

## 2 General Framework

---

The purpose of this chapter is to provide the framework for the economic processes, where the processes are designed to represent the uncertainty in the market. This necessitates (1) the framework of volatility models, which, in the proposed approach are (2) built on stock prices that have (3) to be sampled and pre-processed first.

Therefore, the chapters content is threefold: Section 2.1 gives an introduction to the general classes of volatility models and the related definitions of volatility. Section 2.2 gives the methodological framework of asset pricing and the link from asset prices to latent volatility estimation. Finally, Section 2.3 presents data sampling schemes and pre-processing techniques to sample and correct for errors in raw data.

### 2.1 Classes of Volatility Models

The term volatility refers to the variability of an underlying stochastic process, usually the variability of a given time series. In particular, the volatility is unobservable but can be estimated. The most commonly used estimators can be classified in estimates of stochastic and non-stochastic volatility models.

#### 2.1.1 Time-Invariant Volatility Models

Time-invariant volatility models assume the volatility  $\sigma$  to be constant within a single and specific time window. Two of them, the historical and the implicit volatility, will be explained below. The historical volatility is an explicit estimation of the volatility, while the implied volatility is the implicit result of the Black-Scholes differential equation.

*Historical volatility* The term historical refers to the length of the for estimation considered history  $n$ . The volatility estimate  $\sigma$  is simply the standard deviation  $s$  of the past  $n$  returns. The advantage of historical volatility models is in the simplicity of implementation, while the disadvantage is the arbitrary choice of the historical parameter  $n$ .

*Implied volatility* The implied volatility refers to a concept rather than to a model. The implied volatility proposed by Black and Scholes (1973) is the esti-

mate of the non-observable parameter  $\sigma$  in the Black-Scholes differential equation such that the market value of the considered option, depending on the time and the underlying stock price value, equals the theoretical value of the option.

Please note, that the implied volatility model belongs to stock prices as well as the associated option prices. The options market, however, often shows an immense lack in liquidity what results in insufficient sporadic data sets of option prices. It seems therefore inadvisable to use implied volatility models for ultra-high-frequency volatility estimates, due to the discrepancy in the data granularities of stock and associated option prices.

### 2.1.2 Stochastic Volatility Models

A general overview of the following described class of stochastic volatility (SV) models, autoregressive conditional heteroscedasticity (ARCH) models and generalized autoregressive conditional heteroscedasticity (GARCH) models is given in Bauwens et al. (2012).

Stochastic volatility (SV) models suppose a non-constant but conditional volatility. The stochastic volatility is estimated on the observed trajectory of a hidden stochastic process.

A very early approach of stochastic volatility models is the product process of Taylor (1982, formula (3)). The product is formed by a non-negative process  $\sigma_t$  and a second i.i.d. process  $z_t$  having zero mean and unit variance describing the level of the process:

$$\begin{aligned} y_t - \mu &= \sigma_t z_t \\ \log \sigma_t^2 &= \omega + \beta \log \sigma_{t-1}^2 + v_t \quad v_t \sim N(0, \sigma_u^2), \end{aligned}$$

with  $\mu$  the unconditional mean of  $y_t$ . Taylor supposes furthermore  $\sigma_s$  to be independent of  $z_t$  for all  $s$  and  $t$ . The random variable  $v_t$  declaring the shocks is expected to be i.i.d. and to be uncorrelated with  $z_t$ . The factor  $\beta$  declares the process for volatility clusters. Today, practitioners often simply assume  $z_t$  to be standard normal distributed, i.e.  $z_t \sim N(0, 1)$ . However, any other symmetric distribution, e.g. the t-distribution, or a symmetric mixture of distributions, would do as well, as the variability is declared by  $\sigma_t$ .

*Conditional volatility* A second class of stochastic and conditional volatility models interpret the conditional variation of the underlying process  $y_t$  as a

function of the observable history. A famous representative of the time conditional stochastic volatility models is the autoregressive conditional heteroscedasticity (ARCH) model of Engle (1982, formula (18)). The model assumes the mean of the process  $y_t$  to be declared by  $x_t\beta$ , a linear combination of exogenous and endogenous variables embedded in the information set  $\psi_{t-1}$  available at time  $t - 1$ :

$$\begin{aligned} y_t|\psi_{t-1} &\sim N(x_t\beta, h_t), \\ \varepsilon_t &= y_t - x_t\beta, \\ h_t &= \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \dots + \alpha_q\varepsilon_{t-q}^2, \end{aligned}$$

with  $q$  being the order of the ARCH process,  $\alpha$  and  $\beta$  vectors of unknown parameters, and  $\alpha_0 > 0$  and  $\alpha_1, \alpha_2, \dots, \alpha_p \geq 0$  for regularity. Note, that the ARCH model is often written in the more intuitive form:

$$\begin{aligned} y_t - \mu_t = \varepsilon_t &= \sigma_t z_t \quad z_t \sim N(0, 1), \\ \sigma_t^2 &= \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \dots + \alpha_q\varepsilon_{t-q}^2. \end{aligned}$$

In this form, it becomes obvious that Engle's ARCH model is closely related to Taylor's SV model. The main distinction of SV to ARCH models is the knowledge of the conditional volatility  $\sigma_t$  within an given information set  $\psi_{t-1}$ . While  $\sigma_t|\psi_{t-1}$  is known in ARCH models, this condition is unknown and an unobserved random variable in general SV models. Engle's main idea is, hence, to consider the conditional variance of random model errors as a dependency of historical realized errors.

The generalization of Engle's concept, to model the conditional variance not only as a dependency of the time series history, but also of his own history, is given in Bollerslev (1986). This model is called generalized autoregressive conditional heteroscedasticity (GARCH) model.

Stochastic volatility models are generally more flexible and allow a more natural economic interpretation than GARCH models. Empirical applications are, however, dominated by GARCH instead of SV approaches. This is due to the circumstance that in practise GARCH model estimation is often less complex than an equivalent SV model estimation.

Bauwens et al. (2012, p.33) note, that stochastic volatility models *"...are essentially parametric and usually designed to estimate the daily, weekly, or monthly volatility using data sampled at the same frequency."* Note, that Bauwens et al. are not talking about high-frequency data measured in seconds, minutes, or hours.

The authors note further, that: *"Since French et al. (1987) [...] econometricians have considered using data sampled at a very high frequency to compute ex-post measures of volatility at a lower frequency."*

The daily, weekly, or monthly volatility is not under investigation in this thesis. However, the objects under investigation are high-frequency volatility shocks, making ultra-high-frequency data the data of interest and the realized volatility described below the estimator of interest.

*Realized volatility* The main idea of realized volatility models is to make use of each single variation on a high-frequency or ultra-high-frequency scale. The canonical estimator of the realized volatility is the sum of the squared first differences of the logarithmic asset price process  $X_t$ :

$$RV = \sum_{i=1}^{n_t} (X_{t,i} - X_{t,i-1})^2$$

with  $i$  indicating the  $i$ th observation at period  $t$ , e. g. a day or an hour, and  $n_t$  the total number of observations in period  $t$ .

The origin of the idea, to simply sum up squared realizations to estimate the volatility, is unknown, but it dates back at least to the suggestion of Merton (1980).

### 2.1.3 Integrated Volatility

In contrast to the realized volatility, which is defined in discrete time, the deterministic integral of the quadratic stochastic process  $\sigma_s$

$$IV = \int_0^t \sigma_s^2 ds \quad (2.1)$$

is defined in continuous time. This integral of the quadratic volatility process  $\sigma_s$  along the time interval  $[0, t]$  is called the *integrated volatility* ( $IV$ ). A detailed discussion of the theoretical issues about the integrated volatility is provided below, in Section 2.2.2.

It is worth mentioning that the widely-used GARCH and SV models, belonging to the class of parametric models, infer the integrated volatility. RV models, on the other hand, belonging to the class of non-parametric models, calculate the integrated volatility. Hence, the common denominator of volatility estimation is the latent integrated volatility.

In this thesis, five different realized volatility estimators (introduced in Section 3.2) calculating the integrated volatility will extensively be discussed and evaluated for their overall performance in the final volatility shock monitoring system. The remaining volatility estimators mentioned in this chapter were given in order to understand the connections, but are no longer part of further considerations.

## 2.2 Framework to Model the Integrated Volatility

The given framework follows the standard financial theory on volatility modelling. It is built on the theory of stochastic processes to model asset prices, see Iacus (2008) and Iacus (2011), and the theory of integrated volatility. Further following the approach of Aït-Sahalia and Jacod (2010), it is state of the art to embed a continuous component in high-frequency stock data models. Typically, by making use of a Brownian motion.

### 2.2.1 Modelling Asset Prices

The origin of today's financial mathematics is most probably the dissertation of Bachelier (1900). The fundamental idea of Bachelier is to make use of probabilistic theory to model the motions of asset prices. In particular, to use Brownian motions to evaluate the value of asset options. Today, the *geometric Brownian motion* is the fundamental model to describe the motion of asset prices. A mathematical concept to model the motion of asset prices using stochastic processes is as follows:

*Probability Space* First, assume a probability space  $(\Omega, \mathcal{A}, P)$ . This construct describes the potential outcome of random experiments and consists of three parts:

1. A non-empty sample space  $\Omega$ , which is the set of all possible outcomes of a random experiment.
2. A  $\sigma$ -algebra  $\mathcal{A}$  on the set  $\Omega$ , which is a set consisting of countable subsets of  $\Omega$ .
3. An assignment of probabilities to the events of the random experiment, which is a function  $P : \mathcal{A} \rightarrow [0, 1]$  from events to probabilities.

*$\sigma$ -Algebra* Secondly, assume a  $\sigma$ -algebra in the probability space  $(\Omega, \mathcal{A}, P)$ . A  $\sigma$ -algebra is a system of sets  $\mathcal{A}$ , fulfilling the conditions:

1.  $\Omega \in \mathcal{A}$ .
2.  $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$  with  $A^c$  the complementary set of  $A$ .
3.  $A_1, A_2, \dots \in \mathcal{A} \Rightarrow \bigcup_{t \in \mathbb{N}} A_t \in \mathcal{A}$ .

The intuitive function of  $\sigma$ -algebras in the theory of stochastic processes is to describe the potential observable information at each specific time.

*Filtration* Third, assume a filtration. A family of embedded  $\sigma$ -algebras  $(\mathcal{A}_t)_{t \in \mathbb{N}}$  modelling a rising time sequence is called filtration, meaning for all  $s, t \in \mathbb{N}$  with  $s < t$  is  $\mathcal{A}_s \subseteq \mathcal{A}_t$ . Finally, adapt the stochastic process  $\{S_t\}_{t \in \mathbb{N}}$  to the filtration  $(\mathcal{A}_t)_{t \in \mathbb{N}}$ . The intuitive function of a filtration is to guarantee that it is not possible to have more information at an earlier than actual time.

*Martingale* A stochastic process  $\{S_t\}_{t \in \mathbb{N}}$  adapted to the filtration  $(\mathcal{A}_t)_{t \in \mathbb{N}}$  is further called a martingale, if

1.  $E(|S_t|) < \infty$ , and
2.  $E(S_{t+1} | \mathcal{A}_t) = E(S_{t+1} | S_t, \dots, S_1) = S_t$ .

The first condition ensures the existence of the expectation value, the second condition characterizes the martingale property. Simply changing the second condition to

$$\begin{aligned} E(S_{t+1} | \mathcal{A}_t) &\geq S_t \text{ or} \\ E(S_{t+1} | \mathcal{A}_t) &\leq S_t \end{aligned}$$

is called sub-martingale ( $\geq$ ), respectively super-martingale ( $\leq$ ). The stochastic process  $\{S_t\}_{t \in \mathbb{N}}$  is further called semi-martingale, if  $\{S_t\}$  is either a sub- or an super-martingale.

Further details on probability spaces, adapted processes, filtrations, and martingales, are e. g. given in Bauer (2011).

All together, following the suggestion of Back (1991) to model the motion of asset prices, let  $\{S_t\}_{t \geq 0}$  be the price of an asset and assume the latent DGP of  $\{S_t\}$  to be a super-martingale. Corresponding to the Doob-Meyer decomposition theorem, see e. g. Protter (2005, pp.106-117), the existing condition also gives a not necessarily unique decomposition of the price process  $\{S_t\}_{t \geq 0}$  into the sum of a stochastic process of finite variation on the one hand, and a martingale on the other hand.

Now consider a short time interval  $dt$ . The changing of the asset price process in the interval  $[t, t + dt)$  is accordingly equal to  $dS_t = S_{(t+dt)} - S_t$ , having returns  $r$  given by the proportion of  $dS_t$  to  $S_t$ :

$$r = dS_t / S_t.$$

The Doob-Meyer decomposition property transmits directly to this return process, giving

$$r = \text{deterministic part} + \text{stochastic part.} \quad (2.2)$$

The first part, the deterministic component, is generally related to the risk free interest rate and the deterministic return in  $dt$  is  $\mu dt$ . The process is of finite variation for constant returns in each infinitesimal small time interval  $[t, t + dt)$ :

$$\text{deterministic contribution} = \mu dt. \quad (2.3)$$

The second part, the stochastic component, on the other hand, relates to the stochastic variation of the asset with non-predictable shocks. The shocks are nevertheless typically assumed to be Gaussian, i. e. symmetric with zero mean:

$$\text{stochastic contribution} = \sigma dB_t, \quad (2.4)$$

where  $B_t$  is a standard Brownian motion or standard Wiener process fulfilling  $dB_t = B_{(t+dt)} - B_t$  and  $dB_t \sim N(0, dt)$ . A Brownian motion is furthermore a martingale and hence conform with the statement of the Doob-Meyer decomposition theorem. Fitting the equations (2.3) and (2.4) in the return process (2.2) results in:

$$dS_t/S_t = \mu dt + \sigma dB_t$$

which is equivalent to the stochastic differential form:

$$dS_t = \mu S_t dt + \sigma S_t dB_t. \quad (2.5)$$

Setting the processes  $\{\mu S_t\} = \{\mu_t\}$  and  $\{\sigma S_t\} = \{\sigma_t\}$ , gives

$$dS_t = \mu_t dt + \sigma_t dB_t, \quad (2.6)$$

which is simply the differential form of the general Itô process

$$S_t = S_0 + \mu \int_0^t S_u du + \sigma \int_0^t S_u dB_u.$$

with  $S_0$  as the start value of the process, see e. g. Iacus (2011, pp.8-9). Please note, that any Itô process can also be interpreted as a generalized Brownian motion with random drift and volatility.

Note further, that the *geometric Brownian motion* is the process  $S_t$  that solves formula (2.5). The differential of this process can principally be build by Riemann integration theory, but due to the non-differentiability of the Brownian motion, the differential does not exist in a single point. This warrants the use of stochastic integrals, e. g. the integration terminology of Itô. For a general introduction to stochastic integration theory, see e. g. Chung and Williams (2014).

### 2.2.2 Modelling Integrated Volatility

Taken together, each asset motion  $S_t$ , and therefore also the logarithm of the asset prices  $X_t = \log S_t$ , can be represented by the stochastic differential equation (2.6), which is:

$$dX_t = \mu_t dt + \sigma_t dB_t. \quad (2.7)$$

Thus, the logarithmic price changes are represented by the differential  $dX_t$ , determined by

1. a real-valued and continuous process with finite variation  $\mu_t$ ,
2. a strictly positive Càdlàg process  $\sigma_t$  (any stochastic process is called Càdlàg, if (a) each trajectory  $t \rightarrow X_t$  is right-sided continuous in each point  $t$  a.s. and (b) the left-sided limits exist), and
3. a standard Brownian motion or Wiener process  $B_t$ .

The changing of the logarithmic price process  $X$  within the interval  $(0, t]$  is consequently:

$$dX_t = r_t = \int_0^t \mu_s ds + \int_0^t \sigma_s dB_s. \quad (2.8)$$

Due to the short considered time intervals (e. g. one day or one hour), the deterministic component in (2.8),  $\int_0^t \mu_s ds$ , is supposed to be zero. The stochastic component in (2.8),  $\int_0^t \sigma_s dB_s$ , however, is not calculated evaluating the Itô integral, but by the quadratic variation of the Itô process, which is:

$$[X, X]_t = \int_0^t \sigma_s^2 ds. \quad (2.9)$$

Please note, that every Itô process has the quadratic variation (2.9). This fact is also known as *the representation theorem of the quadratic variation of stochastic processes* (to be discussed in Section 2.2.3). Note in particular the transition from the stochastic integral  $\int(\cdot)dB_s$  to the deterministic integral  $\int(\cdot)ds$  when using the quadratic variation.

What is decisive, however, is that the quadratic variation of the Itô process (2.9) corresponds to the *integrated volatility* of formula (2.1) on page 14. This integrated volatility is the object of interest, either over one or successive periods of time. The task in forecasting high-frequency volatility shocks is therefore, to estimate the latent integrated volatility (2.1) in real-time. This task demands to use intraday data.



### 2.2.3 Quadratic Variation of Stochastic Integrals

The link from stochastic integrals of the form  $\int_{t-1}^t \sigma_s dB_s$  (Itô integrals) to deterministic integrals of the form  $\int_{t-1}^t \sigma_s^2 ds$  (Riemann integrals) is based on the representation theorem of *continuous local martingales as stochastic integrals with respect to Brownian motions*, which is part of semi-martingale process theory. Please note the importance of this theorem in order to model stochastic volatilities.

The representation theorem states (in an abbreviated version) that each stochastic integral  $\int_0^t \sigma_s dB_s$  along a standard Brownian motion  $B$  with a measurable, adapted process  $\sigma$  fulfilling  $P(\int_0^t \sigma_s^2 ds < \infty) = 1$  for every  $0 \leq t < \infty$  (meaning  $\sigma$  to be local bounded in the full domain) is a continuous local martingale with quadratic variation  $\int_0^t \sigma_s^2 ds$ , which is a continuous function of  $P$  a.s. The full theorem including the proof is e.g. given in Karatzas and Shreve (2007, pp.170-173).

However, the representation theorem further assumes complete markets, fulfilling the following conditions:

1. Neither regional, objective, temporal nor personal preferences.
2. Perfect market transparency.
3. Homogeneity of goods.
4. Unlimited fast reactions of all market participants to changes.

The implications in the context of this thesis are: (1) all investors will identically evaluate the impact of unexpected news, (2) all investors have full access to all news arrivals at the same moment, (3) the impact of homogeneous news is identical, and (4) there is no lagged dynamic in the return process of the assets.

Due to the restrictive assumptions, complete markets do obviously not exist. Stock markets, however, apply to be the markets which are next to the requested pure competition, see Harrison and Pliska (1981). Therefore, the sufficient attainment of complete markets to make use of the representation theorem can be regarded as fulfilled. But, inefficiencies in the considered empirical price processes are nevertheless expected.

## 2.3 Sampling and Pre-processing Intraday Data

The umbrella term *intraday data* envelopes all data sampled within a day, independent of the sampling frequency and independent of the sampling scheme.

The literature on intraday data is hence not clear in the definition of sampling frequencies and sampling schemes. The statistical properties of the in Section 3.2 proposed integrated volatility estimators to determine formula (2.9), however, depend on sampling frequencies and sampling schemes.

### 2.3.1 Sampling Schemes

Sampling schemes are rules of data recording. The main classification of sampling schemes is due to the concept of time, see Oomen (2005, 2006) and Griffin and Oomen (2008):

1. *Calendar Time Sampling*: Sampling in calendar time means to sample on an equidistant calendar time scale, for instance every five minutes. Public available stock price data are typically calendar time data.
2. *Business Time Sampling*: Sampling in business time means to sample in event time, but in predefined distances. The sampling frequency in business time follows an intensity process, e. g. a function describing when to sample very often or to sample moderately.
3. *Transaction Time Sampling*: Sampling in transaction time means to sample in event time, but not necessarily in predefined or equidistant distances. The sampling frequency in transaction time records every single transaction. Transaction time sampling provides the most available information.
4. *Tick Time Sampling*: Tick time sampling corresponds to the sampling scheme in transaction time, but with censoring all zero returns. Tick time samples are records of price changes, and hence based on informations, not on transactions.

One main argument to use transaction or tick time sampling is market intensity. Both sampling schemes are not necessarily equidistant, allowing a flexible sampling intensity. The principle is to sample more data in active than in calm market situations, contrary to sampling for instance once a minute.

Two corresponding analyses are provided by Oomen (2005) and Oomen (2006). Their analysis is built on the assumption, that each observable data point consists of the combination of a true but latent data point plus a random data point, called microstructure noise (defined in Section 3.1.2 on page 31). Their central finding is: The mean squared error (MSE) between observed and true data in realized volatility estimates (to be discussed in Section 3.2 on page 32) is lower on business or transaction time samples than on calendar time samples.

Forecasting High-Frequency Volatility Shocks

An Analytical Real-Time Monitoring System

Kömm, H.

2016, XXIX, 171 p. 19 illus., Softcover

ISBN: 978-3-658-12595-0