Abstract

The methods described in this paper are general enough to work with practically any stochastic process. The methods will be presented on the time process of PRIBOR3M. The task of this paper is to estimate the transition density of the process $f(r_{t+1}|r_t)$ with no assumptions on the distribution family nor the parameters of the distribution. We estimate independently the univariate marginal density and the two-variate joint density. Then we get the estimate of the transition density by dividing the joint density with the marginal density. To estimate the joint and marginal density, we will use the so called kernel densities. We can use the estimates to investigate the moments of the transition density. Another possibility is to use the estimate to simulate the stochastic process.

1 Problem formulation

The methods described in this paper are general enough to work with practically any stochastic process. The methods will be presented on the time process of PRIBOR3M. The process is in figure 1. We denote the stochastic process $r_t$. The task of this paper is to estimate the transition density of the process $f(r_{t+1}|r_t)$ with no assumptions on the distribution family nor the parameters of the distribution. We estimate the transition density as

$$
\hat{f}(r_{t+1}|r_t) = \frac{\hat{f}_j(r_{t+1}, r_t)}{\hat{f}_m(r_t)},
$$

where $\hat{f}_j(r_{t+1}, r_t)$ is the joint distribution and $\hat{f}_m(r_t)$ is the marginal (unconditional) density of the process. So we estimate independently the univariate marginal density and the two-variate joint density. To estimate the joint and marginal density, we will use so called kernel densities.

Figure 1: PRIBOR3M daily observations from 16.12.1995 to 29.5.2006.
2 Basics of the kernel density estimates

2.1 From histogram to the kernel density

We start with the univariate kernel densities. The simplest density estimate is a histogram (normalized so the total surface of the bars is equal to one). To construct a histogram, we divide the interval covered by the data values into equal sub-intervals, known as ‘bins’. Every time, a data value falls into a particular sub-interval, then a block of size 1 by the bin width is placed on top of it.\(^2\). The idea of kernel densities is similar. To avoid the dependence of the histogram on the choice of the endpoints of the bins we can alternatively construct a density estimate centering the blocks over the data points rather than fixing the end points of the blocks (see fig. 2 left).

![Figure 2: Left](image)

![Figure 2: Right](image)

Figure 2: **Left:** In the above modified histogram, we place a block of width 1 and height 1/12 (the dotted boxes) as they are 12 data points, and then add them up. The density estimate is the solid curve. The data points are represented by crosses on the x-axis. **Right:** Gaussian kernel density. The data points are represented by crosses on the x-axis. (Both pictures taken from [2])

The box kernel density estimate, the simplest kernel density estimate, is discontinuous and dependent on the bin width (similarly to the simple histogram). We can easily alleviate the first problem by placing different "building blocks" over the data points than a box. The "building blocks" are called kernels.\(^3\) The most common is the Gaussian kernel (see fig. 2 right). So for Gaussian kernel density estimator we get (dropping the time indexes)

\[
\hat{f}_m(r) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{r - r_i}{h} \right),
\]

where

\[
K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2},
\]

\(r_i\) are the observed values, \(n\) is the number of the observed values and \(h\) is the so called bandwidth. Bandwidth, in this case, is the standard deviation of the Gaussian kernel and plays an analogical role as the bin width of the histogram.

2.2 Optimizing the bandwidth

Choosing the bandwidth is a crucial problem. If the bandwidth is too big, the kernel density is over-smoothed and does not reflect some important properties of the unknown density. On the other hand, if the bandwidth is too small, the kernel density is undersmoothed displaying too many details. We face

\(^3\)Note the difference between the kernel ("the building block") and the kernel density (the density estimator).
a trade-off between variance and bias. We would surely like to keep both variance and bias small, but increasing \( h \) will lower the variance while it will raise the bias (decreasing \( h \) will do the opposite). As of yet, there is no universally accepted approach to the problem of choosing the optimal bandwidth. The most widely used way of placing a measure on the global accuracy of an estimator of the density is the mean integrated square error (MISE), i.e. the sum of the integrated variance and integrated squared bias. Minimizing the MISE over \( h \) gives the optimal bandwidth, which represents a compromise between over- and undersmoothing. ([3]) While minimizing the MISE is computationally relatively demanding and time-consuming, there have been introduced so-called "rules of thumb". The most used among them was suggested by Silverman. ([4])

\[
\hat{h}_{opt} = 1.06 \min \left( \frac{\hat{s}}{1.34}, R \right) n^{-0.2},
\]

where \( R \) is the sample interquartile range.

Another possibility would be to use method called Cross Validation (CV). CV is an accuracy method for choosing the smoothing parameter \( h \). The most common is the likelihood CV. The log likelihood is averaged over each choice of omitted \( X_i \) and give the score function, which is to maximize over \( h \)

\[
CV(h) = n^{-1} \sum_{i=1}^{n} \log \left( \hat{f}_{-i}(X_i) \right),
\]

### 2.3 Adaptive kernel density

The problem of the fixed bandwidth is (especially when estimating long tailed distributions) that it may result in undersmoothing in areas with only sparse observations (very small and very high interest rates) while oversmoothing in areas with high density of observations. The solution is to use different bandwidths for each observation. The kernel density estimate is then

\[
\hat{f}_m(r) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i} K \left( \frac{r - r_i}{h_i} \right),
\]

We used the method called adaptive two-stage estimator as proposed in [1]. This algorithm is based on constructing local bandwidth factor \( \lambda_i \). The local bandwidth factors have unit mean and multiply the global fixed bandwidth \( h \).

\[
h_i = \lambda_i h
\]

Henceforth \( h \) controls the overall degree of smoothing while the \( \lambda_i \) stretches or shrinks the sample points bandwidths to adapt to the density of the data (description taken from [6]). \( h \) can be estimated with any of the common fixed bandwidth methods. The local bandwidth factors \( \lambda_i \) should be proportional to the square root of the unknown density function at the sample points:

\[
\lambda_i = (G/f(r_i))^{0.5},
\]

where \( G \) is geometric mean of \( f(r_i) \) \( i = 1, 2, \ldots, n \). Unknown \( f(r_i) \) is usually approximated by pilot density estimate based on fixed \( h \).

### 3 Estimation of the marginal density \( f_m(r_t) \)

For the univariate marginal density we used adaptive bandwidth with pilot estimate using constant bandwidth determined by the rule of thumb \( (h = 0.0105) \). The resulting estimate is in figure 3. It is obvious that parametric distributions would not work well to estimate this density. On the other hand, a
Figure 3: Gaussian kernel density estimate of the marginal density $f_m(r_t)$. 

A typical problem connected with using kernel densities to estimate densities on bounded domain (interest rate exist only on positive domain) is the boundary bias. Using the adaptive bandwidth suppresses most of the boundary bias at the lower tail area. Unfortunately some negative values of $r_t$ close to zero have still nonzero density.

4 Generalization to the multivariate case

Estimating the joint density $f_j(r_{t+1}, r_t)$ is more challenging because the density is two-variate. Univariate kernel densities can be generalized to multivariate kernel densities using multivariate kernels $K_d$ (two dimensional normal density in our case). For this purpose we can not use the usual multiplicative kernel

$$K_2(u_1, u_2) = K(u_1) \cdot K(u_2),$$

because the correlation between both variables (autocorrelation with lag 1) is extreme ($\rho(r_{t+1}, r_t) = 0.9967$). In this case we have to use the whole covariance matrix as the bandwidth (both the standard deviations and the covariance). We will denote (dropping the time indexes) $r = (r_{t+1}, r_t)$ and the two-dimensional data vector $r_i$. The generalization of the formula (2) for the multivariate case is

$$\hat{f}_j(r_{t+1}, r_t) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\text{det}(H)} K_d [H^{-1}(r - r_i)],$$

where $H$ is the bandwidth matrix.

4.1 Bandwidth matrix optimization

The algorithms for optimizing the bandwidth matrix are similar as in univariate case. Again, we can use the rule of thumb, which is derived analogically as in the univariate case. The formula is (see [5]):

$$\hat{H} = \left( \frac{4}{d+2} \right)^{1/(d+4)} \hat{\Sigma}^{1/2} n^{-1/(d+4)},$$

where $\hat{\Sigma}^{1/2}$ is the matrix of square root elements of the sample covariance matrix and $d$ is the density dimension.

The log likelihood CV works in the multivariate case as well. The log likelihood CV score function is the same as that used above in the univariate case.
Because of the unequal density of the data, the adaptive bandwidth was needed again. We can generalize the formula (6):

\[
\hat{f}_j(r) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\text{det}(H_i)} K_d \left[ H_i^{-1}(r - r_i) \right],
\]

where

\[
H_i = \lambda_i H
\]

\[
\lambda_i = \left( \frac{G}{f(r)} \right)^{0.5}
\]

5 Estimation of the joint density \( f_j(r_{t+1}, r_t) \) and the transition density \( f(r_{t+1} | r_t) \)

As the pilot estimate in (14), we used a normalized three-dimensional histogram. The bin width of the histogram was equal to the "rule of thumb" estimate of the fixed bandwidth. Figure 4 shows the kernel density estimate of the joint density \( f_j(r_{t+1}, r_t) \). Not surprisingly, we can see that most of the density is concentrated around the diagonal. Most of the intra-day differences were close to zero. Higher differences (although some of them were observed) are very improbable.

The kernel density estimate of the transition density \( f(r_{t+1} | r_t) \) is in figure 5. The density was estimated from the formula (1). Again, we can see that most of the density is concentrated around the diagonal. That means, that the transition of the process to values distant from \( r_t \) is highly improbable (but the density in the areas distant from the diagonal is not exact zero).

We can use the estimates to investigate the moments of the transition density by numerically integrating them. For example we can forecast the value of the process in the time \( t + 1 \) as \( E(r_{t+1} | r_t) \) or the risk of the process \( \sigma(r_{t+1} | r_t) \) if \( r_t \) is known.

Another possibility would be to use the transition density estimate to simulate the process using some of the advanced non uniform random number generators and count risk measures, like Value at Risk (VaR).

Figure 4: Kernel density estimate of the joint density \( f_j(r_{t+1}, r_t) \).

\( \text{Normalized so the total volume of the bars is equal to one} \)
Figure 5: Kernel density estimate of the transition density $f(r_{t+1}|r_t)$.

References


Pavel Zimmermann
University of Economics Prague
Department of Statistics and Probability
W. Churchill Sq. 4
130 67 Prague 3
Czech Republic
zimmerp@vse.cz

Martin Cicha
University of Economics Prague
Department of Statistics and Probability
W. Churchill Sq. 4
130 67 Prague 3
Czech Republic
cicham@vse.cz