Autoregressive (AR), GBM Process & Volatility Estimators

Theory, Estimation and Diagnostics

Cacciamani Alberto, Visconti Tommaso

December 1, 2025

Introduction

- Autoregressive (AR) models are fundamental in time series analysis.
- Capture dependence of current value on past values.
- Useful for forecasting and understanding dynamics of stochastic processes.

Definition of AR(p)

- General form:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

- $\epsilon_t \sim WN(0, \sigma^2)$, i.i.d.
- ϵ_t represents the unpredictable component (white noise).

Stationarity Definition

- In general, an AR(p) is stationary if the roots of the characteristic polynomial lie outside the unit circle.

For an autoregressive process AR(p), the characteristic polynomial is defined as:

$$\phi(z) = 1 - a_1 z - a_2 z^2 - \cdots - a_p z^p$$

So, it is stationary if:

$$|z_i| > 1 \quad \forall i$$

This condition ensures that the effect of past shocks decays over time and the process does not diverge.

Stationarity Consequences

As a consequence of stationarity:

1. The **mean** is constant:

$$\mathbb{E}[y_t] = \frac{a_0}{1 - a_1 - \cdots - a_p}$$

2. The **variance** is finite and constant (for AR(1) we have the following, for p>1 solve Yule-Walker system):

$$\mathrm{Var}(y_t) = \frac{\sigma^2}{1 - a_1^2}$$

3. The **autocovariance** depends only on the lag k:

$$\gamma_k = \mathrm{Cov}(y_t, y_{t-k})$$

Supplement: Yule-Walker Equations

The Yule-Walker equations relate the autocovariances of a stationary AR(p) process to its parameters.
For an AR(p) process:

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t, \quad \epsilon_t \sim WN(0, \sigma^2)$$

The equations are:

$$\begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_p \end{bmatrix}$$

- Can be solved to obtain the AR parameters ϕ_i from sample autocovariances.
- Variance of the noise:

$$\sigma^2 = \gamma_0 - \sum_{i=1}^{p} \phi_i \gamma_i$$

Autocovariance and Autocorrelation

- The **autocovariance function** at lag k is defined as:

$$\gamma_k = \text{Cov}(y_t, y_{t-k}) = \mathbb{E}[(y_t - \mu)(y_{t-k} - \mu)]$$

  where $\mu = \mathbb{E}[y_t]$ is the mean of the process.
- Note that $\gamma_0 = \text{Var}(y_t)$ is the variance of the process.
- The **autocorrelation function** (ACF) at lag k is:

$$\rho_k = \frac{\gamma_k}{\gamma_0}$$

  which measures the correlation between y_t and y_{t-k}.
- For a stationary AR process, γ_k and ρ_k depend only on the lag k, not on t.

Fitting AR(p) OLS: Idea principale

- The OLS method estimates the parameters by **minimizing the sum of squared errors**:

$$S(a_0, a_1, \ldots, a_p) = \sum_{t=p+1}^{T} \left(y_t - \hat{y}_t\right)^2 = \sum_{t=p+1}^{T} \left(y_t - \sum_{i=0}^{p} a_i y_{t-i}\right)^2$$

- Goal: find parameters $a_0, \ldots, a_p$ that minimize this function.

Fitting AR(p) OLS: Derivation

- Take the derivative of the sum of squared errors w.r.t. each parameter:

$$\frac{\partial S}{\partial a_i} = -2 \sum_{t=p+1}^{T} \left(y_t - \sum_{j=0}^{p} a_j y_{t-j}\right) y_{t-i}$$

- Set each derivative to zero (first-order condition):

$$\frac{\partial S}{\partial a_i} = 0 \quad \forall i = 0, \ldots, p$$

- Solving this system gives the OLS estimates.

Fitting AR(p) OLS: Matrix Form

- Define the response vector Y and design matrix X:

$$Y = \begin{bmatrix} y_p \\ y_{p+1} \\ \vdots \\ y_T \end{bmatrix}, \quad X = \begin{bmatrix} 1 & y_{p-1} & \cdots & y_0 \\ 1 & y_p & \cdots & y_1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & y_{T-1} & \cdots & y_{T-p} \end{bmatrix}$$

- OLS solution:

$$\hat{a} = (X^\top X)^{-1} X^\top Y$$

- Provides the estimates of all AR parameters at once.

Maximum Likelihood Estimation (MLE)

- The MLE finds parameter values that maximize the probability of observing the given data.

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta \mid \text{data})$$

or equivalently using the log-likelihood:

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta \mid \text{data})$$

- MLE provides consistent and asymptotically efficient estimates under regularity conditions.

AR(p) with Gaussian Innovations

For an AR(p) process:

$$y_t = a_0 + a_1 y_{t-1} + \cdots + a_p y_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2)$$

The log-likelihood function is:

$$\ell(a_0, \ldots, a_p, \sigma^2) = -\frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=p}^{T}(y_t - \mu_t)^2$$

where

$$\mu_t = a_0 + a_1 y_{t-1} + \cdots + a_p y_{t-p}$$

AR(p) with Student-t Innovations

$$y_t = a_0 + a_1 y_{t-1} + \cdots + a_p y_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim t_\nu(0, \sigma^2)$$

Log-likelihood:

$$\ell(a_0, \ldots, a_p, \sigma^2, \nu) = \sum_{t=p}^{T} \log \left[\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma^2}} \left(1 + \frac{(y_t - \mu_t)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} \right]$$

- MLE requires numerical optimization (e.g., `scipy.optimize.minimize`).
- As $T \to \infty$, MLEs converge to true parameters under regularity conditions.
- Student-t innovations handle heavy-tailed shocks, unlike Gaussian.

Residual Analysis

- Residuals:

$$\hat{\epsilon}_t = y_t - \hat{y}_t$$

- Purpose: assess model fit and detect misspecification
- Check distribution:
  - Histogram: visually inspect normality
  - QQ-plot: compare residual quantiles to theoretical distribution
- Check moments:
  - Mean ≈ 0
  - Variance $\approx \sigma^2$
- Check autocorrelation:
  - ACF of residuals should be near zero for all lags
  - Significant autocorrelation indicates unmodeled structure

Practical Considerations

- Diagnostics are crucial:
  - Residual plots
  - Stationarity checks (unit root tests, ACF decay)
  - Information criteria (AIC, BIC) for model selection
- Limitations:
  - AR models assume linearity and stationarity
  - May not capture heavy tails or volatility clustering
- For heavy-tailed shocks, consider Student-t innovations or GARCH extensions

The Underlying Price Process (GBM)

We assume that the asset price S_t follows a **Geometric Brownian Motion** (GBM), described by the Stochastic Differential Equation (SDE):

$$dS_t = \mu S_t dt + \sigma S_t dW_t$$

where:

- μ: Constant drift (expected return).
- σ: Constant volatility.
- W_t: Standard Wiener process.

The analytical solution used for simulation is:

$$S_t = S_0 \exp\left(\left(\mu - \frac{1}{2}\sigma^2\right) t + \sigma W_t\right)$$

Statistical Validation of GBM Process

Firstly, we needed to compute some statistics and test on the GBM Process in order to have statistical theoretical consistency. So, we need to verify that the Log-Returns $r_t = \ln(S_t/S_{t-1})$ follow a Normal Distribution:

$$r_t \sim \mathcal{N}\left((\mu - \frac{1}{2}\sigma^2)\Delta t, \sigma^2 \Delta t\right)$$

To do that we computed the first four moments (Mean, Variance, Skewness, Kurtosis) and performed the **Jarque-Bera test** to validate the normality assumption.

Jarque-Bera Test

The **Jarque-Bera (JB) Test** tests whether sample data have the skewness and kurtosis matching thst of a normal distribution.

Hypotheses:

- H_0: The data is normally distributed $(S = 0, K = 3)$.
- H_1: The data is NOT normally distributed.

Test Statistic:

$$JB = \frac{T}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

where:

- T is the sample size.
- S is the sample skewness.
- K is the sample kurtosis.

Asymptotic Distribution:

$$JB \xrightarrow{d} \chi^2_{(2)}$$

Realized Volatility Verification

In order to further validate the simulation, we also need to compute the **Realized Volatility (RV)** for each of the N generated paths.

$$RV = \sum_{t=1}^{N} r_t^2$$

We verify that the mean of the empirical Realized Volatilities across all paths converges to the theoretical expected variance (Integrated Volatility) used in the simulation parameters ($\sigma^2 T$).

Martingale Property Test

Furthermore, we then check that the discounted price process $\tilde{S}_t = e^{-\mu t} S_t$ must be a martingale under the measure $\mathbb{P}$:

$$\mathbb{E}[\tilde{S}_t | \mathcal{F}_0] = S_0$$

We verify this property in two ways:

- **Visual Inspection**: Checking if the mean of the simulated discounted paths converges to the initial price S_0.
- **Statistical Test**: Performing a T-test on the final values to ensure there is no significant bias.

Two-Scales Realized Volatility (TSRV)

We introduce **microstructure noise**: $\epsilon_t \sim \mathcal{N}(0, \eta^2)$ to the efficient price X_t (log-price) to simulate real-world high-frequency data:

$$Y_t = X_t + \epsilon_t$$

We then compare two estimators used for Volatility:

- **Realized Volatility (RV)**: Which is biased and inconsistent under noise.
- **TSRV**: Which corrects for the bias using subsampling (Fast and Slow time scales).

Optimization of Sparse Parameter (K)

Let us observe how the TSRV Estimator is built:

$$\widehat{\langle X, X \rangle}_T = \underbrace{\left(1 - \frac{\bar{n}}{n}\right)^{-1}}_{\text{Correction}} \left(\underbrace{[Y, Y]_T^{(avg)}}_{\text{Slow Scale}} - \frac{\bar{n}}{n} \underbrace{[Y, Y]_T^{(all)}}_{\text{Fast Scale}} \right)$$

As seen in (Zhang, Mykland, Aït-Sahalia, 2005), the TSRV estimator depends on a sparsity parameter K (or n_{sparse}).
To do so, we perform a **Grid Search** to find the optimal K that minimizes the Mean Squared Error (MSE) against the true Integrated Volatility.

Optimization of Sparse Parameter (K)

The performance of the TSRV estimator depends critically on the choice of the time scale parameter K (or n_{sparse}).

The Bias-Variance Trade-off

The parameter K balances two competing errors:

1. **Noise Reduction (Bias)**: A larger K effectively filters out more microstructure noise (reducing bias).
2. **Information Loss (Variance)**: A larger K reduces the number of data points available for the "Slow Scale" estimator, increasing the estimator's variance.

Optimization Approach (Grid Search)

The **Optimization Approach** used is the **Grid Search**.
To find the optimal K, we minimize the Mean Squared Error (MSE) against the True Integrated Volatility (IV_{true}):

$$K_{opt} = \arg\min_{K} \mathbb{E}\left[(\widehat{IV}_{TSRV}(K) - IV_{true})^2\right]$$

Note: Zhang et al. (2005) derived an asymptotic optimal rate of $K \sim cn^{2/3}$.

Consistency Check: Volatility Signature Plot

We then want to test the consistency of the TSRV Estimator, through a specific visual graph.
The **Volatility Signature Plot** analyzes the behavior of realized volatility estimators as the sampling frequency N increases (i.e., as $\Delta t \to 0$).

- **Realized Volatility (RV):**
  - Inconsistent. As $N \to \infty$, the estimator diverges.
  - It ends up estimating the variance of the microstructure noise rather than the price volatility.
- **Two-Scales RV (TSRV):**
  - Consistent. Remains stable and converges to the True Integrated Volatility.

Consistency Check: Volatility Signature Plot

In this section we then study the theoretical asymptotics directly from the paper Zhang et al., 2005.
We find that, as the number of observations $n \to \infty$:

$$RV_{all} \xrightarrow{p} \underbrace{\int_0^T \sigma_t^2 dt}_{\text{True IV}} + \underbrace{2n\mathbb{E}[\epsilon^2]}_{\text{Noise (Explodes)}}$$

$$TSRV \xrightarrow{p} \int_0^T \sigma_t^2 dt \quad \text{(Noise is canceled)}$$

Example - High-Frequency Simulation Parameters

To replicate microstructure noise effects, we simulate a **tick-by-tick** environment (1-second resolution).

Simulation Settings

- **Trading Duration (s):** 23,400 seconds (6.5 hours).
- **Time Horizon (T):** 1 Day.
- **Time Step (dt):** $\frac{1}{252 \times 23400} \approx 1.69 \times 10^{-7}$.
- **Drift (μ):** 0.05 (5% annualized).
- **Volatility (σ):** 0.20 (20% annualized).
- **Scenarios (N_{paths}):** 10 independent paths.

This granular approach allows us to test the asymptotic properties of the estimators.

Example - Validation 1: Log-Returns Normality

We validate the simulation by analyzing the Log-Returns:

$$r_t = \ln(S_t) - \ln(S_{t-1})$$

For a GBM, r_t must be Normally distributed:

$$r_t \sim \mathcal{N}\left((\mu - \frac{1}{2}\sigma^2)\Delta t, \sigma^2 \Delta t\right)$$

Diagnostics performed:

- **Skewness:** Empirically close to 0.
- **Excess Kurtosis:** Empirically close to 0.
- **Jarque-Bera Test:** Checking the null hypothesis of Normality.

Example - Validation 2: Martingale Property

The discounted price process $\tilde{S}_t = e^{-\mu t} S_t$ must be a martingale under the physical measure $\mathbb{P}$.

$$\mathbb{E}[\tilde{S}_t | \mathcal{F}_0] = S_0$$

- **Visual Test:** The average of simulated paths should converge to the horizontal line $y = S_0$.
- **Statistical Test:** We perform a T-test on the final values $\tilde{S}_T$:

$$H_0 : \mu_{\tilde{S}_T} = S_0$$

Example - Microstructure Noise

Real high-frequency prices are contaminated by market microstructure noise (bid-ask bounce, discrete trading). We model the *observed price* Y_t as:

$$Y_t = X_t + \epsilon_t$$

where:

- X_t: True efficient log-price (latent).
- $\epsilon_t \sim \mathcal{N}(0, \eta^2)$: i.i.d. noise component.

Problem: Under noise, the standard **Realized Volatility (RV)** is biased and inconsistent (diverges as $N \to \infty$).

Example - Two-Scales Realized Volatility (TSRV)

To correct the bias, we use the TSRV estimator (Zhang, Mykland, Aït-Sahalia, 2005), which subsamples the data at two frequencies:

$$\widehat{\langle X, X \rangle}_T = \underbrace{\left(1 - \frac{\bar{n}}{n}\right)^{-1}}_{\text{Correction}} \left(\underbrace{[Y, Y]_T^{(avg)}}_{\text{Slow Scale}} - \frac{\bar{n}}{n} \underbrace{[Y, Y]_T^{(all)}}_{\text{Fast Scale}} \right)$$

- **Fast Scale:** Uses all data (estimates noise).
- **Slow Scale:** Average of RVs on subsampled grids (reduces noise).
- $\bar{n} \approx (n - K + 1)/K$.

Example - Optimization of Sparsity Parameter (K)

The performance of TSRV depends on the parameter K (or n_{sparse}). We perform a **Grid Search** to minimize the MSE against the True Integrated Volatility.

$$K_{opt} = \arg\min_{K} \mathbb{E}\left[(\text{TSRV}_K - IV_{true})^2\right]$$

Findings:

- Small K: Insufficient noise removal (High Bias).
- Large K: Too much data discarded (High Variance).
- Optimal K: Balances the trade-off (found empirically via simulation).

Example - Consistency Check: Volatility Signature Plot

The **Volatility Signature Plot** compares the estimators as sampling frequency N increases.

Expected Behavior

1. **RV (Realized Volatility):** Should diverge linearly upwards (estimating infinite variance of the noise).
2. **TSRV:** Should remain stable and converge to the True Integrated Volatility (IV_{true}).

This confirms that TSRV is a consistent estimator suitable for high-frequency data analysis.

Conclusion

- We successfully simulated a realistic HF environment using GBM.
- Statistical tests confirmed the robustness of the simulation engine (Normality, Martingale).
- In the presence of microstructure noise, RV proved to be unreliable.
- **TSRV** effectively filtered out the noise, providing a consistent estimate of the true volatility, provided K is optimally selected.

References

- Klaus Neusser, *Time Series Econometrics*.
- L. Zhang, P. A Mykland & Y. Aït-Sahalia, *A Tale of Two Time Scales*
- F. Corsi & R. Renò, *Discrete-Time Volatility Forecasting With Persistent Leverage Effect and the Link With Continuous-Time Volatility Modeling*