L06: Volatility Modeling & High-Frequency Finance

Extended Slides – BSc Digital Finance Course

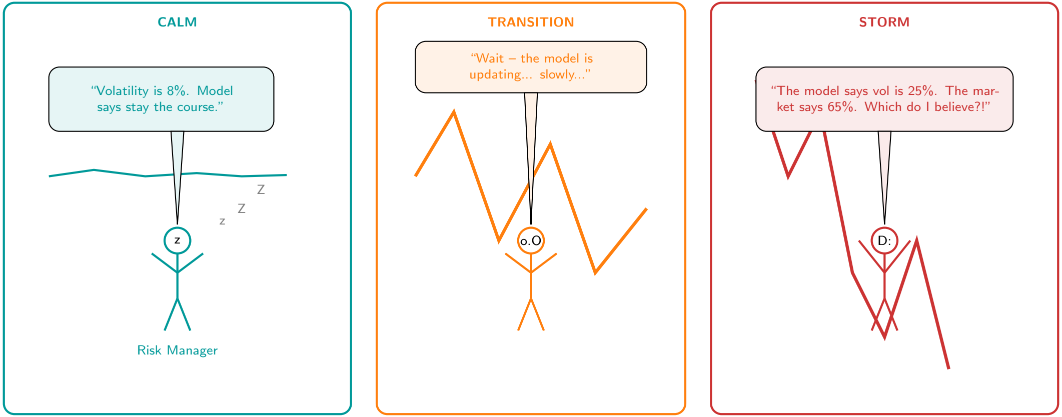
Digital Finance

What Will You Be Able to Do After This Lecture?

By the end of this extended lecture, you will be able to:

1. Derive and interpret the GARCH(1,1) conditional variance equation, including stationarity and persistence conditions
2. Estimate GARCH parameters via maximum likelihood in Python and interpret diagnostics
3. Compare GARCH, EGARCH, and GJR-GARCH for capturing asymmetric volatility (leverage effect)
4. Compute realized volatility from intraday data and explain the bias-variance tradeoff at high frequencies
5. Extract implied volatility from option prices and interpret the volatility surface (smile, skew, term structure)
6. Price variance swaps, test for random-walk departures, and apply volatility models to risk management

Six objectives: formal models (1,3), Python (2,4), applied evaluation (5,6).



The calm is when models work. The storm is when you need them. They are never both at the same time.

How Does One Equation Capture the Fact That Large Moves Predict Large Moves?

Let $r_t = \mu + \varepsilon_t$ where $\varepsilon_t = \sigma_t z_t$, $z_t \sim (0, 1)$. The **GARCH(1,1)** conditional variance:

$$\sigma_t^2 = \omega + \alpha\,\varepsilon_{t-1}^2 + \beta\,\sigma_{t-1}^2, \qquad \omega > 0,\ \alpha \geq 0,\ \beta \geq 0$$

Stationarity condition: $\alpha + \beta < 1$. The **unconditional variance** is $\bar{\sigma}^2 = \omega/(1 - \alpha - \beta)$.

Persistence: The half-life of a volatility shock is $h = \ln 2/\ln(1/(\alpha + \beta))$. Typical equity values: $\alpha \approx 0.05$, $\beta \approx 0.92 \Rightarrow h \approx 23$ days.

Multi-step forecast (assuming $E[\varepsilon_{t+k}^2 | \mathcal{F}_t] = \sigma_{t+k}^2$):

$$\sigma_{t+k|t}^2 = \bar{\sigma}^2 + (\alpha + \beta)^{k-1}(\sigma_{t+1|t}^2 - \bar{\sigma}^2)$$

The forecast **mean-reverts** to $\bar{\sigma}^2$ at rate $(\alpha + \beta)^k$. For $k \to \infty$, $\sigma_{t+k|t}^2 \to \bar{\sigma}^2$.

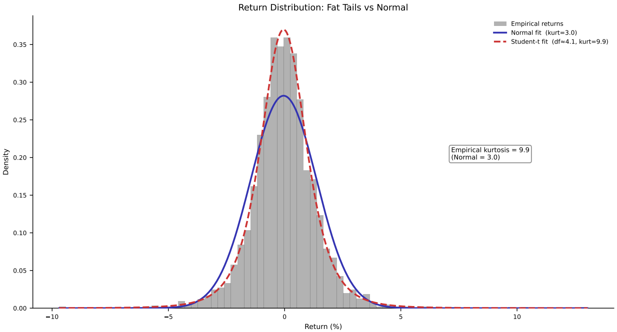
GARCH kurtosis (under normal innovations):

$$\kappa = 3 \cdot \frac{1 - (\alpha + \beta)^2}{1 - (\alpha + \beta)^2 - 2\alpha^2} > 3$$

Even with normal z_t, GARCH produces fat tails because the time-varying variance inflates the unconditional return distribution.

GARCH(1,1): clustering, mean reversion, and fat tails from three parameters and one feedback loop.

Why Does the Normal Distribution Fail at the Tails of Financial Returns?



Reading the overlay

- The **histogram** shows empirical returns simulated from a Student-t with $\nu = 4$ degrees of freedom – a stylized fact for daily equity returns
- The **purple line** is the normal fit. It captures the center but dramatically underestimates the tails
- The **red dashed line** is the Student-t fit. It matches both center and tails because it allows for excess kurtosis
- **Kurtosis:** Empirical kurtosis far exceeds 3.0 (the normal value), confirming fat tails
- **Practical impact:** A 4-sigma daily loss occurs roughly 10x more often than the normal predicts. VaR estimates based on normality are dangerously optimistic

Implication: GARCH with normal innovations captures clustering but still underestimates tail risk. Use Student-t or GED innovations for accurate VaR.

The normal fits the center but misses the tails. Fat tails are the norm.

How Do You Estimate GARCH Parameters from Return Data?

```python
1 import numpy as np
2 from scipy.optimize import minimize
3
4 def garch_loglik(params, returns):
5     """Negative log-likelihood for GARCH(1,1)."""
6     omega, alpha, beta = params
7     T = len(returns)
8     sigma2 = np.zeros(T)
9     sigma2[0] = returns.var()    # init
10    for t in range(1, T):
11        sigma2[t] = (omega
12            + alpha * returns[t-1]**2
13            + beta * sigma2[t-1])
14    # Gaussian log-likelihood
15    ll = -0.5 * np.sum(
16        np.log(2*np.pi*sigma2)
17        + returns**2 / sigma2)
18    return -ll                   # minimize
19
20 x0 = [1e-6, 0.05, 0.90]
21 res = minimize(garch_loglik, x0,
22     args=(returns,), method='L-BFGS-B',
23     bounds=[(1e-8,1),(1e-8,1),(1e-8,1)])
```

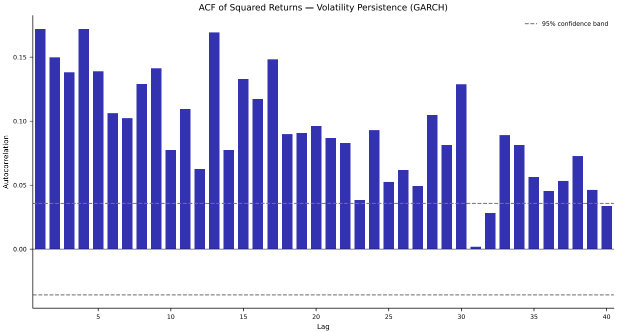
Code walkthrough

- **Lines 4–5:** Three parameters – ω (baseline), α (shock), β (persistence)
- **Line 8:** Initialize σ_0^2 with the sample variance – a standard heuristic
- **Lines 9–12:** The GARCH recursion: each day's variance blends yesterday's shock with yesterday's forecast
- **Lines 14–16:** Gaussian log-likelihood $\ell = -\frac{1}{2}\sum[\ln(2\pi\sigma_t^2) + r_t^2/\sigma_t^2]$
- **Line 17:** Return negative (scipy minimizes)
- **Lines 19–22:** L-BFGS-B with positivity bounds ensures $\omega, \alpha, \beta > 0$
- **After fitting:** Check $\alpha + \beta < 1$ (stationarity) and compute half-life

Diagnostics: After estimation, check that standardized residuals $z_t = r_t/\sigma_t$ are i.i.d. with no remaining autocorrelation in z_t^2.

GARCH MLE: optimization as a recursion – each day's variance feeds the next.

Why Do Squared Returns Show Autocorrelation When Raw Returns Do Not?



The volatility fingerprint

- **Raw returns** show near-zero autocorrelation at all lags – consistent with weak-form efficiency
- **Squared returns** (r_t^2) show strong, slowly decaying autocorrelation – the signature of volatility clustering
- **Interpretation:** Returns are unpredictable in direction but highly predictable in magnitude. You cannot forecast the sign, but you can forecast the variance
- **Decay rate:** The ACF decays approximately as $(\alpha + \beta)^k$, matching the GARCH persistence. Slow decay ($\alpha + \beta \approx 0.97$) means volatility shocks persist for weeks
- **Long memory:** Some series show hyperbolic ACF decay rather than exponential, motivating FIGARCH and HAR-RV models

Test: If r_t^2 ACF is insignificant at all lags, there is no clustering and GARCH adds nothing.

ACF of r_t^2 is the fingerprint of volatility clustering – the empirical justification for GARCH.

How Does EGARCH Capture the Asymmetry That Symmetric GARCH Misses?

Nelson (1991) proposed **EGARCH** to model the **leverage effect** – negative returns amplify volatility more than positive returns of equal size.

EGARCH(1,1) log-variance specification:

$$\ln(\sigma_t^2) = \omega + \alpha\big(|z_{t-1}| - E[|z_{t-1}|]\big) + \gamma\, z_{t-1} + \beta\, \ln(\sigma_{t-1}^2)$$

where $z_t = \varepsilon_t / \sigma_t$ are standardized residuals.

Key properties:

- The γ term captures asymmetry: $\gamma < 0$ means negative returns ($z_{t-1} < 0$) increase $\ln(\sigma_t^2)$ more
- Modeling $\ln(\sigma_t^2)$ guarantees $\sigma_t^2 > 0$ without parameter constraints
- No stationarity constraint on individual parameters (unlike GARCH: $\alpha + \beta < 1$)

GJR-GARCH alternative (Glosten, Jagannathan, Runkle, 1993):

$$\sigma_t^2 = \omega + (\alpha + \gamma\, \mathbf{1}_{\varepsilon_{t-1}<0})\, \varepsilon_{t-1}^2 + \beta\, \sigma_{t-1}^2$$

where $\mathbf{1}_{\varepsilon_{t-1}<0}$ is the indicator for negative returns. The leverage effect adds $\gamma\, \varepsilon_{t-1}^2$ only after down moves.

EGARCH uses log-variance (positivity guaranteed); GJR-GARCH adds a leverage indicator. Both capture asymmetry differently.

Which Stylized Facts Does Each Model Capture – and Which Does It Miss?

Stylized Fact	Const.	GARCH	EGARCH	GJR	SV
Volatility clustering	–	Yes	Yes	Yes	Yes
Fat tails (excess kurt.)	–	Partial	Partial	Partial	Yes
Leverage effect	–	–	Yes	Yes	Partial
Mean reversion	–	Yes	Yes	Yes	Yes
Long memory	–	–	–	–	Partial
Regime switching	–	–	–	–	–
Microstructure noise	–	–	–	–	–

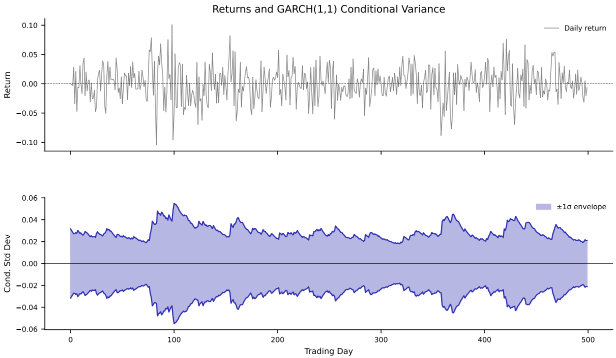
Reading the table

- **Constant vol** captures nothing – it is the baseline against which all models improve
- **GARCH** captures clustering and mean reversion but misses asymmetry and underestimates tails
- **EGARCH/GJR** add the leverage effect but still assume a single regime and short memory
- **Stochastic volatility (SV)** models vol as a latent process, capturing more tail behavior but requiring MCMC or particle filter estimation
- **No single model captures all facts.** Regime switching and microstructure noise require dedicated extensions (MS-GARCH, HAR-RV)

Model selection: Match the model to the stylized facts that matter most for your application. Risk management needs tails; option pricing needs asymmetry.

No model captures all stylized facts. Match assumptions to application: tails for VaR, asymmetry for options.

What Does the GARCH Variance Path Look Like Through Calm and Crisis?



Reading the variance path

- The **upper panel** shows daily returns – notice how large returns cluster together in bursts
- The **lower panel** shows the GARCH conditional variance σ_t^2 responding to shocks and decaying back toward the unconditional mean
- **Spike-and-decay:** Each large return causes a spike in conditional variance that decays exponentially with half-life $h = \ln 2 / \ln(1/(\alpha + \beta))$
- **Regime persistence:** During volatile episodes, the variance stays elevated for weeks because the feedback loop ($\beta \approx 0.92$) sustains it
- **Mean reversion:** Between crises, variance slowly returns to the unconditional level $\bar{\sigma}^2 = \omega/(1 - \alpha - \beta)$

Use case: The variance path directly feeds into VaR calculations, option pricing, and portfolio allocation – it is the core output of the model.

The GARCH variance path spikes with shocks and decays back – capturing the rhythm of risk.

How Does the EWMA Estimator Trade Off Responsiveness Against Stability?

```python
1 import numpy as np
2
3 def ewma_volatility(returns, lam=0.94):
4     """EWMA variance estimator (RiskMetrics).
5
6     Args:
7         returns: array of log returns
8         lam: decay factor (0.94 = daily)
9     Returns:
10        sigma2: conditional variance array
11    """
12    T = len(returns)
13    sigma2 = np.zeros(T)
14    sigma2[0] = returns[:20].var() # init
15    for t in range(1, T):
16        sigma2[t] = (lam * sigma2[t-1]
17            + (1 - lam) * returns[t-1]**2)
18    return sigma2
19
20 # Effective window: 1/(1-lam) = ~17 days
21 vol_94 = ewma_volatility(returns, lam=0.94)
22 vol_97 = ewma_volatility(returns, lam=0.97)
```

Code walkthrough

- **EWMA** is GARCH with $\omega = 0$ and $\alpha + \beta = 1$ (IGARCH). It has no mean reversion – shocks persist forever
- **Line 14:** Initialize with 20-day sample variance. The choice affects early estimates but washes out after $\sim 3/(1-\lambda)$ observations
- **Lines 15–17:** One-parameter recursion: $\sigma_t^2 = \lambda\sigma_{t-1}^2 + (1-\lambda)r_{t-1}^2$
- **$\lambda = 0.94$:** RiskMetrics daily standard. Effective window ≈ 17 days. Reacts quickly but is noisy
- **$\lambda = 0.97$:** Smoother, effective window ≈ 33 days. Less reactive but more stable

EWMA vs GARCH tradeoff: EWMA is simpler (one parameter) but has no mean reversion. GARCH adds $\omega > 0$, giving a long-run anchor. For short horizons, EWMA is often sufficient.

EWMA: one parameter, one recursion, no mean reversion. It is GARCH without the anchor.

How Does Maximum Likelihood Find the Best GARCH Parameters?

Given returns $\{r_1, \ldots, r_T\}$ and GARCH(1,1) conditional variances $\{\sigma_1^2, \ldots, \sigma_T^2\}$:

Log-likelihood (Gaussian innovations):

$$\ell(\theta) = -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\left[\ln(\sigma_t^2) + \frac{r_t^2}{\sigma_t^2}\right], \qquad \theta = (\omega, \alpha, \beta)$$

Student-t log-likelihood (for fat tails, $\nu > 2$ degrees of freedom):

$$\ell(\theta, \nu) = T\left[\ln\Gamma\left(\frac{\nu+1}{2}\right) - \ln\Gamma\left(\frac{\nu}{2}\right) - \frac{1}{2}\ln((\nu-2)\pi)\right]$$
$$- \frac{1}{2}\sum_{t=1}^{T}\left[\ln(\sigma_t^2) + (\nu+1)\ln\left(1 + \frac{r_t^2}{(\nu-2)\sigma_t^2}\right)\right]$$

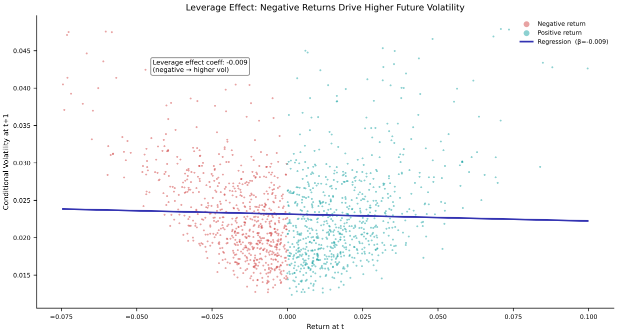
Information criteria for model selection:

$$\text{AIC} = -2\ell(\hat{\theta}) + 2k, \qquad \text{BIC} = -2\ell(\hat{\theta}) + k\ln T$$

where k is the number of parameters. Use AIC/BIC to compare GARCH vs EGARCH vs GJR.

MLE finds the parameters making returns most probable. Student-t adds ν and dramatically improves tail fit.

Why Do Negative Returns Amplify Future Volatility More Than Positive Ones?



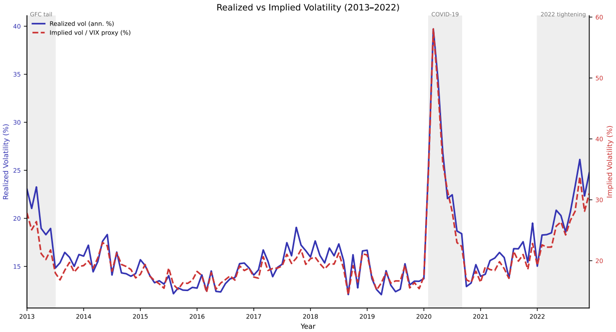
Reading the leverage plot

- Each dot plots today's return (x-axis) against tomorrow's absolute return (y-axis)
- **Asymmetry:** The left side (negative returns) shows higher subsequent absolute returns than the right side – the leverage effect
- **Why "leverage"?** When stock prices fall, the firm's debt-to-equity ratio rises mechanically, making the equity riskier. Black (1976) first documented this
- **Magnitude:** The effect is economically significant – a -3% return typically predicts 1.5x the volatility of a $+3\%$ return
- **Model implication:** Symmetric GARCH assigns equal weight to positive and negative shocks. EGARCH and GJR-GARCH correct this by adding an asymmetry parameter

The leverage effect is one of the strongest arguments for asymmetric GARCH models in equity markets.

The leverage effect: rising debt-to-equity amplifies equity risk after down moves. Panic amplifies falls more than greed lifts.

Why Does Implied Volatility Almost Always Exceed Realized Volatility?



The volatility risk premium

- The **upper line** (implied vol) almost always exceeds the **lower line** (realized vol). The gap is the **volatility risk premium** (VRP)
- **VRP ≈ 2–4% annualized:** Option sellers earn this premium for bearing volatility risk. It is the compensation for being short gamma
- **Crisis inversion:** During extreme events, realized vol can briefly exceed implied vol – the market was caught by surprise
- **Mean reversion:** Both lines tend to converge after spikes, with implied vol leading (it reacts instantly) and realized vol lagging (it accumulates)
- **Trading signal:** The VRP is the foundation of variance swap trading and systematic vol-selling strategies

Key insight: Implied vol is a biased forecast of future realized vol – the bias is the price of volatility insurance.

The implied–realized vol gap is the price of insurance. Sellers earn this premium; buyers pay it.

When Should You Choose GARCH, EGARCH, or GJR-GARCH?

	GARCH(1,1)	EGARCH(1,1)	GJR-GARCH
Equation	$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$	$\ln\sigma_t^2 = \omega + \alpha\|z\| + \gamma z + \beta \ln\sigma_{t-1}^2$	$\sigma_t^2 = \omega + (\alpha + \gamma I_{<0})\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$
Asymmetry	No	Yes (γ)	Yes (γ)
Positivity	Requires $\omega, \alpha, \beta \geq 0$	Automatic (log)	Requires constraints
Parameters	3	4	4
Best for	FX, commodities	Equities (strong leverage)	Equities (moderate leverage)

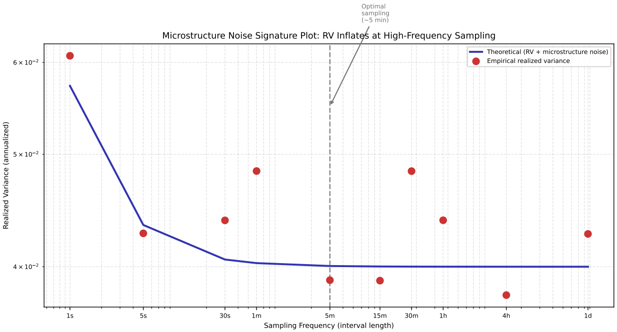
Decision framework

- **Start with GARCH(1,1):** It is the workhorse. If the Ljung-Box test on standardized squared residuals shows no remaining autocorrelation, it is sufficient
- **Test for asymmetry:** Regress z_t^2 on $\mathbf{1}_{z_{t-1}<0}$. If significant, switch to EGARCH or GJR
- **EGARCH vs GJR:** EGARCH is better when positivity constraints bind (small ω). GJR is simpler to interpret and estimate
- **Use AIC/BIC:** Compare in-sample fit. Use out-of-sample forecast evaluation (Mincer-Zarnowitz regression) for model selection
- **Asset class matters:** Equities show strong leverage effects; FX and commodities typically do not

Rule of thumb: For equity risk management, use GJR-GARCH with Student-t innovations. For FX, plain GARCH(1,1) often suffices.

GARCH is the default; EGARCH and GJR add asymmetry for one extra parameter. Asset class determines the choice.

Why Does Realized Volatility Explode When You Sample Too Frequently?



The signature plot

- The x-axis shows sampling frequency (from 1-second to 1-day); the y-axis shows realized volatility
- **At low frequencies** (right side): RV converges to the "true" integrated variance but has high sampling error (few observations)
- **At high frequencies** (left side): RV explodes upward due to microstructure noise (bid-ask bounce, discrete prices, stale quotes)
- **The sweet spot** (5-minute to 15-minute): RV balances noise and sampling error. This is the standard in academic research
- **Noise signature:** The upward curve at high frequencies is diagnostic – if absent, the asset is sufficiently liquid for tick-level estimation

Implication: More data is not always better. The optimal sampling frequency depends on the noise-to-signal ratio of the specific asset.

The signature plot is the first diagnostic for any RV study: where noise overwhelms signal and estimates stabilize.

How Does Quadratic Variation Connect High-Frequency Returns to True Volatility?

Let $p(t)$ be the log-price process. The **integrated variance** over $[0, T]$ is:

$$IV = \int_0^T \sigma^2(s)\, ds$$

Realized variance (the estimator): sample n equally-spaced returns $r_i = p(t_i) - p(t_{i-1})$:

$$RV_n = \sum_{i=1}^{n} r_i^2 \xrightarrow{p} IV \quad \text{as } n \to \infty \quad \text{(no noise)}$$

With microstructure noise $p^*(t_i) = p(t_i) + \eta_i$, $\eta_i \sim (0, \sigma_\eta^2)$:

$$E[RV_n] = IV + 2n\sigma_\eta^2$$

The bias $2n\sigma_\eta^2$ grows linearly with n – more observations mean more noise.

Two-scale realized variance (Zhang, Mykland, Aït-Sahalia, 2005):

$$TSRV = RV_n^{(\text{slow})} - \frac{\bar{n}}{\bar{m}} RV_n^{(\text{fast})}$$

where $RV^{(\text{slow})}$ is computed at sparse frequency and $RV^{(\text{fast})}$ at tick frequency. The subtraction cancels the noise bias.

RV converges to integrated variance in theory; noise derails it in practice. Two-scale estimation corrects the bias.

How Do You Compute Realized Volatility from Tick Data in Practice?

```python
1 import numpy as np
2
3 def realized_vol(prices, freq_min=5):
4     """5-min realized volatility estimator.
5
6     Args:
7         prices: tick-level price array
8         freq_min: sampling interval (minutes)
9     Returns:
10        rv: annualized realized volatility
11    """
12    # Sample at freq_min intervals
13    n = len(prices)
14    step = max(1, freq_min)
15    sampled = prices[::step]
16    # Log returns
17    log_ret = np.diff(np.log(sampled))
18    # Realized variance (sum of squared returns)
19    rv_daily = np.sum(log_ret**2)
20    # Annualize (252 trading days)
21    rv_annual = np.sqrt(rv_daily * 252)
22    return rv_annual
23
24 rv5  = realized_vol(tick_prices, freq_min=5)
25 rv15 = realized_vol(tick_prices, freq_min=15)
```

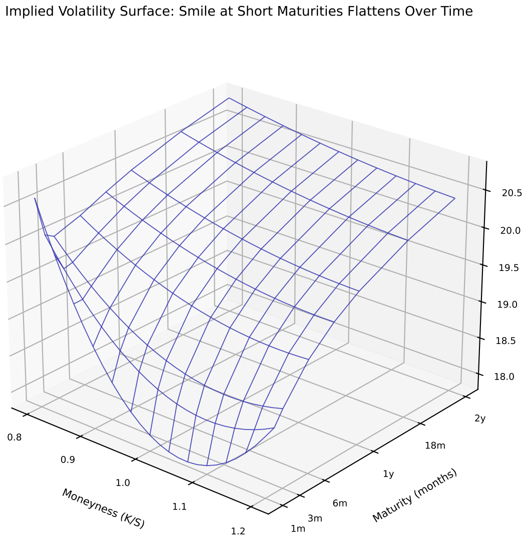
Code walkthrough

- **Lines 12–14:** Subsample tick prices at the desired frequency. 5-minute is the academic standard; 15-minute is more robust for illiquid assets
- **Line 16:** Log returns at the sampled frequency. Using log returns ensures additivity
- **Line 18:** Realized variance = sum of squared returns. This is the discrete approximation to integrated variance $\int \sigma^2(s)\, ds$
- **Line 20:** Annualize by multiplying daily RV by 252 (trading days), then take the square root
- **Lines 22–23:** Compare 5-min and 15-min estimates. If they differ substantially, microstructure noise is significant at the higher frequency

Extension: For noisy data, implement the two-scale estimator (TSRV) by computing RV at both tick and sparse frequencies and subtracting the noise component.

Tick-data RV is model-free. Sampling frequency controls the bias-variance tradeoff; 5 minutes is the standard.

What Does the Full Implied Volatility Surface Look Like?



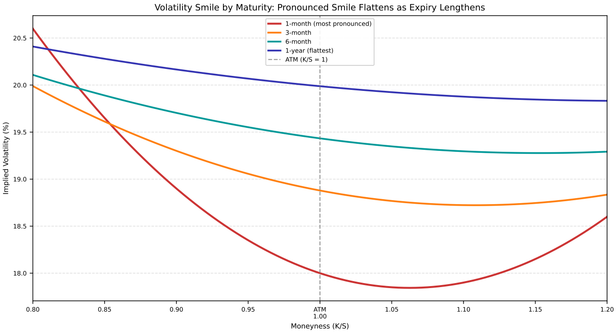
Reading the surface

- **Moneyness** (K/S): ATM at 1.0; OTM puts left; OTM calls right
- **Maturity:** Short-dated (front) show a pronounced smile; long-dated (back) flatten
- **Smile:** IV is higher for OTM than ATM – the market prices tail risk above BS
- **Skew:** OTM puts have higher IV – crash protection is expensive
- **Term structure:** Short-maturity ATM vol varies (contango vs backwardation)

Black-Scholes assumes a flat surface. Every deviation violates BS assumptions.

The implied vol surface summarizes the market's risk expectations; its shape reveals the fear structure.

How Does the Volatility Smile Change Across Maturities?



Smile dynamics

- **Short maturity** (1-month): The smile is steep and V-shaped. OTM puts and calls are both expensive relative to ATM. The market prices near-term tail events aggressively
- **Medium maturity** (6-month): The smile flattens and shifts to a skew – OTM puts remain expensive but OTM calls become cheaper
- **Long maturity** (1-year+): The curve is nearly flat. Over long horizons, the market expects mean reversion in volatility, so extreme-strike options are less overpriced
- **The skew flattening rule:** Smile curvature decays approximately as $1/\sqrt{T}$ – this is consistent with stochastic volatility models

Practical use: Cross-maturity smile dynamics inform calendar spread trading and volatility surface interpolation for exotic pricing.

The smile flattens with maturity: short-dated tail risk is acute; long-dated options embed mean reversion.

How Do You Extract the Market's Fear from a Single Option Price?

The **Black-Scholes** call price is $C_{BS}(S, K, T, r, \sigma)$. Given an observed market price C_{mkt}, **implied volatility** σ_{imp} solves:

$$C_{BS}(S, K, T, r, \sigma_{imp}) = C_{mkt}$$

No closed-form solution exists. Use **Newton-Raphson** iteration:

$$\sigma_{n+1} = \sigma_n - \frac{C_{BS}(\sigma_n) - C_{mkt}}{\text{Vega}(\sigma_n)}, \qquad \text{Vega} = S\sqrt{T}\,\phi(d_1)$$

where $d_1 = [\ln(S/K) + (r + \sigma^2/2)T]/(\sigma\sqrt{T})$ and $\phi(\cdot)$ is the standard normal PDF.

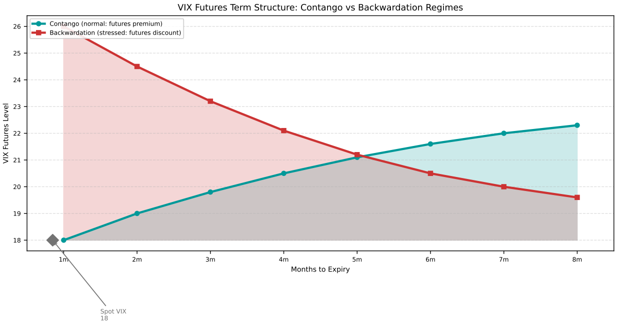
Convergence: Newton-Raphson converges quadratically (3–5 iterations typically) because Vega is smooth and positive for vanilla options.

Implied vol as a quoting convention: Traders quote options in implied vol, not price, because:

- Vol is comparable across strikes, maturities, and underlyings
- A 25% implied vol call on Apple and a 25% implied vol call on Microsoft "cost" the same in risk units
- The vol quote separates the trader's view on volatility from mechanical factors (interest rates, dividends)

Implied vol inverts Black-Scholes: given the price, solve for vol. It is biased upward by the volatility risk premium.

What Does the VIX Futures Curve Tell You About the Market's Fear Timeline?



Reading the term structure

- **Contango** (teal): Futures prices rise with maturity. The market expects volatility to stay low now but assigns a premium to longer-dated uncertainty. This is the normal state ($\sim$80% of the time)
- **Backwardation** (red): Futures prices fall with maturity. Spot VIX is elevated (crisis) and the market expects volatility to mean-revert downward. Rare but signals acute fear
- **The shaded ribbon** shows the VRP embedded in each maturity – wider ribbon means larger risk premium
- **Roll yield:** In contango, long VIX futures positions lose money as contracts roll toward lower spot – the systematic cost of being long volatility

Trading implication: Vol-selling strategies harvest the contango roll yield. They profit 80% of the time but suffer catastrophic losses during backwardation episodes.

Contango: fear priced into the future (normal). Backwardation: fear is here now (crisis signal).

How Do You Extract the Volatility Smile from Option Prices?

```python
1 import numpy as np
2 from scipy.stats import norm
3
4 def bs_call(S, K, T, r, sigma):
5     d1 = (np.log(S/K)+(r+sigma**2/2)*T
6          ) / (sigma * np.sqrt(T))
7     d2 = d1 - sigma * np.sqrt(T)
8     return S*norm.cdf(d1)-K*np.exp(
9         -r*T)*norm.cdf(d2)
10
11 def implied_vol(C_mkt, S, K, T, r,
12                 tol=1e-8, maxiter=50):
13     """Bisection method for implied vol."""
14     lo, hi = 0.01, 3.0
15     for _ in range(maxiter):
16         mid = (lo + hi) / 2
17         if bs_call(S,K,T,r,mid) < C_mkt:
18             lo = mid
19         else:
20             hi = mid
21         if hi - lo < tol:
22             break
23     return mid
24
25 strikes = np.linspace(90, 110, 21)
26 ivs = [implied_vol(p,100,K,0.25,0.02)
27       for K, p in zip(strikes, prices)]
```

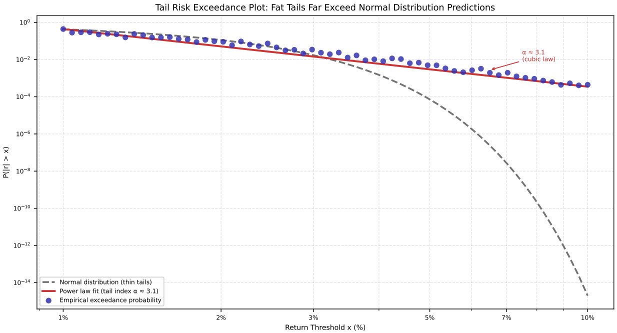
Code walkthrough

- **Lines 4–9:** Standard Black-Scholes call formula. d_1, d_2 are the familiar log-moneyness terms
- **Lines 11–23:** Bisection root-finder. Simpler than Newton-Raphson (no Vega needed) and guaranteed to converge if C_{mkt} is within no-arbitrage bounds
- **Line 14:** Search bounds $[0.01, 3.0]$ cover all plausible implied vols (1% to 300%)
- **Lines 15–21:** Standard bisection: if the model price is too low, the true vol must be higher (increase lower bound), and vice versa
- **Lines 25–27:** Loop over strikes to extract the full smile. The resulting $\sigma_{imp}(K)$ curve reveals the market's tail risk pricing

Extension: For speed, switch to Newton-Raphson with Vega (quadratic convergence) or use the Jaeckel (2015) rational approximation.

Implied vol is root-finding: find the vol making Black-Scholes match the price. Bisection is robust; Newton-Raphson is fast.

How Much Does the Normal Distribution Underestimate Tail Risk?



Reading the exceedance plot

- **Log-log axes:** Both axes use logarithmic scales to reveal tail behavior. A power-law tail appears as a straight line; a normal tail curves downward sharply
- **Normal line** (gray dashed): Drops off exponentially fast. At 5% daily moves, the normal predicts near-zero probability
- **Empirical dots** (purple): Fall much more slowly than the normal predicts. Large moves are orders of magnitude more frequent
- **Power-law fit** (red): Tail index $\alpha \approx 3$ (the cubic law for equities). This means the probability of extreme moves decays as $P(|r| > x) \propto x^{-3}$
- **Practical impact:** A 6-sigma event under normality has probability $\sim 10^{-9}$. Under the power law, it is $\sim 10^{-3}$ – a millionfold underestimate

The normal is a lie in the tails. Power-law tails are the reality – they explain why "impossible" events recur.

How Do You Parameterize the Volatility Surface for Pricing and Interpolation?

The implied volatility surface $\sigma_{imp}(K, T)$ can be parameterized as a function of **log-moneyness** $m = \ln(K/F)$ and maturity T:

SVI (Stochastic Volatility Inspired) parameterization (Gatheral, 2004):

$$w(m) = a + b\left(\rho(m - c) + \sqrt{(m - c)^2 + \sigma_{svi}^2}\right)$$

where $w = \sigma_{imp}^2 T$ is the total implied variance, and $a, b, \rho, c, \sigma_{svi}$ are free parameters.

No-arbitrage constraints:

- Calendar spread: $\partial w / \partial T \geq 0$ (total variance must increase with maturity)
- Butterfly spread: $\partial^2 C / \partial K^2 \geq 0$ (call prices are convex in strike)
- These constrain the SVI parameters and prevent arbitrage in the fitted surface

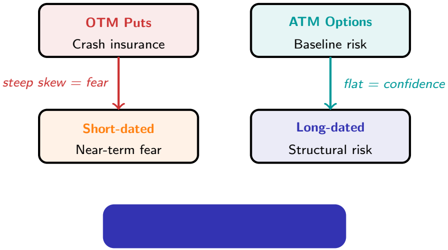
SABR model alternative (Hagan et al., 2002):

$$\sigma_{imp}(K) \approx \frac{\alpha}{(FK)^{(1-\beta)/2}} \cdot \frac{z}{x(z)} \cdot \left[1 + \left(\frac{(1-\beta)^2}{24}\frac{\alpha^2}{(FK)^{1-\beta}} + \frac{\rho\beta\nu\alpha}{4(FK)^{(1-\beta)/2}} + \frac{2-3\rho^2}{24}\nu^2\right) T\right]$$

where $z = \frac{\nu}{\alpha}(FK)^{(1-\beta)/2}\ln(F/K)$ and $x(z) = \ln\frac{\sqrt{1-2\rho z+z^2}+z-\rho}{1-\rho}$.

SVI fits the cross-sectional smile; SABR models dynamics. Both enforce no-arbitrage constraints that interpolation violates.

What Does Each Region of the Volatility Surface Tell You About Market Fear?



Four regions, four signals

- **OTM puts (high vol):** The left wing of the smile is crash insurance. When it steepens, the market is pricing higher probability of a large downward move. Post-1987, this wing has never fully flattened
- **ATM options (baseline):** The level of ATM implied vol reflects the market's consensus expectation of near-term realized vol. It sets the "center of gravity" for the surface
- **Short-dated (steep smile):** Near-term options show the steepest smile because short-dated tail events are most acute. The market prices imminent risk aggressively
- **Long-dated (flat smile):** Over long horizons, mean reversion smooths out extremes. The smile flattens, reflecting lower implied tail probabilities

Monitoring tool: Changes in each region signal different risk narratives – steepening skew means crash fear; flattening term structure means complacency.

The vol surface maps fear: crash risk (OTM puts), baseline (ATM), near-term anxiety (short dates), structural risk (long dates).

How Do You Price a Bet on Future Volatility Without a Volatility Model?

A **variance swap** pays the difference between realized and fixed variance: Payoff $= N_{\text{var}}(\sigma_R^2 - K_{\text{var}})$

Fair strike (model-free, Breeden-Litzenberger):

$$K_{\text{var}} = \frac{2}{T}\left[\int_0^F \frac{P(K)}{K^2}\,dK + \int_F^\infty \frac{C(K)}{K^2}\,dK\right]$$

where F is the forward price, $P(K)$ and $C(K)$ are OTM put and call prices. The integral weights all strikes by $1/K^2$, giving disproportionate weight to deep OTM options (tail risk).

VIX construction (CBOE):

$$\text{VIX}^2 = \frac{2}{T}\sum_i \frac{\Delta K_i}{K_i^2} e^{rT} Q(K_i) - \frac{1}{T}\left(\frac{F}{K_0} - 1\right)^2$$

This is the discrete version of the variance swap formula. VIX is $\sqrt{K_{\text{var}}} \times 100$.

Key insight: The variance swap strike is model-free – it depends only on option prices, not on any volatility model. This is why VIX is called the "fear gauge": it aggregates all option-implied tail information into one number.

Variance swaps and VIX are model-free. They integrate over all OTM options, weighting tails heavily via the $1/K^2$ kernel.

How Do You Test Whether Returns Follow a Random Walk?

```python
1 import numpy as np
2
3 def variance_ratio(returns, q):
4     """Lo-MacKinlay variance ratio test.
5
6     Args:
7         returns: array of log returns
8         q: holding period (days)
9     Returns:
10        vr: variance ratio VR(q)
11        z_stat: test statistic
12    """
13    T = len(returns)
14    mu = returns.mean()
15    var1 = np.sum((returns - mu)**2)/(T-1)
16    # q-period overlapping returns
17    rq = np.array([returns[t:t+q].sum()
18                   for t in range(T-q+1)])
19    varq = np.sum((rq-q*mu)**2)/(T-q+1)
20    vr = varq / (q * var1)
21    # Asymptotic SE under iid null
22    se = np.sqrt(2*(2*q-1)*(q-1)/(3*q*T))
23    z_stat = (vr - 1) / se
24    return vr, z_stat
```

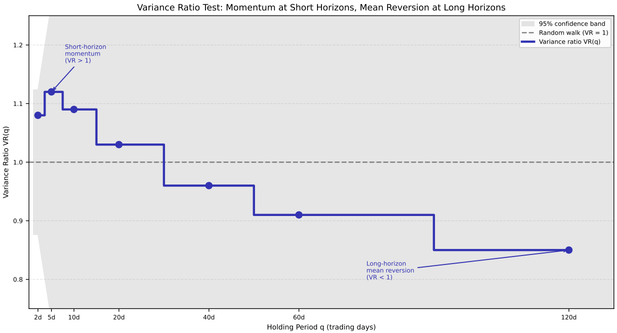
Code walkthrough

- **The idea:** Under a random walk, the variance of q-period returns equals q times the 1-period variance. If $VR(q) \neq 1$, returns are not i.i.d.
- **Line 14:** 1-period variance (sample variance of daily returns)
- **Lines 16–18:** Overlapping q-period returns and their variance. Overlapping maximizes power
- **Line 19:** $VR(q) = \text{Var}(r_q)/(q \cdot \text{Var}(r_1))$
- **Lines 21–22:** Asymptotic standard error under the i.i.d. null (Lo-MacKinlay, 1988)
- $VR > 1$: Positive autocorrelation (momentum)
- $VR < 1$: Negative autocorrelation (mean reversion)

Typical equity result: $VR > 1$ at short horizons (5–20 days), $VR < 1$ at long horizons (60–120 days). The random walk fails at both ends.

The variance ratio detects random walk departures. VR$>$1: momentum. VR$<$1: mean reversion. Both reject efficiency.

At Which Horizons Do Returns Depart from the Random Walk?



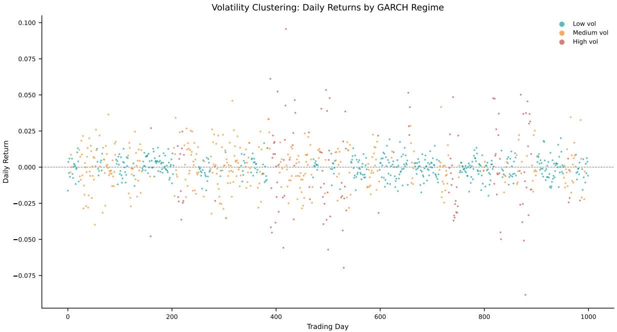
Reading the step plot

- The **purple step function** shows $VR(q)$ at each holding period. The random walk null is $VR = 1$ (gray dashed line)
- **Short horizons** (2–10 days): $VR > 1$, indicating positive autocorrelation. Returns exhibit short-term momentum – trends persist for days
- **Long horizons** (40–120 days): $VR < 1$, indicating negative autocorrelation. Returns exhibit long-term mean reversion – overreactions are eventually corrected
- **The confidence band** (gray shade) shows the 95% interval under the null. Points outside the band reject the random walk at 5% significance
- **Transition zone** (10–40 days): $VR \approx 1$. At medium horizons, momentum and mean reversion roughly cancel out

Implication for volatility: Short-horizon momentum amplifies volatility clustering; long-horizon mean reversion moderates it.

Markets are neither random nor fully predictable: momentum rules the short run, mean reversion rules the long run.

Where Does Volatility Cluster – and What Triggers Regime Switches?



Revisiting the clustering scatter

- Each dot represents a single day's return, colored by its GARCH regime (low, medium, high volatility)
- **Within-regime persistence:** Teal dots (calm) form long continuous runs; red dots (storm) cluster in short, intense bursts. This is the visual manifestation of $\beta \approx 0.92$
- **Transition sharpness:** The boundary between calm and storm regimes is abrupt – often a single large return triggers the switch
- **Recovery asymmetry:** The transition from storm back to calm is gradual (weeks), while the transition from calm to storm is sudden (days)
- **GARCH captures persistence** but misses the sharp regime boundaries. Markov-switching GARCH (Hamilton, 1989) models the transitions explicitly

For risk management: The danger zone is the calm-to-storm transition – your model is most wrong exactly when you need it most.

Single-regime GARCH struggles at regime transitions. The transitions – not the levels – are where risk managers get hurt.

How Do Institutions Trade Volatility as a Standalone Asset Class?

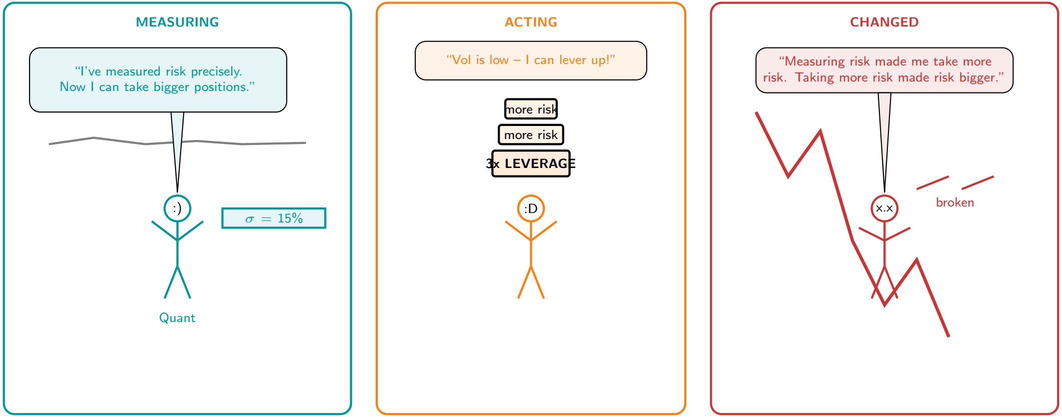
Strategy	Mechanism	Risk	Return
Short variance swap	Sell realized variance, earn VRP	Unlimited upside loss	3–5% pa
Long straddle	Buy ATM call + put, profit from large moves	Theta decay	Tail gains
VIX calendar spread	Long back-month, short front-month VIX	Curve inversion	Carry
Dispersion trade	Short index vol, long single-stock vol	Correlation spike	Correlation risk premium
Gamma scalping	Delta-hedge option, profit from realized > implied	Transaction costs	Realized–implied gap

Five strategies, one common thread

- **Short variance swap:** The purest vol trade. Earns the VRP (implied > realized) most months but faces catastrophic losses when realized vol spikes above the fixed strike
- **Long straddle:** Profits from large moves in either direction. The cost is daily theta decay – you must be right about magnitude, not direction
- **Dispersion:** Exploits the fact that index implied vol embeds a correlation risk premium above the vol of individual stocks
- **Gamma scalping:** Makes money when the underlying moves more than the market expected (realized > implied)

Common risk: All vol strategies are exposed to vol-of-vol – the variance of variance. This second-order risk is hardest to hedge and most dangerous in crises.

Volatility is tradeable. Each strategy harvests a different premium, but all share exposure to the variance of variance.



Measuring risk changes risk. The map is never the territory. And the territory fights back.