Adaptive Location and Scale Estimation with Kernel-Weighted Averages

Michael Pokojovy

Department of Mathematics & Statistics and School of Data Science
Old Dominion University

mpokojovy@odu.edu
https://github.com/mpokojovy

Fall 2024 MAA MD-DC-VA Section Meeting
November 2, 2024
ODU, Norfolk, VA

Abstract

A wide variety of location and scale estimators have been developed for light-tailed distributions. Despite indisputable importance in business, finance, cybersecurity, etc., statistical estimation and inference in the presence of heavy tails have received less attention in the literature. We adopt the Kernel-Weighted Average (KWA) approach to location and scale estimation and present a set of extensive comparisons with five prominent competitors. Unlike nonparametric kernel density estimation, the optimally tuned bandwidth for KWA estimators does not necessarily converge to zero as sample size grows. We also perform a large-scale Monte Carlo simulation to search for the optimal bandwidth that minimizes the mean squared error (MSE) of KWA location and scale estimators with simulated samples from Student's t-distribution with degrees of freedom (df) $1, 2, \dots 30$. We further develop an adaptive technique to estimate the df that best match the observed samples using Cramér-von Mises test of goodness-of-fit. Unlike many existing methodologies, our approach is data-driven and exhibits excellent statistical performance. To illustrate this, we apply it to three real-world financial datasets containing daily closing prices of AMC Entertainment (AMC), GameStop (GME) and Meta Platforms (META) stocks to calibrate a geometric random walk model with Student's t log-increments. This is a joint work with Su Chen (University of Nebraska Medical Center), Andrews T. Anum (University of Memphis) and John Koomson (The University of Texas at El Paso).

Pokojovy, M., Chen, S., Anum, T. A. and Koomson, J. (2024). Adaptive Location and Scale Estimation with Kernel Weighted Averages. *Communications in Statistics – Simulation and Computation*, 1-20. doi:10.1080/03610918.2024.2408622

Pokojovy, M. (2024). `https://github.com/mpokojovy/KWA1D`

Outline

1. Introduction
2. Location and Scale Estimation
3. Optimal Bandwidth Selection
4. Simulation Study
5. Application to Stock Market Data
6. Conclusions

Introduction

- Measures of location and scatter describe central tendency and dispersion of a random variable.
- Based on the first two moments, statistical inference methods such as one/two-sample z-/t test and corresponding confidence intervals, ANOVA and regression analysis, were developed and applied to many fields.
- These classical statistical inference methods strongly rely on normality assumptions.
- The sample mean and sample variance, perform poorly for skewed, heavy-tailed and multi-modal distributions.

Introduction – Cnt'd (1)

- Since moments, such as the first moment (i.e, expected value or mean), typically do not exist for random variables with heavy-tailed distribution, classical parameter estimation methods may fail to provide reasonable results.
- For heavy-tailed data when the expected value and variance do not exist, it is impossible to make reasonable statistical inference based on sample moments.
- Nonparametric counterparts such as sample median, HL and MCD are recommended as alternatives.
- Some of these estimators, such as trimmed mean, produce biased results in hypothesis testing and confidence interval estimation.

Introduction – Cnt'd (2)

- Ahmad (1982) proposed a nonparametric kernel density functional estimation (KDFE) method for location and scale parameters of distributions from a location-scale family $f(x) = \frac{1}{\sigma} f_0(\frac{x-\mu}{\sigma})$ with known base density function $f_0(x)$.
- Chen (2020) rewrote the location parameter of Ahmad (1982) by eliminating unknown density functionals $\int f^2(x)\mathrm{d}x$ and $\int xf^2(x)\mathrm{d}x$ and proposed a new location estimator referred to as kernel weighted average (KWA).
- However, Chen (2020) did not provide
    1. Estimators of scale parameters proposed by Ahmad (1982).
    2. A constructive way to choose the optimal bandwidth for KWA.
- We adopt the KWA approach to scale estimation for samples from both light- and heavy-tailed distributions and also propose a practically amenable approach to compute the optimal bandwidth for both location and scale estimation.

Location and Scale Estimation

- Consider a continuous random variable X with pdf $f(x)$.
- Let $\big(f(x \mid \mu, \sigma)\big)_{(\mu,\sigma)}$ be a regular location-scale family defined as

$$f(x|\mu, \sigma) = \frac{1}{\sigma} f_0\left(\frac{x-\mu}{\sigma}\right) \quad \text{for } x \in \mathbb{R} \tag{1}$$

with location and scale parameters $\mu \in \mathbb{R}$ and $\sigma > 0$, respectively, and a baseline probability density function (pdf) assumed symmetric around 0.

Location and Scale Estimation – Cnt'd (1)

- Generally, for a square-integrable random variable X with a probability density $f(x) = f(x|\mu, \sigma)$ from a location-scale family defined in Equation (1), the location and scale parameters can be expressed as usual population mean and standard deviation

$$\mu = \int x f(x) \mathrm{d}x, \quad \sigma = \left(\int (x - \mu)^2 f(x) \mathrm{d}x \right)^{1/2}. \tag{2}$$

- Since $f^2(x)$ typically exhibits a faster asymptotic decay than $f(x)$, the location parameter for heavy-tailed distributions can alternatively be expressed as

$$\mu = \frac{\int x f^2(x) \mathrm{d}x}{\int f^2(x) \mathrm{d}x}.$$

Location and Scale Estimation – Cnt'd (2)

- The integral for the scale parameter expressed in terms of the baseline pdf is given by

$$\int (x-\mu)^2 f^2(x)\mathrm{d}x = \sigma^2 \int z^2 f_0^2(z)\mathrm{d}z.$$

- Solving for σ, we get

$$\sigma^2 = \frac{\int (x-\mu)^2 f^2(x)\mathrm{d}x}{\int z^2 f_0^2(z)\mathrm{d}z} \times \frac{\int f_0^2(x)\mathrm{d}x}{\int f^2(x)\mathrm{d}x}.$$

or, equivalently,

$$\sigma = \left(\frac{\int z^2 f_0^2(z)\mathrm{d}z}{\int f_0^2(z)\mathrm{d}z}\right)^{-1/2} \left(\frac{\int (x-\mu)^2 f^2(x)\mathrm{d}x}{\int f^2(x)\mathrm{d}x}\right)^{1/2}.$$

Location and Scale Estimation – Cnt'd (3)

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with probability density function $f(x|\mu, \sigma)$ as in Equation (1). Adopting the kernel functional estimation procedure of Chen (2020), we define the KWA estimators of μ and σ via

$$\hat{\mu}_{h_\mu} = \frac{\sum_{i<j} \frac{x_i + x_j}{2} K_{h_\mu}(x_i - x_j)}{\sum_{i<j} K_{h_\mu}(x_i - x_j)}, \qquad (3)$$

$$\hat{\sigma}_{h_\mu, h_\sigma} = C_{n, h_\mu, h_\sigma} \left(\frac{\sum_{i<j} \frac{(x_i - \hat{\mu}_{h_\mu})^2 + (x_j - \hat{\mu}_{h_\mu})^2}{2} K_{h_\sigma}(x_i - x_j)}{\sum_{i<j} K_{h_\sigma}(x_i - x_j)} \right)^{1/2} \qquad (4)$$

  where $K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$ for a regular kernel $K(u)$ and $h_\mu > 0$, $h_\sigma > 0$ are two bandwidths. The kernel $K(u) = (2\pi)^{-1/2} \exp(-u^2/2)$ is selected as the standard Gaussian probability density function.

- The scaling number C_{n, h_μ, h_σ} is a correction factor selected to render the scale estimator asymptotically unbiased.

Optimal Bandwidth Selection

- We adopt the KWA approach to scale estimation for samples from both light- and heavy-tailed distributions and also propose a practically amenable approach to compute the optimal bandwidth for both location and scale estimation.
- We designed and performed a large-scale Monte Carlo simulation to search for the optimal bandwidth that minimizes the mean squared error (MSE) of KWA location and scale estimators with synthetic samples simulated from the (standard) Student's t_ν-distribution with various degrees of freedom (df) $\nu = 1, 2, 3, 4, 5, 10, 20$ and 30.

Optimal KWA Location Bandwidth Selection

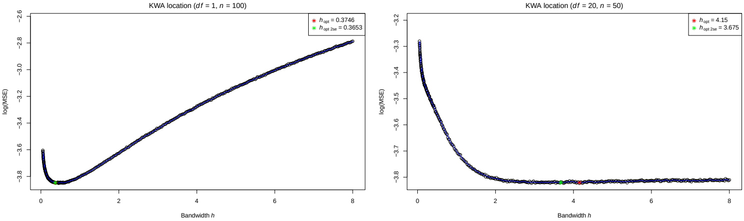

Figure 1: Optimal KWA location bandwidth selection

Optimal KWA Scale Bandwidth Selection

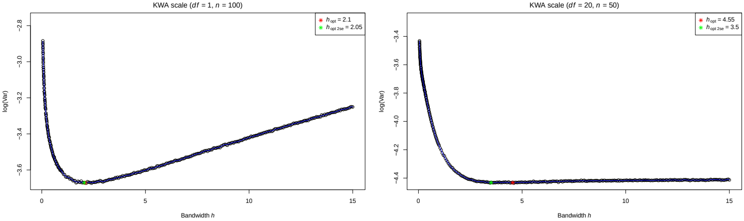

Figure 2: Optimal KWA scale bandwidth selection

Simulation Settings

- For each df and sample size pair (ν, n) with $\nu = 1, 2, 3, 4, 5, 10, 20, 30, \infty$ and $n = 30, 50, 100, 200, 500$, we generated $N = 100{,}000$ samples of size n each from the standard Student's t_ν distribution.
- Competing estimators
  1. location: sample mean, MCD, HL, $\hat{\mu}_{Q_n}$.
  2. scale: sample standard deviation, MCD, HL, Q_n, IQR.
- For each sample, we computed the location and scale estimates $\hat{\mu}_i$ and $\hat{\sigma}_i$ using our KWA location and scale estimators as well as the competing estimators.

$$\widehat{\mathrm{MSE}}(\hat{\mu}) = \frac{1}{N}\sum_{i=1}^{N}\hat{\mu}_i^2,$$

$$\widehat{\mathrm{Var}}\big(\log(\hat{\sigma})\big) = \frac{1}{N-1}\sum_{i=1}^{N}\big(\log(\hat{\sigma}_i)\big)^2 - \frac{N}{N-1}\left(\frac{1}{N}\sum_{i=1}^{N}\log(\hat{\sigma}_i)\right)^2.$$

Simulation Results

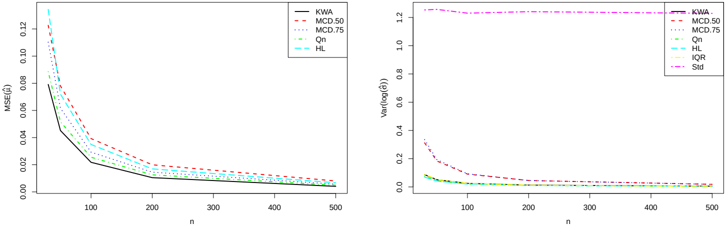

Figure 3: Errors of location and scale estimates for $df = 1$.

Simulation Results – Cnt'd (1)

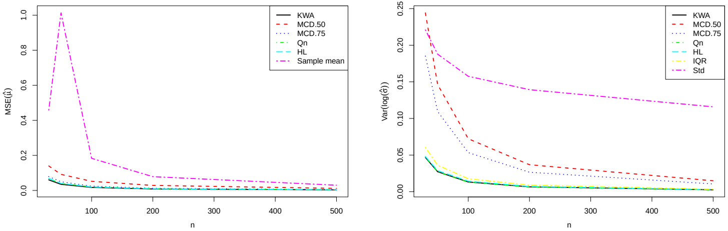

Figure 4: Errors of location and scale estimates for $df = 2$.

Simulation Results – Cnt'd (2)

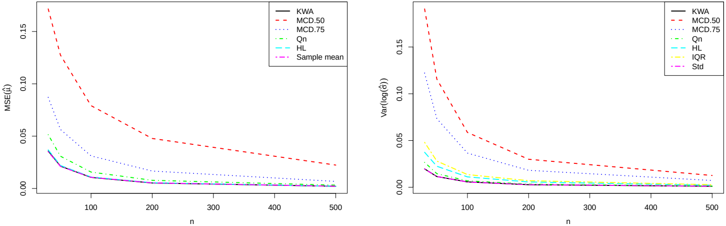

Figure 5: Errors of location and scale estimates for $df = 30$.

Application to Stock Data

- We considered the daily closing prices of AMC Entertainment Holdings, Inc. (AMC), GameStop Corp. (GME) and Meta Platforms, Inc. (META, formerly FB) stocks.

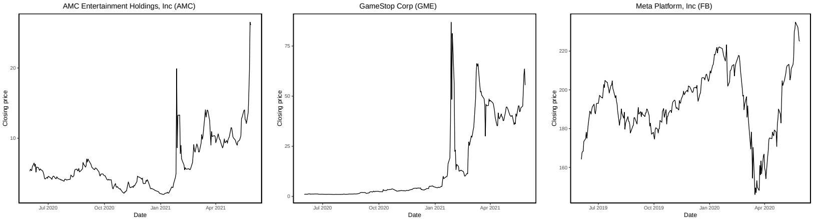

Figure 6: Daily closing prices.

Application to Stock Data – Cnt'd (1)

- The geometric Brownian motion is commonly used to model stock prices and other types of real-world time series with log-normal increments.
- Let $\big(S(t)\big)_{t\geq 0}$ be a geometric Brownian motion with drift r and volatility ω observed over an equispaced time grid $\{t_0, t_1, \dots, t_n\}$ such that $t_k - t_{k-1} \equiv \Delta t$ is a positive constant.
- Letting $x(t_k)$ denote the log-differences of $S(t_k)$'s

  $$x(t_k) := \ln(S(t_k)) - \ln(S(t_{k-1})) = \Big(r - \frac{1}{2}\omega^2\Big)\Delta t + \omega\big(W(t_k) - W(t_{k-1})\big)$$

  with a standard Brownian motion $\big(W(t)\big)_{t\geq 0}$
- $x(t_k) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with mean $\mu = (\Delta t)\Big(r - \frac{1}{2}\omega^2\Big)$ and variance $\sigma^2 = (\Delta t)\omega^2$.

Application to Stock Data – Cnt'd (2)

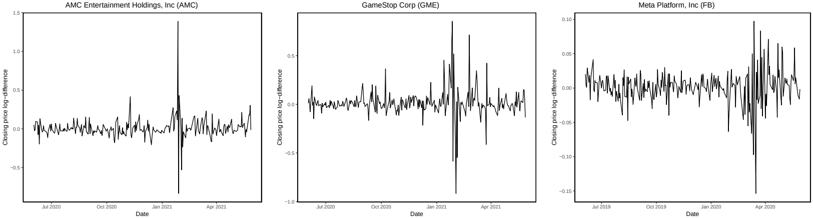

Figure 7: Log-differenced daily closing prices.

Application to Stock Data – Cnt'd (3)

- Therefore, following Pokojovy and Anum (2022), we replace the Gaussianity assumption with the more general assumption that $x(t_k)$'s follow Student's location-scale t_ν-distribution

$$x(t_k) \overset{\text{i.i.d.}}{\sim} t_\nu(\mu, \sigma) \text{ with location } \mu \text{ and scale } \sigma \tag{5}$$

  for some df $\nu \in [1, \infty]$.
- This assumption also allows for Gaussianity of the log-increments when $\nu = \infty$.

Application to Stock Data – Cnt'd (4)

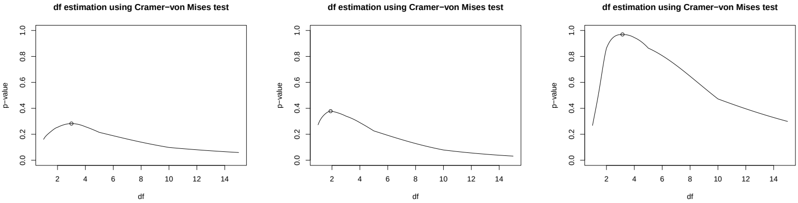

Figure 8: Degree of freedom ν estimation based on $x(t_k)$'s.

Application to Stock Data – Cnt'd (5)

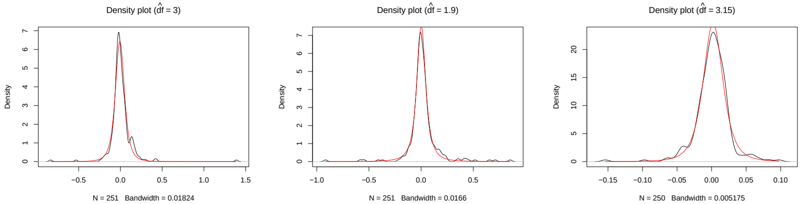

Figure 9: Kernel density (black) vs fitted Student's $t_{\hat{\nu}}$ density plots for the log-increments $x(t_k)$.

Summary & Conclusions

- Adopting the kernel weighted average technique, we proposed a new scale estimator termed as KWA scale estimator akin to the KWA location estimator known in the literature.
- For both KWA location and scale estimators, we performed extensive simulations to compute the optimal bandwidth for datasets for a Student's t_ν location-scale family.
- For more general symmetric distributions, we developed an adaptive data-driven technique based on Cramér-von Mises goodness-of-fit test to optimally select the bandwidth for our KWA location and scale estimators.
- In sum, the optimally tuned KWA location and scale estimators were demonstrated to offer excellent estimation quality and perform reliably across a wide spectrum of tail weights.

Thank you for your attention!

Questions? Comments?