Data Mining in Finance: Current Advances and Future Challenges

Eric Paquet, Herna Viktor and Hongyu Guo

Abstract Data mining has been successfully applied in many businesses, thus aiding managers to make informed decisions that are based on facts, rather than having to rely on guesswork and incorrect extrapolations. Data mining algorithms equip institutions to predict the movements of financial indicators, enable companies to move towards more energy-efficient buildings, as well as allow businesses to conduct targeted marketing campaigns and forecast sales. Specific data mining success stories include customer loyalty prediction, economic forecasting, and fraud detection. The strength of data mining lies in the fact that it allows for not only predicting trends and behaviors, but also for the discovery of previously unknown patterns. However, a number of challenges remain, especially in this era of big data. These challenges are brought forward due to the sheer Volume of today's databases, as well as the Velocity (in terms of speed of arrival) and the Variety, in terms of the various types of data collected. This chapter focuses on techniques that address these issues. Specifically, we turn our attention to the financial sector, which has become paramount to business. Our discussion centers on issues such as considering data distributions with high fluctuations, incorporating late arriving data, and handling the unknown. We review the current state-of-the-art, mainly focusing on model-based approaches. We conclude the chapter by providing our perspective as to what the future holds, in terms of building accurate models against today's business, and specifically financial, data.

Keywords Financial data · Time series · Data streams · Volatility · Stochastic · Marginalisation · Path integral · Bayesian learning · Energy load forecasting

E. Paquet · H. Guo
National Research Council of Canada, Building M-50,
1200 Montreal Road, Ottawa, Canada
e-mail: hongyu.guo@nrc-cnrc.gc.ca

E. Paquet
e-mail: eric.paquet@nrc-cnrc.gc.ca

E. Paquet · H. Viktor (✉)
School of Electrical Engineering and Computer Science, University of Ottawa,
800 King Edward Road, Ottawa, Canada
e-mail: hviktor@uottawa.ca

N. Japkowicz and J. Stefanowski (eds.), *Big Data Analysis: New Algorithms for a New Society*, Studies in Big Data 16, DOI 10.1007/978-3-319-26989-4_7

1 Introduction

Data mining has been successfully applied to many businesses, thus aiding managers to make informed decisions that are based on facts, rather than having to rely on guesswork and incorrect extrapolations. Data mining algorithms allow companies to explore the trends in terms of sales, to predict the movements of financial indicators, and to construct energy-aware buildings, amongst others. Specific data mining (or business analytics) success stories include customer loyalty prediction and sales forecasting, fraud detection, estimating the correlations between stocks and predicting the movements of financial markets. Case studies show that the strength of data mining lies in the fact that it allows for not only predicting trends and behaviors, but also for the discovery of previously unknown patterns in business data.

Making predictions and building trading models are central goals for financial institutions. It is no surprise that this was one of the earliest areas of the application of modern machine learning techniques to real world problems. In this sector, a number of unique challenges need to be addressed. These challenges are brought forward due to the sheer Volume, Velocity (in terms of speed of arrival) and the potential Variety, of the data. In addition, another issue here is that we aim to build an accurate model against uncertain, rapidly changing, and often rather unpredictable, data. That is, the financial sector continuously processes millions, if not trillions, of transactions. For example, the values of stocks are updated at regular intervals, typically every few seconds. These markets require the use of advanced models in order to facilitate trend spotting and to provide some financial trajectory. Ideally, in this scenario, we require just-in-time adaptive models that are accurate even as the data changes, due to concept drifts.

There are many unknowns associated with such financial data, which makes the construction of data mining models a major challenge. Here, analyzing and understanding what attributes and parameters we *do not know* is crucial in order for us to create accurate and meaningful predictions. This fact limits the application of traditional data-driven algorithms, in that we often cannot make assumptions about data distributions or types of relationships. The typical non-parametric way used by most data mining algorithms, to search a large data set to see whether any patterns are exhibited in that set, has limited applicability in a financial setting. Here, the data are susceptible to drift, arrive at a fast rate, may contain late-arriving data, and have parameters that are difficult to estimate. Thus, this type of traditional analysis and model construction may not be ideal when aiming to construct models against big data in finance, where the number of unknowns (and in essence the randomness) is high. Rather, the use of stochastic, model-based approaches comes to mind.

This chapter addresses the above-mentioned issues associated with Volume, Velocity and Variance in big data, while focusing on the financial sector. To this end, we review the state-of-the-art in terms of techniques to mine stocks, bonds, and interest rates. We note that Bayesian approaches have had some success, in which unknown values are integrated out (marginalized) over their prior probability of occurrence. We further describe the special considerations that need to be taken into account when building models against such a vast amount of uncertain and fast-arriving data. Our discussion centers on issues such as handling data distributions with high fluctuations, modeling the unknown, handling potentially conflicting information, and considering boundary conditions (i.e. the prices of the stocks when acquired and sold or the initial and final interest rates) following a path integral approach. We conclude the chapter by providing our perspective as to what the future holds.

We begin this chapter, in Sects. 2 and 3, by setting the stage and by discussing the complexities associated with building predictive models for financial data that are high in Volume and Variety. Section 4 reviews the concepts of bonds and interests rates, while Sect. 5 presents the Black-Scholes model for interest rates. In Sect. 6, we explore the Heath-Jarrow-Molton model for predicting the forward-value of a bond. Next, in Sect. 7, we turn our attention to this issue of Variety, and we discuss the use of social media and non-traditional data sources during model building. Finally, Sect. 8 concludes the chapter and presents our views on the way forward.

2 Business, Finance and Big Data

Our level of indebtedness is unprecedented in history. Whether we like it or not, the finance sector, in general, and the debt sector, in particular, has become paramount to business. In 1965, corporations in the United States of America (US) were earning 12.5 % of their revenues from the financial sector while 50 % of their revenues were coming from manufacturing. In 2007, just before the financial meltdown of 2008, this tendency was completely inverted with 35 % of US corporations' revenues earned from the financial sector, while only 12 % were earned from domestic manufacturing. As a matter of fact, the fraction of corporate earnings from the financial sector has grown more than 400 % over the last 60 years [1].

By all means, finance is big: big by the Volume, Velocity, and Variety of data involved, big by the corresponding amount of money involved (trillions of $), and big by its influence on our lives. Just to present an order of magnitude, on 13 November 2014, a normal trading day, 708,118,734 financial instruments were traded for a total value of $26,847,016,206 at the New York Stock Exchange (NYSE) of which, 641,044 financial instruments were traded with algorithmic programs [2]. (Note that a financial instrument may be defined as a trade-able asset of any kind; either cash, evidence of an ownership interest in an entity, or a contractual right to receive or deliver cash or another financial instrument. For each financial instrument, we keep track of its value as it evolves over time. The market data for a particular instrument would include the identifier of the instrument and where it was traded such as the ticker symbol and exchange code plus the latest bid and ask price and the time of the last trade. It may also include other information such as volume traded, bid and offer sizes, and static data about the financial instrument that may have come from a variety of sources. That is, these massive data streams are in essence time series data.)

It follows that making predictions and building trading models are central goals for financial institutions. For example, a number of researchers have studied the problem of forecasting the volatility of stock markets, through the use of neural networks, decision trees, cluster analysis, and so on [3]. In contrast to econometric approaches, the data-driven modeling approach used in many data mining algorithms makes few assumptions about data distributions or types of relationships. In this framework few (if any) parameters need to be estimated. Neither is there an assumed model form. Instead, the standard non-parametric approach proceeds by searching the data set to see whether any patterns are exhibited in that set. If the patterns found meet certain minimum requirements, then the pattern is recorded for further inspection. The usefulness of the methodology is judged by looking at new data to see whether these patterns also occur there. If so, we say that the data mining model is robust and has found a pattern that holds over time.

However, following a data-driven only approach, as discussed above, may not be ideal when aiming to construct models against big data in finance, in which the number of unknowns, due to the essential randomness, is high. Also, this train-then-test method does not work well for financial data streams that are susceptible to concept drift. To this end, the focus of this chapter is on building models against big data in finance, using a path integral approach. We primarily focus our attention on stocks, bonds, and interest rates from a big data perspective. Stochastic models for the stocks' prices and for the forward rates are introduced. From the knowledge of the probability distribution associated with the noise, it is possible to marginalize our uncertainty about the prices and the rates and to make useful predictions. The lack of knowledge may be leveraged through a framework rooted in the path integral formalism. We show that a thorough understanding of what we *don't know* is instrumental in such a process. In the next sections, we address stock prices, and we then extend our previous analysis to bonds.

3 Finance and Data Mining: Diving into the Unknown

Stock prices and interest rates are time series data that arrive in massive volumes, are fast changing and potentially infinite [3]. In the financial sector, researchers aim to create just-in-time models in order to find similar or regular patterns, to identify trends, to detect sudden concept drifts and to spot outliers, from such big data.

An important task is to find similar series, using either subsequence matching or whole sequence matching [4]. For example, *Selective MUSCLES* as introduced in [5], is an efficient and scalable method for on-line mining for co-evolving time sequences. In their method, they use subset selection and exponential forgetting in order to scale their system up. In addition, trend analysis is often used in order to both gain insights into the underlying forces that generate time series and to predict the future [6]. Here, four main types of analysis are of importance [3]. Firstly, we are interested in modeling long-term movements, e.g. the trend in the behavior of a stock or market over a long period of time. Secondly, there is the study of cyclical movements, which

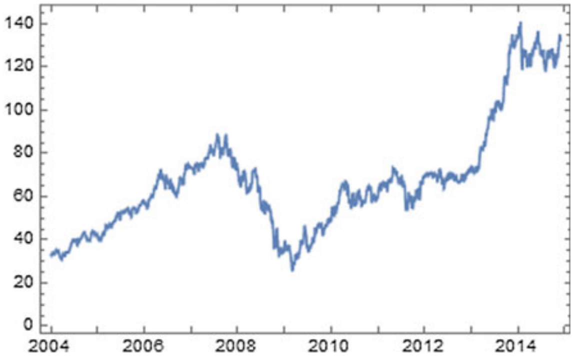

Fig. 1 Boeing stock movement over the last 10 years on the NYSE

refers to long-term oscillations that may or may not be periodic. Thirdly, seasonal drifts refer to variations that are typically calendar related. For example, there may be an increase in food prices traded out of season. In this case, the seasonal movements are typically very similar from year to year, and we are interested in utilizing this knowledge. The fourth type of movement refers to sporadic motions due to random or chance events, such as a volcanic eruption that disrupts air traffic or some unexpected socio-economic turmoil. These type of movements are also known as sudden concept drift, and the challenge here is to react fast, in order to update the models.

It is often said, in jest, that there are two certainties in life: death and taxes. Finance, on the other hand, is the kingdom of uncertainty, which makes trend analysis a challenge. If it would not be the case, risk-free and high-return investments, would be common place. As we all know, this is far from being the case. In order to obtain knowledge from this type of data stream, we often approach the problem by first making a certain number of hypotheses that could be validated subsequently from historical financial series. These hypotheses, once structured, constitute a model. A question which needs to be thoroughly considered is the following: What do we already know and what information may be utilized?

As an example, Fig. 1 shows the long term movement of the Boeing stocks on the New York Stock Exchange (NYSE) in terms of the value at the time of closure, from 1 January 2004 until 1 December 2014. In Fig. 2, we depict the behavior of the Baskem stocks on the NYSE over the same period of time. The figures show the difference in long term behavior between these two equities, with both experiencing a downturn in the 2008–2009 period.

We further know that stock prices and interest rates are volatile. There may be a function that characterizes such volatility, but its precise form is currently out of reach. We also know that the statistical properties associated with stock prices and interest rates are drifting. Such a concept drift could also be characterized by a function of unknown nature. Furthermore, the fact that stock prices and interest rates are intrinsically uncertain, points toward the existence of random fluctuations (noise). These fluctuations may be characterized by a Gaussian, Lévy, or truncated Lévy probability distribution according to the importance devolved to large fluctua-

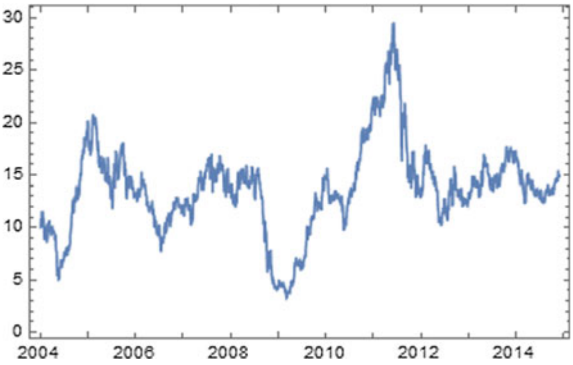

Fig. 2 Baskem stock movement over the last 10 years on the NYSE

tions [7, 8]. The multidimensional functional Gaussian distribution, for instance, is entirely characterized by its mean and covariance. The model relates all these disparate elements into a common and organic framework. As will be discussed below, in the case of stock prices and interest rates, the model is either a differential or a finite difference equation that relates the target variable (stock price or interest rate), the drift function, the volatility function and the stochastic random fluctuations. For instance, the stock prices may be described by the Black-Scholes model, while the interest rates are depicted by the Heath-Jarrow-Morton model, as explained in the following sections.

Nevertheless, the above-mentioned models are plagued with unknowns. The reasons are threefold. Firstly, as we pointed out, the nature of the functions associated to the drift and the volatility are unknown. Secondly, the boundary conditions are, in all likelihood, unknown. Recall that, by boundary conditions, we mean the price of the stock when acquired and sold or the initial and final interest rate. Finally, stocks and interest rates follow a specific financial trajectory in the sense that the time series associated with these financial instruments take precise values at every time t.

The Bayesian framework has been widely used to study such data. One of the main reasons why Bayesian methods have been so successful is their ability to incorporate information from different sources and also address complex estimation problems. Bayesian methods are based on the principle that probability is subjective, in that the degree of belief may be updated as new information, or data, are acquired [9]. Here, the beliefs based on the current knowledge is referred to as the prior probabilities and the posterior probability represents the updated beliefs. To this end, the Bayesian framework has been used for portfolio allocation [10], asset pricing models [11], and for volatility models [12].

That is, one of the best approaches for marginalizing the unknown functions, boundary conditions and financial trajectories is to be found in the Bayesian framework, as will be illustrated throughout this chapter. Here, unknown values are integrated out (marginalized) over their prior probability of occurrence. Consequently, a model associated with a financial instrument may be constructed as follow. Firstly, unknown functions are associated with the instrument's drift and volatility. The drift, the volatility, and the stochastic fluctuations are combined into a differential equation, which characterized the temporal evolution of the underlying time series. Then, the precise nature of the stochastic fluctuations is determined. For instance, the probability distribution associated with the fluctuation may be a multidimensional Gaussian, which means that it is entirely parameterized by its mean and its covariance. The differential equation acts as a constraint on the probability distribution associated with the noise. The constraint is imposed with a Dirac delta function, a generalized function, or distribution, which is zero everywhere, except at the origin and for which the integral over the entire integration domain is equal to one [13]. The unknown noise and financial trajectory are then integrated out or marginalized. That means that at each instant associated with the time series, the value of the financial instrument and of the corresponding stochastic fluctuation are integrated. If unknown, the boundary conditions must also be marginalized. These calculations allow performing predictions of statistical nature about financial instrument such as their expectation and dispersion.

As will be explained below, if the financial process unfolds as a fair game (or so-called Martingale), it is possible to express the drift as a function of the volatility. In the following, we will analyze the computational aspect of models related to stock prices and interest rates.

4 Finance: A Fair Game … Most of the Time

A bond is an instrument of debt, while a treasury bond is an instrument of debt with no risk of default. The money is lent in exchange of an interest over the capital, which is the cost for borrowing money. An important concept associated with bonds is the forward interest rate [1]. The forward interest rate, also called the forward rate, $f(t, \tau)$ is the agreed upon future interest rate, at time $t < \tau$, for an instantaneous loan at future time τ. It is typically calculated using a yield curve. For example, the yield on a three-month treasury bill, six months from now, represent a forward rate. The value of a bond is related to the forward interest by

$$B\left(T_i, T_f\right) \equiv e^{-\int_{T_i}^{T_f} d\tau\ f(t,\tau)}, \tag{1}$$

where $B\left(T_i, T_f\right)$ is the value of a bond at time T_i, maturing at time T_f. Bonds have a remarkable property that is shared by other financial instruments and which is called the fundamental theorem of finance. The fundamental theorem of finance states that financial processes follow a martingale [14]. A martingale is a model of a fair game in which the knowledge of past events never helps predict the mean of the future earnings. Mathematically, it may be formulated as follows:

$$E\left[B^{(k+1)}\left|B^{(1)}, B^{(2)}, \ldots, B^{(k)}\right.\right] = B^{(k)}, \tag{2}$$

At first sight, it seems that the martingale is a rather mild condition. However, as we shall see later, its importance is fundamental in constraining financial models. That is, martingales exclude the possibility of winning strategies based on game history, and thus they are a model of fair games. As will be seen in the next section, the martingale is a strong condition which allows determining the deterministic drift associated with a financial instrument [15]. For instance, in the case of the Black-Scholes (BS) model, the drift is entirely determined by the spot rate, irrespectively of the underlying data as a consequence of the martingale condition. This is something that would be difficult to conclude when following a purely data-driven approach. (Note that the spot rate refers to the price quoted for immediate settlement on a commodity, a security or a currency. The spot rate, also called the spot price, is based on the value of an asset at the moment of the quote.)

To this end, in the next section, we explain how to model evolving equities or stocks. This discussion presents a first step towards the task of modeling interest rates.

5 Black-Scholes Model and Path Integrals or How to Handle the Unknown for Stocks

Before addressing the bond and the forward rate, we consider a rather simpler process, namely the evolution of the price of a financial instrument representing a set of equities or stocks $\{S_i\}_{i=1}^N$. As stated above, this topic has received much attention in the area of time-series data mining [3]. Traditionally, the Black-Scholes model has been used to construct a model of the price evolution of N stocks with a stochastic process [16]:

$$\frac{dS_i(t)}{dt} = \alpha_i S_i(t) + \sigma_i S_i(t) R_i(t), \tag{3}$$

where $S_i(t)$ is the price of stock i at time t, α_i is the deterministic drift associated with stock i, σ_i is the deterministic volatility associated with stock i, while the Gaussian white noise $R(t)$ has a mean and a variance given by

$$\begin{gathered} E[R_i(t)] = 0, \\ E\left[R_i(t) R_j(t')\right] = \rho_{ij}\,\delta\left(t - t'\right), \\ T_i \leq t, t' \leq T_f. \end{gathered} \tag{4}$$

where ρ_{ij} is the estimated correlation in between the various stocks. We shall address the evaluation of the drift and the volatility later in this chapter. For the time being, we concentrate on the stocks per se, and we just mention here that the drift and the volatility may be estimated from historical time series data. The BS model is rather intuitive. That is, we know that the prices of stocks tend to drift; we know that the prices of equities are volatile and we know that the fluctuations associated with financial instruments are of a stochastic nature. The BS model is one of the simplest models that combine all these requirements.

Nevertheless, there are many unknowns associated with financial simulations. For instance, given a financial instrument, the initial and the final value (boundary conditions) of this instrument are unknown. As a matter of fact, this is true for any intermediate state of the instrument. All the intermediate states form a so-called financial trajectory. The only problem, so to say, is that the exact nature of this trajectory is entirely unknown. As it stands, the situation seems rather insoluble. Most of what we know is unknown, but the fact that we know what is unknown, shall prove itself to be crucial. The real question is how should we leverage the unknown? The best answer is that we should consider any possible evolution or path of the stocks. We are not allowed to discard any trajectory, because we do not have any information from which such an action could be justified. What is required is a method to weight the various trajectories in order to extract the expected behaviour of the underlying financial instrument. The weight of a given trajectory may be associated to its probability of occurrence. The white noise, as defined by the previous equation, has a Gaussian distribution. This implies that the probability distribution, as Bayesian prior, associated with a specific noise trajectory is given by:

$$\mathcal{D}R\ \Pr[R] = \frac{\mathcal{D}R\ e^{S[R]}}{Z}, \tag{5}$$

where $S[R]$ is the time integral of the Lagrangian $\mathcal{L}[R]$:

$$S[R] = \int_{T_i}^{T_f} dt\ \mathcal{L}[R]. \tag{6}$$

The latter is a functional that assigns probabilities for the occurrence of the various realisations of the noise and is defined by the quadratic function:

$$\mathcal{L}[R] = -\frac{1}{2}\sum_{i,j=1}^{N} R_i\ \rho_{ij}^{-1}\ R_j. \tag{7}$$

Here, ρ_{ij} is the deterministic factor associated with the correlation in between the various stocks. $\mathcal{D}R$ is the path integral measure, that is, the integration or Bayesian marginalization over all unknown intermediate states along every possible trajectory

$$\int \mathcal{D}R = \prod_{t=T_i}^{T_f}\prod_{i=1}^{N}\int_{-\infty}^{\infty} dR_i\,(t) \tag{8}$$

and Z is a normalization factor known as the partition function.

The integral over every possible state or trajectory is known as a path integral [13]. The probability distribution associated with Eq. (7) is clearly Gaussian, although it is somewhat different from the distribution we are familiar within the sense that it does not involve a variable but a function: in occurrence the stochastic noise. This is why we don't refer to functions, but to functionals (function of a function). From now on, we shall consider the logarithmic of the stock price

$$z_i \equiv \ln S_i. \tag{9}$$

As stated earlier, the financial trajectories associated with the stocks are governed by Eq. (4). The Bayesian probability associated with a specific trajectory is given by

$$\mathcal{D}z\mathcal{D}R\ \Pr[z, R] = \frac{\mathcal{D}z\mathcal{D}R \prod_{t=0}^{T} \prod_{i=1}^{N} \delta\left(\frac{\partial z_i(t)}{\partial t} + \alpha_i - \frac{1}{2}\rho_{ii}\sigma_i^2 + \sigma_i R_i(t)\right) e^{S[R]}}{Z}, \tag{10}$$

where the Dirac delta distribution ensures that the stochastic equation of motion associated with the equities, here the BS model, is always satisfied. The partition function Z, or normalization factor, is obtained by integrating the probability distribution over all possible values of the stocks and of the random variables. The mathematical expectation (mean) of any financial instrument $\mathcal{O}$ is obtained by weighing each occurrence of the financial instrument by its corresponding probability

$$\int \mathcal{D}z\mathcal{D}R\ \mathcal{O}[z, R] \Pr[z, R]. \tag{11}$$

Because the probability distribution associated with the noise is quadratic, we may easily integrate or marginalize the noise out of the equation and obtain a closed-form expression that depends only on the stocks. It follows that closed-form expressions play an important role in the big data era. In finance, these expressions have been successfully used for the pricing of especially exotic derivative products [17]. Indeed, the integration measure typically involves thousands of dimensions. Consequently, for the sake of computational stability and efficiency, numerical evaluation should be strictly restricted to those dimensions that could not be treated analytically. We finally obtained for the expectation value of a given functional of an underlying commodity

$$E[\mathcal{O}[z]] = \frac{1}{Z_{BS}} \int \mathcal{D}z\ e^{S_{BS}[z]}\ \mathcal{O}[z], \tag{12}$$

where the action, the Lagrangian, the partition function and the integration measure are given respectively by [18]

$$S_{BS}\left[z\right] = \int\limits_{T_i}^{T_f} dt\ \mathcal{L}_{BS}\left[z\right],$$

$$\mathcal{L}_{BS}\left[z\right] = -\frac{1}{2}\sum_{i,j=1}^{N}\left[\frac{\frac{\partial z_i(t)}{\partial t}+\alpha_i-\frac{1}{2}\rho_{ii}\sigma_i^2}{\sigma_i}\right]\rho_{ij}^{-1}\left[\frac{\frac{\partial z_j(t)}{\partial t}+\alpha_j-\frac{1}{2}\rho_{jj}\sigma_j^2}{\sigma_j}\right] \tag{13}$$

and

$$Z_{BS} = \int \mathcal{D}z\ e^{S_{BS}[z]},$$

$$\int \mathcal{D}z = \prod_{t=T_i}^{T_f}\prod_{i=1}^{N}\int\limits_{-\infty}^{\infty} dz_i\left(t\right).$$

For instance, the functional $\mathcal{O}\left[\cdot\right]$ in question may be an option. An option is a contract that gives the buyer the right, but not the obligation, to buy or sell an underlying asset or instrument at a specified strike price P on or before a specified date. The seller has the corresponding obligation to fulfill the transaction if the buyer (owner) exercises the option. The buyer pays a premium to the seller for this right. For example, options are often used by electricity generators and retailers to protect from price or cost volatility [19]. One type of option that is used in such a setting is the so-called flexibility-of-delivery option, which permits the contract holder to receive any amount of power within a certain range for defined time periods.

Options valuation is a topic of ongoing research in academic and practical finance, due to its importance in financial markets, their complexity and the large Volume of options being exercised. Options contracts have been known for many centuries, however both trading activity and academic interest increased when, as from 1973, options were issued with standardized terms and traded through a guaranteed clearing house at the Chicago Board Options Exchange [16]. Today many options are created in a standardized form and traded through clearing houses on regulated options exchanges, while other over-the-counter options are written as bilateral, customized contracts between a single buyer and seller, one or both of which may be a dealer or market-maker. Options are part of a larger class of financial instruments known as derivatives.

There are a number of ways to model an option. For example, if an investor acquires an Asian option, then the pay-off function depends on the average price of the stock during a given time interval:

$$\mathcal{O}_A\left[z\right] = \max\left(P,\ \frac{1}{\Delta t}\int\limits_{\Delta t} dt\ g\left[z\left(t\right)\right]\right), \tag{14}$$

where $g\left[\cdot\right]$ is an agreed upon functional.

We can still further improve our model. We know (prior information), from the fundamental theorem of finance [20], that stocks follow a martingale which was defined earlier in Eq. (2–3). That is, it is known that knowledge of past events cannot help us to predict hte mean of future yields. If we compute the mathematical expectation associated with the martingale with Eq. (15) we obtain $\alpha_i = r$. This confirms that, as stated earlier, the drift is entirely determined by the spot rate and is not an independent parameter of the model as might have been expected earlier.

Still, the integration over all possible prices of the stock is not a trivial operation. The value of a stock is a time series, which is updated at regular interval, typically every few seconds. Let us assume that we want to calculate the expectation value of a stock over a period of one week and that the price of the stock is updated every 15 s with a typical trading session lasting from 9:30 until 16:00 local time. Thus the integration measure consists of 7,800 dimensions! This is clearly a big data problem, which is reminiscent of the curse of dimensionality. Nonetheless, such an integral may be calculated efficiently with a Monte Carlo approach, known as the Metropolis-Hasting (MH) algorithm [21]. Instead of systematically integrating over the whole integration domain, the latter is explored with a Markovian process which randomly samples the realizations of the stock. Given a value of the stock $z^{(k)}$, a new value is randomly generated according to

$$z^{(k+1)} = z^{(k)} + R, \tag{15}$$

where R is a Gaussian white noise. The new occurrence of the stock is accepted (or rejected) with probability

$$A\left(z^{(k)} \to z^{(k+1)}\right) = \min\left(1, \exp\left(S\left[z^{(k)}\right] - S\left[z^{(k+1)}\right]\right)\right), \tag{16}$$

where the action S was defined earlier in Eq. (16). This means that the new value is always accepted if its probability of occurrence is higher than the previous one. However, it is nevertheless accepted with a probability that is otherwise equal to $\exp\left(S\left[z^{(k)}\right] - S\left[z^{(k+1)}\right]\right)$. The expectation of a function of the stock is then obtained as the average of this function over the sampled values of the stock

$$E\left[\mathcal{O}\left[z\right]\right] \approx \frac{1}{\left(k_{\max} - k_{\min}\right)} \sum_{k} \mathcal{O}\left[z^{(k)}\right]. \tag{17}$$

The MH algorithm allows for a more efficient sampling of the integration domain and prevents from integrating over trajectories that have a negligible probability of occurrence. These trajectories tend, generally speaking, to introduce a detrimental numerical noise [21]. In the next section, we extend our previous analysis to the modeling of forward interest rates and bonds based on the well-known Heath-Jarrow-Morton model.

6 Heath-Jarrow-Morton Model and Path Integrals or How to Leverage Our Ignorance About Interest Rates

The case of a bond, and of the underlying forward rate, is slightly more complicated than the case of a stock. As we saw earlier, the value of a bond is determined by the forward interest rate (cf. Eq. (1)) which is unknown. The forward interest rate depends on both the present time and the future time. Forward interest rates are typically modeled with a stochastic process known as the Heath-Jarrow-Morton (HJM) model [22]. The HJM model is very similar in nature to the BS model (Eq. 4) except that the drift and the volatility are not constant but depend on the current (calendar) time t and on the future time τ:

$$\alpha \Rightarrow \alpha\,(t,\tau) \tag{18}$$

$$\sigma \Rightarrow \sigma\,(t,\tau)\,.$$

It follows that the forward rate is governed by the following stochastic equation:

$$\frac{\partial f\,(t,\tau)}{\partial t} = \alpha\,(t,\tau) + \sigma\,(t,\tau)\;R\,(t)\,, \tag{19}$$

where the white noise $R\,(t)$ was defined earlier [23]. Following the same approach as for the equities (or stocks), the Bayesian probability associated with a specific trajectory of the forward rate is equal to

$$\mathcal{D}f\mathcal{D}R\;\Pr[f,R] = \frac{\mathcal{D}f\mathcal{D}R\prod_{(t,\tau)\in\mathcal{T}}\delta\left(\frac{\partial f(t,\tau)}{\partial t} - \alpha\,(t,\tau) - \sigma\,(t,\tau)\;R\,(t)\right)e^{S[R]}}{Z}, \tag{20}$$

where the temporal domain $\mathcal{T}$ is defined as

$$\mathcal{T} \Rightarrow t \in \left[T_i, T_f\right] \quad \cap \quad \tau \in [t, t+T_H]\,, \tag{21}$$

where T_H is the investment horizon: the time, during which an investment may be performed. If the white noise is integrated out, one obtains, for the mathematical expectation, a closed-form expression similar to Eqs. (15) and (16). As in the case of the BS model, one may apply the fundamental theorem of finance and demonstrate that the drift is not an independent quantity but is related to the volatility by [23]

$$\alpha\,(t,\tau) = \sigma\,(t,\tau)\int_t^\tau d\tau'\,\sigma\left(t,\tau'\right). \tag{22}$$

It then follows that the path integration may be performed with the MH algorithm. This allows for the computation of the expected value of a bond and its standard deviation, together with other quantities of interest. In other words, this calculation enables one to determine whether a specific investment is worthwhile, in addition to evaluating the concomitant risk level and the level of uncertainty.

Note that historic investment data may further be extracted, for instance, from historical yield curves. Consequently, financial institutions may choose to combine model-driven and data-driven approaches. A data-driven approach is particularly suitable when handling late arriving data [24], such as those which result from a manipulation of the interest rates. In follows that the sheer Volume of data requires greater sophistication of statistical techniques in order to obtain accurate results. In particular, recent research has shown that the number of false correlations increases as the data Volume and dimensionality increases [9]. The reader should further notice that the state-of-the art algorithms based on economic theory typically point to long-term investments opportunities as based on trends in historical data. The task to produce efficient results supporting a short-term investment strategy still poses a challenge for current predictive models [8]. Thus, a number of research challenges remain, in this era where financial institutions are increasingly embracing big data analytics.

7 A Word About Variety

In the above-mentioned discussions, we focused our attention on financial data that is high in Volume and Velocity. However, in order to capitalize on the big data opportunity, enterprises should also embrace Variety, that is different types of data from a wide range of fields, including documents, e-mail, web pages, social media forums data, smart devices data, and sensor data, amongst others. This Variety characteristic associated with big data presents rich information for knowledge discovery.

Such Variety may aid the learning processes from different observation angles, and allows exploring correlation across domains and fields. The financial sector is especially susceptible to changes due to socio-economic factors. It then follows that the use of social media data may provide role-players with a competitive advance. For example, recent studies have shown that the evaluation of large-scale Twitter feeds may be used to accurately predict stock market indicators for markets such as Dow Jones, NASDAQ, and S&P 500 [25, 26]. Specifically, the results in [25] indicate that the accuracy of Dow Jones Industrial Average (DJIA) closing predictions can be significantly improved by the inclusion of specific public mood dimensions.

As another example, we turn our attention to the case of Smart Cities, which has increasingly become of importance in the financial sector [27]. Energy usage costs accounts for approximately 19 % of total expenditures for a typical building in the US [28]. In the European Union (EU), buildings account for approximately 40 % of final energy consumption in 2008 [29]. To this end, both governments as well as the owners of commercial buildings have moved to time-of-use pricing and are exploring ways to balance demand and response signals [27]. Smart energy consumption models, however, heavily relies on accurate short-term load predictions.

In order to generate accurate energy load forecasts, a number of factors from a variety fields need to be taken into account. For instance, the building's routine schedules such as the office hours and daily occupancy information present useful knowledge on how the building is occupied. This knowledge thus provides basic energy usage patterns. Also, the weather condition throughout a day (e.g. the hourly temperatures) is strongly correlated to a building's energy consumption curve [30]. Another factor is the pricing fluctuations, which are further complicated by uncertain energy price policies and uncertainty about fossil fuel prices. Real-time pricing quotes from power grid utilities can force a building to dramatically change its energy consumption behaviors. In addition, related social events (e.g. local sport activities and political news) can significantly shift the energy usage and pricing patterns. Also, recall from above, that electricity supplies and consumers increasingly make use of options in order to optimize their financial gains [19].

Another important data source for accurate energy load prediction is the building's daily operations. For example, actions being taken to reshape energy usage curves have a significant impact on the building's short term energy load. Consider a building with an energy storage unit. After having initial short-term predictions, the building managers often aim to reduce buildings' energy usage during peak energy demand periods, which often impost high-energy usage rates for consumers and large load demand for utility grids. In such scenarios, energy storages such as an ice bank, chiller, boiler, and battery, etc., are often used. An ice bank, for instance, is typically used to build ice in summer when the electricity is cheap, and the ice is then used to cool the building, rather than using electricity, when the price of the energy load is high. In order to have accurate short-term load forecasts, features or sensors related to such reschedulable energy-intensive units have to be taken into account. In short, integrating difference sources of data into the learning will allow the mining methods to figure out key components which impact the energy consumption. In particular, it enables the learning algorithm to explore the multiple interconnected data so that important data or attributes (factors) are not excluded [30].

8 Conclusions

This chapter focused on recent advances in data mining in the financial sector, within the context of big data. The Volume and Verocity of such massive datasets, as well as the large number of unknowns and volatility, led to the use of model-driven approaches. To this end, this review mainly centred around model-based approaches currently used when analyzing stocks, bonds, and interest rates. We also turned our attention to the issue of Variety, and briefly reviewed current advances in terms of using social media data to augment and strengthen current predictions.

When the amount of data is relatively small or when the framework in which they evolve is either well understood or deterministic, it is legitimate to primarily use our prior knowledge about the data and not to pay too much attention about what we *don't know*. In such a setting, the use of data-driven modeling approaches, following the standard training, testing, and validation model construction process holds value. The situation is quite different in a big data framework in which our prior knowledge about the data is often rather marginal and has to be supplemented with an assumed model form. Another issue that needs mention is that we often also need to handle late arriving data, i.e. there is a need to incorporate retroactive data as they arrive. Many of these models must be stochastic in order to make allowance for the random nature of the underlying data. As demonstrated, what should be determined carefully is what we *don't know* and how such drawbacks may be marginalized. The path integral approach provides an efficient and coherent framework to marginalize the unknown. This is possible since it considers every possibility and weighs them according to their probability of occurrence, which may be determined from the concomitant model. Despite the fact that the amount of data is big, it does not mean that closed-form expressions are outdated. As a matter of fact, they are more essential than ever, especially in order to reduce the massive dimensionality associated with the problem.

We further believe that the current surge in the area of data stream mining [31] may hold the key to build accurate, just-in-time models to be used by the financial sector. That is, adaptive learning algorithms that build incremental models from asynchronous streams have much application in the financial sector [24]. Specifically, techniques for building dynamic probabilistic models for streaming data [32] have shown to produce high quality results against data that both contain temporal trends and are susceptible to noise and unknowns. Indeed, these types of models may yet prove to be ideal for exploring financial data.

References

1. Baaquie, B.E.: Interest Rates and Coupon Bonds in Quantum Finance. Cambridge University Press, Cambridge (2010)
2. NYSE Market Data, Data Products, Product Summaries. http://www.nyxdata.com/Data-Products/Product-Summaries. Accessed 14 Nov 2014
3. Han, J., Kamber, M.: Data Mining Concepts and Techniques, 2nd edn. Morgan Kauffman (2008)
4. Shasha, D., Zhu, Y.: High Performance Discovery In Time Series: Techniques and Case Studies. Springer (2004)
5. Yi, B.K., Sidiropoulos, N., Johnson, T., Jagadish, H.V., Faloutsos, C., Biliris, A.: Online data mining for co-evolving time sequences. In: Proceedings of 2000 International Conference on Data Engineering (ICDE 2000), pp. 13–22. San Diego, CA (2000)
6. Shumway, R.H., Stoffer, D.S.: Time Series Analysis and Its Applications. Springer (2005)
7. Oshaug, C.A.J.: Lvy Processes and Path Integral Methods with Applications in the Energy Markets. Norwegian University of Science and Technology, Thesis (2011)
8. Paquet, E., Viktor, H.L., Guo, H.: Learning in the presence of large fluctuations: a study of aggregation and correlation, new frontiers in mining complex patterns. LNCS **7765**, 49–63 (2013)
9. Rachev, S.T., et al.: Bayesian Methods in Finance. Wiley, Hoboken (2008)
10. McNeil., A.J., Wendin, J.P.: Bayesian inference for generalized linear mixed models of portfolio credit risk. J. Empirical Finan. **14**(2), 131–147 (2007)

11. Garlappi, L., Uppal, R., Wang, T.: Portfolio selection with parameter and model uncertainty: a multi-prior approach. Rev. Finan. Stud. Oxford J. **20**(1), 41–81 (2007)
12. Jacquier, E., Polson, N.G., Rossi, P.E.: Bayesian analysis of stochastic volatility models. J. Busin. Econ. Statis. **20**(1), 69–87 (2002)
13. Masujima, M.: Path Integral Quantization and Stochastic Quantization. Springer, Berlin (2009)
14. Campbell, J.Y., Low, A.S., Mackinlay, A.C.: The Econometric of Financial Markets. Princeton University Press, Princeton (1997)
15. Lloyds fined 218m over Libor rate rigging scandal. http://www.bbc.com/news/business-28528349?print=true. Accessed 21 Dec 2014
16. Merton, R.C.: Continuous Time Finance. Blackwell Publishing, Oxford (1990)
17. Lemmens, D., Wouters, M., Tempere, J.: Path integral approach to closed-form option pricing formulas with applications to stochastic volatility and interest rate models. Phys. Rev. E **78**, 016101-1–016101-8 (2008)
18. Baaquie, B.E.: Quantum Finance: Path Integrals and Hamiltonians for Options and Interest Rates. Cambridge University Press (2004)
19. Dalakouras, G.V., Kwon, R.H., Pardalos, P.M.: Semidefinite programming approaches for bounding Asian option prices. In: Computational Methods in Financial, Engineering, pp. 103–116 (2008)
20. Devreese, J.P.A., Lemmens, M., Tempere, J.: Path integral approach to asian options in the BlackâĂŞcholes model. Physica A: Statis. Mech. Appl. **384**, 780–788 (2010)
21. Binder, K., Heermann, D.W.: Monte Carlo Simulation in Statistical Physics: An Introduction. Springer, Berlin (2010)
22. Heath, D., Jarrow, R., Morton, A.: Bond pricing and the term structure of interest rates: a new methodology for contingent claim valuation. Econometrica **60**, 77–105 (1992)
23. Baaquie, B.E.: Financial modeling and quantum mathematics. Comput. Math. Appl. **65**, 1665–1673 (2013)
24. Krempl, G., Z̃liobaite, I., Brzezinski, D., Hullermeier, E., Last, M., Lemaire, V., Stefanowski, J.: Open challenges for data stream mining research. ACM SIGKDD Explor. Newslett. **16**(1), 1–10 (2014)
25. Bollen, J., Mao, H.: Twitter mood as a stock market predictor. Computer **44**(10), 91–94 (2011)
26. Zhang, X., Fuehres, H., Gloor, P.: Predicting stock market indicators through twitter ("I hope it is not as bad as I fear"). Procedia: Social Behav. Sci. **26**, 55–62 (2011)
27. Ferrer, J.N., Olivero, S., Medarova-Bergstorm, K., Rizos, V.: Financing models for smart cities. November 2013. http://eu-smartcities.eu/sites/all/files/GuidelineFinancingModelsforsmartcities-january.pdf. Accessed 15 Dec 2014
28. NationalGridUS, Managing Energy Costs in Office Buildings. https://www.nationalgridus.com/non_html/shared_energyeff_office.pdf. Accessed 15 Dec 2014
29. European PPP Expertise Centre (EPEC): Guidance on energy efficiency in public buildings. http://www.eib.org/epec/resources/epec_guidance_ee_public_buildings_en.pdf. Accessed 15 Dec 2014
30. Guo, H.: Modeling short-term energy load with continuous conditional random fields. In: ECML/PKDD 2013 machine learning and knowledge discovery in databases. In: Lecture Notes in Computer Science, vol. 8188, pp. 433–448 (2013)
31. Bosnić, Z., Demsar, J., Kes̃pret, G., Rodrigues, P.P., Gama, J., Kononenko, I: Enhancing data stream predictions with reliability estimators and explanation. Eng. Appl. Artif. Intell. **34**, 174-188 (2014)
32. Kanagal, B., Deshpande, A.: Online filtering, smoothing and probabilistic modeling of streaming data. In: IEEE 24th International Conference on Data Engineering, 2008, ICDE 2008, pp. 1160–1169 (2008)