Nonparametric Curve Estimation

Jaime Frade

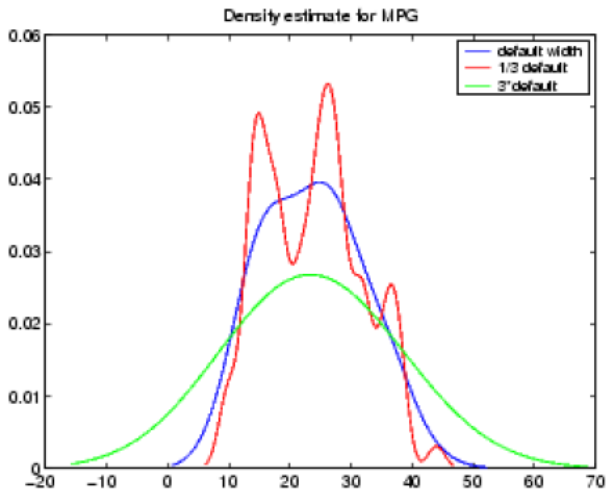
Department of Statistics
Florida State Univeristy

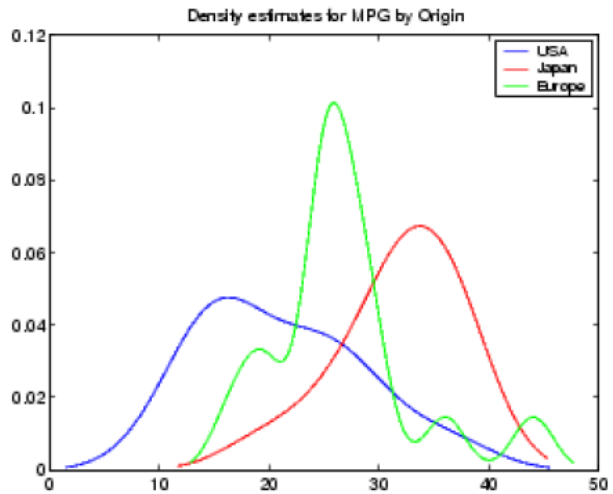
Victor Patrangenaru
STA5334

December 16, 2008

Nonparametric Density Estimation

Probabilty 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 properities can ascertiantined from the density plot.





Goal

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

From there, will illustrate four basic kernels, as well as an adaptive kernel function, to estimate the distribution of returns on a certian asset. Simulation will be completed.

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

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

for $X_i = x_i$, $i = 1, \ldots, n$. The *kernel function* K represents how the probabilty 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 analoguous to the bin width in a histogram.

From lecture notes, 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 (1), the *bandwidth* satisfies

$$h \equiv h_n \longrightarrow \text{as } n \to \infty \tag{2}$$

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

$$\widehat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i), \tag{3}$$

where K_h is the density of hZ. From (3), using the following definition for K_h below, will obtain (1).

$$K_h(y) = \frac{1}{h} K\left(\frac{y}{h}\right) \tag{4}$$

The kernel function K has five important properities

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

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

$$\begin{aligned} \text{MISE} &= \mathbb{E}\left(\int (f(x) - \hat{f}(x))^2 dx\right) \\ &= \int \text{Bias}^2(\hat{f}(x))dx + \int \mathbb{V}ar(\hat{f}(x))dx \end{aligned} \tag{5}$$

To find a density estimator that minimizes tehe 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,

$$\begin{aligned} \text{Bias} &= \frac{\sigma_K^2}{2} f''(x) + O(h_n^4) \\ \text{Var}\left(\hat{f}(x)\right) &= \frac{f(x)R(K)}{nh_n} + O(n^{-1}) \end{aligned} \tag{6}$$

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 tradeoff. 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 charateristics of the middle quartiles of the data.

By subsituting the bias and variance into the formula for (5), minimize MISE with

$$h_n^* = \left(\frac{R(K)}{\sigma_K^4 R(f')} \right)^{\frac{1}{5}} n^{-\frac{1}{5}}$$

From here, still can choose $K(x)$ and insert a *representative* density for $f(x)$ to solve for the bandwidth.

Epanechnickov (1969) showed that, upon subsituting $f(x) = \phi(x)$, the kernel that minimizes MISE is

$$K_E(x) = \begin{cases} -\frac{3}{4}(1 - x^2) & |x| \leq 1 \\ 0 & \text{if } |x| > 1 \end{cases}$$

The resulting bandwidth becomes $h^* \approx 1.06\hat{\sigma}n^{-\frac{1}{5}}$, where $\hat{\sigma}$ is the sample standard deviation. The choice relies on the approximation for σ for $f(x)$, can obtain different answers.

Adapative kernels were derived to alleviate this problem. If use mroe general smoothing function tied to the density at x_j, could generalize the denisty estimator as

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_{n,i}} K\left(\frac{x - x_i}{h_{n,i}} \right) \tag{7}$$

where h_n is a decreasing function of n, under $|\hat{f}(x) - f(x)| \xrightarrow{P} 0$

In a majority of practical applications the parameter of smoothness is unknown. After all, nonparametric estimation is typically a first glance at the data at hand. Thus the aim of one paper I researched is to discuss estimates that are adaptive to an unknown smoothness, i.e., data-driven. While the MISE of adaptive estimators can attain the minimax rate, that is, it is not necessary to know the parameter, (can visualize in data). See Efromovich (1999) pg 281.

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 certainity that the portfolio manager will not lose more than the value of the Var of the portfolio in the next N days.

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.

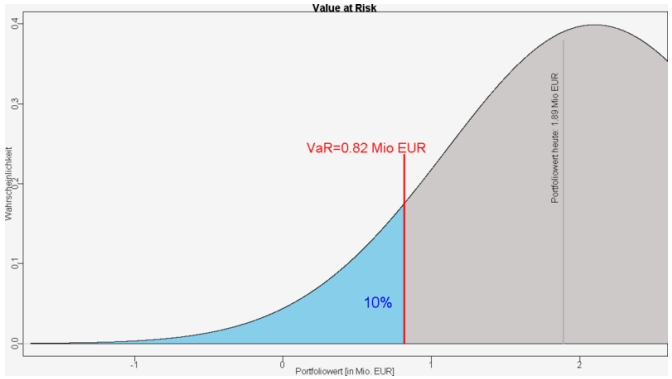


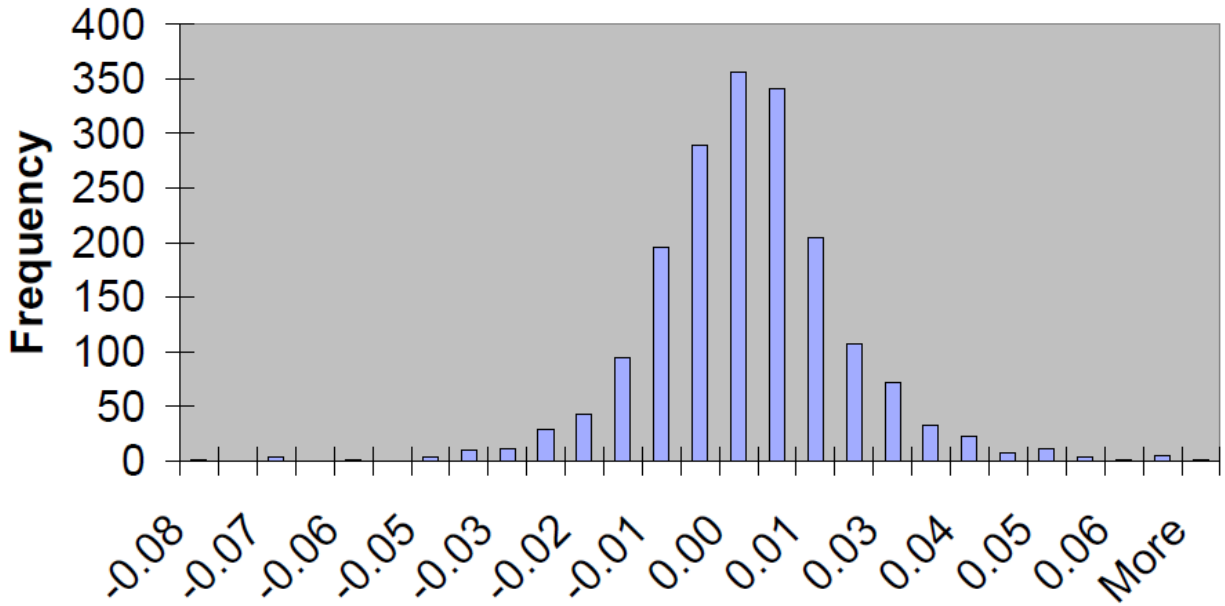
Figure: Illustration of the 10% Value at Risk with normally distibuted portfolio value

Collected historical daily last price for 10 year Treasury notes, 3 month Treasury Bills, and 1 month Treasury bills.

- Treasury bills: (T-bill) A short-term debt obligation backed by the U.S. government with a maturity of less than one year. (maturity is one year or less)
- Treasury notes: (T-notes) similary as above, just mature in two to ten years

Goal:

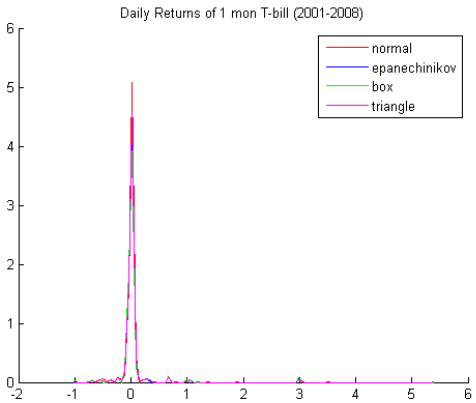
Stochastic dynamics of stock prices is commonly described by a geometric (multiplicative) Brownian motion, which gives a log-normal pdf for returns. However, numerous observations show that the tails of the PDF decays lower than the lognormal distribution predicts (the so-called fat-tails effect). Which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

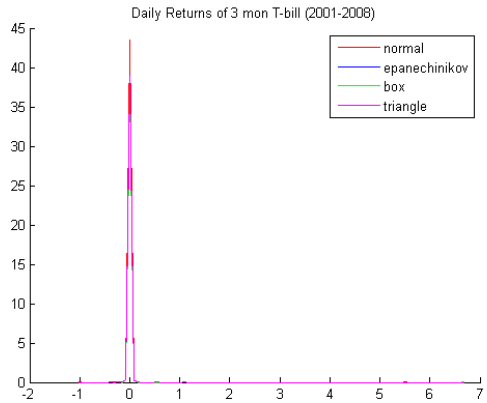


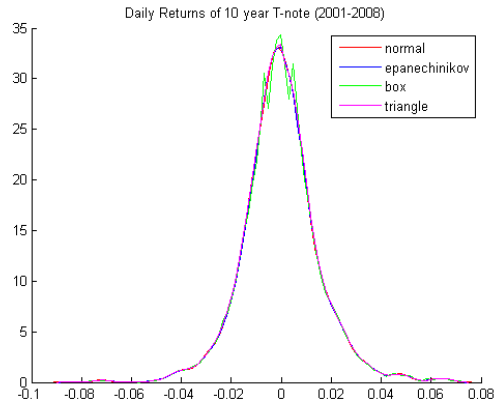
ksdensity

I specified a kernel function of four pre-selected built-in functions in matlab statistical toolbox, 'normal', 'epanechinikov', 'box', and 'triangle', which are all scaled to have standard deviation equal to one, so the bandwidth parameter means roughly the same thing regardless of kernel function.

The default estimator is based on a Normal kernel, Box, Triangle, and Epanechnikov. The next figure shows how the normal kernel compares to the each other kernel. The optimal bandwidth for 1 month T-bills (0.0036), 3 month T-bills (0.0027), and 10 year T-notes (0.0028) was determined.







The density estimates are roughly comparable, but the box kernel produces a density that is rougher than the others.

Futher computational research needs to be completed with the above obtained density curve estimations. Now that a density curve has been approximated, a computational calculation of the area under the curve can done by using a Monte Carlo approach.

- A Course in Mathematical Statistics and Large Sample Theory, by Rabi Bhattacharya and Victor Patrangenaru, 2009, Springer.
- 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.