Non-parametric estimation of time-homogeneous diffusion processes

German Shâma Wache [1] Patrice Takam Soh [2]

[1]MSc AI For Science, AIMS South Africa [2]Professor, University of Yaounde 1, Cameroon

shama@aims.ac.za, takamsoh@fac-uy1.com

Motivation

In financial and economic systems, interest rates, exchange rates, market prices, macroeconomic factors and more are generally modeled by a time-homogeneous diffusion process, i.e. a stochastic process $\{X\}_{t\geq 0}$ driven by a SDE (stochastic differential equation) of the form:

$$dX_t = a(X_t)dt + b(X_t)dW_t \tag{1}$$

The difficulty with this modelling lies in the lack of knowledge about parametric forms of the drift function $a : \mathbb{R} \longrightarrow \mathbb{R}$ and the diffusion function $b : \mathbb{R} \longrightarrow \mathbb{R}$. Hence the need to estimate them.

Background

Two main trends:

1. The first trend consisted of proposing non-parametric estimators of the drift a and the diffusion b based on continuous-time observations. Some authors: Geman (1979), Tuan (1981), Soulier (1998), etc.
2. The second and current trend consist of proposing non-parametric estimators of a and b based on discrete-time observations. Some estimatorauthors: Nicolau (2003), Ziao and Hong (2015), Fabienne comte (2019), etc.

Having obtained non-parametric estimators for a and b, it is important to check their performance by studying their asymptotic properties (consistency and asymptotic normality). In contrast to the standard assumptions made in the literature [2] on this last question (stationarity, markovianity and ergodicity of the process $\{X_t\}_{t\geq 0}$), we were able to prove these results under much weaker assumptions.

Assumptions

$\mathcal{A}_1$: $\{X_t\}_{t\geq 0}$ is stationary with density f

$\mathcal{A}_2$: $\{X_t\}_{t\geq 0}$ is ρ-mixing, i.e. the ρ-mixing coeficient satisfies

$$\rho(k) := \sup_{\substack{X\in L^2(\mathcal{F}^t_{-\infty}) \\ Y\in L^2(\mathcal{F}^{+\infty}_{t+k})}} |corr(X, Y)| \underset{k\to+\infty}{\longrightarrow} 0$$

where $\mathcal{F}^t_{-\infty} = \sigma(X_s : s \leq t)$ and $\mathcal{F}^{+\infty}_{t+k} = \sigma(X_s : s \geq t + k)$.

Observations

$n + 1$ $(n \geq 1)$ observations $X_0, X_\Delta, \ldots, X_{n\Delta}$ of the process $\{X_t\}_{t\geq 0}$ at the respective discrete times $t_0 = 0, t_1 = \Delta, \ldots, t_n = n\Delta$ are made, equidistant in time, with step $\Delta = \Delta_n > 0$. Non-parametric estimators $\widehat{a}_n$ and $\widehat{b}^2_n$ of a and b^2 respectively will be functions of these observations.

Method: Nadaraya-Watson

Let X and Y be real random variables. We assume we have n observations $(X_0, Y_0), \ldots, (X_{n-1}, Y_{n-1})$ of the couple (X, Y).

Definition

A **kernel** is any non-negative and integrable application $K : \mathbb{R} \longrightarrow \mathbb{R}$ satisfying $\int_{-\infty}^{+\infty} K(u)du = 1$.

Definition

The **regression function of Y on X** is the function $r : X(\Omega) \longrightarrow \mathbb{R}$ defined by: $r(x) = \mathbb{E}(Y|X = x), \quad \forall x \in X(\Omega)$.

Nadaraya-Watson estimator of the regression function

The Nadaraya-Watson estimator [1] of the regression function r is given by:

$$\widehat{r}_n(x) = \frac{\sum_{i=0}^{n-1} Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=0}^{n-1} K\left(\frac{X_i - x}{h}\right)}, \quad \forall x \in X(\Omega)$$

where $h = h_n > 0$ is a small positive quantity called bandwith and K is a kernel well chosen.

Proposition

1. $a(x) = \mathbb{E}\left[\dfrac{X_{t+\Delta} - X_t}{\Delta}\middle| X_t = x\right] + O(\Delta)$
2. $b^2(x) = \mathbb{E}\left[\dfrac{(X_{t+\Delta} - X_t)^2}{\Delta}\middle| X_t = x\right] + O(\Delta), \qquad \forall x \in X_0(\Omega).$

Corollary (Non-parametric estimators of a and b^2)

1. $\widehat{a}_n(x) = \dfrac{\sum_{i=0}^{n-1} \frac{X_{t_{i+1}} - X_{t_i}}{\Delta} K\left(\frac{X_{t_i} - x}{h}\right)}{\sum_{i=0}^{n-1} K\left(\frac{X_{t_i} - x}{h}\right)}$
2. $\widehat{b}^2_n(x) = \dfrac{\sum_{i=0}^{n-1} \frac{(X_{t_{i+1}} - X_{t_i})^2}{\Delta} K\left(\frac{X_{t_i} - x}{h}\right)}{\sum_{i=0}^{n-1} K\left(\frac{X_{t_i} - x}{h}\right)}, \qquad \forall x \in X_0(\Omega).$

Asymptotic properties of our estimators

1. **Consistency:** $\widehat{a}_n(x) \underset{n\to+\infty}{\xrightarrow{\mathbb{P}}} a(x)$ and $\widehat{b}^2_n(x) \underset{n\to+\infty}{\xrightarrow{\mathbb{P}}} b^2(x)$.
2. **Asymptotic normality:**

$\sqrt{nh\Delta}[\widehat{a}_n(x) - a(x)] \underset{n\to+\infty}{\xrightarrow{d}} \mathcal{N}\left(0, \dfrac{b^2(x)R(K)}{f(x)}\right)$ and

$\sqrt{nh}[\widehat{b}^2_n(x) - b^2(x)] \underset{n\to+\infty}{\xrightarrow{d}} \mathcal{N}\left(0, \dfrac{3b^4(x)R(K)}{f(x)}\right)$. In order to establish that asymptotic normality, we had to develop a new CLT presented below.

Theorem 1 (A new central limit theorem)

Let $\{U_i = U_{n,i}\}_{i=1,\ldots,n}$ be a random variables sequence. Suppose that:

1. There exists a measurable function $g_n : \mathbb{R}^2 \longrightarrow \mathbb{R}$ depending on n such that $U_i = g_n(X_{t_{i-1}}, X_{t_i}), \quad \forall i \in \{1, \ldots, n\}$
2. $U_i \underset{n\to+\infty}{=} o\left(\dfrac{1}{h\Delta}\right), \quad \forall i \geq 0$
3. $\mathbb{E}[U_i] = o(1)$ and $\mathbb{V}ar[U_i] = s^2 + o(1), \quad \forall i \geq 0$

Then, $\dfrac{1}{\sqrt{n}} S_n := \dfrac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \underset{n\to+\infty}{\xrightarrow{d}} \mathcal{N}(0, s^2)$.

Application: Modelling U.S Dollars to Euro spot exchange rate

We have a sample of data of sizedW 6385 $\{Z_t\}_{t=1,\ldots,6385}$ provided by *Board of governors of the Federal reserve system (US)*. These data represent daily U.S Dollars to Euro spot exchange rate from January 04, 1999 to June 23, 2023. A stationarity test shows that the time serie $\{Z_t\}_{t=1,\ldots,6385}$ is not stationary, but its differenced serie $\{X_t = Z_t - Z_{t-1}\}_{t=2,\ldots,6384}$ is so. Using the data of that latter serie and plotting our non-parametric estimators, the one of the drift shows the trend of an oblique straight line with negative slope meanwhile the one of the diffusion shows the trend of a horizontal line, so that the differenced serie can be model with a **Vasicek** SDE: $dX_t = \alpha(\theta - X_t)dt + \sigma dW_t$. Calibrating the parameters α, θ and σ using the **Euler's scheme** for discretisation and the **least square estimation method**, we end up with the following **Vasicek model:**

$$dX_t = 253.6566(-1.3682 \times 10^{-5} - X_t)\, dt + 0.4371\, dW_t \tag{2}$$

To validate our model, we perform the Euler's scheme on (2) to predict the rate values for the period from June 24, 2023 to September 08, 2023 and then compare these predicted values with the true values of these rates through performance indicators like: $MSE = 0.052$, $MAE = 0.0638$ and $MAPE = 5.84\%$. The MAPE informs us that, on average, forecasts are off by just 5.84%. This confirms that our model is good.

Conclusion

We've constructed and justified theoretically non-parametric estimators of the unknown terms in SDE (1). We finally applied our results in modelling U.S Dollars to Euro spot exchange rates and came up with the Vasicek model (2).

References

[1] Nadaraya. *On estimating regression.* Theory of Probability and its Applications, 1964.

[2] Xin Wang. *Online Non-parametric Estimation of Stochastic Differential Equations.* 2015.