European Journal of Operational Research 267 (2018) 200–211

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Decision Support

Direct data-based decision making under uncertainty

Bogdan Grechuk[a,*], Michael Zabarankin[b]

[a] Department of Mathematics, University of Leicester, LE1 7RH, UK
[b] Department of Mathematical Sciences, Stevens Institute of Technology, Hoboken, NJ 07030, USA

ARTICLE INFO

Article history:
Received 6 January 2017
Accepted 11 November 2017
Available online 21 November 2017

Keywords:
Time series
Decision making under uncertainty
Mean-variance analysis
Portfolio optimization
Utility theory

ABSTRACT

In a typical one-period decision making model under uncertainty, unknown consequences are modeled as random variables. However, accurately estimating probability distributions of the involved random variables from historical data is rarely possible. As a result, decisions made may be suboptimal or even unacceptable in the future. Also, an agent may not view data occurred at different time moments, e.g. yesterday and one year ago, as equally probable. The agent may apply a so-called “time” profile (weights) to historical data. To address these issues, an axiomatic framework for decision making based directly on historical time series is presented. It is used for constructing data-based analogues of mean-variance and maxmin utility approaches to optimal portfolio selection.

1. Introduction

A typical process of decision making under uncertainty is as follows:

$$\boxed{\text{data}} \to \boxed{\text{uncertainty modeling}} \to \to \boxed{\text{risk preference modeling}} \to \boxed{\text{choice/decision}} \tag{1}$$

Let $\mathcal{X}$ be a set of available (feasible) actions. Scheme (1) can be formally stated as: (i) modeling unknown consequences of every action $X \in \mathcal{X}$ as a random variable (r.v.) $R(X)$, (ii) establishing a numerical representation $U' : \mathcal{R} \to \mathbb{R}$ for agent's preference relation, defined on a space $\mathcal{R}$ of all r.v.'s and (iii) finding best action by maximizing U' with respect to $X \in \mathcal{X}$:

$$\max_{X \in \mathcal{X}} U'(R(X)). \tag{2}$$

What an agent has readily available is only historical/experimental data and his/her preferences towards risk and reward. *The rest is statistical inference from the data about corresponding uncertain outcomes based on various assumptions, which largely depend on the nature of data.* For example, measurements of the length of some object can be reliably assumed to be realizations of independent and identically distributed (i.i.d.) r.v.'s—timing of those measurements can be safely ignored. By the central limit theorem (CLT), the average of a large number of i.i.d. r.v.'s is approximately normally distributed, and consequently, confidence intervals for the true length can be readily obtained.

Merton's well-known model (Merton, 1969) of a financial market assumes that asset prices are stochastic processes with *stationary* and *independent* increments *without jumps*—the only such a stochastic process is a Brownian motion with drift. Cox and Ross (1976) argued that asset prices are, in fact, not continuous processes—they may have jumps. Stochastic processes with stationary and independent increments (and with discontinuous sample paths in general), are called Lévy processes (Sato, 1999) and nowadays are widely-used in modeling of financial markets (Kou, 2002; Madan & Seneta, 1990; Merton, 1976). However, it is commonly acknowledged that

(a) The empirical distributions of rates of return of financial assets are typically *non-symmetric* with left tails being much heavier than right tails (Sheikh & Qiao, 2010).
(b) Increments of actual price processes are *not stationary*, and consequently, Lévy processes cannot be calibrated with real data (Madan, 2010).[1]
(c) “Periods of lower returns are systematically followed by compensating periods of higher returns” (Siegel, 2007) (“mean reversion” phenomenon)—evidence that price increments are *not independent*.

In fact, the above issues with stochastic processes can be “fixed” by time-series models. For example, autoregressive models AR(p)

* Corresponding author.
E-mail addresses: bg83@leicester.ac.uk, grechukbogdan@yandex.ru (B. Grechuk), mzabaran@stevens.edu (M. Zabarankin).

[1] Sato processes (Sato, 1991), whose increments are independent but not necessarily stationary, can be used instead.

https://doi.org/10.1016/j.ejor.2017.11.021
0377-2217/

assume that the current value of asset's rate of return depends on p previous ones, moving-average models MA(q) involve last q values of a stochastic error, autoregressive-moving-average models ARMA(p, q) generalize AR(p) and MA(q), whereas ARIMA models generalize ARMA(p, q), and are suitable to describe a wide range of non-stationary processes (Brockwell & Davis, 2016). However, any time-series model is merely another approximation of the historical data and its parameters are subject to estimation errors.

The discrepancy between a real-life phenomenon and its model is called *model error*—in contrast to approximation error, which can be resolved by simply increasing the sample size, the model error implies that *an increase of observations of asset rates does not directly translate into the accuracy/precision in estimation of the probability distributions of the rates.* There are various existing approaches that address model uncertainty. For example, bootstrapping (Bradley & Tibshirani, 1994) generates different scenarios for the variable of interest from a given time series, robust optimization (Bertsimas, Brown, & Caramanis, 2011) assumes that probabilities in question belong to certain intervals, whereas dual characterization of risk and deviation measures (Artzner, Delbaen, Eber, & Heath, 1999; Rockafellar, Uryasev, & Zabarankin, 2006a) relies on risk envelopes, which can be viewed as sets of distortions of an underlying probability measure, see Lesnevski, Nelson, and Staum (2007). Notably, Pflug, Pichler, and Wozabal (2012) showed that the naive $1/n$ investment strategy could be optimal in portfolio selection when model uncertainty is high. Savage (1972) suggested to study decisions as functions from some state space Ω to a set of outcomes $\mathcal{Y} \subset \mathbb{R}$, which are now known as *Savage acts*. This approach involves no probability measure on Ω—a critical feature that gave rise to various Savage-act versions of the expected utility theory (EUT) (Casadesus-Masanell, Klibanoff, & Ozdenoren, 2000; Gul & Pesendorfer, 2014). For example, Gilboa and Schmeidler (1989) proposed to study preference relations over acts, i.e., "functions from states of nature into finite-support distributions over a set of deterministic outcomes". In this case, the agent ends up with the same optimization problem (2), where R is a functional from $\mathcal{X}$ to the set $\mathcal{A}$ of all acts, and $U' : \mathcal{A} \to \mathbb{R}$ is a numerical representation of Gilboa and Schmeidler's preference relation. Of course, the list of existing approaches goes far beyond these examples, see e.g. Ben-Tal, Ghaoui, and Nemirovski (2009), Calafiore (2007), Chan, Karceski, and Lakonishok (1999), and Wozabal (2012) for alternative approaches and Marinacci (2015) and Gilboa and Marinacci (2016) for recent surveys.

In fact, accurately modeling of outcomes of real-life actions in the context of any of these theories is difficult. For example, modeling of financial portfolio returns in terms of Gilboa–Schmeidler acts (Gilboa & Schmeidler, 1989) includes forecasting of a set of finite-support distributions, and therefore, could, in fact, be harder than that in terms of r.v.'s. The main problem with uncertainty modeling is that, contemplating a choice among several alternatives, an agent ponders what alternative he/she would be most benefited from *in the future*, while the only available information is often the data representing historical performances of those alternatives *in the past*.

In view of failure of common statistical assumptions in application to a stock market (Madan, 2010; Sheikh & Qiao, 2010; Siegel, 2007) and in view of sensitivity of optimal decisions (portfolios) to errors in estimation of probability distributions of financial assets (Grechuk & Zabarankin, 2017; Kondor, Pafka, & Nagy, 2007), this work aims to identify intertemporal principles for comparing historical time series of asset rates of return and to develop an axiomatic framework for a rational decision making in portfolio theory on the space of historical time series. For example, an agent may postulate that if A always outperformed B in the past, then $A \succ B$, even though better past performance does not guarantee better future performance.

The idea of making decisions based directly on historical data is not new,[2] but it has received relatively little attention in economic and financial literature. Gilboa and Schmeidler (1995,2001) introduced a case-based decision theory, which makes decisions based on past experience in *similar* situations.[3] In a financial market setting, this theory would identify the moment in the past when the market behavior was most similar to the current one and would prescribe to invest all money into the financial asset which had the highest rate of return in that "similar" situation. However, it is not clear what "similarity measure" to use, and the resulting investment strategy may contradict the diversification principle. There are other objections for the use of direct data-based decision making in portfolio selection:

(i) Information such as recent market trends and news about particular companies may provide valuable insights for selecting a financial portfolio.
(ii) The future may have little in common with the past, for instance, due to unique events such as BREXIT.
(iii) New financial assets lack historical data, but it is unlikely that agents would view stocks, say, of a new bank and a startup IT company similarly.

However, incorporating news and other non-quantitative information, e.g. a recent hire of a highly regarded CEO, into a mathematical model requires human participation and is, therefore, expensive and slow. In contrast, calibrating stochastic models based only on historical data can be fully automated and performed in milliseconds, which is particularly valuable for high-frequency trading. Thus, *if* the choice of optimal portfolio is based on some uncertainty modeling, which in turn uses historical data only, *then* the uncertainty modeling stage could be omitted, and decisions could be made based on data *directly*.

The contribution and organization of this work are as follows. Section 2 introduces the notion of time profile and discusses numerical representation of time series. Section 3 introduces intertemporal principles of rational choice. Section 4 reinterprets the mean-variance and maxmin utility analyses in the context of direct data-based decision making. Section 5 concludes the work. Appendix A contains proofs of key results in Section 3 and Appendix B provides an axiomatic foundation for a data-based analogue of the EUT.

2. Time profiles and numerical representation of time series

Let $\mathcal{T} = \{s_1, \ldots, s_T\}$ be a finite set of discrete time moments $s_1 < \cdots < s_T$ in the past, and let $x_1, \ldots, x_T$ be corresponding rates of return of some financial asset. Since $x_1, \ldots, x_T$ encode a *time structure* and are not realizations of i.i.d. r.v.'s, the agent would unlikely view $x_1, \ldots, x_T$ as equally valuable data and may assign them corresponding weights $q_1, \ldots, q_T$ of historical data "*depreciation*" to be collectively referred to as *time profile Q*. For example, the agent may postulate that ratio q_t/q_{t+1} is a constant $q \in (0, 1]$ independent of t, which implies that

$$q_t = q_T q^{T-t}, \quad t = 1, \ldots, T. \tag{3}$$

Alternatively, $q_1, \ldots, q_T$ can be chosen to be proportional to the (normalized) autocorrelation profile of the asset—if for some time

[2] In psychology, Gorban, Rossiyev, and Dorrer (1995) proposed a neural-network-based approach for making recommendations based on the questionnaire data *directly* that avoids intermediate stage of patient descriptions in terms of "measurements of an individuality".

[3] Gilboa and Schmeidler (1995) argued that in a search for a nanny for their child, a couple faces a lot of uncertainty about how each candidate would perform and can hardly define "states of the world" that would adequately model the situation not to mention accurately forecasting of probabilities of each state.