Forecasting semi-stationary processes and statistical arbitrage

Si Bao, Shi Chen, Wei An Zheng & Yu Zhou

Taylor & Francis
Taylor & Francis Group

Check for updates

Forecasting semi-stationary processes and statistical arbitrage

Si Bao, Shi Chen, Wei An Zheng and Yu Zhou

School of Statistics, East China Normal University, Shanghai, People's Republic of China

ABSTRACT

If a financial derivative can be traded consecutively and its terminal payoffs can be adjusted as the sum of a bounded process and a stationary process, then we can use the moving average of the historical payoffs to forecast and the corresponding errors form a generalised mean reversion process. Thus we can price the financial derivatives by its moving average. One can even possibly get statistical arbitrage from certain derivative pricing. We particularly discuss the example of European call options. We show that there is a possibility to get statistical arbitrage from Black–Scholes's option price.

1. Introduction

In economics, arbitrage is the act of exploiting price differences between two or more markets: building a portfolio of deals upon the imbalance, expecting to get the spread profit. In academia, arbitrage refers to the cost of the possibility obtaining a risk-free profit after the transaction is completed. For example, when there is an arbitrage opportunity, there will be a low buy and sell high immediately.

The practice of 'statistical arbitrage' actually came together with gambling. Suppose that a dice is rolled continuously and a gambler may lose 1 dollar each time when the result is '1' or wins 1 dollar otherwise. Denote by S_n the gambler's gain after n tries, then S_n/n is approximately $\frac{2}{3}$ for large n. That is the law of large numbers in probability theory on which based the theory of mathematical statistics. Furthermore the strong law of large numbers tells us that when $n \to \infty$, S_n/n approaches to $\frac{2}{3}$ with probability 1. However, that is not an arbitrage opportunity for this gambler as he still has a chance to lose no matter how large n is. Morgan Stanley started in the early 1980s (see Gregory van Kipnis' foreword to Pole, 2007) to apply 'statistical arbitrage' to get profit from the stock market. Nevertheless, the returns in the stock market are not independent as rolling dices, so many mathematical models were created, of which most would be possibly never published due to profitability. In Pole (2007), Andrew Pole showed many important historical examples, 'rules', and structural models of statistical arbitrage. In academic use, an arbitrage is risk-free; in common use, as in statistical arbitrage, it may refer to expected profit, though losses may occur, and in practice, there are always risks in arbitrage, some minor (such as fluctuation of prices decreasing profit margins), some major (such as

devaluation of a currency or derivative). In academic use, an arbitrage involves taking advantage of differences in price of a single asset or identical cash flows; in common use, it is also used to refer to differences between similar assets (relative value or convergence trades), as in merger arbitrage.

There have been a few more general academic definitions of statistical arbitrage since that time on. Hogan, Jarrow, and Warachka (2002) defined 'Statistical Arbitrage' by four conditions that the discounted cumulative value $v(t)$ should satisfy: (1) $v(0) = 0$; (2) $\lim_{t \to \infty} E[v(t)] > 0$; (3) $\lim_{t \to \infty} \mathrm{Var}(v(t))/t = 0$; (4) $\lim_{t \to \infty} P[v(t) < 0] = 0$. Their Condition (3) excluded the cases where $v(t)$ is the sum of independent identically distributed outcomes, and thus excluded our dice game and most of casino games. According to Lo (2010), 'Statistical Arbitrage' refers to highly technical short-term mean-reversion strategies involving large numbers of securities (hundreds to thousands, depending on the amount of risk capital), very short holding periods (measured in days to seconds), and substantial computational, trading and information technology (IT) infrastructure. Wang and Zheng (2014) simplified the definition to repeatedly trading a basket of assets according to the same algorithm and get an accumulated profit with statistically stable positive rate. Moreover, the concept of the ergodic theorem of stationary process was applied into the practice of statistical arbitrage in Wang and Zheng (2014). We will further this discussion to derivative pricing in this paper. We will discuss the time series with bounded trend component, of which the increments have some nice property as mean reversion.

The 'mean reversion' phenomena is a special case of stationarity, which suggests that prices and returns

CONTACT Si Bao ✉ christmasbao001@163.com

eventually move back towards the mean or average. This mean or average can be the historical average of the price or return or another relevant average such as the growth in the economy or the average return of an industry. A typical example of mean reversion phenomena is the increments of the Ornstein–Uhlenbeck process (see, e.g., Karatzas & Shreve, 1987, p. 358). However, the Ornstein–Uhlenbeck process comes together with Gaussian law which could not be justified by some real market data. Thus we will define a generalised discrete time mean reversion process as the increments of a semi-stationary process which is the sum of a bounded process and a stationary process. The time average of a generalised mean reversion process vanishes faster than that of an ordinary stationary processes with null mean. Therefore a generalised mean reversion process is more stable than an ordinary stationary one. We show in the next section that if we use the moving average to estimate a time series with bounded trend component, then the time average of the error vanishes in inverse proportion to the time length. As its application, we give an example of statistical arbitrage from Black–Scholes' option price based on estimate of payoffs with generalised mean reversion errors.

We introduce the concept of generalised mean reversion process as the increment process of a semi-stationary process in Section 2. We use Section 3 to explain some basic facts of option pricing. We use Section 4 to study the options of the ETFs which track the main stock indices in the US market. Finally we show that one may get statistical arbitrage from Black–Scholes model in the last section.

2. Forecasting with MA

Given a discrete time stochastic process $\{X(t)\}_{t=0,1,2,\dots}$, it is very important to give an estimate $\tilde{X}(t+1)$ for $X(t+1)$ based on the observed information $\{X(0), X(1), \dots, X(t)\}$. Certainly, the selection of $\tilde{X}(t+1)$ depends on the selection of the norm of the error $X(t+1) - \tilde{X}(t+1)$. When we need to estimate continuously for $t = m+1, m+2, \dots, N$, then the most popular ones are the sum of l_p−norms ($p \geq 1$) of errors

$$\sum_{t=m+1}^{N} \{E[|X(t) - \tilde{X}(t)|^p]\}^{1/p}$$

or the similar ones. However, in many financial applications, an investor is more interested in the l_p-norm of the accumulated sum:

$$\left\{ E[|\sum_{t=m+1}^{N} (X(t) - \tilde{X}(t))|^p] \right\}^{1/p}$$
$$= \left\{ E[|\sum_{t=m+1}^{T} X(t) - \sum_{t=m+1}^{N} \tilde{X}(t)|^p] \right\}^{1/p}. \quad (1)$$

The economic reason is very simple. If $X(t+1)$ is the value of certain financial derivative at time $t+1$ and $X(t+1) - \tilde{X}(t+1)$ is the loss caused by estimate $\tilde{X}(t+1)$, then an investor is more interested in the accumulated lose $\sum_{t=m+1}^{N}(X(t) - \tilde{X}(t))$ rather than the sum $\sum_{t=m+1}^{N} |X(t) - \tilde{X}(t)|$. For the error norm (1), the simplest estimate is to use the moving average (ab. MA)

$$M^{(m)}(t) = \frac{X(t) + X(t-1) + \cdots + X(t-m+1)}{m} \quad (2)$$

to forecast $X(t+1)$, which is called the MA model in time series analysis. We will impose a condition 'semi-stationary' on $\{X(t)\}$ under which (1) will be bounded in T for some $p \geq 0$.

We say that $\{X(t)\}$ is a 'strongly stationary' process, if for each positive a, the processes $\{X(t+a)\}$ and $\{X(t)\}$ obey the same probability law. Thus a sequence of independent identically distributed random variables is a strongly stationary process. In the previous example of continuously rolling dice, if we denote $X(t) = -1$ when the tth outcome is '1', and $X(t) = 1$ otherwise, then $\{X(t)\}$ is a strongly stationary process. The weak form is the weakly stationary process. $\{X'(t)\}_t$ is a 'weakly stationary' process if: (i) $E[X'(t)]$ is a constant; (ii) $\text{Cov}(X'(t), X'(t+a)) = \text{Cov}(X'(0), X'(a))$ for each t. Certainly, if a strongly stationary process $\{X(t)\}$ has its second moments, then $\{X(t)\}$ is weakly stationary. The most important property of a strongly stationary process is the Birkhoff's theorem (Loeve, 1977, p. 76). The time-average $(X(1) + X(2) + \cdots + X(N))/N$ of a strongly stationary process $\{X(t)\}$ converges with probability 1. Furthermore, when $\{X(t)\}$ has the so-called 'ergodic' property, then the limit is just its mean.

Definition 2.1: We call $\{X(t)\}$ a strongly (or weakly) semi-stationary process if there is a bounded process $\{A(t)\}$ such that $\{X(t) - A(t)\}$ is a strongly (or weakly, respectively) stationary process.

It is easy to see that for a semi-stationary process $\{X(t)\}$, $X(T)/T \to 0$ with the rate $1/T$ and the time average

$$\frac{\Delta X(1) + \Delta X(2) + \cdots + \Delta X(N)}{N}$$

of the increments $\Delta X(t) = X(t+1) - X(t)$ of a semi-stationary process $\{X(t)\}$ vanishes at the rate $1/N$ which is faster than that of an ordinary sequence of independent identically distributed random variables, which is $1/\sqrt{N}$.

A time series $\{x(t)\}$ is often to be considered as a sequence of data which can be written as

$$x(t) = m(t) + S(t) + \epsilon(t),$$

where $m(t)$ is the deterministic trend component, $S(t)$ is the seasonal component and $\epsilon(t)$ is a weakly stationary process (see Wang & Zheng, 2014, p. 101, for example). Since $S(t)$ is periodic, it is bounded. If $m(t)$ is also bounded, then $\{x(t)\}$ is a weakly semi-stationary process. We are going to use MA to forecast a weakly semi-stationary time series and the cumulated error has mean vanishes at the rate $1/T$.

The increment of a semi-stationary process has the properties which are very similar to the so-called mean reversion process. The Mean Reversion phenomena in stock prices have been studied for more than three decades. This theory suggests that prices and returns eventually move back towards the mean or average. This mean or average can be the historical average of the price or return or another relevant average such as the growth in the economy or the average return of an industry (see, e.g., Ansley, Spivey, & Wrobleski, 1977; Fama & French, 1988; Mukherji, 2011; Poterba & Summers, 1988). Its typical mathematical model is the Ornstein–Uhlenbeck process of which the increments can be written as

$$\mathrm{d}X(t) = \theta(\mu - X(t))\,\mathrm{d}t + \sigma\,\mathrm{d}W(t), \qquad (3)$$

where μ, $\theta(>0)$ and $\sigma(>0)$ are constants. $\mathrm{d}X(t)$ and $\mathrm{d}W(t)$ are the increments of $X(t)$ and Brownian motion $W(t)$ respectively. Since the above equation is not taught in ordinary text books of probability theory, we briefly explain its meaning here. One considers $\mathrm{d}W(t)$ as the noise and σ as its magnitude. $(X(t) - \mu)$ is the distance from the mean μ. So this process has a drift $\theta(\mu - X(t))$ push its path back to the mean. If we denote $X(t) - \mu = Y(t)$, then

$$\mathrm{d}Y(t) = -\theta Y(t)\,\mathrm{d}t + \sigma\,\mathrm{d}W(t),$$

which has its integral form (see Karatzas & Shreve, 1987, p. 358).

$$Y(t) = Y(0)\exp\{-\theta t\} + \sigma\int_0^t \exp\{\theta(s-t)\}\,\mathrm{d}W(s).$$

Thus

$$X(t) = (X(0) - \mu)\exp\{-\theta t\}$$
$$+ \sigma\int_0^t \exp\{\theta(s-t)\}\,\mathrm{d}W(s) + \mu,$$

which is strongly stationary if $X(0)$ has Gaussian distribution with mean μ and variance $\sigma^2/2\theta$. However, it is known that this process does not precisely match a lot of data experimentally. Therefore we need a more general definition. If we fix $\{Y(t)\}$ and set $X(t) = Y(t) + \mu$, then $X(t)$ satisfies (3) and they have same increments:

$$X(t+1) - X(t)$$
$$= Y(t+1) - Y(t), \quad (t = 0, 1, 2, \ldots).$$

Hence, from the uniqueness in law of the Ornstein–Uhlenbeck process (for fixed initial stationary distribution, μ, θ and σ), we get the following.

Lemma 2.2: *For fixed θ and σ, the probability law of the increments $\{X(t+1) - X(t)\}_{t=0,1,2,\ldots}$ of stationary Ornstein–Uhlenbeck process do not depend on the mean μ.*

The above Lemma inspired us to introduce the following.

Definition 2.3: Let $\{U(t)\}_{t=1,2,\ldots}$ be a stochastic process. If there is a strongly (or weakly) semi-stationary process $\{X(t)\}_{t=0,1,2,\ldots}$ such that $U(t) = \Delta X(t)$, $(t = 1, 2, \ldots)$, then $\{U(t)\}$ is called a generalised strongly (or weakly, respectively) Mean Reversion process and $\{X(t)\}$ is its integrated process. Furthermore, if the integrated process is strongly (or weakly) stationary, then $\{U(t)\}$ is called strongly (or weakly, respectively) Mean Reversion process.

Thus the time average

$$\frac{U(1) + \cdots + U(N)}{N} = \frac{X(N) - X(0)}{N} \to 0$$

with rate $1/N$.

We have easily that

$$\sum_{t=1}^{m} \frac{t}{m}X(t) + \sum_{t=m+1}^{N} [X(t) - M^{(m)}(t-1)]$$

$$= \sum_{t=1}^{m} \frac{t}{m}X(t) + \sum_{t=m+1}^{N} \left[X(t) - \frac{1}{m}\sum_{i=1}^{m} X(t-i) \right]$$

$$= \sum_{t=1}^{m} \frac{t}{m}X(t) + \sum_{t=m+1}^{N} X(t)$$

$$- \frac{1}{m}\sum_{i=1}^{m}\sum_{t=m+1}^{N} X(t-i)$$

$$= \sum_{i=1}^{m} \frac{m+1-i}{m}X(N+1-i)$$

$$= \sum_{t=1}^{m} \frac{t}{m}X(N-m+t).$$

Thus we have

Lemma 2.4:

$$\sum_{t=m+1}^{N} [X(t) - M^{(m)}(t-1)]$$

$$= \sum_{i=1}^{m} \frac{i}{m}X(N-m+i) - \sum_{i=1}^{m} \frac{i}{m}X(i).$$

Since $\sum_{i=1}^{m}(i/m)X(t - m + i)$ is a finite linear combination of the values at different time of a strongly (or weakly) stationary processes, which is also strongly (weakly, respectively) stationary, we get

Theorem 2.5: *When $\{X(t)\}$ is a strongly (or weakly) semi-stationary process, the error process $\{X(t + 1) - M^{(m)}(t)\}$ is a generalised strongly (or weakly, respectively) mean reversion process and the accumulate error is strongly (or weakly, respectively) semi-stationary such that*

$$\frac{\sum_{t=m+1}^{N}(X(t) - M^{(m)}(t - 1))}{N} \to 0$$

with rate $1/N$.

Therefore, if $\{X(t)\}$ is a sequence of financial derivatives in the market, which is semi-stationary, then we can use $\{M^{(m)}(t - 1)\}$ to forecast its values with errors forming a generalised mean reversion process. Hence the errors have mean nearly 0 in long run. In particular, we will consider option prices in the next section.

The ergodic theorem for stationary process has its weak point. It cannot be applied to an arbitrary subsequence of a stationary process. For example, when X is a standard Gaussian random variable, then $\{X, -X, X, -X, \ldots\}$ is a strongly stationary process. However, one can easily choose a subsequence which has no convergent time average. Indeed, if we choose $\{X, -X, -X, -X, X, X, X, X, X, X, X, X, X, X, -X, -X, \ldots\}$, then its time average will take values $X/2$ and $-(X/2)$ repeatedly. Therefore, we need the following.

Lemma 2.6: *Suppose that $\{X_j\}$ is a weakly stationary sequence. If $\{X_{j_i}\}$ is a subsequence and $\min_i\{|j_i - j_{i+1}|\} \geq J$, then for any $\epsilon > 0$,*

$$P\left[\left|\frac{1}{N}\sum_i^N X_{j_i} - E[X_1]\right| \geq \epsilon\right]$$

$$\leq \frac{1}{\epsilon^2}\left[\frac{\mathrm{Var}(X_1)}{N} + \sup_{i \geq J}|\mathrm{Cov}(X_1, X_{1+i})|\right].$$

Proof.

$$E\left|\frac{1}{N}\sum_i^N X_{j_i} - E[X_1]\right|^2$$

$$\leq \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}(X_{j_i}) + \frac{2}{N^2}\sum_{i<k}^{N}\mathrm{Cov}(X_{j_i}, X_{j_k})$$

Thus we get the result by simplification and the classical Chebyshev's inequality in probability theory.

3. Price of option

The basic idea of Black–Scholes–Merton's theory (Black & Scholes, 1973; Merton, 1973) is that the option price of an asset depends only on the current price $S(t)$ of the asset, mature time, strike price, volatility and risk-free interest rate. It is also known that the real traded prices of various options in the market are quite different from those theoretical ones, so that the concepts of 'implied volatility' and 'stochastic volatility' were introduced (Canina & Figlewski, 1993; Chrisensen & Prabhala, 1998; Dumas, Fleming, & Whaley, 1998; Heston, 1993; Poon & Granger, 2003) as attempts to fill the gap between the theoretical ones and the real market ones. In the last two decades, many mathematicians and statisticians introduced a lot of more sophisticated stochastic models in order to describe more accurately the movement of stock price and its option prices. Among them, the volatility study always attracts the main interests.

In the basic Black–Scholes's model, stock share price is assumed to be a geometric Brownian motion. That is,

$$S(t) = S(0)\exp\left\{\sigma W(t) - \left(\frac{\sigma^2}{2} - \mu\right)t\right\}.$$

The main mathematical tool in Black–Scholes's theory is Cameron–Martin–Girsanov's theorem (Karatzas & Shreve, 1987, p. 190) which states that up to a bounded time T, the induced probability measures corresponding to different μ are equivalent. Therefore one can choose the probability measure which makes the geometric Brownian motion a martingale ($\mu = 0$) to get the call option price c through taking the mathematical expectation of

$$(S(T) - K)^+ = c + \int_0^T H(t)\,dS(t),$$

where $H(t)$ is the hedge, K is the strike price and T is the mature time. Therefore the call option price is just the mathematical expectation of $(S(T) - K)^+$ with respect to the risk neutral ($\mu = 0$) probability measure. However, some limit property cannot be kept when $T \to \infty$. For example, $\log S(T)/T \to -(\sigma^2/2 - \mu)$ depending on μ. Therefore in long run, some results deduced from real market price might be different to those deduced from the risk neutral one.

Let us fix a positive integer T. The payoff of a_i shares of European call option with mature time at $(i + 1)T$ and strike price K_i ($i = 0, 1, 2, 3, \ldots$) is

$$a_i V_{i+1} = a_i(S((i + 1)T) - K_i)^+.$$

Denote by Q_i the option price per share paid at time iT, then a buyer's profit at time $(i + 1)T$ will be

$$a_i C_{i+1} = a_i V_{i+1} - a_i Q_i$$
$$= a_i(S((i + 1)T) - K_i)^+ - a_i Q_i.$$

If we can find a sequence of $\{(K_i, a_i)\}_i$ such that $\{a_i V_{i+1}\}_i$ form a semi-stationary sequence, then we may apply Theorem 2.5 to use its moving averages to

estimate the payoffs $\{a_i V_{i+1}\}_i$ based on the market data and taken as $\{a_i Q_i\}$ such that the error terms $\{a_i C_i\}$ are generalised mean reversion.

In order to get such a stationary sequence, we may choose (at time iT) $K_i = kS(iT)$ and $a_i = 1/S(iT)$ where k is a prefixed positive percentage constant. Denote by q_i the cost (at time iT) for purchasing a_i shares of option, then the terminal payoff and the buyer's profit will be

$$v_{i+1} = \left(\frac{S((i+1)T)}{S(iT)} - k \right)^+ \quad \text{and}$$

$$c_{i+1} = \left(\frac{S((i+1)T)}{S(iT)} - k \right)^+ - q_i,$$

which can also be considered the corresponding values percentage to the original asset prices. In practice, in order to get profit percentage continuously, we may fix a large constant C and just trade $C/S(jT)$ shares of options with strike price $kS(jT)$ (both rounded out to the nearest adequate digits). In Black–Scholes's model, $\{S((i+1)T)/S(iT)\}_i$ are just the exponential functions of the increments of Brownian motion, which are independent and identically distributed Gaussian random variables. So the price q_i is the mean of $(S((i+1)T)/S(iT) - k)^+$ under the risk-neutral measure, which is a constant. Therefore $\{v_i\}_i$ and $\{c_i\}_i$ are both strongly stationary processes in Black–Scholes' theory. Furthermore, the error terms $\{c_i\}_i$ are independent identically distributed in Black–Scholes' formula, of which the time average converges slower than that of generalised mean reversion process.

By Lemma 2.6, if $\mathrm{Cov}(v_1, v_j) \to 0$, and $\{v_{j_k}\}_k$ is a subsequence such that $\min_k\{j_{k+1} - j_k\}$ is sufficiently large, then $1/N \sum_i^N v_{j_i}$ will be close to $E[v_1]$. Its special case is the logarithmic return of Heston model (Heston, 1993), which is strongly stationary and ergodic, $\mathrm{Cov}(c_1, c_N) \to 0$ (when $N \to \infty$).

In Black–Scholes–Merton's theory, the only undetermined factor is the volatility σ. In practice, the investors quite often use the near term sample variance to estimate the volatility at time t. That is,

$$\hat{\sigma}^2(t) = \frac{1}{n-1} \sum_{j=t-n}^{t-1} \left[(\log S(j+1) - \log S(j)) \right.$$

$$\left. - \frac{1}{n} \sum_{i=t-n}^{t-1} (\log S(i+1) - \log S(i)) \right]^2 \quad (4)$$

for large n. However, if we use the above $\hat{\sigma}(n)$ with large n for a unified volatility in Black–Scholes formula, the result would be not good. So very often the investors use smaller n to estimate the volatility of $S(t)$ near t, which is just Heston's price with $\rho = 0$ (Heston, 1993; Poon & Granger, 2003).

From the above discussion, we have four different candidates for option prices: (1) the percentage prices

$\{q_i\}$ are the moving averages $\{M(iT)\}$; (2) $\{q_i\}$ are the historical call option prices in the real market; (3) $\{q_i\}$ are Black–Scholes' prices with unified volatility $\hat{\sigma}$; (4) $\{q_i\}$ are Black–Scholes' prices with moving volatility $\hat{\sigma}(t)$. We will use Market data to show that the cumulated error is stationary in Case (1). We will also illustrate the cumulated errors in the other three cases for comparison.

4. Statistics with real market data

The strong stationarity of logarithmic return is implied in the most popular mathematical models for asset prices. If we choose the trading volume inversely proportional to the current price, then the consecutive payoffs will be strongly stationary as discussed in the previous section. Thus we can use the moving averages $\{M(iT)\}$ to estimate a this sequence and the errors will be mean reversion according to Theorem 2.5. In this section, we use the data of ETFs DIA, QQQ, SPY for main stock indices in the US market to illustrate the cumulated buyer's profit percentage, which is also the cumulated percentage errors, to compare with that of Black–Scholes'.

We consider the buyers successive profits

$$c_{i+1} = \left(\frac{S((i+1)T)}{S(iT)} - k \right)^+ - q_i$$

for the prices $\{q_i\}$ defined in the four cases listed at the end of the previous section, which are also the cumulated errors of the corresponding option pricing.

The first thing is to test the stationarity. We have checked by the ADF test with Matlab software package and the hypothesis that '$\{c_{i+1}\}$ has a unit root' is rejected with 95% of confidence, for the trading data of SPY (the ETF tracing S&P500), QQQ (the ETF tracing NASDAQ) and DIA (the ETF tracing DOW) under various k and T. In other words, $\{c_{i+1}\}$ can be statistically considered as a stationary process. For example, when we take $S(t)$ as the price of SPY, take $k = 0.98$ and $T = 7$, Figure 1 shows the successive weekly trading profit sequence $\{c_{i+1}\}$ when $q_i = M(iT)$ ($m = 6$) from 6th January 1995 through the end of July 2015.

We also found that if q_i is the weekly expired percentage option prices in the real market (with $kS(jT)$ rounded out to the nearest strike price in the market), $\{c_{i+1}\}$ still pass the ADF test for SPY, QQQ and DIA data with various k and T.

The parameter k may assume any value. For simplicity, we just list the results for $k = 0.98$, 1 and 1.02 here. We discuss two time scales for T: (a) weekly expired options; (b) monthly expired options. For weekly expired options, we consider to trade call options between the closing time of neighbouring Fridays. Thus if we purchase $1/S(jT)$ $(j = 1, 2, \ldots, n)$ unit of call option at time jT with mature time $(j+1)T$ and strike price $K = kS(jT)$, the profit by time $(j+1)T$
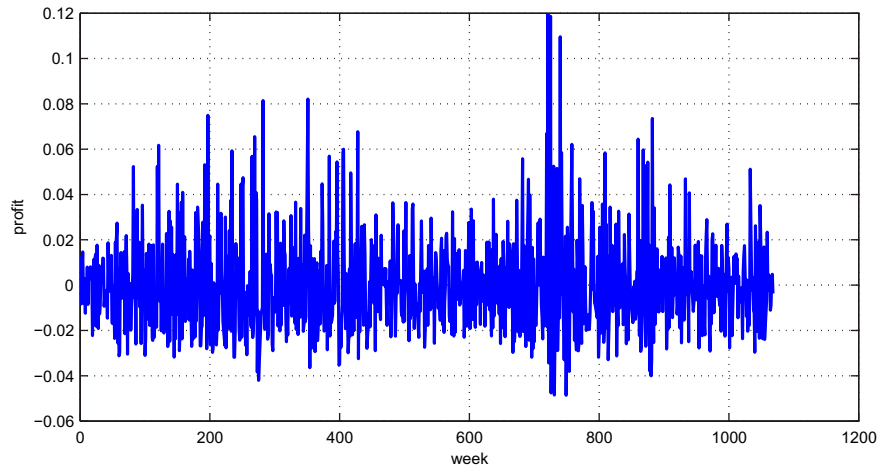
Figure 1. The successive profit sequence $\{X(j)\}$ from 6th January 1995.

Figure 2. The daily price of SPY from 3rd January 1995 through 2nd July 2015.

will be c_{j+1}, and the cumulated profit by time $(j+1)T$ is $\sum_{i=1}^{j} c_{i+1}$. In that case, we can take $T = 7$ calendar days except for very few holidays. Similarly, when we consider the monthly expired options, we trade the options at the 3rd Friday of neighbouring months. In that case, actually $T = 30$ or 31 (except for February). We compare the cumulated profits of the following four prices:

Line 1: $\{q_i\}$ are the moving average $\{M(iT)\}$;
Line 2: $\{q_i\}$ are the historical call option price in the real market;
Line 3: $\{q_i\}$ are Black–Scholes' price with unified volatility $\hat{\sigma}$;
Line 4: $\{q_i\}$ are Black–Scholes' price with moving volatility $\{\hat{\sigma}(iT)\}$.

We should mention some technical points here: (a) when there was no trade at the closing time, we use the mean of bid and ask price as our last price; (b) the strike prices $\{kS(jT)\}$ are rounded out to the nearest available strike prices in the market; (c) all our data are percentage priced; (d) the parameter n in (4) is chosen according to the usual method applied in the real

market (Poon & Granger, 2003) and the selection of m in (2) is less crucial. Actually we found that there is no much differences when m is chosen between 2 and 12.

4.1. SPY analysis

Figure 2 shows the daily price of SPY from 3rd January 1995 through 2nd July 2015.

We have SPY trading data of monthly expired options from 21st January 2005 through 20th March 2015, and its trading data of weekly expired options from 1st July 2011 through 27th March 2015. Thus we can show our cumulative profits of corresponding options in those two periods respectively.

4.1.1. SPY results under different k values

Figures 3–5 show the cumulated profits of both monthly and weekly expired options of SPY under different k values.

4.2. QQQ analysis

Figure 6 shows the daily price of QQQ from 10th March 1999 through 2nd July 2015.

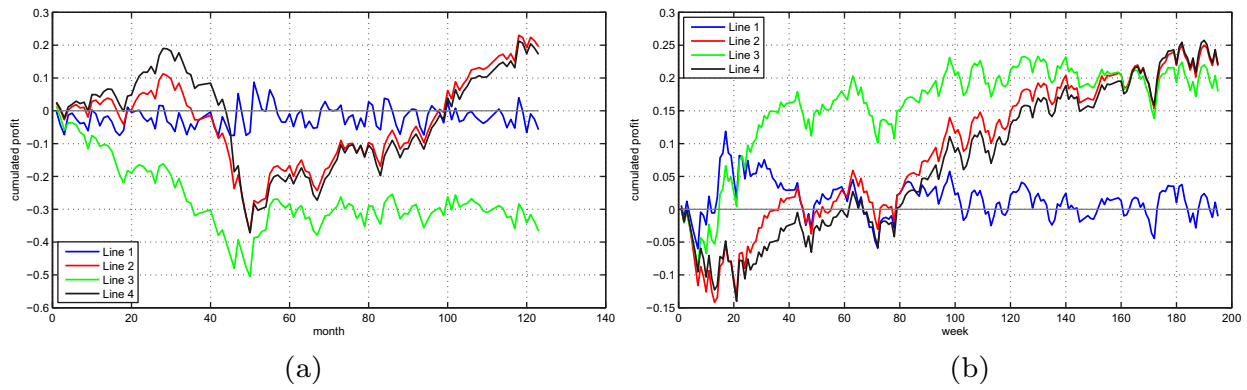We have QQQ trading data of monthly expired options from 15th February 2002 through 17th April

Figure 3. The cumulated profits of SPY options when $k = 0.98$: (a) monthly expired options and (b) weekly expired options.
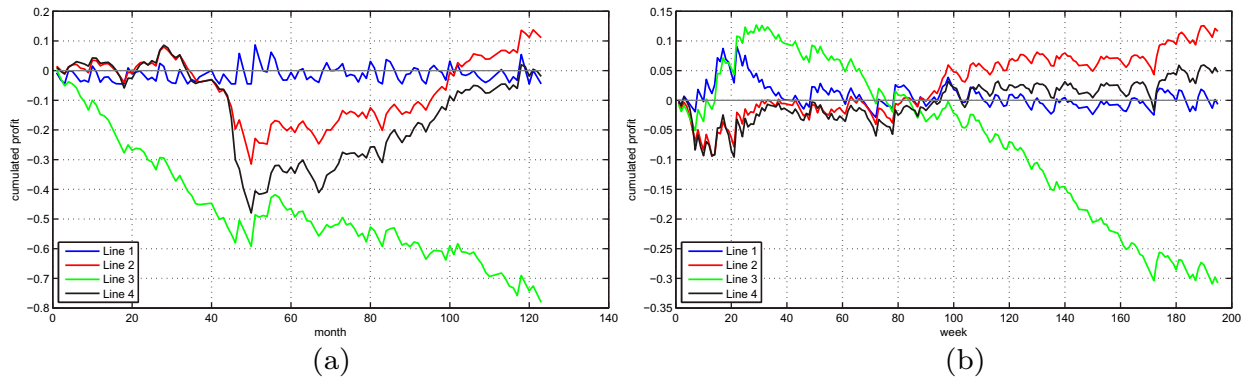

Figure 4. The cumulated profits of SPY options when $k = 1$: (a) monthly expired options and (b) weekly expired options.
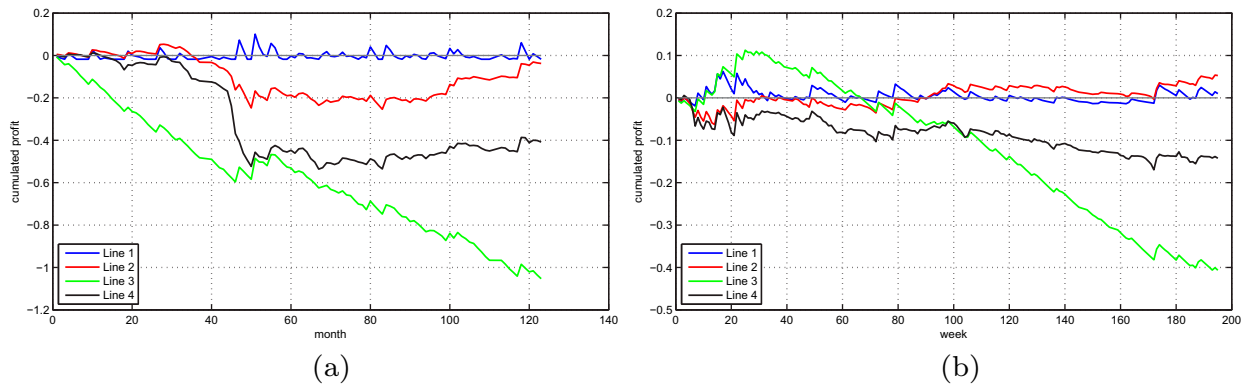

Figure 5. The cumulated profits of SPY options when $k = 1.02$: (a) monthly expired options and (b) weekly expired options.
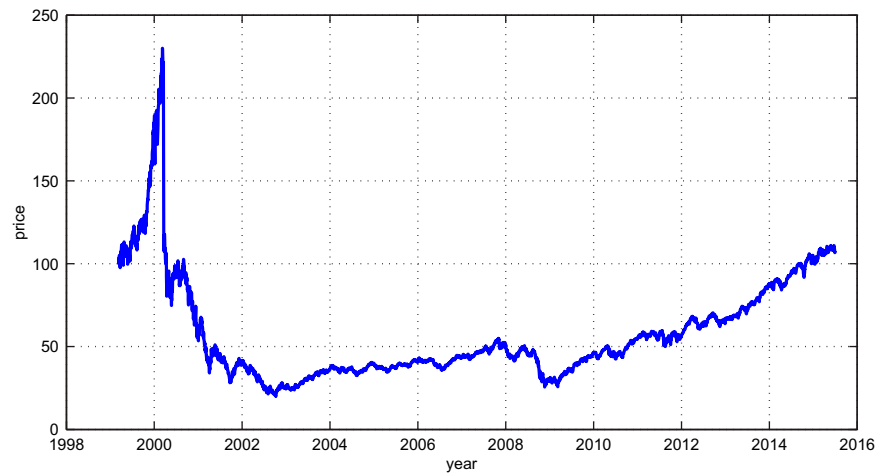

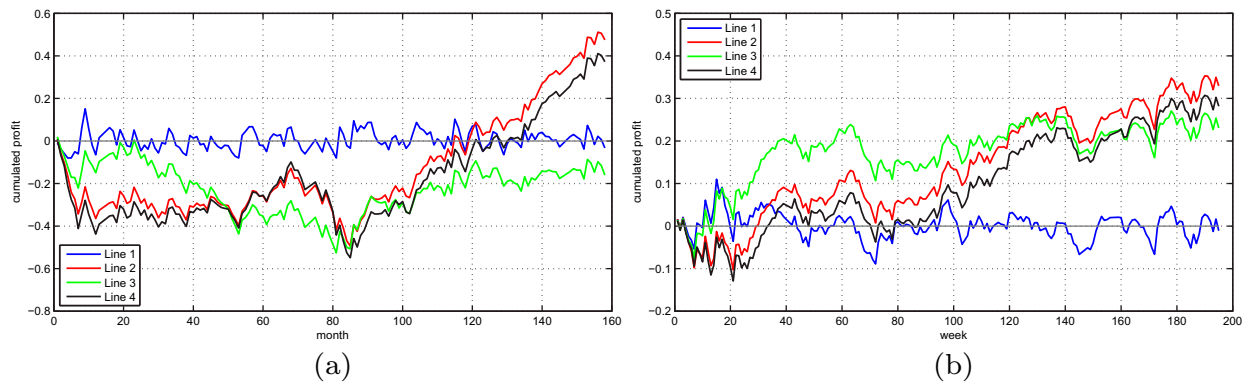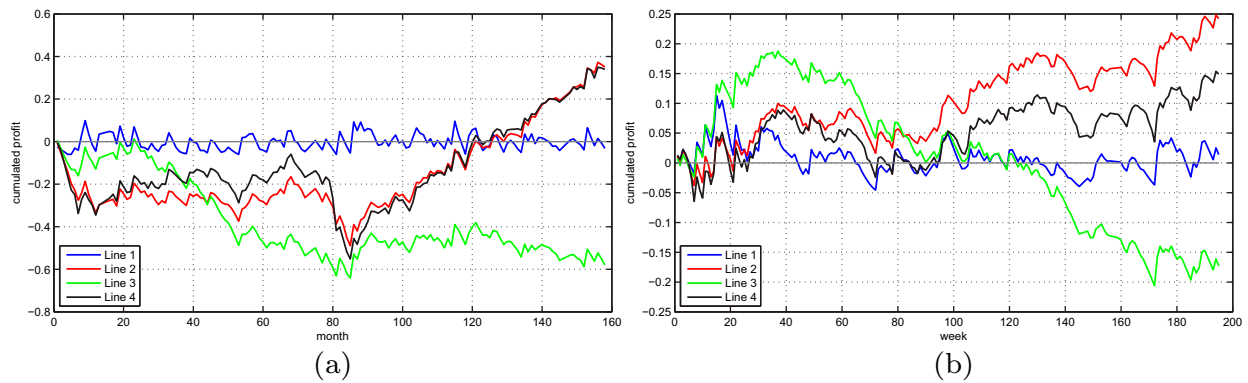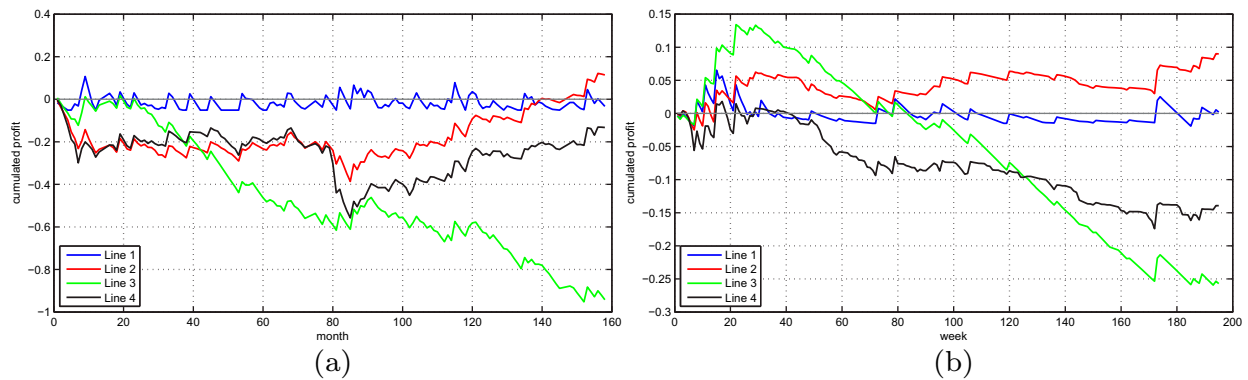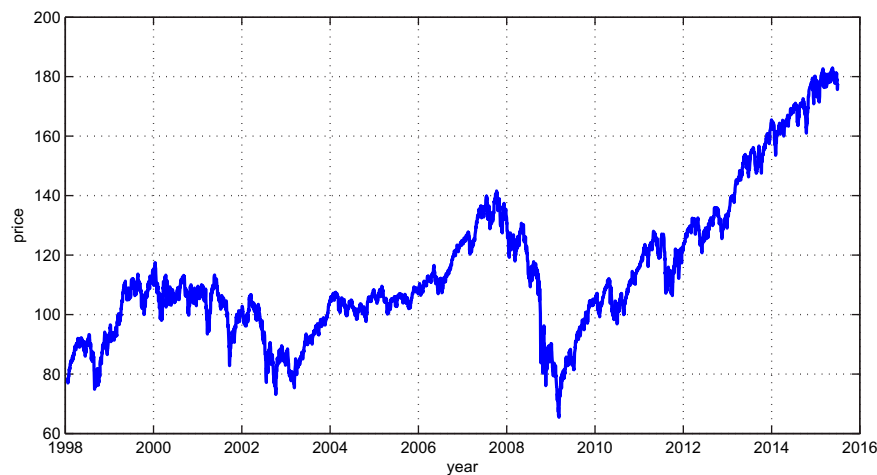**Figure 6.** The daily price of QQQ from 10th March 1999 through 2nd July 2015.
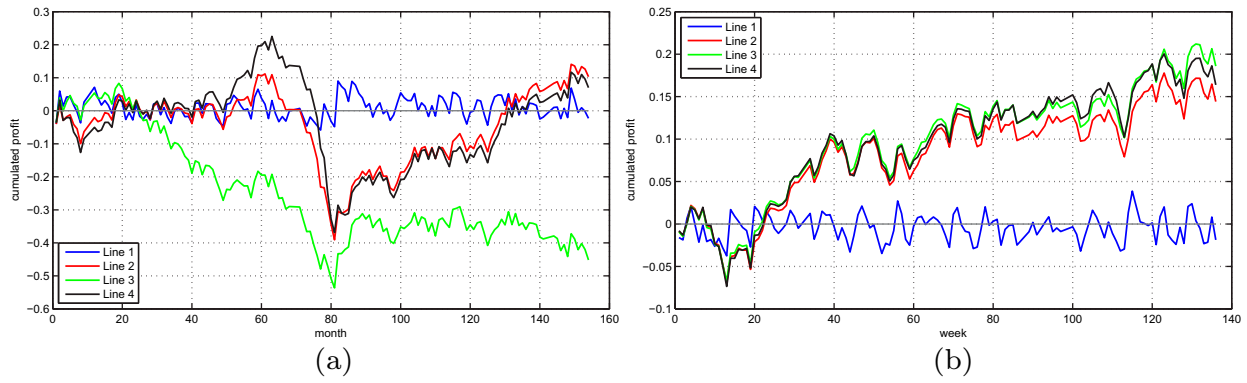
Figure 7. The cumulated profits of QQQ options when $k = 0.98$: (a) monthly expired options and (b) weekly expired options.

Figure 8. The cumulated profits of QQQ options when $k = 1$: (a) monthly expired options and (b) weekly expired options.

Figure 9. The cumulated profits of QQQ options when $k = 1.02$: (a) monthly expired options and (b) weekly expired options.

Figure 10. The daily price of DIA from 20th January 1998 through 2nd July 2015.

Figure 11. The cumulated profits of DIA options when $k = 0.98$: (a) monthly expired options and (b) weekly expired options.
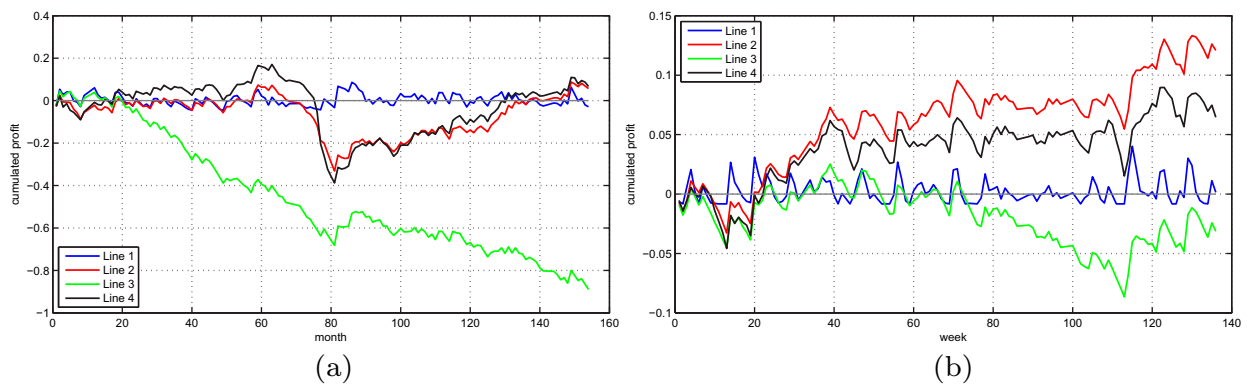

Figure 12. The cumulated profits of DIA options when $k = 1$: (a) monthly expired options and (b) weekly expired options.
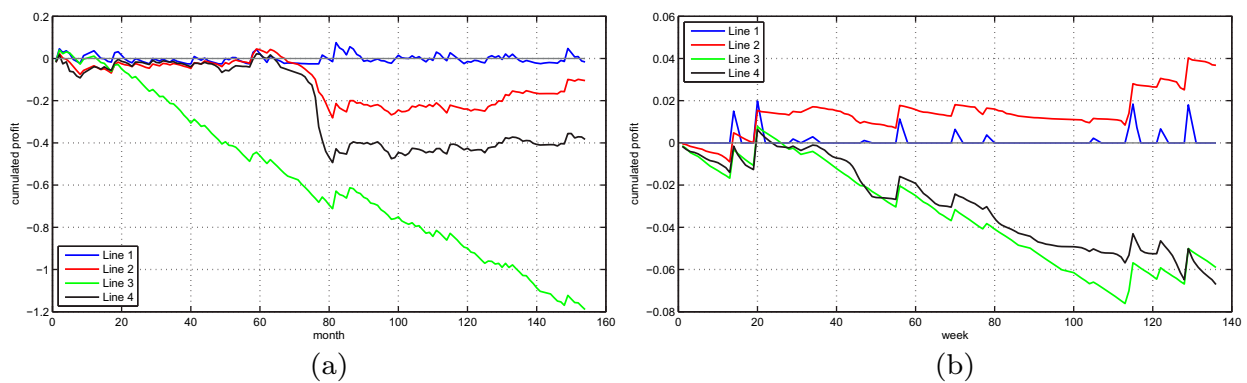

Figure 13. The cumulated profits of DIA options when $k = 1.02$: (a) monthly expired options and (b) weekly expired options.
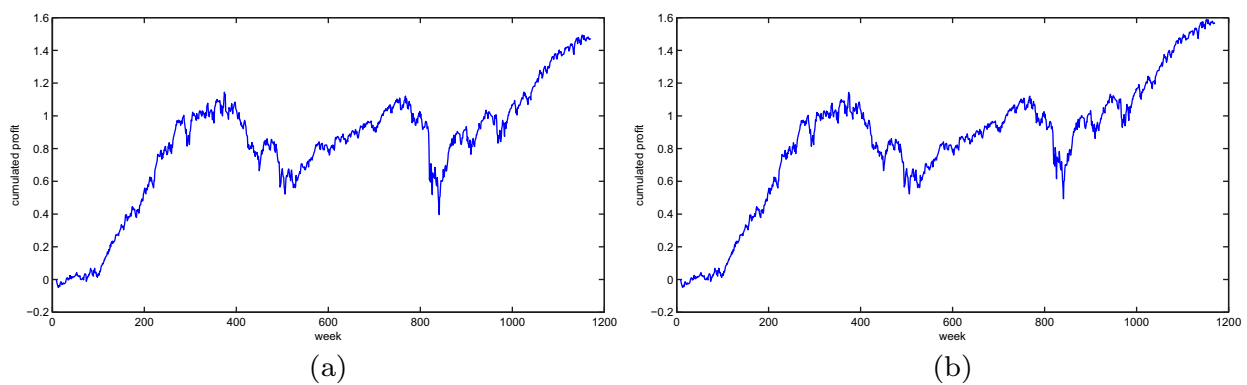

Figure 14. Cumulated gain against Black–Scholes price: (a) when $k = 0$ and (b) when $k = 0.9$.
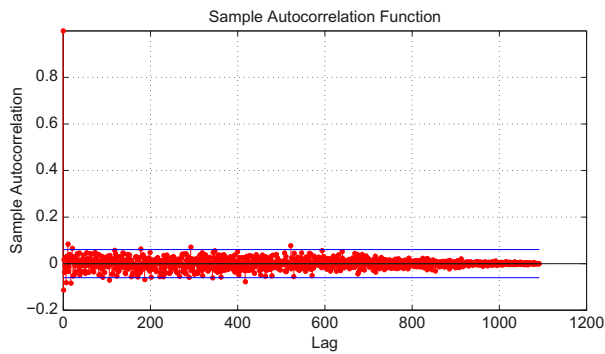
Figure 15. The autocorrelation function.

2015, and trading data of weekly expired options from 1st July 2011 through 27th March 2015.

4.2.1. QQQ results under different k values

Figures 7–9 show the cumulated profits of both monthly and weekly expired options of SPY under different k values.

4.3. DIA analysis

Figure 10 shows the daily price of DIA from 20th January 1998 through 2nd July 2015.

We have DIA trading data of monthly expired options from 21st June 2002 through 17th April 2015, and trading data of weekly expired options from 10th August 2012 through 27th March 2015.

4.3.1. DIA results under different k values

Figures 11–13 show the cumulated profits of both monthly and weekly expired options of SPY under different k values.

5. Possibility of statistical arbitrage

From the above data analysis, we can easily find that the errors of our moving average pricing of option are mean reversion and the cumulated error is stationary. When the strike price is sufficiently low, then the asset price will be always above the strike price and the call option will be always executed. So that the buyer at Black–Scholes' option price can always take statistical arbitrage in long run. Let us take $T = 7$. If one investor continuously purchase at time iT call option for $1/S(iT)$ share expired at $(i + 1)T$ according to Black–Scholes' price, then Figure 14 shows his cumulated profit in the last 22 years (01/29/1993–07/02/2015) when $k = 0$ and $k = 0.9$.
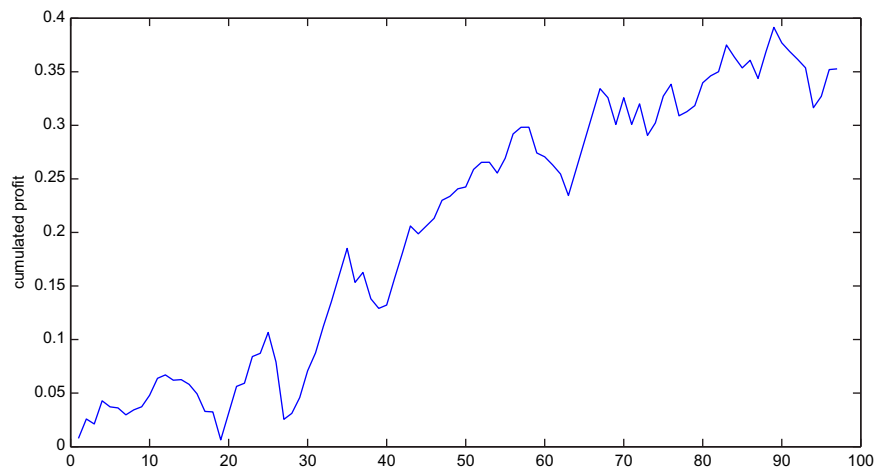

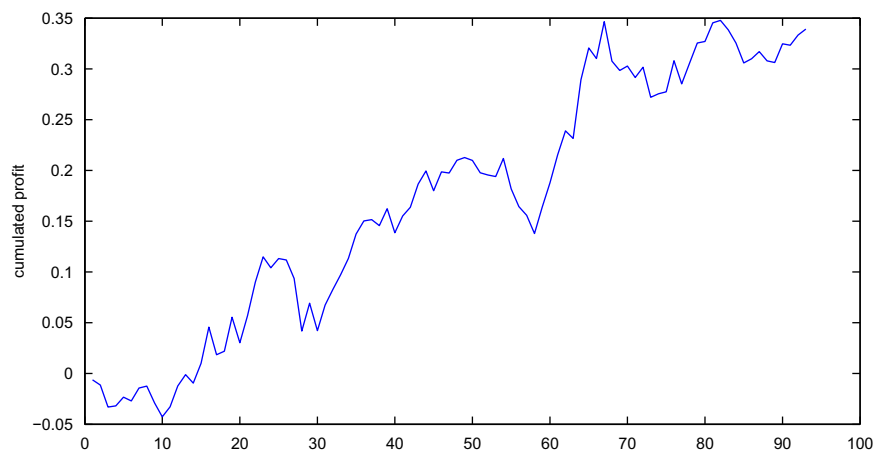**Figure 16.** Cumulated gain against Black–Scholes price.

Figure 17. Cumulated gain against Black–Scholes price with stochastic volatility.

However, there are still many problems left. For example, what happens if only some of the profits $\{c_{i+1}\}$ occur? That is, the partial sum of a strongly stationary sequence may not have a convergent mean if its terms are not uncorrelated as shown in Section 2. However, we can apply Lemma 2.6 to construct statistical arbitrages if the covariance function of $\{c_{i+1}\}$ tends to 0 quick enough.

We show here another more practical example to get statistical arbitrage from Black–Scholes' call option. The following graph shows that when $q_i = M(iT)$, $m = 6$ and $T = 7$, the sample autocorrelation function of c_{i+1} tends to 0 (Figure 15).

Thus if c_{i_k} is a subsequence such that $\min_k\{i_{k+1} - i_k\}$ is sufficiently large, we can apply Lemma 2.1 to get their mean sufficiently close to 0. Denote by $Q(iT)$ the Black–Scholes' price. Our strategy is: (a) when $M(iT) - Q(iT) > 0.01$ and the previous trade was made at least 9 weeks ago, buy $1/S(iT)$ of call option at $Q(iT)$; (b) when $Q(iT) - M(iT) > 0.0005$ and the previous trade was made at least 9 weeks ago, sell $1/S(iT)$ call option at $Q(iT)$. Figure 16 shows our cumulated profit and Figure 17 gives our profit against Black–Scholes price with moving volatility. Thus we can get statistical arbitrage from the mean, which is similar to the case of high-frequency trading (see Wang & Zheng, 2014).

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Si Bao is now working in Xiangcai security Co., LTD. She studied for her Ph.D. in School of Statistics from ECNU.

Shi Chen is a data scientist of PayPal Holding Inc. He received his Ph.D in statistics from ECNU in 2017.

Wei An Zheng is Professor of ECNU and Professor Emeritus of University of California, Irvine, USA.

Yu Zhou works in Guotai Junan Securities. He received his Ph.D in statistics from ECNU in 2016.

References

Ansley, C. F., Spivey, W. A., & Wrobleski, W. J. (1977). On the structure of moving average processes. *Journal of Econometrics*, 6, 121–134.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3), 637–654.

Canina, L., & Figlewski, S. (1993). The informational content of implied volatility. *The Review of Financial Studies*, 6(3), 659–681.

Chrisensen, B. J., & Prabhala, N. R. (1998). The relation between implied and realized volatility. *Journal of Financial Economics*, 50(2), 125–150.

Dumas, B., Fleming, J., & Whaley, R. E. (1998). Implied volatility functions: Empirical tests. *The Journal of Finance*, 53(6), 2059–2106.

Fama, E. F., & French, K. R. (1988). Dividend yields and expected stock returns. *Journal of Financial Economics*, 22, 3–25.

Heston, S. L. (1993). A closed form solution for options with stochastic volatility with applications to bond and currency option. *The Review of Financial Studies*, 6(2), 327–343.

Hogan, S., Jarrow, R., & Warachka, M. (2002). *Statistical arbitrage and tests of market efficiency*. Singapore: Singapore Management University Pre-Prints.

Karatzas, I., & Shreve, S. E. (1987). *Brownian motion and stochastic calculus*. Berlin, Germany: Springer-Verlag.

Lo, A. (2010). *Hedge funds: An analytic perspective* (Revised and Expanded ed, pp. 260). New Jersey, USA: Priceton University Press.

Loeve, M. (1977). *Probability theory II*. Berlin, Germany: Springer-Verlag.

Merton, R. C. (1973). A rational theory of option pricing. *The Bell Journal of Economics and Management Science*, 4(1), 141–183.

Mukherji, S. (2011). Are stock returns still mean-reverting? *Review of Financial Economics*, 20, 22–27.

Pole, A. (2007). *Statistical arbitrage*. Hoboken, New Jersey, USA: John Wiley & Sons Inc.

Poon, S. H., & Granger, C. W. J. (2003). Forecasting volatility in financial markets: a review. *Journal of Economic Literature*, XLI, 478–539.

Poterba, J. M, & Summers, L. H. (1988). Mean reversion in stock prices. *Journal of Financial Economics*, 22, 27–59.

Wang, Z. D., & Zheng, W. A. (2014). *High frequency trading and probability theory*. Singapore: World Scientific.