Volly the Volatility Trader

Anchor Labs

February 2021

1 Problem Setup

Volly, the volatility trader runs proprietary trading algorithms in various crypto-currency markets. Volly's favorite market is the BTC/USDT pair, which has very high volume offering good liquidity and generally high volatility. We model the daily performance of Volly's algorithms as an output $O_t \in \{-1, +1\}$. A simple stochastic model for the performance of Volly's algorithms is a Bernoulli random variable with two separate regimes: low volatility days and high volatility days. We can treat the daily returns as a suitable proxy for intraday volatility leading to the following model:

$$O_t \mid |R_t| \leq \alpha = \begin{cases} +1, \text{ w.p} & \theta_L \\ -1, \text{ w.p } & 1 - \theta_L \end{cases} \tag{1}$$

$$O_t \mid |R_t| > \alpha = \begin{cases} +1, \text{ w.p} & \theta_H \\ -1, \text{ w.p } & 1 - \theta_H \end{cases} . \tag{2}$$

Here R_t is the daily arithmetic returns at time t as

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}. \tag{3}$$

Volly's algorithms generally perform well on on days with large price swings and and perform worse on days of smaller price swings, so we expect $\theta_L < \theta_H$. In this exercise, we use this model to estimate the performance of a particular market makers Profit and Loss (PNL) defined here as the cumulative algorithm output over time.

2 Questions

You are given a file containing the following columns: (1) Time; (2) BTC/USDT price; (3) Volly's aggregate PNL.

1. Load the provided file and visualize the following elements:
   (a) Price series: BTC/USDT price vs. time
   (b) Returns series: use arithmetic returns as opposed to geometric.
   (c) PNL series vs time.
   (d) Daily output series O_t.
2. Given returns R_t and daily algorithm output O_t, propose a methodology to estimate the following parameters: $\lambda = \{\alpha, \theta_L, \theta_H\}$. Formulate a reasonable objective function.
3. Propose a methodology for arrive at the parameters of interest.
4. (a) How does your methodology scale with the length of the time-series: give an $O(T)$ estimate? (b) Do you believe your methodology is optimal in terms of parameter estimation and time complexity? (c) How would you modify your algorithm if the data was not a single file but distributed between various machines. (This would not be done on daily data but much more granular series).