Bachelor’s thesis

Machine Learning & High Frequency Time Series

Predicting short-term regime shifts via Hidden Markov Models

Egil Furnes
egilsf@uio.no

Department of Mathematics
Faculty of Mathematics and Natural Sciences

17th February 2026

Abstract

High-frequency foreign exchange (FX) markets exhibit complex behaviors that challenge standard linear time series models. While daily returns often appear random, tick-by-tick data reveals significant short-term dependence driven by market microstructure effects. This project investigates the modeling of time-varying short-term dependence in FX data using a Regime-Switching Autoregressive model formulated as a Hidden Markov Model (HMM). By decomposing the observed price into a latent efficient price and microstructure noise, we apply pre-averaging techniques to mitigate noise before estimating dependence regimes. We aim to capture the switching behavior between noise-dominated periods and momentum-driven price runs, comparing the HMM approach against static baseline models and online adaptive estimators.

Acknowledgements

I would like to thank my supervisor, Associate Professor Gudmund Horn Hermansen (Department of Mathematics, University of Oslo), for his guidance and support in the fields of Statistics and Data Science throughout this project.

1 Introduction

Financial time series are well known to exhibit non-stationarity, volatility clustering, and structural change. While standard statistical learning techniques, such as those described by Hastie et al. (2009), provide robust frameworks for general data analysis, high-frequency financial data requires specialized treatment. As noted by Hautsch (2012), the analysis of tick-by-tick data introduces unique challenges related to the discreteness of price changes and the irregular spacing of transactions.

The presence of time-varying dependence in financial returns has been a subject of extensive research. Hamilton (1989) introduced a seminal framework for modeling regime changes in economic time series, and further expanded it in 1944 (Hamilton 1994). However, in the context of high-frequency market structure, Lo and MacKinlay (1990) and Engle (1982) identified that short-horizon return autocorrelation is not constant. Modern approaches often require robust estimation techniques; Jacod et al. (2017) and Aït-Sahalia and Jacod (2014) provide critical methodologies for separating microstructure noise from the efficient price. To model the latent states driving these market dynamics, we rely on the Hidden Markov Model (HMM) framework detailed by Rabiner (1989), while also considering the broader financial time series context provided by Tsay (2010).

1.1 Bid–ask bounce

The bid-ask bounce is a fundamental microstructure phenomenon where transaction prices oscillate between the bid and ask quotes even if the underlying asset value remains unchanged. This bouncing effect induces a negative autocorrelation in returns at the highest frequencies, creating a "noise" regime that can obscure genuine price trends.

1.2 Order flow imbalance

Order flow imbalance refers to the disparity between aggressive buy and sell orders. When order flow is heavily skewed in one direction, it can deplete liquidity at the best quote, causing the price to move. Unlike the mean-reverting bid-ask bounce, strong order flow imbalance can induce short-term momentum or positive serial dependence in returns.

1.3 Asynchronous trading

Asynchronous trading occurs because assets trade at different frequencies and times. Lo and MacKinlay (1990) demonstrated that this non-synchronicity can induce spurious cross-autocorrelations between assets. In the context of a single FX pair, irregular arrival times of information and trades contribute to the stochastic nature of the volatility and dependence structure we aim to model.

2 Method

Let P_τ denote the observed transaction price at tick time τ. The observed price can be decomposed as shown in Equation 1:

$$P_\tau = P_\tau^* + \eta_\tau \tag{1}$$

where P_τ^* is the latent efficient price and η_τ represents microstructure noise. The pre-averaged price is defined in Equation 2 as

$$\bar{P}_t = \frac{1}{k}\sum_{j=0}^{k-1} P_{t-j} \tag{2}$$

where t now indexes the last tick in each averaging block. Returns are then constructed as shown in Equation 3:

$$r_t = \bar{P}_t - \bar{P}_{t-1} \tag{3}$$

2.1 Baseline Auto-regressive Model

As a benchmark, we consider a single-regime auto-regressive model of order one, given by Equation 4:

$$r_t = \mu + \phi r_{t-1} + \varepsilon_t \tag{4}$$

where μ is the unconditional mean return, ϕ measures linear serial dependence and ε_t is an innovative term with $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.

2.2 Hidden Markov Model with Regime-Switching Auto-regressive Dynamics

Let $S_t \in \{1, \dots, K\}$ denote an unobserved discrete-time Markov chain representing the dependence regime at time t. The transition probabilities are defined in Equation 5:

$$P(S_t = j \mid S_{t-1} = i) = p_{ij} \tag{5}$$

with $p_{ij} \geq 0$ and $\sum_{j=1}^{K} p_{ij} = 1$ for all i.

Conditional on the regime $S_t = k$, returns follow an auto-regressive process described by Equation 6:

$$r_t = \mu_k + \phi_k r_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_k^2) \tag{6}$$

Here μ_k is the regime-specific mean, ϕ_k captures the strength of the short-term dependence in regime k and σ_k^2 is the regime-specific innovation variance.

Regimes with ϕ_k close to zero are interpreted as noise-dominated periods, while regimes with larger positive ϕ_k correspond to short-term momentum or price returns.

2.3 Modeling volatility

Allowing σ_k^2 to vary across regimes captures abrupt changes in market uncertainty. As an extension, conditional heteroskedasticity can be introduced via a regime-dependent ARCH specification, shown in Equation 7:

$$\sigma_{k,t}^2 = \omega_k + \alpha_k \varepsilon_{t-1}^2 \tag{7}$$

where $\omega_k \geq 0$ and $\alpha_k \geq 0$. Building on this, volatility persistence can be incorporated through a regime-dependent GARCH specification, see Equation 8:

$$\sigma_{k,t}^2 = \omega_k + \alpha_k \varepsilon_{t-1}^2 + \beta_k \sigma_{k,t-1}^2, \tag{8}$$

where $\beta_k \geq 0$. This extension allows volatility shocks to decay gradually over time within each regime rather than affecting only the next period.

2.4 Data

Data was retrieved from HistData.com (2026).

3 Result

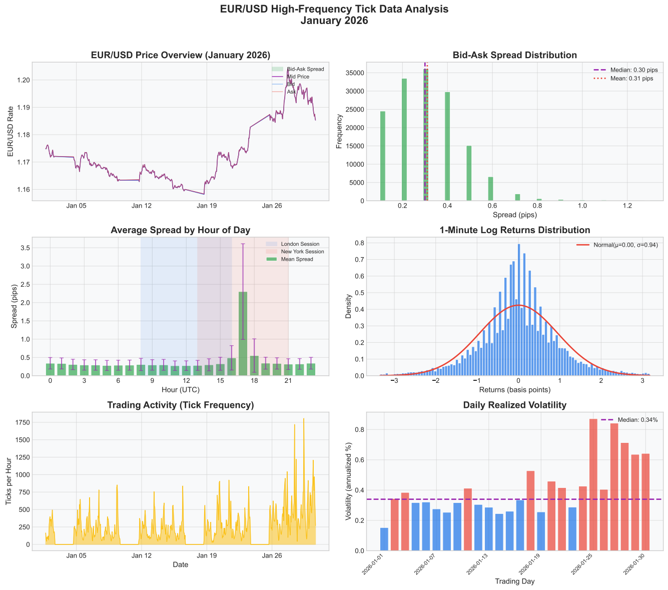

Figure 1: Exploratory data analysis of EUR/USD tick data for January 2026. The panels display the price evolution and bid-ask spread distribution (top); intraday spread seasonality highlighting the London and New York sessions, alongside the leptokurtic distribution of 1-minute log returns compared to a Gaussian benchmark (middle); and hourly trading activity frequency alongside annualized daily realized volatility (bottom).

4 Project Plan

- **February:** Literature review on HMMs and high-frequency econometrics. Data acquisition from HistData (EUR/USD) and implementation of pre-averaging techniques to handle microstructure noise.
- **March:** Implementation of the baseline single-regime Autoregressive (AR) model. Development of the Hidden Markov Model (HMM) with regime-switching AR dynamics in Python/R.
- **April:** Extension of the model to include regime-dependent volatility (ARCH/GARCH effects). Calibration and validation of the model against the baseline using in-sample and out-of-sample testing.
- **May (1–15):** Final interpretation of regime-switching results (identifying noise vs. momentum states), final write-up of the thesis, and preparation of the code repository.

References

Aït-Sahalia, Yacine and Jean Jacod (2014). *High-Frequency Financial Econometrics.* Princeton University Press. URL: https://press.princeton.edu/books/hardcover/9780691161433/high-frequency-financial-econometrics.

Engle, Robert F. (1982). 'Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation'. In: *Econometrica* 50.4, pp. 987–1007. DOI: 10.2307/1912773. URL: https://doi.org/10.2307/1912773.

Hamilton, James D. (1989). 'A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle'. In: *Econometrica* 57.2, pp. 357–384. DOI: 10.2307/1912559. URL: https://doi.org/10.2307/1912559.

— (1994). *Time Series Analysis.* Princeton University Press. URL: https://press.princeton.edu/books/hardcover/9780691042893/time-series-analysis.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman (2009). *The Elements of Statistical Learning.* 2nd ed. Springer. ISBN: 978-0387848570.

Hautsch, Nikolaus (2012). *Econometrics of Financial High-Frequency Data.* Berlin, Heidelberg: Springer. ISBN: 978-3-642-21925-2. DOI: 10.1007/978-3-642-21925-2.

HistData.com (2026). *Free Forex Historical Data.* Accessed 2026-02-09. HistData.com. URL: https://www.histdata.com/ (visited on 09/02/2026).

Jacod, Jean et al. (2017). 'Between Data Cleaning and Inference: Pre-averaging and Robust Estimators of the Efficient Price'. In: *Journal of Econometrics* 196.1, pp. 1–22. DOI: 10.1016/j.jeconom.2016.08.014. URL: https://doi.org/10.1016/j.jeconom.2016.08.014.

Lo, Andrew W. and A. Craig MacKinlay (1990). 'An Econometric Analysis of Nonsynchronous Trading'. In: *Journal of Econometrics* 45.1-2, pp. 181–211. DOI: 10.1016/0304-4076(90)90098-E. URL: https://doi.org/10.1016/0304-4076(90)90098-E.

Rabiner, Lawrence R. (1989). 'A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition'. In: *Proceedings of the IEEE* 77.2, pp. 257–286. DOI: 10.1109/5.18626. URL: https://doi.org/10.1109/5.18626.

Tsay, Ruey S. (2010). *Analysis of Financial Time Series.* 3rd ed. Wiley. URL: https://www.wiley.com/en-us/Analysis+of+Financial+Time+Series%2C+3rd+Edition-p-9780470414354.