Design and analysis of momentum trading strategies

Richard J. Martin*

January 19, 2023

Abstract

We give a complete description of the third-moment (skewness) characteristics of both linear and nonlinear momentum trading strategies, the latter being understood as transformations of a normalised moving-average filter (EMA). We explain in detail why the skewness is generally positive and has a term structure.

This paper is a synthesis of two papers published by the author in RISK in 2012, with some updates and comments.

Introduction

Trend-following, or momentum, strategies have the attractive property of generating trading returns with a positively skewed statistical distribution. Consequently, they tend to hold on to their profits and are unlikely to have severe 'drawdowns'. They are very scalable and are employed in most asset classes—most traditionally in futures, where they are a favourite strategy among CTA ('commodity trading advisor') firms, but also in OTC markets—and by both buy- and sell-side practitioners.

The basic premise behind momentum is to buy what has been going up recently, and sell what has been going down. In other words, if recent returns have been positive then future ones are more likely to be positive, and similarly with negative. Systematic strategies formalise this notion by (i) measuring momentum, essentially by smoothing out recent returns to obtain a signal that is not too rapidly-varying, and (ii) having a law that turns this signal into a trading position, i.e. how may contracts or what notional to have on. Put this way, the ideas that only the finest minds can understand CTA strategies, or how the theory of statistics is of central importance to their construction, or how one needs to have been steeped in managed futures for many years to build a workable strategy, are seen to be self-serving and pretentious—a conclusion implicitly arrived at, even if not thus expressed, by other authors.

Before talking about skewness we may as well deal with the first moment, that is to say the expected return. it is important to understand that this is an entirely separate matter. Any statement about this depends on markets exhibiting momentum, i.e. serially-correlated returns. This can be ascribed to the way information is disseminated into markets or of behavioural characteristics of market participants. However, it is entirely subjective and is a matter of believing that markets will continue to behave in the way that they have done in the past. In contrast, as we analyse in detail here, even if market returns exhibit no serial correlation, during which period the strategy will produce no average return, the *trading* returns of a momentum strategy will still have positive third moment. What is interesting is that the skewness characteristic is a product of the design of

*Dept. of Mathematics, Imperial College London, South Kensington, London SW1 2AZ, UK. Email: richard.martin1@imperial.ac.uk

the strategy, whereas in long-only equities and credit the (negative) skewness is an intrinsic feature of the asset class that has to be tolerated and risk-managed.

Positive skewness results from the way positions are taken. Suppose that we look at the trading returns from one particular instrument (perhaps US Tsy bond futures) of a particular period (perhaps one week) over a long period of history. Let us group these returns by the size of the underlying position. The magnitude of P&L will typically be larger when the magnitude of the position is larger, but crucially it will typically be positive too. This is because momentum strategies typically run bigger positions *when they have already made money*—as opposed to reversion strategies that follow the opposite principle. In statistical language the full distribution of P&L is a mixture of distributions of different mean and variance: the components with a higher variance have positive mean and that is a recipe for positive skewness.

Studies on the subject have generally been empirical (for a good overview see e.g. [21] and references therein, and [2] for a general introduction to technical trading). However, there is a decent literature on quantitative aspects. The first work was by Acar [1] who derived a variety of results in discrete time using different forecasting models and also ascertained that the distribution of momentum trading returns has positive skewness. Potters & Bouchaud [20] consider a particular type of momentum strategy and derive rigorous results about its performance. Bruder et al. [4], and an even longer extension by Jusselin et al. [11], devote considerable effort to deriving moving-average filters from an underlying model. In practice, however, this seems to create more problems than it solves because the assumed model may be wrong: better, we think, is to use a convenient definition of moving-average filter, here the exponentially-weighted moving average (EMA) as it is easily calculated by recursion, and design a strategy using those. Then in 2012 two papers were published in RISK by the author, which considerably broadened the scope of the subject. Both focused on the third-moment characteristics of momentum models. The first [19] dealt with linear models, by which we mean a signal proportional to a momentum signal obtained by applying exponential smoothing to the market returns. The second [15] showed how to deal with strategies defined as nonlinear transformations of momentum signals, within the same framework. This paper is a synthesis of these two. More recently Dao [5] focuses on the connection between convexity, option-like characteristics, and momentum strategies.

In building momentum strategies, two main considerations are important. The first relates to backtesting, in other words finding what worked best in the past and assuming that it will continue to do so. Part of the problem with this is that it is too reliant on historical data: if left unchecked, it wastes an inordinate amount of time in fitting and overfitting, mainly because different models typically produce almost identical historical performance. The second relates to design: that is to say, without regard to the past, force the strategy to have certain statistical properties when the market behaves in predefined ways. One idea, which we consider in depth, is when markets are not trending (and so market returns are uncorrelated). Another is when the market does trend in a way that it did in a particular historical scenario, such as gold in the first decade of this century, or in the first 18 months from January 2019—some designs behave differently from others. There is to an extent a trade-off between these considerations: better positive skewness and better performance in certain trending scenarios may be obtained at the expense of worse average historical performance and vice versa. This necessitates subjective decision, and despite the great effort of systematic trading firms to claim that there is no discretion in the implementation of their systems—nowadays assisted by the smoke-screens of statistical theory and 'machine learning'—inevitably there must be.

An incidental conclusion from reading [5] is that the SG CTA index is very easily replicated, giving the lie to the contentions, blithely trotted out by the CTA industry, that barriers to entry are so high, that the subject can only be understood by those with years of experience, that a cohort of

PhDs are required to build strategies, and that proprietary execution algorithms are important—the last of these is clearly nonsense given the low speed at which the replicating strategy in [5] trades (see Figure 7 in that paper). In fact, rather than clever trade execution being important for momentum strategies, it is the reverse that is true: momentum is an important ingredient in trade execution, as over short time scales many financial time series exhibit momentum. Further, the relatively poor performance of the SG CTA index since the end of the Global Financial Crisis, together with the underlying simplicity of momentum trading, should make investors question whether CTAs' management fees are justified, as well as how big an asset allocation they should receive by comparison with standard investments in equities and fixed income.

A consequence of positive skewness is that the proportion of winning trades may well be negative [20]. Small trading losses are common, but occasional big gains are produced when the strategy levers itself into a trend. The longevity of trend-following funds suggests that this characteristic has served them well over the years, pointing to the conclusion that the oft-asked question "What is your fraction of winning trades?" is misleading. The link between moments and proportion of winning trades can be formalised with the Cornish–Fisher expansion, which estimates for a random variable Y,

$$\mathbf{P}(Y > \mathbf{E}[Y]) \approx \frac{1}{2} - \frac{\kappa_3}{6\sqrt{2\pi}}$$

where κ_3 is the coefficient of skewness of Y (see note[1]).

We show that the skewness of the returns depends strongly on their period[2], so that even if the one-day returns have no skewness, longer-period returns may be skewed. This may at first seem curious, and arises because successive daily *trading* returns are not independent. Thus it is possible to obtain a skewed distribution by adding non-skewed random variables, if those variables are appropriately dependent.

This paper is arranged as follows. We begin with linear strategies (§2) and give a complete exposition of 'skew theory' for them. We show how to calculate this skewness as a function of the return period, by simple application of residue calculus. The skewness of the M-period trading return depends on M: it rises to a maximum at a period proportional to the typical response-time of the trending indicator, and then drops as $M^{-1/2}$ (eqs. 15, 17, 18/19). We then test on some real data and find reasonable correspondence. Finally we analyse a particular hybrid linear strategy that is not pure momentum and derive a simple condition that ensures positive long-term skewness of returns.

An interlude on the option-like nature of trend-following ensues (§3) which is a natural consequence of §2. This has since been treated in an excellent paper by Dao et al. [5] who point out that in effect one is buying long-dated options and selling short-dated ones. This explains neatly why momentum strategies suffer badly from whipsawing, when short-date volatility is high, and perform well when the market moves steadily in one direction. The effect is important because it causes momentum strategies to hold on to most of their previous profits during the periods where they are not making money. As pointed out by Till & Eagleeye [21], this "long-option behaviour" distinguishes them from other strategies that tend to have higher Sharpe ratio, the implication being that the higher Sharpe is a form of remuneration for negative skewness. By the same token,

[1] Take for example the exponential distribution: the exact probability of exceeding the mean is $e^{-1} \approx 0.368$; the Gram-Charlier approximation gives $\frac{1}{2} - 2/(6\sqrt{2\pi}) \approx 0.367$, which is very close. On the other hand there may be considerable divergence for other distributions: for example the third moment may not even exist, or one may have a right-skewed distribution with negative third moment but positive higher odd moments. Nonetheless, the above approximation works 'better than one might suppose'.

[2] By the M-period trading return we mean the gain in P&L between date n and date $n + M$. This is to be distinguished very carefully from the *market* return which simply means the change in price of the traded asset over that period. We are assuming a discrete time model with the time increment being 1 day.

as pointed out in a different context in [17], positive skewness can result in longer drawdown times than strategies with zero skew, depending on what period of history is being considered—so time spent in drawdown is not necessarily a good measure of strategy performance.

We then move on to nonlinear strategies in §4. An arbitrary nonlinear function of several momentum factors (of different speeds) would be very difficult to analyse, so we opt for nonlinearly transforming each momentum factor first, and then the position is a weighted sum of the transformed factors. We extend the work on skewness of trading returns, studying the effect of the nonlinear transformation. This analysis is primarily a matter of algebra, and we derive new results (27;28,29) for the term structure of the skewness of trading returns. For some useful instances of the model (30,31,32) we can evaluate these expressions in closed form, making for easy computation. It turns out that the nature of the transformation is very important and can cause the positive skewness to disappear or even become negative. For example, one simple transformation is the binary construction with a position of $+1$ or -1 according as the momentum factor is positive or negative—but, as we show here, there are good reasons to suppose that it is not optimal. We conclude with some general remarks about the optimal design.

1 Basics and model setup

We have already mentioned *returns* and now formalise this notion. Simply, the return is either the change in price, or the relative change in price, the former being $X_{n+1} - X_n$ and the latter $(X_{n+1} - X_n)/X_n$ which is approximately the difference in $\ln X$ between the two time points. The latter can only be applied when prices are positive and is therefore inapplicable to asset classes such as interest-rate swaps[3], but it is the most natural definition for equities, bond futures and most commodities. We say 'most' commodities because a major upset occurred in the front WTI oil contract in April 2020 when it went negative [16]. On the other hand the former definition is most natural for interest-rate futures. We use the former definition in the ensuing algebra but this choice is not critical to the theoretical development.

We define U_{n+1} to be the return per unit volatility[4] for the asset X between time n and $n+1$, i.e.

$$U_{n+1} = \frac{X_{n+1} - X_n}{\hat{\sigma}_n}, \qquad \hat{\sigma}_n = \mathbf{E}_n[(X_{n+1} - X_n)^2]^{1/2} \tag{1}$$

with $\mathbf{E}_n$ denoting an expectation conditional on $\mathcal{F}_n$, the information known up to and including time n. The reason for dividing by $\hat{\sigma}$ is that we wish U_n to be nondimensional and appropriately normalised.

Following the general principle that one should bet a number of contracts (or contract notional) inversely proportional to the contract volatility[5], we define the position in the asset to be $\varphi_n/\hat{\sigma}_n$ at time n, where φ_n is to be a function of any or all of the U's up to and including U_n. One must

[3] As a swap at inception has zero PV. If one wants to consider the momentum of the underlying swap rate, rather than the contract PV, then this is quite sensible, though not quite the same as the contract PV includes the effects of carry and rolldown, not just the changes in par swap rate. Nonetheless one should still use absolute changes for the obvious reason that interest rates can go negative; also the absolute variation in EUR and USD rates as examples has not been strongly spot-dependent over the last 25 years or so.

[4] Also known as the risk-adjusted return.

[5] Essentially to keep a reasonably constant level of risk on: if the same position is held while the volatility rises substantially, one is likely to break one's market risk limits. More formally this follows from stochastic control theory, see e.g. [3, §14]. Writing μ for the drift and σ for the volatility of the traded asset, the optimal position always emerges as μ/σ^2 multiplied by a few factors that pertain to the exact setup (utility, etc). If the dimensionless trading signal S is representative of the risk-adjusted return μ/σ, the position is S/σ.

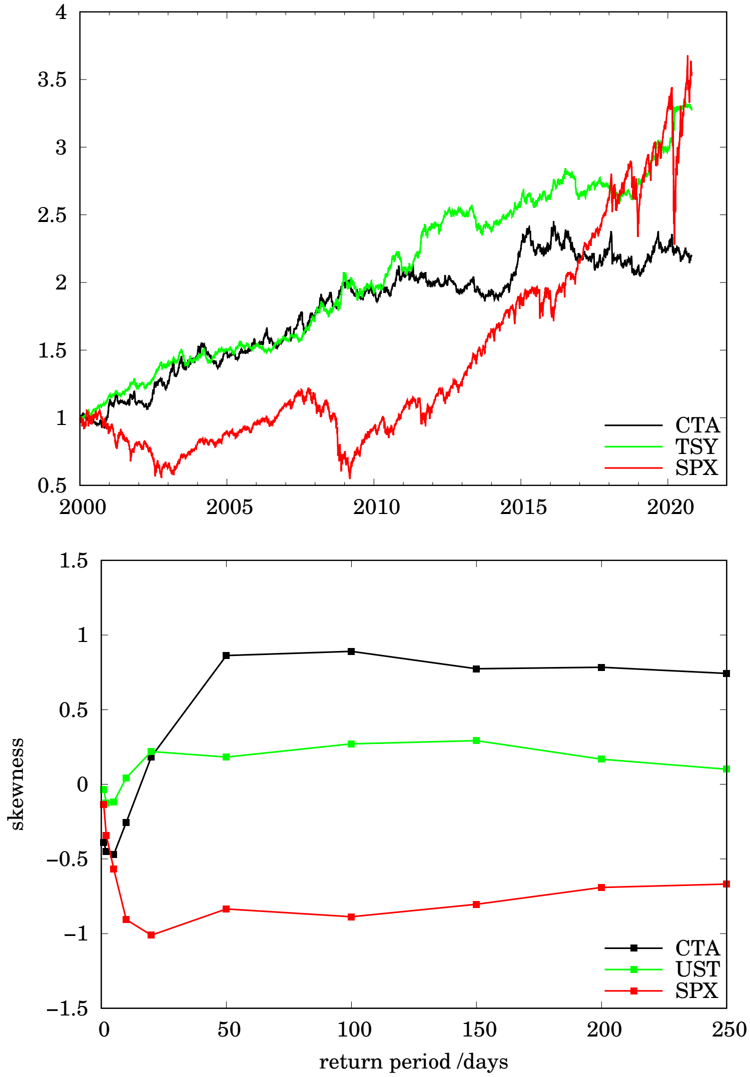


Figure 1: Performance, or 'total return', and skewness of market returns, for three indices: SG CTA index; US Treasuries (7–10y bucket); equities (SPX). Data source for top plot: Bloomberg.

have $\varphi_n \in \mathcal{F}_n$, otherwise the strategy can cheat by looking ahead. Clearly the P&L arising from the period between time n and time $n+1$ is $\varphi_n U_{n+1}$.

For much of this paper we assume that the risk-adjusted returns U_n are i.i.d.[6] and of zero mean. Thus, we are studying the behaviour of the strategy under the assumption that it is not generating any expected return (Potters & Bouchaud do the same). We also assume that U_n has zero third moment, so that its first three moments are 0,1,0. Thus although U_n has no skewness, we are about to show that the trading return may have skewness. We make no further distributional assumptions about the (U_n).

[6] The raw returns need not be, because of stochastic volatility, which is another reason for dividing off by an estimate of the stdev of the asset return.

The period-M trading return is defined as

$$Y_n^{(M)} = \sum_{k=0}^{M-1} \varphi_{n+k} U_{n+k+1}. \qquad (2)$$

The first moment of the trading return is clearly zero. The second moment is given by[7]

$$\left\langle (\varphi_0 U_1 + \cdots + \varphi_{M-1} U_M)^2 \right\rangle .$$

Let us consider this expectation on expansion as a product. The cross-terms all vanish because each contains a U_{n+1} term multiplied by a term in $\mathcal{F}_n$. This leaves the squared terms, which give simply

$$M\langle\varphi^2\rangle$$

(as $\langle U^2\rangle = 1$). The proportionality in M is a consequence of the trading returns being uncorrelated (note that we have *not* said 'independent').

The third moment of the trading return is given by

$$\left\langle (\varphi_0 U_1 + \cdots + \varphi_{M-1} U_M)^3 \right\rangle .$$

Expanding this as a product, we obtain four types of terms:

(i) $\varphi_{n_1} U_{n_1+1} \varphi_{n_2} U_{n_2+1} \varphi_{n_3} U_{n_3+1}$ with $n_1 < n_2 < n_3$;

(ii) $\varphi_n^2 U_{n+1}^2 \varphi_m U_{m+1}$ with $n < m$;

(iii) $\varphi_n^2 U_{n+1}^2 \varphi_m U_{m+1}$ with $n > m$;

(iv) $\varphi_n^3 U_{n+1}^3$.

The independence of the U's and the assumptions about their moments show that (i), (ii) and (iv) all vanish. We are therefore left with (iii), which can be written

$$3 \sum_{0 \le m < n < M} \langle \varphi_n^2 \varphi_m U_{m+1} \rangle. \qquad (3)$$

As φ_n^2 may depend on U_{m+1}, this expression is not necessarily zero, though it *must* be zero when $M = 1$ (as the sum in (3) is empty: the cause is symmetry of the market returns). The factor of 3 comes from the three ways of permuting the indices in (iii).

It is worth mentioning that if the third moment of the market returns is nonzero then this is likely to influence the skewness of the trading returns. 'Risk markets' such as equities and credit typically have returns that have positive expected return but negative skewness. Positivity of the expected return will make a momentum signal have on average a positive allocation, that is, $\mathbf{E}[\varphi_n] > 0$. The terms listed as type (iv) above will then cause the skewness of the short-term trading returns to be negative. In fact this is visible in Figure 1.

[7] $\langle\cdot\rangle$ denotes realisation average.

2 Linear strategies

2.1 Notation and definitions

We now specialise these results to linear strategies, by which we mean

$$\varphi_n = \sum_{j=0}^{\infty} a_j U_{n-j}. \tag{4}$$

Linear strategies have several advantages. They are easily constructed, for example through EMAs ($a_j \propto \alpha^j$) which can be implemented recursively: see Appendix A. They are also easily added, so that one can combine momentum of different periods (or even have negative weights on momentum of certain periods, so that one may capture counter-trending behaviour). Finally, analysis is reasonably straightforward, and the moments of the trading returns can be captured using the coefficients (a_j) alone.

We shall need the autocovariance function of the impulse response:

$$R_k^a = \sum_{j=0}^{\infty} a_j a_{j+k}, \qquad k \geq 0 \tag{5}$$

and also the system function, i.e. the z-transform of the weights:

$$A(z) = \sum_{j=0}^{\infty} a_j z^{-j}, \qquad z \in \mathbb{C}. \tag{6}$$

This is bounded for $|z| \geq 1$. A linear combination of EMAs always has a rational system function and its poles are usually a key part of the design and analysis. For a general account, refer to [9].

The simplest example is the single-EMA case, which we will call 'EMA1': $a_j = \alpha^{j+1}$, and $A(z) = \frac{\alpha}{1-\alpha z^{-1}}$. This arises as the difference between the spot price and an EMA of past prices, and is used by Potters & Bouchaud in [20]. The decay-factor α is linked to the effective period of the EMA, N, by $\alpha = 1 - N^{-1}$, so that the EMA becomes progressively slower, or more highly smoothed, as $\alpha \to 1$.

Another important example is the difference of two expressions of this form, which we call 'EMA2': $a_j = \frac{\alpha^{j+1} - \beta^{j+1}}{\alpha - \beta}$ and $A(z) = \frac{1}{(1-\alpha z^{-1})(1-\beta z^{-1})}$. This arises as the difference between two EMAs of prices, a common device in technical analysis[8]. It has less day-to-day variation than EMA1, on account of being the difference of two smoothed prices, or equivalently a double (rather than single) EMA of the returns.

It is convenient to define a class of models that we call SPRZ ('simple poles, regular at zero'). The precise conditions are: $A(z)$ bounded in $|z| > 1 - \varepsilon$ for some $\varepsilon > 0$; the only singularities to be simple poles; and $A(z)$ regular at the origin. These should be thought of as mild analytical conditions that enable the ready application of residue calculus; models with multiple poles can be understood as limiting cases of models with simple poles as the poles coalesce.

2.2 Note on continuous time formulation

This subsection may be omitted at a first reading.

[8] See for example [10, §9] and also in many online articles on tea-leaf reading, e.g. `www.stockcharts.com/school`.

The definitions of the EMA relate neatly to a continuous-time setting. For a process X_t we can define any time-invariant linear system as

$$\mathcal{K}[X]_t = \int_{-\infty}^{t} K(t-s)\,dX_s$$

where K is commonly known as the kernel. If X_t is a unit Brownian motion then the variance of $\mathcal{K}[X]_t$ is

$$\|\mathcal{K}\|^2 = \int_0^\infty K(t)^2\,dt,$$

which we call the square-norm.

An EMA1 is then the difference between X and its exponentially-weighted moving average, which is an exponential smooth of the returns:

$$X_t - \int_{-\infty}^{t} \dot{\alpha} e^{\dot{\alpha}(s-t)} X_s\,ds = \int_{-\infty}^{t} e^{\dot{\alpha}(s-t)} dX_s \tag{7}$$

and its square-norm is $1/2\dot{\alpha}$ (of course $\dot{\alpha} > 0$).

An EMA2 is the difference of two of these, with $\dot{\beta} > \dot{\alpha}$:

$$\int_{-\infty}^{t} \left(e^{\dot{\alpha}(s-t)} - e^{\dot{\beta}(s-t)}\right) dX_s \tag{8}$$

and its square-norm is $(\dot{\alpha}-\dot{\beta})^2\big/2\dot{\alpha}\dot{\beta}(\dot{\alpha}+\dot{\beta})$. The limit $\dot{\beta} \to \dot{\alpha}$ obviously only makes sense if we divide by $\dot{\beta} - \dot{\alpha}$ first, giving

$$\int_{-\infty}^{t} (t-s) e^{\dot{\alpha}(s-t)}\,dX_s \tag{9}$$

and its square-norm is $1/4\dot{\alpha}^3$. This can also be written as the EMA of X minus the double-EMA (EMA of the EMA) of X. We call this 'EMA2='.

An important practical aspect of continuous-time signals is the notion of path-length, defined for a process Y_t to be

$$\frac{1}{T}\int_0^T |dY_t|.$$

This is a concern because it relates to the rate at which money is lost in proportional transaction costs. Infinite path length results in an infinite rate of loss, unless the problem is obviated. For a Brownian motion the path length is infinite as $|dY_t| = O(\sqrt{dt})$. Now if X_t is a Brownian motion and we pass it through a linear system of kernel $\mathcal{K}$, then what is the path length of $\mathcal{K}[X]_t$? We see that

$$\int_{-\infty}^{t+dt} K(t+dt-s)\,dX_s - \int_{-\infty}^{t} K(t-s)\,dX_s = K(0)\,dX_t + dt \cdot \int_{-\infty}^{t} K'(t-s)\,dX_s$$

and the first term will generate infinite path length unless $K(0) = 0$. The second term (without the dt) is Normally distributed of zero mean and variance $\|\mathcal{K}'\|^2$, so its expected absolute value is $(2\|\mathcal{K}'\|^2/\pi)^{1/2}$. The conclusion is that **the path length is finite for EMA2, infinite for EMA1**. This does not rule out the use of EMA1, as the theory of trading under proportional transaction costs is reasonably well-established (see e.g. [18, 13, 14] and references therein), but it does suggest that in trading systems EMA2 is preferable.

Now it may be advantageous to minimise the path-length, subject to two conditions: (i) the variance of the output is to be unity (as otherwise the solution would be $K \equiv 0$), and (ii) the average lookback period is to be fixed (as otherwise we could take a normalised EMA2 and allow both speeds to tend to zero). The latter constraint can be implemented in various ways, but a felicitous one turns out to be

$$\text{Minimise} \int_0^\infty K'(t)^2\,dt \text{ s.t. } \int_0^\infty K(t)^2\,dt = 1 \text{ and } \int_0^\infty t^{-1}K(t)^2\,dt = 1/\tau \tag{10}$$

where τ, of units time, is a given parameter. One boundary condition is $K(0)=0$, and we require sensible behaviour at ∞. This is a standard type of variational calculus problem and gives rise to the ODE

$$K''(t) + (\lambda + \mu t^{-1})K(t) = 0 \tag{11}$$

where λ, μ are Lagrange multipliers. It is an easy exercise to see that $te^{-\dot{\alpha}t}$ is one solution of this[9], and so in this particular sense EMA2= is an optimal choice of momentum filter.

2.3 Second and third moments

It is immediate that the second moment of the M-period trading return is MR_0^a. For the third moment, we have to find the $U_{m+1}U_j$ term ($j \neq m+1$) in φ_n^2 in the expression (3). This is

$$2 \sum_{k=0, k\neq n-m}^{\infty} a_{n-m} a_k U_{m+1} U_{n-k+1}.$$

This now has to be multiplied by $3\varphi_m U_{m+1}$ and the expectation taken. Thus it is necessary to look for any overlap between U_{n-k+1} and φ_m, and so in the k-summation we only need terms with $k \geq n-m$, and exclude the others. The resulting expression emerges as

$$3 \sum_{0\leq m<n<M} 2 \sum_{k=n-m}^{\infty} a_{n-m-1} a_k a_{m-n+k} = 6 \sum_{0\leq m<n<M} a_{n-m-1} R_{n-m}^a = 6 \sum_{k=1}^{M-1} (M-k) a_{k-1} R_k^a. \tag{12}$$

If we understand a pure momentum, or trend-following, strategy to be one in which all the (a_j) are positive, then by (12) the trading returns must be positively skewed. By the same token a counter-trending strategy, with the a's negative, has a negatively skewed return distribution even if the market returns are symmetrical.

We can z-transform (12), i.e. multiply by z^{-M} and sum from $M=1$ to ∞, to get

$$G_3(z) = \frac{6z}{(z-1)^2} \sum_{k=1}^{\infty} a_{k-1} R_k^a z^{-k}. \tag{13}$$

From the presence of a double pole in $G_3(z)$ at $z=1$ we deduce that the third moment is asymptotically $6M\sum_{k=1}^\infty a_{k-1}R_k^a$ as $M\to\infty$. Recalling that the second moment is linear in M, we deduce that as a function of the return period M the skewness starts from zero, reaches an extremum somewhere and decays as $M^{-1/2}$.

The $M^{-1/2}$ asymptotic is, intriguingly, the same as that observed in Lévy processes. However, the origin of the skewness is completely different. With a Lévy process, it arises because the one-period returns are asymmetrical but independent. Here, they are symmetrical but not independent!

[9] Depending on the Lagrange multipliers. There an infinity of solutions, and the 'sensible' ones are of the form polynomial × exponential. The polynomials in question are related to the Laguerre polynomials.

2.4 Further analysis of the third moment

The asymptotic third moment (without the $6M$ prefactor) is $\sum_{k=1}^{\infty} a_{k-1}R_k^a$ which is also equal to

$$\frac{1}{2\pi\mathrm{i}} \oint_{|z|=1} A(z)A(z^{-1})^2 \, dz$$

(to see this, write the integrand as a product of Taylor series; the integral pulls out the z^{-1} term). This expression can be calculated using residue calculus, if we restrict ourselves to the SPRZ case, as

$$\sum_j \rho_j A(\alpha_j^{-1})^2.$$

Meanwhile the second moment is MR_0^a, and

$$R_0^a = \frac{1}{2\pi\mathrm{i}} \oint_{|z|=1} A(z)A(z^{-1})z^{-1} \, dz = A(0)^2 + \sum_j \rho_j \alpha_j^{-1} A(\alpha_j^{-1}). \tag{14}$$

Collecting the results together, we deduce that the skewness of M-period trading returns, for large M, is

$$\kappa_3^{(M)} \sim \frac{6\sum_j \rho_j A(\alpha_j^{-1})^2}{\Big(A(0)^2 + \sum_j \rho_j \alpha_j^{-1} A(\alpha_j^{-1})\Big)^{3/2} M^{1/2}}. \tag{15}$$

In the EMA1 case we immediately obtain

$$\kappa_3^{(M)} \sim \frac{6\alpha}{(1-\alpha^2)^{1/2}M^{1/2}} \sim 3\sqrt{2}\left(\frac{N}{M}\right)^{1/2}$$

where the right-hand expression is obtained by assuming that $N = (1-\alpha)^{-1}$ is not small. In the EMA2 case, the poles are at α, β and are of residue α^2, $-\beta^2$, and the result is, after a little algebra,

$$\kappa_3^{(M)} \sim \frac{6(\alpha+\beta)(1-\alpha\beta)^{1/2}}{(1-\alpha^2)^{1/2}(1-\beta^2)^{1/2}(1+\alpha\beta)^{1/2}M^{1/2}} \sim 3\sqrt{2}\left(\frac{N_\alpha + N_\beta}{M}\right)^{1/2}.$$

We can also return to (12) to get the exact third moment, not just the long-term asymptotic. To do this, we write (12) in terms of $A(z)$, as

$$\frac{6}{(2\pi\mathrm{i})^3} \sum_{k=1}^{M-1} (M-k) \sum_{j=0}^{\infty} \oint \oint \oint A(y)y^{j-1}A(z)z^{j+k-1}A(w)w^{k-2} \, dw \, dy \, dz$$

in which the contours for w- and y-integrals are of radius $1-\varepsilon$ and the contour for z is just $|z|=1$ (the need for this will presently become apparent). The j-summation and the y-integral can be done immediately (the placement of the contours causes $|yz|<1$, which is necessary for convergence of the sum; in doing the y-integral, expand the contour out to ∞ and pick up the residue at $y = 1/z$ on the way). Next do the k-summation using the identity

$$\sum_{k=1}^{m-1} (m-k)r^{k-1} \equiv \frac{r^m - 1 + m(1-r)}{(1-r)^2}$$

to arrive at

$$\frac{6}{(2\pi\mathrm{i})^2} \oint \oint A(z^{-1})A(z)A(w) \frac{(wz)^M - 1 + \overbrace{M(1-wz)}^{*}}{(1-wz)^2} w^{-1} \, dw \, dz.$$

The marked term exactly generates the large-M result we have already obtained, once the w-integral is done (again, by expanding the contour out to ∞ and picking up the residue at $w = 1/z$ on the way). The remaining part can be calculated by collapsing the w-contour around all the singularities inside the unit circle (note that no singularity arises from the $(1 - wz)^2$ term in the denominator, as $|wz| < 1$), and then collapsing the z-contour. In the SPRZ case, we finally obtain the third moment as

$$6M \sum_j \rho_j A(\alpha_j^{-1})^2 - 6A(0) \sum_j \rho_j A(\alpha_j^{-1}) - 6 \sum_{j,k} \rho_j \rho_k \alpha_j^{-1} A(\alpha_k^{-1}) \frac{1 - \alpha_j^M \alpha_k^M}{(1 - \alpha_j \alpha_k)^2}. \tag{16}$$

For EMA1, the exact expression for the skewness is therefore

$$\kappa_3^{(M)} = \frac{6\alpha}{(1 - \alpha^2)^{1/2} M^{1/2}} \left(1 - \frac{1 - \alpha^{2M}}{1 - \alpha^2} M^{-1} \right). \tag{17}$$

This rises from zero to a peak and then rolls off as $O(M^{-1/2})$ (see Figure 2a). The maximum skew[10] is roughly 2.1–2.4, and occurs for period $M \approx 1.1N$ (recall $\alpha = 1 - N^{-1}$).

For EMA2, the exact expressions for the second and third moments are

$$\mu_2^{(M)} = \frac{M(1 + \alpha\beta)}{(1 - \alpha\beta)(1 - \alpha^2)(1 - \beta^2)} \tag{18}$$

$$\begin{aligned} \mu_3^{(M)} = {} & \frac{6M(\alpha + \beta)(1 + \alpha\beta)}{(1 - \alpha\beta)(1 - \alpha^2)^2(1 - \beta^2)^2} \\ & + \frac{6\alpha^3(1 - \alpha^{2M})}{(\alpha - \beta)^2(1 - \alpha^2)^3(1 - \alpha\beta)} + \frac{6\beta^3(1 - \beta^{2M})}{(\alpha - \beta)^2(1 - \beta^2)^3(1 - \alpha\beta)} \\ & - \frac{6\alpha\beta^2(1 - \alpha^M \beta^M)}{(\alpha - \beta)^2(1 - \beta^2)(1 - \alpha\beta)^3} - \frac{6\alpha^2\beta(1 - \alpha^M \beta^M)}{(\alpha - \beta)^2(1 - \alpha^2)(1 - \alpha\beta)^3} \end{aligned} \tag{19}$$

and $\kappa_3^{(M)} = \mu_3^{(M)} / \big(\mu_2^{(M)}\big)^{3/2}$ as usual. This is qualitatively similar to EMA1 (see Figure 2b). The maximum skew is around 2.1, and occurs at period $M \approx 1.7(N_\alpha + N_\beta)$, provided N_α and N_β are not too far apart. In the extreme case where either of the N's is equal to 1, we recover EMA1. The limit $\beta \to \alpha$ is well-behaved, but algebraically messy and omitted here.

In essence, (16) telescopes the various geometric series that are implicit in the calculation of (12), and allows it to be done with an amount of computational effort independent of M.

2.5 Empirical results

For a demonstration using real data we use two datasets: the CHFUSD futures and the S&P500 futures[11]. For risk-adjusting the returns we use a 20-day EMA of squared price changes to estimate the volatility ($\hat{\sigma}_n$ in the definition of U_n). We are using an EMA2 with $N = 20, 40$.

It is worth recalling the assumptions that we made in deriving our formulae: (i) independence of the risk-adjusted returns (U_n); (ii) symmetry of their distribution up to the third moment. In practice the first clearly does not hold, because it implies that momentum strategies do not generate positive expected return, whereas the evidence is that on average they do. That means that when we examine real data, the observed skewness of returns may well not equal the theoretical result,

[10] Unless N is small we can approximate (17) as $(3/\sqrt{2x^3})(e^{-2x} - 1 + 2x)$, with $x = M/N$; the maximum of this function is ≈ 2.41 and occurs at $x \approx 1.07$.

[11] Bloomberg: `SF1 Curncy` and `SP1 Index`. These are the front contracts, rolled 10 days before expiry to create generic series. Data range: 01-Jan-90 to 31-Dec-09.

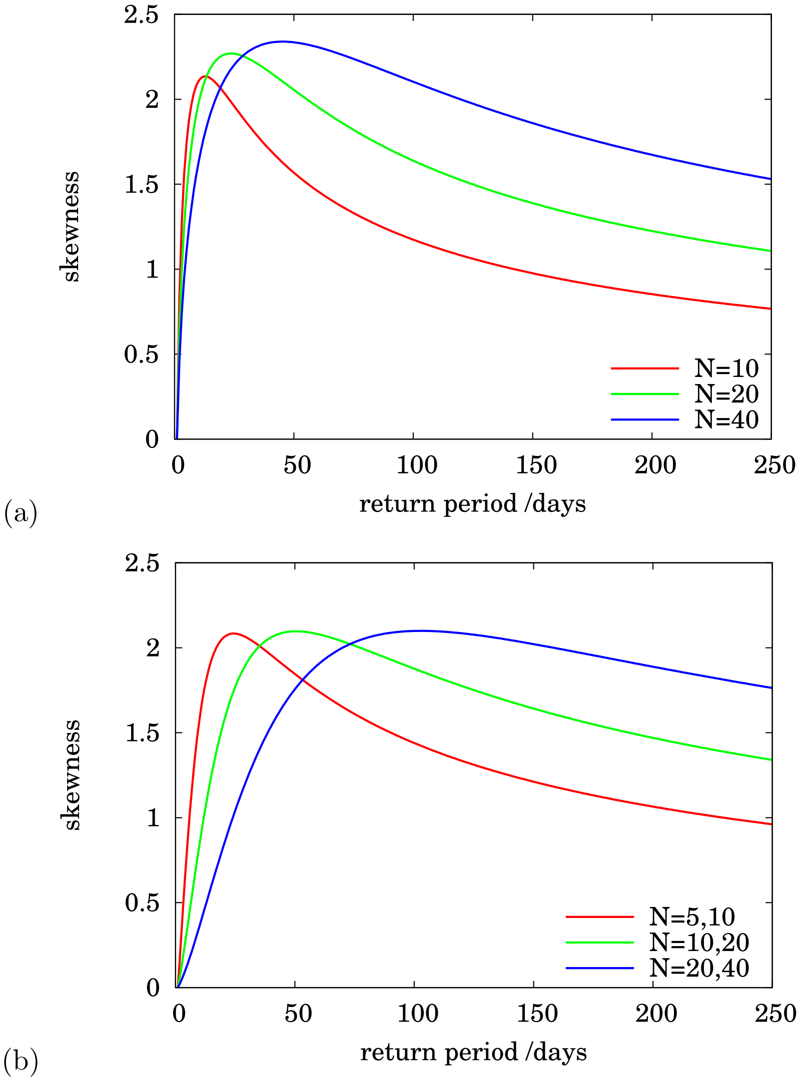


Figure 2: Skewness of trading returns, as a function of period, for (a) EMA1 type model, (b) EMA2 type model. Note the characteristic shape.

by virtue of the mean being different. We therefore plot the central skew (third central[12] moment divided by $\frac{3}{2}$ power of the second central moment—the usual definition) and also the 'non-central' skew (third moment about zero divided by $\frac{3}{2}$ power of the second moment about zero). If the effect of trending is to generate a slightly positive expected return but keep the other moments roughly equal, then the non-central skew will be fractionally higher than the central skew. As to (ii), we know that equity markets occasionally have very negative returns.

Figure 3 shows the results for the two markets, superimposing also the theoretical result from Figure 2b. In spite of the deficiencies in the modelling assumptions the agreement is not bad and the general shape is right. The short-term skewness for the equity market is nonzero because of the asymmetry of the market returns; the higher long-term skewness is best ascribed to the particularly good trending behaviour in the mid-1990s generating high trading returns. The skewness of the

[12] Central moment = moment about the mean.

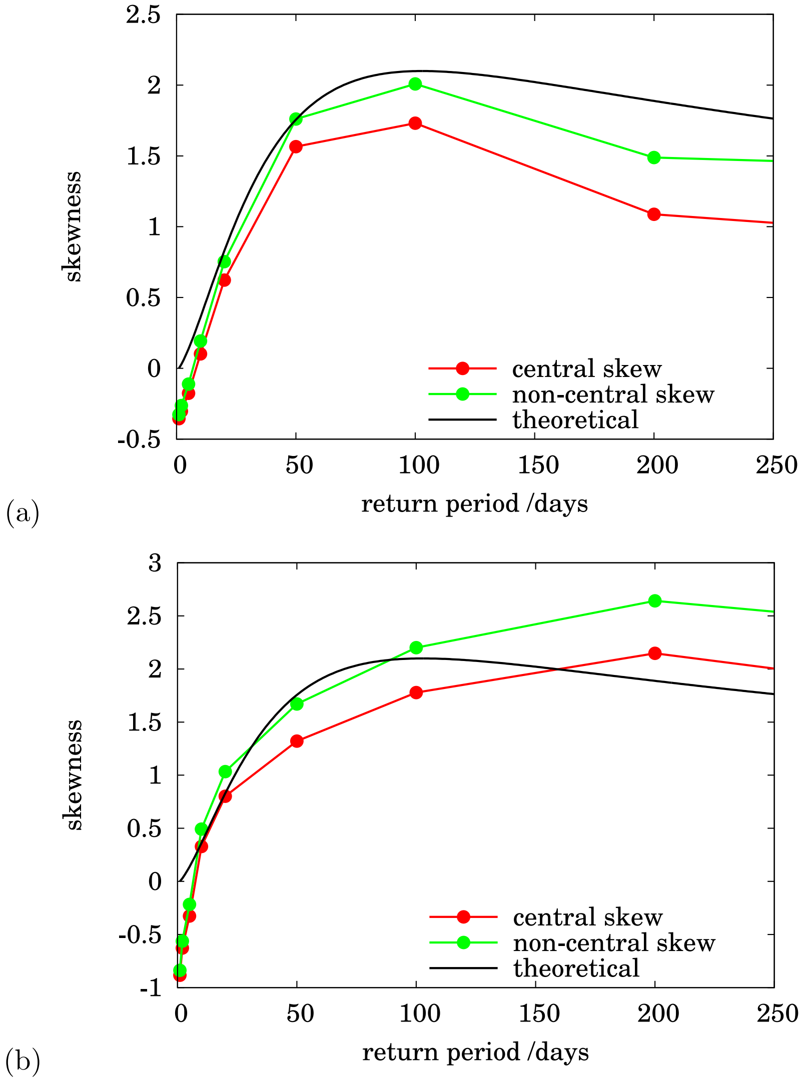


Figure 3: Skewness of trading returns, as a function of period, for (a) CHFUSD, (b) S&P500 futures; theoretical result also shown. $N = 20, 40$.

trading returns is far higher than that of the underlying markets (i.e. of the U_n's): the latter is (to within 0.1) typically about 0.0 for CHFUSD and -0.2 for S&P500. This shows that the skewness comes entirely from the momentum strategy. The Gram-Charlier formula for the probability of exceeding zero is modified to $\Phi(\mathfrak{r}) - \kappa_3/(6\sqrt{2\pi})$ in the presence of nonzero first cumulant (expected return); here $\mathfrak{r} = \kappa_1 \big/ \kappa_2^{1/2}$ is the Sharpe ratio and Φ is the Normal c.d.f. For horizons M in the range 100–200 days the Sharpe ratio of each is roughly $+0.2$ and the skewness is around 2, so this gives the probability of exceeding zero as about 0.45, which corresponds well with the empirical value—note that it is less than one-half.

2.6 Hybrid linear models

Suppose that a strategy has a trend-following and a counter-trending characteristic, as would happen if its weights were obtained from a linear combination of EMA2's, with opposite signs.

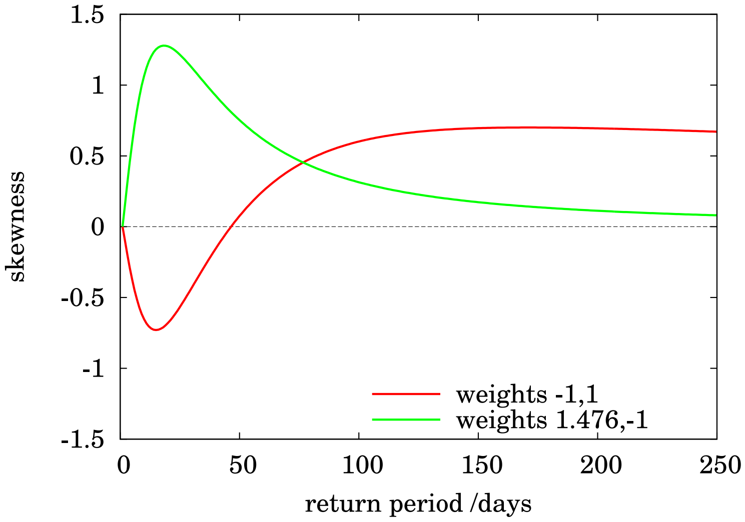


Figure 4: Skewness of trading returns, as a function of period, for hybrid model with both trending and counter-trending behaviour, in two cases.

It may be desirable to ensure that the long-term skewness remains positive, as this is associated with longevity of the strategy. There are two situations in which this arises. In what is basically a trend-following strategy, it is desired:

(i) to make small bets on short-term reversion without this upsetting the behaviour if a longer-term trend occurs;

(ii) to make a small bet against very long-term trends on the supposition that what goes up must eventually come down (or vice versa), provided this bet is not too large.

In the first case the weights on most recent returns will be negative; in the second, it is the weights on the distant past that will be negative. The idea is to make sure that they are not too negative, in a sense to be made precise.

We have a model of the form

$$A(z) = \frac{\lambda_F(\alpha_F - \beta_F)}{(1-\alpha_F z^{-1})(1-\beta_F z^{-1})} + \frac{\lambda_S(\alpha_S - \beta_S)}{(1-\alpha_S z^{-1})(1-\beta_S z^{-1})}$$

where λ_F and λ_S are the multipliers on the fast and slow components. Positive asymptotic skewness is ensured by (15):

$$\sum_j \rho_j A(\alpha_j^{-1})^2 > 0 \tag{20}$$

where there are now four poles $\alpha_1 = \alpha_F$, $\alpha_2 = \beta_F$, $\alpha_3 = \alpha_S$, $\alpha_4 = \beta_S$. Thus

$$\rho_1 = \alpha_F^2 \lambda_F, \qquad A(\alpha_1^{-1}) = \frac{\lambda_F(\alpha_F - \beta_F)}{(1-\alpha_F^2)(1-\beta_F\alpha_F)} + \frac{\lambda_S(\alpha_S - \beta_S)}{(1-\alpha_S\alpha_F)(1-\beta_S\alpha_F)}$$

and similarly for the other three. The LHS of (20) is a homogeneous cubic in λ_F, λ_S, which will factorise as

$$\mathcal{P}(\lambda_F, \lambda_S) = (\lambda_F - \zeta_1\lambda_S)(\lambda_F - \zeta_2\lambda_S)(\lambda_F - \zeta_3\lambda_S)$$

where the ζ's are functions of the four poles. It is possible to identify the coefficients of λ_F^3, $\lambda_F^2\lambda_S$, $\lambda_F\lambda_S^2$, λ_S^3 as functions of the poles, then evaluate them and factorise the cubic by the Cardano-Tartaglia formula. However for practical purposes one might just as well write a numerical routine for LHS(20) and find the roots ζ_i numerically. One root ζ_1 has to be real, and the other two are likely to be complex because we expect $\mathcal{P}$ to be strictly increasing in λ_F and in λ_S: raising either weight should enhance the trending behaviour and hence the asymptotic skewness.

As a particular example, let $N_{\alpha_F} = 5$, $N_{\beta_F} = 10$, $N_{\alpha_S} = 20$, $N_{\beta_S} = 40$. Then $\zeta_1 \approx -1.476$, and the condition for positive asymptotic skewness is simply[13]

$$\lambda_F + 1.476\,\lambda_S > 0.$$

This being so, it is easily incorporated into an optimisation as a 'style constraint'. Figure 4 shows the results for two examples, (i) $\lambda_F = -1$, $\lambda_S = 1$, so the short-term behaviour is counter-trending and generates negative skewness; (ii) the critical case $\lambda_F = 1.476$, $\lambda_S = -1$, where now just enough long-term counter-trending behaviour is added to make the asymptotic skew zero at leading order. These exemplify the cases (i), (ii) discussed above. The results were obtained using (16) again, which is not laborious despite there being four poles (so that the double summation has sixteen terms): it is preferable to Monte Carlo simulation, which even with a few hundred thousand simulations generates noticeable uncertainty.

3 Option-like nature of trend-following

As pointed out in [21], trend-following strategies are often thought to have a long-option-type payoff on account of the positive skewness. For linear strategies this can be formalised as follows. The M-period trading return is

$$Y_n^{(M)} = \mathbf{u}'\mathbf{\Gamma}\mathbf{u}, \qquad \mathbf{u} = \begin{bmatrix} U_{n+M} & U_{n+M-1} & \cdots \end{bmatrix}'$$

where the symmetric matrix $\mathbf{\Gamma}$ is given by

$$\mathbf{\Gamma} = \frac{1}{2}\begin{bmatrix} 0 & a_0 & a_1 & a_2 & \cdots & & \\ a_0 & \ddots & \ddots & \ddots & \ddots & \ddots & \\ a_1 & \ddots & 0 & a_0 & a_1 & a_2 & \cdots \\ a_2 & \ddots & a_0 & 0 & 0 & 0 & \cdots \\ \vdots & \ddots & a_1 & 0 & 0 & 0 & \cdots \\ & \ddots & a_2 & 0 & 0 & 0 & \cdots \\ & & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{array}{l} \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} M \text{ rows} \\ \\ \\ \\ \\ \end{array} . \tag{21}$$

The moments of $Y_n^{(M)}$ relate to the spectrum of $\mathbf{\Gamma}$, and direct calculation reveals

$$\left\langle Y_n^{(M)} \right\rangle = \operatorname{tr}(\mathbf{\Gamma}) = 0, \quad \left\langle \left(Y_n^{(M)}\right)^2 \right\rangle = 2\operatorname{tr}(\mathbf{\Gamma}^2), \quad \left\langle \left(Y_n^{(M)}\right)^3 \right\rangle = 8\operatorname{tr}(\mathbf{\Gamma}^3).$$

Writing $\mathbf{\Gamma}$ in terms of its eigenvalues γ_j and normalised eigenvectors $\mathbf{e}_j$, we have an expression that is a weighted sum (weights adding to zero) of 'orthogonal quadratic bets', i.e. squared linear combinations of returns, which are like straddle payoffs but have constant convexity:

$$Y_n^{(M)} = \sum_j \gamma_j (\mathbf{e}_j \cdot \mathbf{u})^2. \tag{22}$$

[13] Because the other two terms in $\mathcal{P}$ multiply to give a quadratic that is always positive, and hence of no consequence.

Now $\mathrm{tr}(\mathbf{\Gamma}^r) = \sum_j \gamma_j^r$, so the moments of $Y_n^{(M)}$ relate to the moments of the eigenvalue distribution. It is easy to see the rank of $\mathbf{\Gamma}$ is $\leq 2M$ (and is $M+1$ in the EMA1 case as then the rows after the Mth are linear multiples of each other), which limits the number of nonzero eigenvalues to $2M$. The interpretation of all this is that a positively skewed strategy has a small number of large positive eigenvalues and a larger number of smaller negative ones. This generates a small number of large positive-convexity bets and a larger number of smaller negative-convexity bets, which is where the positive skewness comes from. Dao et al. [5] make the same point, but emphasise the important point that the positive-convexity bets are long-dated options and the negative-convexity bets short-dated ones. Thus in situations where the long-term volatility is elevated and the short-term volatility is low, momentum strategies work well.

3.1 *Full distribution of trading returns

This subsection may be omitted at a first reading.

If we want to know the full distribution of trading returns, we need to make an assumption about the full distribution of the market returns, whereas until now we have only used the first three moments.

We can use the ideas of the previous section to compute the full distribution of trading returns, exactly as is done by Acar [1, Ch.3] using generating functions. The moment-generating function of the M-period return is

$$F_M(s) := \left\langle \exp(sY^{(M)}) \right\rangle = \left\langle \exp\left(s(\varphi_0 U_1 + \cdots + \varphi_{M-1} U_M) \right) \right\rangle .$$

Let us assume that the (U_n) are Normally distributed. Then for a linear model

$$F_M(s) = [\det(\mathbf{I} - 2s\mathbf{\Gamma})]^{-1/2}$$

with $\mathbf{\Gamma}$ as above. (See e.g. [7] for details on quadratic transformations of Normal variables.)

Before proceeding further we should note that the above expression is not very helpful because it requires the manipulation of $\mathbf{\Gamma}$ which is an infinite matrix. Let us therefore evaluate *ab initio* the expression

$$\left\langle \exp\left(s(\varphi_0 U_1 + \cdots + \varphi_{M-1} U_M) \right) \right\rangle .$$

Conditioning on $(U_1, \ldots, U_M)$, effectively fixing those values, we have (inside the exponential) a linear combination of $U_0, U_{-1}, \ldots$, added to another expression that is a function of $(U_j)_{j=1}^M$ only. In the first part the coefficients are

$$\begin{array}{rl} U_0 : & s(a_0 U_1 + a_1 U_2 + \cdots + a_{M-1} U_M) \\ U_{-1} : & s(a_1 U_1 + a_2 U_2 + \cdots + a_{M-2} U_M) \end{array}$$

and so on. These variables can then be integrated out to give the expression

$$\exp\left\{ \tfrac{1}{2} s^2 \sum_{l=-1}^{\infty} \left(\sum_{k=1}^{M} a_{k+l} U_k \right)^2 \right\} = \exp\left\{ \tfrac{1}{2} s^2 \sum_{l=-1}^{\infty} \sum_{j,k=1}^{M} a_{j+l} a_{k+l} U_j U_k \right\}.$$

The other part of the expression depends on the U's through pairwise products, i.e. $U_1 U_2$, etc, and is easily seen to be

$$\exp\left\{ s \sum_{1 \leq j < k \leq M} a_{k-j-1} U_j U_k \right\}.$$

Next we multiply these two expressions to obtain

$$F_M(s) = \left\langle \exp \mathbf{u}'\widehat{\mathbf{G}}(s)\mathbf{u} \right\rangle, \qquad \mathbf{u} = \begin{bmatrix} U_1 & U_2 & \cdots & U_M \end{bmatrix}',$$

with

$$\widehat{G}_{jk}(s) = \tfrac{1}{2}s^2 \sum_{l=-1}^{\infty} a_{j+l}a_{k+l} + \tfrac{1}{2}s\mathbf{1}_{j\neq k}a_{|k-j|-1}. \tag{23}$$

Finally integrate $\mathbf{u}$ out, to get the more manageable

$$F_M(s) = [\det(\mathbf{I} - 2\widehat{\mathbf{G}}(s))]^{-1/2} \tag{24}$$

instead—the dimension of $\widehat{\mathbf{G}}$ is just M. Again the infinite summation in the expression for $\widehat{G}_{jk}$ can be done easily enough by residue calculus in the SPRZ case.

The moments and cumulants can now be obtained by performing a Taylor series of, respectively, $F_M(s)$ and $\log F_M(s)$ around the origin, using standard identities for expanding the determinants. The derivation is laborious, however. To obtain an approximation to the distribution of $Y^{(M)}$, we may use Fourier transform inversion or, as discussed in [7], saddlepoint methods, which work well on this problem (see also [12] for a general discussion).

4 Nonlinear strategies

As in the first part of the paper we write the position as $\varphi_n/\hat{\sigma}_n$. Clearly the P&L arising from the period between time n and time $n+1$ is $\varphi_n U_{n+1}$. We define a *momentum factor* V_n to be a moving average of the U's, with weights $(a_j)_{j=0}^{\infty}$. The position, however, will no longer be simply proportional to V_n, but instead transformed using a nonlinear 'activation[14] function' ψ:

$$V_n = \sum_{j=0}^{\infty} a_j U_{n-j}; \qquad \varphi_n = \psi(V_n). \tag{25}$$

We use the autocovariance function R_k^a as before but now for convenience we stipulate that V_n have unit variance, which is to say $R_0^a = 1$. Then $|R_k^a| \leq 1$ (by Cauchy-Schwarz), and the inequality will be strict in all practical examples. Importantly, when designing the activation function ψ we know that the typical scale of variation of its input is unity.

In the EMA1 model, we have

$$a_j = (1-\alpha^2)^{1/2}\alpha^j, \qquad R_k^a = \alpha^k,$$

differing trivially from what we had in §2 by a scaling factor so that $R_0^a = 1$. For EMA2,

$$a_j = \left(\frac{(1-\alpha^2)(1-\beta^2)(1-\alpha\beta)}{1+\alpha\beta} \right)^{1/2} \frac{\alpha^{j+1} - \beta^{j+1}}{\alpha - \beta}, \qquad R_k^a = \frac{\alpha^{k+1}(1-\beta^2) - \beta^{k+1}(1-\alpha^2)}{(\alpha-\beta)(1+\alpha\beta)}.$$

If $\beta = \alpha$ then this is well-defined and is simply a double-EMA of the returns.

This completes our description of the model setup: in summary, the position taken is a nonlinear transformation of smoothed vol-adjusted returns, divided by the volatility of the underlying. It is specified by the EMA parameter(s) and the function ψ. A general system can then be built by taking a linear combination of these models with different speeds.

[14] Following neural network parlance.

4.1 Skewness of nonlinear models

We make the same assumptions as before about the nature of the risk-adjusted returns U_n, that is they are i.i.d. and their first three moments are 0,1,0. Then the second moment of the trading return is

$$\left\langle \big(Y_n^{(M)}\big)^2 \right\rangle = M\langle\varphi^2\rangle \tag{26}$$

and the third moment is

$$\left\langle \big(Y_n^{(M)}\big)^3 \right\rangle = 3\sum_{k=1}^{M-1}(M-k)H_k, \qquad H_{n-m} = \langle \varphi_n^2 \varphi_m U_{m+1}\rangle. \tag{27}$$

Whereas in §2 we evaluated this summation directly, we can no longer do this and therefore take a slightly more roundabout route. Note first that the triple $(\varphi_n^2, \varphi_m, U_{m+1})$ is a simple function of the triple $(V_n, V_m, U_{m+1}) = (Z_1, Z_2, Z_3)$ say, which has a trivariate Normal distribution with covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & R^a_{n-m} & a_{n-m-1} \\ R^a_{n-m} & 1 & 0 \\ a_{n-m-1} & 0 & 1 \end{bmatrix}.$$

The determinant and inverse of $\mathbf{\Sigma}$ are

$$\Delta = 1 - a^2_{n-m-1} - R^2_{n-m}, \qquad \mathbf{\Sigma}^{-1} = \frac{1}{\Delta}\begin{bmatrix} 1 & -R^a_{n-m} & -a_{n-m-1} \\ -R^a_{n-m} & 1-a^2_{n-m-1} & a_{n-m-1}R^a_{n-m} \\ -a_{n-m-1} & a_{n-m-1}R^a_{n-m} & 1-(R^a_{n-m})^2 \end{bmatrix}.$$

Writing the expectation $\langle \varphi_n^2 \varphi_m U_{m+1}\rangle$ as an integral we get[15]

$$H_{n-m} = \frac{1}{(2\pi)^{3/2}\Delta^{1/2}} \iiint_{-\infty}^{\infty} \exp(-\tfrac{1}{2}\mathbf{z}'\mathbf{\Sigma}^{-1}\mathbf{z})\psi(z_1)^2\psi(z_2)z_3\, dz_1\, dz_2\, dz_3$$

and then integrate over z_3 (effectively, integrating U_{m+1} out). After a fair amount of algebra, the whole lot tidies up to give an expression as an expectation over the distribution of (Z_1, Z_2) which is bivariate Normal with unit marginals and correlation $\rho = R^a_k$:

$$\begin{aligned} H_k &= 2a_{k-1}\left\langle \psi(Z_1)^2\psi(Z_2)\frac{Z_1-\rho Z_2}{2(1-\rho^2)}\right\rangle\Big|_{\rho=R^a_k} && (28)\\ &= 2a_{k-1}\left\langle \psi(Z_1)\psi'(Z_1)\psi(Z_2)\right\rangle\big|_{\rho=R^a_k}, && (29)\end{aligned}$$

where the last line follows from integrating by parts w.r.t. Z_1, assuming differentiability of ψ.

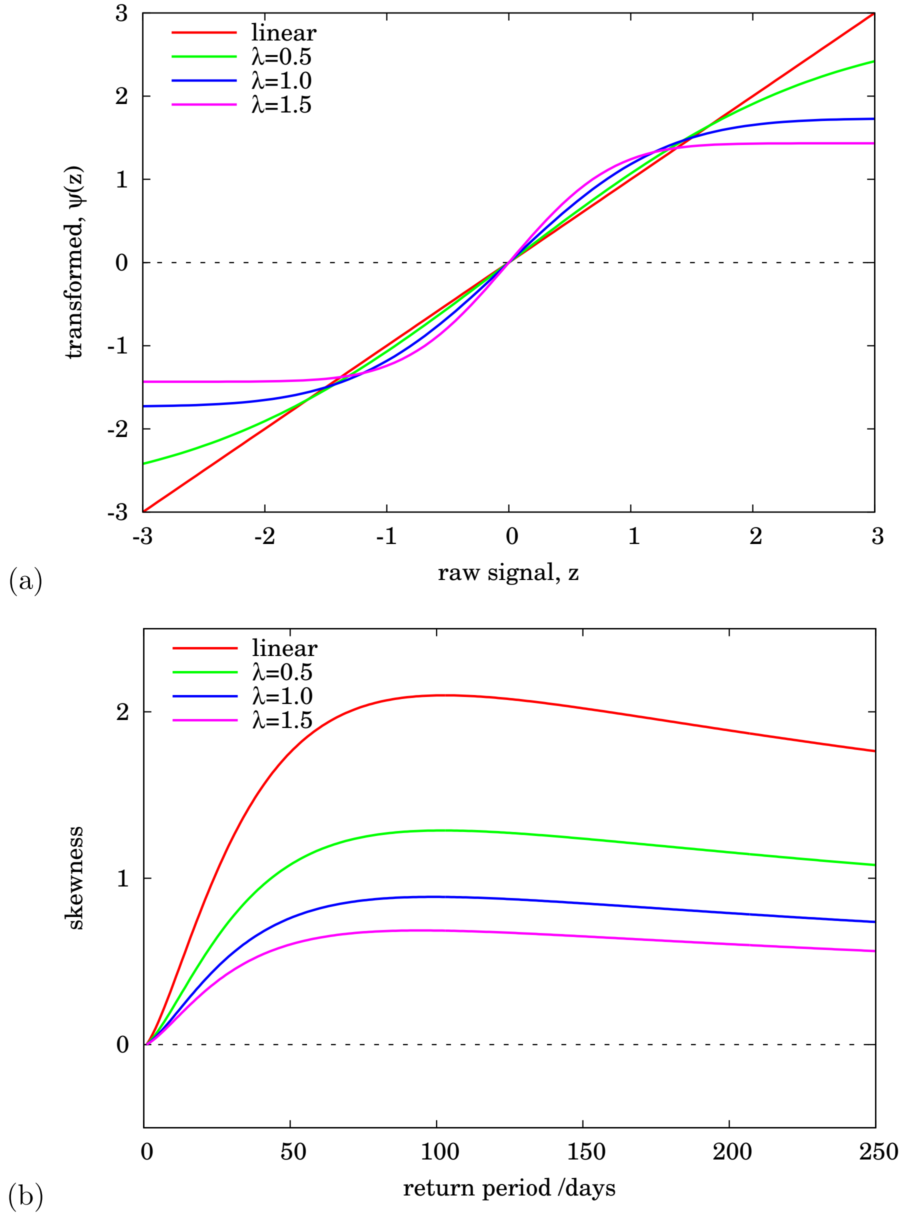
In the linear case, $\psi(z) = z$, we have $H_k = 2a_{k-1}R^a_k$, and hence the third moment is $6\sum_{k=1}^M (M-k)a_{k-1}R^a_k$ as we saw earlier. When ψ is not linear, we have to calculate H_k for each k and then do the sum by explicit calculation. That said, it is possible to find functions ψ of the 'right shape' for which the double integral implicit in (29) can be done in closed form. We discuss some next. The integration is done over Z_2 first conditionally on Z_1, so that $Z_2 \sim N(\rho Z_1, 1-\rho^2)$, and then Z_1 is integrated out. Some helpful identities used in the calculations are given in the Appendix. Furthermore, the summations can be evaluated by recursion in M. In detail, writing $\Theta_M = 3\sum_{k=1}^{M-1}(M-k)H_k$ and $S_M = \sum_{k=1}^{M-1}H_k$, we initialise $S_1 = 0$, $\Theta_1 = 0$, and for $M \geq 2$ we have $S_M = S_{M-1} + H_{M-1}$ and $\Theta_M = \Theta_{M-1} + 3S_M$.

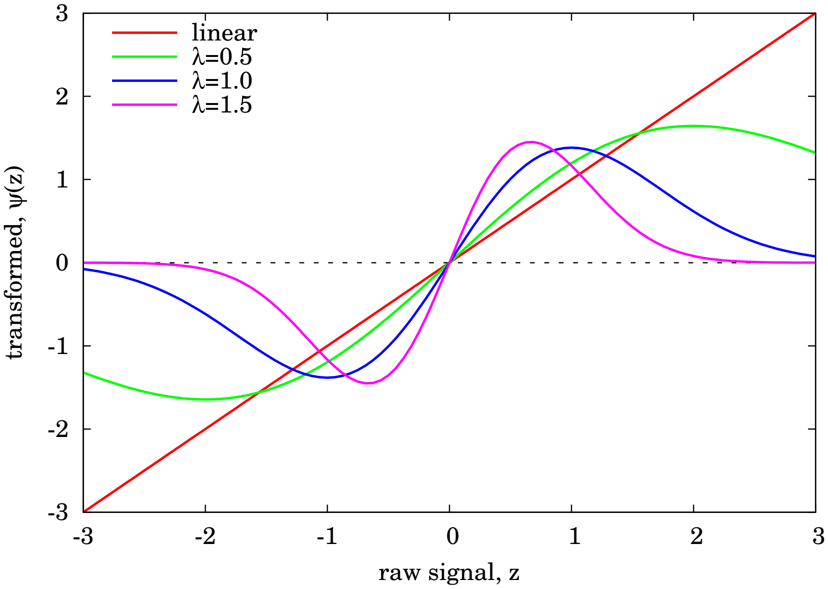
Remember that in obtaining the skewness we need to divide the third moment by the $\frac{3}{2}$ power of the second, so it is convenient to deal with functions normalised to $\left\langle \psi(Z)^2 \right\rangle = 1$.

[15] Notation $\mathbf{z}$ just means the column vector $(z_1, z_2, z_3)'$.

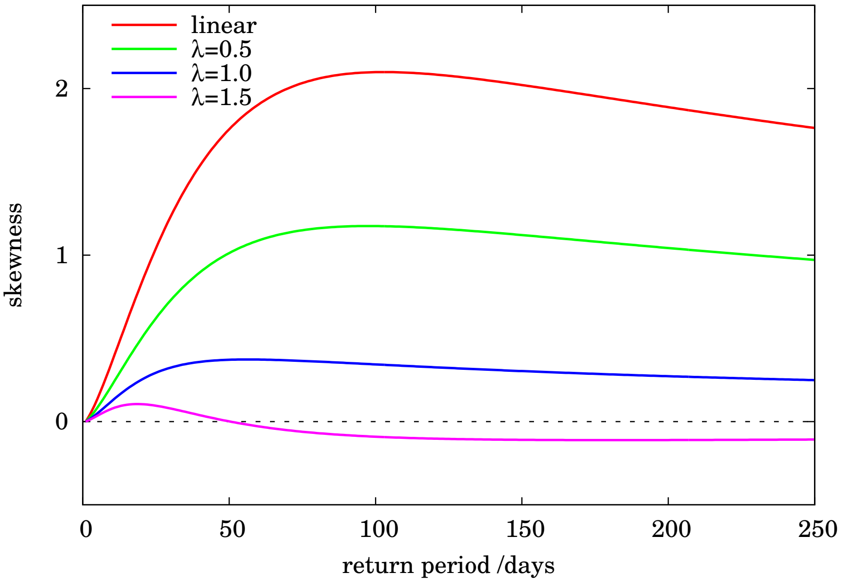
4.2 Examples

We discuss three functions, drawn in Figure 5.




linear
λ=0.5
λ=1.0
λ=1.5
3
2
1
0
-1
-2
-3
transformed, ψ(z)
-3
-2
-1
0
1
2
3
raw signal, z

(c)


linear
λ=0.5
λ=1.0
λ=1.5
2
1
0
skewness
0
50
100
150
200
250
return period /days

(d)

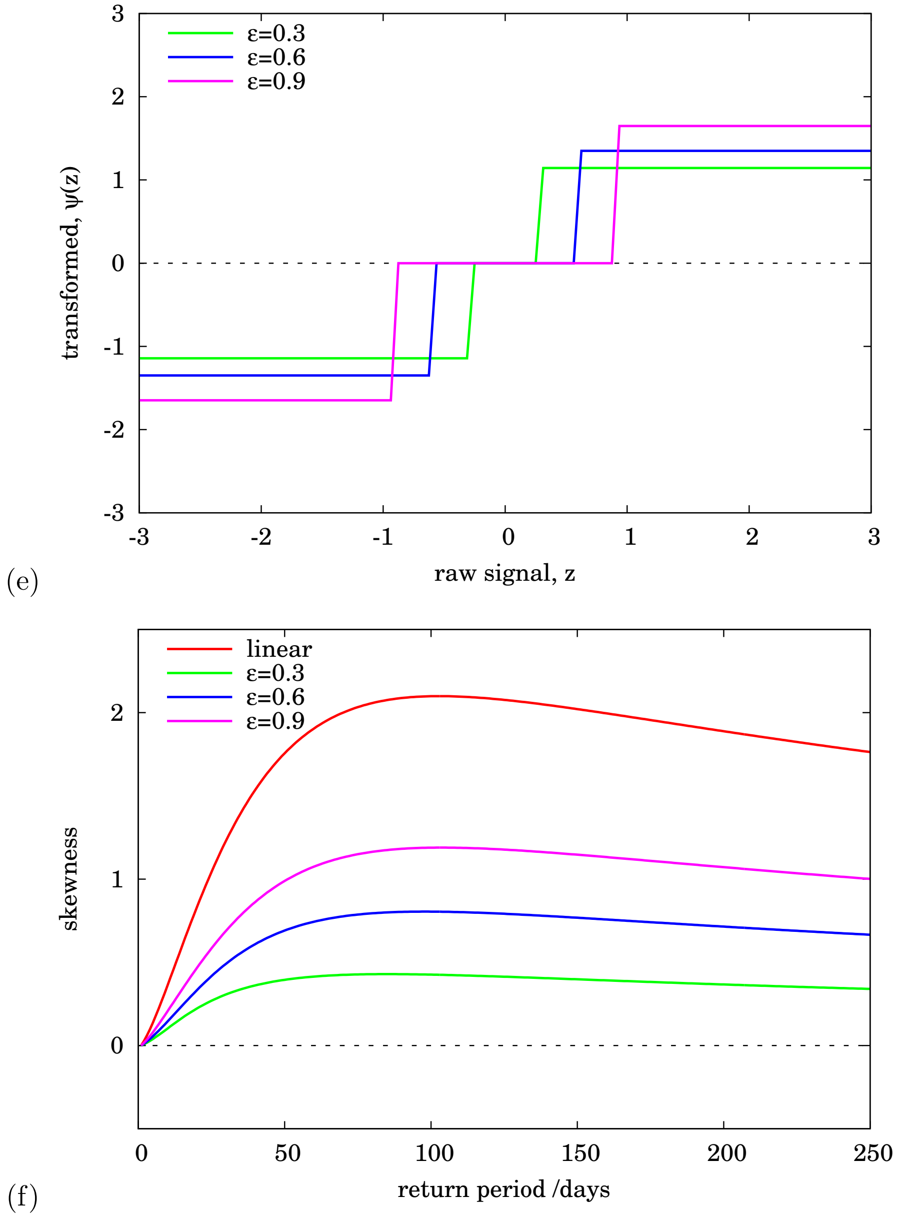


Figure 5: Activation functions and their associated term structure of skewness: (a,b) sigmoids with $\lambda = 1.0, 2.0$; (c,d) reverting sigmoids with $\lambda = 0.5, 1.0, 1.5$; (e,f) double-step with $\varepsilon = 0.3, 0.6, 0.9$. Using EMA2 type momentum signal throughout, with $N = 20, 40$.

4.2.1 Simple sigmoid, $\psi(z) = c_\lambda \cdot \big(2\Phi(\lambda z) - 1\big)$

In effect, this caps the position when the magnitude of the the momentum signal is large. We have

$$H_k = 2a_{k-1} c_\lambda^3 \frac{(2/\pi)^{3/2}\lambda}{\sqrt{1+\lambda^2}} \arctan\left(\frac{\lambda^2 \rho / \sqrt{1+\lambda^2}}{\sqrt{1 + 3\lambda^2 + 2(1-\rho^2)\lambda^4}} \right) \Bigg|_{\rho = R_k^a} \tag{30}$$

and for normalisation we require $c_\lambda = \big(\frac{2}{\pi} \arctan \frac{\lambda^2}{\sqrt{1+2\lambda^2}}\big)^{-1/2}$. We obtain in the limit $\lambda \to 0$ that $H_k = 2a_{k-1} R_k^a$ which is as expected the linear result.

4.2.2 Reverting sigmoid, $\psi(z) = c_\lambda \cdot z e^{-\lambda^2 z^2/2}$

The behaviour of this one is more nonlinear in the sense that it begins *reducing* the position when the momentum gets too high, ultimately to zero if the momentum is strong enough. The rationale for this is that a very strong trend might be more susceptible to reversing (market overbought/oversold), justifying a reduction in position. The maximum positions are held when $z = \pm\lambda^{-1}$. We have

$$H_k = 2a_{k-1} c_\lambda^3 \frac{\rho\big(1 - (1-\rho^2)\lambda^4\big)}{\big(1 + 3\lambda^2 + 2(1-\rho^2)\lambda^4\big)^{5/2}} \Bigg|_{\rho = R_k^a} \tag{31}$$

and for normalisation we require $c_\lambda = (1+2\lambda^2)^{3/4}$. Notice that H_k becomes negative if λ is high enough: this is not surprising because $\psi'(Z)$ is negative for $|Z| > \lambda^{-1}$, wherein the model is betting against the trend. As expected $\lambda \to 0$ gives back the linear result again.

4.2.3 Double-step, $\psi(z) = c_\lambda \cdot (\mathbf{1}_{z>\varepsilon} - \mathbf{1}_{z<-\varepsilon})$

This is $+1$ if the momentum is positive enough, -1 if negative enough, and zero at intermediate levels, with the width of the 'dead zone' being $2\varepsilon > 0$. This is similar to the one considered by Potters & Bouchaud. We have[16]

$$H_k = 2a_{k-1} c_\lambda^3 \phi(\varepsilon) \left(\Phi\left(\varepsilon \sqrt{\frac{1+\rho}{1-\rho}} \right) - \Phi\left(\varepsilon \sqrt{\frac{1-\rho}{1+\rho}} \right) \right) \Bigg|_{\rho = R_k^a} \tag{32}$$

and for normalisation we require $c_\lambda = \big(2\Phi(-\varepsilon)\big)^{-1/2}$.

From this it can be seen that as $\varepsilon \to 0$, creating a binary response $\psi(z) = \operatorname{sgn}(z)$, the skewness vanishes[17]. One way of understanding this is to see that as the position is always of magnitude 1, it does not have the characteristic—associated with positive momentum—of increasing the position when the P&L is positive. A less precise explanation is that during periods in which the market is not trending, the strategy loses money rather quickly because it is buying and selling the same size of position as it holds when a trend has been detected. By contrast, the other two (sigmoidal) functions that we have just examined only trade a small size until a trend is established, resulting in P&L distribution with more, but smaller, losses, and fewer, but bigger, gains: that is where their positive skewness comes from. A recent piece on FX strategy [6] makes this point somewhat

[16] Do not use these results for $\varepsilon < 0$.

[17] This can also be seen from the sigmoidal case when $\lambda \to \infty$, as the argument in the arctan() term goes to zero. The result can also be seen directly from (28), because $\psi(Z_1)^2 = 1$ a.s. and and $Z_1 - \rho Z_2$ is independent of Z_2, so the expectation decouples into a product of two expectations each of which is zero. Technical point: The behaviour as $\rho \to \pm 1$ is awkward. In fact, this limit is irregular and the result as $\rho \to 1$ is not the same as that for $\rho = 1$; similarly for -1. However, these two values of ρ cannot occur in our problem.

tangentially. The authors use the simple step function and note that the performance is excellent when the market is trending but very poor otherwise. In the light of what we have just said, we are not surprised by this observation.

As ε is raised, notice that the skewness rises without limit, which seems rather good: however, if ε is too high then the algorithm hardly ever trades, so practicalities dictate $\varepsilon \lesssim 1.5$.

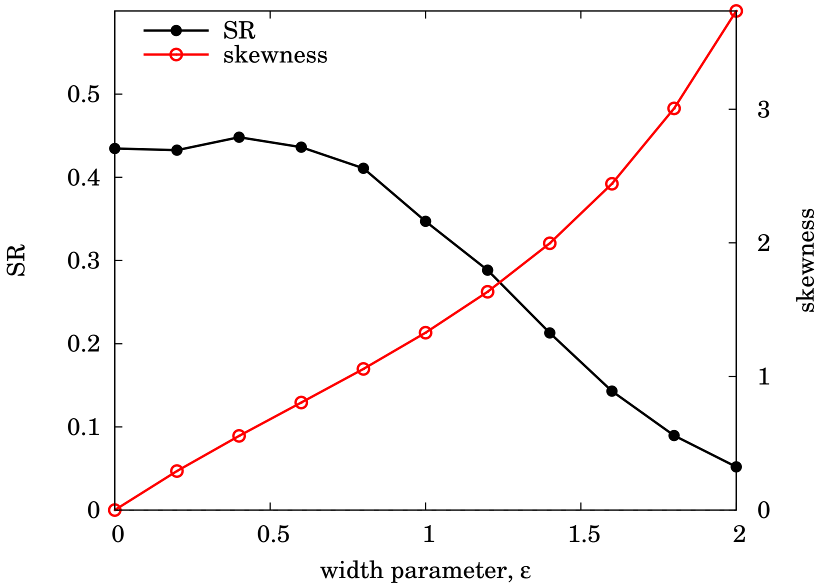


Figure 6: Sharpe ratio and skewness as a function of ε for double-step type activation function, showing performance dropoff for $\varepsilon > 0.6$.

$\frac{w_R}{w_S} \setminus \lambda$	0.25	0.50	0.75	1.00	1.25	1.50	1.75
∞	0.39; 1.77	0.45; 1.17	0.48; 0.69	0.48; 0.34	0.45; 0.09	0.40; −0.09	0.35; −0.23
5.0	0.38; 1.74	0.44; 1.19	0.48; 0.75	0.48; 0.43	0.46; 0.21	0.43; 0.04	0.39; −0.09
2.4	0.38; 1.73	0.43; 1.21	$\boxed{0.47; 0.80}$	0.48; 0.50	0.47; 0.29	0.45; 0.14	0.42; 0.02
1.5	0.38; 1.71	0.43; 1.22	0.46; 0.83	0.48; 0.56	0.47; 0.37	0.46; 0.22	0.43; 0.11
1	0.38; 1.70	0.42; 1.23	0.46; 0.87	0.47; 0.62	0.47; 0.43	0.46; 0.30	0.45; 0.19
0.67	0.38; 1.69	0.41; 1.24	0.45; 0.91	0.47; 0.67	0.47; 0.50	0.46; 0.37	0.45; 0.28
0.4	0.37; 1.67	0.41; 1.25	0.44; 0.94	0.46; 0.73	0.46; 0.57	0.46; 0.46	0.46; 0.37
0.2	0.37; 1.65	0.40; 1.27	0.43; 0.99	0.45; 0.80	0.46; 0.66	0.46; 0.56	0.46; 0.48
0	0.37; 1.63	0.39; 1.29	0.41; 1.05	0.43; 0.89	0.44; 0.77	0.45; 0.69	0.45; 0.62

Figure 7: Sharpe ratio and skewness as a function of λ and the ratio w_R/w_S for compound sigmoidal activation function. In each pair of numbers, the first is the SR, and the second is the skewness.

4.3 Skewness computation

We show in Figure 5 the term structure of skewness for the different examples given above: sigmoid (a,b), reverting sigmoid (c,d), double-step (e,f). The linear result is overlaid for comparison. The precise choice of momentum crossover does not affect the main conclusion, and we have used EMA2 with $N = 20, 40$ throughout. Using a faster or slower momentum measure simply stretches or compresses the graph in a horizontal direction, as it did in the linear models.

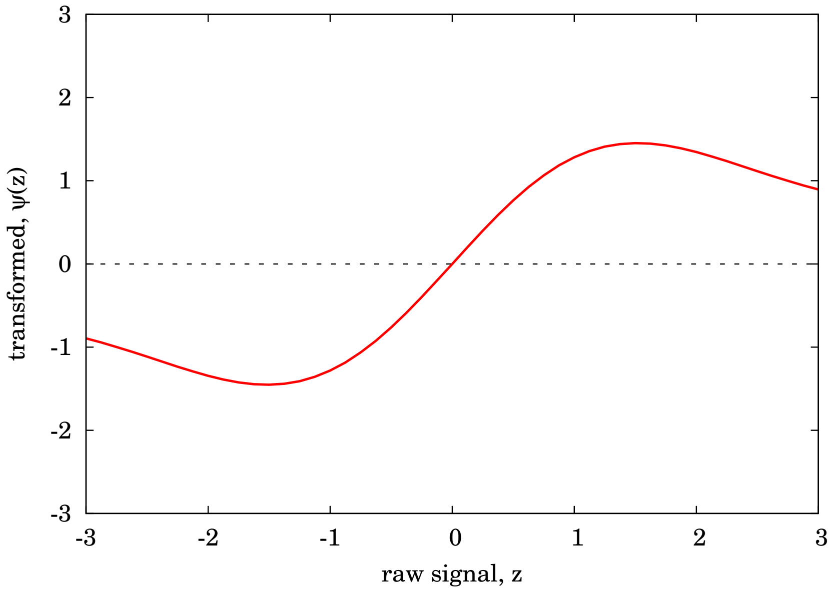


Figure 8: Sketch of the activation function highlighted in Figure 7 (parameters: $w_R = 0.71$, $w_S = 0.30$, $\lambda = 0.75$).

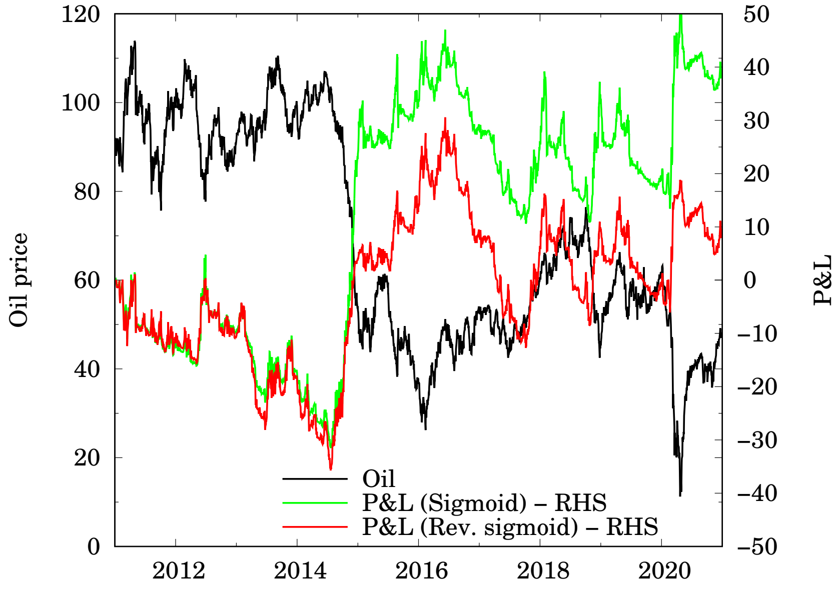


Figure 9: Trend-following oil over a decade, using the sigmoid and reverting sigmoid for activation function ($\lambda = 0.75$, $N_1 = 10$, $N_2 = 20$). The reverting sigmoid fails to capitalise on the full selloff in late 2014.

It is apparent from the results that as the activation function becomes progressively less linear, the main effect is to compress the graph in a vertical direction, so that the maximum skewness is reduced. With the reverting sigmoid, the graph can be affected much more, to the extent of becoming negative when λ is high enough: we predicted this earlier when remarking that H_k could become negative as a result of the activation function being decreasing over much of its domain, so the model spends a lot of time incrementally trading against the trend rather than with it. (In fact for $\psi(z) = ze^{-\lambda^2 z^2/2}$ the critical λ, above which the skewness is no longer everywhere positive, is around 1.3. An explanation is in the Appendix.)

The general conclusion so far is that any capping effect in the activation function will cause the trading returns to be less positively skewed, and any reverting effect will exacerbate this reduction in skewness. From the perspective of skewness alone, these effects should be avoided as much as possible. However, they may well be justified by reason of risk management and/or expected return, so we consider these next.

4.4 Empirical analysis and expected return

Analysis of the expected return is a totally different proposition because there are no theoretical guidelines at all. One can only adopt an empirical approach, seeing what has worked in the past, and relying on it continuing to do so.

We need to decide what objective function is to be maximised, and the most natural thing to do is to maximise the Sharpe ratio (SR) of the trading strategy, i.e. use an objective function that directly relates to trading model performance. As the SR is the expected return divided by the volatility, we will be penalising any effect that increases volatility without generating enough extra return. Taking a range of futures contracts across different asset classes (stocks, bonds, FX, commodities) and a range of EMA2 periods (5 vs 10 days, 10 vs 20 days, etc.), we have run trading simulations over the available history, which is typically 20 years or more, and calculated the Sharpe ratio; this gives a list of Sharpe ratios, one for each contract and speed. For simplicity we are going to use the same activation function across all contracts and speeds. We then average the list of Sharpe ratios and use this as our performance indicator, to be maximised[18].

We first examine the double-step activation function. Here there is only one parameter to adjust, namely ε, the half-width of the 'dead zone'. Figure 6 shows the performance as a function of ε. It is not surprising that the SR drops off as ε becomes large, because the strategy hardly ever has a position on and can never make any money. What is interesting is that the performance for $\varepsilon < 0.6$ is so flat. Thus from the perspective of expected return, one may as well choose any $\varepsilon < 0.6$.

However, when we overlay the conclusion about skewness, we can sharpen this deduction. As the skewness has a term structure, we look at one return-period (M) throughout: we choose $M = 100$ for convenience, this being the top of the curve for a linear activation function when the EMA periods are 20,40 (see Figure 5(b,d,f)). The skewness is also shown in Figure 6, and clearly it increases with increasing ε, so from that perspective alone we prefer ε as high as possible. If we can push ε up to about 0.6 without decreasing the SR, and in doing so can have positive skewness as well, then we should do just that. So this is our first conclusion about design of nonlinear momentum strategies: the blue line in Figure 5(e), $\varepsilon = 0.6$, is a good construction.

Next we turn to the sigmoidal functions that we introduced earlier, and take a linear combination of them:

$$\psi(z) = w_S\big(\tfrac{2}{\pi}\arctan\tfrac{\lambda^2}{\sqrt{1+2\lambda^2}}\big)^{-1/2}\big(2\Phi(\lambda z) - 1\big) + w_R(1+2\lambda^2)^{3/4}\, ze^{-\lambda^2 z^2/2}$$

with weights $w_{R,S} > 0$. To normalise the weighted function we enforce the elliptical constraint

$$w_S^2 + 2w_S w_R \cos\delta + w_R^2 = 1,$$

[18] The degree of temporal and cross-asset-class diversification that can be obtained is governed by what is commonly known as 'breadth' and explained in detail by Grinold [8].

where $\cos\delta$, the correlation between the R- and S- signals, is given by

$$\cos\delta = \frac{\lambda(1+2\lambda^2)^{1/4}}{1+\lambda^2}\left(\arctan\frac{\lambda^2}{\sqrt{1+2\lambda^2}}\right)^{-1/2}.$$

The effective number of parameters is now two: the horizontal scaling λ and the ratio w_R/w_S. The results are shown numerically in the table of Figure 7.

The general picture is that the performance surface is rather flat. Provided one avoids the far left (where the function is too linear and suffers from putting on too much risk when momentum is high) or the top right (where it reverts to zero too quickly when momentum is high), any of the pairs $(\frac{w_R}{w_S}, \lambda)$ would do reasonably well.

Again, we overlay the conclusions about skewness, simplifying as before by using the skewness of $M = 100$-day returns. We know that higher skewness arises from a low value of λ and from w_R/w_S small i.e. little reversive behaviour. This means going as far as possible to the left of the table, and steering well clear of the top right. Going to the left does lose performance (SR), so some trade-off is required. One particular example is highlighted ($w_R = 0.71$, $w_S = 0.30$, $\lambda = 0.75$) and sketched in Figure 8. It is seen that this does not yield the highest SR, but it is close to the maximum. It does not revert as strongly as the pure $ze^{-\lambda^2 z^2/2}$ does (top row of table), and is likely to be preferable on account of its better third-moment characteristics.

We now return to the discussion at the outset about designing strategies that perform well in specific scenarios. Figure 9 shows the results for the sigmoid and reverting sigmoid, with $\lambda = 0.75$, $N_1 = 10$, $N_2 = 20$, over ten years, for the oil market. It is clear that the reverting sigmoid does substantially less well. This is because the trend is strong and persists for a long time, so the reverting behaviour of the activation function causes severe underperformance. Yet the third column of Figure 7 suggests that the reverting sigmoid (top row) on average performs better than the sigmoid (bottom row). Part of the selling-point of CTA strategies is their ability to produce 'alpha' in scenarios such as the selloff in oil (and other commodities, and associated equities) in late 2014. It follows that one should make sure that the strategy does well in such scenarios, rather than simply relying on what has produced the best historical SR. This example also corroborates our earlier remark about calibration being sensitive to data history and therefore subjective. Were the oil selloff in late 2014 absent from the calibration, one would arrive at different conclusions about the optimal model.

As a final comment, we see that the smoother activation functions offer only a small improvement in Sharpe ratio over the double-step. However, we have not considered transaction costs, and models that generate sudden large trades can be more difficult to run a large amount of money on. Models that take positions more gradually are therefore easier to handle in trading. They are also easier to handle in backtesting, because a slight change in the definition of the momentum oscillator, which is the input to the activation function, can for a discontinuous function make a huge difference to the simulated position: even minor changes to the strategy can produce unpredictable results.

That said, the characteristic of the double-step function, that it waits until the momentum is above a certain level before trading, may be worthy of further investigation. Thus one aims for a function that is zero for $0 < |x| < \varepsilon$, then rises smoothly until a maximum is reached, and then rolls off slowly and asymptotes to a level above zero.

5 Conclusions and final remarks

We have shown how to analyse the behaviour of a variety of trend-following models by particular reference to the skewness of the distribution of trading returns. To do this we have needed only the first three moments of the market returns, thus keeping the modelling quite general. As regards linear models the most important formulae are (14,16) giving the second and third moments of the trading returns in an elegant application of residue calculus. Pure momentum (trending) strategies generate positive skewness even though the market returns might be totally symmetrical. The skewness depends on the return period and has a characteristic term structure which we have derived, illustrated and verified with real data. Hybrid strategies, with trending and counter-trending behaviour, may exhibit a more complex term structure of skewness, and we have shown how to analyse a general linear model.

We have investigated 'nonlinear momentum strategies' from different angles, understanding the Sharpe ratio and the skewness—in essence, the first and third moments—of their trading returns. The former was investigated empirically, and the latter mathematically. We have also pointed out that it may be wise to consider the behaviour of a strategy in specific scenarios, especially if they are a *raison d'être* of momentum trading. Specific conclusions about optimal design are given in the text, but two salient ones are repeated here.

First, the common practice of forming a momentum signal from moving averages and then making a 'binary bet' on it, $+1$ or -1 according as momentum is positive or negative, is not a good construction; we make other comments about discontinuous models in the next paragraph. It can be improved by waiting until the signal reaches a threshold before trading. Secondly, although it is a good idea to reduce—not just cap—the position when the momentum is very high, it is not advisable to reduce it too much. If too aggressive, the reducing effect can undesirably cause negatively-skewed trading returns, and much extra trading: when the momentum finally begins to fall, one ends up buying back the position and then dumping it again as the momentum goes back to zero. Also, the reducing effect loses profit badly when the trend is large and prolonged (Figure 9).

One area that we have not touched upon is the design and analysis of the opposite type of strategy, i.e. mean reversion. We can obtain a reversion model from a momentum model simply by multiplying by -1. This of course makes the skewness negative, but we do not want it to be as negative as possible. Therefore the optimal design will not simply be the reverse of what we have done here. Instead, something like the reverting sigmoid, now written as $\psi(x) = -c_\lambda x e^{-\lambda^2 x^2/2}$, is likely to be a good idea. When the market deviates from the reversion level a long way, some risk is taken off, which is likely to be beneficial.

A common explanation for the positive skewness is that it arises from the strategy having positive convexity, as mentioned for example in [5]. This is partly true, and we have explained its origin in §3, but in fact there is more to it than that, as will be discussed in forthcoming work.

A Momentum, price and return (discrete time)

We establish the link between momentum crossovers of prices (total returns), and a weighted sum of returns. The continuous-time analogue is very simple and has been dealt with in eq.(7) et seq.

An elementary view of the simple moving average (EMA1) is that is it the difference between the current price X_n and a weighted average of previous prices X_{n-j}. Define the average as

$$\mathcal{E}[X]_n = (1-\alpha)\sum_{j=0}^{\infty} \alpha^j X_{n-j}$$

which obeys the update recurrence

$$\mathcal{E}[X]_n = \alpha\mathcal{E}[X]_{n-1} + (1-\alpha)X_n.$$

The EMA1 is

$$\begin{aligned} X_n - \mathcal{E}_\alpha[X]_n &= X_n + \sum_{j=0}^{\infty} \alpha^{j+1} X_{n-j} - \sum_{j=0}^{\infty} \alpha^j X_{n-j} \\ &= \sum_{j=0}^{\infty} \alpha^{j+1}(X_{n-j} - X_{n-1-j}), \qquad (33) \end{aligned}$$

a weighted sum of returns with weight α^{j+1}. Note there is no $(1-\alpha)$ in it and the sum of the weights is not unity. It admits the simple update formula

$$X_n - \mathcal{E}_\alpha[X]_n = \alpha\big(X_{n-1} - \mathcal{E}_\alpha[X]_{n-1}\big) + \alpha(X_n - X_{n-1}), \qquad (34)$$

and hence is "$\alpha\times$ previous value plus $\alpha\times$ most recent return".

An EMA2 is the difference of two of these. In price terms it is usually formulated as "fast MA minus slow MA", which gives

$$\begin{aligned} \mathcal{E}_\alpha[X]_n - \mathcal{E}_\beta[X]_n &= \big(X_n - \mathcal{E}_\beta[X]\big) - \big(X_n - \mathcal{E}_\alpha[X]_n\big) \\ &= \sum_{j=0}^{\infty} (\beta^{j+1} - \alpha^{j+1})(X_{n-j} - X_{n-j-1}). \qquad (35) \end{aligned}$$

In context $\alpha < \beta$; equivalently $\alpha = 1 - N_\alpha^{-1}$ and $\beta = 1 - N_\beta^{-1}$ and $N_\alpha < N_\beta$.

When normalising a weighted sum, the premise is that the returns are uncorrelated and of unit variance. If the weights are (w_j) then the variance of the output is simply $\|w\|^2 = \sum_{j=0}^{\infty} w_j^2$, so one must divide by the square root of that quantity if the variance of the output is to be unity. For EMA1 and EMA2 this is easily obtained by summing the geometric series, and we find respectively that $\|w\|^2 = \frac{\alpha^2}{1-\alpha^2}$ for EMA1 and $\|w\|^2 = \frac{(\alpha-\beta)^2(1+\alpha\beta)}{(1-\alpha^2)(1-\beta^2)(1-\alpha\beta)}$ for EMA2, as used in §4.

B Formulary

B.1 Expectation formulae

If $(Z_1, Z_2) \sim N_2(\rho)$ (the bivariate Normal distribution with $N(0,1)$ marginals and correlation ρ) then

$$\mathbf{E}\big[f(Z_1,Z_2)e^{-a_1^2 Z_1^2/2}e^{-a_2^2 Z_2^2/2}\big] = \frac{1}{\sqrt{D}}\widehat{\mathbf{E}}\big[f(\sqrt{D_2/D}Z_1, \sqrt{D_1/D}Z_2)\big]$$

where under $\widehat{\mathbf{E}}$, $(Z_1, Z_2) \sim N_2(\widehat{\rho})$, with

$$\widehat{\rho} = \frac{\rho}{\sqrt{D_1 D_2}}, \quad D_i = (1-\rho^2)a_i^2 + 1, \quad D = (1-\rho^2)a_1^2 a_2^2 + a_1^2 + a_2^2 + 1$$

The following results are of use in obtaining (30–32). For $Z \sim N(0,1)$,

$$\begin{aligned}
\left\langle Z^{2n} e^{-b^2 Z^2/2} \right\rangle &= (2n-1)!!(1+b^2)^{-(2n+1)/2} \\
\langle \phi(a+bZ) \rangle &= \frac{1}{\sqrt{1+b^2}} \phi\left(\frac{a}{\sqrt{1+b^2}}\right) \\
\langle \Phi(a+bZ) \rangle &= \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \\
\langle Z\phi(a+bZ) \rangle &= \frac{-ab}{(1+b^2)^{3/2}} \phi\left(\frac{a}{\sqrt{1+b^2}}\right) \\
\langle Z\Phi(a+bZ) \rangle &= \frac{b}{\sqrt{1+b^2}} \phi\left(\frac{a}{\sqrt{1+b^2}}\right) \\
\langle \Phi(aZ)\Phi(bZ) \rangle &= \frac{1}{2\pi} \arctan\left(\frac{ab}{\sqrt{1+a^2+b^2}}\right) + \frac{1}{4}
\end{aligned}$$

The last result follows from differentiating both sides w.r.t. a, and using integration by parts on $\langle Z\phi(aZ)\Phi(bZ) \rangle$.

B.2 Skewness of reverting sigmoid activation function

We justify why the skewness is always positive for $|\lambda| \lesssim 1.3$. It was derived for linear models by means of z-transforms that for large M,

$$\left\langle \left(Y_n^{(M)}\right)^3 \right\rangle \sim 3M \sum_{k=1}^{\infty} H_k.$$

We are going to calculate this infinite sum, at least approximately, in the EMA1 case. By (31) the condition for positivity of the above expression is

$$\sum_{k=1}^{\infty} 2\alpha^k \frac{\alpha^k \left(1 - (1-\alpha^{2k})\lambda^4\right)}{\left(1 + 3\lambda^2 + 2(1-\alpha^{2k})\lambda^4\right)^{5/2}} > 0.$$

Write $\alpha^{2k} = u$ and approximate the sum as an integral over u, to give

$$\int_0^1 \frac{u\left(1 - (1-u)\lambda^4\right)}{\left(1 + 3\lambda^2 + 2(1-u)\lambda^4\right)^{5/2}} \frac{du}{u} > 0.$$

Upon doing the integral and tidying up, one ends up with

$$2 + 9\lambda^2 + 7\lambda^4 - 8\lambda^6 > 0;$$

so $\lambda^2 < 1.65$. This does not explain rigorously what goes on in the pre-asymptotic region when M is not large, and it uses EMA1 rather than EMA2, but the above analysis seems sufficient.

References

[1] E. Acar. *Economic Evaluation of Financial Forecasting.* PhD thesis, City University, London, 1992.

[2] E. Acar and S. Satchell. *Advanced Trading Rules.* Butterworth, 2002.

[3] T. Björk. *Arbitrage Theory in Continuous Time.* Oxford University Press, 1998.

[4] B. Bruder, T.-L. Dao, J.-C. Richard, and T. Roncalli. Trend filtering methods for momentum strategies. `www.ssrn.com`, 2011.

[5] T.-L. Dao, T.-T. Nguyen, C. Deremble, Y. Lempérière, J.-P. Bouchaud, and M. Potters. Tail protection for long investors: Trend convexity at work. `arXiv:1512.08037`, 2016.

[6] D. Bloom et al. Momentum strategies in FX. Technical report, HSBC Global Research, 27th Feb 2012.

[7] A. Feuerverger and A. C. M. Wong. Computation of value-at-risk for nonlinear portfolios. *J. of Risk*, 3(1):37–55, 2000.

[8] R. C. Grinold and R. N. Kahn. *Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk.* McGraw-Hill, New Jersey, 1999.

[9] S. Haykin. *Modern Filters.* Macmillan, 1989.

[10] A. F. Herbst. *Analyzing and Forecasting Futures Prices.* Wiley, 1992.

[11] P. Jusselin, E. Lezmi, H. Malongo, C. Masselin, T. Roncalli, and T.-L. Dao. Understanding the momentum risk premium: An in-depth journey through trend-following strategies. `www.ssrn.com`, 2017.

[12] R. J. Martin. Saddlepoint methods in portfolio theory. In A. Lipton and A. Rennie, editors, *The Oxford Handbook of Credit Derivatives*, chapter 15. Oxford University Press, 2011. `arXiv:1201.0106`.

[13] R. J. Martin. Optimal multifactor trading under proportional transaction costs. `arXiv:1204.6488`, 2012.

[14] R. J. Martin. Universal trading under proportional transaction costs. *RISK*, 27(8):54–59, 2014. `arXiv:1603.06558`.

[15] R. J. Martin and A. Bana. Nonlinear momentum strategies. *RISK*, 25(11):60–65, 2012.

[16] R. J. Martin and A. Birchall. Black to Negative: Embedded optionalities in commodities markets. `arXiv:2006.06076v2`, 2020.

[17] R. J. Martin and M. J. Kearney. Time since maximum of Brownian motion and asymmetric Lévy processes. *J. Phys. A: Math. Theor.*, 51:275001, 2018.

[18] R. J. Martin and T. Schöneborn. Mean reversion pays, but costs. *RISK*, 24(2):84–89, 2011. Full vsn at `arXiv:1103.4934`.

[19] R. J. Martin and D. Zou. Momentum trading: 'skews me. *RISK*, 25(8):40–45, 2012.

[20] M. Potters and J.-P. Bouchaud. Trend followers lose more often than they gain. *Wilmott Magazine*, pages 58–63, Nov/Dec 2005. Also at `arXiv:0508104`.

[21] H. Till and J. Eagleeye. A Hedge Fund Investor's Guide to Understanding Managed Futures. Technical report, EDHEC-Risk Institute, 2011. `www.edhec-risk.com`.