbers, the e.c.f. $\hat{\varphi}_n(t)$ converges to $\varphi(t)$ almost surely for any fixed t, i.e. it is a strongly consistent estimator for any fixed t. We also have $\mathbb{E}(\hat{\varphi}_n(t)) = \varphi(t)$, and $\mathrm{Var}(\hat{\varphi}_n(t)) = (1 - |\varphi(t)|^2)/n$ with a proof given in Appendix D.

The almost sure convergence of the e.c.f. is uniform on any fixed bounded interval (see Feuerverger and Mureika (1977)); that is, for fixed $T < \infty$, we have

$$\mathbb{P}\left\{\lim_{n\to\infty} \sup_{|t|\le T} |\hat{\varphi}_n(t) - \varphi(t)| = 0\right\} = 1.$$

However, the convergence of the e.c.f. $\hat{\varphi}_n(t)$ to $\varphi(t)$ is not generally uniform on the whole real line. The reason was given in Feuerverger and Mureika (1977). Since the e.c.f. $\hat{\varphi}_n(t)$ is a trigonometric polynomial, it is almost periodic and hence must approach its supremum value $\hat{\varphi}_n(0) = 1$ arbitrarily often as $|t| \to \infty$. On the other hand, we have $\varphi(t) \to 0$ as $|t| \to \infty$ when the distribution $F(x)$ is absolutely continuous. Hence, in general, we do not have $\sup_t |\hat{\varphi}_n(t) - \varphi(t)| \to 0$ as $n \to \infty$; in other words, the convergence of the e.c.f. $\hat{\varphi}_n(t)$ to $\varphi(t)$ is not generally uniform. However, there is one exception that, when $F(x)$ is purely discrete, the strong consistency of $\hat{\varphi}_n(t)$ is uniform over $\mathbb{R}$. Csörgő and Totik (1983) further showed that $\hat{\varphi}_n(t)$ converges to $\varphi(t)$ uniformly over $\mathbb{R}$ if and only if $F(x)$ is purely discrete.

In Feuerverger and Mureika (1977), it was also showed that for any distribution function whose singular part has characteristic function vanishing at infinity, it

holds that

$$\mathbb{P}\left\{\lim_{n\to\infty}\sup_{|t|\leq T_n}|\hat{\varphi}_n(t)-\varphi(t)|=0\right\}=1 \tag{2.48}$$

if $T_n \to \infty$ such that $T_n = O((n/\log n)^{1/2})$. This result was improved by Csörgő (1981a) where it was shown that (2.48) holds for any distribution function if $T_n \to \infty$ such that $T_n = o((n/\log\log n)^{1/2})$. Then this result was extended to the multivariate case by Csörgő (1981b). Later Csörgő and Totik (1983) improved the result again and showed that for any characteristic function $\varphi(\cdot)$ and any dimension, (2.48) holds if $T_n \to \infty$ such that $\lim_{n\to\infty}(\log T_n)/n = 0$, or equivalently, $T_n = \exp(n/G_n)$ for any sequence $\{G_n\}$ such that $G_n \to \infty$. Moreover, it was shown that this rate $T_n = \exp(o(n))$ is the best possible rate in the general case for the almost sure convergence in (2.48). In addition, as discussed in Csörgő (1985), if we specify the rate at which G_n goes to infinity, then we should be able to derive the rate at which $\sup_{|t|\leq\exp(n/G_n)}|\hat{\varphi}_n(t)-\varphi(t)|$ converges to zero almost surely.

In our problem of estimation of jump frequency, although the variable $V = IX$ follows a mixture of absolutely continuous and discrete distributions, the observable variable $Y = Z + V$ is absolutely continuous, which follows from Theorem C.0.4 in Appendix C and the fact that Z is absolutely continuous. Hence, the convergence of the e.c.f. $\hat{\varphi}_Y(t)$ to $\varphi_Y(t)$ is not uniform. In this section we consider replacing the e.c.f. by a uniformly consistent estimator of the characteristic function.

To obtain a uniformly consistent estimator of the characteristic function, typically there are two approaches. The first one is to **truncate** the e.c.f., i.e.

$$\varphi_n^*(t) := \begin{cases} \hat{\varphi}_n(t) & \text{for } |t| \leq T_n \\ 0 & \text{for } |t| > T_n, \end{cases} \tag{2.49}$$

where $T_n \to \infty$ and $\lim_{n\to\infty}(\log T_n)/n = 0$. The following result was presented in Lebedeva and Ushakov (2007).

Lemma 2.4.1. *$\varphi_n^*(t)$ in (2.49) is strongly consistent uniformly on the whole real line.*

The shortcoming of this estimator is that its realizations are never characteristic functions, since every characteristic function is uniformly continuous on the whole real line (see Lukacs 1970, page 15). However, this truncation method is used frequently in the context of de-convolution, which will be discussed in details in Chapter 5.

The second approach is to use the **kernel method**, which is based on the same smoothing idea as in the kernel method used for density estimation. Before we define the kernel estimator of the characteristic function, we shall recall the definition of the kernel estimator of a density function. Let $X_1, X_2, ..., X_n$ be a sequence of i.i.d. random variables. Then the *kernel estimator of the density function f of a random variable X* is defined by

$$\hat{f}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \tag{2.50}$$

where K is a *kernel function* and h_n is the *window width* (depending on the sample size n). The uniform consistency of the kernel density estimator is discussed in, for example, Nadaraya (1965) and Silverman (1978).

In the following we introduce the kernel estimator of the characteristic function. Its definition is implicit in many places, for example, Lebedeva and Ushakov (2007).

Definition 2.4.2. The *kernel estimator of the characteristic function of a random variable X* is defined as the characteristic function of the kernel estimator (as in (2.50)) of the density function of X.

The following lemma states that the the characteristic function of the kernel estimator is the multiplication of the e.c.f. and the Fourier transform of the kernel function. It can be found in, for example, Lebedeva and Ushakov (2007). We can also prove this result directly as follows.

Lemma 2.4.3. *The kernel estimator of the characteristic function of a random variable X, denoted by $\tilde{\varphi}_n(t)$, can be represented as*

$$\tilde{\varphi}_n(t) = \hat{\varphi}_n(t) \cdot K^{ft}(h_n t) \tag{2.51}$$

where $\hat{\varphi}_n(t) = \frac{1}{n}\sum_{j=1}^{n} e^{itx_j}$ is the e.c.f. of the variable X, $K^{ft}(t)$ is the characteristic function of the kernel function K in (2.50) and h_n is the window width in (2.50).

Proof. By Definition 2.4.2, $\tilde{\varphi}_n(t)$ is the characteristic function of the kernel estimator of the density function of X. Thus we have

$$\begin{aligned}
\tilde{\varphi}_n(t) &= \int_R e^{itx} \cdot \hat{f}(x)dx \\
&= \int_R e^{itx} \cdot \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) dx \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_R e^{it(h_n y + X_i)} K(y)dy \\
&= \frac{1}{n} \sum_{i=1}^{n} e^{itX_i} \cdot \int_R e^{ih_n ty} K(y)dy \\
&= \hat{\varphi}_n(t) \cdot K^{ft}(h_n t).
\end{aligned}$$

Remark 2.4.4. Note that the truncated estimator $\varphi_n^*(t)$ in (2.49) can actually be viewed as a special case of the kernel estimator $\tilde{\varphi}_n(t)$ in (2.51) for some particular choice of the kernel function, called the *sinc kernel*. More details on this kernel are presented in Section 5.1.1.

The uniform consistency of the kernel estimator of the characteristic function has been discussed in, for example, Feuerverger and Mureika (1977) and Lebedeva and Ushakov (2007). In the following we present the most recent result given by Lebedeva and Ushakov (2007). It provides the necessary and sufficient conditions under which the kernel estimator is a strongly consistent estimator of the characteristic function uniformly over the whole real line.

Theorem 2.4.5. *[Lebedeva and Ushakov (2007)] Let $K^{ft}(t)$ be the characteristic function of an arbitrary absolutely continuous distribution. Denote $\tilde{\varphi}_n(t;h) = \hat{\varphi}_n(t)K^{ft}(h_n t)$, which is the characteristic function of the kernel estimator of the density of the underlying distribution. Then*

$$\sup_t |\tilde{\varphi}_n(t;h) - \varphi(t)| \to 0 \quad \textit{almost surely, as } n \to \infty,$$

if and only if

$$h_n \to 0 \quad \textit{as } n \to \infty,$$

and

$$\frac{-\log h_n}{n} \to 0 \quad \textit{as } n \to \infty.$$

Now let us come back to our problem. The estimator $\hat{\alpha}(t)$ defined in (2.24) involves the ratio of two characteristic functions:

$$\hat{\alpha}(t) = 1 - \Re\left(\frac{\hat{\varphi}_Y(t)}{\varphi_Z(t)}\right) = 1 - \Re\left(\frac{\hat{\varphi}_Y(t)}{e^{i\mu t - \frac{1}{2}\sigma^2 t^2}}\right).$$

The denominator is in the form of the c.f. of a normal distribution. We would like to replace the current numerator, the e.c.f. of Y, by a uniformly consistent estimator of the c.f. of Y. The objective is to see whether this would improve the estimation of the jump ratio α. To do this, we could use the truncated estimator $\varphi_n^*(t)$ in (2.49) or the kernel estimator $\tilde{\varphi}_n(t)$ in (2.51). The following are some findings:

- If we replace $\hat{\varphi}_Y(t)$ by the truncated estimator $\varphi_n^*(t)$, which is truncated at T_n, then when $t > T_n$ the numerator becomes exactly zero and thus $\hat{\alpha}(t) = 1$. When $t < T_n$ there is no change to the e.c.f.. Therefore, using the truncated estimator does not seem to lead to any improvement of the original estimator which uses the e.c.f..

- Now we consider using the kernel estimator $\tilde{\varphi}_n(t)$ in (2.51). If $K^{ft}(h_n t)$ converges to zero faster than the term $e^{-\frac{1}{2}\sigma^2 t^2}$ in the denominator, then $\hat{\alpha}(t)$

would converges to one as t increases, not the true value α. If $K^{ft}(h_n t)$ converges to zero slower than $e^{-\frac{1}{2}\sigma^2 t^2}$, then $\hat{\alpha}(t)$ would still oscillates with an increasing amplitude as t increases. Thus, the selection of t would pose the same difficulties as when we use the e.c.f..

Thus, the approach of replacing the e.c.f. by a uniformly consistent estimator of the c.f. is not promising. To verify this, in the following we discuss the implementation of the kernel estimator $\tilde{\varphi}_n(t)$. For the selection of the kernel function K in the context of density estimation, different choices have been discussed in the literature, including the Epanechnikov kernel and the Gaussian kernel (i.e. standard normal density function). However, Silverman (1986) shows that the efficiency of different kernels does not change much from one to another. Now it is widely recognized that the quality of density estimates is determined primarily by the choice of the bandwidth rather than that of the kernel (see Ait-Sahalia (1996)). Therefore, we may choose the Gaussian kernel due to its desirable properties.

For the selection of the window width h_n, different methods were suggested by Silverman (1986). Without any prior information, we use the so-called "reference" method (see page 45-48 of Silverman (1986)) where the window width is chosen by

$$h_n = (4/3)^{1/5}\sigma n^{-1/5} = (4/3)^{1/5}(2\Phi^{-1}(0.75))^{-1}Rn^{-1/5} \approx 0.7852Rn^{-1/5} \qquad (2.52)$$

with R denoting the *interquartile range* of the observations. The interquartile range is defined as the difference between the upper and lower quartiles. It is easy to verify that this window width satisfies the conditions given by Theorem 2.4.5.

Now the kernel estimator of the characteristic function of Y is given by

$$\tilde{\varphi}_Y(t) := \hat{\varphi}_Y(t) \cdot K^{ft}(h_n t)$$

where $\hat{\varphi}_Y(t) = \frac{1}{n}\sum_{j=1}^{n} e^{itY_j}$ and $K^{ft}(t)$ is the characteristic function of the Gaussian kernel, i.e. $K^{ft}(t) = e^{-t^2/2}$. Therefore, a new estimator of the characteristic

function of V can be represented as

$$\begin{aligned}
\tilde{\varphi}_V(t) &= \tilde{\varphi}_Y(t)/\varphi_Z(t) \\
&= (\hat{\varphi}_Y(t)/\varphi_Z(t)) \cdot K^{ft}(h_n t) \\
&= \frac{1}{n}\sum_{j=1}^{n} e^{itY_j} \cdot e^{-i\mu t + \frac{1}{2}\sigma^2 t^2} \cdot e^{-h_n^2 t^2/2} \\
&= e^{\frac{1}{2}(\sigma^2 - h_n^2)t^2} \frac{1}{n}\sum_{j=1}^{n} e^{i(Y_j - \mu)t} \\
&= e^{\frac{1}{2}(\sigma^2 - h_n^2)t^2} \left\{ \frac{1}{n}\sum_{j=1}^{n} \cos((Y_j - \mu)t) + i \cdot \frac{1}{n}\sum_{j=1}^{n} \sin((Y_j - \mu)t) \right\},
\end{aligned} \tag{2.53}$$

and thus the new set of estimators of α based on the kernel method is given by

$$\begin{aligned}
\tilde{\alpha}(t) &:= 1 - \Re(\tilde{\varphi}_V(t)) \\
&= 1 - e^{\frac{1}{2}(\sigma^2 - h_n^2)t^2} \cdot \frac{1}{n}\sum_{j=1}^{n} \cos((Y_j - \mu)t).
\end{aligned} \tag{2.54}$$

The only difference between the new estimator $\tilde{\alpha}(t)$ and the original estimator $\hat{\alpha}(t)$ is that there appears an extra term $e^{-h_n^2 t^2/2}$, which comes from $K^{ft}(h_n t)$.

Remark 2.4.6. From equation (2.54) we may infer that this kernel estimator will not work as well as we may expect for our problem. Some explanations are given as follows. The kernel estimator of a c.f. was originally introduced in the literature to estimate the c.f. of an observable random variable, and can be understood as a "smoothed version" of the e.c.f., and thus has some advantages, such as the "uniform consistency" and "being integrable". However, in our problem we would like to estimate the c.f. of an un-observable random variable V which is in the form of a ratio of c.f.'s with the denominator being the c.f. of a normal distribution and the numerator an estimator of the c.f. of an observable variable. Thus the estimator $\tilde{\alpha}(t)$ in (2.54) does not have the same advantages over $\hat{\alpha}(t)$ in (2.24) as the kernel estimator over the e.c.f. for an observable random variable. This is just some intuitive explanation, and we will verify it by the following analysis.

The bias and variance formulas of the new estimator $\tilde{\alpha}(t)$ can be found using the same techniques as demonstrated in the proof of Proposition 2.2.1.

Proposition 2.4.7. *The expectation of the kernel estimator $\tilde{\alpha}(t)$ defined in (2.54) is given by*

$$\mathbb{E}(\tilde{\alpha}(t)) = 1 - e^{-\frac{1}{2}h_n^2 t^2}[1 - \alpha + \alpha R(t)] \tag{2.55}$$

and the variance of $\tilde{\alpha}(t)$ is given by

$$\begin{aligned} Var(\tilde{\alpha}(t)) &= e^{-h_n^2 t^2} \cdot Var(\hat{\alpha}(t)) \\ &= \frac{1}{2n} e^{-h_n^2 t^2} \left(e^{\sigma^2 t^2} + e^{-\sigma^2 t^2} \left[1 - \alpha + \alpha R(2t)\right] - 2\left[1 - \alpha + \alpha R(t)\right]^2 \right). \end{aligned} \tag{2.56}$$

We can see that, by using the kernel method, the variance decreases for all t, but the difference is small when t is small. However, the bias may increase.

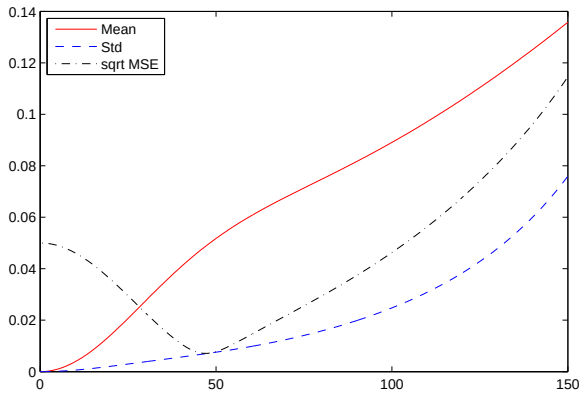
For the selection of t, we recommended Method 3, because it is better than Method 1 and more objective than Method 2. This still holds here. However, Method 3 depends on the shape of the curve $\hat{\alpha}(t)$ for t belonging to a neighborhood of zero at which the variance of $\hat{\alpha}(t)$ is still small, and thus using the kernel method may not show much advantage, if it does not create more error.

Example 2.4.8 (Merton Model)**.** We use the same parameters as in Example 2.1.2. We repeat 100,000 times. Applying Silverman's method (2.52) to choose h_n, we have the mean value of h_n is 2.9×10^{-3} and the standard deviation is 6.9×10^{-5}.

Figure 2.27 shows the mean, st.dev. and MSE of $\tilde{\alpha}(t)$ using the formulas in Proposition 2.4.7, with h_n replaced by its mean value, i.e. 0.0029. Comparing with Figure 2.3, the standard deviation decreases slightly (in the range of t we consider), but the bias increases significantly when $t > 50$ and the mean curve is no longer flat as t increases.

We apply Method 3 to select t. As before, we set $t_{\max} = 100$. Note that there are 122 samples out of 100,000 that has flag=0. This can be explained by Figure

Figure 2.27: Mean, std and MSE of $\tilde{\alpha}(t)$ [Merton]



2.27 where the mean curve has an increasing trend so that the first stationary point increases on average.

Table 2.15: Estimation results using kernel c.f. [Merton]

	Mean	Std	25% quantile	Median	75% quantile	MAD
t_{M3}	78.5406	15.1963	66.8000	76.8000	92.7000	11.9000
$\hat{\alpha}(t_{M3})$	0.0722	0.0186	0.0594	0.0703	0.0831	0.0117

Table 2.15 show the results. Comparing with Table 2.7 where the e.c.f is used, the upward bias is significantly larger, and the variance is also larger. If we plot the histograms, we can see that t_{M3} has a higher "peak" in the last bin and $\hat{\alpha}(t_{M3})$ has a heavier right tail than the ones in Figure 2.15 for which the e.c.f. was used.

To summarize, as uniformly consistent estimators of the c.f. of Y, neither the truncation method nor the kernel method with the bandwidth h_n defined by (2.52) shows any improvement for the estimation of jump frequency.

Therefore, in the following we continue using the e.c.f. as the estimator of the c.f. of Y to estimate the jump frequency.

Chapter 3

Estimation of the Jump Frequency when the Diffusion Parameters are Known: Asymptotics when Data Frequency Increases

In Chapter 2 we discussed the estimation of the jump frequency when the diffusion parameters are known and the data frequency is fixed. In this chapter, we still assume that the diffusion parameters are known, but let the data frequency change. In finance, the high frequency data are often available.

As before, let Δ be the length of the time interval (in years) between two consecutive observations. For daily data, we have $\Delta = 1/250$. When there are 10 observations per day, we have $\Delta = 1/2500$. Notice that if the jump intensity (i.e. the expected number of jumps per year) remains the same, then the jump ratio (i.e. the expected proportion of jump observations among all observations) will decrease when data frequency increases. This is a result of defining jump frequency in a

continuous-time model. For example, suppose there are 12.5 jumps on average per year, which corresponds to $\lambda = 12.5$ in the continuous model, then, by equation (2.5): $\alpha = \lambda\Delta$, the jump ratio is $\alpha = 12.5/250 = 5\%$ for the daily data, but $\alpha = 0.5\%$ for the frequency of 10 observations per day.

As discussed in the beginning of Chapter 2, there are two different types of asymptotics that we may consider for the proposed estimator. The first type is fixing the data frequency and increasing the time horizon, which has been considered in Chapter 2. In this chapter, we investigate the second type: fixing the time horizon and increasing the data frequency (i.e. letting $\Delta \to 0$). In this case, our set-up approaches the situation where continuous trajectories of the financial model are available.

This chapter is organized as follows. In Section 3.1 we present some preliminary results on the properties of the estimator when the data frequency changes. The estimation of jump intensity in the high frequency case is discussed in Section 3.2. In Section 3.3 we compare our method with the maximum likelihood method in the case when $\Delta \to 0$.

3.1 Preliminary Results

For the case when the data frequency Δ changes, we can obtain the bias and variance of our estimator by simply making the following substitutions into Proposition 2.2.1:

1. $\sigma = \sigma_0\sqrt{\Delta}$, where σ_0 is the yearly volatility of the diffusion component, due to equation (2.17).

2. $\alpha = \lambda\Delta$, where λ represents the intensity of the Poisson process, i.e. the expected number of jumps per year, due to equation (2.5).

3. $n = T/\Delta$, where T is the time horizon, i.e. the number of years of the available data.

Note that the function $R(t) \equiv \Re(\varphi_X(t)$ remains the same because the jump size distribution remains the same when the data frequency changes. As mentioned in Remark 2.2.4, the parameter $\mu = \mu_0\Delta$ does not enter the bias or variance formula. Based on the above, we obtain the following result.

Corollary 3.1.1. *For each t, the bias of $\hat{\alpha}(t)$ is*

$$bias(\hat{\alpha}(t)) = \mathbb{E}[\hat{\alpha}(t)] - \alpha = -\lambda\Delta \cdot R(t), \tag{3.1}$$

and the variance of $\hat{\alpha}(t)$ is

$$Var(\hat{\alpha}(t)) = \frac{\Delta}{2T}\left(e^{\sigma_0^2\Delta\cdot t^2} + e^{-\sigma_0^2\Delta\cdot t^2}\left[1 - \lambda\Delta + \lambda\Delta \cdot R(2t)\right] - 2\left[1 - \lambda\Delta + \lambda\Delta \cdot R(t)\right]^2\right). \tag{3.2}$$

Note that for any fixed T and t, we have $\text{bias}(\hat{\alpha}(t)) \to 0$ and $\text{Var}(\hat{\alpha}(t)) \to 0$ as $\Delta \to 0$. However, the jump ratio α also decreases. We use the following example to illustrate the situation.

In the following simulation studies, we consider the frequency of 10 observations per day, i.e. $\Delta = 1/2500$, and refer it as "high frequency".

Example 3.1.2 (Merton Model)**.** We use the same parameters as in Example 2.1.2 but replace the daily frequency by the "high frequency", i.e. $\Delta = 1/250$ by $\Delta = 1/2500$. Thus, the following parameters are the same:

$$\mu_0 = 0.1, \sigma_0 = 0.2; \lambda = 12.5; \mu_X = 0, \sigma_X = 3 \times 0.2 \times \sqrt{1/250} \approx 3.79\%.$$

However, since $\mu = \mu_0\Delta$, $\sigma = \sigma_0\sqrt{\Delta}$, $\alpha = \lambda\Delta$ and $n = T/\Delta$, the following parameters are changed to:

$$\mu = 4 \times 10^{-5}, \sigma = 0.4\%, \alpha = 0.5\%, n = 25000.$$

Figure 3.1: Mean, std and MSE of $\hat{\alpha}(t)$ [Merton, High freq]

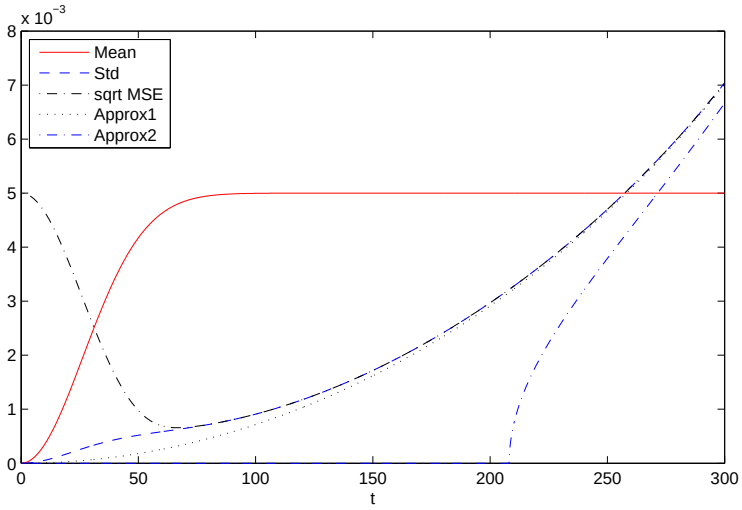


Figure 3.1 plots the expectation, standard deviation and square root of the MSE of $\hat{\alpha}(t)$*. The lowest point of the MSE corresponds to $t_{opt} = 66.7578; \text{MSE}_{min} = 4.3233 \times 10^{-7}$. At this point we also have Mean $= 0.0048$; SD $= 6.2573 \times 10^{-4}$. Comparing with Example 2.2.6, where the daily frequency was used, the optimal value of t increases. The standard deviation and the MSE decreases dramatically. However, the true value of α also decreases, so it is difficult to compare the performance of the estimators.

Instead we may consider comparing the ratio $SD(\hat{\alpha})/\alpha$ or $\sqrt{\text{MSE}(\hat{\alpha})}/\alpha$. This leads to the following approach: instead of considering the jump ratio α, we consider another parameter, the jump intensity parameter λ, which remains the same when the data frequency changes.

*The two approximations, based on (2.40) and (2.41), of the standard deviation of $\hat{\alpha}(t)$ are also shown in this figure.

3.2 Estimation of the Jump Frequency

In this section we consider the estimation of the jump intensity λ implied by the Poisson process. This parameter remains the same when the data frequency changes. Using the relation (2.5), we have $\lambda = \frac{\alpha}{\Delta}$, and thus we may define an estimator of λ by

$$\hat{\lambda}(t) := \frac{\hat{\alpha}(t)}{\Delta} \tag{3.3}$$

where $\hat{\alpha}(t)$ is defined by (2.24) and the selection of t is the same as discussed in Chapter 2. Then all the previous results about $\hat{\alpha}(t)$ can be applied to $\hat{\lambda}(t)$.

Based on Corollary 3.1.1, the following result provides the bias and variance of $\hat{\lambda}(t)$ and their limits as $\Delta \to 0$. Note that the limiting case of $\Delta \to 0$ is equivalent to the case when the continuous trajectories of the financial model are available and all jumps become visible, since the increment of the diffusion component goes to zero.

Corollary 3.2.1. *(i) The bias of $\hat{\lambda}(t)$ defined by (3.3) is*

$$bias(\hat{\lambda}(t)) = -\lambda \cdot R(t) \tag{3.4}$$

and the variance of $\hat{\lambda}(t)$ is

$$Var(\hat{\lambda}(t)) = \frac{1}{2T\Delta}\left(e^{\sigma_0^2\Delta\cdot t^2} + e^{-\sigma_0^2\Delta\cdot t^2}\left[1-\lambda\Delta+\lambda\Delta\cdot R(2t)\right] - 2\left[1-\lambda\Delta+\lambda\Delta\cdot R(t)\right]^2\right). \tag{3.5}$$

(ii) As $\Delta \to 0$, we have

$$\lim_{\Delta\to 0} bias(\hat{\lambda}(t)) = -\lambda \cdot R(t) \tag{3.6}$$

and

$$\lim_{\Delta\to 0} Var(\hat{\lambda}(t)) = \frac{\lambda(3+R(2t)-4R(t))}{2T}. \tag{3.7}$$

From the above result we see that, as $\Delta \to 0$, the bias of $\hat{\lambda}(t)$ for a fixed value of t is unchanged when the data frequency increases. However, as we explain later, when the data frequency increases we may choose a larger value of t to decrease the bias.

After letting $\Delta \to 0$, we may further let $t \to \infty$ to obtain the following result. Obviously, the bias will be removed.

Corollary 3.2.2. *The bias of $\hat{\lambda}(t)$ defined by (3.3) satisfies*

$$\lim_{t\to\infty}\left(\lim_{\Delta\to 0} bias(\hat{\lambda}(t))\right) = 0, \tag{3.8}$$

$$\lim_{t\to\infty}\left(\lim_{\Delta\to 0} Var(\hat{\lambda}(t))\right) = \frac{3\lambda}{2T}, \tag{3.9}$$

and thus

$$\lim_{t\to\infty}\left(\lim_{\Delta\to 0} MSE(\hat{\lambda}(t))\right) = \frac{3\lambda}{2T}. \tag{3.10}$$

Remark 3.2.3. One observation from (3.7) and (3.9) is that the variance of $\hat{\lambda}(t)$ as $\Delta \to 0$ (for fixed t or letting $t \to \infty$) will not converge to zero, but to some constant. This can be explained by the fact that even in the case when a continuous trajectory is available and we can simply count the total number of jumps (i.e. we have the trajectory of the Poisson process), the total number of jumps divided by the length of the time horizon is still only an estimate of the jump intensity, but not exactly equal to it.

Remark 3.2.4. After an examination of the formulas (3.6) and (3.7), we see that the results in Corollary 3.2.2 still hold if we replace "the limit of t going to infinity" by "for a large t", or more exactly, "for $t > t_0$" where t_0 is smallest point starting from which $R(t)$ stabilize at zero, i.e.

$$t_0 := \inf\{t > 0 : R(s) \approx 0, s \geq t\}. \tag{3.11}$$

This is because to obtain the results in Corollary 3.2.2 by using (3.6) and (3.7), we only need $R(t)$ to get close to zero.

Remark 3.2.5. From formula (3.10) and Remark 3.2.4, we see that, as $\Delta \to 0$, the curve of $\mathrm{MSE}(\hat{\lambda}(t))$ tends to stabilize at level $3\lambda/(2T)$ for $t > t_0$, where t_0 is defined in (3.11). This also explains why in Table 3.1 (below) the "optimal" value of t increases more slowly when Δ becomes smaller. Moreover, as $\Delta \to 0$, the solution to the minimization of MSE may not be robust, since the curve of $\mathrm{MSE}(\hat{\lambda}(t))$ tends to be flat for $t > t_0$.

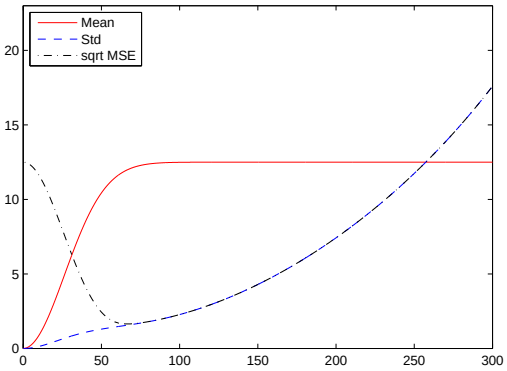
Before proceeding to the simulation studies, we present an inequality which can be easily proved by equation (3.7). It is the same as the one in Lemma C.0.6 in Appendix C.

Corollary 3.2.6. *Let $\varphi(t)$ be the characteristic function of a random variable, and $R(t) \equiv \Re(\varphi(t))$, then*

$$3 + R(2t) - 4R(t) \geq 0, \quad t \in \mathbb{R}.$$

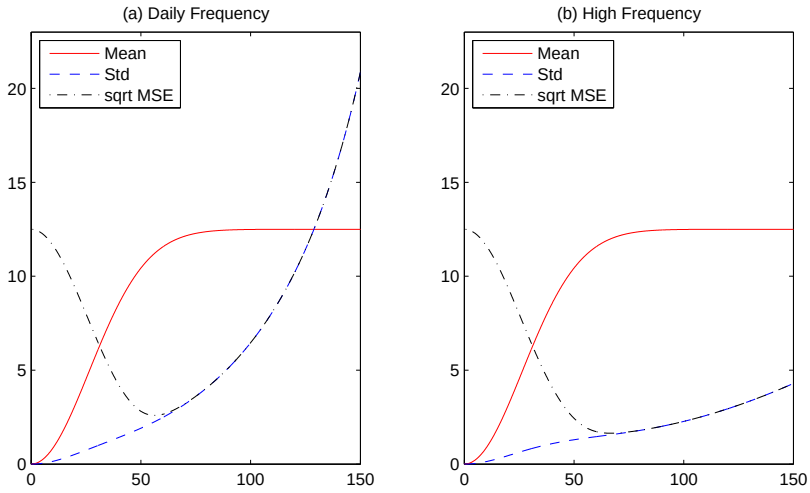
Example 3.2.7 (Merton Model)**.** We use the same parameters as in Example 3.1.2 for high frequency case, i.e. $\Delta = 1/2500$. Figure 3.2 shows the expectation, standard deviation and MSE of $\hat{\lambda}(t)$. We can see that it has the same shape as in Figure 3.1 for $\hat{\alpha}(t)$.

Figure 3.2: Mean, std and MSE of $\hat{\lambda}(t)$ [Merton, High freq]



For comparison, we put together this graph and the one for the daily case into Figure 3.3, where both focus on the parameter λ. We can see that the variance of $\hat{\lambda}(t)$ decreases as data frequency increases. However, this is less obvious for $t \in [0, t_0]$ where t_0 is defined in (3.11). In this figure, we have $t_0 \approx 75$. The explanation has been provided in Remark 3.2.5.

Figure 3.3: Comparison between daily frequency and high frequency [Merton]



Let us look at the "optimal" value of t which minimizes the MSE of $\hat{\lambda}(t)$. When $\Delta = 1/250$, the optimal value is $t_{opt} = 56.2789$, and at this point we have $\text{MSE}_{min} = 6.7295$, Mean $= 11.2224$ and SD $= 2.2577$. When $\Delta = 1/2500$, the optimal value is $t_{opt} = 66.7578$, and at this point $\text{MSE}_{min} = 2.7020$, Mean $= 11.9951$ and SD $= 1.5643$. Comparing with the t_{opt} found in Examples 2.2.6 and 3.1.2 for $\hat{\alpha}(t)$, we see that the optimal value of t for a fixed Δ stays the same, no matter which parameter we consider: the ratio α or the intensity λ. On the other hand, when data frequency increases, the optimal value of t increases, and the bias and variance of $\hat{\lambda}(t)$ at this optimal point both decreases, i.e. the performance of the estimator improves.

Table 3.1: Relation between the optimal value of t and data frequency

Δ	t_{opt}	MSE_{min}	Mean	St. Dev.
1/250 = 4e-3	56.2789	6.7295	11.2224	2.2577
1/2500 = 4e-4	66.7578	2.7020	11.9951	1.5643
4e-5	74.9933	1.9851	12.2822	1.3920
4e-6	80.0683	1.8793	12.3764	1.3653
4e-7	81.6397	1.8665	12.3971	1.3623
4e-8	81.8593	1.8651	12.3997	1.3620
4e-9	81.8811	1.8650	12.3999	1.3620

Note: The time horizon is 10-year. The true value of the intensity is λ=12.5. The rounding error might appear when Δ is very small.

Now we continue using the Merton model to investigate how the optimal value of t depends on the data frequency Δ. From Table 3.1, we can see that, as $\Delta \to 0$, the optimal value t_{opt} increases, and the bias and variance of $\hat{\lambda}(t_{opt})$ both decrease. Therefore, we would have better estimates of λ when the high frequency data is available. However, the rate of t_{opt} increasing becomes slower as Δ decreases. See Remark 3.2.5 for explanations.

In the following we consider the estimation of jump intensity for Merton model.

Example 3.2.8 (Merton Model)**.** We use the same parameters as in Example 3.1.2 for the high frequency case, i.e. $\Delta = 1/2500$. For the selection of t, Method 3 is applied.

Figure 3.4 shows the histogram of $\hat{\lambda}(t_{M3})$ and t_{M3}. We see that $\hat{\lambda}(t_{M3})$ seems closer to be normally distributed, comparing with the daily case which shows an obvious heavy right tail. Table 3.2 shows the statistics of the results. Comparing with the daily case (see Table 2.8 for results on λ, the scenario with $T = 10$ and $\lambda = 12.5$), we can see that

- the mean value of t_{M3} increases and its variance decreases;
- the bias and variance of $\hat{\lambda}(t_{M3})$ both decrease.

That is, the estimation results are better than in the daily case.

Figure 3.4: Histograms of $\hat{\alpha}(t_{M3})$ and t_{M3} when $\sigma_X \approx 3.79\%$ [Merton, High freq]

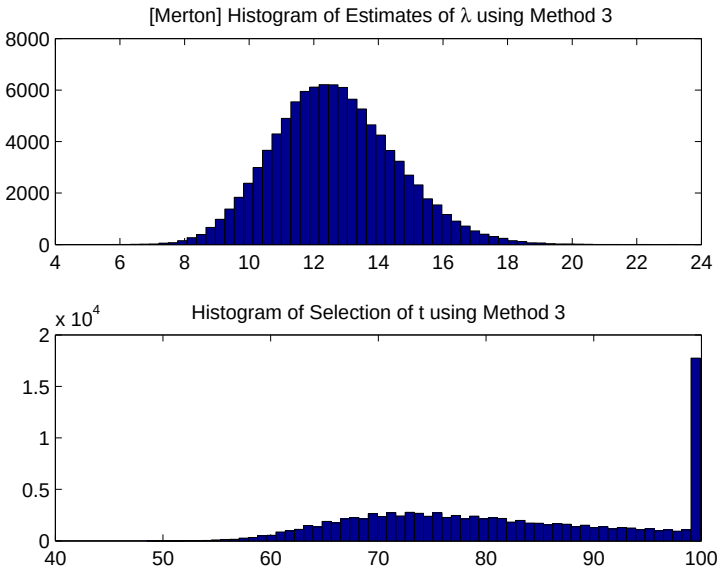


Table 3.2: Estimation results for $\hat{\lambda}(t_{M3})$ when $\sigma_X \approx 3.79\%$ [Merton, High freq]

	Mean	Std	25% quantile	Median	75% quantile	MAD
t_{M3}	81.6465	12.4997	71.4000	80.1000	93.2000	10.2000
$\hat{\lambda}(t_{M3})$	12.6562	1.8936	11.3411	12.5570	13.8719	1.2600

In the following, we increase σ_X, while keeping all the other parameters the same, and see how the high frequency influences the results for different σ_X.

When $\sigma_X = 0.1$, Figure 3.5 and Table 3.3 show the results for the high frequency case. For the daily case, the histogram was given in Figure 2.16, and the statistics

can be found from Table 2.8, but we repeat the statistics in Table 3.4 with more details. We can see that the results of $\hat{\lambda}(t_{M3})$ become better in both bias and variance when the data frequency increases[†].

Figure 3.5: Histograms of $\hat{\lambda}(t_{M3})$ and t_{M3} when $\sigma_X = 0.1$ [Merton, High freq]

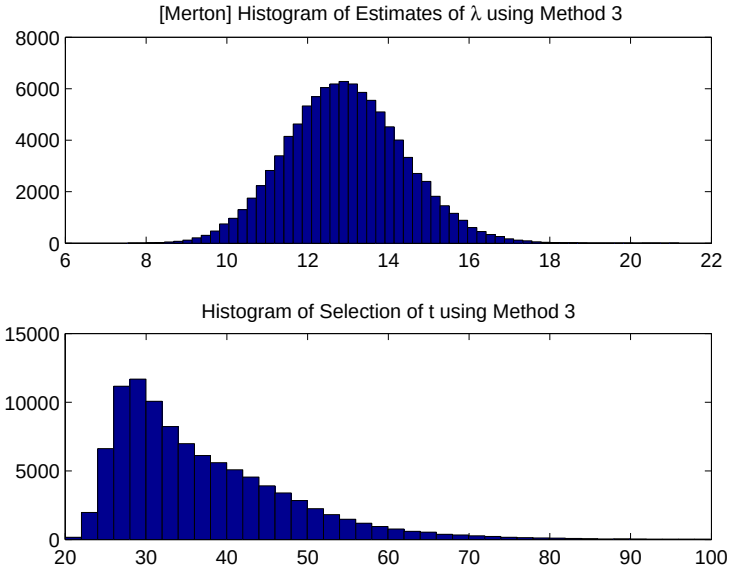


Table 3.3: Estimation results for $\hat{\lambda}(t_{M3})$ when $\sigma_X = 0.1$ [Merton, High freq]

	Mean	Std	25% quantile	Median	75% quantile	MAD
t_{M3}	36.9208	10.8609	28.8000	34.0000	42.5000	6.2000
$\hat{\lambda}(t_{M3})$	12.9381	1.4555	11.9453	12.9039	13.8929	0.9729

When $\sigma_X = 0.6$, Figure 3.6 and Table 3.5 show the results for the high frequency case. For the daily case, the histogram was given in Figure 2.17, and the statistics

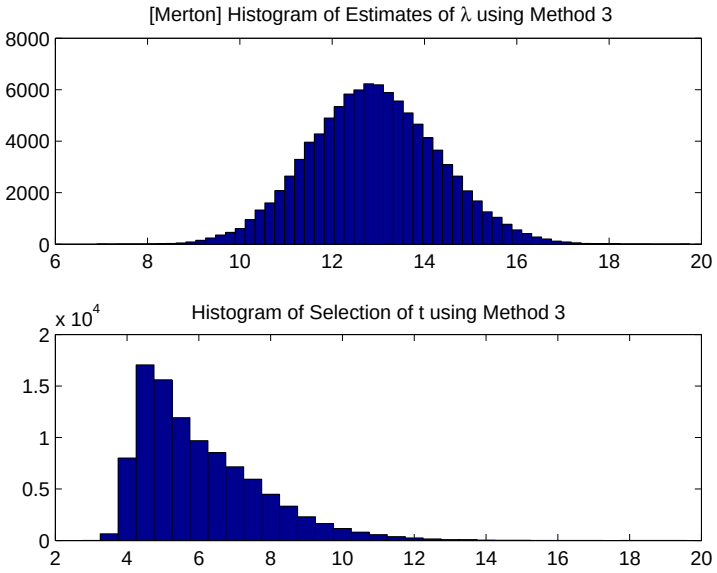
[†]It shows that in this case the mean value of t_{M3} decreases when data frequency increases, but this is only because t_{M3} has a heavier tails in the daily case than in the high frequency case. We can also see that the median of t_{M3} increases from 33.6 to 34.0 when data frequency increases.

Table 3.4: Estimation results for $\hat{\lambda}(t_{M3})$ when $\sigma_X = 0.1$ [Merton, Daily freq]

	Mean	Std	25% quantile	Median	75% quantile	MAD
t_{M3}	37.6227	13.1890	28.3000	33.6000	42.8000	6.3000
$\hat{\lambda}(t_{M3})$	13.2106	2.0834	11.8603	12.9844	14.2451	1.1841

can be found from Table 2.8, but we repeat the statistics in Table 3.6 with better accuracy. We can see that the results of $\hat{\lambda}(t_{M3})$ do improve when the data frequency increases, but the improvement is very limited. See Remark 3.2.5 for explanations, noting that the daily frequency is already "relatively high" for the case $\sigma_X = 60\%$.

Figure 3.6: Histograms of $\hat{\lambda}(t_{M3})$ and t_{M3} when $\sigma_X = 0.6$ [Merton, High freq]



In conclusion, the estimation results become better when the data frequency increases. However, after Δ drops below some level, the improvement will be quite limited.

Table 3.5: Estimation results for $\hat{\lambda}(t_{M3})$ when $\sigma_X = 0.6$ [Merton, High freq]

	Mean	Std	25% quantile	Median	75% quantile	MAD
t_{M3}	6.0363	1.6968	4.7000	5.6000	7.0000	1.0000
$\hat{\lambda}(t_{M3})$	12.9051	1.3947	11.9544	12.8826	13.8366	0.9408

Table 3.6: Estimation results for $\hat{\lambda}(t_{M3})$ when $\sigma_X = 0.6$ [Merton, Daily freq]

	Mean	Std	25% quantile	Median	75% quantile	MAD
t_{M3}	6.0245	1.6928	4.7000	5.6000	7.0000	1.0000
$\hat{\lambda}(t_{M3})$	12.9042	1.3743	11.9657	12.8800	13.8143	0.9244

3.3 Comparison with the MLE

As we mentioned earlier, when $\Delta \to 0$ our set-up approaches the case when the continuous trajectory is available. In that case, all the jumps can be identified without any error, no matter how small the jump size is, i.e. the trajectory of the Poisson process is available. Thus, in order to estimate the intensity λ, we only need to work with the Poisson process. Under this situation, we may consider the maximum likelihood estimator (MLE) to estimate the jump intensity λ of the Poisson process.

In this section we compare the proposed estimator with the MLE. Note that this may not be a "fair" comparison between methods, because our method also works for the low frequency cases, but the MLE does not.

In the following we derive the MLE of the jump frequency λ for a Poisson process. Let T be the time horizon. To find the likelihood function, we partition the time interval $[0, T]$ into m sub-intervals of equal lengths: $[0, \Delta)$, $[\Delta, 2\Delta)$, ..., $[(i-1)\Delta, i\Delta)$, ..., $[(m-1)\Delta, m\Delta)$, with $m\Delta = T$. When Δ is small, there is at

most one jump in one time interval.[‡] Denote

$$B_i := \begin{cases} 1 & \text{if there is one jump during } [(i-1)\Delta, i\Delta); \\ 0 & \text{if there is no jump during } [(i-1)\Delta, i\Delta), \end{cases}$$

for $i = 1, 2, ..., m$. Using the basic properties of Poisson processes and the condition that Δ is small we have

$$B_i \sim \text{Bernoulli}(p)$$

where $p = \lambda \cdot \Delta$. Then the likelihood function can be written as

$$L(B_1, ..., B_m; p) = \prod_{i=1}^{m} p^{B_i}(1-p)^{1-B_i}$$

and the log-likelihood function is given by

$$l(B_1, ..., B_m; p) = \ln p \cdot \sum_{i=1}^{m} B_i + \ln(1-p) \cdot \left(m - \sum_{i=1}^{m} B_i \right).$$

Differentiating with respect to p and equating the score function to zero, we have

$$\frac{\partial l}{\partial p} = \frac{1}{p} \cdot \sum_{i=1}^{m} B_i + \frac{-1}{1-p} \cdot \left(m - \sum_{i=1}^{m} B_i \right) = 0.$$

By solving it we obtain the MLE of p as

$$\tilde{p} = \frac{\sum_{i=1}^{m} B_i}{m}.$$

Since $p = \lambda \cdot \Delta$, we obtain the MLE of λ by

$$\tilde{\lambda} = \frac{\tilde{p}}{\Delta} = \frac{\sum_{i=1}^{m} B_i}{m \cdot \Delta} = \frac{\sum_{i=1}^{m} B_i}{T}.$$

This result is intuitive. The expectation of the MLE is

$$\mathbb{E}(\tilde{\lambda}) = \frac{m\mathbb{E}(B_i)}{T} = \frac{m\lambda\Delta}{T} = \frac{\lambda T}{T} = \lambda, \tag{3.12}$$

[‡]This can be made more rigorous by adding the error term and later verifying that the impact of the error term on the estimation results is negligible in the limit.

and the variance is

$$\mathrm{Var}(\tilde{\lambda}) = \frac{m\mathrm{Var}(B_i)}{T^2} = \frac{m\lambda\Delta(1-\lambda\Delta)}{T^2} = \frac{\lambda(1-\lambda\Delta)}{T}. \tag{3.13}$$

When the partition becomes finer, i.e. $\Delta \to 0$, we have

$$\lim_{\Delta\to 0} \mathrm{Var}(\tilde{\lambda}) = \frac{\lambda}{T}. \tag{3.14}$$

Note that $\sum_{i=1}^{m} B_i$ is the total number of intervals where there is one jump occurring. When $\Delta \to 0$, $\sum_{i=1}^{m} B_i$ describes the total number of jumps in the time interval $[0, T)$. Thus $\sum_{i=1}^{m} B_i$ follows a Poisson distribution with parameter λT. Using the mean and variance of this Poisson distribution, the results (3.12) and (3.14) can also be easily recovered.

Comparing (3.9) with (3.14), we get the following result.

Proposition 3.3.1. *The asymptotic relative efficiency (as $\Delta \to 0$) of our estimator $\hat{\lambda}(t)$ in (3.3) with a large t over the MLE is about 67%.*

However, our method has the advantage of not requiring $\Delta \to 0$, i.e. our method works for low frequency data; while the MLE method assumes that $\Delta = 0$. To provide a more specific example, let us consider the Merton model with $T = 10$ and $\lambda = 12.5$:

- MLE: In this case, the variance of MLE is $\lambda/T = 12.5/10 = 1.25$.
- Asymptotic result based on our method: From Corollary 3.2.2, we have $\lim_{t\to\infty}\left(\lim_{\Delta\to 0}\mathrm{Var}(\hat{\lambda}(t))\right) = \frac{3\lambda}{2T}$ and the corresponding st.dev is $\sqrt{\frac{3\lambda}{2T}} = \sqrt{1.875} = 1.3693$.
- Low frequency result based on our method: Consider the daily frequency, i.e. $\Delta = 1/250$. From Table 2.8, when $\sigma_X = 60\%$, we have $\mathrm{Mean}(\hat{\lambda}(t_{M3})) = 12.9042$ which is close to true λ, and $\mathrm{Std}\hat{\lambda}(t_{M3})) = 1.3743$ which is close to

the asymptotic result 1.3693. When $\sigma_X = 10\%$, the result still looks good. When $\sigma_X \approx 3.79\%$, corresponding to the triple of daily volatility, the result is still acceptable.

In Chapter 6, we compare our method with some other existing methods.

Chapter 4

Estimation of the Jump Frequency when the Diffusion Parameters are Unknown

In this chapter, we remove the assumption that the diffusion parameters μ_0 and σ_0 are known. Then the parameters $\mu = \mu_0\Delta$ and $\sigma = \sigma_0\sqrt{\Delta}$ in the normal distribution of Z are also unknown. Thus, to estimate the jump frequency, we need first estimate the parameters μ and σ.

The cdf of the observable variable Y is given by equation (2.14):

$$F_Y = (1-\alpha)\Phi_Z + \alpha F_G.$$

This is exactly of the same form as the *α-contaminated normal distribution* discussed in the context of *robust statistics*. For the contamination ratio α, a typical value for the financial daily data is around 5% (i.e. the jump intensity $\lambda = 12.5$); see Kou (2002). When higher frequency data is available, the value of α becomes smaller. So we can use the robust statistics to estimate the diffusion parameters.

After obtaining estimates of the diffusion parameters, we can apply the methods discussed in the previous chapters to estimate the jump frequency.

4.1 Robust Estimation of the Diffusion Parameters

We first briefly review the robust methods for the estimation of location parameter in Section 4.1.1, and the estimation of scale parameter in Section 4.1.2. A detailed review of the robustness criterion and different types of robust estimators is provided in Appendix A. In Section 4.1.3 we discuss the simultaneous robust estimation of both location and scale parameters.

4.1.1 Location Estimator

The M-, L- and R- estimators are the three popular types of robust estimators. Among them, the M- estimator is the one used mostly widely.

M-estimators

The *M-estimator T_n of a location parameter* is defined as a solution to the following minimization problem

$$\sum_{i=1}^{n} \rho(X_i - \mu) := min,$$

and if $\rho(\cdot)$ is differentiable with an absolutely continuous derivative $\psi(\cdot)$, then T_n solves the equation

$$\sum_{i=1}^{n} \psi(X_i - \mu) = 0. \tag{4.1}$$

Let us define a functional $T(F)$ on the space of distributions as the solution to $\int \psi(x - T(F))dF(x) = 0$, if it exists and is unique. Then the influence function of $T(F)$ is given by

$$IF(x; T, F) = \frac{\psi(x - T(F))}{\int \psi'(y)dF(y)}.$$

If $T(F)$ is uniquely defined, then T_n is strongly consistent at F, i.e. $T_n \to T(F)$ almost surely as $n \to \infty$.

In our problem, to estimate the drift parameter in the diffusion component of the jump-diffusion model, we look for an M-estimator of the location parameter of a distribution not very far from the normal distribution, but possibly containing an ϵ (α in our problem) proportion of non-normal data. More precisely, this distribution should belong to the family of ϵ-*contaminated normal distributions*:

$$\mathcal{F}_\epsilon = \{F : F = (1 - \epsilon)\Phi + \epsilon H\}.$$

If H is a symmetric or close to symmetric distribution, then we may use the *Huber estimator* proposed by Huber (1964). When H is a distribution close to be symmetric, we may also consider the Huber estimator.

Example 1: Huber Estimator

The ρ- function of the *Huber estimator* is given by

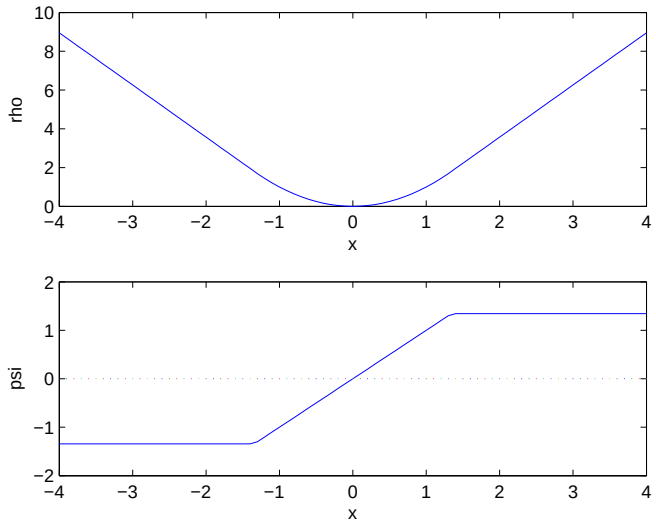
$$\rho(x) = \begin{cases} x^2 & \text{if } |x| \leq k \\ 2k|x| - k^2 & \text{if } |x| > k \end{cases}$$

with derivative $2\psi(x)$, where the ψ-function is given by

$$\psi(x) = \begin{cases} x & |x| \leq k, \\ \text{sign}(x) \cdot k & |x| > k \end{cases} \tag{4.2}$$

with $k > 0$ a fixed constant. The ρ- and ψ- functions are plotted in Figure 4.1. The Huber estimator is very popular in the context of robust estimation. It has

Figure 4.1: Huber ρ- and ψ-functions



a bounded influence function proportional to ψ_H, the breakdown point $\epsilon = \frac{1}{2}$, the global sensitivity $\gamma = \frac{k}{2F(k)-1}$, and the tail-behavior measure $\lim_{a\to\infty} B(a, T_n, F) = \frac{1}{2}$ both for distributions with exponential and heavy tails. Thus, it is a robust estimator of the center of symmetry, with a relatively low sensitivity to extreme observations. As Huber proved in 1964, this estimator is minimaximally robust for the contaminated normal distribution.

For the Huber estimator, typically we choose $k = 1.345$, which corresponds to 95% efficiency at the normal distribution. The following proposition helps to understand the Huber estimator in an intuitive way.

Proposition 4.1.1. *The Huber estimator is equivalent to a "winsorized mean".*

Proof. We have

$$\psi(x-\theta) = \begin{cases} -k & \text{if } x-\theta < -k, \\ x-\theta & \text{if } |x-\theta| \le k, \\ k & \text{if } x-\theta > k. \end{cases}$$

We need to solve the equation $\sum_{i=1}^{n} \psi(x_i - \theta) = 0$. Since the function Ψ is monotone, the solution is unique. We have

$$\begin{aligned}\sum_{i=1}^{n} \psi(x_i - \theta) &= \sum_{|x_i-\theta|\le k} (x_i - \theta) + \sum_{x_i > \theta + k} (k) + \sum_{x_i < \theta - k} (-k) \\ &= \sum_{|x_i-\theta|\le k} (x_i - \theta) + \sum_{x_i > \theta + k} (\theta + k - \theta) + \sum_{x_i < \theta - k} (\theta - k - \theta).\end{aligned}$$

Assume $\hat{\theta}$ is the solution and let

$$Y_i := \begin{cases} \hat{\theta} - k & \text{if } x_i < \hat{\theta} - k, \\ x_i & \text{if } |x_i - \hat{\theta}| \le k, \\ \hat{\theta} + k & \text{if } x_i > \hat{\theta} + k. \end{cases}$$

Then $\sum_{i=1}^{n} (Y_i - \hat{\theta}) = 0$, i.e.

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Thus we proved that Huber estimator is equivalent to the "winsorized mean". However, the solution $\hat{\theta}$ is involved in the construction of Y_i's, so an iterative procedure is necessary to evaluate the Huber estimator. ⊠

When the ψ- function is monotone, the estimator is called a *monotone M-estimator*. The Huber estimator is one of them. There is another class of M-estimators, called the **redescending M-estimators**, whose ψ- functions vanish outside some interval. It has been found that redescending estimators sometimes outperform the Huber estimator. For example, the redescending M-estimators are more efficient than the Huber estimator for some symmetric and heavy-tailed distributions. This is because redescending estimators completely reject gross outliers, while the Huber estimator treats them in the same way as moderate outliers. Examples of redescending ψ-functions include the *Hample function*, the *Andrews sinus*

function, the *Tukey bisquare or biweight function* and the *Cauchy function*. In the following we consider the Tukey bisquare estimator as an example.

For our problem, a jump increment is a convolution of the diffusion increments and another jump size distribution. Therefore, the variance of the jump increments is much larger than the variance of the diffusion increments. Hence, redescending M-estimators might perform better than monotone M-estimators.

Example 2: Tukey Bisquare Estimator

The ρ- and ψ- functions of the *Tukey bisquare estimator* are given by

$$\rho(x) = \begin{cases} 1 - \left[1 - \left(\frac{x}{k}\right)^2\right]^3 & \text{if } |x| \leq k \\ 1 & \text{if } |x| > k \end{cases} \tag{4.3}$$

with the derivative $\rho'(x) = 6\psi(x)/k^2$, where

$$\psi(x) = \begin{cases} x\left[1 - \left(\frac{x}{k}\right)^2\right]^2 & \text{if } |x| \leq k \\ 0 & \text{if } |x| > k. \end{cases} \tag{4.4}$$

These functions are plotted in Figure 4.2. Note that ψ is everywhere differentiable and vanishes outside $[-k, k]$.

For the bisquare estimator, the values of k and the corresponding efficiencies (ratio of asymptotic variance of the MLE and asymptotic variance of the proposed estimator) at the normal distribution are given in Table 4.1. In the following simulations we will use k=4.685, which corresponds to 95% efficiency.

4.1.2 Scale Estimator

For our problem, the estimation of the scale parameter is very important, since the estimator is very sensitive to this parameter. More explanation will be provided

Figure 4.2: Bisquare ρ- and ψ- Functions

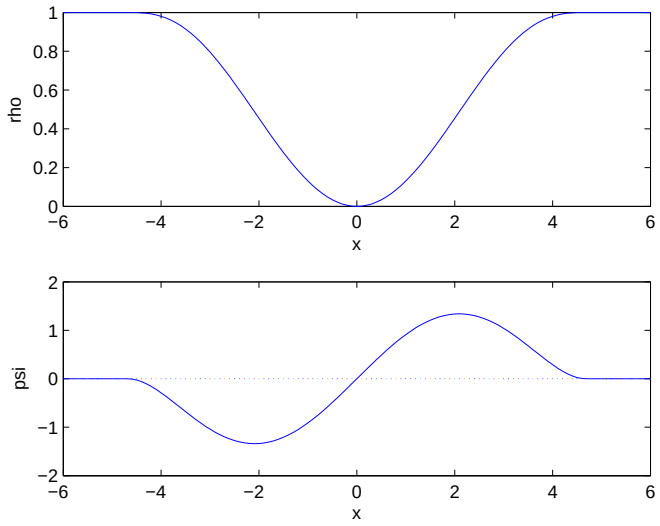


Table 4.1: Choices of k in Tukey bisquare estimator

Efficiency	0.80	0.85	0.90	0.95
k	3.14	3.44	3.88	4.68

at the end of this section. In the following we first introduce some simple but widely used robust estimators of the scale, and then discuss some M-estimators of the scale parameter. Simulation results are also shown.

1. Median absolute deviation (MAD)

MAD is defined as the median of the absolute deviations from the data's median:

$$\text{MAD} := \text{median}_i \left(\, |X_i - \text{median}_j(X_j)| \, \right).$$

This is the most widely used robust estimator of the scale parameter. In order to use MAD as a consistent estimator of the standard deviation of normal distributions,

we have to take

$$\hat{\sigma} = \text{MAD}/(\Phi^{-1}(3/4)) = 1.4826 * \text{MAD}.$$

2. Interquartile range (IQR)

The interquartile range is defined as the difference between the 75th percentile and the 25th percentile of a sample, i.e.

$$\text{IQR} := Q_3 - Q_1,$$

where Q_3 and Q_1 are the 75% and 25% percentiles. For a symmetric distribution (so the median equals the average of the first and third quartiles), half the IQR equals the MAD. The IQR has a breakdown point of 25%. To make IQR a consistent estimator of the standard deviation of normal distributions, we let

$$\hat{\sigma} = \text{IQR}/(2 * \Phi^{-1}(3/4)) = 0.7413 * \text{IQR}.$$

3. Q_n and S_n

These two estimators were proposed by Croux and Rousseeuw (Croux and Rousseeuw (1992), Rousseeuw and Croux (1993)). To have a consistent estimator of the standard deviation of normal distributions, we let

$$\hat{\sigma} = Q_n/(\sqrt{2} * \Phi^{-1}(5/8)) = 2.2191 * Q_n,$$

where

$$Q_n \approx \text{the first quartile of} \{d_{ij} : d_{ij} = |x_i - x_j|, \quad i < j\}.$$

Note that the total number of d_{ij}'s is $\binom{n}{2}$. The other estimator, S_n, is given by

$$\hat{\sigma} = S_n := 1.1926 * \text{med}_i(\text{med}_j(|x_i - x_j|)).$$

The main motivation for these estimators is the weaknesses of the MAD. Firstly, the MAD is not very efficient at Gaussian distributions (37%). Secondly, it does not consider the skewness, since it computes a symmetric statistic about a location estimate. Note that the computational burden for both Q_n and S_n are relatively heavy compared with that for MAD and IQR.

4. Gini Mean Difference

The *Gini mean difference* is defined by

$$G_n := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |X_i - X_j|.$$

It also involves relatively heavy computation compared with MAD and IQR.

5. "Shorth"

It was proposed by Martin and Zamar (1989, 1993). The "shorth" is defined as the shortest interval containing at least half of the data. That is, $b-a$ is the scale estimator such that $\mathbb{P}(a \leq X \leq b) = 0.5$ and $b-a$ is minimized. This estimator was shown to have the property of minimax bias under some circumstance (see Martin and Zamar (1993)).

In the following we derive the "shorth" based on a normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$. Since this pdf is a decreasing function of $(|x-\mu|)$, the shorth is a symmetric interval centered at μ, i.e. $[\mu - c, \mu + c]$ for some $c > 0$. Then we may write

$$\mathbb{P}(\mu - c \leq X \leq \mu + c) = 0.5.$$

Equivalently,

$$\mathbb{P}\left(-\frac{c}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{c}{\sigma}\right) = \Phi\left(\frac{c}{\sigma}\right) - \Phi\left(-\frac{c}{\sigma}\right) = 2\Phi\left(\frac{c}{\sigma}\right) - 1 = 0.5.$$

So we have

$$\Phi\left(\frac{c}{\sigma}\right) = \frac{3}{4},$$

or

$$c = \Phi^{-1}(3/4) \cdot \sigma.$$

Therefore, the shortest interval is $[\mu - \Phi^{-1}(3/4)\,\sigma, \quad \mu + \Phi^{-1}(3/4)\,\sigma]$, which is the same as the interquartile for normal distributions, and its length is $2\Phi^{-1}(3/4)\,\sigma$. Now suppose that, for a given set, $\hat{s}_h$ represents the length of the shortest interval containing half of the data. Then the (standardized) estimate of the scale parameter σ is

$$\hat{\sigma}_h = \hat{s}_h/(2\Phi^{-1}(3/4)).$$

This is almost the same as the interquartile estimator (for normal distributions, their expectations are the same, however, when applied to data set, they could be slightly different).

6. M-estimator of a scale parameter

We first introduce the definition of χ- functions.

Definition 4.1.2. A χ-*function* denotes a function χ such that

1. $\chi(x)$ is a nondecreasing function of $|x|$;
2. $\chi(0) = 0$;
3. $\chi(x)$ is increasing for $x > 0$ such that $\chi(x) < \chi(\infty)$;
4. If χ is bounded, it is also assumed that $\chi(\infty) = 1$.

In general, any estimator $\hat{\sigma}$ satisfying an equation of the form

$$\frac{1}{n}\sum_{i=1}^{n}\chi\left(\frac{x_i}{\hat{\sigma}}\right) = \delta \tag{4.5}$$

where χ is a χ-*function* and δ is a positive constant, will be called an *M-estimator of scale*. Equivalently, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\chi\left(\frac{x_i}{\hat{\sigma}}\right)-\delta\right)=0. \tag{4.6}$$

This representation includes a wide class of scale estimators corresponding to different choices of ρ. For example, in the Huber estimator, we may take $\chi=\psi^2$.

7. Estimators based on logarithmic transformation

A scale estimator can be transformed into a location parameter as follows. Consider equation (4.6), and let $y_i=\log|x_i|$, $\mu=\log\sigma$ and $\psi(t)=\chi(e^t)-\delta$. Then

$$\frac{1}{n}\sum_{i=1}^{n}\left(\chi\left(\frac{|x_i|}{\sigma}\right)-\delta\right)=\frac{1}{n}\sum_{i=1}^{n}(\chi(e^{y_i-\mu})-\delta)=\frac{1}{n}\sum_{i=1}^{n}\psi(y_i-\mu)=0,$$

where we used the fact that by Definition 4.1.2 $\chi(\cdot)$ is an even function. Thus, the problem of selection of an M-estimator of scale is transformed to a similar problem for location parameter. More specifically, we obtain an M-estimator of scale by

$$\hat{\sigma}=e^{\hat{\mu}}$$

where $\hat{\mu}$ solves the equation $\frac{1}{n}\sum_{i=1}^{n}\psi(y_i-\hat{\mu})=0$.

When applying this log method, the data set after log transformation may not have an obvious center point, even when the original data is drawn from a symmetric distribution.

In the following we apply some of the above methods to the Merton and Kou models using simulated data.

Example 4.1.3. With the same parameter setting as in Example 2.1.2 for the Merton model and Example 2.1.3 for the Kou model, Table 4.2 and Table 4.3 show estimation results of the scale parameter for each model. The true value of

the volatility parameters is $\sigma_0 = 0.2$. In the two tables, "Log 1" corresponds to applying median to the obtained location parameter after the log transformation, and "Log 2" corresponds to applying the Huber Proposal 2 to the obtained location parameter. We repeat 3,000 times for each scenario.

Table 4.2: Estimation results for scale parameter [Merton]

Methods	MAD	IQR	Log 1	Log 2
$\hat{\sigma_0}$	0.2080 (0.0049)	0.2081 (0.0049)	0.1404 (0.0033)	0.1266 (0.0027)

Table 4.3: Estimation results for scale parameter [Kou]

Methods	MAD	IQR	Log 1	Log 2
$\hat{\sigma_0}$	0.2073 (0.0048)	0.2074 (0.0048)	0.1399 (0.0033)	0.1261 (0.0027)

From the results we can see that taking the logarithmic transformation does not lead to good results. MAD and IQR already provide acceptable estimates of the scale parameter. However, they are not efficient enough for the estimation of α. This is because $\hat{\alpha}(t)$ is very sensitive to $\hat{\sigma}$, which can be seen from the expression of $\hat{\alpha}(t)$ in (2.24) and the variance formula of $\hat{\alpha}(t)$ in Proposition 2.2.1. In the following we investigate more efficient estimators of both location and scale parameters.

4.1.3 Simultaneous Estimation of the Location and Scale Parameters

In our problem, both the location and scale parameters are unknown and important. An accurate estimation of the scale parameter is of particular importance, because it appears in the exponents of the estimator $\hat{\alpha}(t)$ and of the variance formula of $\hat{\alpha}(t)$. Therefore, a small error in the scale estimate may potentially result

in a large estimation error of the jump frequency. In this section, we look at the simultaneous estimation of the location and scale parameters.

Equations (4.1) and (4.5) define the M- estimators for the location and scale. If we put them together and consider the studentized version, then we would obtain the following estimating functions

$$\begin{cases} \sum_{i=1}^{n} \psi\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) = 0; \\ \frac{1}{n}\sum_{i=1}^{n} \chi\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) = \delta \end{cases} \tag{4.7}$$

(see Maronna et al. (2006)). Now we solve this system of equations. Let

$$W_1(x) := \begin{cases} \psi(x)/x & \text{if } x \neq 0, \\ \psi'(0) & \text{if } x = 0. \end{cases} \tag{4.8}$$

Then the first equation of (4.7) becomes

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \mu}{\sigma}\right) = \sum_{i=1}^{n} W_1\left(\frac{x_i - \mu}{\sigma}\right) \cdot \frac{x_i - \mu}{\sigma} = 0,$$

which equivalently can be rewritten as

$$\mu = \frac{\sum_{i=1}^{n} W_1\left(\frac{x_i - \mu}{\sigma}\right) \cdot x_i}{\sum_{i=1}^{n} W_1\left(\frac{x_i - \mu}{\sigma}\right)}.$$

Thus, given a starting value $\hat{\mu}_0$, the k-th iteration for μ is

$$\hat{\mu}_{k+1} = \frac{\sum_{i=1}^{n} W_1\left(r_{ki}\right) \cdot x_i}{\sum_{i=1}^{n} W_1\left(r_{ki}\right)}.$$

where $r_{ki} := \frac{x_i - \hat{\mu}_k}{\hat{\sigma}_k}$.

For the scale parameter, let

$$W_2(x) := \begin{cases} \chi(x)/x^2 & \text{if } x \neq 0, \\ \chi''(0) & \text{if } x = 0. \end{cases} \tag{4.9}$$

Then the second equation of (4.7) becomes

$$\frac{1}{n}\sum_{i=1}^{n}\chi\left(\frac{x_i-\mu}{\sigma}\right)=\frac{1}{n}\sum_{i=1}^{n}W_2\left(\frac{x_i-\mu}{\sigma}\right)\cdot\left(\frac{x_i-\mu}{\sigma}\right)^2=\delta,$$

which implies that

$$\sigma^2=\frac{1}{n\delta}\sum_{i=1}^{n}W_2\left(\frac{x_i-\mu}{\sigma}\right)\cdot(x_i-\mu)^2.$$

Thus, given a starting value $\hat{\sigma}_0^2$, the k-th iteration for σ^2 is

$$\hat{\sigma}_{k+1}^2=\frac{\hat{\sigma}_k^2}{n\delta}\sum_{i=1}^{n}W_2(r_{ki})\cdot r_{ki}^2.$$

where $r_{ki}=\frac{x_i-\hat{\mu}_k}{\hat{\sigma}_k}$ as defined above.

Putting them together, we have the k-th iteration for the two parameters:

$$\begin{cases}\hat{\mu}_{k+1}=\dfrac{\sum_{i=1}^{n}W_1\left(r_{ki}\right)\cdot x_i}{\sum_{i=1}^{n}W_1\left(r_{ki}\right)}\\ \hat{\sigma}_{k+1}^2=\dfrac{\hat{\sigma}_k^2}{n\delta}\displaystyle\sum_{i=1}^{n}W_2(r_{ki})\cdot r_{ki}^2\end{cases}\tag{4.10}$$

where $r_{ki}=\frac{x_i-\hat{\mu}_k}{\hat{\sigma}_k}$, and $W_1(\cdot)$ and $W_2(\cdot)$ are given by (4.8) and (4.9).

From equation (4.10) we can see that the location estimate is expressed as a weighted mean, and the scale estimate can be interpreted as a weighted RMS (root mean square) estimate.

In the following we first review some existing estimators of the form (4.7), and then look at some new estimators. The simulation results are presented at the end of this section.

1. Huber's Proposal 2

According to Huber's Proposal 2, we choose $\chi = \psi^2$. That is,

$$\begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \psi\left(\frac{x_i - \mu}{\sigma}\right) = 0 \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \psi^2\left(\frac{x_i - \mu}{\sigma}\right) = \delta, \end{cases} \tag{4.11}$$

where $\psi(x)$ is given in equation (4.2).

To evaluate δ, we have $\delta = \mathbb{E}_\Phi[\psi^2(X)]$, where the standard normal distribution is applied. That is,

$$\begin{aligned} \delta = \mathbb{E}_\Phi[\psi^2(X)] &= \int_{\mathbb{R}} \psi^2(x)\phi(x)dx \\ &= \int_{-k}^{k} x^2\phi(x)dx + \int_{-\infty}^{-k} k^2\phi(x)dx + \int_{k}^{\infty} k^2\phi(x)dx \\ &= 2\int_{0}^{k} x^2\phi(x)dx + 2k^2\int_{k}^{\infty}\phi(x)dx \\ &= -2k\phi(k) + 2\left(\Phi(k) - \frac{1}{2}\right) + 2k^2(1 - \Phi(k)) \\ &= -2k\phi(k) + 2(1 - k^2)\Phi(k) + 2k^2 - 1, \end{aligned} \tag{4.12}$$

where $\phi(\cdot)$ are $\Phi(\cdot)$ are the pdf and cdf of a standard normal distribution.

For a given k, we calculate δ based on the above formula, and then apply the iterative method given by equation (4.10) till the procedure converges (i.e. the difference between the results in two consecutive iterations reaches a pre-defined tolerance level).

2. Tukey Bisquare

For Tukey bisquare estimator, we can use the same ψ as in equation (4.4) and choose $\chi = \rho$ as in equation (4.3). Then

$$\delta = \mathbb{E}_\Phi[\chi(X)] = (\frac{6}{k^2} - \frac{18}{k^4} + \frac{30}{k^6} - 2)\cdot\Phi(k) + (-\frac{2}{k} + \frac{8}{k^3} - \frac{30}{k^5})\cdot\phi(k) - (\frac{3}{k^2} - \frac{9}{k^4} + \frac{15}{k^6} - 2).$$

Similarly, we use the iterative procedure given by (4.10) to find the estimates of both location and scale.

3. Method "HT1": Huber + Trimming 1

As discussed earlier, since the variance of the jump increments is much larger than the variance of the diffusion increments, it is better that we use a redescending M-estimators rather than a monotone M-estimator. For this reason, we revise the Huber estimator by combining it with the trimming method.

We revise Huber's estimator by letting

$$\psi(x) = \begin{cases} x & \text{if } |x| \le k_1 \\ \text{sign}\,(x) \cdot k_1 & \text{if } k_1 < |x| \le k_2 \\ 0 & \text{if } |x| > k_2. \end{cases} \tag{4.13}$$

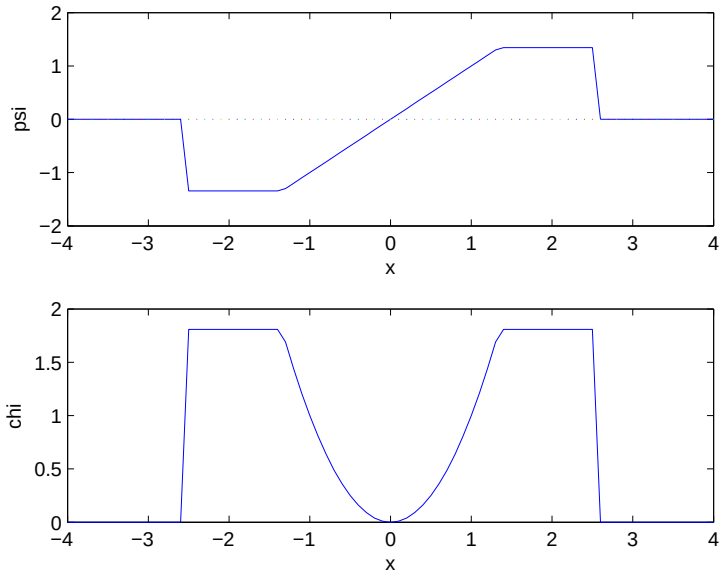
We also choose $\chi = \psi^2$. The plots of ψ and χ are shown in Figure 4.3. Since we use a redescending ψ-function, any large values will be rejected, which seems reasonable for our problem where large jump values should be removed for the estimation of diffusion parameters.

The value of k_1 is set the same as in Huber estimator, i.e. $k_1 = 1.345$. We choose $k_2 = 2.576$ which corresponds to 1%-trimming for a standard normal distribution.

To evaluate δ we have

$$\begin{aligned} \delta = \mathbb{E}_\Phi[\psi^2(X)] &= \int_{\mathbb{R}} \psi^2(x)\phi(x)dx \\ &= \int_{-k_1}^{k_1} x^2\phi(x)dx + \int_{k_1}^{k_2} k_1^2\phi(x)dx + \int_{-k_2}^{-k_1} k_1^2\phi(x)dx \\ &= 2\int_0^{k_1} x^2\phi(x)dx + 2k_1^2\int_{k_1}^{k_2}\phi(x)dx \\ &= -2k_1\phi(k_1) + 2\left(\Phi(k_1) - \frac{1}{2}\right) + 2k_1^2[\Phi(k_2) - \Phi(k_1)] \\ &= -2k_1\phi(k_1) + 2(1-k_1^2)\Phi(k_1) + 2k_1^2\Phi(k_2) - 1, \end{aligned} \tag{4.14}$$

Figure 4.3: Proposed ψ- and χ- Functions



In the plot, $k_1 = 1.345, k_2 = 2.576$.

where $\phi(\cdot)$ are $\Phi(\cdot)$ are the pdf and cdf of a standard normal distribution. Then we use equation (4.10) to perform the iterations till the procedure converges.

4. Method "HT2": Huber + Trimming 2

This method "HT2" is based on the same idea as the method "HT1": combining the Huber estimator with the trimming method, but the procedure is slightly different. The following are the steps for the method HT2:

- **Step 1**: Set the initial values: $\mu^{(0)}$ and $\sigma^{(0)}$.
- **Step 2**: Trim the data (by the pre-defined level) and then apply the Huber's Proposal 2 to the trimmed data to obtain new estimates $\mu^{(1)}$ and $\sigma^{(1)}$.
- **Step 3**: Repeat Steps 1 and 2 till it converges.

By trimming the data, we mean "deleting" a certain percentage of observations from each end of the data, which is not equivalent to "assuming them zeros" and putting the zeros back into the data set. Note that Step 2 involves another layer of iterations, since Huber's Proposal 2 itself is an iterative procedure. The tolerance level is set the same as for the outside iteration (i.e. the tolerance level in Step 3).

We set the degree of trimming corresponding to 1% in the normal data, that is, the observation x_i with $(x_i - \hat{\mu}_0)/\hat{\sigma}_0 > 2.576$ will be trimmed. We still use $k = 1.345$ when we apply Huber's Proposal 2.

Note that the two methods, HT1 and HT2, are different. Suppose they start from the same initial values $\hat{\mu}_0$ and $\hat{\sigma}_0$. Then after one iteration, the obtained $\hat{\mu}_1$ is the same, but $\hat{\sigma}_1$ are different. This is because in the step of calculating $\hat{\sigma}^2_{k+1}$ based on the second formula of (4.10), the factors n and δ in the denominator are different for the two methods. For the method HT1, n corresponds to the total number of observations and δ is the one defined in (4.14); while for method HT2, n corresponds to the number of observations after trimming and δ is the one defined in (4.12).

5. Method "MADT": MAD + Trimming

It is the same procedure as in method "HT2", but we replace the Huber's Proposal 2 in Step 2 by using the median and the MAD.

6. Method "IQRT": IQR + Trimming

It is the same procedure as in method "HT2", but we replace the Huber's Proposal 2 in Step 2 by using the median and the IQR.

Example 4.1.4. With the same parameter setting as in Example 2.1.2 for Merton model and Example 2.1.3 for Kou model, Table 4.4 and Table 4.5 show the results

of simultaneous estimation of the location and scale for Merton model and Kou model based on the above six methods.

Table 4.4: Simultaneous estimation of location and scale [Merton]

	Huber	Bisquare	HT1	HT2	MADT	IQRT
$\hat{\mu_0}$	0.1002	0.1003	0.0998	0.1002	0.0998	0.0998
	(0.0680)	(0.0848)	(0.0701)	(0.0700)	(0.0848)	(0.0848)
$\hat{\sigma_0}$	0.2094	0.2205	0.2007	0.2004	0.2007	0.2007
	(0.0038)	(0.0054)	(0.0047)	(0.0041)	(0.0054)	(0.0054)

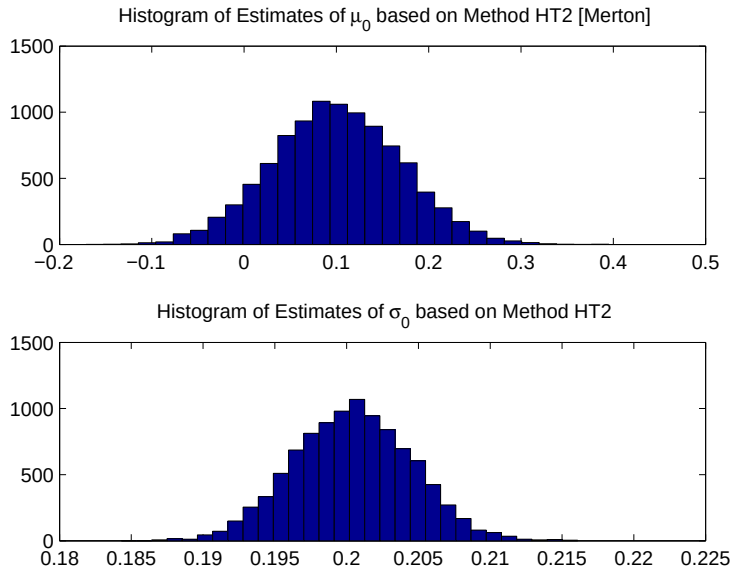
Table 4.5: Simultaneous estimates of location and scale [Kou]

	Huber	Bisquare	HT1	HT2	MADT	IQRT
$\hat{\mu_0}$	0.0057	0.0538	0.0767	0.0736	0.0767	0.0767
	(0.0682)	(0.0851)	(0.0701)	(0.0701)	(0.0851)	(0.0851)
$\hat{\sigma_0}$	0.2087	0.2197	0.2002	0.1999	0.2002	0.2002
	(0.0037)	(0.0053)	(0.0045)	(0.0040)	(0.0053)	(0.0053)

In Table 4.5, we see that the location estimates are always biased downward. This is because for Kou model the parameters are set such that there are more negative jumps and, moreover, the average size of negative jumps is larger than the average size of positive jumps. For the Merton model, this problem does not exist since we set $\mu_X = 0$, i.e. the jump observations and the diffusion observations have the same "center".

From Table 4.4 and Table 4.5 we see that the method "HT2" performs the best. In Figure 4.4 we plot the histograms of the estimates of μ_0 and σ_0 based on this method for the Merton model. It shows a bell shape. It is the same case for the Kou model. In the following we use the method "HT2" to estimate the location and scale parameters of the diffusion component and then estimate the jump frequency.

Figure 4.4: Histograms of $\hat{\mu}_0$ and $\hat{\sigma}_0$ based on Method HT2 [Merton]



4.2 Estimation of the Jump Frequency

We propose the following procedure to estimate the jump ratio α or the jump intensity λ:

Step 1 Use the method "HT2" to obtain estimates, $\hat{\mu}$ and $\hat{\sigma}$, of the location parameter μ and the scale parameter σ.

Step 2 To select t, apply Method 3 proposed in Section 2.3. Denote the selected value by t_{M3}.

Step 3 The estimate of α is $\hat{\alpha}(t_{M3})$, where $\hat{\alpha}(t)$ is defined by

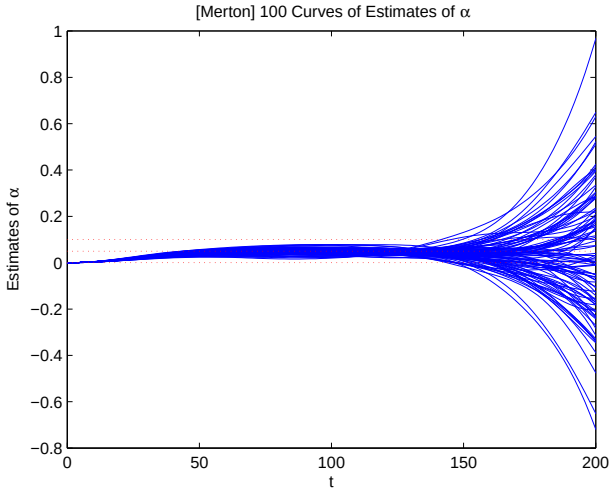
$$\begin{aligned}\hat{\alpha}(t) &:= 1 - \Re(\hat{\varphi}_Y(t)/\hat{\varphi}_Z(t)) \\ &= 1 - e^{\frac{1}{2}\hat{\sigma}^2 t^2} \cdot \frac{1}{n}\sum_{j=1}^{n} \cos((Y_j - \hat{\mu})t), \quad t > 0,\end{aligned} \tag{4.15}$$

as an analogy to (2.24). Then the estimate of λ is obtained by

$$\hat{\lambda}(t) := \hat{\alpha}(t)/\Delta.$$

Example 4.2.1 (Merton and Kou models)**.** With the same parameter setting as in Example 2.1.2 for Merton model, Figure 4.5 shows 100 simulated curves of $\hat{\alpha}(t)$ defined in (4.15). It is interesting to see that, as t increases, the variance of $\hat{\alpha}(t)$ seems to first increase, then decrease and then increase again. For the Kou model, it sometimes shows the same pattern.

Figure 4.5: 100 Curves of $\hat{\alpha}(t)$ when μ and σ unknown [Merton]



In the following, we provide a similar result to Proposition 2.2.1. More specifically, we derive the mean and variance formulas, where instead of the true values μ and σ we use values $\hat{\mu}$ and $\hat{\sigma}$. Since $\hat{\mu}$ and $\hat{\sigma}$ are not necessarily equal to μ and σ, the result below shows that when the diffusion parameters must be estimated, it is more difficult to analyze the behavior of the estimator $\hat{\alpha}(t)$.

Proposition 4.2.2. *Suppose $\hat{\mu}$ and $\hat{\sigma}$ in (4.15) are estimated values (instead of estimators) of μ and σ, respectively. Then the expectation and variance of $\hat{\alpha}(t)$*

defined by (4.15) are given as follows

$$\mathbb{E}\left[\hat{\alpha}(t)\right] = 1 - e^{\frac{1}{2}(\hat{\sigma}^2-\sigma^2)t^2}A(t) \tag{4.16}$$

and

$$Var\left(\hat{\alpha}(t)\right) = \frac{1}{2n}\left[e^{\hat{\sigma}^2t^2} + e^{(\hat{\sigma}^2-2\sigma^2)t^2}A(2t) - 2e^{(\hat{\sigma}^2-\sigma^2)t^2}A^2(t)\right] \tag{4.17}$$

where

$$A(t) := [(1-\alpha) + \alpha R(t)]\cos((\mu-\hat{\mu})t) - \alpha I(t)\sin((\mu-\hat{\mu})t),$$

and $R(t) := \Re(\varphi_X(t))$, $I(t) := \Im(\varphi_X(t))$.

Proof. The techniques used in this proof are similar to the ones used in the proof of Proposition 2.2.1. The expectation of $\hat{\alpha}(t)$ is

$$\begin{aligned}\mathbb{E}\left[\hat{\alpha}(t)\right] &= \mathbb{E}\left[1 - e^{\frac{1}{2}\hat{\sigma}^2t^2}\cdot\frac{1}{n}\sum_{j=1}^{n}\cos((Y_j-\hat{\mu})t)\right] \\ &= 1 - e^{\frac{1}{2}\hat{\sigma}^2t^2}\mathbb{E}\left[\cos((Y-\hat{\mu})t)\right].\end{aligned}$$

Let $\varphi_{Y-\hat{\mu}}(t)$ denote the characteristic function of the random variable $Y-\hat{\mu}$. Then

$$\begin{aligned}\mathbb{E}\left[\cos((Y-\hat{\mu})t)\right] &= \Re\left(\varphi_{Y-\hat{\mu}}(t)\right) \\ &= \Re\left(e^{-i\hat{\mu}t}\varphi_Y(t)\right) \\ &= \Re\left(e^{-i\hat{\mu}t}\cdot e^{i\mu t-\frac{1}{2}\sigma^2t^2}\cdot((1-\alpha)+\alpha\varphi_X(t))\right) \\ &= e^{-\frac{1}{2}\sigma^2t^2}\cdot\{\cos((\mu-\hat{\mu})t)\cdot[(1-\alpha)+\alpha R(t)] - \sin((\mu-\hat{\mu})t)\cdot\alpha I(t)\}\end{aligned} \tag{4.18}$$

Thus, the expectation formula (4.16) is proved. Similarly, the variance of $\hat{\alpha}(t)$ can also be found by using (4.18) and the formula $\cos^2 x = (1+\cos 2x)/2$. That is,

$$\begin{aligned}
\mathrm{Var}\left(\hat{\alpha}(t)\right) &= \mathrm{Var}\left(1 - e^{\frac{1}{2}\hat{\sigma}^2 t^2} \cdot \frac{1}{n}\sum_{j=1}^{n} \cos((Y_j - \hat{\mu})t)\right) \\
&= \frac{1}{n} e^{\hat{\sigma}^2 t^2} \cdot \mathrm{Var}\left(\cos((Y - \hat{\mu})t)\right) \\
&= \frac{1}{n} e^{\hat{\sigma}^2 t^2} \cdot \left\{\mathbb{E}\left[\cos^2((Y - \hat{\mu})t)\right] - \left(\mathbb{E}\left[\cos((Y - \hat{\mu})t)\right]\right)^2\right\} \\
&= \frac{1}{n} e^{\hat{\sigma}^2 t^2} \cdot \left\{\frac{1}{2}\left[1 + \mathbb{E}\left[\cos((Y - \hat{\mu})2t)\right]\right] - \left(\mathbb{E}\left[\cos((Y - \hat{\mu})t)\right]\right)^2\right\}.
\end{aligned}$$

After some algebra we obtain the desired result. ⊠

If we let $\hat{\mu} = \mu$ and $\hat{\sigma} = \sigma$, then equations (4.16) and (4.17) simplify to equations (2.27) and (2.29), i.e. we recover the formulas presented in Proposition 2.2.1 when the diffusion parameters are known.

Example 4.2.3 (Merton model)**.** We use the same parameters as in Example 2.1.2. For the selection of t, we use Method 3 with $t_{\max} = 100$.

Figure 4.6: Histograms of estimates [Merton]

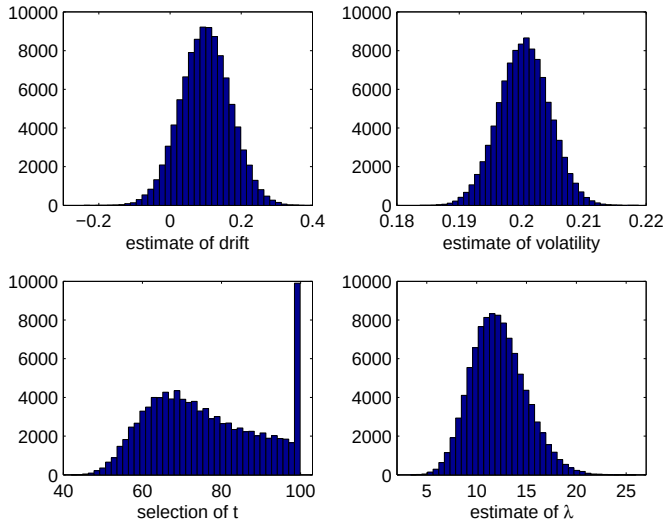


Figure 4.6 shows the histograms of the estimates. The estimates of the μ_0 and σ_0 seems to be normally distributed, but the estimates of λ shows a slightly heavy right tail.

Table 4.6: Estimation results using Method 3 for different scenarios [Merton]

n		$\sigma_X \approx 3.79\%$		$\sigma_X = 10\%$		$\sigma_X = 60\%$	
		$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$
1000	$\hat{\mu_0}$	.1000(.1112)	.0998(.1156)	.0994(.1105)	.1007(.1145)	.0998(.1101)	.1000(.1128)
	$\hat{\sigma_0}$	.2003(.0065)	.2045(.0069)	.1984(.0064)	.2004(.0067)	.1970(.0063)	.1973(.0065)
	t_{M3}	73.23(15.88)	67.17(11.48)	37.74(14.75)	36.01(10.92)	5.75(1.77)	5.97(1.71)
	$\hat{\lambda}_{M3}$	12.56(4.22)	20.82(4.31)	14.04(3.18)	26.02(3.51)	13.15(2.19)	25.92(3.02)
2500	$\hat{\mu_0}$	.1000(.0702)	.0999(.0732)	.1001(.0701)	.1004(.0726)	.1000(.0695)	.0999(.0717)
	$\hat{\sigma_0}$	.2004(.0041)	.2046(.0043)	.1985(.0040)	.2005(.0042)	.1970(.0040)	.1974(.0041)
	t_{M3}	76.19(13.83)	66.47(8.16)	40.74(15.30)	37.33(10.42)	6.06(1.72)	6.27(1.69)
	$\hat{\lambda}_{M3}$	12.21(2.72)	20.44(2.69)	13.74(2.16)	25.57(2.20)	12.93(1.38)	25.60(1.90)
10000	$\hat{\mu_0}$	.1000(.0351)	.1000(.0366)	.1000(.0350)	.0999(.0363)	.0999(.0347)	.0998(.0357)
	$\hat{\sigma_0}$	.2005(.0020)	.2046(.0022)	.1986(.0020)	.2006(.0021)	.1971(.0020)	.1974(.0020)
	t_{M3}	79.60(11.12)	65.37(3.72)	47.28(17.62)	39.02(9.65)	6.52(1.72)	6.69(1.69)
	$\hat{\lambda}_{M3}$	11.96(1.39)	20.24(1.33)	13.59(1.36)	25.18(1.10)	12.73(0.69)	25.31(0.95)

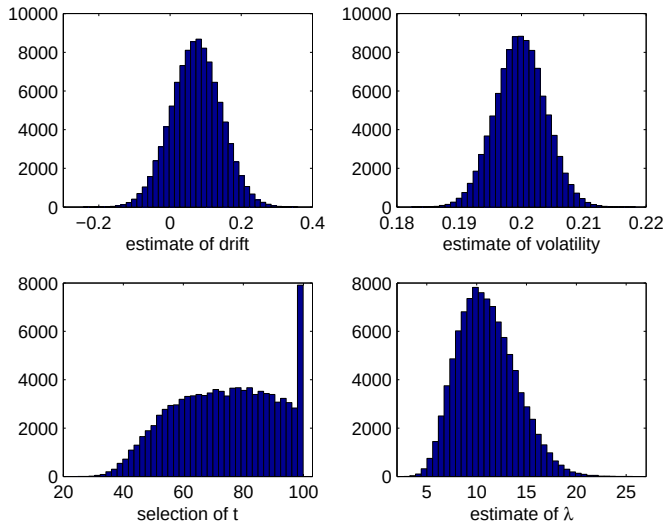
Table 4.6 summarizes the estimation results for different scenarios of parameter settings. In each scenario, Method 3 with $t_{\max} = 100$ is used for the selection of t. The following are some findings:

- Overall, the results are acceptable. The bias of $\hat{\lambda}$ is less than 19% of the true value for every scenario.
- For the estimation of σ_0, we see that the mean value of $\hat{\sigma_0}$ increases when λ increases, or when σ_X decreases, or even when n increases. This implies that there is still room for improvement in the robust estimation of location and scale. For example, we may consider some "adaptive" method to choose the parameters in the robust estimation.

- Comparing this table with Table 2.8 where the diffusion parameters are assumed known, we see that the selected value t_{M3} tends to be smaller when $\hat{\sigma_0}$ is biased up and larger when $\hat{\sigma_0}$ is biased down (except for the first column), so that the product $\hat{\sigma_0} \cdot t_{M3}$ is corrected to some degree*. However, the mean value of $\hat{\lambda}_{M3}$ is still smaller when $\hat{\sigma_0}$ is biased up, and larger when $\hat{\sigma_0}$ is biased down. By the expectation formula of $\hat{\alpha}(t)$ in (4.16), this means that the correction to the product $\hat{\sigma_0} \cdot t_{M3}$ by t_{M3} is still not enough to eliminate the effect of the bias of $\hat{\sigma}_0$ on the estimation of λ.

- From Table 4.6 we also see that, $\hat{\lambda}_{M3}$ decreases when n increases. There are two reasons for this: firstly, it is the same pattern as in Table 2.8 where the diffusion parameters are assumed known; secondly, the mean value of $\hat{\sigma_0}$ increases when n increases.

Figure 4.7: Histograms of estimates [Kou]



*The product $\hat{\sigma_0} \cdot t_{M3}$ plays a critical role in the estimation of α. This can be seen from the definition of $\hat{\alpha}(t)$ given by (4.15), or the expectation and variance formulas in Proposition 4.2.2.

Example 4.2.4 (Kou model)**.** We use the same parameters as in Example 2.1.3 for Kou model. Figure 4.7 shows the histogram of the estimates.

Table 4.7 shows the estimation results for different scenarios of parameter settings. The findings are similar to the ones in the Merton case.

Table 4.7: Estimation results using Method 3 for different scenarios [Kou]

		$\eta_+ = 1/0.02$, $\eta_- = 1/0.04$		$\eta_+ = 1/0.1$, $\eta_- = 1/0.2$	
		$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$
n=1000	$\hat{\mu_0}$	0.0736 (0.1106)	0.0438 (0.1145)	0.0957 (0.1103)	0.0913 (0.1139)
	$\hat{\sigma_0}$	0.1997 (0.0064)	0.2032 (0.0067)	0.1979 (0.0064)	0.1993 (0.0066)
	t_{M3}	66.47 (19.45)	64.36 (15.33)	16.61 (6.03)	18.25 (5.40)
	$\hat{\lambda}_{M3}$	11.14 (4.33)	18.29 (4.40)	11.39 (2.48)	22.81 (3.44)
n=2500	$\hat{\mu_0}$	0.0739 (0.0699)	0.0435 (0.0723)	0.0962 (0.0696)	0.091 (0.0720)
	$\hat{\sigma_0}$	0.1998 (0.0040)	0.2033 (0.0043)	0.198 (0.0040)	0.1994 (0.0042)
	t_{M3}	73.94 (16.55)	68.45 (12.92)	19.41 (5.76)	21.14 (5.32)
	$\hat{\lambda}_{M3}$	11.05 (2.93)	17.96 (2.82)	11.65 (1.64)	23.21 (2.26)
n=10000	$\hat{\mu_0}$	0.0737 (0.0350)	0.0438 (0.0362)	0.0959 (0.0349)	0.0913 (0.0360)
	$\hat{\sigma_0}$	0.1999 (0.0020)	0.2034 (0.0021)	0.198 (0.0020)	0.1994 (0.0021)
	t_{M3}	83.47 (12.61)	72.27 (9.42)	24.95 (6.28)	26.56 (5.58)
	$\hat{\lambda}_{M3}$	11.03 (1.62)	17.64 (1.40)	12.06 (0.94)	23.78 (1.22)

To summarize, in this chapter we applied some robust methods to estimate the diffusion parameters, and then, using Method 3 for selection of t, we estimated the jump frequency. The results of our simulation study show that the proposed estimation method of the jump frequency yields reasonably accurate estimates of this parameter. In the next chapter, we first apply this method to estimate the jump frequency, and then propose a method to further estimate the jump size distribution.

Chapter 5

Estimation of the Jump Size Distribution

In previous chapters, we have discussed the estimation of the jump frequency. In this chapter we propose a method for the estimation of the jump size distribution.

The form of the characteristic function of the jump size X given by (2.10) suggests that as its estimator we can use

$$\hat{\varphi}_X(t) = \frac{\frac{\hat{\varphi}_Y(t)}{\hat{\varphi}_Z(t)} - (1-\hat{\alpha})}{\hat{\alpha}}. \tag{5.1}$$

As explained below, $\hat{\varphi}_X(t)$ indeed can be used in practice:

- We can first obtain robust estimates $\hat{\mu}$ and $\hat{\sigma}$, and then the characteristic function of Z can be estimated by $\hat{\varphi}_Z(t) = e^{i\hat{\mu}t - \frac{1}{2}\hat{\sigma}^2 t^2}$.
- An estimate $\hat{\alpha}$ of jump frequency can be obtained by using the method discussed in previous chapters. Note that $\hat{\alpha}$ in (5.1) is a fixed value (instead of a function of t).

- The characteristic function of the observable variable Y can be estimated by using its empirical characteristic function or other estimators of the characteristic function, like the ones that we have discussed in Section 2.4.

Equation (5.1) involves a ratio of two characteristic functions: $\hat{\varphi}_Y(t)/\hat{\varphi}_Z(t)$. As discussed in Chapter 2, this ratio can be interpreted as an estimator of the characteristic function $\hat{\varphi}_V(t)$ of the random variable $V = Y - Z$, where Y and Z are independent. Suppose we use the e.c.f. to estimate the c.f. of Y. Then, similarly to equation (2.22), we have

$$\begin{aligned}\hat{\varphi}_V(t) &= \hat{\varphi}_Y(t)/\hat{\varphi}_Z(t) \\ &= \frac{1}{n}\sum_{j=1}^{n} e^{itY_j}/e^{i\hat{\mu}t-\frac{1}{2}\hat{\sigma}^2t^2} \\ &= e^{\frac{1}{2}\hat{\sigma}^2t^2}\left\{\frac{1}{n}\sum_{j=1}^{n}\cos((Y_j-\hat{\mu})t) + i\cdot\frac{1}{n}\sum_{j=1}^{n}\sin((Y_j-\hat{\mu})t)\right\},\end{aligned} \tag{5.2}$$

where the real and imaginary parts involve sums of trigonometric functions.

After we have the characteristic function of the jump size, we might want to find the density function of the jump size, since in most applications we are more interested in obtaining the density function. Then it becomes a Fourier inversion problem. An introduction to the "conventional" Fourier inverse transform is given in Section 5.1.2.

However, our problem is not a "conventional" Fourier inversion. As we have already noticed, $\hat{\varphi}_V(t)$ in (5.2) is a ratio of two characteristic functions where the numerator is an empirical characteristic function but the denominator is in the form of the characteristic function of a normal distribution, and thus the denominator converges to zero much faster than the numerator does. Therefore, $\hat{\varphi}_V(t)$ in (5.2) is unbounded and not integrable. In this situation, some regularization procedure needs to be taken before we apply the Fourier inverse transform. This is a common issue in the so-called *de-convolution* problem. The monograph by Meister (2009)

provides a good exposition of the de-convolution problem. We will briefly review this topic in Section 5.1.1.

In Section 5.2, we discuss the estimation of jump size distribution. Simulation studies are given in Section 5.3, and a real data example is provided in Section 5.4.

5.1 Introduction

5.1.1 De-convolution Problem

Many facts presented in this section are based on Meister (2009).

Our goal is to estimate a function f while we observe values from the distribution

$$h = f * G = \int f(x-y)dG(y),$$

that is, the convolution of f and some probability distribution G. Here we focus on the case when G is absolute continuous and thus has a density function g. Then we may write

$$h = f * g = \int f(x-y)g(y)dy. \tag{5.3}$$

This density corresponds to the following model:

$$Y = X + \epsilon, \tag{5.4}$$

where $Y \sim h, X \sim f$ and $\epsilon \sim g$. The function g is often called the *error density* or the *blurring density*. Here Y is the observable variable, but the distribution of X is what we are interested in.

The characteristic function techniques are used commonly in the de-convolution problem since the characteristic function transforms the convolution operator into the multiplication operator. More concretely, (5.3) is equivalent to

$$\varphi_Y = \varphi_X \cdot \varphi_\epsilon,$$

where φ_Y, φ_X and φ_ϵ are the characteristic functions of Y, X and ϵ, respectively. Assuming that distribution of ϵ is known (or can be estimated), we can use the following steps to estimate f:

Step 1: Estimate φ_Y based on observed data, and denote the estimator by $\hat{\varphi}_Y$.

Step 2: Divide $\hat{\varphi}_Y(t)$ by $\varphi_\epsilon(t)$, leading to the estimator $\hat{\varphi}_X(t)$.

Step 3: Regularize $\hat{\varphi}_X$ so that its Fourier inverse transform $\hat{f}$ exists. Take $\hat{f}$ as the de-convolution estimator of f.

This procedure looks straightforward; however, the regularization in the last step involves many mathematical efforts.

For Step 1, suppose we estimate φ_Y by using the empirical characteristic function:

$$\hat{\varphi}_{Y,1}(t) = \frac{1}{n}\sum_{j=1}^{n} \exp(itY_j). \tag{5.5}$$

Then an estimator of $\varphi_X(t)$ is given by

$$\hat{\varphi}_{X,1}(t) = \hat{\varphi}_{Y,1}(t)/\varphi_\epsilon(t) = \frac{1}{n}\sum_{j=1}^{n} \exp(itY_j)/\varphi_\epsilon(t), \tag{5.6}$$

assuming that φ_ϵ vanishes nowhere. For each fixed t, $\hat{\varphi}_{X,1}(t)$ is an unbiased and strongly consistent estimator of $\varphi_X(t)$, since the e.c.f. $\hat{\varphi}_{Y,1}(t)$ is an unbiased and strongly consistent estimator of $\varphi_Y(t)$. We are interested in the density function f of X, so we shall apply the Fourier inversion to $\hat{\varphi}_{X,1}(t)$. Using the standard Fourier inverse formula (5.12) presented in Section 5.1.2, a naive estimator of f is given by

$$\hat{f}_{\text{naive}}(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-itx}\hat{\varphi}_{X,1}(t)dt.$$

However, this estimator is not well-defined because $\hat{\varphi}_{X,1}(t)$ is neither integrable nor square-integrable over $\mathbb{R}$. (Note that the Fourier transform can be defined on $L_1(\mathbb{R})$ or $L_2(\mathbb{R})$; see Appendix B). The tail behaviors of $\hat{\varphi}_{X,1}(t)$ and $\varphi_X(t)$ are

significantly different: $\hat{\varphi}_{X,1}(t)$ oscillates with amplitude going to infinity as $t \to \infty$, while $|\varphi_X(t)| < 1$ for all $t \in \mathbb{R}$ and $\varphi_X(t) \to 0$ as $t \to \infty$.

Therefore, we need to *regularize* $\hat{\varphi}_{X,1}(t)$ before the Fourier inversion is applied. One of the most popular regularization methods is the *kernel method*. Another method is the *truncation method*. In the following we discuss both methods.

Kernel Method

The kernel method has been introduced in Section 2.4. Let $K(x)$ be a kernel function and $b > 0$ be the window width. Then the kernel density function of the observable variable Y in (5.4) is given by

$$\hat{h}(x) = \frac{1}{nb} \sum_{j=1}^{n} K\left(\frac{x - Y_j}{b}\right).$$

Under some condition on K, e.g. $K \in L_1(\mathbb{R}) \cap L_2(\mathbb{R})$, the Fourier transform of $\hat{h}(x)$ exists and is given by

$$\hat{\varphi}_{Y,2}(t) := \hat{\varphi}_{Y,1}(t) \cdot K^{ft}(bt),$$

where $\hat{\varphi}_{Y,1}(t)$ is the e.c.f. given by (5.5), and K^{ft} is the Fourier transform of the function K. This can be seen from Lemma 2.4.3. Then we obtain a second empirical version of $\varphi_X(t)$ by

$$\begin{aligned} \hat{\varphi}_{X,2}(t) = \hat{\varphi}_{Y,2}(t)/\varphi_\epsilon(t) &= \hat{\varphi}_{Y,1}(t) K^{ft}(bt)/\varphi_\epsilon(t) \\ &= K^{ft}(bt) \frac{1}{n} \sum_{j=1}^{n} \exp(itY_j)/\varphi_\epsilon(t). \end{aligned} \tag{5.7}$$

The only difference between $\hat{\varphi}_{X,1}(t)$ and $\hat{\varphi}_{X,2}(t)$ is the deterministic term $K^{ft}(bt)$. Adding this term can be viewed as a regularization for $\hat{\varphi}_{X,1}(t)$. We hope $\hat{\varphi}_{X,2}(t) \in L_1(\mathbb{R}) \cap L_2(\mathbb{R})$ so that its Fourier inverse would exist. However, this does not hold for all kernel functions (and window widths). For example, if K is the Gaussian kernel, ϵ is normally distributed with variance σ_ϵ^2 and $\sigma_\epsilon > b$, then $\hat{\varphi}_{X,2}(t)$ is not

integrable or square integrable. There is one class of kernel functions whose Fourier transforms are bounded and compactly supported, which includes the *sinc-kernel*:

$$K(x) = \frac{\sin x}{\pi x} \tag{5.8}$$

with the Fourier transform given by

$$K^{ft}(t) = 1_{[-1,1]}(t), \tag{5.9}$$

where 1_A denotes the indicator function of a set A. For such a kernel, $\hat{\varphi}_{X,2}(t)$ in (5.7) is supported on $[-1/b, 1/b]$ and bounded whenever $\varphi_\epsilon(t) \neq 0$ (since $\varphi_\epsilon(\cdot)$ is continuous). Then $\hat{\varphi}_{X,2}(t)$ is integrable and square-integrable, so we can apply the Fourier inversion to $\hat{\varphi}_{X,2}(t)$. This leads to the following density estimator

$$\begin{aligned} \hat{f}_{X,2}(x) &= \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \hat{\varphi}_{X,2}(t) dt \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} K^{ft}(bt) \frac{1}{n} \sum_{j=1}^{n} \exp(itY_j)/\varphi_\epsilon(t) dt \end{aligned} \tag{5.10}$$

which is well-defined for any non-vanishing φ_ϵ and bounded and compactly supported K^{ft}. This estimator $\hat{f}_{X,2}(x)$ has been known as the standard *de-convolution kernel density estimator* (See Meister (2009)).

Note that to avoid complex values in the estimator of a density function, we can just take the real part of the estimator $\hat{f}_{X,2}(x)$.

Truncation Method

Another method for the regularization step is *truncation*, where the regularized estimator of $\varphi_X(t)$ is given by

$$\hat{\varphi}_{X,3}(t) = \begin{cases} \hat{\varphi}_{X,1}(t) & \text{if } |t| \leq t_{trunc} \\ 0 & \text{if } |t| > t_{trunc}, \end{cases} \tag{5.11}$$

for some constant $t_{trunc} > 0$, where $\hat{\varphi}_{X,1}(t)$ is defined in (5.6).

This truncation method can be viewed as a sub-class of the kernel method where $K^{ft}(t)$ is an indicator function over a compact set. That is, when the kernel function K is the sinc-kernel and $b = 1/t_{trunc}$, the estimator $\hat{\varphi}_{X,2}(t)$ becomes $\hat{\varphi}_{X,3}(t)$. On the other hand, the kernel method with K^{ft} bounded and compactly supported is essentially the same as the truncation method, as long as there is no other information implying the necessity to modify the middle part of $\hat{\varphi}_{X,1}(t)$, i.e. when $|t| \leq t_{trunc}$. Typically the estimation error of $\hat{\varphi}_{X,1}(t)$ is small when t is close to zero and gets larger when t increases, so we might want to keep the middle part of $\hat{\varphi}_{X,1}(t)$ unchanged. Therefore, the kernel method and the truncation method are equivalent.

Other Methods

There are other methods proposed for the deconvolution problem, for example, the *wavelet-based method* and the *ridge-parameter approach*. In the wavelet-based method, an orthogonal series is used to estimate the density while the coefficients of the orthonormal bases are calculated using the *Parseval's identity*. The ridge-parameter approach was proposed to relax the condition that φ_ϵ vanishes nowhere. More details can be found in Meister (2009).

In our problem, ϵ corresponds to a normal distribution so that the condition that φ_ϵ vanishes nowhere holds true automatically. Thus, there is no need to consider the ridge-parameter approach. Comparing with the wavelet-based method, the kernel method or the truncation method is more convenient for implementation.

Therefore, in the following we shall use the kernel estimator $\hat{f}_{X,2}(x)$ given in (5.10) with K^{ft} bounded and compactly supported, or equivalently, the truncation method in (5.11), as the deconvolution density estimator.

5.1.2 Inverting Transforms of Probability Distributions

This section discusses the calculation of the probability density distributions by numerically inverting characteristic functions. Other transforms, like the Laplace transform and generating functions, can be applied similarly. We can also use this method to recover the cumulative distribution functions. Most of the facts in this section are based on Abate and Whitt (1992).

The following result can be found in many monographs on Fourier transform, or in probability books, for example, Grimmett and Stirzaker (2001, page 189).

Theorem 5.1.1. *If X is continuous with density function f and characteristic function φ then*

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi(t) dt \tag{5.12}$$

at every point x at which f is differentiable.

The cumulative distribution function of X can also be found by

$$F(x_2) - F(x_1) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-itx_1} - e^{-itx_2}}{it} \varphi(t) dt.$$

In most of the applications we are more interested in obtaining the density function rather than the distribution function, so in the following we focus on recovering the density function.

Theorem 5.1.1 implies the integrability of the integrand in equation (5.12). Since $f(x)$ is a real function, we may consider only the real part of the right-hand side of (5.12), i.e.

$$\begin{aligned} f(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \Re(e^{-itx} \varphi(t)) dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} (\Re(\varphi(t)) \cos(tx) + \Im(\varphi(t)) \sin(tx))\, dt \\ &= \frac{1}{\pi} \int_{0}^{\infty} (\Re(\varphi(t)) \cos(tx) + \Im(\varphi(t)) \sin(tx))\, dt, \end{aligned} \tag{5.13}$$

where the last identity is due to the fact that $\Re(\varphi(t))$ is an even function in t and $\Im(\varphi(t))$ is an odd function in t. If we denote

$$g(t) := \frac{1}{\pi}\left(\Re(\varphi(t))\cos(tx) + \Im(\varphi(t))\sin(tx)\right), \tag{5.14}$$

where we suppress the dependence of g on x to simplify the notation, then (5.13) can be written as

$$f(x) = \int_0^\infty g(t)dt.$$

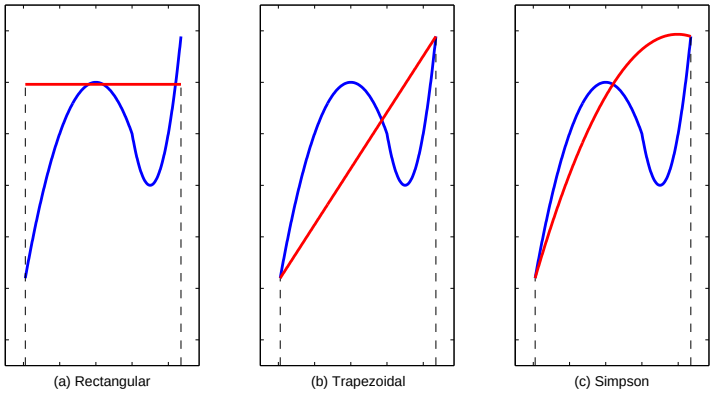
To evaluate this integral, different approximation method can be applied, such as the *rectangular rule* (or *midpoint rule*), the *trapezoidal rule* and the *Simpson's rule*. See Figure 5.1 for the underlying ideas of these three methods how to approximate the area. Note that the Simpson's rule can be viewed as a weighted average of the rectangular rule and the trapezoidal rule. It is known that for smooth functions the Simpson's rule has faster convergence than the trapezoidal rule. However, when the function is periodic, the trapezoidal rule typically converges faster than the other less primitive methods. This can be explained intuitively as follows: when the function is periodic or oscillating, there are about as many sections of the graph that are concave up as concave down, so the errors tend to cancel. See Abate and Whitt (1992) for more details. In our problem, the integrand $g(t)$ exhibits oscillating behavior, so we choose the trapezoidal rule to evaluate the integral.

By the *trapezoidal rule*, the approximation is given by

$$\int_a^b g(u)du \approx h\left[\frac{g(a)+g(b)}{2} + \sum_{k=1}^{n-1} g(a+kh)\right]$$

where h is the mesh size of the partition. This rule also applies when $a = -\infty$ or

Figure 5.1: Different methods to evaluate integrals



$b = \infty$. Using this approximation, we have

$$\begin{aligned} f(x) &= \int_0^\infty g(t)dt \\ &\approx f_h(x) := h\left[\frac{g(0)+g(\infty)}{2} + \sum_{k=1}^{\infty} g(kh)\right] \\ &= \frac{h}{\pi}\left[\frac{1}{2} + \sum_{k=1}^{\infty} \Re(\varphi(kh))\cos(khx) + \sum_{k=1}^{\infty} \Im(\varphi(kh))\sin(khx)\right] \\ &\approx f_{h,N}(x) := \frac{h}{\pi}\left[\frac{1}{2} + \sum_{k=1}^{N} \Re(\varphi(kh))\cos(khx) + \sum_{k=1}^{N} \Im(\varphi(kh))\sin(khx)\right], \end{aligned}$$

where in the third line we used $g(\infty) = 0$ (given that the integral exists). Note that the difference $e_d := |f(x) - f_h(x)|$ is the *discretization error* for function f evaluated at x with mesh size h, and $e_t := |f_h(x) - f_{h,N}(x)|$ is the *truncation error* resulted from considering only the first N terms of the infinite series.

To write it in another way, we let $h = \pi/l$, and replace the notation $f_h(\cdot)$ by

$f_l(\cdot)$, and $f_{h,N}(\cdot)$ by $f_{l,N}(\cdot)$. Then

$$\begin{aligned} f(x) \approx f_l(x) &= \frac{1}{2l} + \sum_{k=1}^{\infty} \frac{1}{l} \Re\left(\varphi\left(\frac{k\pi}{l}\right)\right) \cos\frac{k\pi x}{l} + \sum_{k=1}^{\infty} \frac{1}{l} \Im\left(\varphi\left(\frac{k\pi}{l}\right)\right) \sin\frac{k\pi x}{l} \\ &\approx f_{l,N}(x) = \frac{1}{2l} + \sum_{k=1}^{N} \frac{1}{l} \Re\left(\varphi\left(\frac{k\pi}{l}\right)\right) \cos\frac{k\pi x}{l} + \sum_{k=1}^{N} \frac{1}{l} \Im\left(\varphi\left(\frac{k\pi}{l}\right)\right) \sin\frac{k\pi x}{l}. \end{aligned} \tag{5.15}$$

The last approximation $f_{l,N}(x)$ is a trigonometric polynomial of degree N and period $2l$. The argument x appears only inside the sine and cosine functions, so $\Re(\varphi(\cdot))$ and $\Im(\varphi(\cdot))$ only need to be evaluated N times, regardless of the number of points at which we want the function f to be evaluated. Appropriate values of l and N need to be chosen for the specific problem.

More details on this method of obtaining probability distributions from their transforms, including the analysis on the discretization error and the truncation error, can be found in Abate and Whitt (1992).

5.2 Estimation of the Jump Size Distribution

Now we apply the methods presented in Section 5.1 to estimate the density function $f_X(\cdot)$ of the jump size X.

Equations (5.1) and (5.2) provide an estimator, $\hat{\varphi}_X(t)$, of the characteristic function of the jump size X. However, we need to regularize it before we apply the Fourier inversion, using the method described in Section 5.1.1 for the deconvolution problem. The only difference is that $\hat{\varphi}_X(t)$ in (5.1) is a little more complicated in that it is not just a ratio of characteristic functions, but a transformation of this ratio. However, the transformation is a basic one, and $\lim_{t\to\infty} \varphi_X(t) = 0$, so we can still use the same regularization procedure as for the deconvolution problem.

Among the two methods discussed Section 5.1.1, in our implementation we chose the truncation method given by (5.11).

To use this method, we need to choose the truncation point, t_{trunc}. From Remark 2.2.23 we see that it is better to let t_{trunc} equal to the value of t selected for the estimation of jump frequency. For finite samples, we would use the recommended Method 3 for selection of t, as proposed in Section 2.3.

Let T_n be the selected value of t. Substituting t_{trunc} in (5.11) by T_n, we obtain a regularized version of the estimator of $\varphi_X(t)$ as

$$\hat{\varphi}_{X,3}(t) = \begin{cases} \hat{\varphi}_X(t) & \text{if } |t| \leq T_n \\ 0 & \text{if } |t| > T_n, \end{cases} \tag{5.16}$$

where $\hat{\varphi}_X(t)$ is given in equations (5.1) and (5.2) with $\hat{\alpha} = \hat{\alpha}_n \equiv \hat{\alpha}(T_n)$, and $\hat{\alpha}(t)$ is given in (2.24).

After the regularization step, the Fourier inverse transform is well defined and the estimator of the jump size density f_X is given by

$$\hat{f}_{X,3}(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \hat{\varphi}_{X,3}(t) dt \tag{5.17}$$

where $\hat{\varphi}_{X,3}(t)$ is given in (5.16).

Strong Consistency of the Density Estimator

In the following we show that the proposed density estimator $\hat{f}_{X,3}(x)$ in (5.17) is a consistent estimator of the true jump size density f_X.

Firstly, we introduce a function, called the *imaginary error function*, which is defined by

$$\operatorname{erfi}(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{t^2} dt.$$

This function is defined for all complex arguments x, but in the following we only need the real arguments x.

Theorem 5.2.1. *Consider Model 2 in (2.11) with the assumption (A2-1). Assume that μ, σ and α are known. Suppose the truncating point T_n in (5.16) satisfies that*

(A3-1) *erfi$(\sigma T_n/\sqrt{2})/\sqrt{n} \to 0$, as $n \to \infty$.*

Then the estimator $\hat{f}_{X,3}(x)$ defined by (5.17) satisfies, as $n \to \infty$,

$$\hat{f}_{X,3}(x) \longrightarrow f_X(x)$$

weakly for every $x \in \mathbb{R}$ at which $f(x)$ is continuous.

Proof. Fix an arbitrary x at which $f(x)$ is continuous. We have

$$\begin{aligned}
\hat{f}_{X,3}(x) - f_X(x) &= \frac{1}{2\pi}\int_{\mathbb{R}} e^{-itx}\hat{\varphi}_{X,3}(t)dt - \frac{1}{2\pi}\int_{\mathbb{R}} e^{-itx}\varphi_X(t)dt \\
&= \frac{1}{2\pi}\int_{\mathbb{R}} e^{-itx}\left(\hat{\varphi}_{X,3}(t) - \varphi_X(t)\right)dt \\
&= \frac{1}{2\pi}\int_{|t|\le T_n} e^{-itx}\left(\hat{\varphi}_X(t) - \varphi_X(t)\right)dt \;-\; \frac{1}{2\pi}\int_{|t|> T_n} e^{-itx}\varphi_X(t)dt.
\end{aligned}$$

The second term goes to zero as $n \to \infty$, because $T_n \to \infty$ and the integral $\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi_X(t)dt = f_X(x)$ exists. For the first term,

$$\begin{aligned}
\left|\frac{1}{2\pi}\int_{|t|\le T_n} e^{-itx}\left(\hat{\varphi}_X(t) - \varphi_X(t)\right)dt\right| &\le \frac{1}{2\pi}\int_{|t|\le T_n}\left|e^{-itx}\left(\hat{\varphi}_X(t) - \varphi_X(t)\right)\right|dt \\
&= \frac{1}{2\pi}\int_{|t|\le T_n}\left|\hat{\varphi}_X(t) - \varphi_X(t)\right|dt.
\end{aligned}$$

Therefore, it suffices to show that $V_n \xrightarrow{p} 0$, or equivalently $\mathbb{P}(|V_n| \ge \epsilon) \to 0$ for any $\epsilon > 0$, where

$$V_n := \int_{|t|\le T_n}\left|\hat{\varphi}_X(t) - \varphi_X(t)\right|dt.$$

Since $V_n \ge 0$ and by Markov's inequality $\mathbb{P}(V_n \ge \epsilon) \le \frac{\mathbb{E}(V_n)}{\epsilon}$, we only need to show $E(V_n) \to 0$.

Since μ, σ and λ are assumed known, we have

$$\begin{aligned}\hat{\varphi}_X(t) - \varphi_X(t) &= \frac{\frac{\hat{\varphi}_Y(t)}{\varphi_Z(t)} - (1-\alpha)}{\alpha} - \frac{\frac{\varphi_Y(t)}{\varphi_Z(t)} - (1-\alpha)}{\alpha} \\ &= \frac{\hat{\varphi}_Y(t) - \varphi_Y(t)}{\alpha \cdot \varphi_Z(t)} \\ &= \frac{1}{\alpha}\, e^{-i\mu t + \frac{1}{2}\sigma^2 t^2} \left(\hat{\varphi}_Y(t) - \varphi_Y(t)\right).\end{aligned}$$

By Proposition D.0.9 in Appendix D, we have

$$\mathbb{E}\left[|\hat{\varphi}_Y(t) - \varphi_Y(t)|^2\right] = \mathrm{Var}(\hat{\varphi}_Y(t)) = \frac{1}{n}\left(1 - |\varphi_Y(t)|^2\right) \le \frac{1}{n}.$$

Then

$$\mathbb{E}\left[|\hat{\varphi}_X(t) - \varphi_X(t)|^2\right] = \frac{1}{\alpha^2}\, e^{\sigma^2 t^2}\mathbb{E}\left[|\hat{\varphi}_Y(t) - \varphi_Y(t)|^2\right] \le \frac{1}{\alpha^2 n}\, e^{\sigma^2 t^2}.$$

Thus, by the Cauchy-Schwartz inequality, we have

$$\begin{aligned}\mathbb{E}(V_n) &= \int_{|t|\le T_n} \mathbb{E}\left|\hat{\varphi}_X(t) - \varphi_X(t)\right| dt \\ &\le \int_{|t|\le T_n} \left\{\mathbb{E}\left[|\hat{\varphi}_X(t) - \varphi_X(t)|^2\right]\right\}^{\frac{1}{2}} dt \\ &\le \frac{1}{\alpha\sqrt{n}} \int_{|t|\le T_n} e^{\frac{1}{2}\sigma^2 t^2} dt \\ &= \frac{2\sqrt{2}}{\sigma\alpha} \cdot \frac{1}{\sqrt{n}} \int_0^{\frac{\sigma T_n}{\sqrt{2}}} e^{x^2} dx,\end{aligned}$$

or equivalently,

$$\mathbb{E}(V_n) \le \frac{\sqrt{2\pi}}{\sigma\alpha} \cdot \frac{1}{\sqrt{n}}\, \mathrm{erfi}\left(\frac{\sigma T_n}{\sqrt{2}}\right).$$

By the assumption (A3-1), we obtain $\mathbb{E}(V_n) \to 0$ as $n \to \infty$. Therefore, as $n \to \infty$, $\hat{f}_{X,3}(x) \longrightarrow f_X(x)$ weakly for every x at which $f(x)$ is continuous. $\boxtimes$

Remark 5.2.2. The assumption (A3-1) in the above Theorem 5.2.1 does not depend on the distribution of the jump size. One example of T_n which satisfies this assumption is to let $T_n = \sqrt{\log n^q}$ where $0 < q < 1/\sigma^2$. This can be shown by using the loose bound that $\int_0^x e^{t^2} dt \le x e^{x^2}$ for any $x > 0$.

In the following we discuss the evaluation of the integral in the estimator $\hat{f}_{X,3}(x)$.

Using the Trapezoidal Rule

As discussed in Section 5.1.2, we apply the trapezoidal rule (5.15) to evaluate the integral in the estimator $\hat{f}_{X,3}(x)$ given by (5.17). Then a practical estimator of the density function f_X of jump size is given by

$$\hat{f}_{X,4}(x) = \frac{1}{2l} + \sum_{k=1}^{N} \frac{1}{l} \Re\left(\hat{\varphi}_{X,3}\left(\frac{k\pi}{l}\right)\right) \cos\left(\frac{k\pi x}{l}\right) + \sum_{k=1}^{N} \frac{1}{l} \Im\left(\hat{\varphi}_{X,3}\left(\frac{k\pi}{l}\right)\right) \sin\left(\frac{k\pi x}{l}\right), \tag{5.18}$$

where $\hat{\varphi}_{X,3}$ is given by (5.16). Note that this is a trigonometric polynomial of degree N and period $2l$. In the following we discuss the selection of the parameters N and l.

Selection of the Truncation and Discretization Parameters

Since there are both positive and negative jumps, we may assume that the jump size density $f_X(\cdot)$ is "concentrated" on the interval $[-l_X, l_X]$ for some constant $l_X > 0$ and is negligibly small outside this interval. That is, we are interested in the estimator $\hat{f}_{X,4}(x)$ over the interval $x \in [-l_X, l_X]$. The selection of l_X depends on the size of the increments of the process. For example, we may choose l_X to be a value approximately 1.5 to 2 times of the largest size of the observed increments. Notice that $[-l_X, l_X]$ is a symmetric interval around zero. We may also assume that $f_X(\cdot)$ is concentrated on an asymmetric interval $[-a, b]$ for some $a, b > 0$. Similarly, we can choose a and b approximately equal to 1.5 to 2 times of the largest sizes of the negative and the positive increments, respectively. Then the following steps would be the same as using the symmetric interval (except changing the requirement $l \geq l_X$ stated below by $2l \geq (b + a)$). Therefore, in the following we focus on using the symmetric interval $[-l_X, l_X]$ only.

For the selection of the discretization parameter l and the truncation parameter N, the situation is a little different from the one in the "conventional" Fourier inversion. It is known that in the conventional Fourier inversion, for a fixed l, the larger value of N the better. However, in our problem, the estimated c.f. $\hat{\varphi}_{X,3}$ in (5.16) is truncated. Note that in (5.16) we used the notation T_n for the truncation point. However, to be consistent with the notation in Section 2.3 for the selection of t for finite samples, we change to use the notation t_M when we discuss the implementation issues in the following. That is, $\hat{\varphi}_{X,3}(t) = 0$ for $|t| > t_M$. Then from $\hat{f}_{X,4}(x)$ in (5.18) we can see that N and l should satisfy

$$\frac{\pi N}{l} \leq t_M,$$

because otherwise, for a fixed l, larger N will contribute nothing to $\hat{f}_{X,4}(x)$. To ensure that we use all the information up to t_M, we would rather let

$$\frac{\pi N}{l} = t_M,$$

or equivalently,

$$l = \frac{\pi \cdot N}{t_M}. \tag{5.19}$$

Thus, the ratio N/l is fixed for a given t_M. Now we only need to choose N, since l will then be determined automatically by (5.19). For a fixed value of $x \in [-l_X, l_X]$, a larger value of N means a finer partition of the interval $[0, t_M]$ when using $\hat{f}_{X,4}(x)$ in (5.18) to evaluate the integral in $\hat{f}_{X,3}(x)$ given by (5.17), and thus $\hat{f}_{X,4}(x)$ will approximate $\hat{f}_{X,3}(x)$ better when N is larger. However, note that $\hat{f}_{X,3}(x)$ itself is an estimator of the true density f_X and is random, so a larger value of N might not guarantee a better estimation result of the true density f_X, but instead the result also depends on the properties of the estimator $\hat{f}_{X,3}(x)$.

To clarify our notation, $[-l_X, l_X]$ is the interval over which we want to estimate the density of jump size, and l is the discretization parameter in (5.18). It is required that $l \geq l_X$ so that we can obtain an appropriate density estimator (noticing that

$\hat{f}_{X,4}(x)$ is a periodic function of period $2l$). We can use this requirement to define the smallest number for N that we would consider. That is, by (5.19) and $l \geq l_X$, we have

$$N \geq \frac{t_M \cdot l_X}{\pi}. \tag{5.20}$$

Note that this defines the starting point of N that we would consider, and then we would definitely try larger values of N to check the improvement of the results. Therefore, the procedure of selecting N and l can be summarized as follows:

Step 1 Determine l_X by the size of the observed increments of the process. (For example, let l_X be a value approximately 1.5 to 2 times of the largest size of the observed increments.)

Step 2 Start by letting $N = \frac{t_M \cdot l_X}{\pi}$ (approximately), due to (5.20).

Step 3 The value of l is always determined by (5.19) after setting a value for N. Then proceed with the estimation of jump size distribution, and record the results.

Step 4 Increase N.

Step 5 Repeat Steps 3 and 4, until the estimation results of jump size distribution do not show significant improvement. Then the value of N will be regarded as an appropriate one.

In summary, a practical estimator of the density function f_X of jump size is given by $\hat{f}_{X,4}(x)$ in equation (5.18), where the selection of parameters N and l is discussed in the above procedure.

5.3 Simulation Studies

In this section, we estimate the density function of the jump size in the Merton model and the Kou model using the estimator $\hat{f}_{X,4}(x)$ in (5.18). The robust procedure "HT2" as described in Section 4.1.3 is used to estimate the diffusion parameters. For the selection of t, we apply Method 3 described in Section 2.3, and denote it by t_{M3} as before. Note that this t_{M3} is used in both the estimation of jump frequency and the construction of the regularized estimator $\hat{\varphi}_{X,3}(t)$ in (5.16).

Example 5.3.1 (Merton Model)**.** We use the same parameter settings as in Example 2.1.2, except for changing σ_X from $\sigma_X = 3.79\%$ to $\sigma_X = 10\%$*. For the selection of t using Method 3, the same upper bound $t_{\max} = 100$ is set. We discuss the selection of some parameters as follows:

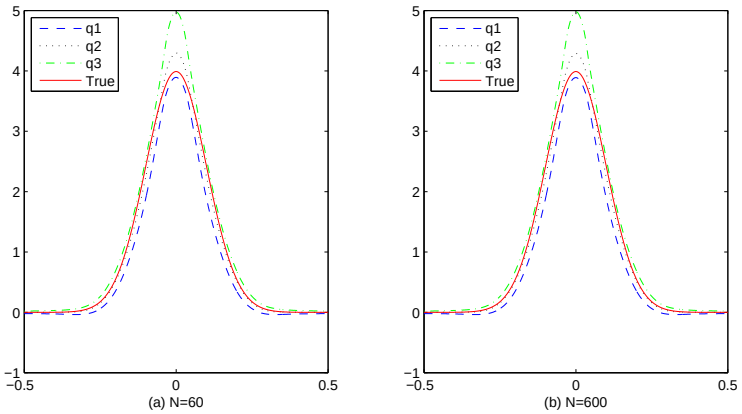
- By looking into the size of the simulated increments, we choose $l_X = 0.5$, which is as large as 5 times the standard deviation of the jump size. That is, we are interested in the estimation of the jump size density over the interval $[-0.5, 0.5]$.
- For the selection of N, we would start by $N = \frac{t_M \cdot l_X}{\pi}$. However, t_M is sample dependent. But $t_M \leq t_{\max}$, so we will just start N by a value close to $\frac{t_{\max} \cdot l_X}{\pi} = \frac{100 \times 0.5}{\pi} \approx 16$, and then increase N.
- Note that l is always determined by (5.19) for every given value of N. Since t_M is sample dependent, l is also sample dependent.

We use $\hat{f}_{X,4}(x)$ to obtain the density estimate of jump size, and repeat 100,000 times. Figure 5.2 shows the *quantile curves* of the density estimates when $N = 60$ in part (a) and $N = 600$ in part (b), where he curves "q1", "q2" and "q3" correspond

*The reason of doing this is to avoid the "confounding" effects of $t_{\max}$ and other factors. See Remark 2.3.8.

to the 25% quantile, the median, and the 75% quantile of the 100,000 density estimates of jumps size. We can see that the difference in the performance of the density estimates between using $N = 60$ and using $N = 600$ is not obvious. Note that this is also true if we simulate only sample.

Figure 5.2: Quantile curves of density estimates for different N [Merton]



The quantile curves provide visual explanations on the performance of the density estimator. However, to quantify the performance, we need to introduce some distance between the density functions. One popular choice is the *total variation (TV) distance*, defined by

$$d(f, g) = \frac{1}{2} \int_{\mathbb{R}} |f(x) - g(x)| dx$$

for any density functions f and g. For our problem, we want to measure the distance between the true density f_X and the estimator $\hat{f}_{X,4}$ over the interval $[-l_X, l_X]$, i.e.

$$d(f_X, \hat{f}_{X,4}) = \frac{1}{2} \int_{-l_X}^{l_X} |\hat{f}_{X,4}(x) - f_X(x)| dx. \tag{5.21}$$

Table 5.1 provides the total variation distance $d(f_X, \hat{f}_{X,4})$ for different values of $N = 20,\ 60,\ 100,\ 200,\ 300,\ 600$. Note that the estimation error includes the

Table 5.1: TV distance between f_X and $\hat{f}_{X,4}$ for different N [Merton]

	$N = 20$	$N = 60$	$N = 100$	$N = 200$	$N = 300$	$N = 600$
$\mathrm{d}(f,\, \hat{f}_{X,4})$	.1077 (.0605)	.1072 (.0600)	.1071 (.0601)	.1069 (.0600)	.1072 (.0602)	.1067 (.0600)

Note: $\sigma_X = 0.1$, $n = 2500, \lambda = 12.5$, $t_{\max} = 100$, $l_X = 0.5$. Repeat 100,000 times. The value of l is determined by (5.19) for each value of N, and is sample dependent because t_M is sample dependent.

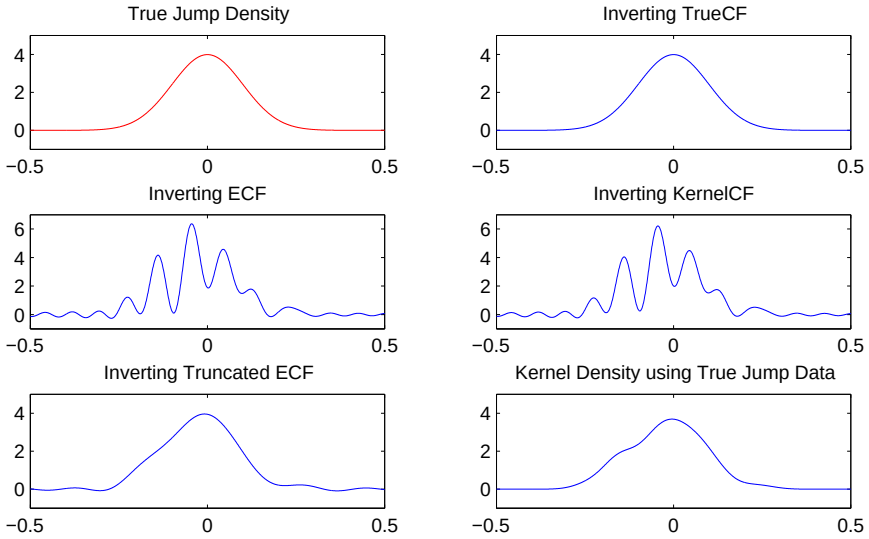
error from every step: the estimation of the diffusion parameters, the estimation of the jump ratio, and the estimation of the jump size distribution. To evaluate the total variation distance (5.21), we partition the interval $[-l_X, l_X]$ into 2,000 equally spaced sub-intervals.[†] From the result we see that the total variation distance decreases in the beginning as N increases. However, the improvement becomes less obvious (or there is no improvement) when N gets large. So in the following we choose $N = 200$.

For a quick comparison, Figure 5.3 shows the estimates of the jump size density using different methods (described below), based on one simulated data set. There are six graphs in this figure, and we explain each of them as follows:

- The graph on the first line and first row, denoted by (1,1), is the true density function of jump size.
- The graph (1,2) shows the density estimate obtained by inverting numerically the true characteristic function $\varphi_X(\cdot)$. Since it is very close to the true density, this verifies the appropriateness of using the trapezoidal rule for the Fourier inversion, i.e. equation (5.15).
- The graph (2,1) shows the density estimate obtained by using the e.c.f., i.e. by inverting $\hat{\varphi}_X(t)$ given in equations (5.1) and (5.2). As expected, the result

[†]In the following simulation studies, we always do this to evaluate the total variation distance.

Figure 5.3: Different density estimates of jump size [Merton]



is not good.

- The graph (2,2) shows the density estimate obtained by using the kernel estimate of the c.f. with the Gaussian kernel, i.e. equation (2.53). The result is similar to the one using e.c.f., because the window width h_n given by (2.52) in the kernel estimate of the c.f. satisfies $h_n < \sigma$ and thus the shape of the kernel estimate is still the same as the shape of the estimate by using the e.c.f.. This also verifies that the "conventional" kernel estimator of the c.f. (e.g. with the Gaussian kernel) does not work well in general de-convolution problems.

- The graph (3,1) is based on the method that we recommend, i.e. using the truncated e.c.f. given by (5.16). The result looks significantly better.

- The graph (3,2) shows the kernel density estimate based on the true jump observations. In practice it is impossible to separate all the jump observations

from the diffusion observations. Therefore, the fact that the results in the last two graphs are comparable shows that our method is working well.

Note that graphs (2,1) and (2,2) are obtained by using the same l as in graph (3,1) but with N increased to twice the size used in graph (3,1). In these cases, we have found that the estimation error increases when N becomes larger. The graph (1,2) are obtained by using the same l and N as in graph (3,1).

For further interest, Figure 5.4 shows the real and imaginary parts of different estimates of the c.f. of jump size, based on one simulated data set. This can be used to explain why our method (using the truncated e.c.f.) is better than the others: from the figure we see that the real part of the truncated e.c.f. is the one closest to the real part of the true c.f.. The real part of the truncated e.c.f. is continuous and equals zero at the point t_M (i.e. the selected value of t) by equations (5.1) and (4.15), and stays at zero for $t > t_M$; while the true c.f. starts to approach zero from the point t_M. For the imaginary parts, the truncated c.f. is also the closest one to the true c.f.

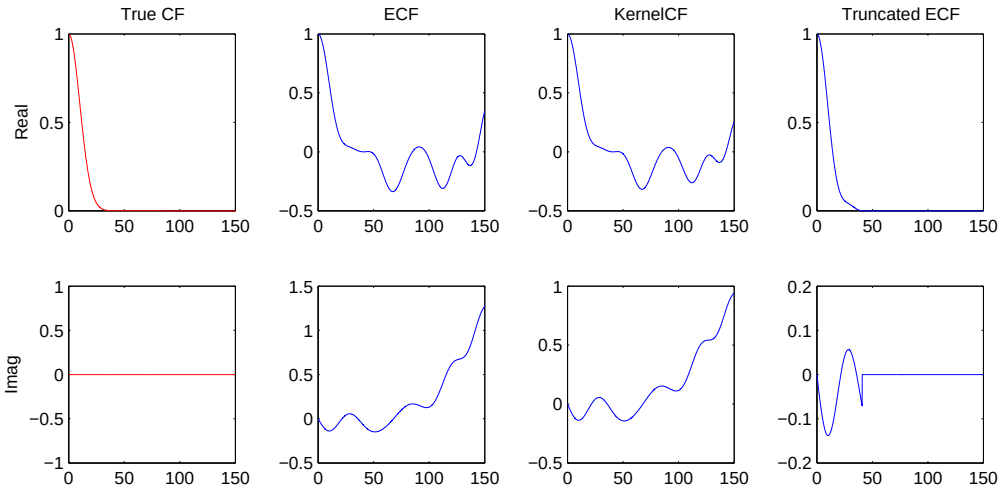
Table 5.2: TV distance between f and $\hat{f}_{X,4}$ for different scenarios [Merton]

	$\sigma_X \approx 3.79\%$		$\sigma_X = 10\%$		$\sigma_X = 60\%$	
n	$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$
1000	.1467(.0653)	.1058(.0409)	.1473(.0750)	.0971(.0436)	.1209(.0501)	.0894(.0365)
2500	.0987(.0408)	.0811(.0280)	.1068(.0600)	.0646(.0290)	.0815(.0334)	.0598(.0242)
10000	.0545(.0214)	.0673(.0161)	.0724(.0484)	.0340(.0147)	.0441(.0179)	.0319(.0128)
40000	.0321(.0120)	.0640(.0084)	.0598(.0433)	.0172(.0069)	.0237(.0096)	.0169(.0067)

Note: $N = 200$, $l_X = 5 \cdot \sigma_X$, $t_{\max} = 100$. Repeat 100,000 times.

Now we look at the estimation results under different scenarios of parameter settings. Table 5.2 shows the total variation distance between f_X and $\hat{f}_{X,4}$ for different values of σ_X, jump intensity λ and sample size n. In the beginning, we use the same analysis as in Table 5.1 to choose an appropriate value of N for every σ_X, and it turns out that $N = 200$ is appropriate for all cases. For the selection

Figure 5.4: Real and imaginary parts of density estimates of jump size [Merton]



of t using Method 3, the upper bound $t_{\max} = 100$ is always set. For the Fourier inversion, we always set $l_X = 5 \cdot \sigma_X$. Repeat 100,000 times. From the results we can see that:

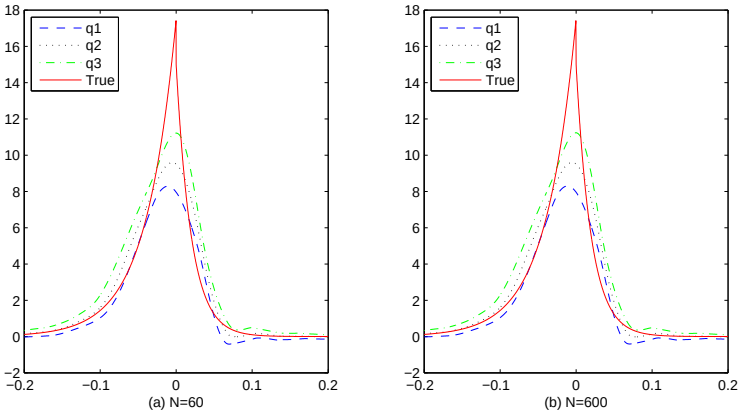
- As n increases (while fixing the other two parameters: σ_X and λ), the TV distance decreases.
- As λ increases, the TV distance decreases. (There are exceptions when $\sigma_X \approx 3.79\%$, $n = 10,000$ and when $\sigma_X \approx 3.79\%$, $n = 40,000$. This might be due to the different degrees of accuracy of the volatility estimator, or the different degrees of influence of $t_{\max}$ on the selection of t, for different scenarios.)
- As σ_X increases, the TV distance decreases. (There is one exception when comparing the first column and the third column. This might be due to the same reasons as given above.)

Example 5.3.2 (Kou Model)**.** We use the same parameters as in Example 2.1.3 for the Kou model. In the following we perform a similar analysis as we did for the Merton model.

By looking at the sizes of the simulated increments, we choose $l_X = 0.2$. For the selection of N, we start N by a value close to $\frac{t_{\max} \cdot l_X}{\pi} = \frac{100 \times 0.2}{\pi} \approx 6$, and then increase N. The value of l is determined by (5.19) for every given N.

Figure 5.5 shows the *quantile curves* of the density estimates when $N = 60$ in part (a) and $N = 600$ in part (b). Similarly to the Merton case, the difference between the performance of the density estimates for $N = 60$ and for $N = 600$ is not obvious.

Figure 5.5: Quantile curves of density estimates for different N [Kou]



One observation from Figure 5.5 is that the middle part of the density estimate is over-smoothed. (Note that the true density is not continuous at zero.) This problem is not caused by the Fourier inversion where the trapezoidal rule is used; see graph (1,2) in Figure 5.6 for a counter-example. Rather it is a result of the fact that the c.f. of the double exponential distribution in the Kou model decays slowly comparing with the normal distribution in Merton model does. When we truncate the e.c.f. at t_{M3}, the true c.f. of the double exponential distribution is not close

enough to zero yet, so we lose the information of the c.f. on the interval $t > t_{M3}$. However, we can not truncate the e.c.f. at a larger value of t because otherwise we will increase the error – this can be verified by the graph (2,1) in Figures 5.6.

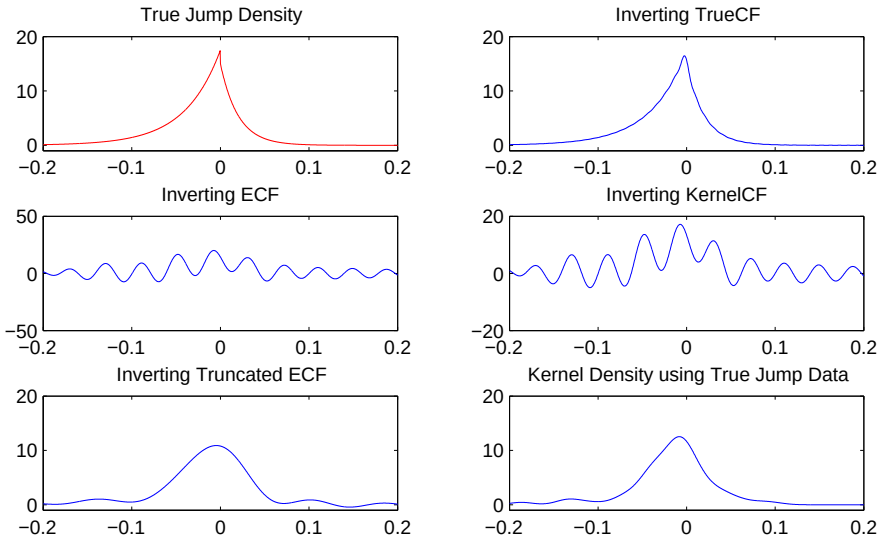
Table 5.3 provides the total variation distance $d(f_X, \hat{f}_{X,4})$ for different values of N. Based on these results, we also choose $N = 200$ for further studies.

Table 5.3: TV distance between f_X and $\hat{f}_{X,4}$ for different N [Kou]

	$N = 10$	$N = 30$	$N = 100$	$N = 200$	$N = 300$	$N = 600$
$\mathrm{d}(f,\ \hat{f}_{X,4})$	.1985 (.0490)	.1916 (.0468)	.1904 (.0468)	.1900 (.0469)	.1897 (.0465)	.1897 (.0469)

Note: $\eta_+ = 1/0.02, \eta_- = 1/0.04, p = 0.3$, $l_X = 0.2$, $n = 2500, \lambda = 12.5$. Repeat 100,000 times. $t_{\max} = 100$.

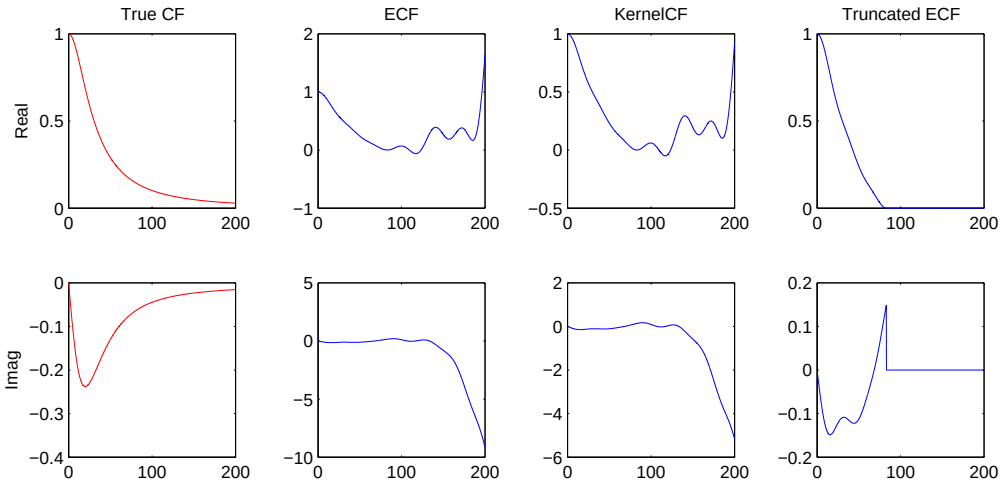
Figure 5.6: Different density estimates of jump size [Kou]



For a quick comparison, Figure 5.6 shows different estimates of the density function of jump size, where one sample is used, similar to what we did for the

Merton model. For this sample, we have $t_{M3} = 83.0$. In graph (3,1), $N = 200$ and $l = N * \pi / t_{M3} = 7.57$. Similarly to the Merton case, the estimates in graphs (2,1) and (2,2) are not good, while the estimate in graph (3,1) is comparable with the kernel density estimate as in graph (3,2). Note that the graphs (2,1) and (2,2) are obtained by using the same l as in graph (3,1) but we have increased N to twice of the size used in graph (3,1). The graph (1,2) is obtained by using the same l as in graph (3,1) but with N increased to 1500.[‡] Such a large N is required to obtain a good density approximation, as in graph (1,2), by inverting the true c.f. of jump size distribution, because the c.f. of the double exponential distribution decays slowly, which can also be verified by comparing the graph (1,1) in Figure 5.7 below for the Kou case with the graph (1,1) in Figure 5.4 for the Merton case.

Figure 5.7: Real and imaginary parts of density estimates of jump size [Kou]



For further interest, Figure 5.7 shows the real and imaginary parts of different estimates of the c.f. of jump size where one sample is used. We see that the real part

[‡]Note that in the Merton case, we used the same N for graphs (1,2) and (3,1).

of the truncated e.c.f. is the one closest to the real part of the true c.f., although it is not as close as for the Merton case.

Table 5.4: TV distance between f and $\hat{f}_{X,4}$ for different scenarios [Kou]

	$\eta_+ = 1/0.02$, $\eta_- = 1/0.04$ $l_X = 0.2$		$\eta_+ = 1/0.1$, $\eta_- = 1/0.2$ $l_X = 1.0$	
d(f, $\hat{f}_{X,4}$)	$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$
$n = 1000$	.2316 (.0735)	.2057 (.0509)	.1804 (.0547)	.1415 (.0398)
$n = 2500$	.1898 (.0467)	.1881 (.0350)	.1312 (.0362)	.1043 (.0275)
$n = 10000$	.1641 (.0291)	.1833 (.0191)	.0820 (.0205)	.0661 (.0165)
$n = 40000$	.1584 (.0193)	.1840 (.0103)	.0533 (.0153)	.0423 (.0101)

Note: $p = 0.3$ is fixed. $N = 200$, $t_{\max} = 100$. Repeat 100,000 times.

Table 5.4 shows the estimation results under different scenarios of parameter settings. Note that $p = 0.3$ is fixed for the different jump size distributions. The situation is similar to the Merton model case. The TV distance decreases as n increases, or λ increases, or the variance of jump size increases.

Table 5.5: TV distance between f and $\hat{f}_{X,4}$ for different scenarios [t distribution]

	$\sigma_X \approx 3.79\%$		$\sigma_X = 10\%$		$\sigma_X = 60\%$	
n	$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$
1000	.2470(.1933)	.2218(.0719)	.1848(.0761)	.1303(.0518)	.1597(.0664)	.1165(.0453)
2500	.1818(.0694)	.1938(.0487)	.1279(.0533)	.0902(.0364)	.1060(.0404)	.0773(.0246)
10000	.1298(.0381)	.1782(.0269)	.0754(.0349)	.0554(.0239)	.0572(.0166)	.0418(.0118)
40000	.1037(.0233)	.1743(.0156)	.0491(.0272)	.0413(.0133)	.0309(.0086)	.0224(.0061)

Note: $N = 200$, $l_X = 5 \cdot \sigma_X$, $t_{\max} = 100$. Repeat 100,000 times.

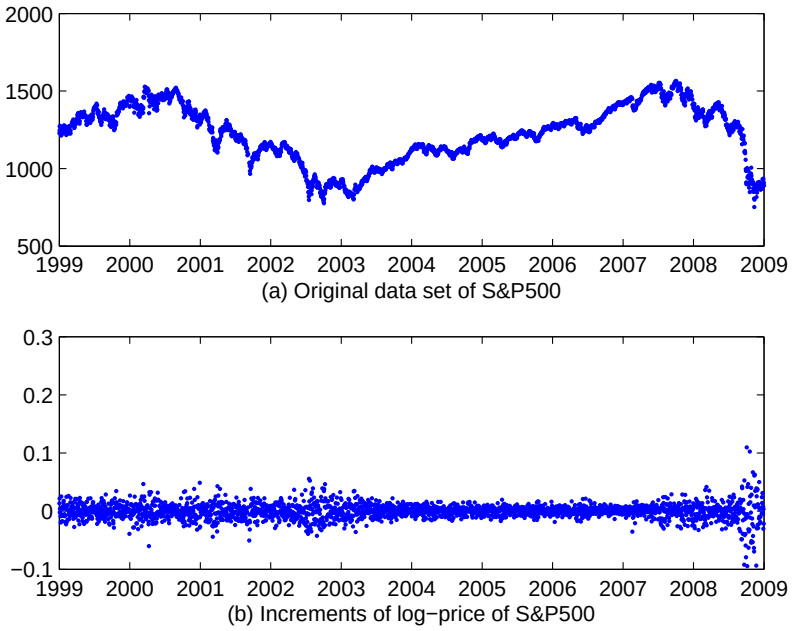
Example 5.3.3 (t distribution)**.** We follow the same procedure as in the Merton model example. Table 5.5 shows the total variation distance between f_X and $\hat{f}_{X,4}$ for different values of σ_X, jump intensity λ and sample size n. From the results,

we can see that the same pattern exists as the one in the Merton model example, although the results are not as good as the one for Merton model.

5.4 Real Data Example

In this section we work with real data. We use the same data set as in Sepp (2012), which is the adjusted closing values[§] of the S&P500 index from January 4, 1999 to January 9, 2009, with a total of 2521 observations. Taking log of the data set, we work with the increments of the log-price. Figure 5.8 shows the scatter plots of the original S&P500 data set in part (a) and the increments of the log-price of S&P500 in part (b).

Figure 5.8: Scatter plots for S&P daily data

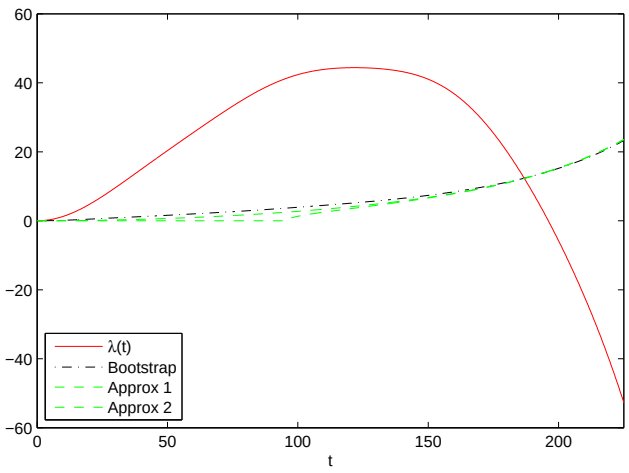


[§]close price adjusted for dividends and splits (as in Yahoo Finance web)x.

In the following we apply our method to this data set, and then compare the obtained results with the results presented in Sepp (2012). In Sepp (2012) the volatility was treated as constant, and we do the same.

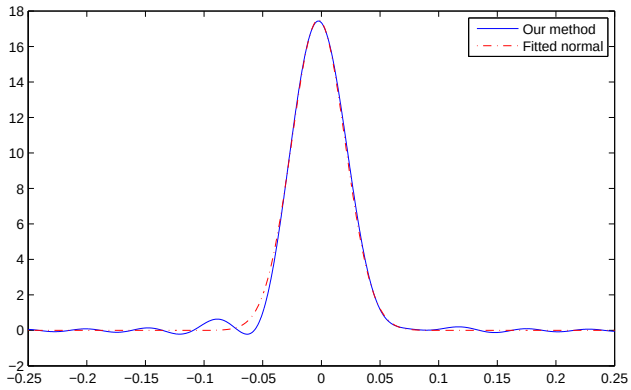
The robust estimates of the diffusion parameters are $\hat{\mu}_0 = 0.0341$ and $\hat{\sigma}_0 = 0.1378$. Figure 5.9 shows the curve of $\hat{\lambda}(t)$, the bootstrapped estimate of $\mathrm{Std}(\hat{\lambda}(t))$, and two "approximations" of $\mathrm{Std}(\hat{\lambda}(t))$ based on (2.40) and (2.41). For the selection of t, Method 3 is used. We set the upper bound $t_{\max} = 200$, with the same reason as given in Example 2.3.11. Then we found that $t_{M3} = 121.6$, $\hat{\lambda}(t_{M3}) = 44.3906$, or equivalently, $\hat{\alpha}(t_{M3}) = 0.1776$. Later we will compare these results with the ones presented in Sepp (2012).

Figure 5.9: The curve of $\hat{\lambda}(t)$ for S&P500 daily data



Now let us look at the estimation of the jump size distribution. Figure 5.10 shows the estimate of the jump size density, based on the method proposed in Section 5.2. Note that we used $l_X = 0.25$, $N = 300$, and $l = N * \pi / t_{M3} = 7.75$. As shown in the figure, this density estimate can be well "fitted" by a normal distribution with $\hat{\mu}_X = -0.0025$ and $\hat{\sigma}_X = 0.0228$. Thus, our estimated model is not far from the Merton model.

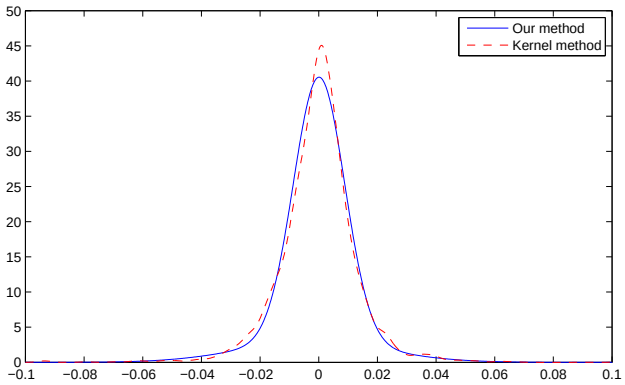
Figure 5.10: Density estimate of jump size [S&P500]



Now we check the performance of the above estimation method. On one hand, we can apply kernel density estimation to the 2520 observations of the increments (of the log-price of S&P500) to obtain a density estimate of the increments. This corresponds to "kernel method" in Figure 5.11. On the other hand, we may plug-in the above estimates $\hat{\mu}_0, \hat{\sigma}_0, \hat{\lambda}$ and the fitted normal distribution for jump size to the Merton model and calculate the density of the increments of the Merton model (or use the mixture model to approximate it). This corresponds to the curve labelled as "our method" in Figure 5.11. Comparing the two density estimates of the increments, we see that our method does a good job, except for the middle part which shows underestimation (relative to the "kernel method"). This may be because the Fourier inversion by the truncation method did not capture the high peak of the middle part of the jump size density, but tends to have a smoothed peak which looks like a normal distribution (see the above Kou model example).

In Table 5.6, we put together our results with the results from Sepp (2012). In Sepp (2012), the jump size was assumed to follow a mixture of normal distributions, although the number of components is to be determined. After fitting to the data, the jump size in the Sepp model follows a mixture of four normal distributions (as

Figure 5.11: Density estimate of increments [S&P500]



introduced in Example 2.3.11). The standard deviations of the four normal distributions are the same, i.e. $\sigma_X = 0.0127$, and the mean of this mixture distribution is found to be $-1.9 * 10^{-4}$. Comparing our results with the results from Sepp (2012), we can see the following:

- The estimates of λ are very close.
- The estimates of the volatility parameter σ_0 in the diffusion are also close.
- Due to different model settings (see the footnote for equation (2.1)), the drift term μ_0 in our model corresponds to $\tilde{\mu}_0 - \sigma_0^2/2$, where $\tilde{\mu}_0$ is the drift term in Sepp (2012). It was assumed $\tilde{\mu}_0 = 0$ in Sepp (2012). Thus, we wrote $\hat{\mu}_0 = 0 - \hat{\sigma}_0^2/2 = -0.0091$ for Sepp (2012).
- The estimates of the jump size density are different – actually it is difficult to compare them, because the distribution of the jump size is assumed a parametric form in the Sepp model, but it is assumed completely unknown in our model.

The obtained results based on our method may be used as prior information for other estimation methods, for example, the ones with parametric model settings.

Table 5.6: Comparison of the estimation results

	Our Method	Sepp (2012)
$\hat{\mu}_0$	0.0341	-0.0091
$\hat{\sigma}_0$	0.1378	0.1348
t_{M3}	121.6	N/A
$\hat{\lambda}(t_{M3})$	44.3906	46.4444
$\hat{\mu}_X$	$-2.5 * 10^{-3}$	$-1.9 * 10^{-4}$
$\hat{\sigma}_X$	0.0228	0.0127

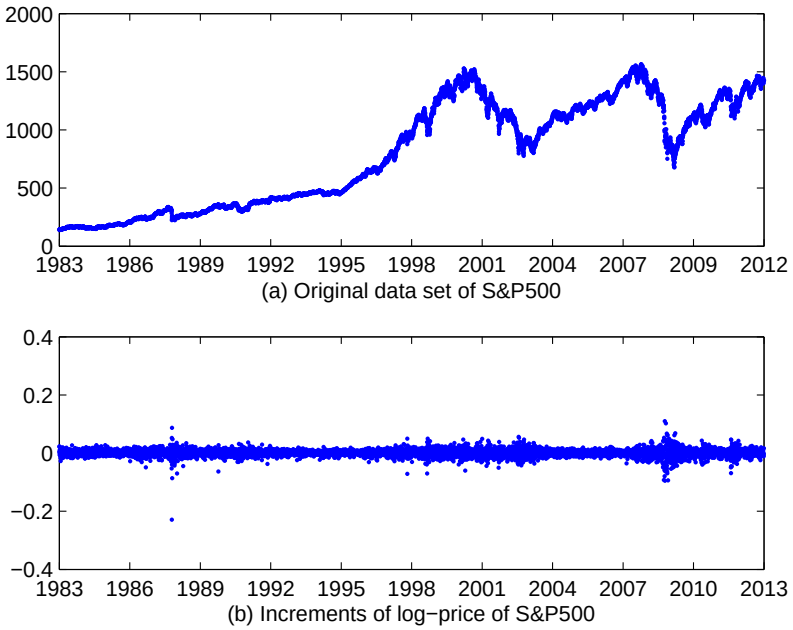
Now let us consider a longer period of S&P 500 daily data in the past three decades from 1983 to 2012, and see some trend in the jump component. Figure 5.12 shows the plot of the original data set in part (a) and the plot of the increments of the log-price in part (b).

Table 5.7: Estimation results for S&P500 daily data

Period	1983-1992	1993-2002	2003-2012	1993-2012	1983-2012
# Year	10	10	10	20	30
$\hat{\mu}_0$	0.1413	0.1439	0.1724	0.1585	0.1552
$\hat{\sigma}_0$	0.1098	0.1188	0.1173	0.1181	0.1144
t_{M3}	157.5	141.1	139.2	139.9	147.7
$\hat{\lambda}(t_{M3})$	36.2071	50.7978	51.1623	51.0299	47.1194
$\hat{\mu}_X$	0.0006	-0.0030	-0.0028	-0.0025	-0.0021
$\hat{\sigma}_X$	0.0162	0.0185	0.0219	0.0200	0.0190

Table 5.7 shows the estimation results for different periods in the 30 years. From

Figure 5.12: Scatter plots for S&P500 daily data [30 years]



the results of $\hat{\lambda}$ and $\hat{\sigma}_X$, we see that the jump rate increases during the past three decades, and the average jump size increases as well.

To summarize, in this chapter we have proposed an estimator of the density function of jump size, and proved its consistency. The results suggest that the proposed estimation method performs slightly better for the Merton model than for the Kou model, which can be explained by the fact that the c.f. of normal distribution converges to zero faster than the c.f. of double exponential distribution does. For the daily data, when the sample size is $n = 2500$ (i.e. time horizon $T = 10$) and the jump intensity is $\lambda = 12.5$, there are 125 jump observations on average in each sample, which is a low number even for classical nonparametric density estimation methods. Moreover, for our problem, we have to first figure out some way to (explicitly or implicitly) "disentangle" the jump observations from the diffusion

observations, which makes the problem more difficult. The presented simulation results, as well as the real data example, show that our method could provide reasonably accurate estimates. Although we can expect that the results would become better when the data frequency increases, our method has the advantage of not requiring high frequency data.

Chapter 6

Comparison with Existing Methods

In this chapter, we compare the method proposed in this thesis (hereafter referred as "our method") with some of the existing methods for the estimation of Lévy measure. In the finite activity case, the estimation of the Lévy measure is equivalent to the estimation of jump intensity and jump size distribution.

In the literature, several such methods have been proposed (see Chapter 1). However, many of them focus on the asymptotic behavior of the estimators and do not provide enough information for practical implementations. In Section 6.1, we look at the method introduced by Mancini (2004) and Gegler and Stadtmüller (2010), which is based on the idea of defining a "threshold" to disentangle the jumps observations from the diffusion observations. In Section 6.2, we compare our method with this threshold method.

6.1 Introduction

In Mancini (2004) a method was proposed to estimate the jump-diffusion models, where a "threshold" was defined to differentiate between the observations with jumps and the ones without jumps. This method was further investigated by Gegler and Stadtmüller (2010), where the implementation of the method was also discussed. Hereafter we refer to this method as the "threshold method". In the following we review this method.

Suppose L_t is a Lévy process with the characteristic triplet $(\mu_0, \sigma_0^2, \nu(dx))$. To ease the notation, we define a function

$$\nu(x) := \nu((-\infty, x]), \quad x \in \mathbb{R},$$

which is regarded as the cumulative Lévy measure. Then the notation ν could have two meanings: the Lévy measure or the cumulative Lévy measure — it should not be difficult to distinguish them in the context. When L_t is a finite-activity Lévy process, ν is equal to the c.d.f, F_X, of jump size X multiplied by the jump intensity, i.e. $\nu(\cdot) = \lambda \cdot F_X(\cdot)$.

Let Δ be the length of time step, and T be the time horizon. Then the increments of the Lévy process are denoted by

$$Y_k = L_{k\Delta} - L_{(k-1)\Delta}, \quad k = 1, 2, ..., n$$

with $n\Delta = T$. Denote by B_n the threshold value such that an increment Y_k is classified as an observation with jump if it satisfies* $|Y_k| > B_n$, and otherwise it is classified as an observation without jump. In the finite activity case, the estimators

*Although not mentioned in the paper by Gegler and Stadtmüller (2010), the following step is carried out in our implementation of the method: if the increments Y_k have a nonzero mean, we first transform the data by subtracting the sample median from each increment.

of the Lévy triplet are defined by

$$\hat{\mu_0} = \frac{1}{T}\sum_{k=1}^{n} Y_k \cdot 1\{|Y_k| \leq B_n\}$$

$$\hat{\sigma_0}^2 = \frac{1}{T}\sum_{k=1}^{n} Y_k^2 \cdot 1\{|Y_k| \leq B_n\} \tag{6.1}$$

and

$$\hat{\nu}(x) = \frac{1}{T}\sum_{k=1}^{n} 1\{-\infty < Y_k \leq x, |Y_k| > B_n\}, \tag{6.2}$$

with

$$B_n = \beta\sqrt{n^{-\alpha}(1+2\alpha)\log n}$$

for $\alpha \in (1/2, 1)$ and $\beta > \sigma_0$. It was shown by Gegler and Stadtmüller (2010) that the above estimators are consistent and asymptotically normally distributed as $T \to \infty$ and $\Delta \to 0$.

In practice, however, it is difficult to choose the threshold B_n for finite samples because the parameters α and β are unknown. For this, Gegler and Stadtmüller (2010) proposed a different method for the selection of B_n to be used in their numerical implementation. The method can be described as follows.

Each of the increments can be decomposed as

$$Y_k = {}^{C}Y_k + {}^{J}Y_k, \quad k = 1, 2, ..., n,$$

where ${}^{C}Y_k$ and ${}^{J}Y_k$ correspond to the continuous and the jump part of Y_k, respectively. Let

$$E_n = \mathbb{E}\left[\#\{k : |{}^{C}Y_k| > B_n, k = 1, 2, ..., n\}\right].$$

That is, E_n is the expected number of misclassifications that a diffusion observation is misclassified as a jump observation. Then

$$E_n = n\gamma_n$$

with

$$\gamma_n = \mathbb{P}(|{}^C Y_k| > B_n). \tag{6.3}$$

This expression for E_n can be justified by using the fact that $\#\{k : |{}^C Y_k| > B_n, k = 1, 2, ..., n\}$ follows a Binomial distribution with parameters n and γ_n.

Since ${}^C Y_k \sim \mathcal{N}(\mu_0\Delta, \sigma_0^2\Delta)$, we have $\gamma_n = 2 - \Phi\left(\frac{B_n+\mu_0\Delta}{\sigma_0\sqrt{\Delta}}\right) - \Phi\left(\frac{B_n-\mu_0\Delta}{\sigma_0\sqrt{\Delta}}\right)$. Note that for small Δ the drift term $\mu_0\Delta$ is negligible when compared with the term $\sigma_0\sqrt{\Delta}$, so we may simplify the expression of γ_n to $\gamma_n = 2 - 2\Phi\left(\frac{B_n}{\sigma_0\sqrt{\Delta}}\right)$. Solving the latter for B_n, we obtain

$$B_n = \sigma_0\sqrt{\Delta} z_{1-\gamma_n/2}, \tag{6.4}$$

where z_α is the α-quantile of a standard normal distribution.

In the paper by Gegler and Stadtmüller (2010), the threshold method was applied to the Merton model with $E_n = 1$, i.e. $\gamma_n = 1/n$. In the beginning, an iteration procedure was applied to estimate the threshold B_n and the volatility σ_0 as follows: as the first step, the sample variance was used to estimate σ_0^2, and then an estimate of B_n was obtained by (6.4). After that, a new estimate of σ_0^2 was obtained by (6.1). The procedure was repeated till it converged[†].

For any positive threshold B_n, it is not possible to estimate the Lévy measure on the interval $[-B_n, B_n]$. To solve this problem, the authors applied the extrapolation technique by introducing a cubic spline; more specifically, fitting two polynomials: p_1 on $[-B_n, 0]$ and p_2 on $[0, B_n]$, each of order three. Then the (cumulative) Lévy

[†]Although not mentioned in that paper, we set the tolerance level for the iteration to be $|\hat{\sigma}_{0,1} - \hat{\sigma}_{0,0}|/\hat{\sigma}_{0,0} < 10^{-5}$.

measure was estimated by

$$\hat{\nu}_c(x) = \begin{cases} \dfrac{1}{T}\displaystyle\sum_{k=1}^{n} 1\{-\infty < Y_k \leq x, |Y_k| > B_n\} & \text{if } x \leq -B_n \\ p_1(x) & \text{if } -B_n < x < 0 \\ p_2(x) & \text{if } 0 \leq x \leq B_n \\ p_2(B_n) + \dfrac{1}{T}\displaystyle\sum_{k=1}^{n} 1\{B_n \leq Y_k \leq x, |Y_k| > B_n\} & \text{if } x > B_n. \end{cases} \tag{6.5}$$

Note that $\hat{\nu}_c(x)$ in (6.5) is a corrected version of $\hat{\nu}(x)$ in (6.2). We let

$$\hat{\lambda}_{un} := \hat{\nu}(\mathbb{R}) \quad \text{and} \quad \hat{\lambda} := \hat{\nu}_c(\mathbb{R}), \tag{6.6}$$

which are the uncorrected and corrected estimators of the jump frequency.

To obtain the eight coefficients in the polynomials p_1 and p_2, the authors used the continuity and the differentiability (first and second orders) of the cumulative Lévy measure at the points $-B_n, 0$ and B_n, which lead to eight equations (see below). Note that to obtain the derivatives of the step function $\hat{\nu}$, they fit polynomialsf_1 and f_2 to $\hat{\nu}$ on $[-\kappa_1 B_n, -B_n]$ and $[B_n, \kappa_2 B_n]$, respectively, where κ_1 and κ_2 are chosen such that 20% of the data on the left of $-B_n$ and to the right of B_n, respectively,

are in those intervals[‡]. The eight equations are as follows:

$$\begin{cases} p_1(0) = p_2(0) \\ p_1'(0) = p_2'(0) \\ p_1''(0) = p_2''(0) \\ p_1(-B_n) = \hat{\nu}(-B_n) \\ p_1'(-B_n) = f_1'(-B_n) \\ p_1''(-B_n) = f_1''(-B_n) \\ p_2'(B_n) = f_2'(B_n) \\ p_2''(B_n) = f_2''(B_n) \end{cases}$$

However, there is no guarantee that the resulting cubic spline is always nondecreasing. Although the authors did not mention any adjustment in their paper, in our implementation we take it constant whenever it decreases.

The numerical implementation of the threshold method is given in the next section.

6.2 Comparison with the Threshold Method

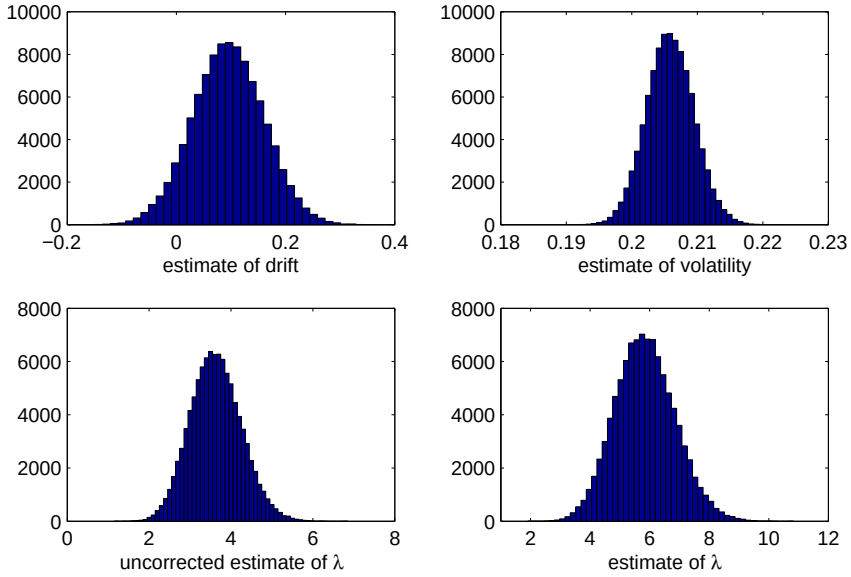
In this section we compare the threshold method with our method, firstly in the estimation of the jump frequency and then in the estimation of the jump size distribution, based on simulated data.

For the estimation of jump frequency, let us first look at the Merton model. We use the same parameters as in Example 2.1.2. Figure 6.1 shows the histograms of

[‡]In our implementation, we have slightly modified this part to include more data to fit f_1 and f_2. Also, the authors did not mention the order of the polynomials f_1 and f_2; in our implementation we used the order of three.

the estimates $\hat{\mu_0}, \hat{\sigma_0}^2$ of the diffusion parameters, the uncorrected estimate $\hat{\lambda}_{un}$ and the corrected estimate $\hat{\lambda}$ of the jump frequency, based on the threshold method. They all look normally distributed. As expected, the average of $\hat{\lambda}$ is greater than the average of the uncorrected $\hat{\lambda}_{un}$. However, since the true value of jump frequency is $\lambda = 12.5$, both $\hat{\lambda}_{un}$ and $\hat{\lambda}$ show significant underestimation. Comparing this figure with Figure 4.6 which presents the histograms based on our method, we see that our method performs much better in the estimation of jump frequency.

Figure 6.1: Histograms of $\hat{\mu_0}, \hat{\sigma_0}^2$, $\hat{\lambda}_{un}$ and corrected $\hat{\lambda}$ based on threshold method



The statistics based on the threshold method can be found in Table 6.1, which includes results for different scenarios of parameter settings for the Merton model.

Note that in our implementation we have fixed $\gamma_n = 1/1000$ (i.e. γ_n does not depend on n) instead of fixing E_n, where γ_n is defined in (6.3) and can be understood as the probability that a diffusion observation is misclassified as a jump observation. By doing this, the "threshold" level would stay at the same level when the sample size n changes, as long as the other parameters remain the same. For Table 6.1,

Table 6.1: Estimation results using threshold method [Merton]

n		$\sigma_X \approx 3.79\%$		$\sigma_X = 10\%$		$\sigma_X = 60\%$	
		$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$
1000	$\hat{\mu_0}$	.0955(.1029)	.0923(.1075)	.0937(.0993)	.0890(.1004)	.0932(.0974)	.0885(.0948)
	$\hat{\sigma_0}$	.2060(.0058)	.2153(.0073)	.1995(.0053)	.2005(.0062)	.1943(.0049)	.1897(.0051)
	$\hat{\lambda}_{un}$	3.62(0.99)	6.50(1.37)	8.68(1.48)	17.06(2.07)	12.12(1.71)	24.04(2.38)
	$\hat{\lambda}$	5.84(1.60)	10.45(2.22)	13.12(2.26)	25.83(3.14)	12.97(1.89)	25.57(2.63)
2500	$\hat{\mu_0}$	.0952(.0648)	.0920(.0682)	.0939(.0629)	.0891(.0634)	.0934(.0615)	.0881(.0600)
	$\hat{\sigma_0}$	.2059(.0036)	.2151(.0046)	.1995(.0034)	.2004(.0039)	.1943(.0031)	.1896(.0032)
	$\hat{\lambda}_{un}$	3.62(0.63)	6.53(0.87)	8.67(0.93)	17.08(1.30)	12.11(1.08)	24.03(1.50)
	$\hat{\lambda}$	5.85(1.02)	10.50(1.42)	13.12(1.43)	25.86(1.99)	12.94(1.21)	25.55(1.66)
10000	$\hat{\mu_0}$	.0953(.0325)	.0921(.0339)	.0938(.0316)	.0891(.0316)	.0933(.0308)	.0882(.0300)
	$\hat{\sigma_0}$	.2058(.0018)	.2149(.0023)	.1995(.0017)	.2003(.0019)	.1943(.0015)	.1896(.0016)
	$\hat{\lambda}_{un}$	3.63(0.32)	6.54(0.44)	8.68(0.47)	17.09(0.66)	12.11(0.54)	24.03(0.74)
	$\hat{\lambda}$	5.85(0.52)	10.51(0.72)	13.12(0.72)	25.87(1.00)	12.93(0.60)	25.54(0.83)

the average values of the threshold B_n are the same for all scenarios in the same column. The average values for the six columns are 0.0433, 0.0453, 0.0420, 0.0422, 0.0409 and 0.0400, respectively.

Comparing Table 6.1 with Table 4.6, which shows the estimation results based on our method, we can draw the following conclusions:

- Our method is significantly better than the threshold method when σ_X is small. For example, when $\sigma_X \approx 3.79\%$ (i.e. triple of daily volatility), the average of the estimates of λ based on our method is around 12.2 when $\lambda = 12.5$ and 20.4 when $\lambda = 25$, but the average based on the threshold method is around 5.8 when $\lambda = 12.5$ and 10.5 when $\lambda = 25$.

- When σ_X becomes larger, both the threshold method and our method perform well and they yield similar results.

Note that the estimation results of the diffusion parameters using the iteration procedure in the threshold method are not as good as the ones using the robust

procedure in our method. However, the threshold method is less sensitive to the estimation of the volatility σ_0 than our method. We have checked that the results of the threshold method would not change much if we replace the iteration procedure by the robust procedure to estimate the diffusion parameters.

Next, we compare the threshold method with our method in the estimation of jump size distribution.

The threshold method results in an estimate of the cumulative distribution function, while our method produces an estimate of the probability density function. To compare them, we will transform the cdf obtained from the threshold method to a pdf, using the "kernel" method. However, the kernel method must be modified since there are no observations available in a neighbourhood of zero (between $-B_n$ and B_n). For this, we propose to simulate "small jumps" following some distribution. Below we provide a description of the procedure that we have used.

Step 1 Put *large jumps* (i.e. jump observations with size larger than or equal to the threshold B_n) in one group. Let m_1 be the number of large jumps.

Step 2 The number of *small jumps* (i.e. jump observations with size smaller than the threshold B_n) is calculated by

$$m_2 = m - m_1, \qquad \text{where} \quad m = \hat{\lambda} \cdot \#\{\text{years}\}.$$

Then simulate m_2 small jumps (between $-B_n$ and B_n) following a distribution that is the same as the distribution of the random variable

$$D = \frac{B_n}{c} \cdot Z \cdot 1_{\{|Z| \le c\}}, \tag{6.7}$$

where Z is a standard normal random variable, c is the $(1 - \frac{m_1}{2m})$ - quantile of the standard normal distribution, and 1_A is an indicator function over A. If $m_2 = 0$, then we skip this step.

Step 3 Combine the large jumps from Step 1 and the small jumps from Step 2 to form the jump data. Then apply the kernel method (e.g. with the Gaussian kernel) to the jump data to estimate the pdf of the jump size.

In the above Step 2 we simulate "small jumps" from the distribution of the random variable D defined in (6.7). By doing this, the resulted jump data "describes" a normal distribution in the Merton model case. More specifically, in the Merton model case, the obtained "large jumps" from Step 1 represents the data of a normal distribution with size larger than B_n and with weight m_1/m, assuming that we can ignore the effect, when the jump size is large, of the convolution of jump size distribution with the diffusion distribution. The obtained "small jumps" from Step 2 represents the data of a normal distribution with size smaller than B_n and with weight $1-(m_1/m)$. Therefore, putting the "large" and "small" jumps together, we obtain a sample representing a normal distribution. Of course, this assumes that the estimation of jump intensity based on the threshold method is good enough.

To implement the above Step 2, an easy way is to use the normal random number generator and apply an appropriate bound. See below for details:

Step 2-1 Generate m observations, $r_1, r_2, ..., r_m$, from the standard normal distribution, and denote it by U.

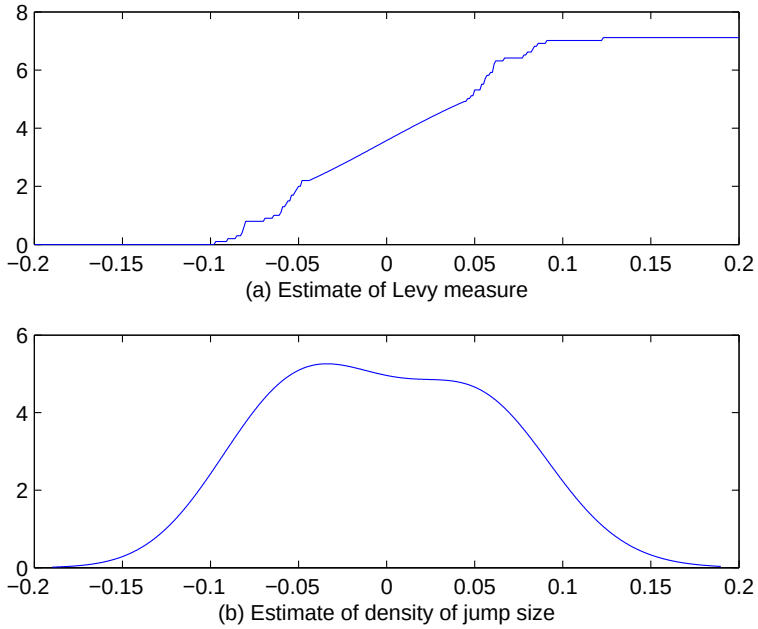
Step 2-2 Let U_1 be the set of m_2 ($m_2 < m$) observations from U with the smallest absolute values. That is, if $|r^{(1)}| < |r^{(2)}| < \cdots < |r^{(m_2)}| < \cdots < |r^{(m)}|$, then $U_1 = \{r^{(1)}, r^{(2)}, ..., r^{(m_2)}\} \subset U$. Denote $a = r^{(m_2)} = \max_{r \in U_1} |r|$.

Step 2-3 Multiply each observation in U_1 by $\frac{B_n}{a}$, and the resulted data will represent the "small jumps".

Now we consider the Merton model with the same parameter settings as in Example 2.1.2. Figure 6.2 shows an estimate of the (cumulative) Levy measure

$\hat{\nu}(x)$ based on the threshold method and an estimate of the jump size density using the transformation method described above, based on one set of simulated data. We found that it happens frequently that the middle part of the estimated density is over-smoothed. This can be explained as follows: to estimate the Lévy measure using the threshold method, the middle part is fit by a polynomial that increases slowly, so that the jump intensity might be underestimated. This phenomenon is seen more often when σ_X is small.

Figure 6.2: An estimate of Lévy measure based on the threshold method



After transforming the cdf of the jump size obtained by using the threshold method into a pdf, we calculate the total variation distance between this (estimated) pdf and the true pdf of jump size. Table 6.2 provides the results of the total variation distances for different scenarios of parameter settings of the Merton model. Similarly as in Chapter 5, we use $[-l_X, l_X]$ with $l_X = 5\sigma_X$ to define the interval over which we estimate the jump density. Repeat 100,000 times for each scenario.

Table 6.2: TV distance for threshold method [Merton]

	$\sigma_X \approx 3.79\%$		$\sigma_X = 10\%$		$\sigma_X = 60\%$	
n	$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$
1000	.3138(.0331)	.3203(.0230)	.0847(.0351)	.0651(.0249)	.0947(.0373)	.0720(.0257)
2500	.2981(.0177)	.3119(.0142)	.0579(.0210)	.0448(.0151)	.0655(.0229)	.0503(.0162)
10000	.2911(.0093)	.3114(.0077)	.0339(.0097)	.0266(.0069)	.0394(.0115)	.0302(.0082)
40000	.2940(.0056)	.3189(.0044)	.0229(.0044)	.0184(.0031)	.0279(.0071)	.0206(.0049)

From this table we can see that the results based on the threshold method become better as σ_X increases. Comparing this table with Table 5.2, which presents the results based on our method, we can see the following:

- When σ_X is small (e.g. $\sigma_X = 3 * 0.2 * \sqrt{1/250} \approx 3.79\%$, which is triple of the daily volatility), our method performs significantly better.
- When σ_X is large (e.g. $\sigma_X = 10\%$, which is approximately 8 times of the daily volatility), the threshold method performs better, if the sample size n is not too large. Note that, when σ_X is large, most of the jump observations are visible, and the threshold method thus works well.
- An important finding is that, as n increases, our method improves faster than the threshold method does. Thus, for large n, our method performs better than the threshold method, even in the case when σ_X is large.

Remark 6.2.1. As stated in Mancini (2004) and Gegler and Stadtmüller (2010), the asymptotic properties of the estimators based on the threshold method relies on the requirement that the time horizon $T \to \infty$ and the data frequency $\Delta \to 0$. When the sample size increases due to larger T but fixed Δ, the threshold B_n stays at the same level, and thus there is no improvement in the estimation of jump frequency (which is verified by Table 6.1). The improvement in the estimation of jump size distribution is only because we observe more "large" jump observations

which can be used to estimate the Lévy measure (or the pdf of jump size based on the modified "kernel" method).

Table 6.3: Estimation results using threshold method [Kou]

		$\eta_+ = 1/0.02$, $\eta_- = 1/0.04$		$\eta_+ = 1/0.1$, $\eta_- = 1/0.2$	
		$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$
n=1000	$\hat{\mu_0}$	0.0483 (0.1020)	-0.0074 (0.1052)	0.0858 (0.0990)	0.0742 (0.0983)
	$\hat{\sigma_0}$	0.2039 (0.0055)	0.2102 (0.0065)	0.1975 (0.0051)	0.1962 (0.0057)
	$\hat{\lambda}_{un}$	3.8354 (0.9993)	7.1568 (1.3891)	9.8618 (1.5659)	19.5082 (2.1768)
	$\hat{\lambda}$	6.1276 (1.6066)	11.4701 (2.2374)	12.4640 (2.0823)	24.5758 (2.8761)
n=2500	$\hat{\mu_0}$	0.0488 (0.0643)	-0.0070 (0.0666)	0.0865 (0.0626)	0.0742 (0.0618)
	$\hat{\sigma_0}$	0.2039 (0.0034)	0.2100 (0.0041)	0.1975 (0.0032)	0.1961 (0.0035)
	$\hat{\lambda}_{un}$	3.8346 (0.6336)	7.1688 (0.8832)	9.8629 (0.9881)	19.5145 (1.3784)
	$\hat{\lambda}$	6.1264 (1.0185)	11.4872 (1.4221)	12.4666 (1.3140)	24.5799 (1.8209)
n=10000	$\hat{\mu_0}$	0.0488 (0.0322)	-0.0064 (0.0333)	0.0862 (0.0313)	0.0745 (0.0311)
	$\hat{\sigma_0}$	0.2038 (0.0017)	0.2099 (0.0020)	0.1975 (0.0016)	0.1961 (0.0018)
	$\hat{\lambda}_{un}$	3.8356 (0.3181)	7.1699 (0.4443)	9.8538 (0.4982)	19.5181 (0.6903)
	$\hat{\lambda}$	6.1284 (0.5115)	11.4893 (0.7152)	12.4567 (0.6619)	24.5820 (0.9100)
n=40000	$\hat{\mu_0}$	0.0489 (0.0161)	-0.0063 (0.0167)	0.0864 (0.0156)	0.0743 (0.0155)
	$\hat{\sigma_0}$	0.2038 (0.0009)	0.2099 (0.0010)	0.1975 (0.0008)	0.1961 (0.0009)
	$\hat{\lambda}_{un}$	3.8244 (0.1543)	7.1941 (0.2213)	9.8409 (0.2553)	19.5409 (0.3498)
	$\hat{\lambda}$	6.1113 (0.2485)	11.5272 (0.3563)	12.4433 (0.3350)	24.6002 (0.4579)

Note: The threshold B_n is the same on average for all scenarios in the same column. The average values for the four columns are 0.0429, 0.0442, 0.0416, 0.0413.

Table 6.4: TV distance for threshold method [Kou]

	$\eta_+ = 1/0.02$, $\eta_- = 1/0.04$ $l_X = 0.2$		$\eta_+ = 1/0.1$, $\eta_- = 1/0.2$ $l_X = 1.0$	
	$\lambda = 12.5$	$\lambda = 25$	$\lambda = 12.5$	$\lambda = 25$
$n = 1000$	.3443 (.0318)	.3439 (.0217)	.1315 (.0383)	.1072 (.0285)
$n = 2500$	.3315 (.0181)	.3393 (.0138)	.0973 (.0249)	.0811 (.0187)
$n = 10000$	.3296 (.0095)	.3420 (.0079)	.0637 (.0126)	.0553 (.0095)
$n = 40000$	.3354 (.0052)	.3475 (.0045)	.0445 (.0062)	.0405 (.0046)

Now we present the estimation results for Kou model based on the threshold

method. Table 6.3 shows the estimation results on the jump frequency. Table 6.4 shows the results on jump size distribution. When comparing with our method, we arrive at exactly the same observations as in the case of the Merton model.

In conclusion, our method works better than the threshold method when the jump sizes are relatively small on average. When the jump sizes are large on average, the threshold method works better if the sample size is not too large. When the sample size is large, our method performs better, even in the case that the sample sizes are large on average.

Chapter 7

Applications: Classification and Testing for a Change of Jump Frequency

In previous chapters, we proposed estimators of the diffusion parameters, the jump frequency and the jump size distribution. Therefore, by now we have all tools required to estimate every component of the jump-diffusion model (or equivalently, the two-component mixture model).

In this chapter we illustrate how to apply these results to solve two problems. The first problem is to classify each increment as the one with or without jump. The second deals with the problem of detecting a change in the jump frequency. In the paper by Lee et. al. (2003), the cusum method was used to test the parameter change. We will apply this method in our simulations.

7.1 Classification of Increments

In this section, we discuss the classification of each increment as the one with or without jump.

As discussed in Chapter 2, under assumption (A1-2), the increments of the jump-diffusion model (2.1) can be approximated by a two-component mixture model (2.13):

$$Y = (1 - I)Z + IG, \tag{7.1}$$

where I is a Bernoulli random variable with

$$\mathbb{P}(I = 1) = \lambda\Delta = \alpha. \tag{7.2}$$

The probability density function of Y is then given by the equation (2.15):

$$f_Y = (1 - \alpha)\phi_Z + \alpha f_G, \tag{7.3}$$

where $f_G = \phi_Z * f_X$. In previous chapters, we proposed estimators of the diffusion parameters (i.e. in the normal density ϕ_Z), the jump frequency α and the jump size density f_X. Using them, we can estimate all the components of the density function f_Y. In this section we consider the problem of classification of increments as the ones with jump and the ones without jump.

Given an observation $Y = y$, we would like to find the probability that it can be classified as the one that comes from the distribution G (corresponding to a jump), i.e. we would like to find $\mathbb{P}(I = 1|Y \in dy)$. To understand the notation, we have

$$\{Y \in dy\} \equiv \{Y \in (y, y + dy)\},$$

where dy is an infinitely small number. Using Bayes' rule, we have

$$\begin{aligned} p(y) &:= \mathbb{P}(I = 1|Y \in dy) \\ &= \frac{\mathbb{P}(I = 1, Y \in dy)}{\mathbb{P}(Y \in dy)} \\ &= \frac{\mathbb{P}(Y \in dy|I = 1)\mathbb{P}(I = 1)}{f_Y(y)dy} \end{aligned} \tag{7.4}$$

where

$$\mathbb{P}(Y \in dy | I = 1) = \mathbb{P}(G \in dy | I = 1) = \mathbb{P}(G \in dy) = f_G(y)dy \tag{7.5}$$

by the independence between I and G. Using equations (7.2) and (7.3) again, we have

$$p(y) = \frac{\alpha f_G(y)}{(1-\alpha)\phi_Z(y) + \alpha f_G(y)}. \tag{7.6}$$

On the other side, the probability that the observation y is generated from the normal distribution Z (i.e. without a jump) is $\mathbb{P}(I = 0 | Y \in dy) = 1 - p(y)$. Thus, the probability $p(y), y \in \mathbb{R}$ provides important information for the classification problem.

Before proceeding to the discussion of the classification procedure, we present an intuitive result describing the limiting behavior of $p(y)$ as $\Delta \to 0$. It states that as $\Delta \to 0$, which corresponds to the case where the continuous trajectory of a model is available, we can identify all the jumps with probability one.

Proposition 7.1.1. *As $\Delta \to 0$, the probability $p(y)$ converges to one if $y \neq 0$ and zero otherwise, i.e.*

$$\lim_{\Delta \to 0} p(y) = \begin{cases} 1, & \text{if} \quad y \neq 0; \\ 0, & \text{if} \quad y = 0. \end{cases}$$

To prove this result, we use the *Dirac delta function* $\delta_0(\cdot)$. It can be defined as a generalized function on the real line such that

$$\delta_0(x) = \begin{cases} +\infty, & x = 0 \\ 0, & x \neq 0 \end{cases} \tag{7.7}$$

and

$$\int_{-\infty}^{\infty} \delta_0(x)\, dx = 1.$$

It can also be interpreted as a singular distribution which is characterized by its cdf being a Heaviside step function $H(x)$.

In our problem, as $\Delta \to 0$,

$$\phi_Z(y) \equiv \phi(y; \mu_0\Delta, \sigma_0^2\Delta) = \frac{1}{\sqrt{2\pi\sigma_0^2\Delta}} \exp\left\{-\frac{(y-\mu_0\Delta)^2}{2\sigma_0^2\Delta}\right\} \longrightarrow \delta_0(y).$$

Then, as $\Delta \to 0$,

$$f_G(y) = (\phi_Z * f_X)(y) = \int_{-\infty}^{\infty} f_X(y-x)\phi_Z(x)dx \longrightarrow \int_{-\infty}^{\infty} f_X(y-x)dH(x) = f_X(y).$$

This follows from the fact that f_G is the limiting distribution of a sum where one of the two components converges to zero (in probability and in distribution). Thus, for any $y \neq 0$, we have

$$\lim_{\Delta\to 0} \frac{(1-\alpha)\phi_Z(y)}{\alpha f_G(y)} = \lim_{\Delta\to 0} \frac{(1-\lambda\Delta)\phi_Z(y)}{\lambda\Delta \cdot f_X(y)} = \frac{1}{\lambda f_X(y)} \cdot \lim_{\Delta\to 0} \frac{\phi_Z(y)}{\Delta} = 0,$$

where for the last step the L'Hospital's rule is used. For $y = 0$, we have

$$\lim_{\Delta\to 0} \frac{(1-\alpha)\phi_Z(y)}{\alpha f_G(y)} = \frac{1}{\lambda f_X(y)} \lim_{\Delta\to 0} \frac{\phi_Z(y)}{\Delta} = \infty.$$

Therefore,

$$\lim_{\Delta\to 0} p(y) = \lim_{\Delta\to 0} \frac{\alpha f_G(y)}{(1-\alpha)\phi_Z(y) + \alpha f_G(y)} = \lim_{\Delta\to 0} \frac{1}{\frac{(1-\alpha)\phi_Z(y)}{\alpha f_G(y)} + 1} = \begin{cases} 1, & \text{if} \quad y \neq 0; \\ 0, & \text{if} \quad y = 0. \end{cases}$$

This completes the proof. ⊠

Consider the random variable $p(Y)$, where $p(\cdot)$ is defined in (7.6) and Y is a random variable with pdf given by (7.3). To formulate our classification procedure, we define an indicator variable by

$$J^*(Y) = \begin{cases} 1 & \text{if } p(Y) \geq \frac{1}{2}, \\ 0 & \text{otherwise,} \end{cases} \tag{7.8}$$

then an observed value y is classified as an observation with jump if $J^*(y) = 1$, and without jump if $J^*(y) = 0$. This intuitive method can be verified to be optimal in some sense as explained below.

Bayes Procedure for Decision Problems

Let us consider a general decision problem. Suppose X is a random variable from a distribution with pdf (in the continuous case) or pmf (i.e. probability mass function in the discrete case) given by $f_\theta, \theta \in \Theta$. Once we observe a value of X, say x, we take an *action*. Let $\mathcal{A}$ denote the set of all feasible actions. A *decision function* δ is any function on $\mathbb{R}$ that takes values in $\mathcal{A}$, more specifically, $\delta(X) : \mathbb{R} \to \mathcal{A}$. A nonnegative function $L : \Theta \times \mathcal{A} \to \mathbb{R}$ represents a *loss function*, in the sense that $L(\theta, a)$ is the loss if an action a is taken when θ is the true parameter value. Note that $L(\theta, \delta(X))$ is a random variable. Let $\mathcal{D}$ be a class of decision functions, then the function R defined on $\Theta \times \mathcal{D}$ by

$$R(\theta, \delta) = \mathbb{E}_\theta L(\theta, \delta(X))$$

is known as the *risk function* associated with δ at θ.

Using a Bayesian approach, we treat θ as a random variable distributed according to a pdf/pmf $\pi(\theta)$ on Θ, and π is called the *prior distribution*. In this framework, the risk function $R(\theta, \delta)$ is viewed as a conditional expected loss given θ. The *Bayes risk* of a decision function δ is defined by

$$R(\pi, \delta) = \mathbb{E}_\pi R(\theta, \delta).$$

The *Bayes solution* or *Bayes rule* is a decision function that minimizes the Bayes risk $R(\pi, \delta)$.

The hypothesis testing problem can be regarded as a special case of the general decision problem. In the following we assume that the set $\mathcal{A}$ contains only two

points, a_0 and a_1, in the way that a_0 corresponds to the acceptance of the null hypothesis $H_0 : \theta = \theta_0$, and a_1 corresponds to the alternative $H_1 : \theta = \theta_1$, and $\Theta = \{\theta_0, \theta_1\}$. Also, suppose the loss function is defined by

$$\begin{cases} L(\theta_0, a_1) = b > 0, \\ L(\theta_1, a_0) = c > 0, \\ L(\theta_1, a_1) = 0, \\ L(\theta_0, a_0) = 0. \end{cases} \tag{7.9}$$

Further, let $\pi(\theta_0) = \pi_0$ and $\pi(\theta_1) = 1 - \pi_0 = \pi_1$. Then the Bayes risk is

$$R(\pi, \delta) = b\pi_0 \mathbb{P}_{\theta_0}\{\delta(X) = a_1\} + c\pi_1 \mathbb{P}_{\theta_1}\{\delta(X) = a_0\}. \tag{7.10}$$

The following result, which can be found in Rohatgi and Saleh (2001), provides the Bayes solution for this case.

Proposition 7.1.2. *Let X be a random variable with pmf/pdf $f_\theta, \theta \in \Theta = \{\theta_0, \theta_1\}$. Let $\pi(\theta_0) = \pi_0$, $\pi(\theta_1) = 1 - \pi_0 = \pi_1$ be the prior pmf on Θ. A Bayes solution for testing $H_0 : X \sim f_{\theta_0}$ against $H_1 : X \sim f_{\theta_1}$, using the loss function (7.9), is to reject H_0 if*

$$\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} \geq \frac{b\pi_0}{c\pi_1}. \tag{7.11}$$

Let us come back to our problem. In the language of hypothesis testing, we consider the test that, given an observation y from the p.d.f (7.3), the null hypothesis is

$$H_0 : y \sim \phi_Z, \qquad \text{i.e. with no jump,}$$

and the alternative is $H_1 : y \sim f_G$, i.e. with jump. Correspondingly, the prior probabilities are $\pi_0 = 1 - \alpha$ and $\pi_1 = \alpha$. Suppose we use the loss function (7.9) with $b = c$. Then by Proposition 7.1.2, the Bayes solution, denoted by δ^*, is to reject H_0 if

$$\frac{f_G(y)}{\phi_Z(y)} \geq \frac{1 - \alpha}{\alpha}, \tag{7.12}$$

or equivalently

$$p(y) \geq \frac{1}{2}.$$

Otherwise, we accept H_0. This is exactly the intuitive classification criterion $J^*(y)$ given in (7.8).

If we further let $b = c \equiv 1$, i.e. the loss function in (7.9) is the 0-1 loss function, then, by (7.10), the Bayes risk of the Bayes solution δ^* is given by

$$\begin{aligned} R(\pi, \delta^*) &= (1-\alpha)\mathbb{P}_{\phi_Z}\left(p(Y) \geq 1/2\right) + \alpha \mathbb{P}_{f_G}\left(p(Y) < 1/2\right) \\ &= \int 1_{\{p(y)\geq 1/2\}}(1-\alpha)\phi_Z(y)dy + \int 1_{\{p(y)<1/2\}}\alpha f_G(y)dy \\ &= \mathbb{E}\left\{1_{\{p(Y)\geq 1/2\}}(1-p(Y)) + 1_{\{p(Y)<1/2\}}p(Y)\right\} \end{aligned} \tag{7.13}$$

or equivalently,

$$R(\pi, \delta^*) = \mathbb{E}\left[\min\left(p(Y), 1-p(Y)\right)\right] \tag{7.14}$$

or

$$R(\pi, \delta^*) = \frac{1}{2} - \frac{1}{2}\,\mathbb{E}[|2p(Y)-1|]. \tag{7.15}$$

Note that in (7.13), $\mathbb{E}\left[1_{\{p(Y)\geq 1/2\}}(1-p(Y))\right]$ is the probability that a diffusion observation is misclassified as a jump observation, and $\mathbb{E}\left[1_{\{p(Y)<1/2\}}p(Y)\right]$ is the probability that a jump observation is misclassified as a diffusion observation.

In the language of classification problem, the *misclassification error* of a classification criterion $J(Y)$ is given by

$$\epsilon := \mathbb{P}\{J(Y) \neq I(Y)\},$$

where

- $I(y) = 1$ means the observation y truly involves a jump;
- $I(y) = 0$ means the observation y does not involve a jump;

- $J(y) = 1$ means the observation y is classified as the one with jump;
- $J(y) = 0$ means the observation y is classified as the one without jump.

When the 0-1 loss function is used for the above Bayes procedure, the Bayes risk is equal to the misclassification error. Then the Bayes solution which minimizes the Bayes risk is exactly the classification criterion which minimizes the misclassification error. Since the Bayes solution δ^* is the same as the intuitive classification criterion J^* given in (7.8), we conclude that the criterion J^* minimizes the misclassification error. Moreover, the minimized misclassification error is equal to the Bayes risk of the Bayes solution, i.e.

$$\epsilon^* := \mathbb{P}\{J^*(Y) \neq I(Y)\} = R(\pi, \delta^*) = \mathbb{E}\left[\min\left(p(Y), 1 - p(Y)\right)\right]. \tag{7.16}$$

Another proof of this form of the misclassification error can be found in Devroye et. al. (1996).

From (7.16) we can further write

$$\begin{aligned} \epsilon^* &= \int_{\mathbb{R}} \min\left\{p(Y), 1 - p(Y)\right\} f_Y(y) dy \\ &= \int_{\mathbb{R}} \min\{\alpha f_G(y), (1-\alpha)\phi_Z(y)\} dy. \end{aligned} \tag{7.17}$$

It is obvious that

$$\epsilon^* \to 0, \qquad \text{as } \Delta \to 0$$

since $\min\{\alpha f_G(y), (1-\alpha)\phi_Z(y)\} \leq \alpha f_G(y)$ and $\int_{\mathbb{R}} \alpha f_G(y) dy = \alpha = \lambda\Delta \to 0$. However, even if the misclassification error ϵ^* converges to zero as $\Delta \to 0$, one reason for this is that the total number of observations goes to infinity, among which the average number of jumps is unchanged and finite. To solve this problem, we can consider the "adjusted" misclassification error:

$$\epsilon^*/\Delta.$$

In the following proposition we show that, for the Merton and Kou models, this "adjusted" misclassification error goes to zero. The convergence rate is also provided.

Proposition 7.1.3. *For the Merton and Kou models, the misclassification error* ϵ^* *defined in (7.16) satisfies*

$$\epsilon^*/\Delta \to 0$$

as $\Delta \to 0$. *Moreover, we have*

$$\lim_{\Delta\to 0} \frac{\epsilon^*}{\Delta\sqrt{\Delta \log \frac{1}{\Delta}}} = \sqrt{3}\sigma_0\lambda[f_X(0-) + f_X(0+)] \tag{7.18}$$

where $f_X(0-)$ *and* $f_X(0+)$ *are the left and right limits of the density function of jump size.*

Proof. See the appendix in the end of this chapter.

Surprisingly, the result for the Kou model is similar to the result for the Merton model. The only difference is that for the Kou model, the density function of jump size is not continuous at point zero, and thus in the result $f_X(0)$ is changed to $(f_X(0-) + f_X(0+))/2$. (Note that $f_X(0-) \neq 0$ and $f_X(0+) \neq 0$ in the Kou model.)

Remark 7.1.4. We conjecture that the results in Proposition 7.1.3 also hold true for any model (7.1) with an absolute continuous jump size distribution whose pdf f_X is bounded (i.e. there exists $M > 0$ such that $f_X(x) \leq M$ for all $x \in \mathbb{R}$), and the left and right limits of f_X at zero exist.

In the following we use numerical procedures to find the theoretical values of the misclassification errors at some fixed frequencies for the Merton and Kou models.

Example 7.1.5 (Merton and Kou Models)**.** We use the same parameter settings as in Example 2.1.2 for the Merton model, and as in Example 2.1.3 for the Kou model.

Note that for the Merton model, where the jump size X follows a normal distribution $\phi(\cdot;\mu_X,\sigma_X^2)$, the convolution is $f_G=\phi_Z*f_X=\phi(\cdot;\mu_0\Delta+\mu_X,\sigma_0^2\Delta+\sigma_X^2)$, which is still a normal distribution. For the Kou model, where X follows a double exponential distribution with pdf in (1.6), the convolution $f_G=\phi_Z*f_X$ is

$$\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma_0^2\Delta}}e^{-\frac{(y-x-\mu_0\Delta)^2}{2\sigma_0^2\Delta}}\left(p\eta_1e^{-\eta_1x}1_{\{x\geq 0\}}+q\eta_2e^{\eta_2x}1_{\{x<0\}}\right)dx$$
$$=p\eta_1e^{\frac{\sigma_0^2\Delta\eta_1^2}{2}+\eta_1(\mu_0\Delta-y)}\Phi\left(\frac{y-\mu_0\Delta-\sigma_0^2\Delta\eta_1}{\sigma_0\sqrt{\Delta}}\right)+q\eta_2e^{\frac{\sigma_0^2\Delta\eta_2^2}{2}+\eta_2(y-\mu_0\Delta)}\Phi\left(\frac{\mu_0\Delta-y-\sigma_0^2\Delta\eta_2}{\sigma_0\sqrt{\Delta}}\right).$$

Table 7.1 shows the theoretical values of the misclassification errors when $\Delta=1/250$ and $\Delta=1/2500$ for the Merton model. Note that both types of misclassification errors are included in this table. To read the table, take the first entry for example: the number 1.700% is the probability $\mathbb{P}$(classified as jump, and jump has occurred), while the number 34.0% inside the parentheses is the probability $\mathbb{P}$(classified as jump, given that jump has occurred).

Table 7.1: Theoretical classification results for Merton model

	$\Delta=1/250$		$\Delta=1/2500$	
	Jump	Non-jump	Jump	Non-jump
Classified as jump	1.700% (34.0%)	0.243% (0.3%)	0.3411% (68.2%)	0.0093% (0.01%)
Classified as non-jump	3.300% (66.0%)	94.757% (99.7%)	0.1589% (31.8%)	99.4907% (99.99%)
Total	5.000% (100%)	95.000% (100%)	0.5000% (100%)	99.5000% (100%)

Let us define

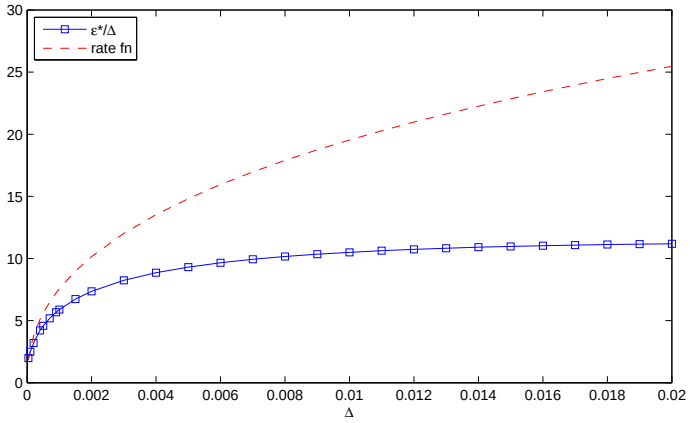
Type-I error := $\mathbb{P}$(classified as jump, given that jump has not occurred);

Type-II error := $\mathbb{P}$(classified as non-jump, given that jump has occurred).

Then as Δ increases, type-I error decreases from 0.3% to 0.01%, and type-II error decreases from 66.0% to 31.8%. This verifies the result stated in Proposition 7.1.3 that the "adjusted" misclassification errors (i.e. ϵ^*/Δ) decreases as the data frequency increases.

Figure 7.1 shows the graph of the "adjusted" misclassification error (i.e. ϵ^*/Δ) with respect to the data frequency Δ. The rate function, $\sqrt{3}\sigma_0\lambda[f_X(0-)+f_X(0+)]\cdot\sqrt{\Delta\log\frac{1}{\Delta}}$, is also plotted (in dashed line) for comparison. It seems that the two curves converge to zero at the same speed. This is consistent with the result in Proposition 7.1.3.

Figure 7.1: Adjusted misclassification error v.s. data frequency [Merton]



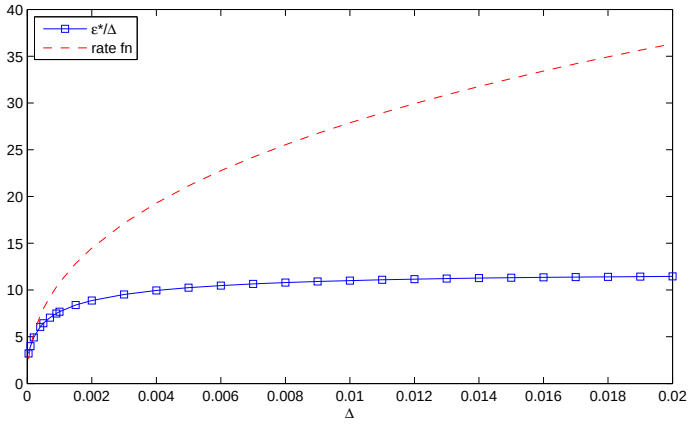
Similarly, Table 7.2 shows the results for Kou model. Figure 7.2 shows the graph of the "adjusted" misclassification error. We arrive at the same conclusion as in the Merton case.

However, in practice $p(y)$ is unknown. Thus, we need to propose an estimator of $p(y)$. Given the estimators $\hat{\mu}$, $\hat{\sigma}$, $\hat{\alpha}$ and $\hat{f}_X$ obtained in the previous chapters, we

Table 7.2: Theoretical classification errors for Kou model

	$\Delta = 1/250$		$\Delta = 1/2500$	
	Jump	Non-jump	Jump	Non-jump
Classified as jump	1.193% (23.9%)	0.171% (0.2%)	0.2686% (53.7%)	0.0108% (0.01%)
Classified as non-jump	3.807% (76.1%)	94.829% (99.8%)	0.2314% (46.3%)	99.4892% (99.99%)
Total	5.000% (100%)	95.000% (100%)	0.5000% (100%)	99.5000% (100%)

Figure 7.2: Adjusted misclassification error v.s. data frequency [Kou]



can estimate $p(Y)$ by

$$\hat{p}(Y) = \frac{\hat{\alpha}\hat{f}_G(Y)}{(1-\hat{\alpha})\hat{\phi}_Z(Y) + \hat{\alpha}\hat{f}_G(Y)}$$

where $\hat{f}_G = \hat{\phi}_Z * \hat{f}_X$. Then we can classify a given observation y as an observation involving a jump if

$$\hat{p}(y) \geq \frac{1}{2} \tag{7.19}$$

and otherwise we classify it as an observation without a jump. In the following we apply this criterion for the classification of the increments.

Example 7.1.6 (Merton and Kou Models)**.** For the Merton model, we use the same parameter settings as in Example 2.1.2. For Kou model, we use the same parameter settings as in Example 2.1.3.

Table 7.3 shows the classification results for the two frequencies for the Merton model, where the numbers are the average of 100,000 repetitions. The results are quite close to the ones in Table 7.1 where the true model was used for calculation.

Table 7.3: Classification results for Merton model

	$\Delta = 1/250$		$\Delta = 1/2500$	
Percentage	Jump	Non-jump	Jump	Non-jump
Classified as jump	1.691%	0.255%	0.3472%	0.0184%
Classified as non-jump	3.310%	94.744%	0.1529%	99.4815%
Total	5.001%	94.999%	0.5001%	99.4999%

Table 7.4: Classification results for Kou model

	$\Delta = 1/250$		$\Delta = 1/2500$	
Percentage	Jump	Non-jump	Jump	Non-jump
Classified as jump	1.656%	0.205%	0.3111%	0.0137%
Classified as non-jump	3.345%	94.794%	0.1889%	99.4863%
Total	5.001%	94.999%	0.5000%	99.5000%

Similarly, Table 7.4 shows the classification results for the Kou model. The numbers are also close to the ones in Table 7.2 for which the true model was used for calculation, although not as close as in the Merton case. This is because

the estimation accuracy of components of the Kou model is not as good as the estimation accuracy of components of the Merton model.

These results confirm that the Bayes classification method based on our estimates of the jump-diffusion model can be used in practice.

7.2 Testing for a Change of Jump Frequency

Changes happen every day and everywhere. An awareness of these changes can help people to avoid unnecessary losses and harness beneficial transitions. The change point analysis (or disorder problems) is widely used in various fields such as quality control, economics, finance, medicine, psychology and literature.

In this section we describe a method to test whether there is a change in the jump frequency. The null hypothesis is

H_0: there is no change in the jump frequency α (or equivalently, λ).

Since $p(Y)$ in (7.6) is defined as a conditional expectation, it follows that $\mathbb{E}(p(Y)) = \mathbb{E}(I) = \alpha$. Therefore we can test the change in the jump frequency α by testing the change in the mean of $p(Y)$.

There exist different methods dealing with the change-point analysis, such as the likelihood ratio, Bayesian analysis, nonparametric analysis and approaches based on information criteria. For our problem, we do not know the distribution of $p(Y)$, so we have to consider a nonparametric approach. One of the most important nonparametric methods is the *cusum test*.

The cusum (cumulative sum) test was initiated by Page (1954, 1955). It has been used widely to detect the change and locate the change point(s) for a parameter. In the beginning the i.i.d. samples were considered, and then the focus moved into the time series context. For example, in the paper by Lee et. al. (2003), the

authors use the cusum method to construct a statistic for testing the parameter change in time series models. It was shown that, under the null hypothesis, this statistic has an asymptotic distribution which is the same as the distribution of the supremum of the square of a standard Brownian bridge. This result holds in a quite general framework. In the following we first state this result, and then apply it to our problem of testing for a change in the jump frequency.

Theorem 7.2.1 (Lee et al. (2003))**.** *Suppose* $\{X_1, X_2, ...\}$ *is a stationary time series, and* $\theta \in \mathbb{R}$ *is the parameter to be tested for change. The null hypothesis is* H_0*:* θ *does not change for* $X_1, X_2, ..., X_n$*. Let* $\hat{\theta}_k$ *be an estimator of* θ *based on* $\{X_1, X_2, ..., X_k\}$*,* $k = 1, 2, ..., n$*. Assume that* $\hat{\theta}_k$ *satisfies*

$$\hat{\theta}_k - \theta = \frac{1}{k}\sum_{j=1}^{k} l_j(\theta) + \frac{1}{\sqrt{k}}\Delta_k, \tag{7.20}$$

where $l_j(\theta)$ *forms stationary martingale differences (with respect to some filtration) with* $\gamma = Var(l_1(\theta))$*, and* Δ_k *satisfies*

$$\max_{1\le k\le n} \frac{\sqrt{k}}{\sqrt{n}}|\Delta_k| = o_P(1).$$

Then, under H_0*, we have*

$$M_n := \max_{1\le k\le n} \frac{k^2}{n\gamma}(\hat{\theta}_k - \hat{\theta}_n)^2 \quad \xrightarrow{d} \quad \sup_{0\le s\le 1}(W^o(s))^2 \tag{7.21}$$

as $n \to \infty$*, where* W^o *is a one-dimensional standard Brownian bridge. We reject* H_0 *if* M_n *is large.*

Note that the distribution of $\sup_{0\le s\le 1}(W^o(s))^2$ cannot be obtained explicitly, but we can use simulation to obtain the empirical quantiles. See Table 7.5 for which 100,000 samples are simulated.*

*In our implementation, we generate the trajectories of Brownian bridge directly, based on $W^o(t) = W(t) - tW(1)$. It is slightly different from the procedure in Lee et al. (2003) where a formula was involved in the simulation of the data and computation of the empirical quantiles.

Table 7.5: Empirical quantiles of supremum of square of a Brownian bridge

	.01	.05	.10	.90	.95	.99
Quantile	0.183	0.256	0.311	1.457	1.794	2.574

For our problem, in order to test the change in the jump frequency α, we will test the change in the mean of $p(Y)$. The difficulty in applying Theorem 7.2.1 is that $p(Y_1), p(Y_2), ..., p(Y_n)$ are not observable, although $Y_1, ..., Y_n$ are observable. However, we can use the methods proposed in Chapters 2-5 to estimate the function p. Based on these estimates, we can define the following sequence:

$$\hat{\theta}_k = \frac{1}{k}\sum_{j=1}^{k} \hat{p}(Y_j) = \frac{1}{k}\sum_{j=1}^{k} \frac{\hat{\alpha}\hat{f}_G(Y_j)}{(1-\hat{\alpha})\hat{\phi}_Z(Y_j) + \hat{\alpha}\hat{f}_G(Y_j)} \tag{7.22}$$

with $\hat{f}_G = \hat{\phi}_Z * \hat{f}_X$. In (7.20), we let $l_j(\theta) = \hat{p}(Y_j) - \theta$ and $\Delta_k = 0$. Then $\gamma = \mathrm{Var}(l_1(\theta)) = \mathrm{Var}(\hat{p}(Y_1))$. The other details of the proposed method are explained in the following in the context of the Merton model.

Example 7.2.2 (Merton model)**.** We use the same parameter setting as in Example 2.1.2, and we generate a sample of size $n = 2500$ from the distribution of Y assuming that there is no change in the jump frequency. Then we do the following:

Step 1 Obtain estimates $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\alpha}$ by using the method discussed in Section 4.2, and obtain an estimate $\hat{f}_X$ by the method discussed in Section 5.2.

Step 2 Substitute the estimated values into $\hat{\theta}_k$ in (7.22). To calculate $\hat{f}_G = \hat{\phi}_Z * \hat{f}_X$, we approximate the convolution by discretizing the integral.

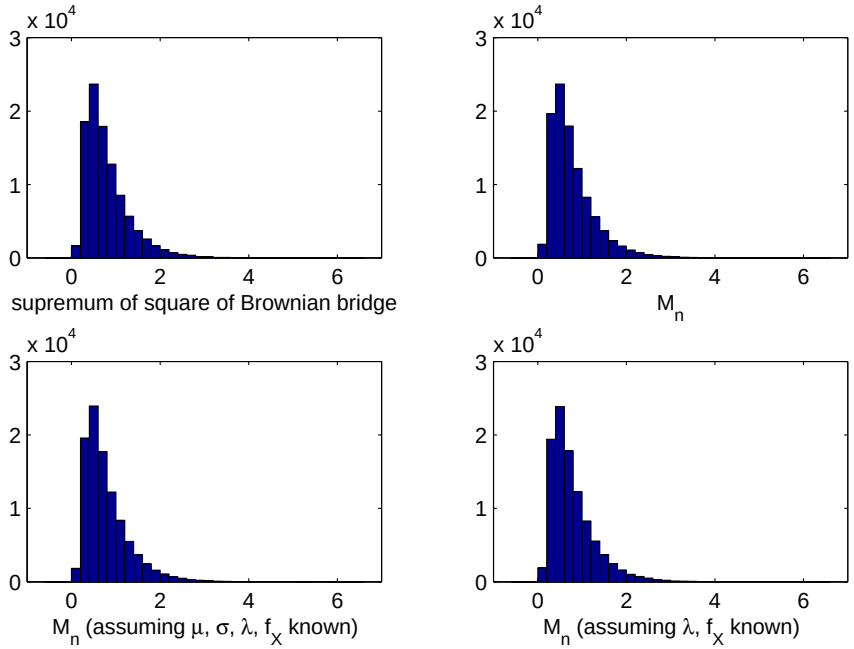
Step 3 For $\gamma = \mathrm{Var}(\hat{p}(Y_1))$, we use the sample variance of $\hat{p}(Y_1), \hat{p}(Y_2), ..., \hat{p}(Y_n)$.

Step 4 Calculate M_n in (7.21).

Repeat the above procedure 100,000 times. Then we obtain 100,000 values of M_n. Based on these, we created Figure 7.3, where

- The plot (1,1) shows the histogram of $\sup_{0\leq s\leq 1}(W^o(s))^2$.
- The plot (1,2) shows the histogram of M_n, for which we use estimates $\hat{\mu}$, $\hat{\sigma}$, $\hat{\alpha}$ and $\hat{f}_X$ to obtain $\hat{\theta}_k$ in (7.22).
- The plot (2,1) shows the histogram of M_n, for which we use true values of μ, σ, α and true density f_X to calculate $\hat{\theta}_k$ in (7.22).
- The plot (2,2) shows the histogram of M_n, for which we use estimates $\hat{\mu}$, $\hat{\sigma}$ and true α, f_X to calculate $\hat{\theta}_k$ in (7.22).

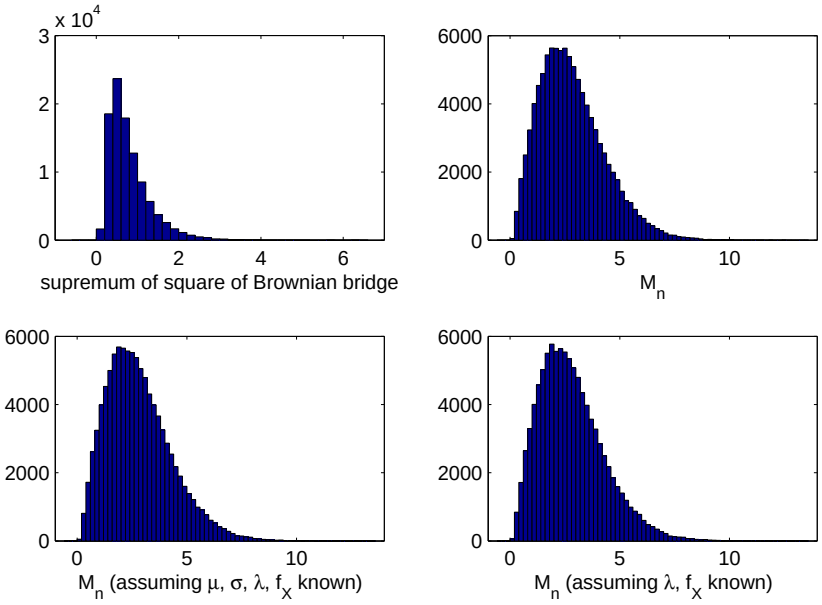
Figure 7.3: Histograms of $\sup_{0\leq s\leq 1}(W^o(s))^2$ and M_n when no change occurs



We can see that the four histograms in Figure 7.3 are quite similar to each other. This verifies that the distribution of $\sup_{0\leq s\leq 1}(W^o(s))^2$ is appropriate to use as the asymptotic distribution under H_0 for testing the change in jump frequency, even if $p(Y_j), j = 1, 2, ..., n$ are unknown and we have to use the estimated values.

Example 7.2.3 (Merton model (when change occurs))**.** Now we investigate on the performance of our method assuming that a change in the jump frequency has occurred. The same time horizon $T = 10$ and the same data frequency (daily) are used. We generate a sample of size $n = 2500$ of the increment Y, assuming that a change in the jump frequency has occurred at the moment corresponding to $n = 1300$, and the jump frequency changes from $\lambda = 12.5$ to $\lambda = 25$.

Figure 7.4: Histograms of $\sup_{0\leq s\leq 1}(W^o(s))^2$ and M_n when a change occurs



We follow the same procedure as described in Example 7.2.2 above. Figure 7.4 shows the histograms where:

- The plot (1,1) is the same as the one in Figure 7.3.
- The other three plots (1,2) (2,1) (2,2) still look similar to each other, but they are different from plot (1,1).

The difference between plot (1,1) and plots (1,2) (2,1) (2,2) is what we expected

because, when a change occurs, the asymptotic distribution of the statistics M_n will be different from the distribution of $\sup_{0\leq s\leq 1}(W^o(s))^2$. The similarity between the three plots (1,2) (2,1) (2,2) verifies that we can use the estimated values $\hat{p}(Y_j), j = 1, 2, ..., n$ when the true values are unknown.

We have applied our method to the Kou model, and the results are very similar to the ones we obtained for the Merton model.

Based on our simulation study we conclude that the proposed cusum method where we estimate the function p has a practical value. Below we illustrate its application in the context of real data.

Real Data Example

We use the same 30-year data set as in Section 5.4: the S&P 500 daily data from 1983 to 2012. There are 7565 observations. The estimation results of the diffusion parameters and the jump component have been given in Section 5.4. Using them we can calculate the estimated values $\hat{p}(Y_j), j = 1, 2, ..., n$. Then we obtain $M_n = 63.2134$, which is much larger than the 99% percentile (2.5738) of the distribution of $\sup_{0\leq s\leq 1}(W^o(s))^2$. Thus, we reject the null hypothesis that there is no change, and conclude that a change occurred in the jump frequency of the S&P 500 daily data from 1983 to 2012.

For the location of the change point, we notice that $M_n = 63.2134$ occurs when $k = 3929$ (see the definition of M_n in (7.21)). It corresponds to the date July 20, 1998.

Appendix

Proof of Proposition 7.1.3: In the proof of Proposition 7.1.1, we have showed

that as $\Delta \to 0$, $f_G = \phi_Z * f_X \to f_X$, $\phi_Z \to \delta_0$ where δ_0 is the Dirac delta function, and moreover,

$$\lim_{\Delta \to 0} \frac{(1-\alpha)\phi_Z(y)}{\alpha f_G(y)} = \begin{cases} 0, & \text{if} \quad y \neq 0; \\ \infty, & \text{if} \quad y = 0. \end{cases}$$

Fix a small Δ. Then there exist $a(\Delta) \leq 0$ and $b(\Delta) \geq 0$ such that

$$\begin{cases} (1-\alpha)\phi_Z(y) \geq \alpha f_G(y) & \text{if } a(\Delta) \leq y < b(\Delta) \\ (1-\alpha)\phi_Z(y) < \alpha f_G(y) & \text{if } y < a(\Delta) \text{ or } y \geq b(\Delta). \end{cases} \tag{7.23}$$

Now we prove the result for the Merton model. To find $a(\Delta)$ and $b(\Delta)$, we need to solve the equation $(1-\alpha)\phi_Z(y) = \alpha f_G(y)$. For the Merton model, the convolution $G = Z + X$ still follows a normal distribution, more specifically, $\mathcal{N}(\mu_0\Delta + \mu_X, \sigma_0^2\Delta + \sigma_X^2)$. Then the equation becomes

$$(1-\lambda\Delta)\frac{1}{\sqrt{2\pi\sigma_0^2\Delta}}\, e^{-\frac{(y-\mu_0\Delta)^2}{2\sigma_0^2\Delta}} = \lambda\Delta \cdot \frac{1}{\sqrt{2\pi(\sigma_0^2\Delta + \sigma_X^2)}}\, e^{-\frac{(y-\mu_0\Delta-\mu_X)^2}{2(\sigma_0^2\Delta+\sigma_X^2)}}$$

or equivalently,

$$\frac{(y-\mu_0\Delta)^2}{2\sigma_0^2\Delta} - \frac{(y-\mu_0\Delta-\mu_X)^2}{2(\sigma_0^2\Delta+\sigma_X^2)} = \log\left(\frac{1-\lambda\Delta}{\lambda\Delta}\sqrt{\frac{\sigma_0^2\Delta+\sigma_X^2}{\sigma_0^2\Delta}}\right).$$

Thus, we obtain the two solutions as

$$y_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

where $a = \sigma_X^2$, $b = 2\Delta(\mu_X\sigma_0^2 - \mu_0\sigma_X^2)$, and

$$c = 2\sigma_0^2\Delta(\sigma_0^2\Delta + \sigma_X^2)\log\left(\frac{1-\lambda\Delta}{\lambda\Delta}\sqrt{\frac{\sigma_0^2\Delta+\sigma_X^2}{\sigma_0^2\Delta}}\right) + \Delta(\mu_X^2\sigma_0^2 + 2\mu_0\mu_X\sigma_0^2\Delta - \mu_0^2\sigma_X^2\Delta).$$

From (7.23), we know that $a(\Delta) = y_1$ and $b(\Delta) = y_2$. As $\Delta \to 0$, we have $a(\Delta) \to 0$ and $b(\Delta) \to 0$. Moreover, using L'Hospital's rule for multiple times, we obtain

$$a(\Delta) \sim -\sqrt{3}\sigma_0\sqrt{\Delta \log\frac{1}{\Delta}}, \qquad b(\Delta) \sim \sqrt{3}\sigma_0\sqrt{\Delta \log\frac{1}{\Delta}}$$

where by "$\sim$" we mean the ratio of the two sides has a limit of one. Note that the only difference between $a(\Delta)$ and $b(\Delta)$ is the sign.

Using (7.23) again, the misclassification error in (7.17) can be written as

$$\begin{aligned}\epsilon^* &= \int_{a(\Delta)}^{b(\Delta)} \alpha f_G(y)dy + \left(\int_{-\infty}^{a(\Delta)} + \int_{b(\Delta)}^{\infty}\right)(1-\alpha)\phi_Z(y)dy \\ &:= I_1 + I_2.\end{aligned}$$

- For I_1: Since $a(\Delta) \to 0$ and $b(\Delta) \to 0$, we have $f_G(y)1_{\{a(\Delta)\le y<b(\Delta)\}} \to f_X(0)$ and $I_1 \to 0$. Using the facts that $f_G(y) \le \max_y f_X(y) < \infty$ and $b(\Delta)-a(\Delta) < \infty$, we apply the bounded convergence theorem to obtain

$$\begin{aligned} I_1 &\sim \lambda f_X(0) \cdot \Delta(b(\Delta) - a(\Delta)) \\ &\sim 2\sqrt{3}\sigma_0\lambda f_X(0) \cdot \sqrt{\Delta^3 \log\frac{1}{\Delta}}. \end{aligned} \tag{7.24}$$

  Note that $f_X(0) \neq 0$ in the Merton model.

- For I_2: we have

$$\begin{aligned} I_2 &= \left(\int_{-\infty}^{a(\Delta)} + \int_{b(\Delta)}^{\infty}\right)(1-\alpha)\phi_Z(y)dy \\ &= (1-\lambda\Delta)\left(\int_{-\infty}^{a(\Delta)} + \int_{b(\Delta)}^{\infty}\right)\frac{1}{\sqrt{2\pi\sigma_0^2\Delta}} e^{-\frac{(y-\mu_0\Delta)^2}{2\sigma_0^2\Delta}} dy \\ &= (1-\lambda\Delta)\left(\int_{-\infty}^{\frac{a(\Delta)-\mu_0\Delta}{\sigma_0\sqrt{\Delta}}} + \int_{\frac{b(\Delta)-\mu_0\Delta}{\sigma_0\sqrt{\Delta}}}^{\infty}\right)\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \\ &= (1-\lambda\Delta)\left[\Phi\left(\frac{a(\Delta)-\mu_0\Delta}{\sigma_0\sqrt{\Delta}}\right) + \Phi\left(\frac{-b(\Delta)+\mu_0\Delta}{\sigma_0\sqrt{\Delta}}\right)\right]. \end{aligned}$$

  Since $a(\Delta)/\sqrt{\Delta} \to -\infty$ and $b(\Delta)/\sqrt{\Delta} \to \infty$, we have $I_2 \to 0$. Moreover,

$$I_2 \sim 2\Phi\left(-\frac{b(\Delta)}{\sigma_0\sqrt{\Delta}}\right) \sim 2\Phi\left(-\sqrt{3\log\frac{1}{\Delta}}\right). \tag{7.25}$$

To identify the convergence rate of ϵ^* going to zero, we need to compare the convergence rates of I_1 and I_2 going to zero. The slower one would dominate the convergence rate of ϵ^*. By the L'Hospital's rule,

$$\begin{aligned}
\lim_{\Delta\to 0}\frac{I_2}{I_1} &= \lim_{\Delta\to 0}\frac{2\Phi\left(-\sqrt{3\log\frac{1}{\Delta}}\right)}{2\sqrt{3}\sigma_0\lambda f_X(0)\cdot\sqrt{\Delta^3\log\frac{1}{\Delta}}} \\
&= \frac{1}{\sqrt{3}\sigma_0\lambda f_X(0)}\cdot\lim_{\Delta\to 0}\frac{\Phi\left(-\sqrt{3\log\frac{1}{\Delta}}\right)}{\sqrt{\Delta^3\log\frac{1}{\Delta}}} \\
&= \frac{1}{\sqrt{3}\sigma_0\lambda f_X(0)}\cdot\lim_{\Delta\to 0}\frac{\phi\left(-\sqrt{3\log\frac{1}{\Delta}}\right)\cdot\frac{d}{d\Delta}\left(-\sqrt{3\log\frac{1}{\Delta}}\right)}{\frac{d}{d\Delta}\sqrt{\Delta^3\log\frac{1}{\Delta}}} \\
&= \frac{1}{\sqrt{3}\sigma_0\lambda f_X(0)}\cdot\lim_{\Delta\to 0}\frac{\frac{1}{\sqrt{2\pi}}\frac{\sqrt{3}}{2}}{\frac{3}{2}\log\frac{1}{\Delta}-\frac{1}{2}} \\
&= 0.
\end{aligned}$$

That is, I_1 converges to zero at a slower rate than I_2 does. Thus, the convergence rate of ϵ^* is the same as the convergence rate of I_1. Using (7.24) we obtain

$$\epsilon^* \sim 2\sqrt{3}\sigma_0\lambda f_X(0)\cdot\sqrt{\Delta^3\log\frac{1}{\Delta}}.$$

Equivalently,

$$\frac{\epsilon^*}{\Delta} \sim 2\sqrt{3}\sigma_0\lambda f_X(0)\cdot\sqrt{\Delta\log\frac{1}{\Delta}} \tag{7.26}$$

which goes to zero as $\Delta\to 0$. This proves (7.18).

For the Kou model, the proof is similar to the one for the Merton case. ⊠

Chapter 8

Future Work

In this chapter, we present some research topics that are natural extensions to what we have investigated so far.

8.1 Extension to Infinite-Activity Lévy Processes

The jump-diffusion processes we have considered in the thesis are Lévy processes with finite activity, i.e. there are finite number of jumps in any finite time interval. We may consider an extension of our method to an infinite-activity Lévy process, for which there are infinite number of jumps in any finite time interval that includes zero.

Let us look at the jump component of the infinite-activity Lévy processes. We may define some threshold, ϵ, to split jumps into "small" and "large" ones. As we know, there are infinitely many "small" jumps in any finite interval. These "small" jumps may play a role similar to the one that the Brownian motion does in a jump-diffusion model. The remaining "large" jumps can be viewed as a compound

Poisson process. Thus, the jump component of an infinite-activity Lévy process can be approximated by a jump-diffusion model.

The Lévy-Khintchine formula stated in Theorem 1.1.2 can be rewritten as follows (basically, replacing one by ϵ).

Theorem 8.1.1 (Lévy-Khintchine)**.** *Let L be a Lévy process with Lévy measure v. Then*

$$\mathbb{E}\big[e^{iuL_t}\big] = e^{t\psi(u)},$$

where

$$\psi(u) = iau - \frac{1}{2}\sigma^2 u^2 + \int_{\{|x|\geq\epsilon\}} (e^{iux} - 1)\, v(dx) + \int_{\{|x|<\epsilon\}} (e^{iux} - 1 - iux)\, v(dx) \quad (8.1)$$

for any $\epsilon > 0$. Moreover, given a, σ^2, v, the corresponding Lévy process is unique in distribution.

From this result, we see that a Lévy process can be decomposed into three parts, the Brownian motion, the "large" jumps and the "small" jumps. For a Lévy process without the Brownian motion component, we have $\sigma^2 = 0$.

We would consider applying the proposed method used for the estimation of the jump-diffusion model to estimate the Lévy measure of size larger than ϵ for an infinite-activity Lévy process. As the data frequency $\Delta \to 0$, we can let the threshold $\epsilon \to 0$ so as to obtain a complete picture of the Lévy measure.

8.1.1 Approximation of Infinite-Activity Lévy Processes

In the following we show that the component of the "small" jumps can be approximated by another Brownian motion. Rydberg (1997) arrived at the same result but with the emphasis on the normal inverse Gaussian processes.

Let us work with the last term in equation (8.1). By applying the Taylor expansion $e^{iux} = 1 + iux - \frac{u^2x^2}{2} - \frac{iu^3x^3}{3!} + O(x^4)$, we obtain

$$\int_{x\in(-\epsilon,\epsilon)} (e^{iux} - 1 - iux)v(dx) = \int_{x\in(-\epsilon,\epsilon)} \left(-\frac{u^2x^2}{2} - \frac{iu^3x^3}{3!} + O(x^4)\right) v(dx).$$

Since $\int(1 \wedge |x|^2)v(dx) < \infty$ holds for any Lévy process (see Theorem 1.1.2) and 1 can be replace by any fixed positive number, we have

$$\delta_\epsilon^2 := \int_{x\in(-\epsilon,\epsilon)} x^2 v(dx) < \infty. \tag{8.2}$$

Then the integral of the first term becomes

$$\int_{x\in(-\epsilon,\epsilon)} \left(-\frac{u^2x^2}{2}\right) v(dx) = -\frac{\delta_\epsilon^2 u^2}{2}.$$

The second term satisfies

$$\begin{aligned}\left|\int_{x\in(-\epsilon,\epsilon)} \left(-\frac{iu^3x^3}{3!}\right) v(dx)\right| &\le \left|\frac{iu\epsilon}{3}\int_{x\in(-\epsilon,\epsilon)} \left(\frac{u^2x^2}{2}\right) v(dx)\right| \\ &= \left|\frac{iu\epsilon}{3}\cdot\frac{\delta_\epsilon^2 u^2}{2}\right| = \frac{\delta_\epsilon^2 u^2}{2}\cdot O(\epsilon),\end{aligned}$$

and the remaining term also satisfies

$$\int_{x\in(-\epsilon,\epsilon)} O(x^4)v(dx) = \frac{\delta_\epsilon^2 u^2}{2}\cdot O(\epsilon^2).$$

Therefore,

$$\int_{x\in(-\epsilon,\epsilon)} (e^{iux} - 1 - iux)v(dx) = -\frac{\delta_\epsilon^2 u^2}{2}\left(1 + O(\epsilon)\right). \tag{8.3}$$

Thus, the "small" jump component can be approximated by a Brownian motion with drift zero and volatility δ_ϵ. Therefore, we obtain the following result.

Proposition 8.1.2. *Let L be a Lévy process with characteristic exponent $\psi(u)$ in (8.1). For a small $\epsilon > 0$, L can be approximated by $\tilde{L}$ with*

$$\mathbb{E}\left[e^{iu\tilde{L}_t}\right] = e^{t\tilde{\psi}(u)},$$

where

$$\tilde{\psi}(u) = iau - \frac{1}{2}(\sigma^2 + \delta_\epsilon^2)u^2 + \int_{\{|x|\geq\epsilon\}} (e^{iux} - 1)\, v(dx), \tag{8.4}$$

and the difference between the characteristic exponents of L and $\tilde{L}$ is

$$\psi(u) - \tilde{\psi}(u) = \delta_\epsilon^2 u^2 \cdot O(\epsilon).$$

This result states that a general Lévy process with Lévy triplet $(a, \sigma^2, v(dx))$ can be approximated by a jump-diffusion model, where the diffusion component is a Brownian motion with drift a and volatility $\sigma^2 + \delta_\epsilon^2$, and the jump component is a compound Poisson process with the measure $v(x)1_{\{|x|>\epsilon\}}$.

8.1.2 Estimation of Infinite-Activity Lévy Processes

After we approximate the infinite-activity Lévy process by a jump-diffusion model, we would like to apply the method used for estimation of the jump-diffusion model to estimate the infinite-activity Lévy process. However, for a fixed data frequency, we can only estimate the Lévy measure with jump size larger than some threshold ϵ (at a given accuracy). Fortunately, as the data frequency $\Delta \to 0$, we can make the threshold $\epsilon \to 0$, and thus obtain a complete picture of the Lévy measure. This is the basic idea for an extension of our method to the infinite-activity Lévy processes.

Let L_t denote a Lévy process with measure v. For $\epsilon > 0$, we define the *jump frequency with size above* ϵ by

$$\lambda_\epsilon := \mathbb{E}\left[\sum_{t\in[0,1]} 1_{\{|\Delta L_t|>\epsilon\}}\right]$$

when the Levy measure v is an even function. Otherwise, we may define

$$\lambda_\epsilon^+ := \mathbb{E}\left[\sum_{t\in[0,1]} 1_{\{\Delta L_t>\epsilon\}}\right], \quad \lambda_\epsilon^- := \mathbb{E}\left[\sum_{t\in[0,1]} 1_{\{\Delta L_t<-\epsilon\}}\right].$$

In the following, without loss of generality, we consider only the case where v is an even function. From the definition, we can see that λ_ϵ is the expected number of jumps with size larger than ϵ per unit time for the Lévy process, i.e. $\lambda_\epsilon = \nu((-\infty, -\epsilon] \cup [\epsilon, \infty))$. Our purpose is to estimate λ_ϵ for a given ϵ. When we have λ_ϵ for every $\epsilon > 0$, we obtain the Levy measure v of L_t.

For a given data frequency Δ, a practical problem is how to select the threshold ϵ. On one hand, the threshold ϵ of the jump size needs to be small enough so that the normal approximation in (8.3) for small jumps can be accurate; on the other hand, the threshold ϵ cannot be too small because otherwise the assumption (A1-2) that $\lambda_\epsilon \Delta$ is small will be violated.

This is an interesting research topic and worth further investigations.

8.2 Other Directions

In Chapter 4, we used some robust procedures to estimate the diffusion parameters. These robust procedures can be further improved by some *adaptive* methods. For example, in the Huber estimator, instead of assigning a value for the parameter k in the beginning, we could use some adaptive way to select this parameter based on the available data.

In Chapter 5, we proved the consistency of the estimator of the density function of jump size, when the diffusion parameters and the jump frequency are assumed known. However, it is desirable to obtain a more general result for which the parameters are assumed unknown. Moreover, we may consider identifying the convergence rate of this estimator. For this, probably we need consider two different types of jump size distributions: super smooth ones, and ordinary smooth ones. Some existing results in the context of de-convolution problems may be of help.

When applying the trapezoidal rule to recover the density function of jump size,

instead of choosing equally spaced points, we may choose points such that they are more dense when t is closer to zero. Because the estimated characteristic function is more accurate when t is closer to zero, such an approach may improve the accuracy of our estimator.

For the estimation of the jump frequency, we only used the information of the real part of the characteristic functions. We are always curious whether the imaginary part could be used in any way for improvement.

Appendices

Appendix A

Review on Robust Statistics

All statistical procedures rely explicitly or implicitly on a number of assumptions. For example, the normality assumption in the least square estimator, F-test and t-test. However, there is no guarantee that these assumptions would be exactly true. Based on this consideration, alternative *robust* procedures have been proposed. By *robustness*, we mean the insensitivity to small deviations from assumptions (see Huber 1981, page 1).

The following are basically the study notes by reading the books by Huber (1981), Jurečková and Picek (2006), and others (see the Bibliography).

A.1 Mathematical Tools

Mathematically, we consider a robust statistics as a statistical functional defined on the space of distribution functions. We are interested in its behavior in a neighborhood of a specific distribution or a model.

Suppose we have a probability space $(\Omega, \mathcal{B}, P)$, where Ω is the sample space, $\mathcal{B}$ is a σ-algebra over Ω (i.e. a collection of events which are closed under the

complement and countable unions of its elements; by "an event" we mean a subset of the sample space) and P is a probability measure (i.e. a set function such that $P : \mathcal{B} \to [0, 1]$; by "set function" we mean a function whose domain is a set).

A *statistical model* is a family $\mathcal{P}$ of probability measures or probability distribution functions, to which our specific distribution P belongs. A *parametric statistical model* is a family $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ of probability measures, each of which is indexed by a unique finite-dimensional *parameter* θ. The parameter θ can take on any value in the *parametric space* $\Theta \subseteq \mathbb{R}^p$. In the one-dimensional case, $p = 1$ and θ is a real number. The triplet $\{\Omega, \mathcal{B}, P_\theta : \theta \in \Theta\}$ denotes a parametric statistical model.

Consider a random variable X with probability distribution P_θ and distribution function F, where $P_\theta \in \mathcal{P}$. The parameter θ, typically viewed as an index for the family $\mathcal{P}$ of distributions, has another interpretation. It can be viewed as a statistical functional $\theta \equiv T(P) : \mathcal{P} \to \mathbb{R}$. Sometimes we also write $\theta \equiv T(F)$. Note that a *statistical functional* is a functional defined on a set of distribution functions; and a *functional* is a function whose domain is a set of functions. The following are some examples:

1. Expected value: $\mu = T(P) = \int_{\mathbb{R}} x dP = \mathbb{E}(X)$
2. Variance: $\sigma^2 = T(P) = \int_{\mathbb{R}} x^2 dP - (\mathbb{E}(X))^2$
3. Median: $\mathrm{m} = T(P) = F^{-1}\left(\frac{1}{2}\right)$

Given a random sample $X_1, X_2, \cdots, X_n$ from the distribution P_θ, the *empirical probability distribution* is defined by

$$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} I[X_i \in A], \quad A \in \mathcal{B}. \tag{A.1}$$

Note that P_n can be viewed as a uniform distribution on the set $\{X_1, X_2, \cdots, X_n\}$. The *empirical distribution function* is defined by

$$F_n(x) = P_n((-\infty, x]) = \frac{1}{n} \sum_{i=1}^{n} I[X_i \leq x], \quad x \in \mathbb{R}. \tag{A.2}$$

Then a natural estimator of the parameter θ is the "plug-in" estimator $T(P_n)$. Examples are given as follows:

1. Expected value: $\hat{\mu} = T(P_n) = \int_{\mathbb{R}} x dP_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
2. Variance: $\hat{\sigma}^2 = T(P_n) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2 = S_n^2$
3. Median: $\hat{\mathrm{m}} = T(P_n) = F_n^{-1}\left(\frac{1}{2}\right)$

We say that an estimator (or a statistical functional) $\hat{\theta}_n = T(P_n)$ is a *Fisher consistent* estimator of parameter θ if it satisfies $T(P) = \theta$. From the robustness point of view, the natural property of Fisher consistency of an estimator is more important than its unbiasedness (i.e. $\mathbb{E}(\hat{\theta}_n) = \theta$).

A good estimator $T(P_n)$ should tend to $T(P)$ as $n \to \infty$, with respect to some type of convergence defined on the space of probability measures. Examples are convergence in probability, convergence in distribution, almost sure convergence and convergence in the mean which corresponds to the large sample bias of an estimator. If $T(P_n)$ converges to $T(P)$ in probability, then we say $T(P_n)$ is *(weakly) consistent*. If $T(P_n)$ converges to $T(P)$ almost surely,then we say $T(P_n)$ is *strongly consistent*. Notice that neither the definition of Fisher consistency or the definition of consistency encompasses the other; the Fisher consistency can describe the finite-sample property while the consistency cares only the limiting or asymptotic behavior.

Since we are interested in studying the behavior of $T(P_n)$ in a neighborhood of P, we consider an expansion of the functional $T(P_n) - T(P)$ of the Taylor type. To do this we need some concepts, such as the distances between probability measures, the continuity and the differentiability of functional T with respect to the considered distance.

The following are some examples of the most widely used distances of probability measures. Let $\mathcal{X}$ be a metric space with metric d, separable and complete, and $\mathcal{B}$

be the σ-field of its Borel subsets. Let $\mathcal{P}$ be the system of all probability measures on the space $(\mathcal{X}, \mathcal{B})$, with two elements P and Q.

1. *Prochorov distance*:

$$d_P(P,Q) = \inf\{\epsilon > 0 : P(A) \leq Q(A^\epsilon) + \epsilon, \forall A \in \mathcal{B}, A \neq \emptyset\}$$

where $A^\epsilon = \{x \in \mathcal{X} : \inf_{y \in A} d(x,y) \leq \epsilon\}$ is a closed ϵ-neighborhood of a non-empty set A.

2. *Lévy distance*: $\mathcal{X} = \mathbb{R}$ is the real line; let F, G be the distribution functions of probability measures P, Q, then

$$d_L(F,G) = \inf\{\epsilon > 0 : F(x-\epsilon) - \epsilon \leq G(x) \leq F(x+\epsilon) + \epsilon, \forall x \in \mathbb{R}\}.$$

3. *Kolmogorov distance*: $\mathcal{X} = \mathbb{R}$ is the real line and F, G are the distribution functions of probability measures P, Q, then

$$d_K(F,G) = \sup_{x \in \mathbb{R}} |F(x) - G(x)|.$$

4. *Total variation distance*:

$$\begin{aligned} d_{TV}(P,Q) &= \sup_{A \in \mathcal{B}} |P(A) - Q(A)| \\ &= \frac{1}{2} \int_{\mathcal{X}} |f - g| d\mu \end{aligned}$$

where the last line holds if P and Q are absolutely continuous w.r.t. μ with the densities f and g.

5. *Hellinger distance*:

$$d_H(P,Q) = \left\{ \int_{\mathcal{X}} \left(\sqrt{dP} - \sqrt{dQ} \right)^2 \right\}^{1/2}.$$

If $f = \frac{dP}{d\mu}$ and $g = \frac{dQ}{d\mu}$ are densities of P and Q with respect to some measure μ, then the Hellinger distance can be rewritten as

$$d_H(P,Q) = \left\{ \int_{\mathcal{X}} \left(\sqrt{f} - \sqrt{g} \right)^2 d\mu \right\}^{1/2} = \left\{ 2 \left(1 - \int_{\mathcal{X}} \sqrt{fg} d\mu \right) \right\}^{1/2}.$$

6. *Lipschitz distance*: Assume that $d(x,y) \leq 1, \forall x, y \in \mathcal{X}$ (we take the metric $d' = \frac{d}{1+d}$ otherwise), then

$$d_{Li}(P,Q) = \sup_{\psi \in \mathcal{L}} \left| \int_{\mathcal{X}} \psi dP - \int_{\mathcal{X}} \psi dQ \right|$$

where $\mathcal{L} = \{\psi : \mathcal{X} \to \mathbb{R} : |\psi(x) - \psi(y)| \leq d(x,y)\}$ is the set of the Lipschitz functions.

7. *Kullback-Leibler divergence*: Let p, q be the densities of probability distributions P, Q with respect to measure μ (Lebesgue measure on the real line or the counting measure), then

$$d_{KL}(Q,P) = \int q(x) \ln \frac{q(x)}{p(x)} d\mu(x)$$

The Kullback-Leibler divergence is not a metric, because it is not symmetric in P, Q and does not satisfy the triangle inequality.

Two commonly discussed derivatives of statistical functions are the *Gâteau derivative* and the *Fréchet derivative*. Their definitions are given in the below. More details on these mathematical tools for robust statistics can be found in Huber (1981).

Definition A.1.1. We say that functional T *is differentiable in the Gâuteau sense in* P *in direction* Q, if there exists the limit

$$T'_Q(P) = \lim_{t \to 0+} \frac{T(P + t(Q-P)) - T(P)}{t}.$$

$T'_Q(P)$ is called the *Gâuteau derivative of functional* T *in* P *in direction* Q.

Remark A.1.2. If we denote $\psi(t) := T((1-t)P + tQ), 0 \leq t \leq 1$, then the Gâuteau derivative is equal to the ordinary right derivative of function ψ at point 0, i.e.

$$T'_Q(P) = \psi'(0+).$$

The Gâuteau derivative of order k is defined similarly, i.e.

$$T_Q^{(k)}(P) = \psi^{(k)}(0+).$$

Some Examples are given as follows:

1. Expected value: $T'_Q(P) = \mathbb{E}_Q X - \mathbb{E}_P X$
2. Variance: $T'_Q(P) = \mathbb{E}_Q X^2 - \mathbb{E}_P X^2 - 2\mathbb{E}_P X \cdot \mathbb{E}_Q X + 2(\mathbb{E}_P X)^2$

Definition A.1.3. We say that functional T *is differentiable in* P *in the Fréchet sense*, if there exists a linear functional $L_P(Q-P)$ such that

$$\lim_{t\to 0} \frac{T(P+t(Q-P)) - T(P)}{t} = L_P(Q-P)$$

uniformly in $Q \in \mathcal{P}, \delta(P,Q) \leq C$ for any fixed $C \in (0, +\infty)$. The lineal functional $L_P(Q-P)$ is called the *Fréchet derivative of functional* T *in* P *in direction* Q.

Remark A.1.4. If T is differentiable in the Fréchet sense, then it is differentiable in the Gâuteau sense, too, i.e. there exists $T'_Q(P), \forall Q \in \mathcal{P}$ and it holds that

$$T'_Q(P) = L_P(Q-P), \forall Q \in \mathcal{P}.$$

The following theorem provides the *asymptotic normality* property of the estimators as statistical functionals. Let δ_x denote the dirac probability measure that assigns probability one to a single point x.

Theorem A.1.5. *Let* T *be a statistical functional, Fréchet differentiable in* P*, and assume that the empirical probability distribution* P_n *satisfies* $\sqrt{n} \cdot d(P_n, P) = O_p(1)$ *as* $n \to \infty$ *for some distance* d*. Then the sequence* $\sqrt{n}(T(P_n) - T(P))$ *is asymptotically normally distributed as* $n \to \infty$*; more specifically,*

$$\sqrt{n}(T(P_n) - T(P)) \longrightarrow N(0, Var_P T'_{\delta_{X_1}}(P)) \quad \textit{in distribution.} \tag{A.3}$$

One example is that the variance estimator: $\sqrt{n}(S_n^2 - \sigma^2) \longrightarrow N(0, \mu_4 - \mu_2^2)$.

Remark A.1.6. If the metric d is, e.g. the Kolmogorov distance, then it automatically satisfies that

$$\sqrt{n} \cdot d(P_n, P) = O_p(1), \quad \text{as } n \to \infty$$

or equivalently,

$$\sqrt{n} \max_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| = O_p(1), \quad \text{as } n \to \infty$$

due to the famous result that

$$\sqrt{n} \max_{0 \le x \le 1} |\hat{F}_n(x) - F(x)| \to \max_{0 \le x \le 1} |B^0(x)| \qquad \text{in distribution.}$$

The scheme to prove the above result is as follows: It is well-known that $\max_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|$ has the distribution-free property, i.e. it has the same distribution for any choice of the continuous distribution function F. Then we may choose the Unif$(0, 1)$ distribution, and it can be shown that $\sqrt{n}(\hat{F}_n(x) - F(x)) \to B^0(x)$ in distribution, where B^0 is a Brownian bridge. Thus we can prove the above result. Additionally, we can use the following formula to identify the exact probability:

$$\mathbb{P}\left\{ \max_{0 \le t \le 1} |B^0(t)| > x \right\} = 2 \sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2 x^2}.$$

A.2 Robustness Criterion

In the following we first introduce the definition of qualitative robustness, and then discuss several different measures of quantitative robustness.

Definition A.2.1. We say that the sequence of statistics (i.e. empirical functionals) $\{T_n\}$ is *qualitatively robust* for probability distribution P, if for any $\epsilon > 0$ there exists a $\delta > 0$ and a positive integer n_0 such that, for all $Q \in P$ and $n \ge n_0$,

$$d(P, Q) < \delta \quad \Longrightarrow \quad d(\mathcal{L}_P(T_n), \mathcal{L}_Q(T_n)) < \epsilon \tag{A.4}$$

where $\mathcal{L}_P(T_n)$ and $\mathcal{L}_Q(T_n)$ denote the probability distributions of T_n under P and Q, respectively.

This definition was first formulated by Hampel (1968, 1971). It uses the idea of the continuity of functional $\{T_n\}$ in a neighborhood of P.

This robustness is only *qualitative*. However, we not only want to see whether the functional T is qualitatively robust or not, but also we want to compare different functionals to see which one is more robust. To do this, we must characterize the robustness with some *quantitative* measure. Different quantitative measures of robustness are introduced in the following.

Influence Function

The Influence Function is one of the most important characteristics of a statistical functional.

Definition A.2.2. The Gâteau derivative of a functional T in distribution P in the direction of the Dirac distribution δ_x is called the *influence function of* T *in* P, i.e.

$$IF(x;T,P) = \lim_{t\to 0+} \frac{T((1-t)P + t\delta_x) - T(P)}{t}. \tag{A.5}$$

Or equivalently,

$$IF(x;T,P) = T'_{\delta_x}(P).$$

The influence function $IF(x;T,P)$ measures the effect of an infinitesimal contamination of functional T by a single value x. Thus a robust functional T is expected to have a bounded influence function. Notice that even T is a qualitatively robust functional it does not guarantee its influence function is bounded.

The following are some properties of the influence function.

1. $\mathbb{E}_P[IF(x;T,P)] = 0$.

2. Suppose T is a Fréchet differentiable functional satisfying $\sqrt{n}\delta(P_n, P) = O_p(1)$, then from Theorem A.1.5 we have

$$\sqrt{n}(T(P_n) - T(P)) \to N(0, Var_P(IF(X_1; T, P))). \tag{A.6}$$

The following are some examples of the influence function.

Example A.2.3. 1. Expected value μ_P: $IF(x; T, P) = x - \mu_P$ (unbounded)

2. Variance σ_P^2: $IF(x; T, P) = (x - \mu_P)^2 - \sigma_P^2$ (unbounded)

3. Median m_P: $IF(x; T, P) = \frac{\text{sign}\,(x - m_P)}{2f(m_P)}$ (bounded)

The most popular quantitative characteristics of robustness based on the influence functions are the *global and local sensitivities*.

Definition A.2.4. The *global sensitivity* of the functional T under distribution P is the value

$$\gamma^* = \sup_{x \in \mathcal{X}} |IF(x; T, P)|. \tag{A.7}$$

Definition A.2.5. The *local sensitivity* of the functional T under distribution P is the value

$$\lambda^* = \sup_{x \neq y} \left| \frac{IF(y; T, P) - IF(x; T, P)}{y - x} \right|. \tag{A.8}$$

Example A.2.6. 1. Expected value: $\gamma^* = \infty, \lambda^* = 1$, which implies that the mean is not robust, but it is not sensitive to the local changes.

2. Variance: $\gamma^* = \infty, \lambda^* = \infty$, which implies that the variance is non-robust both to large and small (local) changes.

3. Median m_P: $\gamma^* = \frac{1}{2f(m_P)}$, $\lambda^* = \infty$, which implies that the median is robust (although it is sensitive to local changes).

Breakdown Point

The breakdown point, introduced by Donoho and Huber in 1983, is another popular quantitative characteristic of robustness. Roughly speaking, the *breakdown point* is the smallest proportion of contamination that may cause an estimator to take on arbitrarily large aberrant values, or that may lead the estimator T_n up to infinity. For example, the breakdown point is 0 for the mean and $\frac{1}{2}$ for the median.

Minimax Bias

Another natural measure of robustness of the functional T is its *maximum bias* defined as

$$b(\mathcal{F}_\epsilon) = \sup_{F \in \mathcal{F}_\epsilon} |T(F) - T(F_0)|, \tag{A.9}$$

where $\mathcal{F}_\epsilon$ is a neighborhood of a fixed distribution F_0. Suppose $\mathcal{F}_\epsilon$ is the widely used ϵ-*contaminated neighborhood*, i.e.

$$\mathcal{F}_\epsilon = \{F : F = (1-\epsilon)F_0 + \epsilon G, \quad G \text{ arbitrary}\}. \tag{A.10}$$

It is relatively easy to find the maximal bias when the statistical functional is *monotone* with respect to the *stochastic ordering* of distribution functions. (The definition of stochastic ordering is as follows: Random variable X with distribution function F is *stochastically smaller* than random variable Y with distribution G if $F(x) \geq G(x), \forall x \in \mathbb{R}$.) The mean and median are examples of monotone statistical functionals. It is obvious that the monotone statistical functional thus attains the maximum bias by

$$b(\mathcal{F}_\epsilon) = \max\left(|T(F_\infty) - T(F_0)|, |T(F_{-\infty}) - T(F_0)|\right), \tag{A.11}$$

where $F_\infty = (1-\epsilon)F_0 + \epsilon\delta_\infty$ and $F_{-\infty} = (1-\epsilon)F_0 + \epsilon\delta_{-\infty}$. Based on this fact, it can be shown that the *median* achieves the smallest maximum bias among all *translation equi-variant* functionals over $\mathcal{F}_\epsilon$. (Note that a functional $T(\cdot)$ is called

translation equi-variant if $T(X_1+c,\cdots,X_n+c)=T(X_1,\cdots,X_n)+c$ for any $c\in\mathbb{R}$. This is a desired property for an estimator of the location parameter.)

Minimaximal Robustness

Another criterion is called the *minimaximal robustness*. The maximum asymptotic variance over a specified class $\mathcal{F}_\epsilon$ of distribution functions can be considered as a measure of robustness of the functional T. Then we look for a functional T_0 which minimizes the maximum asymptotic variance.

Definition A.2.7. Suppose T_n is asymptotically normally distributed, i.e. as $n\to\infty$,

$$\sqrt{n}(T(F_n)-T(F))\to N(0,\sigma^2(F,T)), \tag{A.12}$$

then the estimator

$$T_0=\arg\inf_{T\in\mathcal{T}}\sup_{F\in\mathcal{F}_\epsilon}\sigma^2(F,T) \tag{A.13}$$

where $\mathcal{T}$ is a class of estimators of parameter θ and $\mathcal{F}_\epsilon$ is a neighborhood of the assumed model, is called the *minimaximally robust estimator* in the class $\mathcal{T}$ and for the neighborhood $\mathcal{F}_\epsilon$.

The obtained F_0 is called the *least favorable distribution*, which is the distribution in $\mathcal{F}_\epsilon$ with the smallest Fisher information. Then we have $\sigma^2(F_0,T)\geq\sigma^2(F_0,T_0)\geq\sigma^2(F,T_0)$, i.e. (F_0,T_0) is the saddle point for the asymptotic variance $\sigma^2(F,T)$. In the context of robustness analysis, two widely-used classes of $\mathcal{F}_\epsilon$ are given as follows:

1. ϵ-contamination neighborhood (or gross-error neighborhood):

$$\mathcal{F}_\epsilon=\{F:F=(1-\epsilon)G+\epsilon H,\quad H\text{ arbitrary}\} \tag{A.14}$$

2. Kolmogorov neighborhood:

$$\mathcal{F}_\epsilon = \{F : \sup_{x\in\mathbb{R}} |F(x) - G(x)| \leq \epsilon\}. \tag{A.15}$$

Huber (1964, 1981) solved the problem for both the ϵ-contamination neighborhood with G arbitrary but satisfying some regularity conditions and the Kolmogorov neighborhood with $G = \Phi$. Specifically, the estimator obtained for the family of *ϵ-contaminated normal distributions* is called the **Huber estimator** and it has the form of (A.23) given below. The corresponding least favorable distribution is of the density

$$f_0(x) = \begin{cases} (1-\epsilon)\dfrac{1}{\sqrt{2\pi}}\, e^{-x^2/2} & |x| \leq k, \\ (1-\epsilon)\dfrac{1}{\sqrt{2\pi}}\, e^{k^2/2-k|x|} & |x| > k \end{cases} \tag{A.16}$$

i.e. it is normal in interval $[-k, k]$ and exponential outside.

Finite Sample Minimax Property

There is another *finite sample minimax property* of the Huber estimator (Huber (1968)). Consider the Kolmogorov ϵ-neighborhood of the standard normal distribution, i.e.

$$\mathcal{F} = \{F : \sup_{x\in\mathbb{R}} |F(x) - \Phi(x)| \leq \epsilon\} \tag{A.17}$$

For a fixed $a > 0$, define the inaccuracy measure of an estimator T of θ by

$$\sup_{F\in\mathcal{F},\theta\in\mathbb{R}} P_\theta\left(|T - \theta| > a\right) \tag{A.18}$$

Let T_H be a slightly modified version of the Huber estimator, which depends on ϵ and a. Then T_H minimizes the inaccuracy (A.18) in the family of translation equivariant estimators of θ.

Tail-Behavior Measure

Jurecková (1981) introduced the following *tail-behavior measure* of an equivariant estimator T_n:

$$B(T_n; a) = \frac{-\ln P_\theta(|T_n - \theta| > a)}{-\ln(1 - F(a))}, \quad a > 0 \tag{A.19}$$

The value $B(T_n; a)$ show how many times faster the probability $P_\theta(|T_n - \theta| > a)$ tends to 0 than $1 - F(a)$, as $a \to \infty$. The larger the value $B(T_n; a)$, the better the estimator T_n.

Summary on Measures of Robustness

It is desirable for a robust estimator to have a bounded influence function, small global and local sensitivities, a breakdown point of 1/2, a minimax bias, asymptotically minimax robustness, finite sample minimax robustness and a large value of the tail-behavior measure $B(T_n; a)$. Of course, one estimator may not have all these properties. We need to find a suitable robust estimator based on the specific problem.

A.3 Type of Estimators

In the literature of robust statistics, the M-estimator, L-estimator and R-estimator are three popular types of robust estimators. Recently the *adaptive* combinations of estimators have drawn attention and the idea behind it is simple and intuitive. In the following we briefly review each of these estimators.

M-Estimator

An *M-estimator* T_n is defined as a solution of the minimization problem

$$\sum_{i=1}^{n} \rho(X_i, \theta) := min, \quad \theta \in \Theta$$

where $\rho(\cdot, \cdot)$ is a properly chosen function. If ρ is differentiable in θ with a continuous derivative $\psi(\cdot, \theta) = \frac{\partial}{\partial \theta}\rho(\cdot, \theta)$, then T_n is a root (or one of the roots) of the equation

$$\sum_{i=1}^{n} \psi(X_i, \theta) = 0, \quad \theta \in \Theta. \tag{A.20}$$

Corresponding, the functional $T(F)$ is defined by

$$\int \psi(x, T(F))dF(x) = 0.$$

Some properties of M-estimators are given as follows.

1. The influence function is

$$IF(x; T, F) = \frac{\psi(x, T(F))}{-\int \psi'(y, T(F))dF(y)}$$

   where $\psi'(y, T(F)) = [\frac{\partial}{\partial \theta}\psi(y, \theta)]|_{\theta=T(F)}$.

2. If $T(F)$ is uniquely defined, then T_n is consistent at F: $T_n \to T(F)$ in probability and almost surely.

3. Under some regularity conditions, we have that

$$\sqrt{n}(T_n - T(F)) \to N\left(0, \frac{E_F[\psi^2(x, T(F))]}{(\int \psi'(y, T(F))dF(y))^2}\right).$$

M-Estimator of location parameter

If we are particularly interested in a **location parameter** θ, then we solve

$$\sum_{i=1}^{n} \rho(X_i - \theta) := min, \tag{A.21}$$

or

$$\sum_{i=1}^{n} \psi(X_i - \theta) = 0.$$

The influence function is

$$IF(x; T, F) = \frac{\psi(x - T(F))}{\int \psi'(y) dF(y)}. \tag{A.22}$$

The breakdown point is 0 if $\psi(\cdot)$ is an unbounded function, and $\frac{1}{2}$ if $\psi(\cdot)$ is odd and bounded. The following are two examples of the M-estimator of the location parameter:

- Mean: $\rho(x) = x^2$, $\psi(x) = 2x$
- Median: $\rho(x) = |x|$, $\psi(x) = \text{ sign }(x)$.

Choice of ψ function

If we look for an M-estimator of the location parameter of a distribution belonging to the family of *ϵ-contaminated normal distributions*:

$$\mathcal{F}_\epsilon = \{F : F = (1 - \epsilon)\Phi + \epsilon H, \quad H \text{ symmetric}\},$$

where H runs over symmetric distribution functions, we should use the **Huber estimator** (proposed by Huber (1964)):

$$\rho_H(x) = \begin{cases} x^2 & \text{if } |x| \leq k \\ 2k|x| - k^2 & \text{if } |x| > k \end{cases}$$

and

$$\psi_H(x) = \begin{cases} x & |x| \leq k, \\ k \cdot \text{sign}(x) & |x| > k \end{cases} \tag{A.23}$$

where $k > 0$ is a fixed constant, depending on ϵ though the following identity:

$$2\Phi(k) - 1 + \frac{2\Phi'(k)}{k} = \frac{1}{1 - \epsilon}.$$

The Huber estimator is very popular in the context of robust estimation. It has a bounded influence function proportional to ψ_H, the breakdown point $\epsilon = \frac{1}{2}$, the global sensitivity $\gamma = \frac{k}{2F(k)-1}$, and the tail-behavior measure $\lim_{a\to\infty} B(a, T_n, F) = \frac{1}{2}$ both for distributions with exponential and heavy tails. Thus it is a robust estimator of the center of symmetry, insensitive to the extreme observations. As Huber proved in 1964, this estimator is minimaximally robust for a contaminated normal distribution. An interesting and natural question is whether there exists a distribution F such that the Huber estiamtor is the maximum likelihood estimator of θ for $F(x-\theta)$, i.e. such that ψ_H is the likelihood function for F. Such a distribution really exists, and its density is normal in the interval $[-k, k]$, and exponential outside. See equation (A.16).

The ψ- function for the Huber estimator is monotone, and thus the Huber estimator is one of the so-called *monotone M-estimators*. There is another class of M-estimators, called the **redescending M-estimators**, whose ψ- functions vanish outside some interval. It has been found that redescending estimators sometimes outperform the Huber estimator. For example, the redescending M-estimators are more efficient than the Huber estimator for some symmetric and heavy-tailed distributions. This is because redescending estimators completely reject gross outliers, while the Huber estimator treats them the same as moderate outliers. Also, as an example, the MLE for the student t distribution corresponds to a redescending ψ-function. Examples of the redescending ψ-functions include the *Hample function*, the *Andrews sinus function*, the *Tukey bisquare or biweight function* and the *Cauchy function*.

Studentized M-estimators

We see from equation (A.21) that the M-estimator of a location parameter is *translation equivariant*:

$$T_n(X_1 + c, \cdots, X_n + c) = T_n(X_1, \cdots, X_n) + c,$$

but generally not *scale equivariant*:

$$T_n(cX_1, \cdots, cX_n) = cT_n(X_1, \cdots, X_n).$$

This can be overcome by *studentizing the M-estimators* by a scale statistics S_n and solving the following equation

$$\sum_{i=1}^{n} \psi\left(\frac{X_i - \theta}{S_n}\right) = 0, \quad \theta \in \mathbb{R}. \tag{A.24}$$

To guarantee the translation and scale equivariance of the solution to (A.24), the scale statistics S_n should satisfy the *translation invariance* and the *scale equivariance*.

The M-estimator is the most important robust estimator discussed in the literature. Other types of robust estimators include the L-estimator and the R-estimator.

L-Estimator

An *L-estimator* is a linear combination of order statistics, i.e.

$$T_n = \sum_{i=1}^{n} a_{ni} h(X_{(i)}).$$

Examples are the *median*, *α-trimmed mean*, *α-Winsorized mean*, *midrange* $T_n = \frac{1}{2}(X_{(1)} + X_{(n)})$ and *Sen's weighted mean* for the location parameter; and *sample range* $R_n = X_{(n)} - X_{(1)}$ and *Gini mean difference* $G_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |X_i - X_j|$ for the scale parameter.

R-Estimator

Let R_i be the rank of X_i among a random sample $X_1, \cdots, X_n$, i.e.

$$R_i = \sum_{j=1}^{n} I[X_j \leq X_i], \qquad i = 1, \cdots, n$$

and thus $R_i = nF_n(X_i), i = 1, \cdots, n$ where F_n is the empirical distribution function. The ranks are invariant w.r.t. the class of monotone transformation of observations and the test based on ranks have many advantages, for example, the distribution of the test criterion under the hypothesis of randomness (i.e. if $X_1, \cdots X_n$ are i.i.d.) is independent of the distribution of observations.

An *R-estimator* is typically obtained by an inversion of the rank tests. The *Hodges-Lehmann estimator* is one of the examples which corresponds to the one-sample Wilcoxon test and it is given by

$$T_n = \text{median}\left\{\frac{X_i + X_j}{2} : 1 \leq i \leq j \leq n\right\},$$

i.e. the median of the set of all n^2 pairs. There are other types of R-estimator, for example, the *normal score estimator*; see Huber (1981).

Adaptive Combination of Estimators

We explain the idea of the adaptive combination of estimators by considering the M-estimator of location parameter as an example.

Let $T_n(\delta)$ be a solution of the minimizing problem

$$\sum_{i=1}^{n} \rho\left(\frac{X_i - \theta}{S_n}\right) := \min, \quad \theta \in \mathbb{R}, \tag{A.25}$$

where

$$\rho(x) = \delta\rho_1(x) + (1 - \delta)\rho_2(x), \quad 0 \leq \delta \leq 1$$

with $\rho_1(x)$ and $\rho_2(x)$ two discrepancy functions defining the respective estimators, and S_n is the scale estimator. For example, $\rho_1(x) = |x|$ and $\rho_2(x) = x^2$, then the solution $T_n(\delta)$ is a combination of the sample median and the sample mean.

Let f be the pdf of the observations $X_1, \cdots, X_n$. To decide on the value of δ, we use the following *adaptive* way. Suppose $\sqrt{n}(T_n(\delta) - \theta)$ has an asymptotically normal distribution $\mathcal{N}(0, \sigma^2(\delta, \rho, f))$, and then we choose δ by

$$\delta_0 \equiv \delta_0(\rho, f) = \arg \min_{0 \le \delta \le 1} \sigma^2(\delta, \rho, f).$$

Thus, we obtain the estimator $T_n(\delta_0)$ which minimizing the asymptotic variance of $\sqrt{n}(T_n(\delta) - \theta)$ with respect to δ. This $T_n(\delta_0)$ is an *adaptive combination of two estimators*.

By using the adaptive combination of two estimators, we may be able to diminish the shortages of both methods. For more details on the adaptive combination of estimators, see Dodge and Jurečková (2000) where the regression analysis was emphasized.

List of estimators for location parameter

Here we provide a list of robust estimators for the location parameter.

1. Median: A median is described as the numerical value separating the higher half of a sample, a population, or a probability distribution, from the lower half.

2. α-trimmed mean: It is the mean after discarding $\alpha/2$ portion of a probability distribution, or a sample, at both the high and low ends. Typically α =5% or 10%.

3. α-Winsorized mean: It is the mean after replacing $\alpha/2$ portion of a probability distribution, or a sample, at both the high and low ends with the most extreme remaining values. Typically α =5% or 10%.

4. Hodges-Lehmann estimator: $\hat{\mu} = \text{med}\left\{\frac{X_i+X_j}{2} : 1 \leq i \leq j \leq n\right\}$.

5. Midrange: $\hat{\mu} = (X_{(1)} + X_{(n)})/2$.

6. Midhinge: $\hat{\mu} = (Q_1+Q_3)/2$, where Q_1 and Q_3 are the first and third quartiles.

7. Sen's weighted mean: $\hat{\mu}_{n,k} = \binom{n}{2k+1}^{-1} \sum_{i=1}^{n} \binom{i-1}{k}\binom{n-i}{k} X_{(i)}$, where $0 < k < \frac{n-1}{2}$.

8. Huber estimator.

9. Redescending M-estimator.

10. (Adaptive) combination of estimators.

11. MLE based on the density estimator.

Appendix B

Fourier Transform

The Fourier transform can be defined on $L_1(\mathbb{R})$ or $L_2(\mathbb{R})$. The following results can be found in Meister (2009, page 181 and page 185).

Theorem B.0.1. *Assume that $f \in L_1(\mathbb{R})$ is bounded and continuous at some $x \in \mathbb{R}$; and, in addition, $f^{ft} \in L_1(\mathbb{R})$. Then, we obtain*

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\{-itx\} f^{ft}(t) dt.$$

Let $\mathcal{C}$ be the set of all bounded and continuous functions in $L_1(\mathbb{R})$ with integrable Fourier transform.

Theorem B.0.2. *The Fourier transform on $L_2(\mathbb{R})$, defined by the unique continuation of the Fourier transform on $\mathcal{C}$, is a bijective mapping from $L_2(\mathbb{R})$ to $L_2(\mathbb{R})$. Its reverse mapping is equal to $f \mapsto \frac{1}{2\pi} f^{ft}(-\cdot)$. Further, we have*

$$< f, g >= \frac{1}{2\pi} \left\langle f^{ft}, g^{ft} \right\rangle, \quad \text{for all } f, g \in L_2(\mathbb{R}) \quad \text{(Plancherel's isometry)}$$

and

$$\|f\|_2^2 = \frac{1}{2\pi} \|f^{ft}\|_2^2, \quad \text{for all } f \in L_2(\mathbb{R}) \quad \text{(Parseval's identity).}$$

Appendix C

Characteristic Functions

Properties of the characteristic function can be found in the book by Luckacs (1970).

Theorem C.0.3 (Luckacs (1970), page 15)**.** *Every characteristic function is uniformly continuous on the whole real line.*

Theorem C.0.4 (Luckacs (1970), page 38)**.** *Let $F = F_1 * F_2$ be the convolution of two distributions F_1 and F_2. If one of the components of F is a continuous function, then the symbolic product is also a continuous function. If one of the components of F is absolutely continuous then F is also absolutely continuous.*

The following theorem can be found in Luckacs (1970), page 23, or Grimmett and Stirzaker 2001, page 183.

Theorem C.0.5 (Grimmett and Stirzaker 2001, page 183)**. (a)** *If $\varphi^{(k)}(0)$ exists then*

$$\begin{cases} \mathbb{E}|X^k| < \infty & \text{if } k \text{ is even,} \\ \mathbb{E}|X^{k-1}| < \infty & \text{if } k \text{ is odd.} \end{cases}$$

(b) *If* $\mathbb{E}|X^k| < \infty$ *then*

$$\varphi(t) = \sum_{j=0}^{k} \frac{\mathbb{E}(X^j)}{j!}(it)^j + o(t^k),$$

and so $\varphi^{(k)}(0) = i^k \mathbb{E}(X^k)$.

Lemma C.0.6 (Luckacs (1970), page 56)**.** *Let* $F(x)$ *be a distribution function and* $\varphi(t)$ *its characteristic function. Then* $\Re[1 - \varphi(t)] \geq \frac{1}{4}\Re[1 - \varphi(2t)]$.

Proof. Using the elementary relation that $1 - \cos tx = 2\sin^2 \frac{tx}{2} \geq 2\sin^2 \frac{tx}{2}\cos^2 \frac{tx}{2} = \frac{1}{2}\sin^2 tx = \frac{1}{4}(1 - \cos 2tx)$. $\boxtimes$

By induction we obtain the following result.

Theorem C.0.7 (Luckacs (1970), page 69)**.** *Let* n *be a non-negative integer; then the inequality*

$$\Re[1 - \varphi(t)] \geq \frac{1}{4^n}\Re[1 - \varphi(2^n t)]$$

is satisfied for every characteristic function.

Complex function of a real variable

The characteristic function $\varphi(t)$ is a continuous complex-valued function of a real variable t. Let $\|z\|$ denote the *norm* or *absolute value* of a complex number z.

Proposition C.0.8. *Then we have the following results:*

1. $\int_a^b \Re(\varphi(t))dt = \Re\left(\int_a^b \varphi(t)dt\right)$.
2. $\int_a^b \Im(\varphi(t))dt = \Im\left(\int_a^b \varphi(t)dt\right)$.
3. $\int_a^b \varphi(t)dt = \int_a^b \Re(\varphi(t))dt + \int_a^b \Im(\varphi(t))dt \leq \int_a^b \|\varphi(t\|dt$.

Appendix D

Empirical Characteristic Functions

Suppose the *characteristic function* of a random variable X is given by

$$\varphi(t) = \mathbb{E}[e^{itX}] = \mathbb{E}[costX + isintX] := R(t) + iI(t)$$

where $R(t)$ and $I(t)$ are its real and imaginary parts. Given a random sample $X_1, X_2, \cdots, X_n$, the *empirical characteristic function* is given by

$$\varphi_n(t) = \frac{1}{n}\sum_{j=1}^{n} e^{itX_j}. \tag{D.1}$$

Now we want to find the mean and variance of the e.c.f. $\varphi_n(t)$. Notice that $\varphi_n(t)$ itself can be viewed as a random variable; and it is a complex random variable. The definition for the mean of a complex random variable is generally without doubt. However, more care needs to be taken on the definition of the variance of a complex random variable.

Mean and Variance of a Complex Random Variable

Assume X and Y are real random variables. Let $Z = X + iY$, then Z is a complex random variable. For the definition of the mean value or the expectation, it is somehow straightforward:

$$\mathbb{E}(Z) = \mathbb{E}(X) + i\mathbb{E}(Y).$$

For the definition of the variance, we could make our own. For example, if we naively look at the imaginary unit i as a coefficient, then it may lead to a definition of the variance as follows: $\mathrm{Var}(X+iY) = \mathrm{Var}(X)-\mathrm{Var}(Y)+2i\mathrm{Cov}(X,Y)$. However, this is a complex value, and both the real and imaginary parts could be negative. This definition has no way to match the original meaning of the "variance". As we known, the variance is used to describe the degree of spread of the data from its mean. We could make other definitions to accommodate this property. However, there already exists one definition in the literature which is popularly accepted:

$$\mathrm{Var}(Z) = \mathbb{E}(\|Z - \mathbb{E}(Z)\|^2),$$

where $\|z\|$ denotes the *norm* or *absolute value* of a complex number z. This definition of the variance leads to a single non-negative real value, which matches with the meaning of the variance, and it measures the spread of the length between the origin and the data from the length between the origin and the mean value of the data. It is easy to verify that

$$\mathrm{Var}(Z) = \mathbb{E}(|X - \mathbb{E}(X)|^2 + |Y - \mathbb{E}(Y)|^2) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

More generally, the covariance of two complex random variables Z_1 and Z_2 is defined by

$$\mathrm{Cov}(Z_1, Z_2) = \mathbb{E}\left[(Z_1 - \mathbb{E}Z_1)\overline{(Z_2 - \mathbb{E}Z_2)}\right],$$

where $\overline{z}$ denotes the *complex conjugate* of a complex value z. Then the variance defined above is a special case of the covariance.

Mean and Variance of ECF

Using the above definitions, we obtain the following result.

Proposition D.0.9. *The expectation and the variance of the e.c.f. $\varphi_n(t)$ are given by*

$$\mathbb{E}(\varphi_n(t)) = \varphi(t) \tag{D.2}$$

and

$$Var(\varphi_n(t)) = \frac{1}{n}\left(1 - \|\varphi(t)\|^2\right), \tag{D.3}$$

where $\|z\|$ denotes the norm of a complex number z. Specifically, the variances of the real part and the imaginary part of $\varphi_n(t)$ are given by

$$Var(\Re(\varphi_n(t))) = \frac{1}{n}\left[\frac{1}{2} + \frac{1}{2}R(2t) - R^2(t)\right]$$

and

$$Var(\Im(\varphi_n(t))) = \frac{1}{n}\left[\frac{1}{2} - \frac{1}{2}R(2t) - I^2(t)\right].$$

Proof. The expectation of the e.c.f. is given by

$$\mathbb{E}(\varphi_n(t)) = \mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n} e^{itX_j}\right) = \mathbb{E}[e^{itX_j}] = \varphi(t).$$

The variance of the e.c.f. is given by

$$\begin{aligned}\mathrm{Var}(\varphi_n(t)) &= \mathbb{E}(\|\varphi_n(t) - \varphi(t)\|^2) \\ &= \mathrm{Var}(\Re(\varphi_n(t))) + \mathrm{Var}(\Im(\varphi_n(t)))\end{aligned}$$

where

$$\begin{aligned}\mathrm{Var}(\Re(\varphi_n(t))) &= \frac{1}{n}\mathrm{Var}(\cos(tX_j)) \\ &= \frac{1}{n}\left\{\mathbb{E}[(\cos tX_j)^2] - [\mathbb{E}(\cos tX_j)]^2\right\} \\ &= \frac{1}{n}\left\{\mathbb{E}\left[\frac{1+\cos 2tX_j}{2}\right] - R^2(t)\right\} \\ &= \frac{1}{n}\left\{\frac{1}{2} + \frac{1}{2}R(2t) - R^2(t)\right\}\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}(\Im(\varphi_n(t))) &= \frac{1}{n}\mathrm{Var}(\sin(tX_j)) \\
&= \frac{1}{n}\left\{\mathbb{E}[(\sin tX_j)^2] - [\mathbb{E}(\sin tX_j)]^2\right\} \\
&= \frac{1}{n}\left\{\mathbb{E}\left[\frac{1-\cos 2tX_j}{2}\right] - I^2(t)\right\} \\
&= \frac{1}{n}\left\{\frac{1}{2} - \frac{1}{2}R(2t) - I^2(t)\right\}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathrm{Var}(\varphi_n(t)) &= \mathrm{Var}(\Re(\varphi_n(t))) + \mathrm{Var}(\Im(\varphi_n(t))) \\
&= \frac{1}{n}\left(1 - R^2(t) - I^2(t)\right) \\
&= \frac{1}{n}\left(1 - \|\varphi(t)\|^2\right).
\end{aligned}$$

This ends the proof. $\boxtimes$

Bibliography

[1] Abate, J., Whitt, W. (1992). The Fourier-series method for inverting transforms of probbability distributions. *Queueing Systems*, **10**, 5-88.

[2] Ait-Sahalia, Y. (1996). Testing continuous-time models of the spot interest rate. *The Review of Financial Studies*, **9**, 385-426.

[3] Aït-Sahalia, Y., Jacod, J. (2009). Testing for jumps in a discretely observed process. *The Annals of Statistics*, **37** (1), 184-222.

[4] Aït-Sahalia, Y., Jacod, J. (2010). Is Brownian motion necessary to model high-frequency data? *The Annals of Statistics*, **38** (5), 3093-3128.

[5] Aït-Sahalia, Y., Jacod, J. (2011). Testing whether jumps have finite or infinite activity. *The Annals of Statistics*, **39** (3), 1689-1719.

[6] Barndorff-Nielsen, O.E., Mikosch, T., Resnick, S.I. (2001). *Lévy Processes: Theory and Applications*. Birkhäuser.

[7] Barndorff-Nielsen, O.E., Shephard, N. (2006). Ecomometrics of testing for jumps in financial ecomonics using bipower variation. *Journal of Financial Econometrics*, **4** (1), 1-30.

[8] Basawa, I.V., Brockwell, P.J. (1982). Non-parametric estimation for non-

decreasing Lévy processes. *Journal of the Royal Statistical Society, Series B*, **44** 2, 262-269.

[9] Belomestny, D., Reiß, M. (2006). Spectral calibration of exponential Lévy models. *Finance and Stochastics*, **10**, 449-474.

[10] Bertoin, J. (1996). *Lévy Processes*. Cambridge University Press.

[11] Bordes, L., Delmas, C., Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model where a component is known. *Scand. J. Statist.*, **33**, 733-752.

[12] Bordes, L., Mottelet, S., Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model. *Ann. Statist.*, **34** (3), 1204-1232.

[13] Brant, R. (1984). Approximate likelihood and probability calculations based on transforms. *Ann. Statist.*, **3**, 989-1005.

[14] Breiman, L. (1968). *Probability*. Addison-Wesley Prublishing Company.

[15] Brodsky, B.E., Darkhovsky, B.S. (1993). *Nonparametric methods in change-point problems*. Kluwer Academic Publishers.

[16] Chen, S.X., Delaigle, A., Hall, P. (2010). Nonparametric estimation for a class of Lévy processes. *Journal of Econometrics*, **157**, 257-271.

[17] Chen, J., Gupta, A.K. (2000). *Parametric statistical change point analysis*. Birkhauser, Boston.

[18] Cherubini, U., Lunga, G.D., Mulinacci, S., Rossi, P. (2010). *Fourier Transform Methods in Finance*. Wiley Finance.

[19] Collins, J.R. (1999). Robust M-estimators of scale: minimax bias versus maximal variance. *The Canadian Journal of Statistics*, **27** (1), 81-96.

[20] Comte, F., Genon-Catalot, V. (2009). Nonparametric estimation for pure jump Lévy processes based on high frequency data. *Stochastic Processes and their Applications*, **119**, 4088-4123.

[21] Comte, F., Genon-Catalot, V. (2010). Nonparametric adaptive estimation for pure jump Lévy processes. *Ann. Inst. H. Poincaré Probab. Statist.*, **46** (3), 595-617.

[22] Cont, R., Tankov, P. (2004). *Financial Modelling with Jump Processes*. Chapman & Hall.

[23] Croux, C., Rousseeuw, P.J. (1992). Time-efficient algorithms for two highly robust estimators of scale. *Computational Statistics*, **2**, 411-428.

[24] Csörgő, S. (1981a). Limit bahaviour of the empirical characteristic function *Ann. Probability*, **9** (1), 130-144.

[25] Csörgő, S. (1981b). Multivariate empirical characteristic function *Z. Wahrscheinlichkeitstheorieverw. Gebiete*, **55**, 203-229.

[26] Csörgő, S. (1985). Rate of uniform convergence for the empirical characteristic function *Acta Sci. Math.*, **48**, 97-102.

[27] Csörgő, S., Totik, V. (1983). On how long interval is the empirical characteristic function uniformly consistent? *Acta Sci. Math.*, **45**, 141-149.

[28] Devroye, L., Györfi, L., Lugosi, G.(1996). *A Probabilistic Theory of Pattern Recognition*. Springer.

[29] Dodge, Y., Jurečková, J. (2000). *Adaptive Regression*. Springer.

[30] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*. **7**, 1-26.

[31] Fan, Y., Fan, J. (2011). Testing and detecting jumps based on a discretely observed process. *Journal of Econometrics*, **164**, 331-344.

[32] Feuerverger, A., McDunnough, P. (1981a). On the efficiency of empirical characteristic function procedures. *J. R. Statist. Soc.* B, **1**, 20-27.

[33] Feuerverger, A., McDunnough, P. (1981b). On some fourier methods for inference. *J. R. Amer. Statist. Assoc.*, **76** (374), 379-387.

[34] Feuerverger, A., McDunnough, P. (1984). On statistical transform methods and their efficiency. *Canadian Journal of Statistics*, **4**, 303-317.

[35] Feuerverger, A., Mureika, R.A. (1977). The empirical characteristic function and its applications. *Ann. Statist.*, **5**, 88-97.

[36] Figueroa-López, J.E. (2009). Nonparametric estimation for Lévy models based on discrete-sampling. *Lecture Notes-Monograph Series*, **57**, 117-146.

[37] Figueroa-López, J.E., Houdré, C. (2006). Risk bounds for the non-parametric estimation of Lévy processes. *IMS Lecture Notes - Monograph Series*, **51**, 96-116.

[38] Figueroa-López, J.E., Houdré, C. (2008). Nonparametric estimation for Lévy processes with a view towards mathematical finance. *Preprint*.

[39] Gegler, A. (2011). Estimation of the characteristics of a multivariate Lévy process. *Preprint*.

[40] Gegler, A., Stadtmüler, U. (2010). Estimation of the characteristics of a Lévy process. *Journal of Statistical Planning and Inference*, **140**, 1481-1496.

[41] Grimmett, G., Stirzaker, D. (2001). *Probability and Random Processes*. Oxford University Press, 3rd Edition.

[42] Gugushvili, S. (2009). Nonparametric estimation of the characteristic triplet of a discretely observed Lévy process. *Journal of Nonparametric Statistics*, **21** (3), 321-343.

[43] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons.

[44] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, **58**, 13-30.

[45] Hu, T., Móricz, F., Taylor, R.L. (1989). Strong laws of large numbers for arrays of rowwise independent random variables. *Acta Math. Hung.*, **54** (1-2), 153-162.

[46] Huber, P.J. (1964). Robust estimation of a location parameter *Ann. Math. Statist.*, **35** (1), 73-101.

[47] Huber, P.J. (1981). *Robust Statistics*. Wiley.

[48] Huber, P.J., Ronchetti, E.M. (2011). *Robust Statistics*. John Wiley and Sons, 2nd Edition.

[49] Hunter, D.R., Wang, S., Hettmansperger, T.P. (2007). Inference for Mixtures of Symmetric Distributions. *Ann. Statist.*, **35** (1), 224-251.

[50] Inclán, C., Tiao, F.C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, **89** 427, 913-923.

[51] Itô, K. (1961). *Lectures on Stochastic Processes*. Tata Institute of Fundamental Research. Springer, Berlin.

[52] Jurečková, J., Picek, J. (2006). *Robust Statistical Methods with R*. Chapman and Hall/CRC.

[53] Kappus, J., Reiß, M. (2010). Estimation of the characteristics of a Lévy process observed at arbitrary frequency. *Statistica Neerlandica*, **64**, 314-328.

[54] Karatzas, I., Shreve S.E. (1991). *Brownian Motion and Stochastic Calculus*. Springer, 2nd Edition.

[55] Kolkiewicz, A.W. (2003). Best monotone M-estimators. *The Canadian Journal of Statistics*, **31** (3), 329-347.

[56] Kou, S.G. (2002). A jump-diffusion model for option pricing. *Management Science*, **8**, 1086-1101.

[57] Lebedeva, T.V., Ushakov, N.G. (2007). Uniformly consistent estimators of the characteristic function. *Journal of Mathematical Sciences*, **146** (4), 6008-6015.

[58] Lee, S., Ha, J., Na, O. (2003). The cusum test for parameter change in time series models. *Scandinavian Journal of Statistics*, **30**, 781-796.

[59] Lee, S.S., Hannig, J. (2010). Detecting jumps from Lévy jump diffusion processes. *Journal of Financial Economics*, **96**, 271-290.

[60] Lee, S., Lee, T. (2004). Cusum test for parameter change based on the maximum likelihood estimator. *Sequential Analysis*, **2**, 239-256.

[61] Lee, S., Na, O. (2005a). Test for parameter change in stochastic processes based on conditional least-squares estimator. *J. Multivariate Analysis*, **93**, 375-393.

[62] Lee, S., Na, O. (2005b). Test for parameter change based on the estimator minimizing density-based divergence measures. *Ann. Inst. Statist. Math.*, **3**, 553-573.

[63] Lee, S., Nishiyama, Y., Yoshida, N. (2006). Test for parameter change in diffusion processes by cusum statistics based on one-step estimators. *Ann. Inst. Statist. Math.*, **58**, 211-222.

[64] Lévy, P. (1954). *Processus Stochastiques et Mouvement Brownien*. 2nd edn. Gauthier-Villars, Paris.

[65] Loève, M. (1963). *Probability Theory*. D. Van Nostrand Co., London. Third edition.

[66] Lukacs, E. (1970). *Characteristic Functions*. Griffin, London.

[67] Madan, D.B., Seneta, E. (1987). Chebyshev polynomial approximations and characteristic function estimation. *J. R. Statist. Soc.* B, **49** (2), 163-169.

[68] Mancini, C. (2004). Estimation of the characteristics of the jumps of a general Poisson-diffusion Model. *Scand. Actuarial J.*, **1**, 42-52.

[69] Markatou, M., Horowitz, J.L., Lenth, R.V. (1995). Robust scale estimation based on the empirical characteristic function. *Statistics & Probability Letters*, **25**, 185-192.

[70] Maronna, R.A., Martin, D., Yohai, V.J. (2006). *Robust Statistics: Theory and Methods*. John Wiley & Sons, Ltd.

[71] Martin, R.D., Zamar, R.H. (1989). Asymptotically min-max bias-robust M-estimates of scale for positive random variables. *Journal of the American Statistical Association*, **84**, 494-501.

[72] Martin, R.D., Zamar, R.H. (1993). Bias robust estimates of scale. *Ann. Statist.*, **21**, 991-1017.

[73] McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *Ann. Probab.* , **2** (4), 620-628.

[74] Meister, A. (2009). *Deconvolution Problems in Nonparametric Statistics*. Springer.

[75] Merton, R.C. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science*, **4** (1), 141-183.

[76] Nadaraya, E.A. (1965). On nonparametric estimates of density functions and regression curves. *Theor. Probab. Appl.*, **10** 186-190.

[77] Neumann, M., Reiß, M. (2009). Nonparametric estimation for Lévy processes from low-frequency observations. *Bernoulli*, **15**, 223-248.

[78] Page, E.S. (1954). Continuous inspection schemes. *Biometrika* , **41**, 100-115.

[79] Page, E.S. (1955). A Test for a change in a parameter occurring at an unknown point. *Biometrika* , **42**, 523-527.

[80] Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, **33** (3), 1065-1076.

[81] Rohatgi, V.K., Saleh, A.K. (2001). *An Introduction to Probability and Statistics*. John Wiley & Sons, Second Edition.

[82] Rousseeuw, P.J., Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, **88** (424), 1273-1283.

[83] Rubin, H., Tucker, H.G. (1959). Estimating the parameters of a differential process. *The Annals of Mathematical Statistics*, **30** (3), 641-658.

[84] Rydberg, T.H. (1997). The normal inverse Gaussian Lévy process: simulation and approximation. *Commun. Statist. - Stochastic Models*, **13** 4, 887-910.

[85] Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press, Cambridge.

[86] Sepp, A. (2012). An approximate distribution of delta-hedging errors in a jump-diffusion model with discrete trading and transaction costs. *Quantitative Finance*, **12**, No. 7, 1119-1141.

[87] Sezer, S.O. (2006). Bayesian sequential change-point detection and hypothesis testing problems for compound Poisson and Wiener processes. Ph.D. Thesis, Princeton University.

[88] Shimizu, Y. (2006). Density estimation of Lévy measures for discretely observed diffusion processes with jumps. *J. Japan Statist. Soc.*, **36** (1), 37-62.

[89] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.

[90] Silverman, B.W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.*, **6** (1), 177-184.

[91] Sueishi, N., Nishiyama, Y. (2005). Estimation of Lévy processes in mathematical finance: a comparative study. *Proceedings of MODSIM05*. 953-959.

[92] West, R.W., Ogden, R.T. (1997). Continuous-time estimation of a change-point in a Poisson process. *J. Statist. Comput. Simul.*, **56**, 293-302.

[93] Wu, E.K.H. (1992). Distributions minimizing Fisher information for scale in ϵ-contamination neighbourhoods. *Statistics & Probability Letters*, **14**, 373-383.