Taming impulsive high-frequency data using optimal sampling periods

George Tzagkarakis, Frantz Maurer, J.P. Nolan

Annals of Operations Research
https://doi.org/10.1007/s10479-023-05701-y

ORIGINAL RESEARCH

Taming impulsive high-frequency data using optimal sampling periods

George Tzagkarakis[1,2] · **Frantz Maurer**[3] · **John P. Nolan**[4]

Received: 8 October 2021 / Accepted: 25 September 2023

Abstract
Optimal sampling period selection for high-frequency data is at the core of financial instruments based on algorithmic trading. The unique features of such data, absent in data measured at lower frequencies, raise significant challenges to their statistical analysis and econometric modelling, especially in the case of heavy-tailed data exhibiting outliers and rare events much more frequently. To address this problem, this paper proposes a new methodology for optimal sampling period selection, which better adapts to heavy-tailed statistics of high-frequency financial data. In particular, the novel concept of the degree of impulsiveness (DoI) is introduced first based on alpha-stable distributions, as an alternative source of information for characterising a broad range of impulsive behaviours. Then, a DoI-based generalised volatility signature plot is defined, which is further employed for determining the optimal sampling period. The performance of our method is evaluated in the case of risk quantification for high-frequency indexes, demonstrating a significantly improved accuracy when compared against the well-established volatility-based approach.

Keywords High-frequency indexes · Alpha-stable models · Degree of impulsiveness · Optimal sampling period

✉ George Tzagkarakis
gtzag@ics.forth.gr

Frantz Maurer
frantz.maurer@kedgebs.com

John P. Nolan
jpnolan@american.edu

1 Foundation for Research and Technology - Hellas, Heraklion, Crete, Greece

2 IRGO, EA 4190, University of Bordeaux, Bordeaux, France

3 KEDGE Business School, Talence, France

4 Math/Stat Department, American University, Washington, DC, USA

Published online: 24 November 2023

1 Introduction

High-frequency financial data analysis has experienced an enormous and fast development over the recent years. At the core of modern financial instruments is the instantaneous collection of massive tick-by-tick data from financial markets. The availability of high-frequency data on transactions, quotes and order flow in electronic order-driven markets has revolutionised data processing and statistical modelling techniques for the design of advanced algorithmic trading systems, bringing up new theoretical and computational challenges. Most importantly, high-frequency financial data possesses a complicated structure due to irregularities and roughness caused by a large number of instantaneous changes of the markets and trading noises.

A fundamental question in algorithmic trading is how often one should sample, in order to account for the micro-structure effects. Ané and Geman (2000) employ stochastic time changes for generating virtually perfect normality in high-frequency asset returns, whilst allowing both the time-change and price processes to take the form of jump diffusions. In response to the increase in market fragmentation, due to the considerable changes of market micro-structure in recent years, Delaney (2018) proposed an optimal timing strategy to invest in high-frequency trading technologies. In the seminal work of Aït-Sahalia et al. (2005) it was shown that if micro-structure noise is present but unaccounted for, then the *optimal sampling period* is finite and can be derived in closed form. On the other hand, if the presence of noise is accounted for, modelling the noise term explicitly restores the first-order statistical effect that sampling as often as possible is optimal. A main limitation of their method is that it relies on a fully parametric framework, by assuming that the noise follows a Gaussian distribution. Sample moments of high-frequency returns data recorded at different frequencies were employed in Bandi and Russell (2006) to calculate the optimal sampling period, by minimising a mean-squared error criterion on the realised variance estimator as a function of the sampling period. However, this method cannot be applied to financial time series modelled by a finite mixture model, whilst it does not fully consider the number of parameters involved in a model. To address this issue, Choi and Kang (2014) proposed a finite mixture modelling scheme to calculate the optimal sampling period through a modified likelihood ratio test. However, a finite mixture model with only two components is considered, which may yield inaccurate fitting in the case of highly heavy-tailed returns data. Furthermore, estimation of model parameters is required, by minimising Akaike's information criterion, which may be computationally intractable as the order of the finite mixture model increases. Optimal sampling period for volatility models was estimated by Bhattacharyya et al. (2009) via GARCH-based statistical modelling of high-frequency data. However, the method proposed therein is based on a normality assumption for the distribution of returns, without accounting for the potential occurrence of gross values in the returns data.

From the above, it becomes apparent that the optimal sampling period should be selected on the basis of satisfying a trade-off between accuracy and potential biases due to market micro-structure frictions. To this end, an alternative tool was introduced by Fang (1996) and Andersen et al. (2000) to assess this trade-off, namely, the volatility signature plot. This plot provides a simple graphical diagnostic for calculating the realised volatility of high-frequency financial returns, by characterising different market micro-structures in terms of their volatility signatures. In particular, the patterns of bias injected in realised volatility are identified by sampling progressively more frequently the underlying returns. Despite its computational efficiency, this tool is defined in terms of second-order moments, by employing the so-called price variogram (Haslett, 1997), as well as by assuming a Gaussian distribution

for the noise term. Nevertheless, both assumptions are violated in the case of heavy-tailed data.

The majority of existing methods for high-frequency financial data analysis rely primarily on the controversial use of second-order moments, or equivalently on light-tailed, finite-variance assumptions for the statistics of the data-generating processes, in order to estimate volatility. However, despite the analytic tractability and practical appeal, these assumptions may be problematic when we analyse impulsive data, which give rise to heavy-tailed processes with possibly infinite variance. On the other hand, the presence of large-amplitude samples, which can be of infinite or very large variance, can mask the information content of the time series, especially in neighbouring time instants. This may degrade dramatically the accuracy of subsequent decision making, thus necessitating the design of novel data analysis techniques that are able to adapt to heavy-tailed financial data exhibiting outliers or rare events much more frequently than what a light-tailed distribution dictates.

In this paper, we propose a generalised framework for jointly quantifying the inherent impulsiveness and estimating the optimal sampling period for mitigating the micro-structure effects in high-frequency financial data. To this end, first we introduce the novel concept of the *degree of impulsiveness* (DoI), as an alternative key indicator of the variability of high-frequency data, which complements the well-established concept of volatility. Then, a DoI-based *generalised price variogram* is defined, which adapts to a broad range of impulsive behaviours (i.e., from light-tailed to highly impulsive data), along with the associated *generalised volatility signature plot* that is further used to estimate the optimal sampling period of high-frequency financial data. For this, we rely on the efficiency of *alpha-stable distributions* and fractional lower-order moments (FLOMs) (Nolan, 2020; Samorodnitsky & Taqqu, 1994; Nikias & Shao, 1995), to accurately model the heavy-tailed, possibly infinite-variance, time series data.

Notice that, from a practitioner's viewpoint, processes with infinite variance may sound counter-intuitive, since they give raise to infinite power that does not really exist in real world data. Nevertheless, from a probabilistic perspective, variance is just a measure of how spread out a distribution is. Distributions with infinite variance present fat upper tails that decrease at an extremely slow rate. Intuitively, this means that the distribution will vanish for very large absolute values of the corresponding random variable. Actually, in theory, it never vanishes, and this is precisely the reason we say that the upper tail is "unbounded". The slow decay of probability in this area increases the odds of extreme values (outliers), and other surprising last-minute events at some point in the future. Although the model has infinite variance, this does not imply that the real-world phenomenon being modeled also extends to infinity. It just means that the model is a "good enough" fit to describe the behaviour of the phenomenon under study.

Since most trading and risk management strategies rely on the returns of an asset, hereafter, we employ continuously compounded returns, r_t, over an horizon of τ time units, defined by

$$r_t = \log\left(\frac{v_t}{v_{t-\tau}}\right) = \log(v_t) - \log(v_{t-\tau}) \;\; t = \tau + 1, \ldots, N \tag{1}$$

where v_t denotes an asset's price at time t and $\log(\cdot)$ is the natural logarithm. For instance, when we operate with minute data, $\tau = 1$ corresponds to minute returns, whereas $\tau = 60$ corresponds to hourly returns computed from minute data. In any case, the time unit will be explicitly defined whenever needed, thus the interpretation of τ will be clear.

The rest of the paper is organised as follows: Sect. 2 briefly overviews the main concepts of alpha-stable models. Section 3 describes the data utilised in this study, along with an

assessment of its statistical behaviour. Section 4 introduces the degree of impulsiveness as an additional source of information to the well-established volatility. Section 5 analyses our proposed method for optimal sampling period estimation tailored to high-frequency financial data. Section 6 evaluates the performance of our method on the problem of risk quantification with distinct high-frequency indexes. Finally, Sect. 7 summarises the main outcomes and gives directions for further extensions.

2 Alpha-stable models

In this section, we briefly overview the main concepts and definitions of alpha-stable models, which are at the core of our methodology. The most concrete way to describe an alpha-stable (α-stable) distribution is through its characteristic function, given by Nolan (1997),

$$\varphi(t) = \begin{cases} \exp\left(i\delta t - \gamma^{\alpha}|t|^{\alpha}\left[1 - i\beta\tan\left(\frac{\pi\alpha}{2}\right)\operatorname{sign}(t)\right]\right) & , \ \alpha \neq 1 \\ \exp\left(i\delta t - \gamma|t|\left[1 + i\beta\frac{2}{\pi}\operatorname{sign}(t)\log(|t|)\right]\right) & , \ \alpha = 1 \, . \end{cases} \tag{2}$$

From (2), we see that an α-stable distribution requires four parameters to be fully described, namely, (i) the characteristic exponent $\alpha \in (0, 2]$ (the smaller the α, the heavier the tails of an α-stable density function), (ii) the skewness $\beta \in [-1, 1]$, (iii) the dispersion $\gamma > 0$, and (iv) the location $\delta \in \mathbb{R}$. We will denote α-stable distributions by $S_{\alpha}(\gamma, \beta, \delta)$, and write $X \sim S_{\alpha}(\gamma, \beta, \delta)$ to indicate that X follows an α-stable distribution with parameters $(\alpha, \beta, \gamma, \delta)$.

Without loss of generality, hereafter we assume that $\delta = 0$ for a given time series. This stems from the property that, if $X \sim S_{\alpha}(\gamma, \beta, \delta)$, then $X + c \sim S_{\alpha}(\gamma, \beta, \delta + c)$. In our implementation, the model parameters $(\alpha, \beta, \gamma, \delta)$ are estimated from the given data using the empirical characteristic function (ECF) based method described in Kogon and Williams (1998). We note that all the subsequent numerical calculations involving α-stable densities are performed using the STABLE toolbox.[1]

Due to their algebraic tails (i.e., tails decaying slower than exponential), α-stable distributions lack finite second-order moments. Instead, all moments of order p less than α do exist and are called the *fractional lower-order moments* (FLOMs). In particular, the FLOMs of an α-stable random variable $X \sim S_{\alpha}(\gamma, \beta, 0)$ with the parameterisation given by (2) are obtained in a similar way with the FLOMs of skewed stable distributions calculated in Kuruoglu (2001), as follows,

$$\mathbb{E}\left\{|X|^{p}\right\} = C_{p,\alpha,\beta} \cdot \gamma^{p} \quad p \in (-1, \alpha) \tag{3}$$

where

$$C_{p,\alpha,\beta} = \frac{\Gamma\left(1 - \frac{p}{\alpha}\right)}{\Gamma(1-p)\cos\left(\frac{\pi}{2}p\right)} \cdot \frac{\cos\left(\frac{p\theta}{\alpha}\right)}{|\cos(\theta)|^{p/\alpha}} \tag{4}$$

and $\theta = \arctan\left(\beta\tan\left(\frac{\alpha\pi}{2}\right)\right)$.

Notice the dependence of the above expressions on the parameter p, the order of the FLOM. It holds that all the FLOMs of a stable random variable are equivalent, in the sense that any two of the fractional lower order moments differ by a fixed constant which is independent of the random variable itself (Nikias & Shao, 1995). Although p is a free parameter, motivated by the equivalence property of FLOMs and in order to avoid a trial-and-error preprocessing step for setting p, in the following we select p as a function of α, which is estimated directly

[1] Robust Analysis Inc., STABLE toolbox version 5.3 (http://www.robustanalysis.com).

Table 1 FLOM-based optimal p as a function of the characteristic exponent α

α	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
p_{opt}	0.52	0.56	0.58	0.61	0.64	0.69	0.72	0.76	0.81	0.88	0.98

from the observed data, based on a specific optimisation criterion. In particular, Tzagkarakis et al. (2006) proposed an approach for selecting the optimal p as the one that minimises the standard deviation of a FLOM-based covariation estimator between two α-stable random variables $X \sim S_\alpha(\gamma_X, \beta_X, 0)$ and $Y \sim S_\alpha(\gamma_Y, \beta_Y, 0)$, defined as follows,

$$c_{XY} = \frac{\sum_{i=1}^{N} x_i |y_i|^{p-1} \text{sign}(y_i)}{\sum_{i=1}^{N} |y_i|^p} \gamma_Y^\alpha \tag{5}$$

where

$$\text{sign}(y_i) = \begin{cases} -1 & \text{for } y_i < 0 \\ 0 & \text{for } y_i = 0 \\ 1 & \text{for } y_i > 0 \,. \end{cases}$$

The covariation, c_{XY}, plays an analogous role for α-stable variables to the one played by covariance for light-tailed distributions. Notice that covariance is not defined for α-stable models with $\alpha < 2$.

This approach yields an almost linear relation between α and the optimal value of p, and specifically $p \lesssim \alpha/2$. In addition, if $p < \alpha/2$ the FLOM estimator has a finite variance, which is desirable (Nikias & Shao, 1995). This constraint for the value of p also agrees with the remarks in Kuruoglu (2001), which suggests to set p at the order of $\alpha/10$. In the following, we set the value of p in two steps: (i) linearly interpolating the entries of the lookup Table 1, generated by Tzagkarakis et al. (2006), (ii) multiplying with a correction factor of 0.2 to align with Kuruoglu (2001).

Remark 1 Without loss of generality, in the subsequent analysis we restrict the index of stability to the range $1 < \alpha \leq 2$, which is most frequently encountered in practical applications. The condition of $\alpha > 1$ also yields that, whenever required in the subsequent derivations, the mean of the associated random variable is defined.

3 Data and assessment of heavy-tailed behaviour

In the subsequent analysis, we employ minute closing prices of four major stock indexes worldwide and one cryptocurrency, namely, S&P 500 (in USD) [SPXUSD], NIKKEI 225 (in JPY) [JPXJPY], DAX 30 (in EUR) [GRXEUR], EUROSTOXX 50 (in EUR) [ETXEUR], and Bitcoin (in USD) [BTCUSD], spanning the period from January 2nd, 2016 to December 31st, 2018. The four stock indexes have been downloaded from Google Finance, and the Bitcoin's prices from Bitstamp.

As a first qualitative assessment of the behaviour of our data, a selection of summary statistics, such as the mean, standard deviation, skewness and (excess) kurtosis, are typically presented. Table 2 shows these statistics for the compounded returns of the above five indexes. As it can be seen, their minute means are quite small, whilst minute volatility is similar for the four stock indexes, and an order of magnitude larger for the cryptocurrency. Concerning the

Table 2 Returns summary statistics for the selected five (minute) indexes in the period 02/01/2016–31/12/2018

	SPXUSD	JPXJPY	GRXEUR	ETXEUR	BTCUSD
Mean (%)	2.537e−05	6.619e−06	1.216e−06	−9.602e−06	3.414e−04
Std (%)	0.027	0.040	0.039	0.050	0.832
Min (%)	−1.180	−3.606	−10.703	−13.250	−597.220
Max (%)	1.402	2.297	2.153	2.742	596.405
Skewness	0.080	−0.769	−34.381	−43.542	−1.344
Kurtosis	71.099	137.630	9100	11186	4.980e+05

lowest returns, SPXUSD reaches a minimum on 2016/01/20, whilst JPXJPY and GRXEUR got the lowest return by mid February of 2016, for the period considered. Finally, both ETXEUR and BTCUSD reach a minimum by mid June of 2016. Furthermore, the returns of SPXUSD, JPXJPY and BTCUSD present a relatively small skewness, whilst the returns of GRXEUR and ETXEUR present a highly negative skewness. Most importantly, all five indexes possess a significantly high (excess) kurtosis, thus providing a strong, though not perfect, indication of fat tails in the returns distribution.

Two different approaches are commonly used in order to check whether our data is in the stable domain of attraction: (i) Q-Q plots, which are utilised as a statistical diagnostic that visualises the relationship between the empirical quantiles of a data set and the corresponding theoretical quantiles obtained under a specific distribution; (ii) thickness of the tails of the density function, as expressed by estimated characteristic exponent, α.

Figure 1 shows the Q-Q plots for the minute returns of SPXUSD, GRXEUR and BTCUSD indexes, which present highly distinct degrees of kurtosis (ref. Table 2). Three distribution assumptions are tested, namely, (i) normal, (ii) generalised extreme value, and (iii) α-stable. Clearly, the returns deviate significantly from a normal distribution, as expected, whilst the α-stable model yields an excellent approximation to the empirical quantiles, thus underpinning our choice for this family of heavy-tailed distributions. Similar behaviour is observed for the JPXJPY and ETXEUR indexes.

To further examine the thickness of the tails of the density function for the index returns considered herein, Fig. 2 shows the corresponding estimated characteristic exponent, α, as a function of the horizon τ (in minutes). Although the returns exhibit distinct statistical behaviours for different returns horizons τ, all indexes are characterised by a significantly high impulsiveness, as expressed by the corresponding characteristic exponent values, which are much smaller than 2.

4 Degree of impulsiveness as an additional source of information

Volatility, that is, the standard deviation of asset returns, is a key input for several financial applications like option pricing (Date & Islyaev, 2015) and risk management (McGee & McGroarty, 2017). Nevertheless, volatility is, by definition, a measure of how spread the returns of a given security or market index are about the mean. As such, it does not provide any information regarding the rate of decay of a distribution, which is related to the probability of extreme values for the associated random variable. This is especially important in various financial applications, such as risk quantification and portfolio optimisation. Specifically, heavy-tailed distributions possess heavier tails than an exponential distribution, tending to

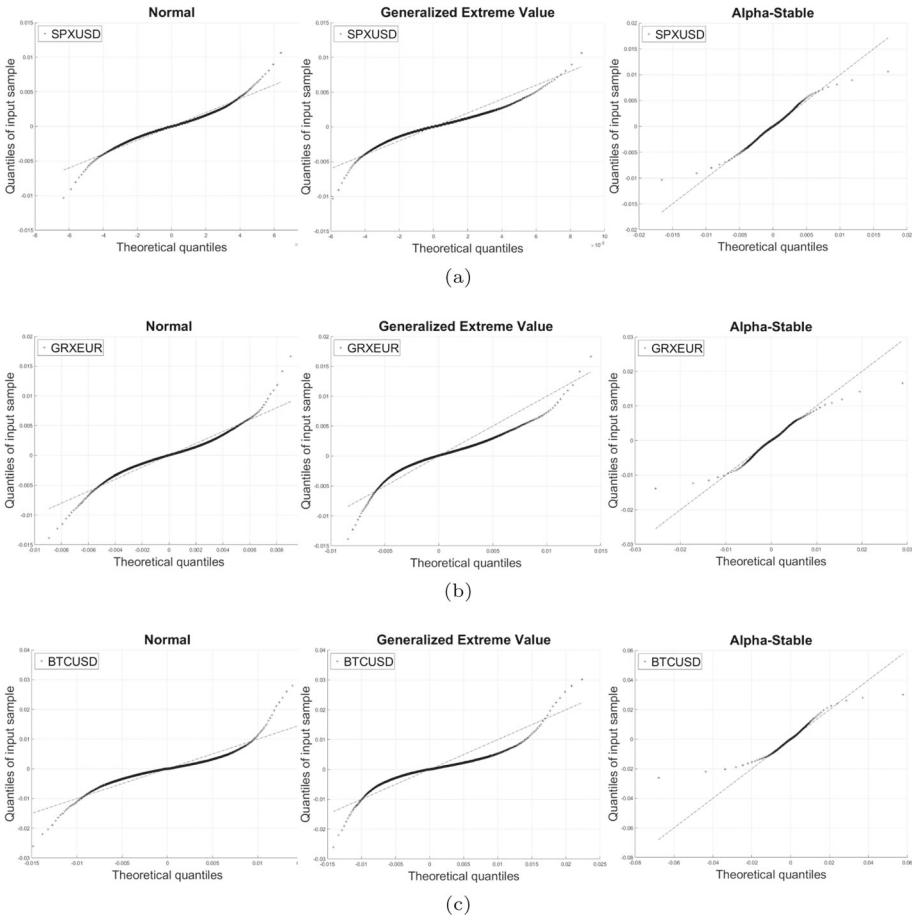

Fig. 1 Q-Q plots of empirical quantiles against theoretical quantiles of (i) normal, (ii) generalised extreme value, and (iii) α-stable distributions, for **a** SPXUSD, **b** GRXEUR, and **c** BTCUSD minute returns

present outliers with very high values. The heavier the tail, the larger the probability that the associated random variable will get one or more disproportionate values in a sample.

To illustrate this, we employ a toy example, as shown in Fig. 3a. In particular, a random series ("Original") of length $N = 1024$ is generated first by drawing samples from a Gaussian distribution, which is then corrupted by random spikes at 5% of the samples ("Impulsive" series). Although the two series have an almost equal volatility (about 3), their empirical probability density function differs significantly, as shown in Fig. 3b. Clearly, the "Impulsive" series is characterised by heavy tails, decreasing at a much slower rate, when compared against the "Original" version. The slow decay of probability in this area increases the odds of very extreme values (outliers), abrupt changes in the distribution, and other unexpected events at some point in the future. This justifies the quantification of impulsiveness, as an important additional source of information, that can complement the well-established volatility.

From physics, it is well known that the amount of energy carried by a wave is related to the amplitude of the wave, which is defined as the maximum amount of displacement from a rest position. In specific, the energy, E, transported by a wave is directly proportional to

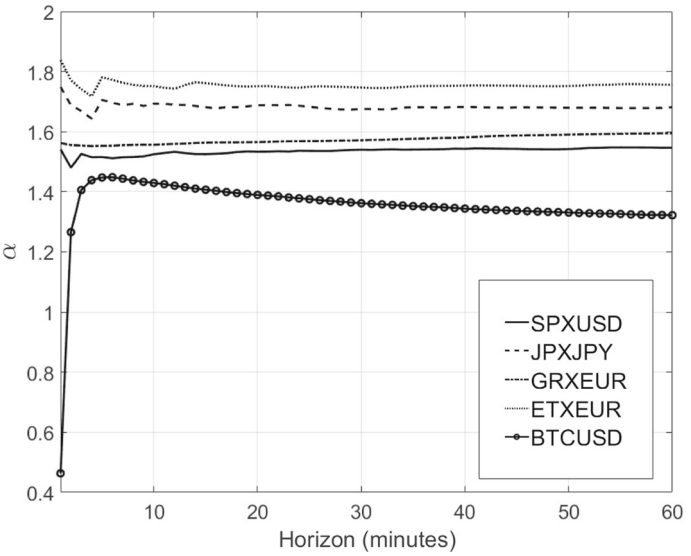

Fig. 2 Estimated characteristic exponent, α, of the returns series for the five indexes, as a function of the horizon $\tau = 1, \ldots, 60$ minutes

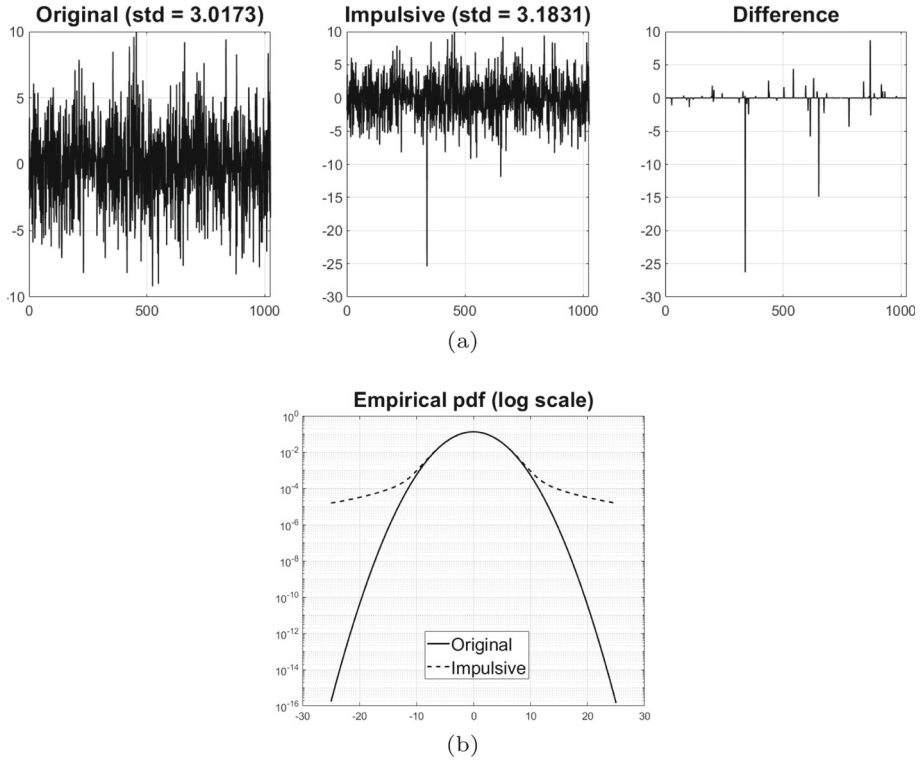

Fig. 3 Difference between volatility and impulsiveness. **a** Original data corrupted by randomly adding spikes; **b** corresponding empirical probability density functions

the square of the amplitude, A of the wave, that is, $E \propto A^2$. As a result, a high-energy wave is characterised by a high amplitude, whilst a low-energy wave is characterised by a low amplitude. Observing Fig. 3a, one can assert that the amplitude of a wave, i.e., a time series in our case, is related to the presence of large samples. Indeed, the amplitude of the "Impulsive" series is much higher than the amplitude of its "Original" counterpart. Thus there is a direct relation between the impulsiveness and the energy of a given time series.

In statistical signal processing terms, the probabilistic average energy of a random variable X is given by

$$E_{\text{avg}}(X) = \mathbb{E}\{X^2\} \,. \tag{6}$$

For a discrete time series of N samples, $\mathbf{x} \in \mathbb{R}^N$, the above equation can be expressed in asymptotic form as follows,

$$\tilde{E}_{\text{avg}}(\mathbf{x}) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} x_i^2 \tag{7}$$

or, in other words, the larger the number of samples N, the closer the average energy will be to the probabilistic average energy.

Nevertheless, (6) (and consequently (7)), which is expressed as a second-order moment, is not valid for heavy-tailed data modelled by α-stable distributions with $\alpha < 2$. A natural extension of the probabilistic average energy definition for α-stable random variables $X \sim S_\alpha(\gamma, \beta, 0)$ is as follows,

$$E_{\text{avg},\,p}(X) = \mathbb{E}\left\{|X|^p\right\} \tag{8}$$

which is precisely the definition of FLOMs. Similarly to (7), for a discrete time realisation $\mathbf{x} \in \mathbb{R}^N$ of $X \sim S_\alpha(\gamma, \beta, 0)$, its asymptotic approximation of the average energy is given by

$$\tilde{E}_{\text{avg},\,p}(\mathbf{x}) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} |x_i|^p \,. \tag{9}$$

In order to account for the differences in scale between two distinct time series, a normalisation can be applied with respect to a "rest position". We define this position to be the average absolute signal, defined by,

$$R_{\text{avg}}(X) = \mathbb{E}\{|X|\} \tag{10}$$

or, in discrete time form,

$$\tilde{R}_{\text{avg}}(\mathbf{x}) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} |x_i| \,. \tag{11}$$

From the above, our proposed *degree of impulsiveness* (DoI) of a random variable $X \sim S_\alpha(\gamma, \beta, 0)$, is defined as the deviation of the probabilistic average energy from the rest position, as follows,

$$\text{DoI}(X) = \frac{R_{\text{avg}}(X)}{\left(E_{\text{avg},\,p}(X)\right)^{1/p}} \overset{(3),(8)}{=} \frac{R_{\text{avg}}(X)}{\left(C_{p,\alpha,\beta}\gamma^p\right)^{1/p}} \tag{12}$$

where the exponent $1/p$ at the denominator of (12) is used to maintain the same unit of measurement as the variable X. The corresponding asymptotic approximation of the DoI is

given by substituting (9) and (11) to (12),

$$\tilde{\mathrm{DoI}}(\mathbf{x}) = \lim_{N\to\infty} \frac{\frac{1}{N}\sum_{i=1}^{N}|x_i|}{\left(\frac{1}{N}\sum_{i=1}^{N}|x_i|^p\right)^{1/p}} . \tag{13}$$

In practice, when the length, N, of the given time series is "small" or when a relatively large number of zeros exist in the samples, then (13) should be preferred (ignoring the limit operator) from a computational perspective. In any other case, (12) is employed for the calculation of the DoI, where the parameters α, β, γ, and $C_{p,\alpha,\beta}$ are estimated directly from the available data $\mathbf{x}$, as described in Sect. 2. The value of p is selected as a function of α, following the process described in that same Section. Finally, adopting a convention that is commonly used in signal processing, hereafter the $\mathrm{DoI}(X)$ and the $\tilde{\mathrm{DoI}}(\mathbf{x})$ will be expressed in decibels (dB), as follows, $\mathrm{DoI}_{\mathrm{dB}}(X) = 20\log_{10}(\mathrm{DoI}(X))$ and $\tilde{\mathrm{DoI}}_{\mathrm{dB}}(\mathbf{x}) = 20\log_{10}\left(\tilde{\mathrm{DoI}}(\mathbf{x})\right)$. Without loss of generality, in the following we employ the DoI definition in (12).

As an illustration of the validity of our proposed DoI indicator, we calculate its value for a set of synthetic signals with varying tail thickness (i.e., characteristic exponents). Specifically, a baseline signal of length $N = 2048$ is generated by drawing samples from a standard Gaussian distribution. Then, a set of outliers drawn from a $S_\alpha(\gamma, 0, 0)$ distribution corrupts the 10% of randomly chosen samples in the baseline signal. The dispersion is fixed at $\gamma = 1.5$ and the characteristic exponent varies in $\alpha \in \{1,\ 1.2,\ 1.4,\ 1.6,\ 1.8\}$. The simulation is repeated for 500 Monte Carlo runs, by keeping the same baseline signal and generating different corrupting outliers in each run. Finally, the average DoI is calculated for each α over all Monte Carlo runs. Figure 4a shows three instances of signals with varying tail thicknesses. As expected, the smaller the α, the more impulsive the signal is (i.e., more abrupt spikes occur), and the larger the DoI value should be. Indeed, as shown in Fig. 4b, the degree of impulsiveness decreases as the tail thickness reduces, which is also consistent with the visual inspection of the three signals on top.

Thus far, we have illustrated the validity of α-stable distributions in accurately modelling the impulsive nature of high-frequency financial returns, and we defined the degree of impulsiveness as a key source of information that complements the well-established volatility. In the following section, we propose a generalised volatility signature plot[2] based on the degree of impulsiveness for selecting the optimal sampling period of high-frequency impulsive data. Before proceeding, for convenience, Fig. 5 summarises the flow diagram of our proposed methodology for DoI-based optimal sampling period selection.

Remark 2 In order to avoid any misunderstanding in the subsequent analysis, we emphasise the difference between the original sampling period that generated the data and the optimal sampling period obtained by our proposed methodology. In particular, the original sampling period refers to the time interval between consecutive asset prices, as reported by a stock exchange. For instance, in the case of minute indexes, the original sampling period is equal to one minute. On the other hand, the optimal sampling period calculated by our method dictates an additional sampling that is applied to the original prices so as to improve the performance of a subsequent task (e.g. risk quantification). For instance, if the original sampling period is equal to one minute and the optimal sampling period is calculated to be equal to 5, this means that the original price series is subsampled by maintaining 1 out of every 5 consecutive samples. This subsampled series is then employed to carry out the subsequent task.

[2] Although volatility is undefined for α-stable models with $\alpha < 2$, we use the term "generalised volatility" exceptionally, in order to be aligned with the conventional volatility signature plot definition.

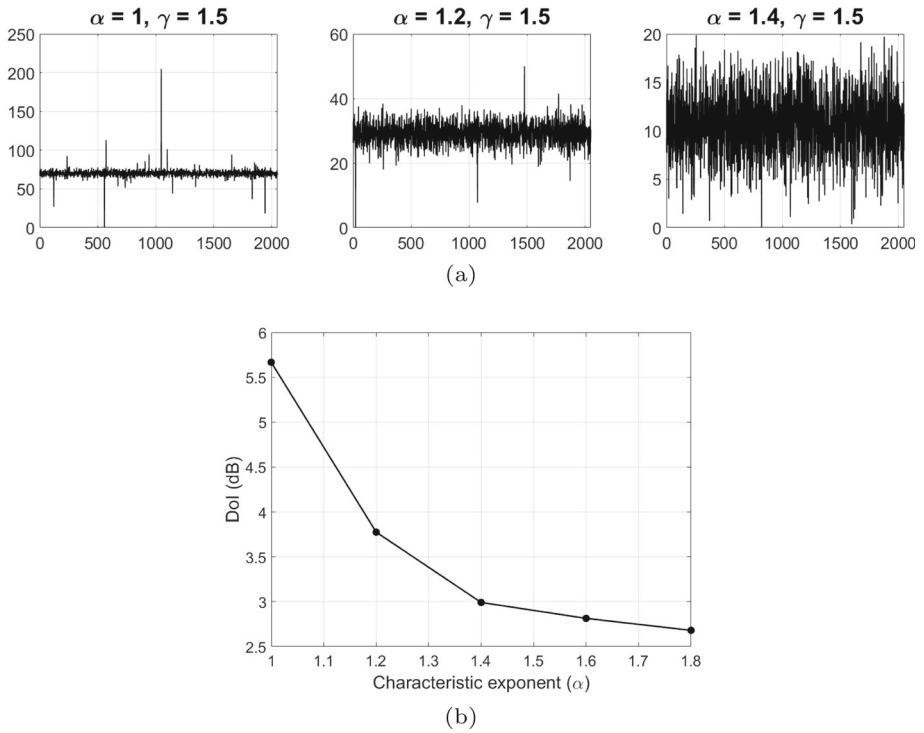

Fig. 4 **a** Instances of signals with varying tail thickness; **b** Average DoI (in dB) over 500 Monte Carlo runs, as a function of α

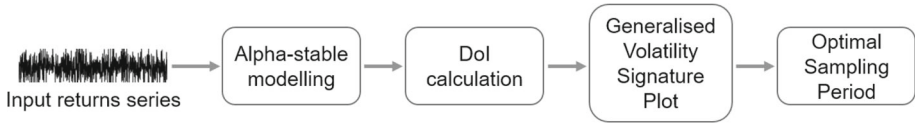

Fig. 5 Flow diagram of proposed methodology for DoI-based optimal sampling period selection

5 Impulsiveness-adaptive optimal sampling period estimation

Under the assumption that price changes have zero mean, then, the conventional price variogram is defined as a function of the returns horizon τ, as follows,

$$V(\tau) = \mathbb{E}\{(v_t - v_{t-\tau})^2\}\,. \tag{14}$$

To account for the generic case of mean and variance non-stationarity for the underlying generating processes of financial time series, Haslett (1997) introduced the following price variogram estimator, whose bias is small and independent of the magnitude of a potential drift in the data,

$$\tilde{V}(\tau) = \frac{1}{2}\frac{1}{(N-\tau-1)}\sum_{t=1}^{N-\tau}(d_{t,\tau} - \bar{d}_\tau)^2 \tag{15}$$

where $d_{t,\tau} = v_t - v_{t-\tau}$, and $\bar{d}_\tau$ is the sample mean of $d_{t,\tau}$ for $t = 1, \ldots, N-\tau$. Note that $\tilde{V}(\tau)$ is simply the sample variance of τ-horizon price differences. An important implication is that the volatility observed by sampling price series at a given horizon τ is itself dependent on that horizon, as follows (Hommes & LeBaron, 2018),

$$\tilde{\sigma}^2(\tau) = \frac{\tilde{V}(\tau)}{v_0^2 \cdot \tau} \tag{16}$$

where v_0 is either the current price or some medium-term average. A plot of $\tilde{\sigma}(\tau)$ versus τ is called a volatility signature plot (VSP). Fang (1996) and Andersen et al. (1999) further exploited the VSP to select the optimal sampling period for high-frequency returns. Specifically, the optimal period is calculated heuristically by identifying the value of τ where the $\tilde{\sigma}(\tau)$ curve begins to flatten out.

Nevertheless, both (14) and (16) are defined in terms of second-order moments of the price series, which may be infinite (undefined) in the case of α-stable distributed data. In order to address this drawback, we generalise the concepts of price variogram and volatility signature plot, so as to perfectly adapt to the underlying heavy-tailed data generating processes.

5.1 Generalised volatility signature plot

In order to adapt to the varying impulsiveness of distinct high-frequency returns series, a *generalised price variogram*, $V_g(\tau)$, is naturally defined by

$$V_g(\tau) = \mathbb{E}\{|v_t - v_{t-\tau}|^p\}\,. \tag{17}$$

By combining (12) with (17), a DoI-based expression is obtained for the generalised price variogram,

$$V_g(\tau) = \left(\frac{R_{\text{avg}}(d_{t,\tau})}{\text{DoI}(d_{t,\tau})}\right)^p \tag{18}$$

where $d_{t,\tau} = v_t - v_{t-\tau}$ is the random variable of τ-horizon price differences. Note that all the parameters involved in the calculation of (18) (i.e., p, α, β, and γ) are estimated from $d_{t,\tau}$ (ref. Section 2).

We also consider an alternative generalised price variogram estimator, defined as the sample central FLOM of price differences, $d_{t,\tau} = v_t - v_{t-\tau}$, as follows,

$$\tilde{V}_g(\tau) = \frac{1}{(N-\tau)} \sum_{t=1}^{N-\tau} |d_{t,\tau} - \bar{d}_\tau|^p \xrightarrow[N\to\infty]{(3),(9)} C_{\hat{p},\hat{\alpha},\hat{\beta}} \hat{\gamma}^{\hat{p}} \tag{19}$$

where $\bar{d}_\tau$ is the sample mean of $d_{t,\tau}$ for $t = 1, \ldots, N-\tau$. Note that the sample mean is defined properly for the α-stable distributed returns due to our constraint of $1 < \alpha \leq 2$ (ref. Remark 1). Furthermore, the parameters $\hat{p}, \hat{\alpha}, \hat{\beta}$ and $\hat{\gamma}$ in the asymptotic expression of $\tilde{V}_g(\tau)$ are estimated directly from the samples $\hat{d}_{t,\tau} = d_{t,\tau} - \bar{d}_\tau$, for $t = 1, \ldots, N-\tau$.

Remark 3 Notice that in the above definitions of the generalised price variogram (ref. (17)–(19)), the parameter p can either take a fixed predefined value, irrespectively of the returns horizon τ, or it can adapt to the statistics of the price differences $d_{t,\tau}$. In the latter case, the α-stable model parameters (p, α, β and γ) depend on τ. In the subsequent implementation, we employ the second approach, that is, the model parameters are estimated from the price differences for each τ, yet, for simplicity, in the corresponding equations we omit their dependence on τ.

Without loss of generality, in the following analysis we employ the alternative generalised price variogram of (19), which is more robust to potential drift in the data. Nevertheless, similar expressions are obtained by replacing $\tilde{V}_g(\tau)$ with $V_g(\tau)$ in the subsequent derivations.

Under the assumption that price changes have zero mean, which is a good approximation on short time scales, then, the generalised price variogram grows at a power law with the returns horizon τ, such that

$$\tilde{V}_g(\tau) = \kappa \cdot \tau^m \quad m \in \mathbb{R} \,. \tag{20}$$

In our implementation, m is estimated from the corresponding $\tilde{V}_g(\tau)$ vs. τ (similarly for $V_g(\tau)$ vs. τ) curve via nonlinear regression model fitting.

In the case of the conventional price variogram, $m = 1$. Furthermore, given the prevalence of multiplicative models for price changes on longer time scales, it has become customary to define the volatility σ in relative terms, even for short timescales, as follows (ref. Hommes and LeBaron (2018)),

$$\kappa = \sigma^2 v_0^2 \,. \tag{21}$$

In order to adapt to the heavy-tailed statistics of high-frequency returns, we define the variability of price changes by

$$\kappa = \tilde{\gamma}^\alpha v_0^\alpha \tag{22}$$

where the dispersion is employed instead of volatility (which, from a statistical perspective, is undefined for α-stable models with $\alpha > 1$), and the index of stability, α, is incorporated to account for the heaviness of the tails.

By combining (20) and (22), and noticing that the dispersion $\tilde{\gamma}$ depends on the returns horizon τ, we obtain the following equation,

$$\tilde{\gamma}(\tau) = \left(\frac{\tilde{V}_g(\tau)}{v_0^\alpha \cdot \tau^m} \right)^{1/\alpha} \,. \tag{23}$$

Then, our proposed *generalised volatility signature plot* (gVSP) is defined as the plot of $\tilde{\gamma}(\tau)$ versus τ.

5.2 Optimal sampling period estimation

Following the methodology of Fang (1996) and Andersen et al. (1999), we calculate the optimal sampling period heuristically by identifying the horizon τ where the gVSP curve, $\tilde{\gamma}(\tau)$ vs. τ, begins to flatten out. In the following, we assume for convenience that $v_0 = 1$ for all indexes. Figure 6a shows the gVSP curves for the five indexes as a function of the horizon $\tau = 1, \ldots, 120$ minutes. From a visual inspection, all curves start to flatten beyond a specific horizon τ.

In order to automate the process of calculating the optimal sampling period as the point τ^* where the gVSP curve, $\tilde{\gamma}(\tau)$ vs. τ, begins to flatten out, we apply a *local standard deviation filtering* scheme. Specifically, for $\tau = 1, \ldots, T$, let $\tilde{\gamma}(\tau)$ be a given gVSP curve, and $\tilde{\gamma}_f(\tau)$ be the corresponding filtered gVSP curve obtained by applying local standard deviation filtering on $\tilde{\gamma}(\tau)$. The filtering is performed simply by calculating the standard deviation in neighborhoods of size K, that is, $\tilde{\gamma}_f(\tau) = \text{std}\{\tilde{\gamma}(\tau - \lfloor K/2 \rfloor), \tilde{\gamma}(\tau - \lfloor K/2 \rfloor + 1), \ldots, \tilde{\gamma}(\tau + \lfloor K/2 \rfloor - 1), \tilde{\gamma}(\tau + \lfloor K/2 \rfloor)\}, \tau = 1, \ldots, T$. Finally, the optimal sampling period is determined as the point τ^* for which $\tilde{\gamma}_f(\tau^*) \leq \sigma_{\text{thr}}$, where σ_{thr} is a predefined threshold standard

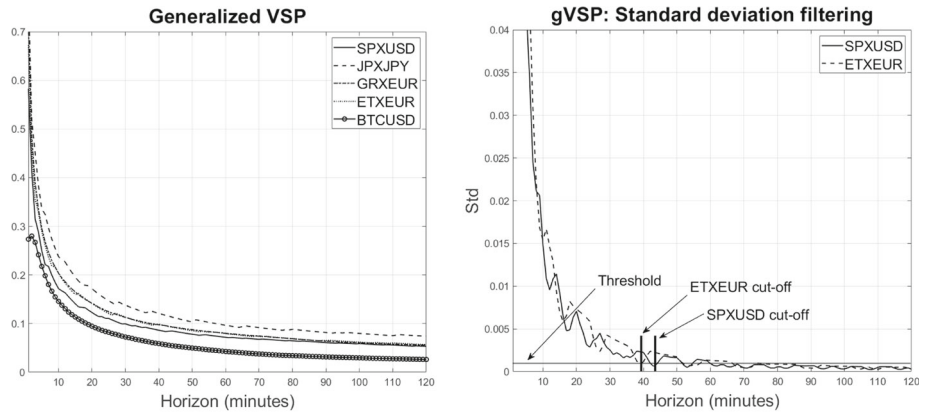

Fig. 6 **a** Generalised VSPs for the five indexes; **b** Optimal sampling period estimation via standard deviation filtering, for SPXUSD and ETXEUR indexes, as a function of the returns horizon $\tau = 1, \ldots, 120$ minutes

deviation. In our implementation, we set $K = 5$ and $\sigma_{\text{thr}} = 0.001$. Figure 6b illustrates the filtered gVSP curves, along with the threshold level and the corresponding optimal sampling periods identified for the SPXUSD (= 42 minutes) and ETXEUR (= 39 minutes) indexes. Note that the optimal period depends on the threshold, nevertheless, our experimental evaluation showed that a threshold at the order of 10^{-3} suffices to identify accurately the plateau's starting point for the five distinct indexes.

6 Empirical evaluation in risk quantification

Since there is not a ground truth for the optimal sampling period, in this section, the performance of our proposed method is evaluated empirically by solving the problem of risk quantification for high-frequency financial indexes. In particular, the optimal sampling period is calculated first by means of (i) the conventional and (ii) our generalised volatility signature plot. Then, the original price series is subsampled according to the optimal sampling period and the corresponding risk is quantified by employing two well-established risk measures, namely, value-at-risk (VaR) and expected shortfall (ES). We emphasize that this work does not intend to focus on the risk quantification problem, which is rather used as a test case to evaluate the performance of our methodology.

Given a random variable of returns r_t, a confidence level $c \in (0, 1)$, and a holding period T_H (i.e., the time period over which losses may occur), the ϑ-level VaR and ES, with $\vartheta = 1-c$, are defined by

$$\Pr\left(r_t \leq -\text{VaR}_t(\vartheta)\right) = \vartheta \quad \text{ES}_t(\vartheta) = -\mathbb{E}\{r_t \mid r_t \leq -\text{VaR}_t(\vartheta)\}\,. \tag{24}$$

Like VaR, ES is universal and can be applied to almost any instrument and underlying source of risk. The Basel Committee (BCBS, 2013) recommends to set $c = 99\%$ (equivalently, $\vartheta = 0.01$) for VaR, and is now proposing to move towards ES with $c = 97.5\%$ (equivalently, $\vartheta = 0.025$) since it theoretically captures better the information contained in the leftmost tail of returns distribution.

In the following, the exponential weighted moving average (EWMA) method is used as a benchmark for the calculation of VaR and subsequently of ES. For convenience, an infinitely large estimation window is typically assumed to approximate the EWMA-based variance via a simple recurrence formula,

$$\hat{\sigma}_t^2 \approx (1-\lambda)(r_{t-1}^2 + \sum_{i=2}^{\infty} \lambda^{i-1} r_{t-i}^2) = (1-\lambda) r_{t-1}^2 + \lambda \hat{\sigma}_{t-1}^2 \,. \tag{25}$$

A value of the decay factor frequently used in practice, and adopted hereafter, is $\lambda = 0.94$ (ref. Morgan & Reuters, 1996). Furthermore, both VaR and ES are calculated on a rolling window fashion. In the following experimentation, the window size is fixed to $w = 2880$ (i.e., 48 h) and the step size is equal to $s = 1$. To assess the predictive accuracy of the VaR and ES forecasts based on the optimal sampling periods calculated with the conventional and our generalised volatility signature plot, we adopt well-established backtesting methods, whose performance is compared against the naive approach, that is, no subsampling is applied to the original price series.

Regarding VaR, the number of observed violations ($P_{V,obs}$), is a commonly adopted performance metric. Additionally, we employ the binomial test (Danielsson, 2011), which assesses whether the number of failures is consistent with the VaR confidence level. The test statistic (TStatBin) and the corresponding p-value (PValueBin) of the binomial test are reported. As a third VaR backtesting method, the conditional predictive ability test introduced by Giacomini and White (2006), hereafter denoted as GW test, is employed. The GW test examines the equal conditional predictive ability of the three distinct sampling strategies (i.e., (i) no subsampling, (ii) subsampling using the conventional VSP, and (iii) subsampling using our generalised VSP) for VaR quantification, against the benchmark model. In our case, we assume that the benchmark model yields the true return values. Concerning the VaR measures compared herein, the outperforming method (i.e., sampling strategy) is the one with the highest GW test value and the lower GW p-value, defined below.

$$\text{TStatGW} = \sqrt{N} \frac{\mathbb{E}\{E\}}{\sqrt{\text{NWHAC}(E)}} \tag{26}$$

where N is the number of samples, $E = (E_r - E_b)^2$ with E_r and E_b denoting, respectively, the random variables representing the errors between the reference VaR measure and the true VaR values, and the benchmark VaR measure and the true VaR values. Herein we consider that $E_b = 0$. The function NWHAC(E) denotes the Newey West HAC variance estimator. The associated p-value is defined by

$$\text{PValueGW} = 1 - F_{\chi^2}(\text{TStatGW}^2) \tag{27}$$

where F_{χ^2} is the χ^2 cumulative distribution with one degree of freedom.

As for the ES, we backtest it by employing two distinct performance indicators: (1) the observed versus the actual ϑ level, along with the expected versus the actual ES failures; (2) a nonparametric test (T_{ES}^1) proposed by Acerbi and Szekely (2014), which is free from assumptions on distribution, with greater ability to detect an effect than the VaR test, while also eliminating the need for Monte Carlo simulations for most practical cases. Furthermore, T_{ES}^1 scales the losses by the corresponding ES value based on the unconditional relationship

$\mathrm{ES}_t(\vartheta) = -\mathbb{E}\left\{\frac{r_t \mathbb{1}_t}{\theta}\right\}$, and reports the associated unconditional test statistic,

$$\mathrm{TStatZ}^1_{\mathrm{ES}} = \frac{1}{N\theta}\sum_{t=1}^{N}\frac{r_t \mathbb{1}_t}{\mathrm{ES}_t(\vartheta)} + 1 \tag{28}$$

with $\mathbb{1}(\cdot)$ being the indicator function. Under the assumption that the distributional assumptions for the returns are correct, it holds that $\mathbb{E}\{\mathrm{TStatZ}^1_{\mathrm{ES}}\} = 0$. Negative values of the test statistic indicate risk underestimation. The unconditional test is a one-sided test that rejects when there is evidence that the model underestimates risk. Furthermore, the test rejects the model when the p-value is less than 1 minus the test confidence level. Most importantly, $\mathrm{TStatZ}^1_{\mathrm{ES}}$ turns out to be stable across a range of distributional assumptions for r_t, from thin-tailed up to heavy-tailed distributions.

As a first evaluation, we compare the performance of VaR quantification based on the (i) original minute indexes without subsampling (hereafter denoted by No-S), (ii) subsampled indexes using the optimal sampling period calculated via the conventional VSP (hereafter denoted by VSP-S), and (iii) subsampled indexes using the optimal sampling period calculated via our proposed generalised VSP (hereafter denoted by gVSP-S), by varying the holding period $T_H \in \{360, 720\}$ minutes.

Tables 3 and 4 display the results of the VaR backtesting methods described above, for $T_H = 360$ and $T_H = 720$, respectively. As it can be seen, for the smaller holding period, $T_H = 360$, our method (gVSP-S) achieves a superior performance for all indexes, with an observed level almost equal to the true VaR level. On the other hand, the conventional VSP-S overestimates risk for all indexes, whilst the no-subsampling strategy (No-S) significantly underestimates risk, as revealed by the extremely larger number of VaR violations for the majority of the indexes. The improved performance of our gVSP-S method is also verified through the binomial test, which explicitly accepts the hypothesis that the number of failures is consistent with the VaR confidence level for SPXUSD, JPXJPY, and GRXEUR indexes. Regarding ETXEUR and BTCUSD, although the binomial test rejects the null hypothesis for all the three sampling strategies, we observe that our method underestimates (positive TStatBin) or overestimates (negative TStatBin) VaR at a significantly lower degree than No-S and VSP-S. Furthermore, our proposed method is consistently closer to the corresponding benchmark model, as expressed by the smaller GW test value for all the five indexes.

Concerning the more challenging case of a larger holding period, $T_H = 720$, the performance of the three sampling strategies deteriorates, as expected, since we are estimating events that occur rarely. Nevertheless, our gVSP-S method demonstrates an improved accuracy in quantifying VaR, when compared against No-S and VSP-S, as expressed by the ratio of the observed ($\mathrm{P}_{\mathrm{V,obs}}$) over the expected ($\mathrm{P}_{\mathrm{V,expected}}$) VaR violations, which is closer to one for all except for the ETXEUR index. In this latter case, the No-S strategy is better than the other two alternatives. Regarding the binomial test, although it rejects the null hypothesis in the vast majority of indexes, however, gVSP-S outperforms the other two strategies, as revealed by the significantly smaller TStatBin values, which means that our method neither overestimates nor underestimates VaR so heavily as No-S and VSP-S. Most importantly, gVSP-S is capable of better estimating VaR even for extremely skewed and kurtotic indexes, such as GRXEUR and BTCUSD. Concerning the proximity to the corresponding benchmark model, gVSP-S is consistently outperforming No-S and VSP-S, as expressed by its smaller GW test value for all the five indexes.

In the following, we backtest ES for the three sampling strategies, for $T_H \in \{360, 720\}$ minutes. Tables 5 and 6 show the results of the ES backtesting methods described above, along with the p-value and the critical value of the $\mathrm{T}^1_{\mathrm{ES}}$ test. Table 5 displays the backtesting

Table 3 VaR backtesting for No-S, gVSP-S, and VSP-S, for $T_H = 360$

Index	Method	VaRLevel	ObservedLevel	$P_{V,obs}$	$P_{V,expected}$	Bin	TStatBin	PValueBin	TStatGW	PValueGW
SPXUSD	No-S	0.99	0.984	13124	8149	Reject	55.380	0	6.297	3.025e−10
SPXUSD	gVSP-S	0.99	0.989	187	163	Accept	1.891	0.029	3.293	9.879e−04
SPXUSD	VSP-S	0.99	0.996	378	905	Reject	− 17.618	0	4.338	1.437e−05
JPXJPY	No-S	0.99	0.988	8746	7945	Reject	9.025	0	7.087	1.369e−12
JPXJPY	gVSP-S	0.99	0.989	144	139	Accept	0.393	0.347	3.339	8.398e−04
JPXJPY	VSP-S	0.99	0.993	230	318	Reject	− 4.950	3.702e−07	3.756	1.728e−04
GRXEUR	No-S	0.99	0.975	15881	6333	Reject	120.58	0	5.469	4.511e−08
GRXEUR	gVSP-S	0.99	0.989	132	117	Accept	1.367	0.085	3.032	0.002
GRXEUR	VSP-S	0.99	0.995	234	396	Reject	− 8.1746	1.110e−16	3.649	2.624e−04
ETXEUR	No-S	0.99	0.987	5467	4410	Reject	15.988	0	4.962	6.964e−07
ETXEUR	gVSP-S	0.99	0.992	80	110	Reject	− 2.896	0.001	2.982	0.002
ETXEUR	VSP-S	0.99	0.996	110	260	Reject	− 9.324	0	3.448	5.646e−04
BTCUSD	No-S	0.99	0.998	2480	10624	Reject	− 79.408	0	7.214	5.420e−13
BTCUSD	gVSP-S	0.99	0.986	134	93	Reject	4.250	1.069e−05	2.727	0.006
BTCUSD	VSP-S	0.99	0.998	103	424	Reject	− 15.696	0	3.424	6.154e−04

Note: In Tables 3 and 4 the following notation is used: (i) No-S: original indexes are used without subsampling, (ii) gVSP-S: indexes are optimally subsampled using our proposed gVSP method, and (iii) VSP-S: indexes are optimally subsampled using the conventional VSP method. All test statistics and p-values correspond to a 95% test confidence level

Table 4 VaR backtesting for No-S, gVSP-S, and VSP-S, for $T_H = 720$

Index	Method	VaRLevel	ObservedLevel	$P_{V,obs}$	$P_{V,expected}$	Bin	TStatBin	PValueBin	TStatGW	PValueGW
SPXUSD	No-S	0.99	0.974	21104	8146	Reject	144.290	0	6.329	2.464e−10
SPXUSD	gVSP-S	0.99	0.994	97	163	Reject	−5.189	1.052e−07	3.314	9.215e−04
SPXUSD	VSP-S	0.99	0.991	805	905	Reject	−3.344	4e-04	4.371	1.236e−05
JPXJPY	No-S	0.99	0.985	11462	7942	Reject	39.698	0	7.117	1.105e−12
JPXJPY	gVSP-S	0.99	0.994	86	139	Reject	−4.540	2.811e−06	3.353	7.988e−04
JPXJPY	VSP-S	0.99	0.995	164	318	Reject	−8.665	0	3.762	1.688e−04
GRXEUR	No-S	0.99	0.969	19068	6330	Reject	160.92	0	5.401	6.639e−08
GRXEUR	gVSP-S	0.99	0.992	83	117	Reject	−3.175	7e-04	3.065	0.002
GRXEUR	VSP-S	0.99	0.994	242	395	Reject	−7.762	4.219e−15	3.667	2.458e−04
ETXEUR	No-S	0.99	0.987	5874	4406	Reject	22.211	0	4.799	1.592e−06
ETXEUR	gVSP-S	0.99	0.994	56	110	Reject	−5.187	1.069e−07	3.0174	0.002
ETXEUR	VSP-S	0.99	0.995	124	259	Reject	−8.440	0	3.466	5.278e−04
BTCUSD	No-S	0.99	0.996	3543	10620	Reject	−69.019	0	7.222	5.090e−13
BTCUSD	gVSP-S	0.99	0.991	87	93	Accept	−0.640	0.260	2.725	0.006
BTCUSD	VSP-S	0.99	0.999	39	425	Reject	−18.812	0	3.429	6.047e−04

Table 5 ES backtesting for No-S, gVSP-S, and VSP-S, for $T_H = 360$

Index	Method	VaRLevel	ObservedLevel	Failures	Expected	Unconditional	PValueZ^1_{ES}	TStatZ^1_{ES}	CriticalValueZ^1_{ES}
SPXUSD	No-S	0.99	0.984	13124	8100	Accept	0.897	−0.852	−3.846
SPXUSD	gVSP-S	0.99	0.989	167	150	Accept	1	−0.172	−3.286
SPXUSD	VSP-S	0.99	0.996	378	900	Accept	1	0.587	−4.129
JPXJPY	No-S	0.99	0.988	8736	7900	Accept	0.913	−0.169	−2.025
JPXJPY	gVSP-S	0.99	0.990	97	100	Accept	1	−0.025	−3.819
JPXJPY	VSP-S	0.99	0.993	211	300	Accept	1	0.303	−3.062
GRXEUR	No-S	0.99	0.975	15847	6300	Accept	0.887	−1.945	−6.117
GRXEUR	gVSP-S	0.99	0.989	106	100	Accept	1	−0.082	−4.824
GRXEUR	VSP-S	0.99	0.994	225	350	Accept	1	0.358	−3.966
ETXEUR	No-S	0.99	0.987	5459	4400	Accept	0.841	−0.355	−1.918
ETXEUR	gVSP-S	0.99	0.992	74	100	Accept	1	0.245	−2.815
ETXEUR	VSP-S	0.99	0.996	103	250	Accept	1	0.574	−2.665
BTCUSD	No-S	0.99	0.998	2480	10600	Accept	0.958	0.519	−1.370
BTCUSD	gVSP-S	0.99	0.986	69	50	Accept	1	−0.699	−2.631
BTCUSD	VSP-S	0.99	0.998	100	400	Accept	1	0.732	−3.008

Note: In Tables 5 and 6 the following notation is used: (i) No-S: original indexes are used without subsampling, (ii) gVSP-S: indexes are optimally subsampled using our proposed gVSP method, and (iii) VSP-S: indexes are optimally subsampled using the conventional VSP method. All test statistics and p-values correspond to a 95% test confidence level

Table 6 ES backtesting for No-S, gVSP-S, and VSP-S, for $T_H = 720$

Index	Method	VaRLevel	ObservedLevel	Failures	Expected	Unconditional	PValueZ^1_{ES}	TStatZ^1_{ES}	CriticalValueZ^1_{ES}
SPXUSD	No-S	0.99	0.974	21104	8100	Accept	0.800	− 2.168	− 3.793
SPXUSD	gVSP-S	0.99	0.994	87	150	Accept	1	0.407	− 3.751
SPXUSD	VSP-S	0.99	0.991	805	900	Accept	1	0.078	− 3.955
JPXJPY	No-S	0.99	0.986	11462	7900	Accept	0.888	− 0.596	− 2.006
JPXJPY	gVSP-S	0.99	0.994	53	100	Accept	1	0.479	− 3.375
JPXJPY	VSP-S	0.99	0.995	149	300	Accept	1	0.516	− 2.915
GRXEUR	No-S	0.99	0.969	19045	6300	Accept	0.905	− 2.583	− 6.466
GRXEUR	gVSP-S	0.99	0.993	68	100	Accept	1	0.341	− 4.389
GRXEUR	VSP-S	0.99	0.992	237	350	Accept	1	0.339	− 4.214
ETXEUR	No-S	0.99	0.987	5874	4400	Accept	0.902	− 0.416	− 1.815
ETXEUR	gVSP-S	0.99	0.995	48	100	Accept	1	0.516	− 3.087
ETXEUR	VSP-S	0.99	0.995	117	250	Accept	1	0.546	− 2.469
BTCUSD	No-S	0.99	0.997	3543	10600	Accept	0.933	0.352	− 1.316
BTCUSD	gVSP-S	0.99	0.989	51	50	Accept	1	− 0.114	− 2.854
BTCUSD	VSP-S	0.99	0.999	44	400	Accept	1	0.889	− 2.956

performance for the smaller holding period of $T_H = 360$ minutes. Clearly, our gVSP-S method outperforms significantly the other two sampling strategies, in terms of a more accurate estimation of the realised number of ES failures, for all indexes. Especially for the highly skewed and kurtotic indexes, namely, GRXEUR, ETXEUR and BTCUSD, the No-S and VSP-S methods deviate significantly from the expected number of violations by heavily underestimating and overestimating ES, respectively. On the contrary, our proposed gVSP-S method achieves a close approximation of the tail risk.

As for the larger holding period of $T_H = 720$ minutes, Table 6 shows that the VSP-S method performs better for the SPXUSD index, in terms of the number of identified violations. Our gVSP-S method is the next best performing, whilst No-S significantly underestimates risk. For the ETXEUR index, the strategy of no subsampling achieves a ratio of the observed over the expected number of ES violations that is closer to one, when compared against gVSP-S and VSP-S, both of which achieve a similar performance. These two methods yield an equal performance for the JPXJPY index, with the No-S method heavily underestimating risk, whilst gVSP-S is slightly better in the case of GRXEUR index. Finally, regarding the highly skewed and kurtotic BTCUSD index, our gVSP-S method is better capable of capturing the tail risk, yielding a significantly more accurate approximation of the true ϑ level.

The above results demonstrate the improved capabilities of our proposed generalised method for optimal sampling period calculation, in better adapting to a broad range of impulsive behaviours that are inherent to distinct market indexes. This yields a more accurate quantification of risk measures such as VaR and ES, which also achieves a better trade-off between under-/over-estimation of risk. Regarding the statistics of an index (e.g. skewness, kurtosis) as a potential factor that could affect the estimation performance of VaR and ES based on our proposed sampling strategy, the results do not indicate a clear connection between the two, thus revealing an increased robustness of our methodology.

7 Conclusions and future work

This paper proposes a novel methodology for taming the impulsiveness that is inherent to high-frequency financial returns, via the calculation of a proper optimal sampling period. Our method is grounded on the new concept of the degree of impulsiveness (DoI), as an alternative source of information for characterising the statistical behaviour of such data. Then, a generalised volatility signature plot is defined based on the DoI, demonstrating a better adaptability to a broad range of impulsive behaviours in high-frequency returns. Finally, our generalised volatility signature plot is coupled with a local standard deviation filtering scheme to calculate the optimal sampling period.

An empirical evaluation of our method is performed in the framework of risk quantification. Specifically, a subsampling strategy is applied first to the price series of five distinct minute indexes, followed by the estimation of two well-established risk measures, namely, VaR and ES. Our proposed method for optimal sampling period calculation is compared against the no-subsampling alternative, as well as against the method that calculates the optimal sampling period based on the conventional volatility signature plot, which relies on second-order moments. The experimental results revealed the clear superiority of our proposed method, when compared against the two alternative sampling strategies.

The methodology developed in this paper offers several open research avenues. Linked with financial signal processing, it should be interesting to investigate the performance of our

method in carrying out various challenging financial data processing tasks, such as denoising for high-frequency financial data mining or jump and volatility analysis. Furthermore, an interesting question would be to formulate the problem of finding the optimal FLOM order, p, jointly with a risk (e.g. VaR or ES) minimisation criterion, instead of the lookup table approach described in Sect. 2. Finally, our optimal sampling period calculation method could be meaningfully applied to risk parity portfolio construction and market co-integration analysis based on high-frequency financial data.

Author Contributions Conceptualization: G. Tzagkarakis, F. Maurer; Methodology: G. Tzagkarakis, F. Maurer; Formal analysis and investigation: All authors; Writing—original draft preparation: G. Tzagkarakis, F. Maurer; Writing—review and editing: All authors.

Funding Open access funding provided by HEAL-Link Greece. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability Not applicable.

Code Availability Not applicable.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication All authors have approved the manuscript and give their consent for submission and publication.

References

Acerbi, C., & Szekely, B. (2014). Backtesting expected shortfall. Tech. rep., MSCI Inc., https://www.msci.com/documents/10199/22aa9922-f874-4060-b77a-0f0e267a489b

Aït-Sahalia, Y., Mykland, P. A., & Zhang, L. (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *The Review of Financial Studies, 18*(2), 351–416. https://doi.org/10.1093/rfs/hhi016

Andersen, T. G., Bollerslev, T., & Diebold, F. X., & Labys, P. (1999). (Understanding, optimizing, using and forecasting) Realized volatility and correlation. New York University, Leonard N Stern School Finance Dept Working Paper Series. https://archive.nyu.edu/bitstream/2451/27128/2/wpa99061.pdf

Andersen, T. G., Bollerslev, T., Diebold, F. X., et al. (2000). Great realizations. *Risk, 13*(3), 105–108.

Ané, T., & Geman, H. (2000). Order flow, transaction clock, and normality of asset returns. *The Journal of Finance, 55*(5), 2259–2284.

Bandi, F. M., & Russell, J. R. (2006). Separating microstructure noise from volatility. *Journal of Financial Economics, 79*, 655–692. https://doi.org/10.1016/j.jfineco.2005.01.005

BCBS. (2013). Fundamental review of the trading book: A revised market risk framework. Consultative document, Basel Committee on Banking Supervision, https://www.bis.org/publ/bcbs265.pdf

Bhattacharyya, M., Kumar, M. D., & Kumar, R. (2009). Optimal sampling frequency for volatility forecast models for the Indian stock markets. *Journal of Forecasting, 28*, 38–54. https://doi.org/10.1002/for.1080

Choi, H., & Kang, M. (2014). Optimal sampling frequency for high frequency data using a finite mixture model. *Journal of the Korean Statistical Society, 43*(2), 251–262. https://doi.org/10.1016/j.jkss.2013.09.003

Danielsson, J. (2011). *Financial risk forecasting: The theory and practice of forecasting market risk with implementation in R and Matlab*. Wiley Finance.

Date, P., & Islyaev, S. (2015). A fast calibrating volatility model for option pricing. *European Journal of Operational Research, 243*(2), 599–606. https://doi.org/10.1016/j.ejor.2014.12.031

Delaney, L. (2018). Investment in high-frequency trading technology: A real options approach. *European Journal of Operational Research, 270*(1), 375–385. https://doi.org/10.1016/j.ejor.2018.03.025

Fang, Y. (1996). Volatility modeling and estimation of high-frequency data with Gaussian noise. Phd thesis, MIT, Sloan School of Management, https://dspace.mit.edu/handle/1721.1/11041

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica, 74*(6), 1545–1578. https://doi.org/10.1111/j.1468-0262.2006.00718.x

Haslett, J. (1997). On the sample variogram and the sample autocovariance for non-stationary time series. *Journal of the Royal Statistical Society: Series D (The Statistician), 46*(4), 475–485.

Hommes, C., & LeBaron, B. (2018). *Handbook of computational economics, volume 4: Heterogeneous agent modeling*. Elsevier.

Morgan, J. P., Reuters. (1996). Riskmetrics-technical document. Tech. rep., MSCI Inc., https://cutt.ly/xx9jnoJ, p. 97. Accessed March 26, 2021.

Kogon, S., & Williams, D. (1998). Characteristic function based estimation of stable parameters. In R. Adler, R. Feldman, & M. Taqqu (Eds.), *A practical guide to heavy tailed data Boston* (pp. 311–338). Birkhäuser.

Kuruoglu, E. (2001). Density parameter estimation of skewed α-stable distributions. *IEEE Transactions on Signal Processing, 49*(10), 2192–2201. https://doi.org/10.1109/78.950775

McGee, R. J., & McGroarty, F. (2017). The risk premium that never was: A fair value explanation of the volatility spread. *European Journal of Operational Research, 262*(1), 370–380. https://doi.org/10.1016/j.ejor.2017.03.070

Nikias, C., & Shao, M. (1995). *Signal processing with alpha-stable distributions and applications*. Wiley.

Nolan, J. P. (1997). Numerical calculation of stable densities and distribution functions. *Communications in Statistics-Stochastic Models, 13*(4), 759–774. https://doi.org/10.1080/15326349708807450

Nolan, J. P. (2020). Univariate stable distributions—Models for heavy tailed data. New York: Springer Verlag, chapter 1 online at https://edspace.american.edu/jpnolan

Samorodnitsky, G., & Taqqu, M. (1994). *Stable non-Gaussian random processes: stochastic models with infinite variance*. New York: Chapman & Hall.

Tzagkarakis, G., Beferull-Lozano, B., & Tsakalides, P. (2006). Rotation-invariant texture retrieval with Gaussianized steerable pyramids. *IEEE Transactions on Image Processing, 15*(9), 2702–2718. https://doi.org/10.1109/TIP.2006.877356