Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 36 (2009) 2819–2830
Expert Systems with Applications
www.elsevier.com/locate/eswa

Flexible least squares for temporal data mining and statistical arbitrage

Giovanni Montana [a,*], Kostas Triantafyllopoulos [b], Theodoros Tsagaris [a,1]

[a] Department of Mathematics, Statistics Section, Imperial College London, London SW7 2AZ, UK
[b] Department of Probability and Statistics, University of Sheffield, Sheffield S3 7RH, UK

Abstract

A number of recent emerging applications call for studying data streams, potentially infinite flows of information updated in real-time. When multiple co-evolving data streams are observed, an important task is to determine how these streams depend on each other, accounting for dynamic dependence patterns without imposing any restrictive probabilistic law governing this dependence. In this paper we argue that flexible least squares (FLS), a penalized version of ordinary least squares that accommodates for time-varying regression coefficients, can be deployed successfully in this context. Our motivating application is statistical arbitrage, an investment strategy that exploits patterns detected in financial data streams. We demonstrate that FLS is algebraically equivalent to the well-known Kalman filter equations, and take advantage of this equivalence to gain a better understanding of FLS and suggest a more efficient algorithm. Promising experimental results obtained from a FLS-based algorithmic trading system for the S&P 500 Futures Index are reported.

Keywords: Temporal data mining; Flexible least squares; Time-varying regression; Algorithmic trading system; Statistical arbitrage

1. Introduction

Temporal data mining is a fast-developing area concerned with processing and analyzing high-volume, high-speed data streams. A common example of data stream is a time series, a collection of univariate or multivariate measurements indexed by time. Furthermore, each record in a data stream may have a complex structure involving both continuous and discrete measurements collected in sequential order. There are several application areas in which temporal data mining tools are being increasingly used, including finance, sensor networking, security, disaster management, e-commerce and many others. In the financial arena, data streams are being monitored and explored for many different purposes such as algorithmic trading, smart order routing, real-time compliance, and fraud detection. At the core of all such applications lies the common need to make time-aware, instant, intelligent decisions that exploit, in one way or another, patterns detected in the data.

In the last decade we have seen an increasing trend by investment banks, hedge funds, and proprietary trading boutiques to systematize the trading of a variety of financial instruments. These companies resort to sophisticated trading platforms based on predictive models to transact market orders that serve specific speculative investment strategies.

Algorithmic trading, otherwise known as automated or systematic trading, refers to the use of expert systems that enter trading orders without any user intervention; these systems decide on all aspects of the order such as the timing, price, and its final quantity. They effectively implement pattern recognition methods in order to detect and exploit market inefficiencies for speculative purposes. Moreover, automated trading systems can slice a large trade automatically into several smaller trades in order to hide its impact on the market (a technique called *iceberging*) and lower

* Corresponding author.
E-mail address: g.montana@imperial.ac.uk (G. Montana).
[1] The author is also affiliated with BlueCrest Capital Management. The views presented here reflect solely the author's opinion.

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.01.062

trading costs. According to the financial times, the London stock exchange foresees that about 60% of all its orders in the year 2007 will be entered by algorithmic trading.

Over the years, a plethora of statistical and econometric techniques have been developed to analyze financial data De Gooijer and Hyndma, 2006. Classical time series analysis models, such as ARIMA and GARCH, as well as many other extensions and variations, are often used to obtain insights into the mechanisms that generates the observed data and make predictions Chatfield, 2004. However, in some cases, conventional time series and other predictive models may not be up to the challenges that we face when developing modern algorithmic trading systems. Firstly, as the result of developments in data collection and storage technologies, these applications generate massive amounts of data streams, thus requiring more efficient computational solutions. Such streams are delivered in real-time; as new data points become available at very high frequency, the trading system needs to quickly adjust to the new information and take almost instantaneous buying and selling decisions. Secondly, these applications are mostly exploratory in nature: they are intended to detect patterns in the data that may be continuously changing and evolving over time. Under this scenario, little prior knowledge should be injected into the models; the algorithms should require minimal assumptions about the data-generating process, as well as minimal user specification and intervention.

In this work we focus on the problem of identifying time-varying dependencies between co-evolving data streams. This task can be casted into a regression problem: at any specified point in time, the system needs to quantify to what extent a particular stream depends on a possibly large number of other explanatory streams. In algorithmic trading applications, a data stream may comprise daily or intra-day prices or returns of a stock, an index or any other financial instrument. At each time point, we assume that a target stream of interest depends linearly on a number of other streams, but the coefficients of the regression models are allowed to evolve and change smoothly over time.

The paper is organized as follows. In Section 2 we briefly review a number of common trading strategies and formulate the problem arising in *statistical arbitrage*, thus proving some background material and motivation for the proposed methods. The flexible least squares (FLS) methodology is introduced in Section 3 as a powerful exploratory method for temporal data mining; this method fits our purposes well because it imposes no probabilistic assumptions and relies on minimal parameter specification. In Section 4 some assumptions of the FLS method are revisited, and we establish a clear connection between FLS and the well-known Kalman filter equations. This connection sheds light on the interpretation of the model, and naturally yields a modification of the original FLS that is computationally more efficient and numerically stable. Experimental results that have been obtained using the FLS-based trading system are described in Section 5. In that section, in order to deal with the large number of predictors, we complement FLS with a feature extraction procedure that performs on-line dimensionality reduction. We conclude in Section 7 with a discussion on related work and directions for further research.

2. A concise review of trading strategies

Two popular trading strategies are *market timing* and *trend following*. Market timers and trend followers both attempt to profit from price movements, but they do it in different ways. A market timer forecasts the direction of an asset, going long (i.e. buying) to capture a price increase, and going short (i.e. selling) to capture a price decrease. A trend follower attempts to capture the market trends. Trends are commonly related to serial correlations in price changes; a trend is a series of asset prices that move persistently in one direction over a given time interval, where price changes exhibit positive serial correlation. A trend follower attempts to identify developing price patterns with this property and trade in the direction of the trend if and when this occurs.

Although the time-varying regression models discussed in this work may be used to implement such trading strategies, we will not discuss this further. We rather focus on *statistical arbitrage*, a class of strategies widely used by hedge funds or proprietary traders. The distinctive feature of such strategies is that profits can be made by exploiting statistical *mispricing* of one or more assets, based on the expected value of these assets.

The simplest special case of these strategies is perhaps *pairs trading* (see Elliott, van der Hoek, & Malcolm, 2005; Gatev, Goetzmann, & Rouwenhorst, 2006). In this case, two assets are initially chosen by the trader, usually based on an analysis of historical data or other financial considerations. If the two stocks appear to be tied together in the long term by some common stochastic trend, a trader can take maximum advantage from temporary deviations from this assumed equilibrium[2].

A specific example will clarify this simple but effective strategy. Fig. 1 shows the historical prices of two assets, SouthWest Airlines and Exxon Mobil; we denote the two price time series by y_t and x_t for $t = 1, 2, \ldots$, respectively. Clearly, from 1997 till 2004, the two assets exhibited some dependence: their spread, defined as $s_t = y_t - x_t$ (plotted in the inset figure) fluctuates around a long-term average of about -20. A trading system implementing a pairs trading strategy on these two assets would exploit temporary divergences from this market equilibrium. For instance, when the spread s_t is greater than some predetermined positive constant c, the system assume that the SouthWest Airlines is overpriced and would go short on SouthWest Airlines and long on Exxon Mobil, in some predetermined ratio.

[2] This strategy relies on the idea of *co-integration*. Several applications of co-integration-based trading strategies are presented in Alexander and Dimitriu (2002) and Burgess (2003).