# Trend Filtering Methods for Momentum Strategies*

Benjamin Bruder
Research & Development
Lyxor Asset Management, Paris
benjamin.bruder@lyxor.com

Tung-Lam Dao
Research & Development
Lyxor Asset Management, Paris
tung-lam.dao@lyxor.com

Jean-Charles Richard
Research & Development
Lyxor Asset Management, Paris
jean-charles.richard@lyxor.com

Thierry Roncalli
Research & Development
Lyxor Asset Management, Paris
thierry.roncalli@lyxor.com

December 2011

**Abstract**

This paper studies trend filtering methods. These methods are widely used in momentum strategies, which correspond to an investment style based only on the history of past prices. For example, the CTA strategy used by hedge funds is one of the best-known momentum strategies. In this paper, we review the different econometric estimators to extract a trend of a time series. We distinguish between linear and non-linear models as well as univariate and multivariate filtering. For each approach, we provide a comprehensive presentation, an overview of its advantages and disadvantages and an application to the S&P 500 index. We also consider the calibration problem of these filters. We illustrate the two main solutions, the first based on prediction error, and the second using a benchmark estimator. We conclude the paper by listing some issues to consider when implementing a momentum strategy.

**Keywords:** Momentum strategy, trend following, moving average, filtering, trend extraction.

**JEL classification:** G11, G17, C63.

## 1 Introduction

The efficient market hypothesis tells us that financial asset prices fully reflect all available information (Fama, 1970). One consequence of this theory is that future returns are not predictable. Nevertheless, since the beginning of the nineties, a large body of academic research has rejected this assumption. One of the arguments is that risk premiums are time varying and depend on the business cycle (Cochrane, 2001). In this framework, returns on financial assets are related to some slow-moving economic variables that exhibit cyclical patterns in accordance with the business cycle. Another argument is that some agents are

*We are grateful to Guillaume Jamet and Hoang-Phong Nguyen for their helpful comments.

not fully rational, meaning that prices may underreact in the short run but overreact at long horizons (Hong and Stein, 1997). This phenomenon may be easily explained by the theory of behavioural finance (Barberis and Thaler, 2002).

Based on these two arguments, it is now commonly accepted that prices may exhibit trends or cycles. In some sense, these arguments chime with the Dow theory (Brown *et al.*, 1998), which is one of the first momentum strategies. A momentum strategy is an investment style based only on the history of past prices (Chan *et al.*, 1996). We generally distinguish between two types of momentum strategy:

1. the trend following strategy, which consists of buying (or selling) an asset if the estimated price trend is positive (or negative);

2. the contrarian (or mean-reverting) strategy, which consists of selling (or buying) an asset if the estimated price trend is positive (or negative).

Contrarian strategies are clearly the opposite of trend following strategies. One of the tasks involved in these strategies is to estimate the trend, excepted when based on mean-reverting processes (see D'Aspremont, 2011). In this paper, we provide a survey of the different trend filtering methods. However, trend filtering is just one of the difficulties in building a momentum strategy. The complete process of constructing a momentum strategy is highly complex, especially as regards transforming past trends into exposures – an important factor that is beyond the scope of this paper.

The paper is organized as follows. Section two presents a survey of the different econometric trend estimators. In particular, we distinguish between methods based on linear filtering and nonlinear filtering. In section three, we consider some issues that arise when trend filtering is applied in practice. We also propose some methods for calibrating trend filtering models and highlight the problem of estimator variance. Section four offers some concluding remarks.

# 2 A review of econometric estimators for trend filtering

Trend filtering (or trend detection) is a major task of time series analysis from both a mathematical and financial viewpoint. The trend of a time series is considered to be the component containing the global change, which contrasts with local changes due to noise. The trend filtering procedure concerns not only the problem of denoising; it must also take into account the dynamics of the underlying process. This explains why mathematical approaches to trend extraction have a long history, and why this subject is still of great interest to the scientific community[1]. From an investment perspective, trend filtering is fundamental to most momentum strategies developed in asset management and hedge funds sectors in order to improve performance and limit portfolio risks.

## 2.1 The trend-cycle model

In economics, trend-cycle decomposition plays an important role by identifying the permanent and transitory stochastic components in a non-stationary time series. Generally, the permanent component can be interpreted as a trend, whereas the transitory component may

[1] See Alexandrov *et al.* (2008).

be a noise or a stochastic cycle. Let $y_t$ be a stochastic process. We assume that $y_t$ is the sum of two different unobservable parts:

$$y_t = x_t + \varepsilon_t$$

where $x_t$ represents the trend and $\varepsilon_t$ is a stochastic (or noise) process. There is no precise definition for trend, but it is generally accepted to be a smooth function representing long-term movements:

> "*[...] the essential idea of trend is that it shall be smooth.*" (Kendall, 1973).

It means that changes in the trend $x_t$ must be smaller than those of the process $y_t$. From a statistical standpoint, it implies that the volatility of $y_t - y_{t-1}$ is higher than the volatility of $x_t - x_{t-1}$:

$$\sigma(y_t - y_{t-1}) \gg \sigma(x_t - x_{t-1})$$

One of the major problems in financial econometrics is the estimation of $x_t$. This is the subject of signal extraction and filtering (Pollock, 2009).

Finite moving average filtering for trend estimation has a long history. It has been used in actuarial science since the beginning of the twentieth century[2]. But the modern theory of signal filtering has its origins in the Second World War and was formulated independently by Norbert Wiener (1941) and Andrei Kolmogorov (1941) in two different ways. Wiener worked principally in the frequency domain whereas Kolmogorov considered a time-domain approach. This theory was extensively developed in the fifties and sixties by mathematicians and statisticians such as Hermann Wold, Peter Whittle, Rudolf Kalman, Maurice Priestley, George Box, etc. In economics, the problem of trend filtering is not a recent one, and may date back to the seminal article of Muth (1960). It was extensively studied in the eighties and nineties in the literature on business cycles, which led to a vast body of empirical research being carried out in this area[3]. However, it is in climatology that trend filtering is most extensively studied nowadays. Another important point is that the development of filtering techniques has evolved according to the development of computational power and the IT industry. The Savitzky-Golay smoothing procedure may appear very basic today though it was revolutionary[4] when it was published in 1964.

In what follows, we review the class of filtering techniques that is generally used to estimate a trend. Moving average filters play an important role in finance. As they are very intuitive and easy to implement, they undoubtedly represent the model most commonly used in trading strategies. The moving average technique belongs to the class of linear filters, which share a lot of common properties. After studying this class of filters, we consider some nonlinear filtering techniques, which may be well suited to solving financial problems.

## 2.2 Linear filtering

### 2.2.1 The convolution representation

We denote by $y = \{\ldots, y_{-2}, y_{-1}, y_0, y_1, y_2, \ldots\}$ the ordered sequence of observations of the process $y_t$. Let $\hat{x}_t$ be the estimator of the underlying trend $x_t$ which is by definition an

[2] See, in particular, the works of Henderson (1916), Whittaker (1923) and Macaulay (1931).

[3] See for example Cleveland and Tiao (1976), Beveridge and Nelson (1981), Harvey (1991) or Hodrick and Prescott (1997).

[4] The paper of Savitzky and Golay (1964) is still considered by the *Analytical Chemistry* journal to be one of its 10 seminal papers.

unobservable process. A filtering procedure consists of applying a filter $\mathcal{L}$ to the data $y$:

$$\hat{x} = \mathcal{L}(y)$$

with $\hat{x} = \{\ldots, \hat{x}_{-2}, \hat{x}_{-1}, \hat{x}_0, \hat{x}_1, \hat{x}_2, \ldots\}$. When the filter is linear, we have $\hat{x} = \mathcal{L}y$ with the normalisation condition $\mathbf{1} = \mathcal{L}\mathbf{1}$. If we assume that the signal $y_t$ is observed at regular dates[5], we obtain:

$$\hat{x}_t = \sum_{i=-\infty}^{\infty} \mathcal{L}_{t,t-i} y_{t-i} \tag{1}$$

We deduce that linear filtering may be viewed as a convolution. The previous filter may not be of much use, however, because it uses future values of $y_t$. As a result, we generally impose some restriction on the coefficients $\mathcal{L}_{t,t-i}$ in order to use only past and present values of the signal. In this case, we say that the filter is *causal*. Moreover, if we restrict our study to time invariant filters, the equation (1) becomes a simple convolution of the observed signal $y_t$ with a window function $\mathcal{L}_i$:

$$\hat{x}_t = \sum_{i=0}^{n-1} \mathcal{L}_i y_{t-i} \tag{2}$$

With this notation, a linear filter is characterised by a window kernel $\mathcal{L}_i$ and its support. The kernel defines the type of filtering, whereas the support defines the range of the filter. For instance, if we take a square window on a compact support $[0, T]$ with $T = n\Delta$ the width of the averaging window, we obtain the well-known moving average filter:

$$\mathcal{L}_i = \frac{1}{n}\mathbf{1}\{i < n\}$$

We finish this description by considering the lag representation:

$$\hat{x}_t = \sum_{i=0}^{n-1} \mathcal{L}_i \mathbf{L}^i y_t$$

with the lag operator $\mathbf{L}$ satisfying $\mathbf{L}y_t = y_{t-1}$.

### 2.2.2 Measuring the trend and its derivative

We discuss here how to use linear filtering to measure the trend of an asset price and its derivative. Let $S_t$ be the asset price which follows the dynamics of the Black-Scholes model:

$$\frac{\mathrm{d}S_t}{S_t} = \mu_t \,\mathrm{d}t + \sigma_t \,\mathrm{d}W_t$$

where $\mu_t$ is the drift, $\sigma_t$ is the volatility and $W_t$ is a standard Brownian motion. The asset price $S_t$ is observed in a series of discrete dates $\{t_0, \ldots, t_n\}$. Within this model, the appropriate signal to be filtered is the logarithm of the price $y_t = \ln S_t$ but not the price itself. Let $R_t = \ln S_t - \ln S_{t-1}$ represent the realised return at time $t$ over a unit period. If $\mu_t$ and $\sigma_t$ are known, we have:

$$R_t = \left(\mu_t - \frac{1}{2}\sigma_t^2\right)\Delta + \sigma_t\sqrt{\Delta}\eta_t$$

[5] We have $t_{i+1} - t_i = \Delta$.

where $\eta_t$ is a standard Gaussian white noise. The filtered trend can be extracted using the following equation:

$$\hat{x}_t = \sum_{i=0}^{n-1} \mathcal{L}_i y_{t-i}$$

and the estimator of $\mu_t$ is[6]:

$$\hat{\mu}_t \simeq \frac{1}{\Delta} \sum_{i=0}^{n-1} \mathcal{L}_i R_{t-i}$$

We can also obtain the same result by applying the filter directly to the signal and defining the derivative of the window function as $\ell_i = \hat{\mathcal{L}}_i$:

$$\hat{\mu}_t \simeq \frac{1}{\Delta} \sum_{i=0}^{n} \ell_i y_{t-i}$$

We obtain the following correspondence:

$$\ell_i = \begin{cases} \mathcal{L}_0 & \text{if } i = 0 \\ \mathcal{L}_i - \mathcal{L}_{i-1} & \text{if } i = 1, \ldots, n-1 \\ -\mathcal{L}_{n-1} & \text{if } i = n \end{cases} \tag{3}$$

**Remark 1** *In some senses, $\hat{\mu}_t$ and $\hat{x}_t$ are related by the following expression:*

$$\hat{\mu}_t = \frac{\mathrm{d}}{\mathrm{d}t} \hat{x}_t$$

*Econometric methods principally involve $\hat{x}_t$, whereas $\hat{\mu}_t$ is more important for trading strategies.*

**Remark 2** *$\hat{\mu}_t$ is a biased estimator of $\mu_t$ and the bias increases with the volatility of the process $\sigma_t$. The expression of the unbiased estimator is then:*

$$\hat{\mu}_t = \frac{1}{2}\sigma_t^2 + \frac{1}{\Delta} \sum_{i=0}^{n-1} \mathcal{L}_i R_{t-i}$$

**Remark 3** *In the previous analysis, $\hat{x}_t$ and $\hat{\mu}_t$ are two estimators. We may also represent them by their corresponding probability density functions. It is therefore easy to derive estimates, but we should not forget that these estimators present some variance. In finance, and in particular in trading strategies, the question of statistical inference is generally not addressed. However, it is a crucial factor in designing a successful momentum strategy.*

### 2.2.3 Moving average filters

**Average return over a given period** Here, we consider the simplest case corresponding to the moving average filter where the form of the window is:

$$\mathcal{L}_i = \frac{1}{n} \mathbf{1}\{i < n\}$$

In this case, the only calibration parameter is the window support, i.e. $T = n\Delta$. It characterises the smoothness of the filtered signal. For the limit $T \to 0$, the window becomes a Dirac distribution $\delta_t$ and the filtered signal is exactly the same as the observed signal:

[6] If we neglect the contribution from the term $\sigma_t^2$. Moreover, we consider $\Delta = 1$ to simplify the calculation.

$\hat{x}_t = y_t$. For $T > 0$, if we assume that the noise $\varepsilon_t$ is independent from $x_t$ and is a centered process, the first contribution of the filtered signal is the average trend:

$$\hat{x}_t = \frac{1}{n}\sum_{i=0}^{n-1} x_{t-i}$$

If the trend is homogeneous, this average value is located at $t - (n-1)/2$ by construction. It means that the filtered signal lags the observed signal by a time period which is half the window. To extract the derivative of the trend, we compute the derivative kernel $\ell_i$ which is given by the following formula:

$$\ell_i = \frac{1}{n\Delta}(\delta_{i,0} - \delta_{i,n})$$

where $\delta_{i,j}$ is the Kronecker delta[7]. The main advantage of using a moving average filter is the reduction of noise due to the central limit theorem. For the limit case $n \to \infty$, the signal is completely denoised but it corresponds to the average value of the trend. The estimator is also biased. In trend filtering, we also face a trade-off between denoising maximisation and bias minimisation. The problem is the calibration procedure for the lag window $T$. Another way to determine the optimal parameter $T^\star$ is to take into account the dynamics of the trend.

The above moving average filter can be applied directly to the signal. However, $\hat{\mu}_t$ is simply the cumulative return over the window period. It needs only the first and last dates of the period under consideration.

**Moving average crossovers** Many practitioners, and even individual investors, use the moving average of the price itself as a trend indication, instead of the moving average of returns. These moving averages are generally uniform moving averages of the price. Here we will consider an average of the logarithm of the price, in order to be consistent with the previous examples:

$$\hat{y}_t^n = \frac{1}{n}\sum_{i=0}^{n-1} y_{t-i}$$

Of course, an average price does not estimate the trend $\mu_t$. This trend is estimated from the difference between two moving averages over two different time horizons $n_1$ and $n_2$. Supposing that $n_1 > n_2$, the trend $\mu$ may be estimated from:

$$\hat{\mu}_t \simeq \frac{2}{(n_1 - n_2)\Delta}(\hat{y}_t^{n_2} - \hat{y}_t^{n_1}) \tag{4}$$

In particular, the estimated trend is positive if the short-term moving average is higher than the long-term moving average. Thus, the sign of the trend changes when the short-term moving average crosses the long-term moving average. Of course, when the short-term horizon $n_1$ is one, then the short-term moving average is just the current asset price. The scaling term $2(n_1 - n_2)^{-1}$ is explained below. It is derived from the interpretation of this estimator as a weighted moving average of asset returns. Indeed, this estimator can be interpreted in terms of asset returns by inverting the formula (3) with $\mathcal{L}_i$ being interpreted as the primitive of $\ell_i$:

$$\mathcal{L}_i = \begin{cases} \ell_0 & \text{if } i = 0 \\ \ell_i + \mathcal{L}_{i-1} & \text{if } i = 1, \ldots, n-1 \\ -\ell_{n+1} & \text{if } i = n \end{cases}$$

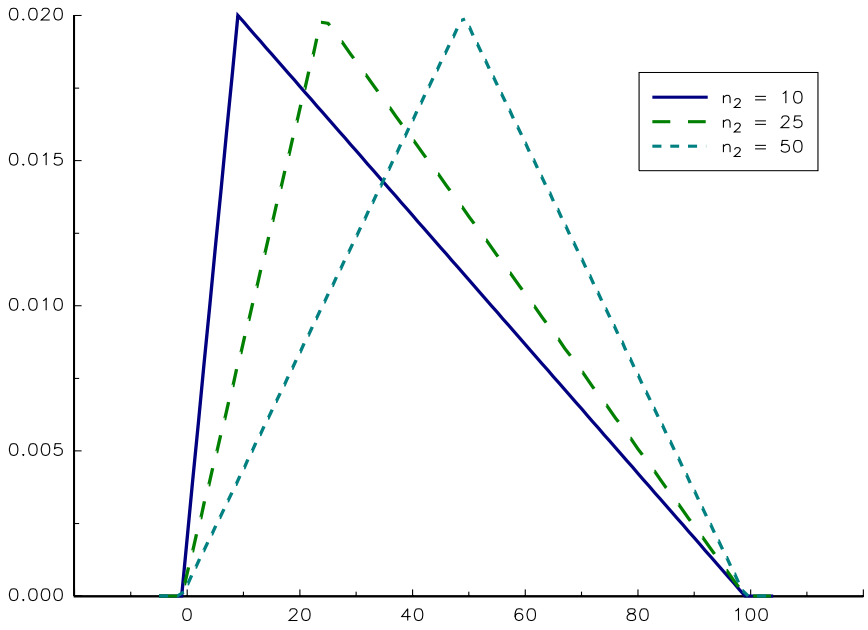
[7] $\delta_{i,j}$ is equal to 1 if $i = j$ and 0 otherwise.

The weighting of each return in the estimator (4) is represented in Figure 1. It forms a triangle, and the biggest weighting is given at the horizon of the smallest moving average. Therefore, depending on the horizon $n_2$ of the shortest moving average, the indicator can be focused toward the current trend (if $n_2$ is small) or toward past trends (if $n_2$ is as large as $n_1/2$ for instance). From these weightings, in the case of a constant trend $\mu$, we can compute the expectation of the difference between the two moving averages:

$$\mathbb{E}\left[\hat{y}_t^{n_2} - \hat{y}_t^{n_1}\right] = \frac{n_1 - n_2}{2}\left(\mu - \frac{1}{2}\sigma_t^2\right)\Delta$$

Therefore, the scaling factor in formula (4) appears naturally.

Figure 1: Window function $\mathcal{L}_i$ of moving average crossovers ($n_1 = 100$)




**Enhanced filters** To improve the uniform moving average estimator, we may take the following kernel function:

$$\ell_i = \frac{4}{n^2}\operatorname{sgn}\left(\frac{n}{2} - i\right)$$

We notice that the estimator $\hat{\mu}_t$ now takes into account all the dates of the window period. By taking the primitive of the function $\ell_i$, the trend filter is given as follows:

$$\mathcal{L}_i = \frac{4}{n^2}\left(\frac{n}{2} - \left|i - \frac{n}{2}\right|\right)$$

We now move to the second type of moving average filter which is characterised by an asymmetric form of the convolution kernel. One possibility is to take an asymmetric window function with a triangular form:

$$\mathcal{L}_i = \frac{2}{n^2}(n - i)\,\mathbf{1}\{i < n\}$$

By computing the derivative of this window function, we obtain the following kernel:

$$\ell_i = \frac{2}{n}\left(\delta_i - \mathbf{1}\left\{i < n\right\}\right)$$

The filtering equation of $\mu_t$ then becomes:

$$\hat{\mu}_t = \frac{2}{n}\left(x_t - \frac{1}{n}\sum_{i=0}^{n-1} x_{t-i}\right)$$

**Remark 4** *Another way to define $\hat{\mu}_t$ is to consider the Lanczos generalised derivative (Groetsch, 1998). Let $f(x)$ be a function. We define the Lanczos derivative of $f(x)$ in terms of the following relationship:*

$$\frac{\mathrm{d}^L}{\mathrm{d}x} f(x) = \lim_{\varepsilon \to 0} \frac{3}{2\varepsilon^3} \int_{-\varepsilon}^{\varepsilon} t f(x+t)\ \mathrm{d}t$$

*In the discrete case, we have:*

$$\frac{\mathrm{d}^L}{\mathrm{d}x} f(x) = \lim_{h \to 0} \frac{\sum_{k=-n}^{n} k f(x+kh)}{2\sum_{k=1}^{n} k^2 h}$$

*We first notice that the Lanczos derivative is more general than the traditional derivative. Although Lanczos' formula is a more onerous method for finding the derivative, it offers some advantages. This technique allows us to compute a "pseudo-derivative" at points where the function is not differentiable. For the observable signal $y_t$, the traditional derivative does not exist because of the noise $\varepsilon_t$, but does in the case of the Lanczos derivative. Let us apply the Lanczos' formula to estimate the derivative of the trend at the point $t - T/2$. We obtain:*

$$\frac{\mathrm{d}^L}{\mathrm{d}t} \hat{x}_t = \frac{12}{n^3} \sum_{i=0}^{n} \left(\frac{n}{2} - i\right) y_{t-i}$$

*We deduce that the kernel is:*

$$\ell_i = \frac{12}{n^3}\left(\frac{n}{2} - i\right)\mathbf{1}\left\{0 \leq i \leq n\right\}$$

*By computing an integration by parts, we obtain the trend filter:*

$$\mathcal{L}_i = \frac{6}{n^3} i\,(n-i)\,\mathbf{1}\left\{0 \leq i \leq n\right\}$$

In Figure 2, we have represented the different functions $\mathcal{L}_i$ given in this paragraph. We may extend these filters by computing the convolution of two or more filters. For exemple, the mixed filter in Figure 2 is the convolution of the asymmetric filter with the Lanczos filter. Let us apply these filters to the S&P 500 index. The results are given in Figure 3 for two values of the window length ($n = 65$ days and $n = 260$ days). We notice that the choice of $n$ has a big impact on the filtered series. The choice of the window function seems to be less important at first sight. However, we should mention that traders are principally interested in the derivative of the trend, and not the absolute value of the trend itself. In this case, the window function may have a significant impact. Figure 4 is the scatterplot of the $\hat{\mu}_t$ statistic in the case of the S&P 500 index from January 2000 to July 2011 (we have considered the uniform and Lanczos filters using $n = 260$). We may also show that this impact increases when we reduce the length of the window as illustrated in Table 1.

Figure 2: Window function $\mathcal{L}_i$ of moving average filters ($n = 100$)

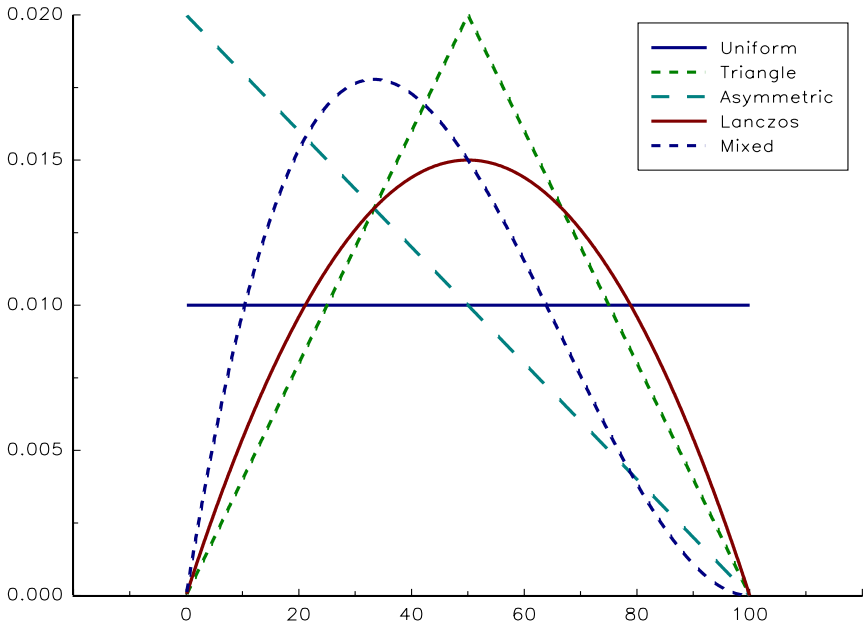



Figure 3: Trend estimate for the S&P 500 index

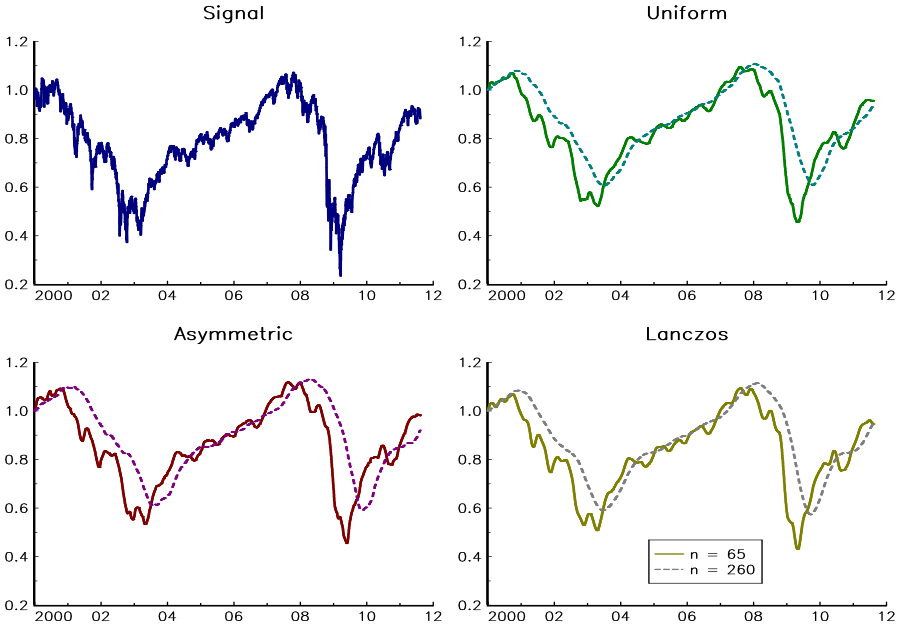
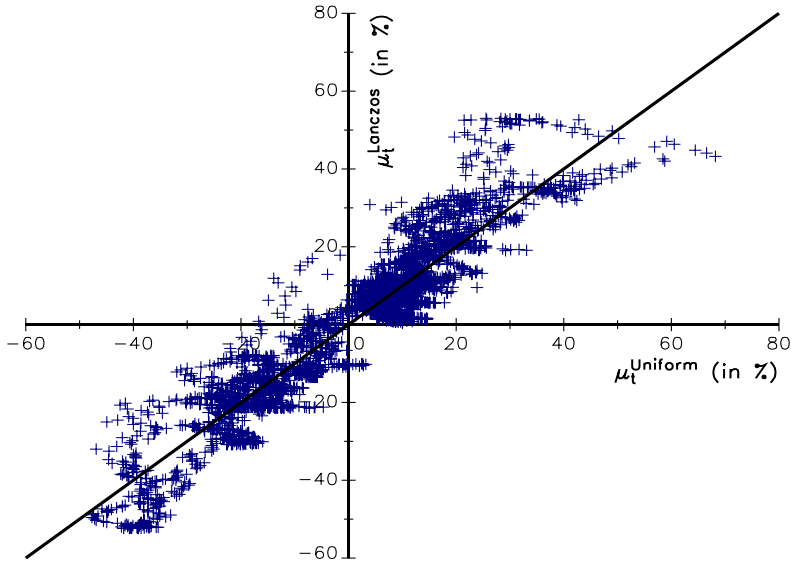


Table 1: Correlation between the uniform and Lanczos derivatives

| $n$ | 5 | 10 | 22 | 65 | 130 | 260 |
|---|---|---|---|---|---|---|
| Pearson $\rho$ | 84.67 | 87.86 | 90.14 | 90.52 | 92.57 | 94.03 |
| Kendall $\tau$ | 65.69 | 68.92 | 70.94 | 71.63 | 73.63 | 76.17 |
| Spearman $\varrho$ | 83.15 | 86.09 | 88.17 | 88.92 | 90.18 | 92.19 |

Figure 4: Comparison of the derivative of the trend




### 2.2.4 Least squares filters

**$L_2$ filtering** The previous Lanczos filter may be viewed as a local linear regression (Burch *et al.*, 2005). More generally, least squares methods are often used to define trend estimators:

$$\{\hat{x}_1, \ldots, \hat{x}_n\} = \arg\min \frac{1}{2} \sum_{t=1}^{n} (y_t - \hat{x}_t)^2$$

However, this problem is not well-defined. We also need to impose some restrictions on the underlying process $y_t$ or on the filtered trend $\hat{x}_t$ to obtain a solution. For example, we may consider a deterministic constant trend:

$$x_t = x_{t-1} + \mu$$

In this case, we have:

$$y_t = \mu t + \varepsilon_t \tag{5}$$

Estimating the filtered trend $\hat{x}_t$ is also equivalent to estimating the coefficient $\mu$:

$$\hat{\mu} = \frac{\sum_{t=1}^{n} t y_t}{\sum_{t=1}^{n} t^2}$$

If we consider a trend that is not constant, we may define the following objective function:

$$\frac{1}{2}\sum_{t=1}^{n}\left(y_t-\hat{x}_t\right)^2+\lambda\sum_{t=2}^{n-1}\left(\hat{x}_{t-1}-2\hat{x}_t+\hat{x}_{t+1}\right)^2$$

In this function, $\lambda$ is the regularisation parameter which controls the competition between the smoothness[8] of $\hat{x}_t$ and the noise $y_t-\hat{x}_t$. We may rewrite the objective function in the vectorial form:

$$\frac{1}{2}\left\|y-\hat{x}\right\|_2^2+\lambda\left\|D\hat{x}\right\|_2^2$$

where $y=(y_1,\ldots,y_n)$, $\hat{x}=(\hat{x}_1,\ldots,\hat{x}_n)$ and the $D$ operator is the $(n-2)\times n$ matrix:

$$D=\begin{bmatrix}1&-2&1&&&&\\&1&-2&1&&&\\&&&\ddots&&&\\&&&1&-2&1&\\&&&&1&2&1\end{bmatrix}$$

The estimator is then given by the following solution:

$$\hat{x}=\left(I+2\lambda D^\top D\right)^{-1}y$$

It is known as the Hodrick-Prescott filter (or $L_2$ filter). This filter plays an important role in calibrating the business cycle.

**Kalman filtering** Another important trend estimation technique is the Kalman filter, which is described in Appendix A.1. In this case, the trend $\mu_t$ is a hidden process which follows a given dynamic. For example, we may assume that the model is[9]:

$$\begin{cases}R_t=\mu_t+\sigma_\zeta\zeta_t\\ \mu_t=\mu_{t-1}+\sigma_\eta\eta_t\end{cases}\tag{6}$$

Here, the equation of $R_t$ is the measurement equation and $R_t$ is the observable signal of realised returns. The hidden process $\mu_t$ is supposed to follow a random walk. We define $\hat{\mu}_{t|t-1}=\mathbb{E}_{t-1}\left[\mu_t\right]$ and $P_{t|t-1}=\mathbb{E}_{t-1}\left[\left(\hat{\mu}_{t|t-1}-\mu_t\right)^2\right]$. Using the results given in Appendix A.1, we have:

$$\hat{\mu}_{t+1|t}=\left(1-K_t\right)\hat{\mu}_{t|t-1}+K_tR_t$$

where $K_t=P_{t|t-1}/\left(P_{t|t-1}+\sigma_\zeta^2\right)$ is the Kalman gain. The estimation error is determined by Riccati's equation:

$$P_{t+1|t}=P_{t|t-1}+\sigma_\eta^2-P_{t|t-1}K_t$$

Riccati's equation gives us the stationary solution:

$$P^*=\frac{\sigma_\eta}{2}\left(\sigma_\eta+\sqrt{\sigma_\eta^2+4\sigma_\zeta^2}\right)$$

The filter equation becomes:

$$\hat{\mu}_{t+1|t}=\left(1-\kappa\right)\hat{\mu}_{t|t-1}+\kappa R_t$$

[8] We notice that the second term is the discrete derivative of the trend $\hat{x}_t$ which characterises the smoothness of the curve.

[9] Equation (5) is a special case of this model if $\sigma_\eta=0$.

with:

$$\kappa = \frac{2\sigma_\eta}{\sigma_\eta + \sqrt{\sigma_\eta^2 + 4\sigma_\zeta^2}}$$

This Kalman filter can be considered as an exponential moving average filter with parameter[10] $\lambda = -\ln(1-\kappa)$:

$$\hat{\mu}_t = \left(1 - e^{-\lambda}\right) \sum_{i=0}^{\infty} e^{-\lambda i} R_{t-i}$$

with[11] $\hat{\mu}_t = \mathbb{E}_t[\mu_t]$. The filter of the trend $\hat{x}_t$ is therefore determined by the following equation:

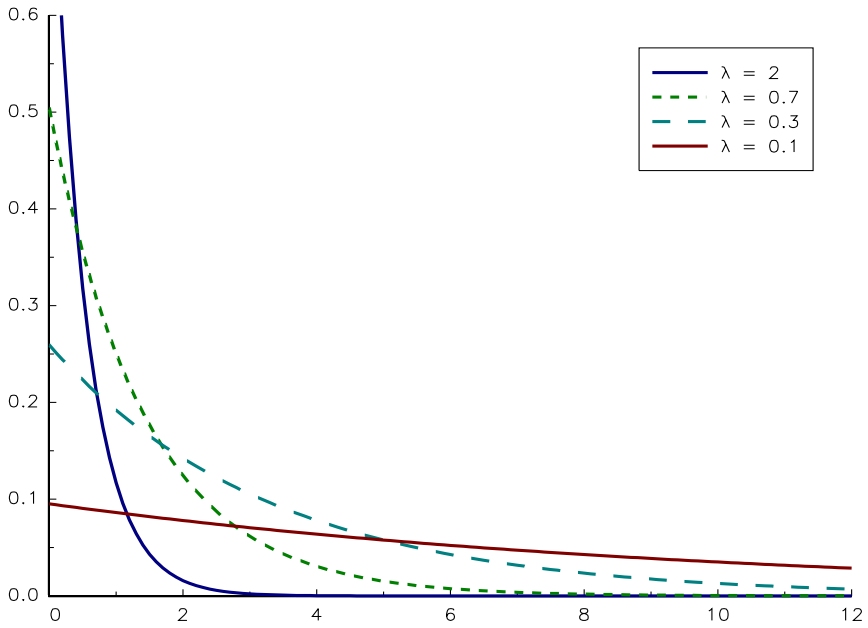
$$\hat{x}_t = \left(1 - e^{-\lambda}\right) \sum_{i=0}^{\infty} e^{-\lambda i} y_{t-i}$$

while the derivative of the trend may be directly related to the observed signal $y_t$ as follows:

$$\hat{\mu}_t = \left(1 - e^{-\lambda}\right) y_t - \left(1 - e^{-\lambda}\right)\left(e^{\lambda} - 1\right) \sum_{i=1}^{\infty} e^{-\lambda i} y_{t-i}$$

In Figure 5, we reported the window function of the Kalman filter for several values of $\lambda$. We notice that the cumulative weightings increase strongly with $\lambda$. The half-life of this filter is approximatively equal to $\lceil \left(\lambda^{-1} - 2^{-1}\right) \ln 2 \rceil$. For example, the half-life for $\lambda = 5\%$ is 14 days.

Figure 5: Window function $\mathcal{L}_i$ of the Kalman filter




[10] We have $0 < \kappa < 1$ and $lambda > 0$.
[11] We notice that $\hat{\mu}_{t+1|t} = \hat{\mu}_t$.

We may wonder what the link is between the regression model (5) and the Markov model (6). Equation (5) is equivalent to the following state space model[12]:

$$\begin{cases} y_t = x_t + \sigma_\varepsilon \varepsilon_t \\ x_t = x_{t-1} + \mu \end{cases}$$

If we now consider that the trend is stochastic, the model becomes:

$$\begin{cases} y_t = x_t + \sigma_\varepsilon \varepsilon_t \\ x_t = x_{t-1} + \mu + \sigma_\zeta \zeta_t \end{cases}$$

This model is called the local level model. We may also assume that the slope of the trend is stochastic, in which case we obtain the local linear trend model:

$$\begin{cases} y_t = x_t + \sigma_\varepsilon \varepsilon_t \\ x_t = x_{t-1} + \mu_{t-1} + \sigma_\zeta \zeta_t \\ \mu_t = \mu_{t-1} + \sigma_\eta \eta_t \end{cases}$$

These three models are special cases of structural models (Harvey, 1989) and may be easily solved by Kalman filtering. We also deduce that the Markov model (6) is a special case of the latter when $\sigma_\varepsilon = 0$.

**Remark 5** *We have shown that Kalman filtering may be viewed as an exponential moving average filter when we consider the Markov model (6). Nevertheless, we cannot regard the Kalman filter simply as a moving average filter. First, the Kalman filter is the optimal filter in the case of the linear Gaussian model described in Appendix A.1. Second, it could be regarded as "an efficient computational solution of the least squares method" (Sorensen, 1970). Third, we could use it to solve more sophisticated processes than the Markov model (6). However, some nonlinear or non Gaussian models may be too complex for Kalman filtering. These nonlinear models can be solved by particle filters or sequential Monte Carlo methods (see Doucet et al., 1998).*

Another important feature of the Kalman approach is the derivation of an optimal smoother (see Appendix A.1). At time $t$, we are interested by the numerical value of $x_t$, but also by the past values of $x_{t-i}$ because we would like to measure the slope of the trend. The Kalman smoother improves the estimate of $\hat{x}_{t-i}$ by using all the information between $t-i$ and $t$. Let us consider the previous example in relation to the S&P 500 index, using the local level model. Figure 6 gives the filtered and smoothed components $x_t$ and $\mu_t$ for two sets of parameters[13]. We verify that the Kalman smoother reduces the noise by incorporating more information. We also notice that the restriction $\sigma_\varepsilon = 0$ increases the variance of the trend and slope estimators.

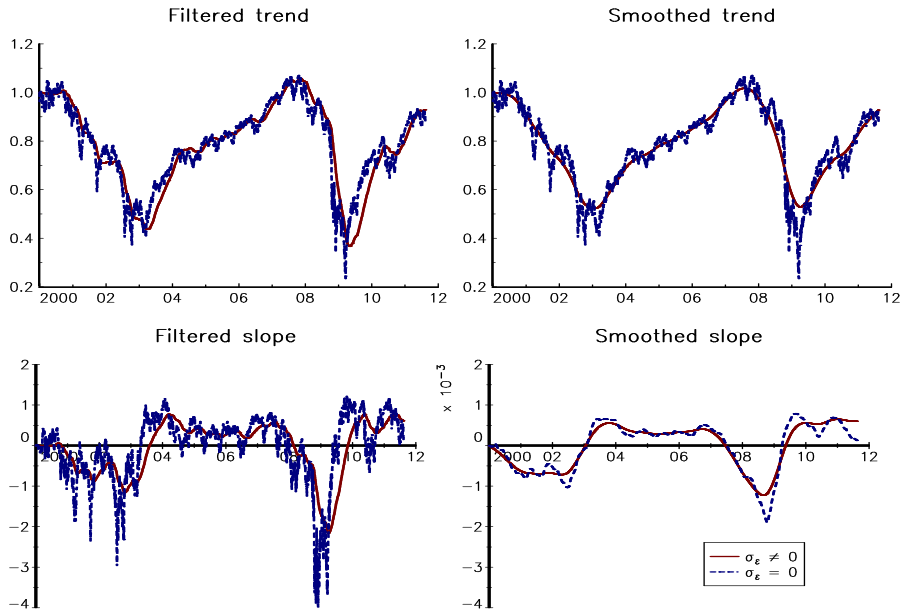
## 2.3 Nonlinear filtering

In this section, we review other filtering approaches. They are generally classed as nonlinear filters, because it is not possible to express the trend as a linear convolution of the signal and a window function.

[12] In what follows, the noise processes are white noise: $\varepsilon_t \sim \mathcal{N}(0,1)$, $\zeta_t \sim \mathcal{N}(0,1)$ and $\eta_t \sim \mathcal{N}(0,1)$.

[13] For the first set of parameters, we assume that $\sigma_\varepsilon = 100\sigma_\zeta$ and $\sigma_\eta = 1/100\sigma_\zeta$. For the second set of parameters, we impose the restriction $\sigma_\varepsilon = 0$.

Figure 6: Kalman filtered and smoothed components




### 2.3.1 Nonparametric regression

In the regression model (5), we assume that $x_t = f(t)$ while $f(t) = \mu t$. The model is said to be parametric because the estimation of the trend consists of estimating the parameter $\mu$. We then have $\hat{x}_t = \hat{\mu} t$. With nonparametric regression, we directly estimate the function $f$, obtaining $\hat{x}_t = \hat{f}(t)$. Some examples of nonparametric regression are kernel regression, loess regression and spline regression. A popular method for trend filtering is local polynomial regression:

$$\begin{aligned} y_t &= f(t) + \varepsilon_t \\ &= \beta_0(\tau) + \sum_{j=1}^{p} \beta_j(\tau)(\tau - t)^j + \varepsilon_t \end{aligned}$$

For a given value of $\tau$, we estimate the parameters $\hat{\beta}_j(\tau)$ using weighted least squares with the following weightings:

$$w_t = \mathcal{K}\left(\frac{\tau - t}{h}\right)$$

where $\mathcal{K}$ is the kernel function with a bandwidth $h$. We deduce that:

$$\hat{x}_t = \mathbb{E}[y_t | \tau = t] = \hat{\beta}_0(t)$$

Cleveland (1979) proposed an improvement to the kernel regression through a two-stage procedure (loess regression). First, we fit a polynomial regression to estimate the residuals $\hat{\varepsilon}_t$. Then, we compute $\delta_t = \left(1 - u_t^2\right) \cdot 1\{|u_t| \leq 1\}$ with $u_t = \hat{\varepsilon}_t / (6 \operatorname{median}(|\hat{\varepsilon}|))$ and run a second kernel regression[14] with weightings $\delta_t w_t$.

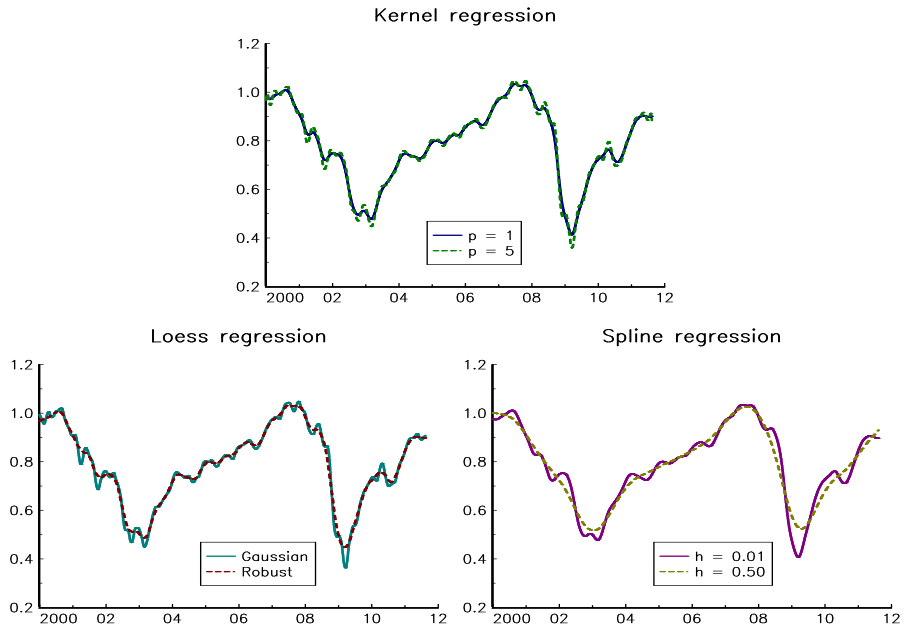
[14] Cleveland (1979) suggests using the tricube kernel function to define $\mathcal{K}$.

A spline function is a $C^2$ function $S(\tau)$ which corresponds to a cubic polynomial function on each interval $[t, t+1[$. Let $\mathcal{SP}$ be the set of spline functions. We then have to solve the following optimisation programme:

$$\min_{S \in \mathcal{SP}} (1-h) \sum_{t=0}^{n} w_t \left(y_t - S(t)\right)^2 + h \int_0^T w_\tau S''(\tau)^2 \, \mathrm{d}\tau$$

where $h$ is the smoothing parameter – $h = 0$ corresponds to the interpolation case[15] and $h = 1$ corresponds to the linear regression[16].

Figure 7: Illustration of the kernel, loess and spline filters




We illustrate these three nonparametric methods in Figure 7. The calibration of these filters is more complicated than for moving average filters, where the only parameter is the length $n$ of the window. With these methods, we have to decide the polynomial degree[17] $p$, the kernel function[18] $\mathcal{K}$ and the smoothing parameter[19] $h$.

### 2.3.2 $L_1$ filtering

The idea of the Hodrick-Prescott filter can be generalised to a larger class of filters by using the $L_p$ penalty condition instead of the $L_2$ penalty. This generalisation was previously

[15] We have $\hat{x}_t = S(t) = y_t$.

[16] We have $\hat{x}_t = S(t) = \hat{c} + \hat{\mu} t$ with $(\hat{c}, \hat{\mu})$ the OLS estimate of $y_t$ on a constant and time $t$ because the optimum is reached for $S''(\tau) = 0$.

[17] For the kernel regression, we use a Gaussian kernel with a bandwidth $h = 0.10$. We notice the impact of the degree of polynomial. The higher the degree, the smoother the trend (and the slope of the trend).

[18] For the loess regression, the degree of polynomial is set to 1 and the bandwidth $h$ is 0.02. We show the impact of the second step which modifies the kernel function.

[19] For the spline regression, we consider a uniform kernel function. We notice that the parameter $h$ has an impact on the smoothness of the trend.

discussed in the work of Daubechies *et al.* (2004) in relation to the linear inverse problem, while Tibshirani (1996) considers the Lasso regression problem. If we consider an $L_1$ filter, the objective function becomes:

$$\frac{1}{2}\sum_{t=1}^{n}\left(y_t-\hat{x}_t\right)^2+\lambda\sum_{t=2}^{n-1}\left|\hat{x}_{t-1}-2\hat{x}_t+\hat{x}_{t+1}\right|$$

which is equivalent to the following vectorial form:

$$\frac{1}{2}\left\|y-\hat{x}\right\|_2^2+\lambda\left\|D\hat{x}\right\|_1$$

Kim *et al.* (2009) shows that the dual problem of this $L_1$ filter scheme is a quadratic programme with some boundary constraints[20]. To find $\hat{x}$, we may also use the quadratic programming algorithm, but Kim *et al.* (2009) suggest using the primal-dual interior point method in order to optimise the numerical computation speed.

We have illustrated the $L_1$ filter in Figure 8. Contrary to all other previous methods, the filtered signal comprises a set of straight trends and breaks[21], because the $L_1$ norm imposes the condition that the second derivative of the filtered signal must be zero. The competition between the two terms in the objective function turns to the competition between the number of straight trends (or the number of breaks) and the closeness to the data. Thus, the smoothing parameter $\lambda$ plays an important role for detecting the number of breaks. This explains why $L_1$ filtering is radically different to $L_2$ (or Hodrick-Prescott) filtering. Moreover, it is easy to compute the slope of the trend $\hat{\mu}_t$ for the $L_1$ filter. It is a step function, indicating clearly if the trend is up or down, and when it changes (see Figure 8).

### 2.3.3 Wavelet filtering

Another way to estimate the trend $x_t$ is to denoise the signal $y_t$ by using spectral analysis. The Fourier transform is an alternative representation of the original signal $y_t$, which becomes a frequency function:

$$y\left(\omega\right)=\sum_{t=1}^{n}y_te^{-i\omega t}$$

We note $y\left(\omega\right)=\mathcal{F}\left(y\right)$. By construction, we have $y=\mathcal{F}^{-1}\left(y\right)$ with $\mathcal{F}^{-1}$ the inverse Fourier transform. A simple idea for denoising in spectral analysis is to set some coefficients $y\left(\omega\right)$ to zero before reconstructing the signal. Figure 9 is an illustration of denoising using the thresholding rule. Selected parts of the frequency spectrum can easily be manipulated by filtering tools. For example, some can be attenuated, and others may be completely removed. Applying the inverse Fourier transform to this filtered spectrum leads to a filtered time series. Therefore, a smoothing signal can be easily performed by applying a low-pass filter, that is, by removing the higher frequencies. For example, we have represented two denoised signals of the S&P 500 index in Figure 9. For the first one, we use a 95% thresholding procedure whereas 99% of the Fourier coefficients are set to zero in the second case. One difficulty with this approach is the bad time location for low frequency signals and the bad frequency location for the high frequency signals. It is then difficult to localise when the trend (which is located in low frequencies) reverses. But the main drawback of spectral analysis is that it is not well suited to nonstationary processes (Martin and Flandrin, 1985, Fuentes, 2002, Oppenheim and Schafer, 2009).

[20] The detail of this derivation is shown in Appendix A.2.
[21] A break is the position where the signal trend changes.

Figure 8: $L_1$ versus $L_2$ filtering

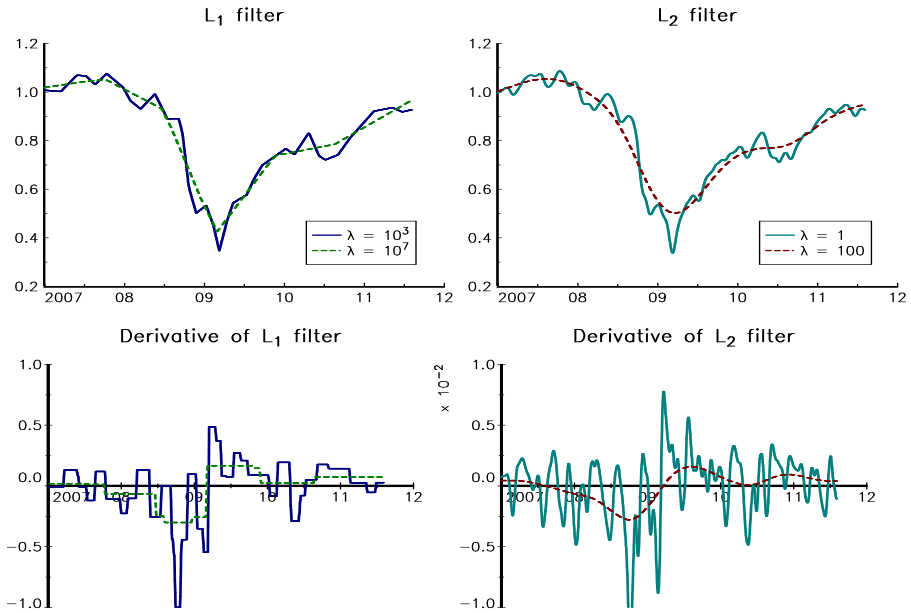
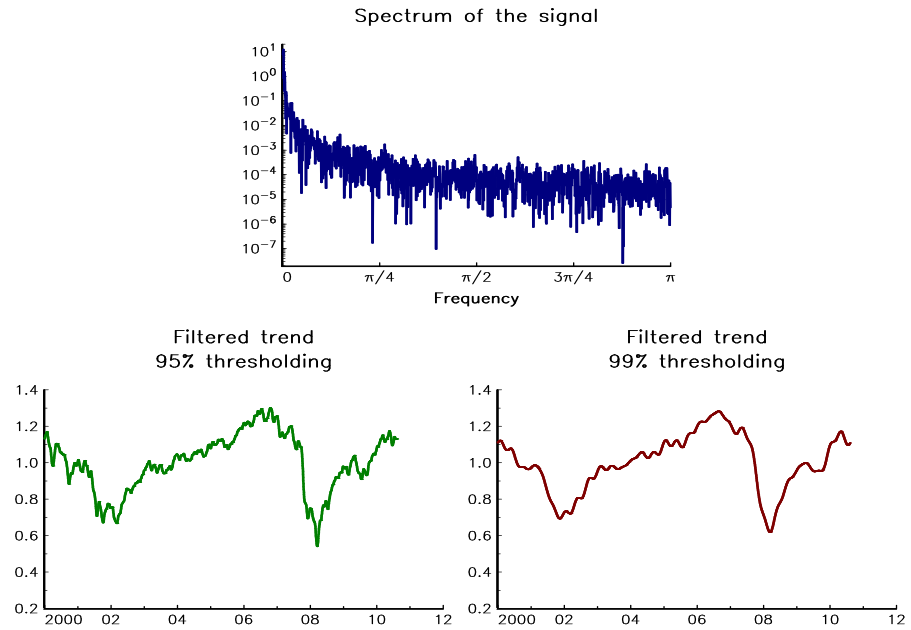



Figure 9: Spectral filtering



A solution consists of adopting a double dimension analysis, both in time and frequency. This approach corresponds to the wavelet analysis. The method of denoising is the same as described previously and the estimation of $x_t$ is done in three steps:

1. we compute the wavelet transform $\mathcal{W}$ of the original signal $y_t$ to obtain the wavelet coefficients $\omega = \mathcal{W}(y)$;
2. we modify the wavelet coefficients according to a denoising rule $D$:
$$\omega^\star = D(\omega)$$
3. We convert the modified wavelet coefficients into a new signal using the inverse wavelet transform $\mathcal{W}^{-1}$:
$$x = \mathcal{W}^{-1}(\omega^\star)$$

There are two principal choices in this approach. First, we have to specify which mother wavelet to use. Second, we have to define the denoising rule. Let $\omega^-$ and $\omega^+$ be two scalars with $0 < \omega^- < \omega^+$. Donoho and Johnstone (1995) define several shrinkage methods[22]:

- Hard shrinkage
$$\omega_i^\star = \omega_i \cdot \mathbf{1}\left\{|\omega_i| > \omega^+\right\}$$
- Soft shrinkage
$$\omega_i^\star = \operatorname{sgn}(\omega_i) \cdot \left(|\omega_i| - \omega^+\right)_+$$
- Semi-soft shrinkage
$$\omega_i^\star = \begin{cases} 0 & \text{si } |\omega_i| \leq \omega^- \\ \operatorname{sgn}(\omega_i)\left(\omega^+ - \omega^-\right)^{-1}\omega^+\left(|\omega_i| - \omega^-\right) & \text{si } \omega^- < |\omega_i| \leq \omega^+ \\ \omega_i & \text{si } |\omega_i| > \omega^+ \end{cases}$$
- Quantile shrinkage is a hard shrinkage method where $w^+$ is the $q^{\text{th}}$ quantile of the coefficients $|\omega_i|$.

Wavelet filtering is illustrated in Figure 10. We have computed the wavelet coefficients using the cascade algorithm of Mallat (1989) and the low-pass and high-pass filters of order 6 proposed by Daubechies (1992). The filtered trend is obtained using quantile shrinkage. In the first case, the noisy signal remains because we consider all the coefficients ($q = 0$). In the second and third cases, 95% and 99% of the wavelet coefficients are set to zero[23].

### 2.3.4 Other methods

Many other methods can be used to perform trend filtering. The most recent include, for example, singular spectrum analysis[24] (Vautard *et al.*, 1992), support vector machines[25] and empirical mode decomposition (Flandrin *et al.*, 2004). Moreover, we notice that traders sometimes use their own techniques (see, *inter alia*, Ehlers, 2001).

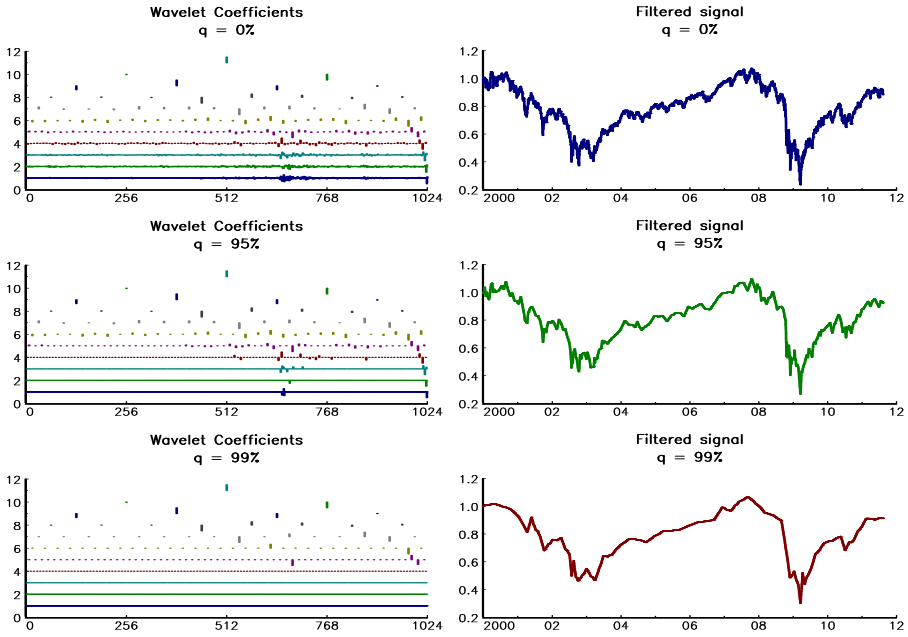
[22] In practice, the coefficients $\omega_i$ are standardised before being computed.

[23] It is interesting to note that the denoising procedure retains some wavelet coefficients corresponding to high and medium frequencies and located around the 2008 crisis.

[24] See Appendix A.5 for an illustration.

[25] A brief presentation is given in Appendix A.4.

Figure 10: Wavelet filtering




## 2.4 Multivariate filtering

Until now, we have assumed that the trend is specific to a financial asset. However, we may be interested in estimating the common trend of several financial assets. For example, if we wanted to estimate the trend of emerging markets equities, we could use a global index like the MSCI EM or extract the trend by considering several indices, e.g. the Bovespa index (Brazil), the RTS index (Russia), the Nifty index (India), the HSCEI index (China), etc. In this case, the trend-cycle model becomes:

$$\begin{pmatrix} y_t^{(1)} \\ \vdots \\ y_t^{(m)} \end{pmatrix} = x_t + \begin{pmatrix} \varepsilon_t^{(1)} \\ \vdots \\ \varepsilon_t^{(m)} \end{pmatrix}$$

where $y_t^{(j)}$ and $\varepsilon_t^{(j)}$ are respectively the signal and the noise of the financial asset $j$ and $x_t$ is the common trend. One idea for estimating the common trend is to obtain the mean of the specific trends:

$$\hat{x}_t = \frac{1}{m} \sum_{j=1}^{m} \hat{x}_t^{(j)}$$

If we consider moving average filtering, it is equivalent to applying the filter to the average filter[26] $\bar{y}_t = \frac{1}{m}\sum_{j=1}^{m} y_t^{(j)}$. This rule is also valid for some nonlinear filters such as $L_1$ filtering (see Appendix A.2). In what follows, we consider the two main alternative approaches developed in econometrics to estimate a (stochastic) common trend.

### 2.4.1 Error-correction model, common factors and the P-T decomposition

The econometrics of nonstationary time series may also help us to estimate a common trend. $y_t^{(j)}$ is said to be integrated of order 1 if the change $y_t^{(j)} - y_{t-1}^{(j)}$ is stationary. We will note $y_t^{(j)} \sim I(1)$ and $(1-L)\, y_t^{(j)} \sim I(0)$. Let us now define $y_t = \left(y_t^{(1)}, \ldots, y_t^{(m)}\right)$. The vector $y_t$ is cointegrated of rank $r$ if there exists a matrix $\beta$ of rank $r$ such that $z_t = \beta^\top y_t \sim I(0)$. In this case, we show that $y_t$ may be specified by an error-correction model (Engle and Granger, 1987):

$$\Delta y_t = \gamma z_{t-1} + \sum_{i=1}^{\infty} \Phi_i \Delta y_{t-i} + \zeta_t \tag{7}$$

where $\zeta_t$ is a $I(0)$ vector process. Stock and Watson (1988) propose another interesting representation of cointegration systems. Let $f_t$ be a vector of $r$ common factors which are $I(1)$. Therefore, we have:

$$y_t = A f_t + \eta_t \tag{8}$$

where $\eta_t$ is a $I(0)$ vector process and $f_t$ is a $I(1)$ vector process. One of the difficulties with this type of model is the identification step (Peña and Box, 1987). Gonzalo and Granger (1995) suggest defining a permanent-transitory (P-T) decomposition:

$$y_t = P_t + T_t$$

such that the permanent component $P_t$ is difference stationary, the transitory component $T_t$ is covariance stationary and $(\Delta P_t, T_t)$ satisfies a constrained autoregressive representation. Using this framework and some other conditions, Gonzalo and Granger show that we may obtain the representation (8) by estimating the relationship (7):

$$f_t = \breve{\gamma}^\top y_t \tag{9}$$

where $\breve{\gamma}^\top \gamma = 0$. They then follow the works of Johansen (1988, 1991) to derive the maximum likelihood estimator of $\breve{\gamma}$. Once we have estimated the relationship (9), it is also easy to identify the common trend[27] $\hat{x}_t$.

[26] We have:

$$\begin{aligned}
\hat{x}_t &= \frac{1}{m}\sum_{j=1}^{m}\sum_{i=0}^{n-1} \mathcal{L}_i y_{t-i}^{(j)} \\
&= \sum_{i=0}^{n-1} \mathcal{L}_i \left(\frac{1}{m}\sum_{j=1}^{m} y_{t-i}^{(j)}\right) \\
&= \sum_{i=0}^{n-1} \mathcal{L}_i \bar{y}_{t-i}
\end{aligned}$$

[27] If a common trend exists, it is necessarily one of the common factors.

### 2.4.2 Common stochastic trend model

Another idea is to consider an extension of the local linear trend model:

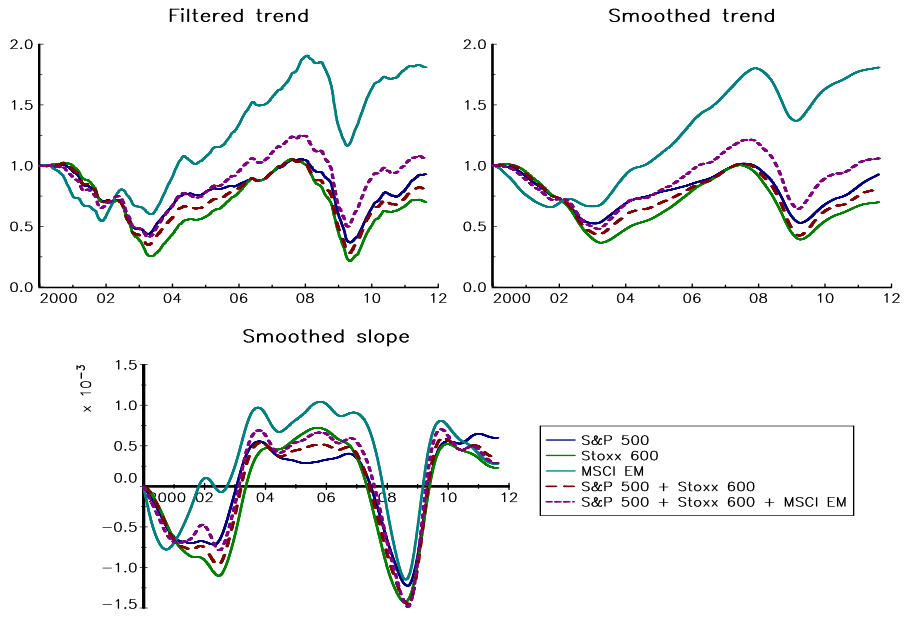
$$\begin{cases} y_t = \alpha x_t + \varepsilon_t \\ x_t = x_{t-1} + \mu_{t-1} + \sigma_\zeta \zeta_t \\ \mu_t = \mu_{t-1} + \sigma_\eta \eta_t \end{cases}$$

with $y_t = \left(y_t^{(1)}, \ldots, y_t^{(m)}\right)$, $\varepsilon_t = \left(\varepsilon_t^{(1)}, \ldots, \varepsilon_t^{(m)}\right) \sim \mathcal{N}(0, \Omega)$, $\zeta_t \sim \mathcal{N}(0,1)$ and $\eta_t \sim \mathcal{N}(0,1)$. Moreover, we assume that $\varepsilon_t$, $\zeta_t$ and $\eta_t$ are independent of each other. Given the parameters $(\alpha, \Omega, \sigma_\zeta, \sigma_\eta)$, we may run the Kalman filter to estimate the trend $x_t$ and the slope $\mu_t$ whereas the Kalman smoother allows us to estimate $x_{t-i}$ and $\mu_{t-i}$ at time $t$.

**Remark 6** *The case $\sigma_\eta = 0$ has been extensively studied by Chang et al. (2009). In particular, they show that $y_t$ is cointegrated with $\beta = \Omega^{-1}\Gamma$ and $\Gamma$ a $m \times (m-1)$ matrix such that $\Gamma^\top \Omega^{-1} \alpha = 0$ and $\Gamma^\top \Omega^{-1} \Gamma = I_{m-1}$. Using the P-T decomposition, they also found that the common stochastic trend is given by $\alpha^\top \Omega^{-1} y_t$, implying that the above averaging rule is not optimal.*

We come back to the example given in Figure 6 page 14. Using the second set of parameters, we now consider three stock indices: the S&P 500 index, the Stoxx 600 index and the MSCI EM index. For each index, we estimate the filtered trend. Moreover, using the previous common stochastic trend model[28], we estimate the common trend for the bivariate signal (S&P 500, Stoxx 600) and the trivariate signal (S&P 500, Stoxx 600, MSCI EM).

Figure 11: Multivariate Kalman filtering




[28] We assume that $\alpha_j$ takes the value 1 for the three signals.

# 3 Trend filtering in practice

## 3.1 The calibration problem

For the practical use of the trend extraction techniques discussed above, the calibration of filtering parameters is crucial. These calibrated parameters must incorporate our prediction requirement or they can be mapped to a commonly-known benchmark estimator. These constraints offer us some criteria for determining the optimal parameters for our expected prediction horizon. Below, we consider two possible calibration schemes based on these criteria.

### 3.1.1 Calibration based on prediction error

One idea for estimating the parameters of a model is to use statistical inference tools. Let us consider the local linear trend model. We may estimate the set of parameters $(\sigma_\varepsilon, \sigma_\zeta, \sigma_\eta)$ by maximising the log-likelihood function[29]:

$$\ell = \frac{1}{2}\sum_{t=1}^{n} \ln 2\pi + \ln F_t + \frac{v_t^2}{F_t}$$

where $v_t = y_t - \mathbb{E}_{t-1}[y_t]$ is the innovation process and $F_t = \mathbb{E}_{t-1}\left[v_t^2\right]$ is the variance of $v_t$. In Figure 12, we have reported the filtered and smoothed trend and slope estimated by the maximum likelihood method. We notice that the estimated components are more noisy than those obtained in Figure 6. We can explain this easily because maximum likelihood is based on the one-day innovation process. If we want to look at a longer trend, we have to consider the innovation process $v_t = y_t - \mathbb{E}_{t-h}[y_t]$ where $h$ is the horizon time. We have reported the slope for $h = 50$ days in Figure 12. It is very different from the slope corresponding to $h = 1$ day.

The problem is that the computation of the log-likelihood for the innovation process $v_t = y_t - \mathbb{E}_{t-h}[y_t]$ is trickier because there is generally no analytic expression. This is why we do not recommend this technology for trend filtering problems, because the trends estimated are generally very short-term. A better solution is to employ a cross-validation procedure to calibrate the parameters $\theta$ of the filters discussed above. Let us consider the calibration scheme presented in Figure 13. We divide our historical data into a training set and a validation set, which are characterised by two time parameters $T_1$ and $T_2$. The size of training set $T_1$ controls the precision of our calibration, for a fixed parameter $\theta$. For this training set, the value of the expectation of $\mathbb{E}_{t-h}[y_t]$ is computed. The second parameter

[29] Another way of estimating the parameters is to consider the log-likelihood function in the frequency domain analysis (Roncalli, 2010). In the case of the local linear trend model, the stationary form of $y_t$ is $S(y_t) = (1-L)^2 y_t$. We deduce that the associated log-likelihood function is:

$$\ell = -\frac{n}{2}\ln 2\pi - \frac{1}{2}\sum_{j=0}^{n-1} \ln f(\lambda_j) - \frac{1}{2}\sum_{j=0}^{n-1} \frac{I(\lambda_j)}{f(\lambda_j)}$$

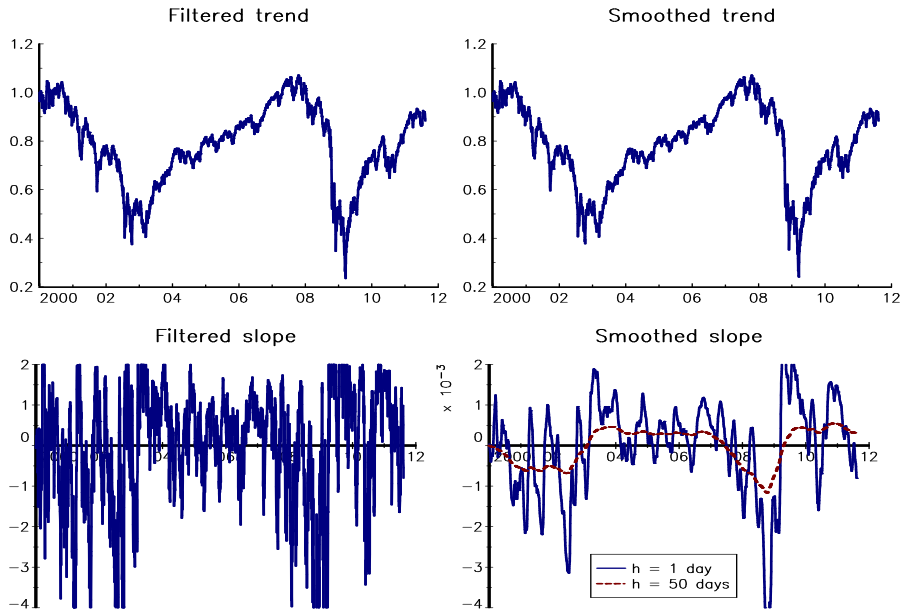
where $I(\lambda_j)$ is the periodogram of $S(y_t)$ and $f(\lambda)$ is the spectral density:

$$f(\lambda) = \frac{\sigma_\eta^2 + 2(1-\cos\lambda)\sigma_\zeta^2 + 4(1-\cos\lambda)^2\sigma_\varepsilon^2}{2\pi}$$

because we have:

$$S(y_t) = \sigma_\eta \eta_{t-1} + \sigma_\zeta (1-L)\zeta_t + \sigma_\varepsilon (1-L)^2 \varepsilon_t$$

Figure 12: Maximum likelihood of the trend and slope components

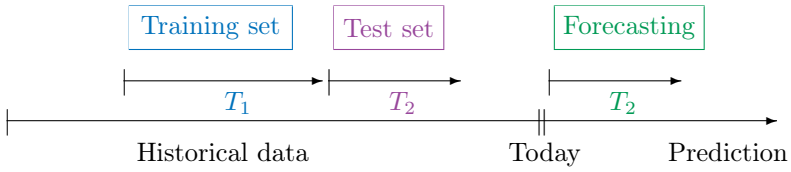



$T_2$ determines the size of the validation set, which is used to estimate the prediction error:

$$e\left(\theta; h\right) = \sum_{t=1}^{n-h} \left(y_t - \mathbb{E}_{t-h}\left[y_t\right]\right)^2$$

This quantity is directly related to the prediction horizon $h = T_2$ for a given investment strategy. The minimisation of the prediction error leads to the optimal value $\theta^\star$ of the filter parameters which will be used to predict the trend for the test set. For example, we apply this calibration scheme for $L_1$ filtering for $h$ equal to 50 days. Figure 14 illustrates the calibration procedure for the S&P 500 index with $T_1 = 400$ and $T_2 = 50$. Minimising the cumulative prediction error over the validation set gives the optimal value $\lambda^\star = 7.03$.

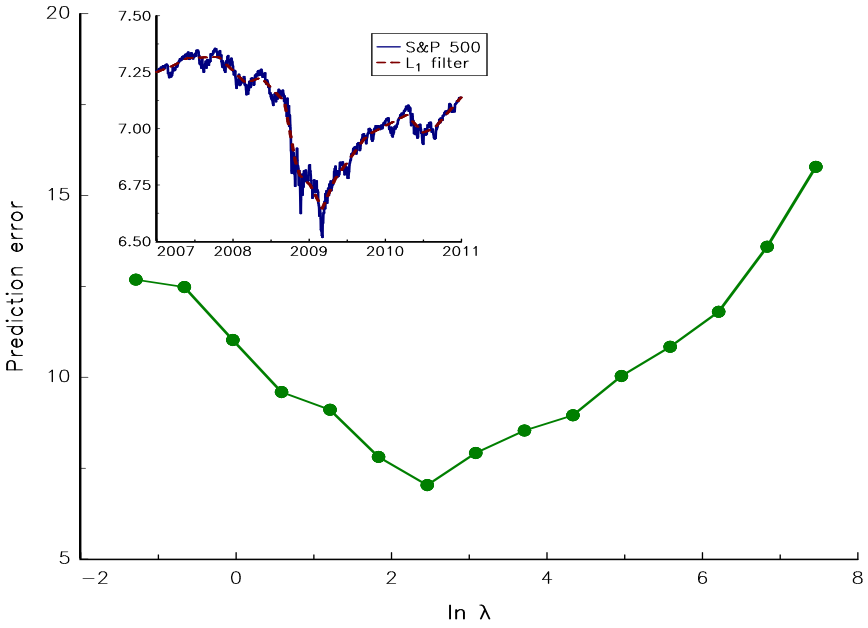
Figure 13: Cross-validation procedure for determining optimal parameters $\theta^\star$




### 3.1.2 Calibration based on benchmark estimator

The trend filtering algorithm can be calibrated with a benchmark estimator. In order to illustrate this idea, we present in this discussion the calibration procedure for $L_2$ filters by

Figure 14: Calibration procedure with the S&P 500 index for the $L_1$ filter




using spectral analysis. Though the $L_2$ filter provides an explicit solution which is a great advantage for numerical implementation, the calibration of the smoothing parameter $\lambda$ is not straightforward. We propose to calibrate the $L_2$ filter by comparing the spectral density of this filter with that obtained using the uniform moving average filter with horizon $n$ for which the spectral density is:

$$f^{\mathrm{MA}}(\omega) = \frac{1}{n^2}\left|\sum_{t=0}^{n-1} e^{-i\omega t}\right|^2$$

For the $L_2$ filter, the solution has the analytical form $\hat{x} = \left(1 + 2\lambda D^\top D\right)^{-1} y$. Therefore, the spectral density can also be computed explicitly:

$$f^{\mathrm{HP}}(\omega) = \left(\frac{1}{1 + 4\lambda\left(3 - 4\cos\omega + \cos 2\omega\right)}\right)^2$$

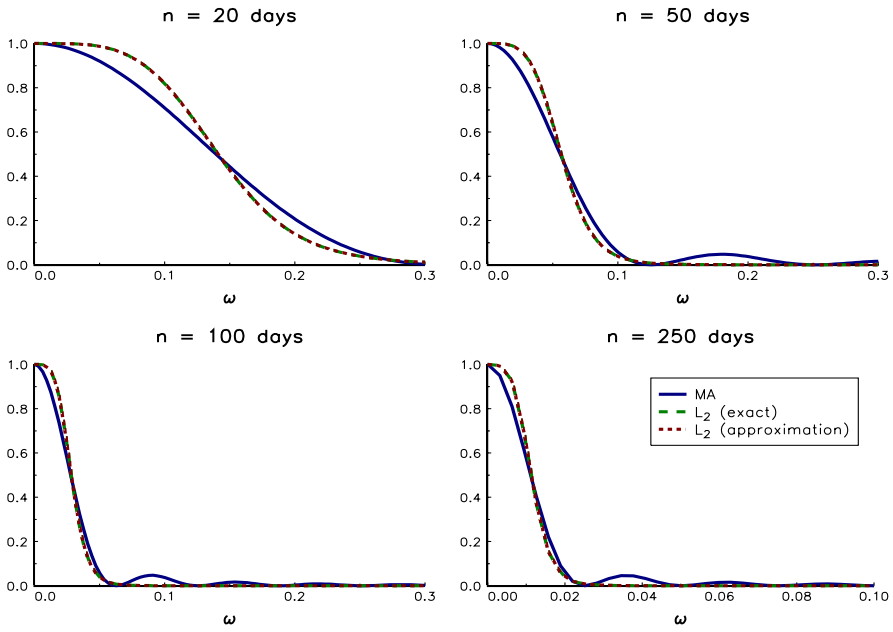
This spectral density can then be approximated by $1/\left(1 + 2\lambda\omega^4\right)^2$. Hence, the spectral width is $(2\lambda)^{-1/4}$ for the $L_2$ filter whereas it is $2\pi n^{-1}$ for the uniform moving average filter. The calibration of the $L_2$ filter could be achieved by matching these two quantities. Finally, we obtain the following relationship:

$$\lambda \propto \lambda_\star = \frac{1}{2}\left(\frac{n}{2\pi}\right)^4$$

In Figure 15, we represent the spectral density of the uniform moving average filter for different window sizes $n$. We also report the spectral density of the corresponding $L_2$ filters. To obtain this, we calibrated the optimal parameter $\lambda^\star$ by least square minimisation. In

Figure 16, we compare the optimal estimator $\lambda^\star$ with that corresponding to $10.27 \times \lambda_\star$. We notice that the approximation is very good[30].

Figure 15: Spectral density of moving average and $L_2$ filters




## 3.2 What about the variance of the estimator?

Let $\hat{\mu}_t$ be the estimator of the slope of the trend. There may be a confusion between the estimator of the slope and the estimated value of the slope (or the estimate). The estimator is a random variable and is defined by a probability distribution function. Based on the sample data, the estimator takes a value which is the estimate of the slope. Suppose that we obtain an estimate of 10%. It means that 10% is the most likely value of the slope given the data. But it does not mean that 10% is the true value of the slope.

### 3.2.1 Measuring the efficiency of trend filters

Let $\mu_t^0$ be the true value of the slope. In statistical inference, the quality of an estimator is defined by the mean squared error (or MSE):

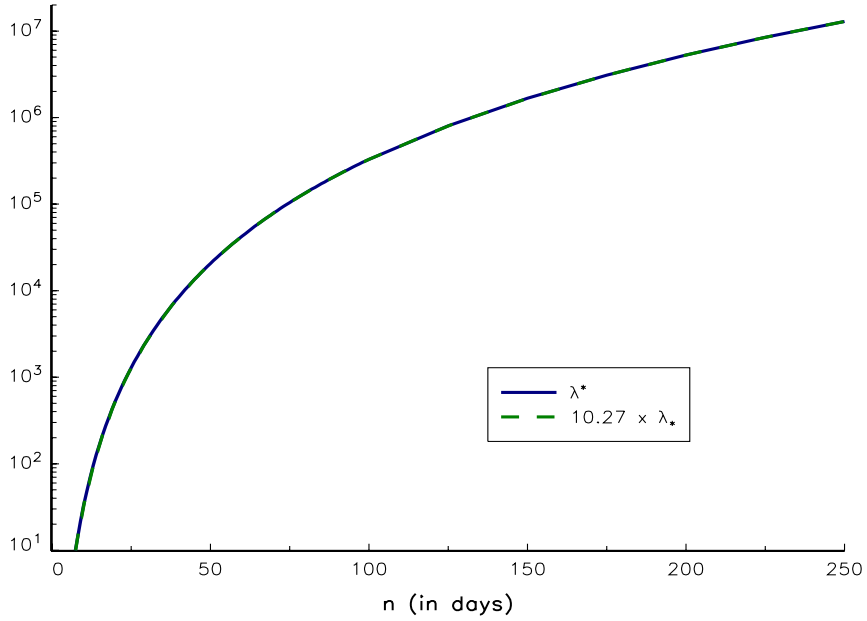
$$\text{MSE}\left(\hat{\mu}_t\right) = \mathbb{E}\left[\left(\hat{\mu}_t - \mu_t^0\right)^2\right]$$

It indicates how far the estimates are from the true value. We say that the estimator $\hat{\mu}_t^{(1)}$ is more efficient than the estimator $\hat{\mu}_t^{(2)}$ if its MSE is lower:

$$\hat{\mu}_t^{(1)} \succ \hat{\mu}_t^{(2)} \Leftrightarrow \text{MSE}\left(\hat{\mu}_t^{(1)}\right) \leq \text{MSE}\left(\hat{\mu}_t^{(2)}\right)$$

[30] We estimated the figure 10.27 using least squares.

Figure 16: Relationship between the value of $\lambda$ and the length of the moving average filter




We may decompose the MSE statistic into two components:

$$\text{MSE}\left(\hat{\mu}_t\right) = \mathbb{E}\left[\left(\hat{\mu}_t - \mathbb{E}\left[\hat{\mu}_t\right]\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}\left[\hat{\mu}_t\right] - \mu_t^0\right)\right]^2$$

The first component is the variance of the estimator $\text{var}\left(\hat{\mu}_t\right)$ whereas the second component is the square of the bias $B\left(\hat{\mu}_t\right)$. Generally, we are interested by estimators that are unbiased $\left(B\left(\hat{\mu}_t\right) = 0\right)$. If this is the case, comparing two estimators is equivalent to comparing their variances.

Let us assume that the price process is a geometric Brownian motion:

$$\mathrm{d}S_t = \mu^0 S_t \,\mathrm{d}t + \sigma^0 S_t \,\mathrm{d}W_t$$

In this case, the slope of the trend is constant and is equal to $\mu^0$. In Figure 17, we have reported the probability density function of the estimator $\hat{\mu}_t$ when the true slope $\mu^0$ is 10%. We consider the estimator based on a uniform moving average filter of length $n$. First, we notice that using filters is better than using the noisy signal. We also observe that the variance of the estimators increases with the parameter $\sigma^0$ and decreases with the length $n$.

### 3.2.2 Trend detection versus trend filtering

In the previous paragraph, we saw that an estimate of the trend may not be significant if the variance of the estimator is too large. Before computing an estimate of the trend, we then have to decide if there is a trend or not. This process is called trend detection. Mann (1945) considers the following statistic:

$$\mathbb{S}_t^{(n)} = \sum_{i=0}^{n-2} \sum_{j=i+1}^{n-1} \text{sgn}\left(y_{t-i} - y_{t-j}\right)$$

Figure 17: Density of the estimator $\hat{\mu}_t$

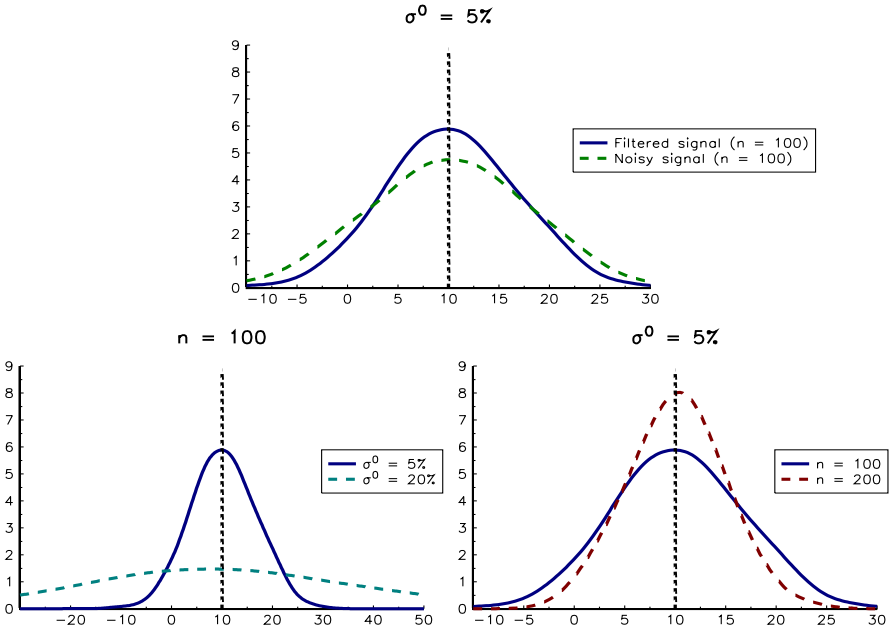
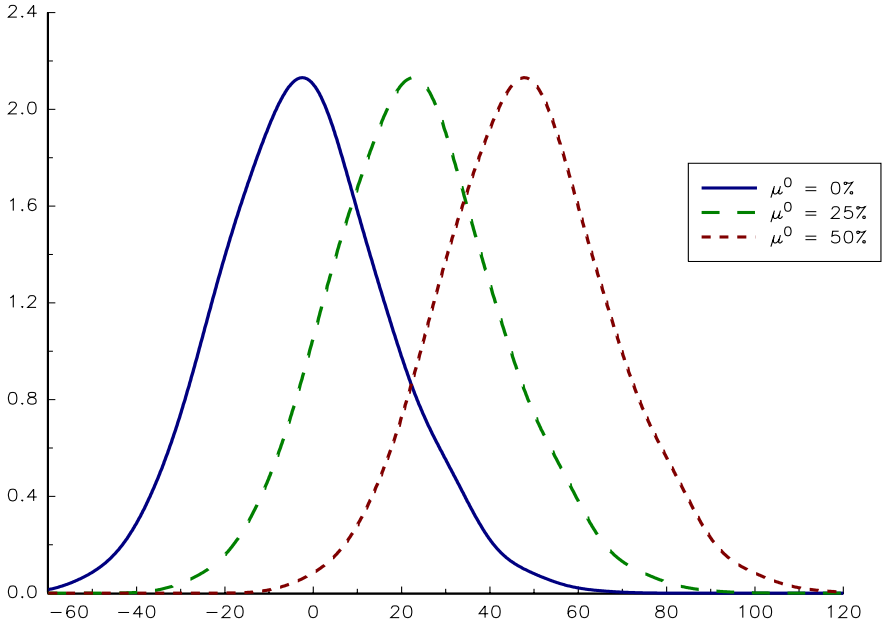



Figure 18: Impact of $\mu^0$ on the estimator $\hat{\mu}_t$



with $\text{sgn}(y_{t-i} - y_{t-j}) = 1$ if $y_{t-i} > y_{t-j}$ and $\text{sgn}(y_{t-i} - y_{t-j}) = -1$ if $y_{t-i} < y_{t-j}$. We have[31]:

$$\text{var}\left(\mathbb{S}_t^{(n)}\right) = \frac{n(n-1)(2n+5)}{18}$$

We can show that:

$$-\frac{n(n+1)}{2} \leq \mathbb{S}_t^{(n)} \leq \frac{n(n+1)}{2}$$

The bounds are reached if $y_t < y_{t-i}$ (negative trend) or $y_t > y_{t-i}$ (positive trend) for $i \in \mathbb{N}^*$. We can then normalise the score:

$$\mathcal{S}_t^{(n)} = \frac{2\mathbb{S}_t^{(n)}}{n(n+1)}$$

$\mathcal{S}_t^{(n)}$ takes the value $+1$ (or $-1$) if we have a perfect positive (or negative) trend. If there is no trend, it is obvious that $\mathbb{S}_t^{(n)} \simeq 0$. Under this null hypothesis, we have:

$$Z_t^{(n)} \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1)$$

with:

$$Z_t^{(n)} = \frac{\mathbb{S}_t^{(n)}}{\sqrt{\text{var}\left(\mathbb{S}_t^{(n)}\right)}}$$

In Figure 19, we reported the normalised score $\mathcal{S}_t^{(n)}$ for the S&P 500 index and different values of $n$. Statistics relating to the null hypothesis are given in Table 2 for the study period. We notice that we generally reject the hypothesis that there is no trend when we consider a period of one year. The number of cases when we observe a trend increases if we consider a shorter period. For example, if $n$ is equal to 10 days, we accept the hypothesis that there is no trend in 42% of cases when the confidence level $\alpha$ is set to 90%.

Table 2: Frequencies of rejecting the null hypothesis with confidence level $\alpha$

| $\alpha$ | 90% | 95% | 99% |
|---|---|---|---|
| $n = 10$ days | 58.06% | 49.47% | 29.37% |
| $n = 3$ months | 85.77% | 82.87% | 76.68% |
| $n = 1$ year | 97.17% | 96.78% | 95.33% |

**Remark 7** *We have reported the statistic $\mathcal{S}_t^{(10)}$ against the trend estimate[32] $\hat{\mu}_t$ for the S&P 500 index since January 2000. We notice that $\hat{\mu}_t$ may be positive whereas $\mathcal{S}_t^{(10)}$ is negative. This illustrates that a trend measurement is just an estimate. It does not mean that a trend exists.*

[31] If there are some tied sequences ($y_{t-i} = y_{t-i-1}$), the formula becomes:

$$\text{var}\left(\mathbb{S}_t^{(n)}\right) = \frac{1}{18}\left(n(n-1)(2n+5) - \sum_{k=1}^{g} n_k(n_k-1)(2n_k+5)\right)$$

with $g$ the number of tied sequences and $n_k$ the number of data points in the $k^{\text{th}}$ tied sequence.

[32] It is computed with a uniform moving average of 10 days.

Figure 19: Trend detection for the S&P 500 index

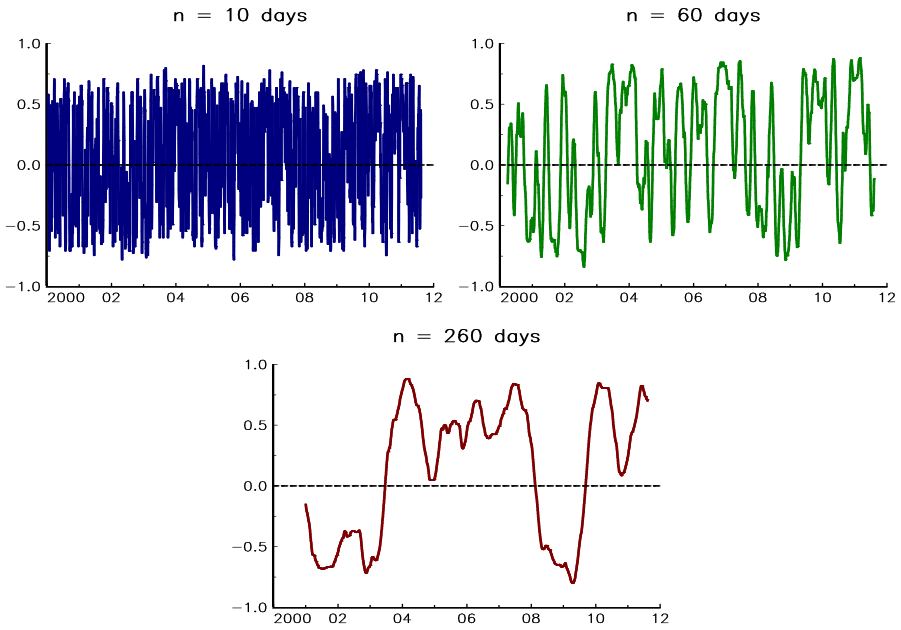
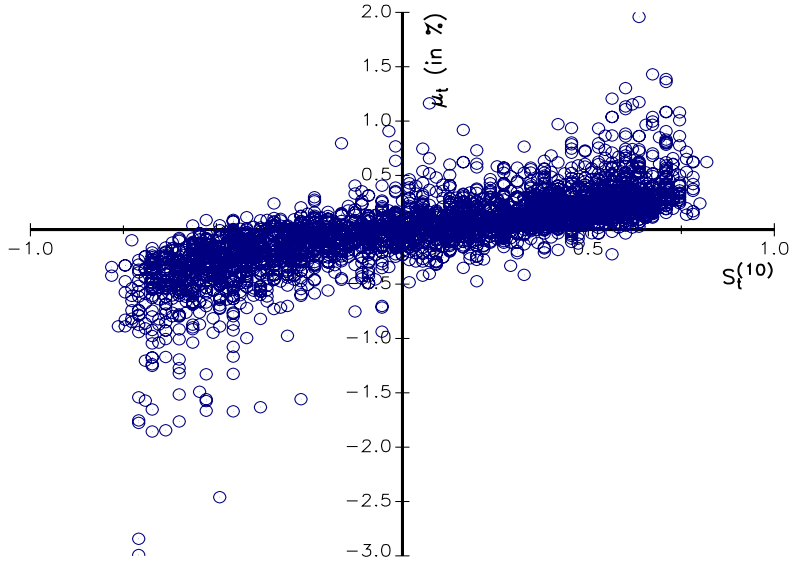


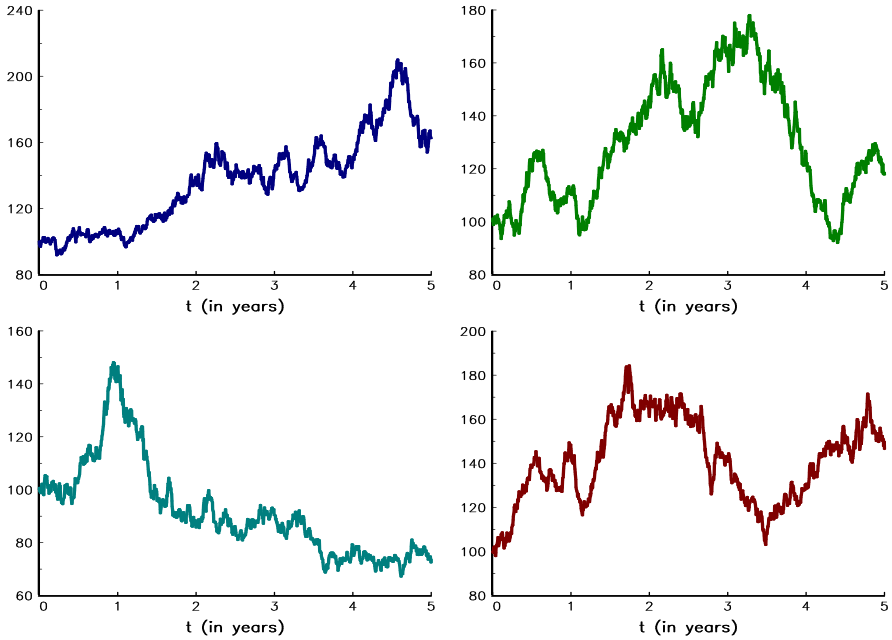

Figure 20: Trend detection versus trend filtering



## 3.3 From trend filtering to trend forecasting

There are two possible applications for the trend following problem. First, trend filtering can analyse the past. A noisy signal can be transformed into a smoother signal, which can be interpreted more easily. An ex-post analysis of this kind can, for instance, clearly separate increasing price periods from decreasing price periods. This analysis can be performed on any time series, or even on a random walk. For example, we have reported four simulations of a geometric Brownian motion without drift and annual volatility of 20% in Figure 21. In this context, trend filtering could help us to estimate the different trends in the past.

Figure 21: Four simulations of a geometric Brownian motion without drift

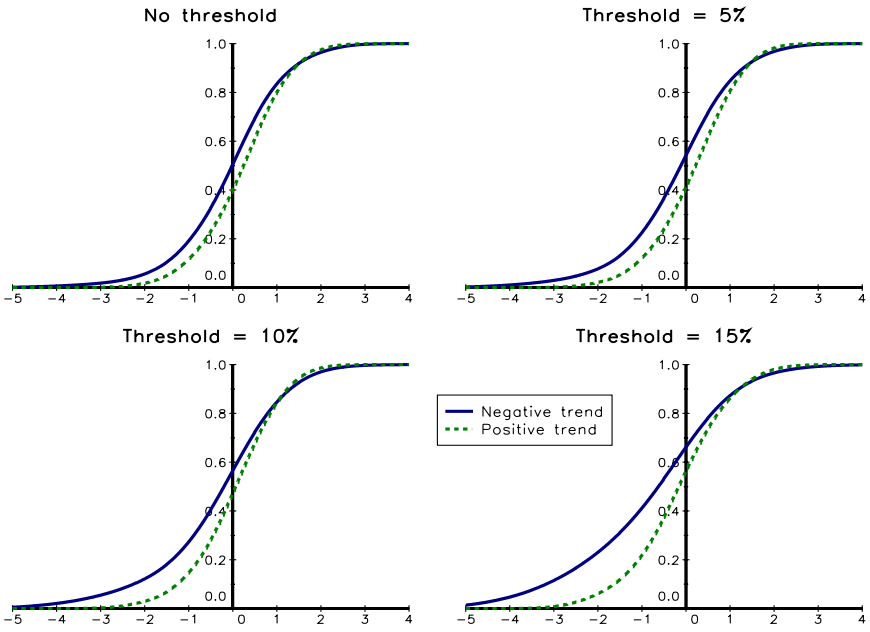



On the other hand, trend analysis may be used as a predictive tool. Prediction is a much more ambitious objective than analysing the past. It cannot be performed on any time series. For instance, trend following predictions suppose that the last observed trend influences future returns. More precisely, these predictors suppose that positive (or negative) trends are more likely to be followed by positive (or negative) returns. Such an assumption has to be tested empirically. For example, it is obvious that the time series in Figure 21 exhibit certain trends, whereas we know that there is no trend in a geometric Brownian motion without drift. Thus, we may still observe some trends in an ex-post analysis. It does not mean, however, that trends will persist in the future.

The persistence of trends is tested here in a simple framework for major financial indices[33]. For each of these indices the average one-month returns are separated into two sets. The first set includes one-month returns that immediately follow a positive three-month return (this is negative for the second set). The average one-month return is computed for each of these two sets, and the results are given in Table 3. These results clearly show

[33]The study period begins in January 1995 (January 1999 for the MSCI EM) and finish in October 2011.

Figure 22: Distribution of the conditional standardised monthly return




that, on average, higher returns can be expected after a positive three-month return than after a negative three-month period. Therefore, observation of the current trend may have a predictive value for the indices under consideration. Moreover, we consider the distribution of the one-month returns, based on past three-month returns. Figure 22 illustrates the case of the GSCI index. In the first quadrant, the one-month returns are divided into two sets, depending on whether the previous three-month return is positive or negative. The cumulative distributions of these two sets are shown. In the second quadrant, we consider, on the one hand, the distribution of one-month returns following a three-month return below $-5\%$ and, on the other hand, the distribution of returns following a three-month return exceeding $+5\%$. The same procedure is repeated in the other quadrants, for a 10% and a 15% threshold. This simple test illustrates the usefulness of trend following strategies. Here, trends seem persistent enough to study such strategies. Of course, on other time scales or for other assets, one may obtain opposite results that would support contrarian strategies.

Table 3: Average one-month conditional return based on past trends

| Trend | Positive | Negative | Difference |
|---|---|---|---|
| Eurostoxx 50 | 1.1% | 0.2% | 0.9% |
| S&P 500 | 0.9% | 0.5% | 0.4% |
| MSCI WORLD | 0.6% | −0.3% | 1.0% |
| MSCI EM | 1.9% | −0.3% | 2.2% |
| TOPIX | 0.4% | −0.4% | 0.9% |
| EUR/USD | 0.2% | −0.2% | 0.4% |
| USD/JPY | 0.2% | −0.2% | 0.4% |
| GSCI | 1.3% | −0.4% | 1.6% |

# 4 Conclusion

The ultimate goal of trend filtering in finance is to design portfolio strategies that may benefit from these trends. But the path between trend measurement and portfolio allocation is not straightforward. It involves studies and explanations that would not fit in this paper. Nevertheless, let us point out some major issues. Of course, the first problem is the selection of the trend filtering method. This selection may lead to a single procedure or to a pool of methods. The selection of several methods raises the question of an aggregation procedure. This can be done through averaging or dynamic model selection, for instance. The resulting trend indicator is meant to forecast future asset returns at a given horizon.

Intuitively, an investor should buy assets with positive return forecasts and sell assets with negative forecasts. But the size of each long or short position is a quantitative problem that requires a clear investment process. This process should take into account the risk entailed by each position, compared with the expected return. Traditionally, individual risks can be calculated in relation to asset volatility. A correlation matrix can aggregate those individual risks into a global portfolio risk. But in the case of a multi-asset trend following strategy, should we consider the correlation of assets or the correlation of each individual strategy? These may be quite different, as the correlations between strategies are usually smaller than the correlations between assets in absolute terms. Even when the portfolio risks can be calculated, the distribution of those risks between assets or strategies remains an open problem. Clearly, this distribution should take into account the individual risks, their correlations and the expected return of each asset. But there are many competing allocation procedures, such as Markowitz portfolio theory or risk budgeting methods.

In addition, the total amount of risk in the portfolio must be decided. The average target volatility of the portfolio is closely related to the risk aversion of the final investor. But this total amount of risk may not be constant over time, as some periods could bring higher expected returns than others. For example, some funds do not change the average size of their positions during period of high market volatility. This increases their risks, but they consider that their return opportunities, even when risk-adjusted, are greater during those periods. On the contrary, some investors reduce their exposure to markets during volatility peaks, in order to limit their potential drawdowns. Anyway, any consistent investment process should measure and control the global risk of the portfolio.

These are just a few questions relating to trend following strategies. Many more arise in practical cases, such as execution policies and transaction cost management. Each of these issues must be studied in depth, and re-examined on a regular basis. This is the essence of quantitative management processes.

# A Statistical complements

## A.1 State space model and Kalman filtering

A state space model is defined by a transition equation and a measurement equation. In the measurement equation, we postulate the relationship between an observable vector and a state vector, while the transition equation describes the generating process of the state variables. The state vector $\alpha_t$ is generated by a first-order Markov process of the form:

$$\alpha_t = T_t \alpha_{t-1} + c_t + R_t \eta_t$$

where $\alpha_t$ is the vector of the $m$ state variables, $T_t$ is a $m \times m$ matrix, $c_t$ is a $m \times 1$ vector and $R_t$ is a $m \times p$ matrix. The measurement equation of the state-space representation is:

$$y_t = Z_t \alpha_t + d_t + \varepsilon_t$$

where $y_t$ is a $n$-dimension time series, $Z_t$ is a $n \times m$ matrix, $d_t$ is a $n \times 1$ vector. $\eta_t$ and $\varepsilon_t$ are assumed to be white noise processes of dimensions $p$ and $n$ respectively. These two last uncorrelated processes are Gaussian with zero mean and respective covariance matrices $Q_t$ and $H_t$. $\alpha_0 \sim \mathcal{N}(a_0, P_0)$ describes the initial position of the state vector. We define $a_t$ and $a_{t|t-1}$ as the optimal estimators of $\alpha_t$ based on all the information available respectively at time $t$ and $t-1$. Let $P_t$ and $P_{t|t-1}$ be the associated covariance matrices[34]. The Kalman filter consists of the following set of recursive equations (Harvey, 1990):

$$\begin{cases} a_{t|t-1} = T_t a_{t-1} + c_t \\ P_{t|t-1} = T_t P_{t-1} T_t^\top + R_t Q_t R_t^\top \\ y_{t|t-1} = Z_t a_{t|t-1} + d_t \\ v_t = y_t - y_{t|t-1} \\ F_t = Z_t P_{t|t-1} Z_t^\top + H_t \\ a_t = a_{t|t-1} + P_{t|t-1} Z_t^\top F_t^{-1} v_t \\ P_t = \left(I_m - P_{t|t-1} Z_t^\top F_t^{-1} Z_t\right) P_{t|t-1} \end{cases}$$

where $v_t$ is the innovation process with covariance matrix $F_t$ and $y_{t|t-1} = \mathbb{E}_{t-1}[y_t]$. Harvey (1989) shows that we can obtain $a_{t+1|t}$ directly from $a_{t|t-1}$:

$$a_{t+1|t} = (T_{t+1} - K_t Z_t) a_{t|t-1} + K_t y_t + (c_{t+1} - K_t d_t)$$

where $K_t = T_{t+1} P_{t|t-1} Z_t^\top F_t^{-1}$ is the matrix of gain. We also have:

$$a_{t+1|t} = T_{t+1} a_{t|t-1} + c_{t+1} + K_t \left(y_t - Z_t a_{t|t-1} - d_t\right)$$

Finally, we obtain:

$$\begin{cases} y_t &= Z_t a_{t|t-1} + d_t + v_t \\ a_{t+1|t} &= T_{t+1} a_{t|t-1} + c_{t+1} + K_t v_t \end{cases}$$

This system is called the innovation representation.

Let $t^\star$ be a fixed given date. We define $a_{t|t^\star} = \mathbb{E}_{t^\star}[\alpha_t]$ and $P_{t|t^\star} = \mathbb{E}_{t^\star}\left[\left(a_{t|t^\star} - \alpha_t\right)\left(a_{t|t^\star} - \alpha_t\right)^\top\right]$ with $t \leq t^\star$. We have $a_{t^\star|t^\star} = a_{t^\star}$ and $P_{t^\star|t^\star} = P_{t^\star}$. The Kalman smoother is then defined by the following set of recursive equations:

$$\begin{aligned} P_t^* &= P_t T_{t+1}^\top P_{t+1|t}^{-1} \\ a_{t|t^\star} &= a_t + P_t^* \left(a_{t+1|t^\star} - a_{t+1|t}\right) \\ P_{t|t^\star} &= P_t + P_t^* \left(P_{t+1|t^\star} - P_{t+1|t}\right) P_t^{*\top} \end{aligned}$$

[34] We have $a_t = \mathbb{E}_t[\alpha_t]$, $a_{t|t-1} = \mathbb{E}_{t-1}[\alpha_t]$, $P_t = \mathbb{E}_t\left[(a_t - \alpha_t)(a_t - \alpha_t)^\top\right]$ and $P_{t|t-1} = \mathbb{E}_{t-1}\left[\left(a_{t|t-1} - \alpha_t\right)\left(a_{t|t-1} - \alpha_t\right)^\top\right]$ where $\mathbb{E}_t$ indicates the conditional expectation operator.

## A.2 $L_1$ filtering

### A.2.1 The dual problem

The $L_1$ filtering problem can be solved by considering the dual problem which is a QP programme. We first rewrite the primal problem with a new variable $z = D\hat{x}$:

$$\begin{array}{ll} \min & \frac{1}{2}\left\|y-\hat{x}\right\|_2^2 + \lambda\left\|z\right\|_1 \\ \text{u.c.} & z = D\hat{x} \end{array}$$

We now construct the Lagrangian function with the dual variable $\nu \in \mathbb{R}^{n-2}$:

$$\mathcal{L}\left(\hat{x}, z, v\right) = \frac{1}{2}\left\|y-\hat{x}\right\|_2^2 + \lambda\left\|z\right\|_1 + \nu^\top\left(D\hat{x}-z\right)$$

The dual objective function is obtained in the following way:

$$\inf_{\hat{x},z}\mathcal{L}\left(\hat{x}, z, \nu\right) = -\frac{1}{2}\nu^\top DD^\top\nu + y^\top D^\top\nu$$

for $-\lambda\mathbf{1} \leq \nu \leq \lambda\mathbf{1}$. According to the Kuhn-Tucker theorem, the initial problem is equivalent to the dual problem:

$$\begin{array}{ll} \min & \frac{1}{2}\nu^\top DD^\top\nu - y^\top D^\top\nu \\ \text{u.c.} & -\lambda\mathbf{1} \leq \nu \leq \lambda\mathbf{1} \end{array}$$

This QP programme can be solved by a traditional Newton algorithm or by interior-point methods, and finally, the solution of the trend is:

$$\hat{x} = y - D^\top\nu$$

### A.2.2 Solving using interior-point algorithms

We briefly present the interior-point algorithm of Boyd and Vandenberghe (2009) in the case of the following optimisation problem:

$$\begin{array}{ll} \min & f_0\left(\theta\right) \\ \text{u.c.} & \left\{\begin{array}{l} A\theta = b \\ f_i\left(\theta\right) < 0 \quad \text{for } i = 1, \ldots, m \end{array}\right. \end{array}$$

where $f_0, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ are convex and twice continuously differentiable and $\text{rank}\left(A\right) = p < n$. The inequality constraints will become implicit if the problem is rewritten as:

$$\begin{array}{ll} \min & f_0\left(\theta\right) + \sum_{i=1}^{m}\mathcal{I}_-\left(f_i\left(\theta\right)\right) \\ \text{u.c.} & A\theta = b \end{array}$$

where $\mathcal{I}_-\left(u\right) : \mathbb{R} \to \mathbb{R}$ is the non-positive indicator function[35]. This indicator function is discontinuous, so the Newton method can not be applied. In order to overcome this problem, we approximate $\mathcal{I}_-\left(u\right)$ using the logarithmic barrier function $\mathcal{I}_-^\star\left(u\right) = -\tau^{-1}\ln\left(-u\right)$

[35] We have:

$$\mathcal{I}_-\left(u\right) = \left\{\begin{array}{ll} 0 & u \leq 0 \\ \infty & u > 0 \end{array}\right.$$

with $\tau \to \infty$. Finally the Kuhn-Tucker condition for this approximation problem gives $r_t(\theta, \lambda, \nu) = 0$ with:

$$r_\tau(\theta, \lambda, \nu) = \begin{pmatrix} \nabla f_0(\theta) + \nabla f(\theta)^\top \lambda + A^\top \nu \\ -\operatorname{diag}(\lambda) f(\theta) - \tau^{-1}\mathbf{1} \\ A\theta - b \end{pmatrix}$$

The solution of $r_\tau(\theta, \lambda, \nu) = 0$ can be obtained using Newton's iteration for the triple $\pi = (\theta, \lambda, \nu)$:

$$r_\tau(\pi + \Delta\pi) \simeq r_\tau(\pi) + \nabla r_\tau(\pi) \Delta\pi = 0$$

This equation gives the Newton step $\Delta\pi = -\nabla r_\tau(\pi)^{-1} r_\tau(\pi)$, which defines the search direction.

### A.2.3 The multivariate case

In the multivariate case, the primal problem is:

$$\begin{array}{ll} \min & \frac{1}{2}\sum_{j=1}^{m} \left\| y^{(j)} - \hat{x} \right\|_2^2 + \lambda \|z\|_1 \\ \text{u.c.} & z = D\hat{x} \end{array}$$

The dual objective function becomes:

$$\inf_{\hat{x},z} \mathcal{L}(\hat{x}, z, \nu) = -\frac{1}{2}\nu^\top DD^\top \nu + \bar{y}^\top D^\top \nu + \frac{1}{2}\sum_{j=1}^{m} \left(y^{(j)} - \bar{y}\right)^\top \left(y^{(j)} - \bar{y}\right)$$

for $-\lambda\mathbf{1} \leq \nu \leq \lambda\mathbf{1}$. According to the Kuhn-Tucker theorem, the initial problem is equivalent to the dual problem:

$$\begin{array}{ll} \min & \frac{1}{2}\nu^\top DD^\top \nu - \bar{y}^\top D^\top \nu \\ \text{u.c.} & -\lambda\mathbf{1} \leq \nu \leq \lambda\mathbf{1} \end{array}$$

The solution is then $\hat{x} = \bar{y} - D^\top \nu$.

### A.2.4 The scaling of the smoothing parameter

We can attempt to estimate the order of magnitude of the parameter $\lambda_{\max}$ by considering the continuous case. We assume that the signal is a process $W_t$. The value of $\lambda_{\max}$ in the discrete case is defined by:

$$\lambda_{\max} = \left\| \left(DD^\top\right)^{-1} Dy \right\|_\infty$$

can be considered as the first primitive $I_1(T) = \int_0^T W_t \, \mathrm{d}t$ of the process $W_t$ if $D = D_1$ ($L_1 - C$ filtering) or the second primitive $I_2(T) = \int_0^T \int_0^t W_s \, \mathrm{d}s \, \mathrm{d}t$ of $W_t$ if $D = D_2$ ($L_1 - T$ filtering). We have:

$$\begin{aligned} I_1(T) &= \int_0^T W_t \, \mathrm{d}t \\ &= W_T T - \int_0^T t \, \mathrm{d}W_t \\ &= \int_0^T (T - t) \, \mathrm{d}W_t \end{aligned}$$

The process $I_1(T)$ is a Wiener integral (or a Gaussian process) with variance:

$$\mathbb{E}\left[I_1^2(T)\right] = \int_0^T (T-t)^2 \, \mathrm{d}t = \frac{T^3}{3}$$

In this case, we expect that $\lambda_{\max} \sim T^{3/2}$. The second order primitive can be calculated in the following way:

$$\begin{aligned}
I_2(T) &= \int_0^T I_1(t) \, \mathrm{d}t \\
&= I_1(T)\, T - \int_0^T t \, \mathrm{d}I_1(T) \\
&= I_1(T)\, T - \int_0^T tW_t \, \mathrm{d}t \\
&= I_1(T)\, T - \frac{T^2}{2} W_T + \int_0^T \frac{t^2}{2} \, \mathrm{d}W_t \\
&= -\frac{T^2}{2} W_T + \int_0^T \left(T^2 - Tt + \frac{t^2}{2}\right) \mathrm{d}W_t \\
&= \frac{1}{2} \int_0^T (T-t)^2 \, \mathrm{d}W_T
\end{aligned}$$

This quantity is again a Gaussian process with variance:

$$\mathbb{E}[I_2^2(T)] = \frac{1}{4} \int_0^T (T-t)^4 \, \mathrm{d}t = \frac{T^5}{20}$$

In this case, we expect that $\lambda_{\max} \sim T^{5/2}$.

## A.3 Wavelet analysis

The time analysis can detect anomalies in time series, such as a market crash on a specific date. The frequency analysis detects repeated sequences in a signal. The double dimension analysis makes it possible to coordinate time and frequency detection, as we use a larger time window than a smaller frequency interval (see Figure 23). In this area, the uncertainty of localisation is $1/\mathrm{d}t$, with $\mathrm{d}t$ the sampling step and $f = 1/\mathrm{d}t$ the sampling frequency. The wavelet transform can be a solution to analysing time series in terms of the time-frequency dimension.

The first wavelet approach appeared in the early eighties in seismic data analysis. The term *wavelet* was introduced in the scientific community by Grossmann and Morlet (1984). Since 1986, a great deal of theoretical research, including wavelets, has been developed. The wavelet transform uses a basic function, called the *mother wavelet*, then dilates and translates it to capture features that are local in time and frequency. The distribution of the time-frequency domain with respect to the wavelet transform is long in time when capturing low frequency events and long in frequency when capturing high frequency events. As an example, we represent some mother wavelets in Figure 24.

The aim of wavelet analysis is to separate signal trends and details. These different components can be distinguished by different levels of resolution or different sizes/scales of detail. In this sense, it generates a phase space decomposition which is defined by two

Figure 23: Time-frequency dimension

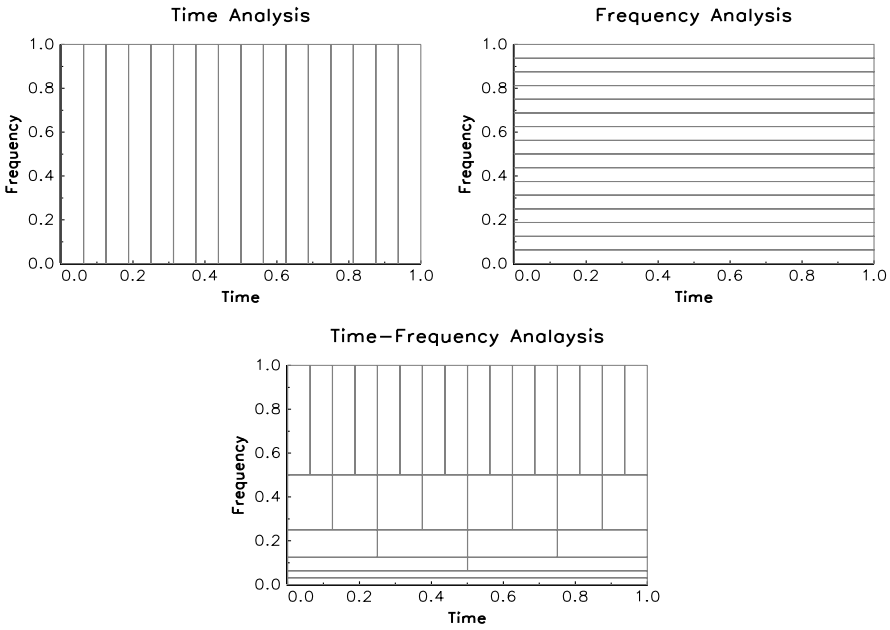
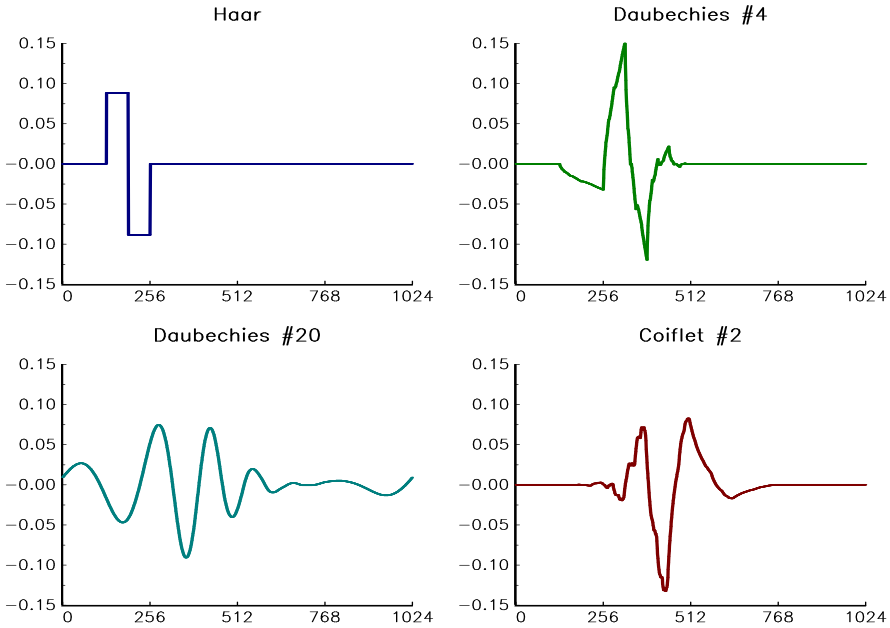



Figure 24: Some mother wavelets



parameters *(scale* and *location)* in opposition to a Fourier decomposition. A wavelet $\psi(t)$ is a function of time $t$ such that:

$$\begin{aligned}\int_{-\infty}^{+\infty} \psi(t)\, \mathrm{d}t &= 0 \\ \int_{-\infty}^{+\infty} |\psi(t)|^2\, \mathrm{d}t &= 1\end{aligned}$$

The continuous wavelet transform is a function of two variables $W(u,s)$ and is given by projecting the time series $x(t)$ onto a particular wavelet $\psi$ by:

$$W(u,s) = \int_{-\infty}^{+\infty} x(t)\, \psi_{u,s}(t)\, \mathrm{d}t$$

with:

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right)$$

which corresponds to the mother wavelet translated by $u$ (location parameter) and dilated by $s$ (scale parameter). If the wavelet satisfies the previous properties, the inverse operation may be performed to produce the original signal from its wavelet coefficients:

$$x(t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} W(u,s)\, \psi(u,s)\, \mathrm{d}u\, \mathrm{d}s$$

The continuous wavelet transform of a time series signal $x(t)$ gives an infinite number of coefficients $W(u,s)$ where $u \in \mathbb{R}$ and $s \in \mathbb{R}^+$, but many coefficients are close or equal to zero. The discrete wavelet transform can be used to decompose a signal into a finite number of coefficients where we use $s = 2^{-j}$ as the scale parameter and $u = k2^{-j}$ as the location parameter with $j \in \mathbb{Z}$ and $k \in \mathbb{Z}$. Therefore $\psi_{u,s}(t)$ becomes:

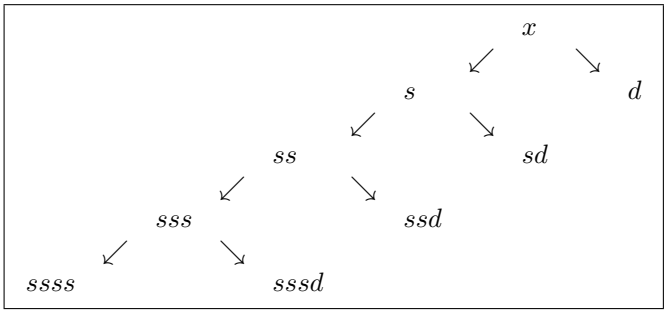
$$\psi_{j,k}(t) = 2^{\frac{j}{2}} \psi\left(2^j t - k\right)$$

where $j = 1, 2, ..., J$ in a $J$-level decomposition. The wavelet representation of a discrete signal $x(t)$ is given by:

$$x(t) = s_{(0)} \phi(t) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^{j-1}} d_{(j),k} \psi_{j,k}(t)$$

where $\phi(t) = 1$ if $t \in [0,1]$ and $J$ is the number of multi-resolution levels. Therefore, computing the wavelet transform of the discrete signal is equivalent to compute the smooth coefficient $s_{(0)}$ and the detail coefficients $d_{(j),k}$.

Introduced by Mallat (1989), the multi-scale analysis corresponds to the following iterative scheme:



where the high-pass filter defines the details of the data and the low-pass filter defines the smoothing signal. In this example, we obtain these wavelet coefficients:

$$W = \begin{bmatrix} ssss \\ sssd \\ ssd \\ sd \\ d \end{bmatrix}$$

Applying this pyramidal algorithm to the time series signal up to the $J$ resolution level gives us the wavelet coefficients:

$$W = \begin{bmatrix} s_{(0)} \\ d_{(0)} \\ d_{(1)} \\ . \\ . \\ . \\ d_{(J-1)} \end{bmatrix}$$

## A.4 Support vector machine

The support vector machine is an important part of statistical learning theory (Hastie *et al.*, 2009). It was first introduced by Boser *et al.* (1992) and has been used in various domains such as pattern recognition, biometrics, etc. This technique can be employed in different contexts such as classification, regression or density estimation (see Vapnik, 1998). Recently, applications in finance have been developed in two main directions. The first employs the SVM as a nonlinear estimator in order to forecast the trend or volatility of financial assets. In this context, the SVM is used as a regression technique with the possibility for extension to nonlinear cases thank to the kernel approach. The second direction consists of using the SVM as a classification technique which aims to define the stock selection in trading strategies.

### A.4.1 SVM in a nutshell

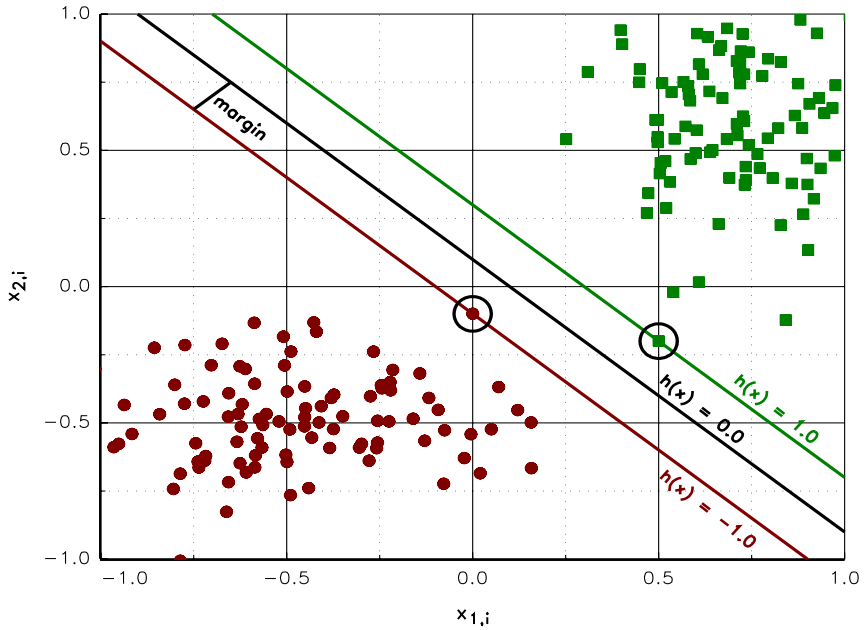
We illustrate here the basic idea of the SVM as a classification method. Let us define the training data set consisting of $n$ pairs of "input/output" points $(x_i, y_i)$ where $x_i \in \mathcal{X}$ and $y_i \in \{-1, 1\}$. The idea of linear classification is to look for a possible hyperplane that can separate $\{x_i \subset \mathcal{X}\}$ into two classes corresponding to the labels $y_i = \pm 1$. It consists of constructing a linear discriminant function $h(x) = w^\top x + b$ where $w$ is the vector of weights and $b$ is called the bias. The hyperplane is then defined by the following equation:

$$H = \{x : h(x) = w^\top x + b = 0\}$$

The vector $w$ is interpreted as the normal vector to the hyperplane. We denote its norm $\|w\|$ and its direction $\hat{w} = w/\|w\|$. In Figure 25, we give a geometric interpretation of the margin in the linear case. Let $x_+$ and $x_-$ be the closest points to the hyperplane from the positive side and negative side. These points determine the margin to the boundary from which the two classes of points $\mathcal{D}$ are separated:

$$m_{\mathcal{D}}(h) = \frac{1}{2}\hat{w}^\top (x_+ - x_-) = \frac{1}{\|w\|}$$

Figure 25: Geometric interpretation of the margin in a linear SVM




The main idea of a maximum margin classifier is to determine the hyperplane that maximises the margin. For a separable dataset, the margin SVM is defined by the following optimisation problem:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{u.c.} \quad & y_i\left(w^\top x_i + b\right) > 1 \quad \text{for } i = 1, \ldots, n \end{aligned}$$

The historical approach to solving this quadratic problem with nonlinear constraints is to map the primal problem to the dual problem:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j \\ \text{u.c.} \quad & \alpha_i \geq 0 \quad \text{for } i = 1, \ldots, n \end{aligned}$$

Because of the Kuhn-Tucker conditions, the optimised solution $(w^\star, b^\star)$ of the primal problem is given by $w^\star = \sum_{i=1}^{n} \alpha_i^\star y_i x_i$ where $\alpha^\star = (\alpha_1^\star, \ldots, \alpha_n^\star)$ is the solution of the dual problem.

We notice that linear SVM depends on input data via the inner product. An intelligent way to extend SVM formalism to the nonlinear case is then to replace the inner product with a nonlinear kernel. Hence, the nonlinear SVM dual problem can be obtained by systematically replacing the inner product $x_i^\top x_j$ by a general kernel $K(x_i, x_j)$. Some standard kernels are widely used in pattern recognition, for example polynomial, radial basis or neural

network kernels[36]. Finally, the decision/prediction function is then given by:

$$f(x) = \operatorname{sgn} h(x) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b\right)$$

### A.4.2 SVM regression

In the last discussion, we presented the basic idea of the SVM in the classification context. We now show how the regression problem can be interpreted as a SVM problem. In the general framework of statistical learning, the SVM problem consists of minimising the risk function $\mathcal{R}(f)$ depending on the form of the prediction function $f(x)$. The risk function is calculated via the loss function $\mathbf{L}(f(x), y)$ which clearly defines our objective (classification or regression):

$$\mathcal{R}(f) = \int \mathbf{L}(f(x), y)\, \mathrm{d}P(x, y)$$

where the distribution $P(x, y)$ can be computed by empirical distribution[37] or an approximated distribution[38]. For the regression problem, the loss function is simply defined as $\mathbf{L}(f(x), y) = (f(x) - y)^2$ or $\mathbf{L}(f(x), y) = |f(x) - y|^p$ in the case of $L_p$ norm.

We have seen that the linear SVM is a special case of nonlinear SVM within the kernel approach. We therefore consider the nonlinear case directly where the approximate function of the regression has the following form $f(x) = w^\top \phi(x) + b$. In the VRM framework, we assume that $P(x, y)$ is a Gaussian noise with variance $\sigma^2$:

$$\mathcal{R}(f) = \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - y_i|^p + \sigma^2 \|w\|^2$$

We introduce the variable $\xi = (\xi_1, \ldots, \xi_n)$ which satisfies $y_i = f(x_i) + \xi_i$. The optimisation problem of the risk function can now be written as a QP programme with nonlinear constraints:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + \left(2n\sigma^2\right)^{-1} \sum_{i=1}^{n} |\xi_i|^p \\ \text{u.c.} \quad & y_i = w^\top \phi(x_i) + b + \xi_i \quad \text{for } i = 1, \ldots, n \end{aligned}$$

In the present form, the regression looks very similar to the SVM classification problem and can be solved in the same way by mapping to the dual problem. We notice that the SVM regression can be easily generalised in two possible ways:

1. by introducing a more general loss function such as the $\varepsilon$-SV regression proposed by Vapnik (1998);

2. by using a weighting distribution $\omega$ for the empirical distribution:

$$\mathrm{d}P(x, y) = \sum_{i=1}^{n} \omega_i \delta_{x_i}(x)\, \delta_{y_i}(y)$$

[36] We have, respectively, $K(x_i, x_j) = \left(x_i^\top x_j + 1\right)^p$, $K(x_i, x_j) = \exp\left(-\|x_i - x_j\|^2 / \left(2\sigma^2\right)\right)$ or $K(x_i, x_j) = \tanh\left(a x_i^\top x_j - b\right)$.

[37] This framework called ERM was first introduced by Vapnik and Chervonenskis (1991).

[38] This framework is called VRM (Chapelle, 2002).

As financial series have short memory and depend more on the recent past, an asymmetric weight distribution focusing on recent data would improve the prediction[39].

The dual problem in the case $p = 1$ is given by:

$$\begin{array}{ll} \max\limits_{\alpha} & \alpha^\top y - \dfrac{1}{2}\alpha^\top K \alpha \\ \text{u.c.} & \begin{cases} \alpha^\top \mathbf{1} = 0 \\ |\alpha| \leq \left(2n\sigma^2\right)^{-1} \mathbf{1} \end{cases} \end{array}$$

As previously, the optimal vector $\alpha^\star$ is obtained by solving the QP programme. We then deduce that $w^\star = \sum_{i=1}^n \alpha_i^\star \phi(x_i)$ and $b^\star$ is computed using the Kuhn-Tucker condition:

$$w^\top \phi(x_i) + b - y_i = 0$$

for support vectors $(x_i, y_i)$. In order to achieve a good level of accuracy for the estimation of $b$, we average out the set of support vectors and obtain $b^\star$. The SVM regressor is then given by the following formula:

$$f(x) = \sum_{i=1}^n \alpha_i^\star K(x, x_i) + b^\star$$

with $K(x, x_i) = \phi(x)\,\phi(x_i)$.

In Figure 26, we apply SVM regression with the Gaussian kernel to the S&P 500 index. The kernel parameter $\sigma$ characterises the estimation horizon which is equivalent to period $n$ in the moving average regression.

## A.5 Singular spectrum analysis

In recent years the singular spectrum analysis (SSA) technique has been developed as a time-frequency domain method[40]. It consists of decomposing a time series into a trend, oscillatory components and a noise.

The method is based on the principal component analysis of the auto-covariance matrix of the time series $y = (y_1, \ldots, y_t)$. Let $n$ be the window length such that $n = t - m + 1$ with $m < t/2$. We define the $n \times m$ Hankel matrix $\mathcal{H}$ as the matrix of the $m$ concatenated lag vector of $y$:

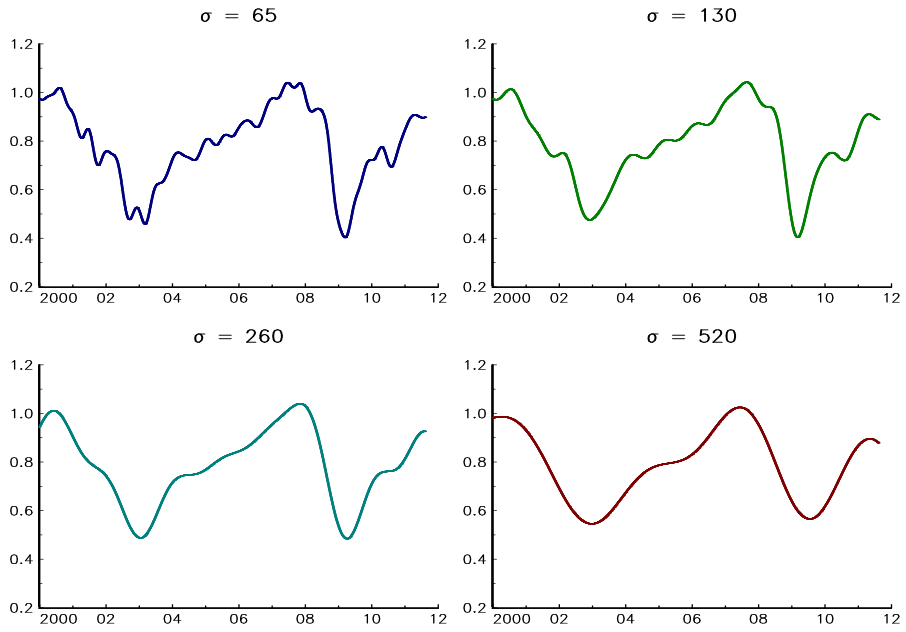
$$\mathcal{H} = \begin{pmatrix} y_1 & y_2 & y_3 & \cdots & y_m \\ y_2 & y_3 & y_4 & \cdots & y_{m+1} \\ y_3 & y_4 & y_5 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & y_{t-1} \\ y_n & y_{n+1} & y_{n+2} & \cdots & y_t \end{pmatrix}$$

We recover the time series $y$ by diagonal averaging:

$$y_p = \frac{1}{\alpha_p} \sum_{j=1}^m \mathcal{H}^{(i,j)} \tag{10}$$

[39] See Gestel *et al.* (2001) and Tay and Cao 2002.
[40] Introduced by Broomhead and King (1986).

Figure 26: SVM filtering




where $i = p - j + 1$, $0 < i < n + 1$ and:

$$\alpha_p = \begin{cases} p & \text{if } p < m \\ t - p + 1 & \text{if } p > t - m + 1 \\ m & \text{otherwise} \end{cases}$$

This relationship seems trivial because each $\mathcal{H}^{(i,j)}$ is equal to $y_p$ with respect to the conditions for $i$ and $j$. But this equality no longer holds if we apply factor analysis. Let $\mathcal{C} = \mathcal{H}^\top \mathcal{H}$ be the covariance matrix of $\mathcal{H}$. By performing the eigenvalue decomposition $\mathcal{C} = V \Lambda V^\top$, we can deduce the corresponding principal components:

$$\mathcal{P}_k = \mathcal{H} V_k$$

where $V_k$ is the matrix of the first $k^{\text{th}}$ eigenvectors of $\mathcal{C}$.

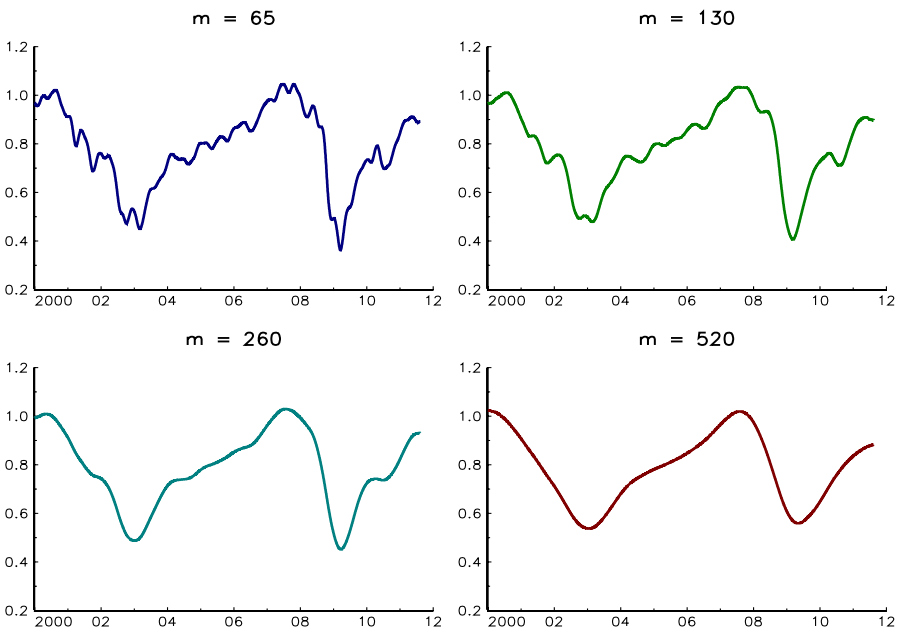
Let us now define the $n \times m$ matrix $\hat{\mathcal{H}}$ as follows:

$$\hat{\mathcal{H}} = \mathcal{P}_k V_k^\top$$

We have $\hat{\mathcal{H}} = \mathcal{H}$ if all the components are selected. If $k < m$, we have removed the noise and the trend $\hat{x}$ is estimated by applying the diagonal averaging procedure (10) to the matrix $\hat{\mathcal{H}}$.

We have applied the singular spectrum decomposition to the S&P 500 index with different lags $m$. For each lag, we compute the Hankel matrix $\mathcal{H}$, then deduce the matrix $\hat{\mathcal{H}}$ using only the first eigenvector ($k = 1$) and estimate the corresponding trend. Results are given in Figure 27. As for other methods, such as nonlinear filters, the calibration depends on the parameter $m$, which controls the window length.

Figure 27: SSA filtering



# References


[1] Alexandrov T., Bianconcini S., Dagum E.B., Maass P. and McElroy T. (2008), A Review of Some Modern Approaches to the Problem of Trend Extraction , *US Census Bureau*, RRS #2008/03.

[2] Antoniadis A., Gregoire G. and McKeague I.W. (1994), Wavelet Methods for Curve Estimation, *Journal of the American Statistical Association*, 89(428), pp. 1340-1353.

[3] Barberis N. and Thaler T. (2002), A Survey of Behavioral Finance, *NBER Working Paper*, 9222.

[4] Beveridge S. and Nelson C.R. (1981), A New Approach to the Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to Measurement of the Business Cycle, *Journal of Monetary Economics*, 7(2), pp. 151-174.

[5] Boser B.E., Guyon I.M. and Vapnik V. (1992), A Training Algorithm for Optimal Margin Classifier, *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pp. 114-152.

[6] Boyd S. and Vandenberghe L. (2009), *Convex Optimization*, Cambridge University Press.

[7] Brockwell P.J. and Davis R.A. (2003), *Introduction to Time Series and Forecasting*, Springer.

[8] Broomhead D.S. and King G.P. (1986), On the Qualitative Analysis of Experimental Dynamical Systems, in Sarkar S. (ed.), *Nonlinear Phenomena and Chaos*, Adam Hilger, pp. 113-144.

[9] Brown S.J., Goetzmann W.N. and Kumar A. (1998), The Dow Theory: William Peter Hamilton's Track Record Reconsidered, *Journal of Finance*, 53(4), pp. 1311-1333.

[10] Burch N., Fishback P.E. and Gordon R. (2005), The Least-Squares Property of the Lanczos Derivative, *Mathematics Magazine*, 78(5), pp. 368-378.

[11] Carhart M.M. (1997), On Persistence in Mutual Fund Performance, *Journal of Finance*, 52(1), pp. 57-82.

[12] Chan L.K.C., Jegadeesh N. and Lakonishok J. (1996), Momentum Strategies, *Journal of Finance*, 51(5), pp. 1681-1713.

[13] Chang Y., Miller J.I. and Park J.Y. (2009), Extracting a Common Stochastic Trend: Theory with Some Applications, *Journal of Econometrics*, 150(2), pp. 231-247.

[14] Chapelle O. (2002), *Support Vector Machine: Induction Principles, Adaptive Tuning and Prior Knowledge*, PhD thesis, University of Paris 6.

[15] Cleveland W.P. and Tiao G.C. (1976), Decomposition of Seasonal Time Series: A Model for the Census X-11 Program, *Journal of the American Statistical Association*, 71(355), pp. 581-587.

[16] Cleveland W.S. (1979), Robust Locally Regression and Smoothing Scatterplots, *Journal of the American Statistical Association*, 74(368), pp. 829-836.

[17] Cleveland W.S. and Devlin S.J. (1988), Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting, *Journal of the American Statistical Association*, 83(403), pp. 596-610.

[18] Cochrane J. (2001), *Asset Pricing*, Princeton University Press.

[19] Cortes C. and Vapnik V. (1995), Support-Vector Networks, *Machine Learning*, 20(3), pp. 273-297.

[20] D'Aspremont A. (2011), Identifying Small Mean Reverting Portfolios, *Quantitative Finance*, 11(3), pp. 351-364.

[21] Daubechies I. (1992), *Ten Lectures on Wavelets*, SIAM.

[22] Daubechies I., Defrise M. and De Mol C. (2004), An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint, *Communications on Pure and Applied Mathematics*, 57(11), pp. 1413-1457.

[23] Donoho D.L. (1995), De-Noising by Soft-Thresholding, *IEEE Transactions on Information Theory*, 41(3), pp. 613-627.

[24] Donoho D.L. and Johnstone I.M. (1994), Ideal Spatial Adaptation via Wavelet Shrinkage, *Biometrika*, 81(3), pp. 425-455.

[25] Donoho D.L. and Johnstone I.M. (1995), Adapting to Unknown Smoothness via Wavelet Shrinkage, *Journal of the American Statistical Association*, 90(432), pp. 1200-1224.

[26] Doucet A., De Freitas N. and Gordon N. (2001), *Sequential Monte Carlo in Practice*, Springer.

[27] Ehlers J.F. (2001), *Rocket Science for Traders: Digital Signal Processing Applications*, John Wiley & Sons.

[28] Elton E.J. and Gruber M.J. (1972), Earnings Estimates and the Accuracy of Expectational Data, *Management Science*, 18(8), pp. 409-424.

[29] Engle R.F. and Granger C.W.J. (1987), Co-Integration and Error Correction: Representation, Estimation, and Testing, *Econometrica*, 55(2), pp. 251-276.

[30] Fama E. (1970), Efficient Capital Markets: A Review of Theory and Empirical Work, *Journal of Finance*, 25(2), pp. 383-417.

[31] Flandrin P., Rilling G. and Goncalves P. (2004), Empirical Mode Decomposition as a Filter Bank, *Signal Processing Letters*, 11(2), pp. 112-114.

[32] Fliess M. and Join C. (2009), A Mathematical Proof of the Existence of Trends in Financial Time Series, in El Jai A., Afifi L. and Zerrik E. (eds), *Systems Theory: Modeling, Analysis and Control*, Presses Universitaires de Perpignan, pp. 43-62.

[33] Fuentes M. (2002), Spectral Methods for Nonstationary Spatial Processes, *Biometrika*, 89(1), pp. 197-210.

[34] Gençay R., Selçuk F. and Whitcher B. (2002), *An Introduction to Wavelets and Other Filtering Methods in Finance and Economics*, Academic Press.

[35] Gestel T.V., Suykens J.A.K., Baestaens D., Lambrechts A., Lanckriet G., Vandaele B., De Moor B. and Vandewalle J. (2001), Financial Time Series Prediction Using Least Squares Support Vector Machines Within the Evidence Framework, *IEEE Transactions on Neural Networks*, 12(4), pp. 809-821.

[36] Golyandina N., Nekrutkin V.V. and Zhigljavsky A.A. (2001), *Analysis of Time Series Structure: SSA and Related Techniques*, Chapman & Hall, CRC.

[37] Gonzalo J. and Granger C.W.J. (1995), Estimation of Common Long-Memory Components in Cointegrated Systems, *Journal of Business & Economic Statistics*, 13(1), pp. 27-35.

[38] Grinblatt M., Titman S. and Wermers R. (1995), Momentum Investment Strategies, Portfolio Performance, and Herding: A Study of Mutual Fund Behavior, *American Economic Review*, 85(5), pp. 1088-1105.

[39] Groetsch C.W. (1998), Lanczo's Generalized Derivative, *American Mathematical Monthly*, 105(4), pp. 320-326.

[40] Grossmann A. and Morlet J. (1984), Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape, *SIAM Journal of Mathematical Analysis*, 15, pp. 723-736.

[41] Härdle W. (1992), *Applied Nonparametric Regression*, Cambridge University Press.

[42] Harvey A.C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press.

[43] Harvey A.C. and Trimbur T.M. (2003), General Model-Based Filters for Extracting Cycles and Trends in Economic Time Series, *Review of Economics and Statistics*, 85(2), pp. 244-255.

[44] Hastie T., Tibshirani R. and Friedman R. (2009), *The Elements of Statistical Learning*, second edition, Springer.

[45] Henderson R. (1916), Note on Graduation by Adjusted Average, *Transactions of the Actuarial Society of America*, 17, pp. 43-48.

[46] Hodrick R.J. and Prescott E.C. (1997), Postwar U.S. Business Cycles: An Empirical Investigation, *Journal of Money, Credit and Banking*, 29(1), pp. 1-16.

[47] Holt C.C. (1959), Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages, *ONR Research Memorandum*, 52, reprinted in *International Journal of Forecasting*, 2004, 20(1), pp. 5-10.

[48] Hong H. and Stein J.C. (1977), A Unified Theory of Underreaction, Momentum Trading and Overreaction in Asset Markets, *NBER Working Paper*, 6324.

[49] Johansen S. (1988), Statistical Analysis of Cointegration Vectors, *Journal of Economic Dynamics and Control*, 12(2-3), pp. 231-254.

[50] Johansen S. (1991), Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models, *Econometrica*, 52(6), pp. 1551-1580.

[51] Kalaba R. and Tesfatsion L. (1989), Time-varying Linear Regression via Flexible Least Squares, *Computers & Mathematics with Applications*, 17, pp. 1215-1245.

[52] Kalman R.E. (1960), A New Approach to Linear Filtering and Prediction Problems, *Transactions of the ASME – Journal of Basic Engineering*, 82(D), pp. 35-45.

[53] Kendall M.G. (1973), *Time Series*, Charles Griffin.

[54] Kim S-J., Koh K., Boyd S. and Gorinevsky D. (2009), $\ell_1$ Trend Filtering, *SIAM Review*, 51(2), pp. 339-360.

[55] Kolmogorov A.N. (1941), Interpolation and Extrapolation of Random Sequences, *Izvestiya Akademii Nauk SSSR, Seriya Matematicheskaya*, 5(1), pp. 3-14.

[56] Macaulay F. (1931), *The Smoothing of Time Series*, National Bureau of Economic Research.

[57] Mallat S.G. (1989), A Theory for Multiresolution Signal Decomposition: The Wavelet Representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7), pp. 674-693.

[58] Mann H.B. (1945), Nonparametric Tests against Trend, *Econometrica*, 13(3), pp. 245-259.

[59] Martin W. and Flandrin P. (1985), Wigner-Ville Spectral Analysis of Nonstationary Processes, *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(6), pp. 1461-1470.

[60] Muth J.F. (1960), Optimal Properties of Exponentially Weighted Forecasts, *Journal of the American Statistical Association*, 55(290), pp. 299-306.

[61] Oppenheim A.V. and Schafer R.W. (2009), *Discrete-Time Signal Processing*, third edition, Prentice-Hall.

[62] Peña D. and Box, G.E.P. (1987), Identifying a Simplifying Structure in Time Series, *Journal of the American Statistical Association*, 82(399), pp. 836-843.

[63] Pollock, D.S.G. (2006), Wiener-Kolmogorov Filtering Frequency-Selective Filtering and Polynomial Regression, *Econometric Theory*, 23, pp. 71-83.

[64] Pollock D.S.G. (2009), Statistical Signal Extraction: A Partial Survey, in Kontoghiorges E. and Belsley D.E. (eds.), *Handbook of Empirical Econometrics*, John Wiley and Sons.

[65] Rao S.T. and Zurbenko I.G. (1994), Detecting and Tracking Changes in Ozone air Quality, *Journal of Air and Waste Management Association*, 44(9), pp. 1089-1092.

[66] Roncalli T. (2010), *La Gestion d'Actifs Quantitative*, Economica.

[67] Savitzky A. and Golay M.J.E. (1964), Smoothing and Differentiation of Data by Simplified Least Squares Procedures, *Analytical Chemistry*, 36(8), pp. 1627-1639.

[68] Silverman B.W. (1985), Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting, *Journal of the Royal Statistical Society*, B47(1), pp. 1-52.

[69] Sorenson H.W. (1970), Least-Squares Estimation: From Gauss to Kalman, *IEEE Spectrum*, 7, pp. 63-68.

[70] Stock J.H. and Watson M.W. (1988), Variable Trends in Economic Time Series, *Journal of Economic Perspectives*, 2(3), pp. 147-174.

[71] Tay F.E.H. and Cao L.J. (2002), Modified Support Vector Machines in Financial Times Series Forecasting, *Neurocomputing*, 48(1-4), pp. 847-861.

[72] Tibshirani R. (1996), Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society*, B58(1), pp. 267-288.

[73] Vapnik V. (1998), *Statistical Learning Theory*, John Wiley and Sons, New York.

[74] Vapnik V. and Chervonenskis A. (1991), On the Uniform Convergence of Relative Frequency of Events to their Probabilities, *Theory of Probability and its Applications*, 16(2), pp. 264-280.

[75] Vautard R., Yiou P., and Ghil M. (1992), Singular Spectrum Analysis: A Toolkit for Short, Noisy Chaotic Signals, *Physica D*, 58(1-4), pp. 95-126.

[76] Wahba G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, 59, SIAM.

[77] Wang Y. (1998), Change Curve Estimation via Wavelets, *Journal of the American Statistical Association*, 93(441), pp. 163-172.

[78] Wiener N. (1949), *Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*, MIT Technology Press and John Wiley & Sons (originally published in 1941 as a Report on the Services Research Project, DIC-6037).

[79] Whittaker E.T. (1923), On a New Method of Graduation, *Proceedings of the Edinburgh Mathematical Society*, 41, pp. 63-75.

[80] Winters P.R. (1960), Forecasting Sales by Exponentially Weighted Moving Averages, *Management Science*, 6(3), 324-342.

[81] Yue S. and Pilon P. (2004), A Comparison of the Power of the t-test, Mann-Kendall and Bootstrap Tests for Trend Detection, *Hydrological Sciences Journal*, 49(1), 21-37.

[82] Zurbenko I., Porter P.S., Rao S.T., Ku J.K., Gui R. and Eskridge R.E. (1996), Detecting Discontinuities in Time Series of Upper-Air Data: Demonstration of an Adaptive Filter Technique, *Journal of Climate*, 9(12), pp. 3548-3560.