Final Project: Nonparametric Density Estimation

Jaime Frade
Dr. Yiyuan She
STA5507: Applied Nonparametrics
Florida State University: Department of Statistics
jfrade@stat.fsu.edu

November 27, 2009

Contents

1 Introduction **2**

2 Details of Data **4**
2.1 Financial Application 4
2.1.1 Setup of Problem 4
2.1.2 Setup of Problem: VaR 4
2.2 Financial Definitions 5
2.3 Goal 6

3 Methodology **7**
3.1 Definitions 7

4 Results **9**
4.1 Computational Results 9
4.1.1 Density Estimation 9
4.1.2 VaR Calculations 10
4.2 Conclusion 10

5 Code **11**

6 References **13**

List of Figures

1.1 Standard and Poor Daily Returns 2

2.1 Illustration of the 10% Value at Risk with normally distibuted portfolio value 5
2.2 Histogram for distribution of returns 6

Chapter 1

Introduction

To obtain an estimate for the distribution of the returns for a given asset or a portfolio. Once obtained, to use the distribution to calculate risk measurement for a given time horizon.

Probability density estimation goes hand in hand with nonparametric estimation of the cumulative distribution. The density function provides a better visual summary of how the random variable is distributed across its support. Skewness, kurtosis, disperseness are just a few properties can ascertained from the density plot.

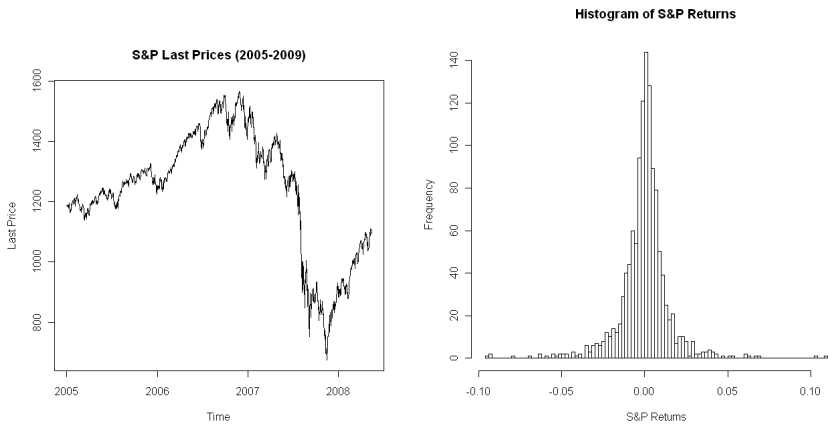


Figure 1.1: Standard and Poor Daily Returns

Unlike using density function, which places probability mass $\frac{1}{n}$ on each observation, the paper focuses on the kernel density estimator that more fairly spreads out the probability mass of each observation, not arbitrarily in a fixed interval, but smoothly around the observation, typically in a symmetric way.

The data for this project will illustrate the usage of basic kernels, to estimate the distribution of returns on a certain asset, as well as a returns for a portfolio.

Using this curve estimation, a certain key risk measure, Value at Risk, which will ultimately provide a single quantitative number summarizing the total risk in portfolio of financial assets, or on a single asset. This measurement is widely used for managers when determining the possible losses for the market risks during certain scenarios.

Chapter 2

Details of Data

2.1 Financial Application

2.1.1 Setup of Problem

In finance, there exists a need to analyze the distribution of returns of various indexes. From these calculations, can obtain measurements of risk. The value-at-risk (VaR) is a measurement which accounts for a confidence interval for the amount of loss one may expect. There exists a percentage of certainty that the portfolio manager will not lose more than the value of the Var of the portfolio in the next N days.

2.1.2 Setup of Problem: VaR

if a portfolio of stocks has a one-day 5% VaR of $1 million, there is a 5% probability that the portfolio will decline in value by more than $1 million over the next day, assuming markets are normal and there is no trading.

The reason for assuming normal markets and no trading, and to restricting loss to things measured in daily accounts, is to make the loss observable. In some extreme financial events it can be impossible to determine losses, either because market prices are unavailable or because the loss-bearing institution breaks up.

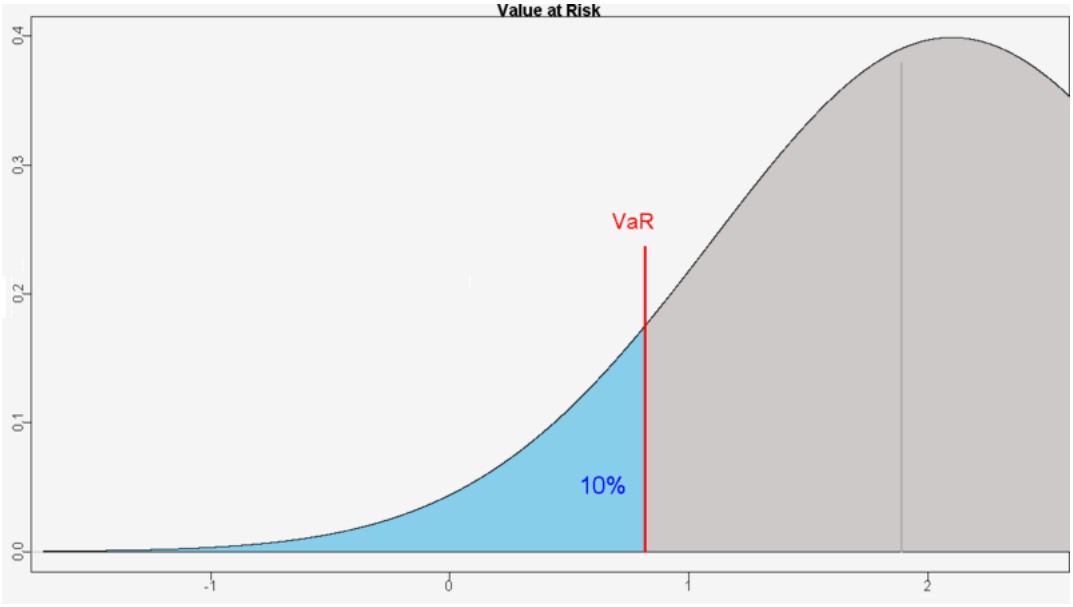


Figure 2.1: Illustration of the 10% Value at Risk with normally distibuted portfolio value

2.2 Financial Definitions

Collected historical daily last price for the following companies below, from 01/2005-11/2009. The daily log return was calculated.

- Standard and Poor's 500 Index: (S&P) A capitalization-weighted index of 500 stocks. The index is designed to measure performance of a broad domestic economy through changes in the aggregate market of 500 stocks representing all major industries.
- NASDAQ: The composite index is a board-based capitalization-weighted index of stocks in all three tiers: Global Select, Global Market, and Capital Markets.
- DOW Jones: The industrial average is a price-weighted average of 30 blue chips stocks that are generally the leaders in their industry.
- AT&T: Communications holding company, provides local and long-distance phone service, wireless and data communications, etc.
- Microsoft: develops, manufactures, licenses, sells, and supports software products. Offers operating system software, etc.
- PetroChina: Explores, develops, and produces crude oil and natural gas.

2.3 Goal

Stochastic dynamics of stock prices is commonly described by a geometric (multiplicative) Brownian motion, which gives a log-normal density distributions for returns. However, numerous observations show that the tails of the PDF decays lower than the lognormal distribution predicts (the so-called fat-tails effect). Which does not provide a substantial estimate for for calculating VaR, basically the area under the lower left curve, when returns are high spreads.

The following provides a histogram of the distribution of returns. Each of the distributions used followed similar shape and symmetry by analyzing the following output. The distributions of returns display a normal shape with outliers causing a relatively fat tails.

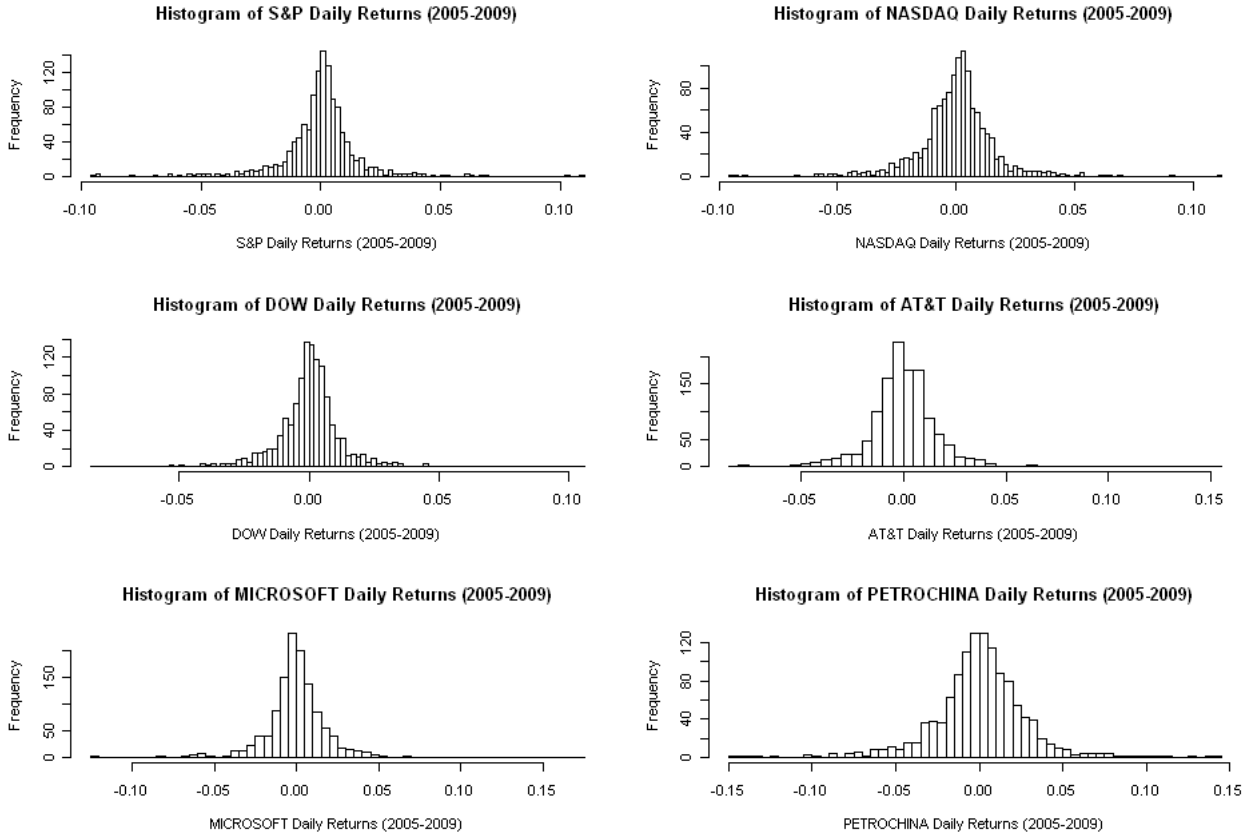


Figure 2.2: Histogram for distribution of returns

Chapter 3

Methodology

3.1 Definitions

With a sample of $X_1, X_2, \ldots, X_n$, write the density estimator

$$\hat{f}(x) = \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_n}\right) \tag{3.1}$$

for $X_i = x_i$, $i = 1, \ldots, n$. The *kernel function* K represents how the probability mass is assigned. For example, for the histogram, in any particular interval, K is constant. The smoothing function h_n is a positive sequence of bandwidths analogous to the bin width in a histogram.

To estimate f, one may use the density of the random variable $\widehat{X}+hZ$, where $\widehat{X}$ has the distribution (conditionally, given $X_1, \ldots, X_n$) and Z is independent of $\widehat{X}$. From (3.1), the *bandwidth* satisfies

$$h \equiv h_n \longrightarrow \text{as } n \to \infty \tag{3.2}$$

$\widehat{X} + hZ$ has the density.

$$\widehat{f_n}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i), \tag{3.3}$$

where K_h is the density of hZ. From (3.3), using the following definition for K_h below, will obtain (3.1).

$$K_h(y) = \frac{1}{h} K\left(\frac{y}{h}\right) \tag{3.4}$$

The kernel function K has five important properties

- $K(x) \geq 0 \quad \forall\, x$
- $K(x) = K(-x) \quad \forall \;\; x > 0$
- $\int K(u)du = 1$
- $\int uK(u)du = 0$
- $\int u^2 K(v)dv = \sigma_k^2 < \infty$

Basic idea is that K controls the shape, h_n controls the spread of the kernel. The accuracy of a density estimator can be evaluated using he mean intergrated squared error, defined as

$$\begin{aligned} \text{MISE} &= \mathbf{E}\left(\int (f(x) - \hat{f}(x))^2 dx\right) \\ &= \int \text{Bias}^2(\hat{f}(x))dx + \int \mathbf{V}ar(\hat{f}(x))dx \end{aligned} \tag{3.5}$$

To find a density estimator that minimizes the MISE under the five constraints, also will assume that $f(x)$ is continuous and twice differentiable, $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. Under these conditions,

$$\begin{aligned} \text{Bias} &= \frac{\sigma_K^2}{2} f''(x) + O(h_n^4) \\ \text{Var}\left(\hat{f}(x)\right) &= \frac{f(x)R(K)}{nh_n} + O(n^{-1}) \end{aligned} \tag{3.6}$$

where $R(g) = \int g(u)^2 du$

Determine the minimum MISE by the choice of h_n. However, choosing h_n to reduce the bias will increase the variance and vice versa, there is a trade off. The choice of the *bandwidth* is important to the construction of $\hat{f}(x)$.

If h is chosen to be small, the minor difference in main part of the density will be apparent. If h is choose to be large, the tails of the distribution are better modeled, but fail to see important characteristics of the middle quartiles of the data.

Using numerical integration from the distributions of the data, as well as, using R packages, VaR calculations were done on the data.

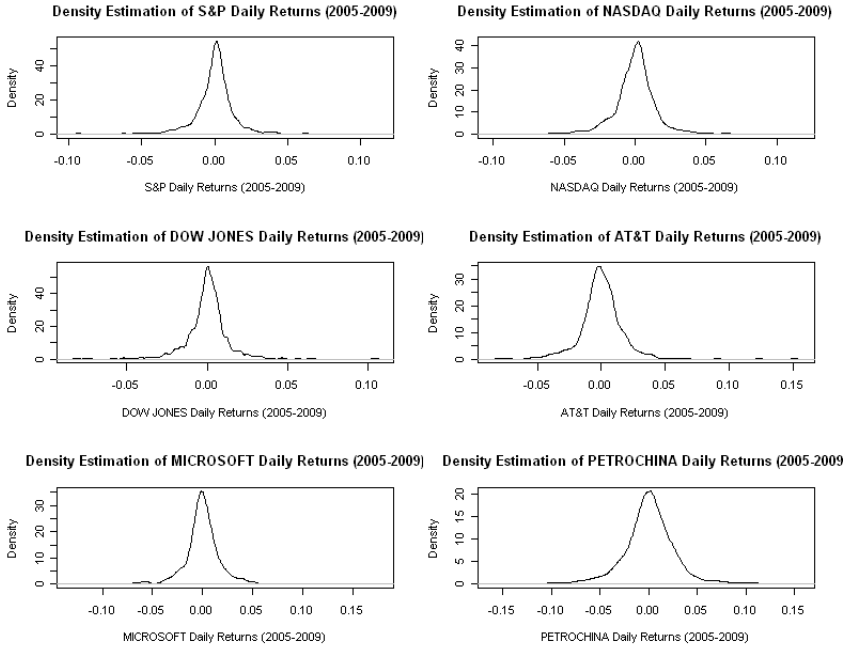
Chapter 4

Results

4.1 Computational Results

I specified a kernel function using the statistical program R 'epanechinikov' which are all scaled to have standard deviation equal to one, so the bandwidth parameter means roughly the same thing regardless of kernel function.

4.1.1 Density Estimation



4.1.2 VaR Calculations

The following is the list of percentiles for one-day 5% VaR, the likelihood that a given portfolio's losses will exceed this certain amount.

VaR	S&P	NASDAQ	DOW	AT&T	Microsoft	PetroChina
5%	39.52153	81.48998	340.4551	1.064806	1.390841	8.99558

The above results can be interpreted as largest loss likely to be suffered on a portfolio solely this position over a holding period of one day. For instance, an investment bank holding that position in the portfolio might report that its portfolio has a 1-day VaR of $39.52 at the 95% confidence level, if invested soley in Standard and Poors.

4.2 Conclusion

The results presented in this paper are highly dependent on historical data. The underlying method taken here was to simulate a density using kernel techniques and the lowest 5% quantile of this distribution is used as an approximation to VaR. Current VaR calculations involve Monte Carlo simulations and Variance-Covariance matrices which account for more robust measures, as well as, more applicable to larger aggregate portfolios. The methods taken here are soley for illustration of nonparametric techniques in a real world application.

Chapter 5

Code

```
datafiles<- read.table("datanodates.csv", header = TRUE, sep = ",", na.string=".")
datareturns<- read.table("datanodates2.csv", header = TRUE, sep = ",", na.string=".")

SPX<-datareturns[2]
CCMP<-datareturns[3]
INDU<-datareturns[4]
T<-datareturns[5]
MSFT<-datareturns[6]
PTR<-datareturns[7]
Y<-data.frame(SPX, CCMP, INDU, T, MSFT, PTR)

x <- seq(as.Date("2005-01-01"), as.Date("2009-10-31"), by = "day")
plot(x[1:length(Y$SPX)],Y$SPX, xlab = "Time", ylab = "Last Price", main = "S&P Price (2005
####################

SPX<-datafiles[1]
CCMP<-datafiles[2]
INDU<-datafiles[3]
T<-datafiles[4]
MSFT<-datafiles[5]
PTR<-datafiles[6]
Y<-data.frame(SPX, CCMP, INDU, T, MSFT, PTR)

par(mfrow = c(3,2))
xname<-"S&P Daily Returns (2005-2009)"
hist(Y$SPX,breaks = "FD",main = paste("Histogram of" , xname),xlab = xname)

xname<-"NASDAQ Daily Returns (2005-2009)"
hist(Y$CCMP,breaks = "FD",main = paste("Histogram of" , xname),xlab = xname)

xname<-"DOW Daily Returns (2005-2009)"
hist(Y$INDU,breaks = "FD",main = paste("Histogram of" , xname),xlab = xname)

xname<-"AT&T Daily Returns (2005-2009)"
```

```
hist(Y$T,breaks = "FD",main = paste("Histogram of" , xname),xlab = xname)

xname<-"MICROSOFT Daily Returns (2005-2009)"
hist(Y$MSFT,breaks = "FD",main = paste("Histogram of" , xname),xlab = xname)

xname<-"PETROCHINA Daily Returns (2005-2009)"
hist(Y$PTR,breaks = "FD",main = paste("Histogram of" , xname),xlab = xname)

################################
par(mfrow = c(3,2))
xname<-"S&P Daily Returns (2005-2009)"
plot(density(Y$SPX, bw="ucv",kernel = "epanechnikov"),main = paste("Density Estimation of"
#X<-seq(min(Y$SPX), max(Y$SPX), length=length(Y$SPX))

xname<-"NASDAQ Daily Returns (2005-2009)"
plot(density(Y$CCMP, bw="ucv",kernel = "epanechnikov"),main = paste("Density Estimation of
#X<-seq(min(Y$CCMP), max(Y$CCMP), length=length(Y$CCMP))

xname<-"DOW JONES Daily Returns (2005-2009)"
plot(density(Y$INDU, bw="ucv",kernel = "epanechnikov"),main = paste("Density Estimation of
#X<-seq(min(Y$INDU), max(Y$INDU), length=length(Y$INDU))

xname<-"AT&T Daily Returns (2005-2009)"
plot(density(Y$T, bw="ucv",kernel = "epanechnikov"),main = paste("Density Estimation of" ,
#X<-seq(min(Y$T), max(Y$T), length=length(Y$T))

xname<-"MICROSOFT Daily Returns (2005-2009)"
plot(density(Y$MSFT, bw="ucv",kernel = "epanechnikov"),main = paste("Density Estimation of
#X<-seq(min(Y$MSFT), max(Y$MSFT), length=length(Y$MSFT))

xname<-"PETROCHINA Daily Returns (2005-2009)"
plot(density(Y$PTR, bw="ucv",kernel = "epanechnikov"),main = paste("Density Estimation of"
#X<-seq(min(Y$PTR), max(Y$PTR), length=length(Y$PTR))

#############
library(VaR)
var1<-VaR.norm(Y$SPX[1:1230], p = 0.95, dt = 1)
var2<-VaR.norm(Y$CCMP[1:1230], p = 0.95, dt = 1)
var3<-VaR.norm(Y$INDU[1:1230], p = 0.95, dt = 1)
var4<-VaR.norm(Y$T[1:1230], p = 0.95, dt = 1)
var5<-VaR.norm(Y$MSFT[1:1230], p = 0.95, dt = 1)
var6<-VaR.norm(Y$PTR[1:1230], p = 0.95, dt = 1)
```

Chapter 6

References

- V. A. Epanechnikov, Non-Parametric Estimation of a Multivariate Probability Density, Theory Probab. Appl. Volume 14, Issue 1, pp. 153-158 (January 1969) Issue Date: January 1969
- Efromovich, S. (1999). Nonparametric Curve Estimation: Methods, Theory and Applications. Springer Series in Statistics. New York, N.Y.: Springer.
- Kvam, P. H., & Vidakovic, B. (2007). Nonparametric statistics with applications to science and engineering. Wiley series in probability and statistics. Hoboken, N.J.: Wiley-Interscience.
- Tapia, R. A., & Thompson, J. R. (1978). Nonparametric probability density estimation. Baltimore: Johns Hopkins University Press.
- Hull, J. (1989). Options, futures, and other derivative securities. Englewood Cliffs, N.J.: Prentice Hall.
- Hull, J. (2007). Risk management and financial institutions. Upper Saddle River, NJ: Pearson Prentice Hall.