Dynamic nonparametric filtering with application to volatility estimation

Ming-Yen Cheng[a] * and Jianqing Fan[b] † and Vladimir Spokoiny[c]

[a]Department of Mathematics, National Taiwan University
Taipei 106, Taiwan

[b]Department of Operation Research and Financial Engineering, Princeton University
Princeton, NJ 08544

[c]Weierstrass-Institute
Mohrenstr. 39, 10117 Berlin, Germany

Problems of nonparametric filtering arises frequently in engineering and financial economics. Nonparametric filters often involve some filtering parameters to choose. These parameters can be chosen to optimize the performance locally at each time point or globally over a time interval. In this article, the filtering parameters are chosen via minimizing the prediction error for a large class of filters. Under a general martingale setting, with mild conditions on the time series structure and virtually no assumption on filters, we show that the adaptive filter with filtering parameter chosen by historical data performs nearly as well as the one with the ideal filter in the class, in terms of filtering errors. The theoretical result is also verified via intensive simulations. Our approach is also useful for choosing the orders of parametric models such as AR or GARCH processes. It can also be applied to volatility estimation in financial economics. We illustrate the proposed methods by estimating the volatility of the returns of the S&P500 index and the yields of the three-month Treasury bills.

1. Introduction

Problems of nonparametric filtering arises frequently in engineering, financial economics, and many other scientific disciplines. Given a time series $\{Y_t\}$, the nonparametric filtering problem is to dynamically predict Y_t based on the observations preceding t. This is a specific problem of the time domain smoothing (see §6.2 of [10]), but allows to use only historical data at each time. A traditional class of nonparametric filters is the moving average filtering which is the average of last m time periods (see Example 1 below). Other classes include exponential smoothing (Example 2 below), kernel smoothing (Example 2), autoregressive filtering (Example 5) and ARCH and GARCH filtering (see §4.2). All of these filters depend on certain parameters, called filtering parameters

*Partially supported by NSC grant 91-2118-M-002-001
†Partially supported by NSF grant DMS-0204329 and a direct allocation RGC grant of the Chinese University of Hong Kong.

in this paper. An interesting and challenging issue is how to choose these parameters so that they are adaptive automatically to the data.

There are basically two versions of filtering parameters, local and global versions. The local version is that at each time point t, we choose the filtering parameters $\widehat{\lambda}_t$, say, to optimize the performance near t. The global version is to set an in-sample period, $[1, T]$, say, then to choose filtering parameters $\widehat{\lambda}$ to optimize the performance in the time interval $[1, T]$, and finally to predict the data in the out-sample period $[T+1, T+n]$, say, using the filtering parameters $\widehat{\lambda}$. The local choice of ideal filtering parameters is more powerful than the global one. However, owing to stochastic errors, data-driven choices of local filtering parameters, while are more flexible, do not necessarily outperform the global choice. However, they are very useful in the situation where there are structural changes of an underlying series over time. The situation is very similar to the local bandwidth and global bandwidth selection in the nonparametric smoothing literature (see e.g. [5] and [7]).

A natural criterion for choosing filtering parameters is the prediction error, since only the historical data have been used in the construction of filters. Because of this, semi-martingale structures remain valid in the computation of the prediction error. This enables us to show, with mild conditions on the time series structure and virtually no assumption on filters, that the resulting adaptive filter performs nearly as well as the ideal choice of filtering parameters. This property is also verified via intensive numerical computation. The nice property encourages us to apply the techniques to volatility estimation in financial econometrics.

The concept of volatility is associated with the notation of risks. It is very critical for portfolio optimization, option pricing and management of financial risks. As shown in Section 4.1, the problem of dynamic prediction of volatility is strongly associated with a filtering problem. In fact, a family of power transform can even be accommodated to estimate volatility (see [15]), with our filtering techniques. This yields a family of volatility estimators: some aim at robustness, while others at efficiency. The family of nonparametric methods compares favorably with GARCH techniques in volatility estimation, from our numerical experiments.

The paper is organized as follows. Section 2 outlines various filtering techniques. Their filtering parameters are selected in Section 3 where the properties of the adaptive filters are investigated both theoretically and empirically. Problems of volatility estimation and their associations with nonparametric filtering are investigated in Section 4.

2. Problems of dynamic filtering

Consider a time series $Y_1, \ldots, Y_T$, which is progressively measurable with respect to a filtration $\mathbb{F} = (\mathcal{F}_t)$ and allows a semi-martingale representation:

$$Y_t = f_t + v_t \varepsilon_t, \tag{1}$$

where f_t and v_t are predictable and the innovations ε_t form a standardized martingale difference. They satisfy

$$\boldsymbol{E}[\varepsilon_t \mid \mathcal{F}_{t-1}] = 0, \qquad \boldsymbol{E}[\varepsilon_t^2 \mid \mathcal{F}_{t-1}] = 1.$$

The function f_t is a trend or a drift series and v_t is the conditional standard deviation or the diffusion of the series. In statistical forecasting, one wishes to estimate the conditional mean f_t based on the past data $Y_1, \ldots, Y_t$. Such a problem is also referred to as filtering or one-step forecasting of the process $\{Y_t\}$.

Depending on the background of applications, as detailed below, filters or predictors depend on some tuning parameters. Suppose that we are given a family of different filters (predictors) $\widehat{f}_{t,\lambda}$ indexed by some parameter λ. Each predictor $\widehat{f}_{t,\lambda}$ estimates the unknown value f_t from the 'past' observations $Y_1, \ldots, Y_{t-1}$. Our goal is to construct one predictor which does the job nearly as well as the best filter in the family $\{\widehat{f}_{t,\lambda}, \lambda \in \Lambda\}$. We use a few examples to illustrate the versatility of the scope of our study.

Example 1 (*Moving average filtering*) A traditional approach to estimate the trend of a time series is the moving average estimator. For every integer m, one defines

$$\widehat{f}_{t,m} = \frac{1}{m} \sum_{s=t-m}^{t-1} Y_s .$$

Here the parameter λ coincides with the window size m. One may consider a family of such predictors for different window sizes and the problem of adaptive estimation is to choose a window size from data.

Example 2 (*Exponential smoothing*) An improvement of the moving average (MA) filtering is the exponential smoothing (ES) which weighs down the observed data from the past. The family of exponential smoothing is defined by

$$\widehat{f}_{t,\lambda} = \frac{1}{1-e^{-\lambda}} \sum_{s<t} e^{-(t-s)\lambda} Y_s, \tag{2}$$

for a positive parameter λ. Formally this estimate $\widehat{f}_{t,\lambda}$ depends on all the past observations, but this dependence decreases exponentially. Our problem becomes to choose the parameter λ from data such that the resulting adaptive estimator performs nearly as well as the ideal exponential filtering among this family.

The MA and ES filtering are a member of the kernel estimator. See, for example, [8] and §6.2 of [10]. In fact, the ES corresponds to the kernel regression estimator in time domain

$$\widehat{f}_{t,h} = \sum_{s<t} K_h(t-s) Y_s \Big/ \sum_{s<t} K_h(t-s), \quad K_h(x) = h^{-1} K(x/h)$$

with the one-sided kernel $K(x) = \exp(-x) I(x > 0)$ and $\lambda = 1/h$. Further discussion on this subject, including bandwidth selection and asymptotic theory, can be found in [11].

Example 3 (*J.P. Morgan's RiskMetrics*) An important measure to gauge the risk of a portfolio is the Value-at-Risk, which is the worst loss to be expected with certain confidence for a given time horizon. See [13]. An important contribution to the calculation of VaR is the RiskMetrics of J.P. Morgan[16]. The method is to first estimate the volatility for holding a portfolio for one day before converting this into the volatility for multiple days and to then compute the quantile of standardized return processes through the assumption that the processes follow a standard normal distribution. Let S_t be the price of a portfolio at time t and $R_t = \log(S_t/S_{t-1})$ be the observed return at time t. The J.P. Morgan estimate of volatility $\widehat{\sigma}_t^2$ for one-period return is

$$\widehat{\sigma}_t^2 = (1-\lambda_0)R_{t-1}^2 + \lambda_0 \widehat{\sigma}_{t-1}^2.$$

By iterating the above formula, it can easily be seen that

$$\widehat{\sigma}_t^2 = (1-\lambda_0)\{R_{t-1}^2 + \lambda_0 R_{t-2}^2 + \lambda_0^2 R_{t-3}^2 + \cdots\}.$$

This is an alternative form of the ES (2) with $\lambda_0 = \exp(-\lambda)$. Our adaptive dynamic filtering is to choose λ_0 from data to ameliorate the performance. Such an approach has been introduced by [8]. Our current study gives further theoretical endorsement of their approach.

The above estimator is basically a discritized approach to estimate the diffusion function $\theta(t)$ in the following geometric Brownian motion $d\log(S_u) = \theta(u)dW_u$ via a local constant approximation. See [8] for derivations and connections.

Example 4 (*Adaptive estimation of volatility*) The problem of estimating v_t in (1) can also be regarded as adaptive filtering problem. Let $\widehat{R}_t = Y_t - \widehat{f}_t$ be the residual from model fitting (1). Then, define a family of the filters for square residuals as

$$\widehat{v}_{t,h}^2 = \sum_{s<t} K_h(t-s)\widehat{R}_s^2 / \sum_{s<t} K_h(t-s).$$

As shown in [9] and [17], the errors in estimating $\widehat{f}_t$ are usually negligible in estimating $\widehat{v}_{t,h}^2$. Hence, our methodology and theory continue to apply.

Example 5 (*Autoregression*) Suppose that the process $\{Y_t\}$ is to be approximated by an autoregressive (AR) equation

$$Y_t = \alpha_1 Y_{t-1} + \ldots + \alpha_p Y_{t-p} + v_t \varepsilon_t$$

where ε_t are conditionally independent innovations, v_t is an unknown predictable process, and $\alpha_1, \ldots, \alpha_p$ are unknown autoregression coefficients. Denote by

$$X_{t,p} = (Y_{t-1}, \ldots, Y_{t-p})^\top \quad \text{and} \quad \alpha = (\alpha_1, \ldots, \alpha_p)^\top.$$

Then, the above autoregressive equation can be written in the form $Y_t = X_{t,p}^\top \alpha + v_t \varepsilon_t$. The least-squares estimate of the parameter α from the observations Y_s for $t_0 \le s < t$ reads as follows:

$$\widehat{\alpha}_{t,p} = \left(\sum_{s=t_0}^{t-1} X_{s,p} X_{s,p}^\top \right)^{-1} \sum_{t=t_0}^{t-1} X_{s,p} Y_s$$

and the corresponding filter $\widehat{f}_{t,p}$ of Y_t is defined as $\widehat{f}_{t,p} = X_{t,p}^\top \widehat{\alpha}_{t,p}$. The problem of adaptive dynamic filtering is to choose a p using the available data before time t such that it performs nearly as well as the ideal choice of p.

Similar idea can be applied to choose the order of GARCH models (see §4.2 of [10] and [12])

3. Choice of filtering parameters

There are two possible choices of filtering parameters. A local choice aims at choosing a λ, which depends on t, such that it optimizes the performance of the filter at each time point t. In other words, we use $\widehat{f}_{t,\widehat{\lambda}_t}$ to estimate f_t, where $\widehat{\lambda}_t$ is selected based the data collected up to $t-1$. A global choice aims at choosing a λ, which is independent of t, such that it optimizes the performance of the filter over an interval, $[t_0, T]$. The local choice of the filtering parameters is more flexible than the global one and the resulting filter is more capable of adapting to the dynamic change of the underlying time series. On the other hand, the local choice of filtering parameters is harder and more variable, since only the local data are involved in choosing the filtering parameters. The problem is very analogous to the local and global bandwidths in nonparametric smoothing, studied, for example, by [5] and [7].

3.1. Global adaptation via minimal prediction error

The performance of any filter can be measured by the sum of squared filtering errors:

$$R(\lambda) = \sum_{t=t_0}^{T} \left(f_t - \widehat{f}_{t,\lambda} \right)^2 ,$$

where T is the length of a time series and t_0 is a predefined time point large enough to avoid the boundary effect. An ideal choice of the parameter λ can be defined as the one that minimizes the global loss:

$$\lambda_{I\!I} = \operatorname*{arginf}_{\lambda \in \Lambda} \sum_{t=t_0}^{T} \left(f_t - \widehat{f}_{t,\lambda} \right)^2 . \tag{3}$$

This choice is ideal since it relies on the unknown target function.

An empirical analog of the filtering error is the prediction error defined as:

$$\rho(\lambda) = \sum_{t=t_0}^{T} \left(Y_t - \widehat{f}_{t,\lambda} \right)^2 .$$

This criterion leads to the following data-driven selection rule:

$$\widehat{\lambda} = \operatorname*{arginf}_{\lambda \in \Lambda} \rho(\lambda) = \operatorname*{arginf}_{\lambda \in \Lambda} \sum_{t=t_0}^{T} \left(Y_t - \widehat{f}_{t,\lambda} \right)^2 . \tag{4}$$

The resulting adaptive filter is given by $\widehat{f}_t = \widehat{f}_{t,\widehat{\lambda}}$. The filtering error of this estimator is given by

$$R(\widehat{\lambda}) = \sum_{t=t_0}^{T} \left(f_t - \widehat{f}_t \right)^2 = \sum_{t=t_0}^{T} \left(f_t - \widehat{f}_{t,\widehat{\lambda}} \right)^2 .$$

An interesting question is how the the quality (filtering error) of the data-driven selector from (4) is compared with the "ideal" selector from (3). We attempt to answer this question in the next section.

3.2. Properties of the adaptive selector

For every $\lambda \in \Lambda$, it holds that

$$\rho(\lambda) = \sum_{t=t_0}^{T} \left(Y_t - \widehat{f}_{t,\lambda}\right)^2 = R(\lambda) + 2\sum_{t=t_0}^{T} \left(f_t - \widehat{f}_{t,\lambda}\right) v_t \varepsilon_t + \sum_{t=t_0}^{T} v_t^2 \varepsilon_t^2 .$$

The last sum in the above decomposition does not depend on λ and hence does not affect the minimization in (4). Hence, minimizing the prediction error $\rho(\lambda)$ corresponds to minimizing the filtering error $R(\lambda)$ plus the cross term

$$S_{cross,\lambda} = 2\sum_{t=t_0}^{T} \left(f_t - \widehat{f}_{t,\lambda}\right) v_t \varepsilon_t .$$

If one could show that this cross term is relatively small, then these two minimization procedures would be nearly equivalent.

To bound the cross term $S_{cross,\lambda}$, we apply a result for martingales from [14]. Define

$$M_{t,\lambda} = \sum_{s=t_0}^{t} \left(f_s - \widehat{f}_{s,\lambda}\right) v_s \varepsilon_s, \quad \text{and} \quad V_{t,\lambda}^2 = \sum_{s=t_0}^{t} \left(f_s - \widehat{f}_{s,\lambda}\right)^2 v_s^2 .$$

Note that for homoscedastic error $v_s \equiv v$, $V_{T,\lambda}^2 = v^2 R^2(\lambda)$. In general, $V_{T,\lambda}$ is of the same order as $R^2(\lambda)$ as long as v_s is bounded from below and above. Since both the drift f_t and the estimator $\widehat{f}_{t,\lambda}$ are predictable processes with respect to the filtration $\mathcal{F}_{t-1}$, $M_{t,\lambda}$ is a square integrable martingale with the quadratic variation $V_{t,\lambda}^2$.

Lemma 1 *Let the innovations ε_t fulfill $\boldsymbol{E} e^{u\varepsilon_t} \le \exp\{u^2/(2a)\}$ for some positive a and all $u \ge 0$. Then, for all $\gamma \ge 1$*

$$\boldsymbol{P}\left(M_{T,\lambda} > \gamma V_{T,\lambda}\,,\, \mathcal{A}_\lambda\right) \le \alpha_\lambda(\gamma)$$

where $\mathcal{A}_\lambda = \left\{\vartheta \le V_{T,\lambda}^2 \le \vartheta B\right\}$ with some deterministic values ϑ, B and $\alpha_\lambda(\gamma) = 4\sqrt{e}(1+\log B)\gamma e^{-\gamma^2/(2a)}$.

Note that the constants ϑ, B may also depend on λ. We suppress this dependence to facilitate the notation. As a corollary of Lemma 1:

$$\sum_{\lambda\in\Lambda} \boldsymbol{P}\left(S_{cross,\lambda} > 2\gamma V_{T,\lambda}\,,\, \mathcal{A}_\lambda\right) \le \sum_{\lambda\in\Lambda} \alpha_\lambda(\gamma). \tag{5}$$

As noted above, $V_{T,\lambda}^2 \le v^2 R(\lambda)$ when $v_t \le v$. It follows that $S_{cross,\lambda} \le 2\gamma v\sqrt{R(\lambda)}, \forall \lambda \in \Lambda$ with a probability at least $1 - \sum_{\alpha\in\Lambda} \alpha(\gamma)$. This yields the following results.

Theorem 1 *It holds for every $\gamma \ge 0$ and every $v > 0$*

$$\boldsymbol{P}\left(\sqrt{R(\widehat{\lambda})} \ge \sqrt{R(\lambda_{I\!I})} + 3v\gamma,\, \mathcal{A}\right) \le \sum_{\lambda\in\Lambda} \alpha_\lambda(\gamma)$$

with $\mathcal{A} = \bigcap_{\lambda\in\Lambda} \mathcal{A}_\lambda \cap \{V_{T,\lambda}^2 \le v^2 R(\lambda)\}$.

Proof. In view of (5) it suffices to show that the inequalities

$$S_{cross,\lambda} \le 2\gamma V_{T,\lambda} \le 2\gamma v\sqrt{R(\lambda)} \qquad \forall \lambda \in \Lambda \tag{6}$$

imply that

$$\sqrt{R(\widehat{\lambda})} \le \sqrt{R(\lambda_{I\!I})} + 3v\gamma/2.$$

To this end, define $\rho^*_\lambda = R(\lambda)+S_{cross,\lambda}$ and denote by $R_\lambda = R(\lambda)$. Then $\widehat{\lambda}$ is the minimizer of ρ^*_λ, while $\lambda_{I\!I}$ is the minimizer of R_λ. The condition $\rho^*_\lambda \le R_\lambda + 2\gamma v\sqrt{R_\lambda}$ implies $\sqrt{\rho^*_\lambda} \le \sqrt{R_\lambda}+\gamma v$. Similarly, (6) implies $\rho^*_\lambda \ge R_\lambda - 2\gamma v\sqrt{R_\lambda}$. Since $\sqrt{x-y} \ge \sqrt{x} - y/\sqrt{x}$ for every positive numbers x, y, it follows that for each $\lambda \in \Lambda$:

$$\sqrt{R_\lambda} - 2\gamma v \le \sqrt{\rho_\lambda} \le \sqrt{R_\lambda} + \gamma v\,.$$

Since $\widehat{\lambda}$ is the minimizer of ρ_λ, $\rho_{\widehat{\lambda}} \le \rho_{\lambda_{I\!I}}$. Therefore,

$$\sqrt{R_{\widehat{\lambda}}} \le \sqrt{\rho_{\widehat{\lambda}}} + 2\gamma v \le \sqrt{\rho_{I\!I}} + 2\gamma v \le \sqrt{R_{I\!I}} + 3\gamma v \tag{7}$$

as required. ∎

Note that $\alpha_\lambda(\gamma) = o(T^{-b})$, when $\gamma > (2ab\log T)^{1/2}$ for a given positive b. Thus, when the number of elements in Λ is of order $O(T^b)$, $\sum_{\lambda\in\Lambda} \alpha_\lambda(\gamma) \to 0$. Thus, the extra term $3v\gamma$ required for adaptation in Theorem 1 is not excessive and is only of order $O\{(\log T)^{1/2}\}$. For example, for a parametric model such as an $AR(p)$ model, the filtering error $|f_t - \widehat{f}_{t,\lambda}|^2$ is typically of order t^{-1} so that $R_{\lambda,T}$ is of order $\sum_{t\le T} t^{-1} \approx \log T$. The extra term $3v\gamma = O\{(\log T)^{1/2}\}$ is negligible. For nonparametric filters like moving average with window size m, the filtering errors are of order $O(T^{-2/5})$ for $m = T^{1/5}$, which are much larger than those of parameter models. Hence, the extra term of order $O\{(\log T)^{1/2}\}$ is also negligible, comparing with $R(\lambda_{I\!I})$. In summary, with probability tending to one, the data-driven filters perform as well as filters with an ideal choice of filtering parameter.

3.3. Local choice

The aforementioned global procedure chooses one filter parameter to fit the whole observed path. Such a method can be efficient in many situations where there are virtually no structural breaks in the observed time series. However, it has a serious drawback of being slow in reacting to spontaneous changes in the structure of the observed process. We illustrate this issue using the moving average filter with window size m. See also Example 6. For a large m, the accuracy of estimation is very good provided that the underlying process f_t is nearly constant within this window. However, if the value f_t changes abruptly at some time point, then the filter with a large m will react to this change with a long delay of order m. On the other hand, a filter with a small m allows for a fast reaction to the sudden changes in structure, but is not as precise and stable as a filter with larger m over stationary regions.

To enhance the flexibility of the family of filters $\{\widehat{f}_{t,\lambda}\}$ to adapt to possible structural changes over time, the parameter λ should be allowed to vary over time. For each given time t, λ_t should be chosen so that it optimizes the performance near the time point t. Following (4), we choose

$$\widehat{\lambda}_t = \operatorname*{arginf}_{\lambda\in\Lambda} \sum_{t=t-M+1}^{t} \left(Y_t - \widehat{f}_{t,\lambda}\right)^2, \tag{8}$$

where M is the size of the neighborhood preceding the time point t, over which we wish to optimize the performance. One can also regard M as another filtering parameter and wishes to choose λ_t and M simultaneously. But, simultaneous choices of M and λ face the challenges of instability and computational cost.

In the local bandwidth selection setting, [7] employed a similar idea. However, the resulting parameters $\{\widehat{\lambda}_t\}$ are smoothed further to enhance the smoothness of the resulting estimate $\widehat{f}_{t,\widehat{\lambda}_t}$. In our time domain smoothing, such a step can be avoided, since the smoothness of $\widehat{f}_{t,\widehat{\lambda}_t}$ in time domain is not a critical visual requirement.

Applying Theorem 1 to the local choice of the filter parameters, we can obtain a similar result on the bound of the filtering errors around the time t. Again, as long as the number of elements in Λ is not excessively large, the performance of the data-driven choice of local filtering parameters is nearly as good as their ideal choice.

3.4. Numerical Results

We illustrate the performance of the global and local choices of filtering parameters via two different classes of underlying processes: piecewise constant processes and autoregressive processes. For the first class, an application of moving average or exponential smoothers is quite reasonable, while the second class is oriented towards autoregressive filtering in Example 5. The effectiveness of each filter can be assessed by the Mean Absolute Filtering Error (MAFE) or the Mean Squared Filtering Error (MSFE):

$$\text{MAFE} = \frac{1}{n}\sum_{t=T+1}^{T+n} |f_t - \widehat{f}_t|, \qquad \text{MSFE} = \frac{1}{n}\sum_{t=T+1}^{T+n} |f_t - \widehat{f}_t|^2,$$

for a post sample of size n. The in-sample period is taken to be $t = 1, \cdots, T$. Since the results are similar by using MAFE and MSFE, we only report the MAFE.

Example 6 Let the process f_t take only two values $\{-1, 1\}$, with transitions between these two states at random stopping times $\tau_1 < \tau_2 < \ldots < \tau_m < \ldots$. These stopping times were generated from a Poisson process with rate $1/\mu$, namely, the intervals $\tau_k - \tau_{k-1}$ were generated from the exponential with mean $\mu = 150$. The observed process is

$$Y_t = f_t + \sigma\varepsilon_t, \quad \varepsilon_t \sim N(0, 1).$$

Figure 1(a) depicts a simulated series of length 500.

To estimate the function f_t, we apply the moving average (MA) and exponential smoothing (ES) methods to estimate the time trend. We first apply the global method to choose the filtering parameters, the window size m in MA and the decay parameter λ in ES. The initial value $t_0 = 101$ is taken. The filtering parameters are chosen to minimize (4) among 15 geometric grids. Figure 1(b) shows the resulting estimates for the realization

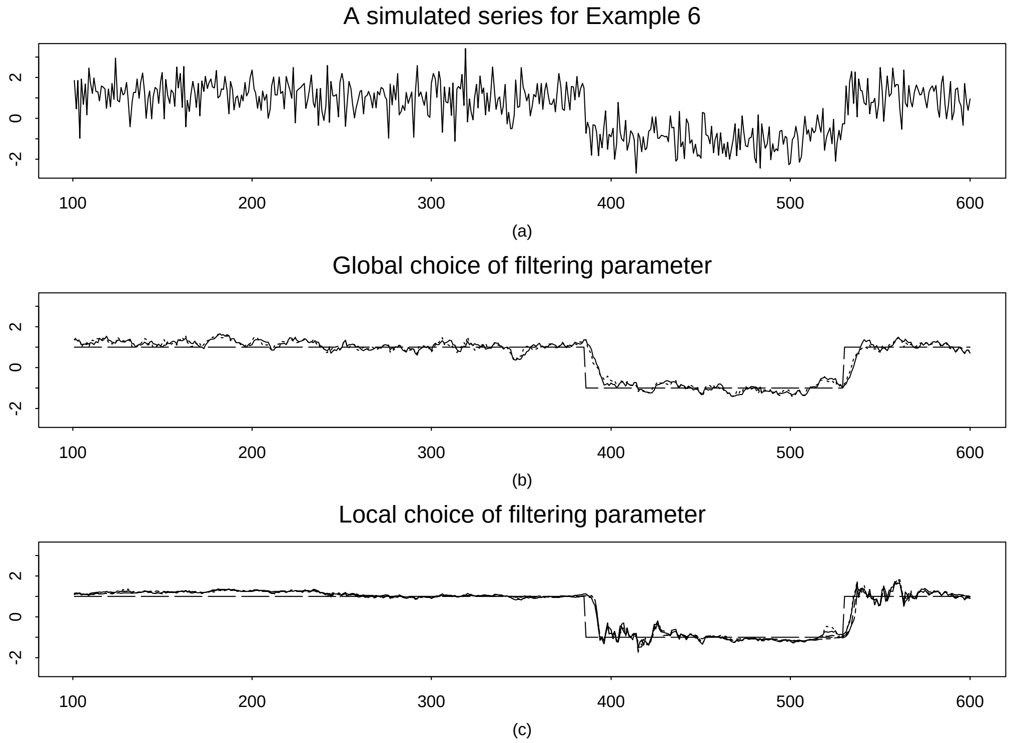

Figure 1. (a) A simulated series from Example 6. (b) Filtered series with global choices of filtering parameters. True f_t — long-dashed curve; MA — solid curve; ES — dotted curve. (c) Filtered series with local choices of filtering parameters. MA — dotted curve; MA ideal — dot-and-dash curve; ES — solid; ES ideal — dashed.

in Figure 1(a). Both the adaptive MA and ES methods recover reasonably well the mean process f_t and detect the jumps in f_t. The jumps in the process f_t force the methods to choose small values of m and λ. As a result, the estimates are somewhat undersmoothed and have rough appearance in the constant regions.

For the signal function f_t in this example, it is reasonable to expect that a large smoothing parameter is used in the first part of the data and a smaller one is applied to the last piece of series. To achieve such a scheme adaptively, we appeal to (8) with $M = 40$. The resulting estimates by using the MA and ES with local choice of filtering parameters are shown in Figure 1(c). To compare with their performance with the ideal choices of the local filtering parameters, which minimize the corresponding local version of (3), Figure 1(c) also depicts the MA and ES estimates using the ideal local filtering parameters. The four estimates are hard to differentiate, which in turn endorses the performance of our adaptive local version of selecting filtering parameters.

Comparing the estimates with the local choices of filtering parameters to those with the global ones, the local version tends to choose larger smoothing parameters for the first part of the series. At the point of the structure break, smaller smoothing parameters are chosen so that "leakages" (biases around the change point) have been reduced by the

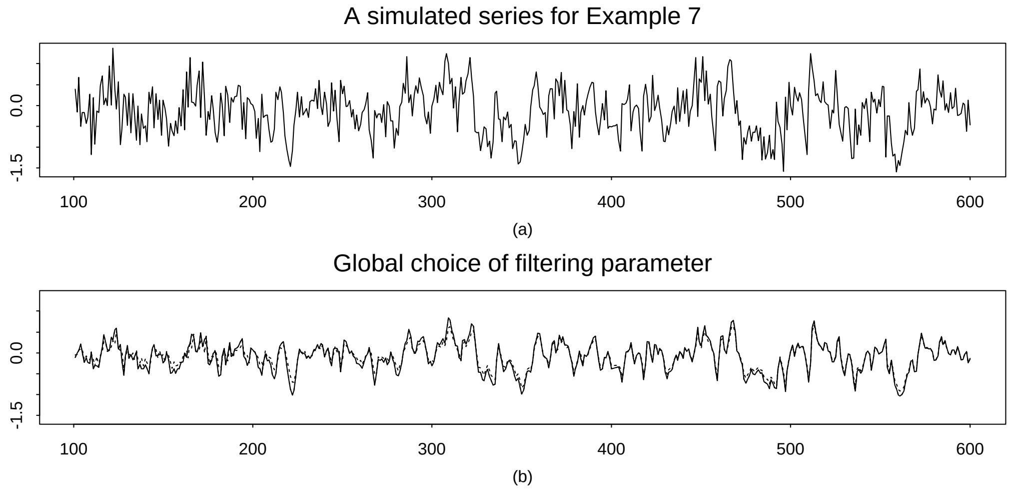

Figure 2. (a) A simulated series from Example 7. (b) Filtered series with global choices of filtering parameters. True f_t — solid curve; filtered series — dotted curve.

local methods, but at the expenses of increasing variability.

The simulation results in terms of MAFE are reported in Table 1. The post-sample size is $n = 500$ and the in-sample period is [1,1000].

Next, we consider an application of the proposed methods to selecting the order of autoregressive processes.

Example 7 We generate a series from the following AR(2) model:

$$Y_t = 0.4\,Y_{t-1} + 0.32\,Y_{t-2} + \varepsilon_t, \qquad \varepsilon_t \sim N(0, 0.5^2).$$

Figure 2(a) depicted a realization of length 600. As in Example 5, the aim is to choose an order p to best predict the series.

Recall the filter $\widehat{f}_{t,p}$ defined in Example 5 is based on an autoregressive model of order p. For the global choice, we choose p to minimize $\sum_{t=101}^{600}\left(Y_t - \widehat{f}_{t,p}\right)^2$, among the set $\mathcal{P} = \{1,\,2,\,4,\,8\}$. For this realization, the above order selection rule yields $\widehat{p} = 2$, which is the same as the true value of p. The resulting filter is depicted in Figure 2(b). The estimate is very well in accordance with the mean process f_t.

The simulation results in terms of MAFE are reported in Table 1.

The next example deals with the situation where the dynamic of an underlying process changes over time. This reflects some extent in the real world, where stochastic dynamics such as stock markets can change over time.

Example 8 We simulated a process from the AR(2) process $Y_t = 0.3\,Y_{t-1} + 0.4\,Y_{t-2} + 0.3\varepsilon_t$ till the time point $t = 450$ and then from the AR(1)-process $Y_t = 0.7\,Y_{t-1} + 0.3\,\varepsilon_t$ after $t = 450$. Here, ε_t is a standard Gaussian noise. Figure 3(a) depicts a realization from the

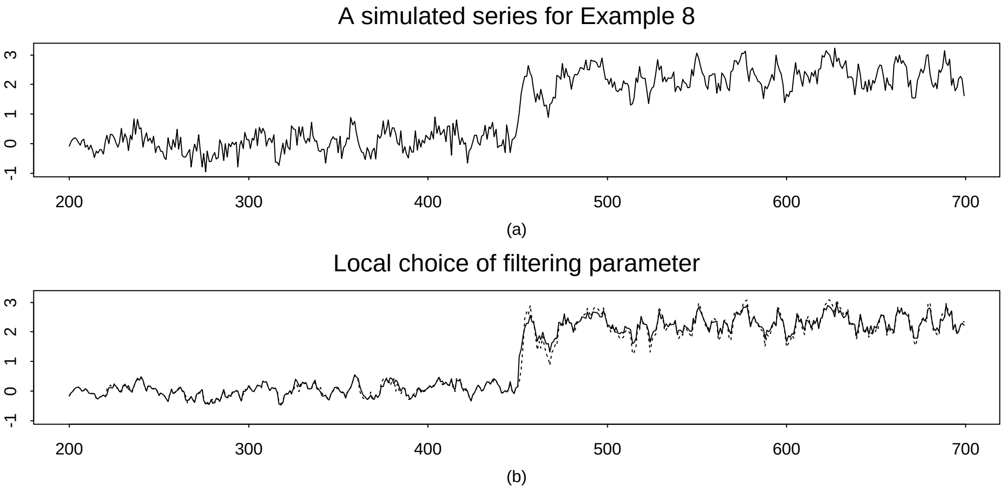

Figure 3. (a) A simulated series from Example 7. (b) Filtered series with global choices of filtering parameters. True f_t — solid curve; filtered series — dotted curve.

model. This model is similar to the thresholding autoregressive model (see [10] and [18]), but the structure change occurs in the time domain rather than the state domain.

To accommodate the possible structural break over time, it is natural to use only a local stretch of data around the time t. Similarly to Example 5, we consider the family of filters

$$\widehat{f}_{t,p,m} = X_{t,p}^{\top} \left(\sum_{s=t-m}^{t-1} X_{s,p} X_{s,p}^{\top} \right)^{-1} \sum_{s=t-m}^{t-1} X_{s,p} Y_s.$$

At each time point, a local selection procedure is applied to choose both p and m to minimize $\sum_{s=t-M+1}^{t} \left(Y_s - \widehat{f}_{s,p,m}\right)^2$. We searched p in $\mathcal{P} = \{1,\ 2,\ 4,\ 8\}$ and m in $\{\,20, 40, 80, 160\,\}$ and took $M = 20$. The result filter is plotted in Figure 3(b). The result illustrates how this procedure works in the stationary region before the change and immediately after it. In particular, the delay between the change and the first moment when the procedure starts to selects a small m can be interpreted as the sensitivity to changes.

We now briefly summarize the simulation results using MAFE. The relative performance of a filter to another one is measured by the ratio of the MAFE of the former filter to the latter. This ratio is independent of the scale of a simulated data. Table 1 summarizes the distributions of these ratios over 500 simulations by computing their mean, SD, the first, second and third quartiles. From the right block of Table 1, one can see easily that the relative performance between the filters with their parameters chosen by data and those using the ideal ones is nearly the same. This is consistent with the theoretical result given by Theorem 1. For each realization, we computed the relative MAFE of the filters with filtering parameters chosen by data to that with ideal filtering parameters. The in-sample period is set to be [1, 1000] and the post-sample period is [1001, 1500].

Table 1
Relative MAFE performance. Empirical mean (first row), sample standard deviation (second row), first quartile (third row), median (fourth row), and third quartile (fifth row) of MAFE ratios.

	Relative to $\widehat{f}_{t,\widehat{p}}$			Relative to ideal counterparts			
	ES global	ES local	AR local	ES global	ES local	AR global	AR local
	0.931	0.743	1.228	1.000	1.076	0.100	1.104
	0.065	0.138	0.069	0.034	0.057	0.0197	0.041
Ex. 6	0.901	0.655	1.177	0.996	1.036	1.000	1.075
	0.936	0.745	1.226	1.000	1.065	1.000	1.101
	0.967	0.845	1.268	1.000	1.111	1.000	1.131
	10.789	14.254	8.200	1.003	1.290	1.089	2.559
	6.534	8.619	4.940	0.010	0.103	0.608	0.449
Ex. 7	6.191	8.450	4.785	1.000	1.217	1.000	2.252
	8.988	11.790	6.912	1.000	1.280	1.000	2.499
	13.453	17.695	10.039	1.004	1.352	1.000	2.810
	1.001	0.952	0.759	1.014	1.149	1.001	1.320
	0.086	0.083	0.081	0.049	0.068	0.010	0.093
Ex. 8	0.941	0.899	0.705	1.000	1.097	1.000	1.251
	0.993	0.945	0.753	1.000	1.138	1.000	1.312
	1.057	1.009	0.809	1.000	1.194	1.000	1.381

The results in the left block of Table 1 summarize the relative performance among 4 different filters: ES global (using $\widehat{\lambda}$), ES local (using $\widehat{\lambda}_t$), AR global (using $\widehat{p}$) and AR local (using $\widehat{p}_t$ and $\widehat{m}_t$). All filters are compared with the AR global filter. This avoids the scale problems, which vary from one simulation to another. For Example 6, the best procedure among 4 competitors is ES local, followed by ES global and AR global. This is consistent with our intuition, since the data were not generated from an AR model, but a piecewise AR model. For Example 7, since the data were generated from an AR(2) model, AR global performs the best, followed by AR local. The performance of the AR local filter can be much better than what we presented here, if we allow the upper bound of m to take a larger value. ES global outperforms the ES local, since the data are stationary. The AR local performs the best for Example 8, since the model is a piecewise AR model. The ES local performs outstandingly, thanks to its flexibility.

4. Applications to volatility estimation

Let $S_1, \ldots, S_T$ be the prices of an asset or yields of a bond. The return of such an asset or bond process is usually described via the conditional heteroscedastic model:

$$R_t = \sigma_t \varepsilon_t \tag{9}$$

where $R_t = \log(S_t/S_{t-1})$, σ_t is the predictable volatility process and ε_t's are the standardized innovations.

4.1. Connections with filtering problems

The volatility is associated with the notion of risks. For many purposes in financial practice, such as portfolio optimization, option pricing and prediction of Value-at-Risk, one would be interested in predicting the volatility σ_t based on the past observations $S_1, \ldots, S_{t-1}$ of the asset process. The distribution of the innovation process can be skewed or have heavy tails. To produce a robust procedure, following [15], we consider the power transformation $Y_t = |R_t|^\gamma$ for some γ. Then, the model (9) can be written in the following semi-martingale form:

$$Y_t = C_\gamma \sigma_t^\gamma + D_\gamma \sigma_t^\gamma \xi_t \equiv f_t + v_t \xi_t \tag{10}$$

with $C_\gamma = \boldsymbol{E}|\varepsilon_t|^\gamma$, $D_\gamma^2 = \operatorname{Var}|\varepsilon_t|^\gamma$ and $\xi_t = D_\gamma^{-1}\left(|\varepsilon_t|^\gamma - C_\gamma\right)$. Mercurio and Spokoiny [15] argued that the choice $\gamma = 1/2$ leads to a nearly Gaussian distribution of the 'innovations' ξ_t, when $\varepsilon_t \sim N(0,1)$. In particular, $\boldsymbol{E} e^{u\xi_t} \le e^{u^2/(2a)}$ with $a \approx 1.005$, a condition in Lemma 1.

Now the original problem is clearly equivalent to estimating the drift coefficient $f_t = C_\gamma \sigma_t^\gamma$ from the 'observations' $Y_s = |R_s|^\gamma$, $s = 1, \ldots, t-1$. The semi-martingale representation (10) is a specific case of the model (1) with $v_t = D_\gamma \sigma_t^\gamma$. Hence, the techniques introduced in Section 3 are still applicable.

There is a large literature on the estimation of volatility. In addition to the famous parametric models such as ARCH and GARCH (see [10] and [12]), stochastic volatility models have also received a lot of attention (see, for example, [1], [2] and [4] and references therein). We here consider only the ARCH and GARCH models in addition to the nonparametric methods (MA and ES) in Section 3.

4.2. Choice of orders of ARCH and GARCH

Commonly used parametric techniques for modeling volatility are ARCH [6] and GARCH [3] models. See [10] and [12] for an overview of the field. In the current context, ARCH model assumes the following autoregressive structure:

$$\boldsymbol{E}\left[Y_t \mid \mathcal{F}_{t-1}\right] = \theta_1 Y_{t-1} + \ldots + \theta_p Y_{t-p}$$

The coefficients $\theta = (\theta_1, \ldots, \theta_p)^\top$ can be estimated by using the least-squares approach:

$$\widehat{\theta}_p = \left(\sum_{s=t_0}^{t-1} X_{s,p} X_{s,p}^\top\right)^{-1} \sum_{s=t_0}^{t-1} X_{s,p} Y_s$$

with $X_{s,p} = (Y_{s-1}, \ldots, Y_{s-p})^\top$. The estimate $\widehat{f}_{t,p}$ is then defined by $\widehat{f}_{t,p} = X_{t,p}^\top \widehat{\theta}_p$. As in Section 3, the order p can be chosen by minimizing the prediction error:

$$\widehat{p} = \operatorname*{arginf}_{p \le p^*} \sum_{s=t_0}^{t} (Y_s - \widehat{f}_{s,p})^2 \tag{11}$$

The upper bound p^* should be sufficiently large to reduce possible approximation errors. To facilitate computation, t in (11) can be replaced by T, the length of the time series in the in-sample period. The approach is a global choice of the order of an ARCH model.

The volatility process σ_t in GARCH(p,q) is modeled as

$$\sigma_t^2 = c_0 + \alpha_1\sigma_{t-1}^2 + \ldots + \alpha_p\sigma_{t-p}^2 + \beta_1 R_{t-1}^2 + \ldots + \beta_q R_{t-q}^2.$$

The coefficients α_j, β_k can be estimated by using the maximum likelihood method. See for example Fan and Yao [10]. The estimates $\widehat{\alpha}_j$ and $\widehat{\beta}_k$ are then used to construct the filter

$$\widehat{f}_{t,p,q} = C_\gamma \left(\sum_{j=1}^{p} \widehat{\alpha}_j \sigma_{t-j}^2 + \sum_{k=1}^{q} \widehat{\beta}_k R_{t-k}^2 \right)^{\gamma/2}.$$

The order (p,q) can be chosen to minimize a quantity that is similar to (11).

GARCH(1,1) is one of most frequently used models in volatility estimation in financial time series. It has been observed to fit well many financial time series. To simplify the computation efforts, we mainly focus on the GARCH(1,1) rather than general GARCH(p,q) in our simulation studies.

4.3. Simulated financial time series

We simulated time series from the volatility model

$$\begin{aligned}
\text{GARCH(1,1)}: &\quad \sigma_t^2 = 0.00005 + 0.85\sigma_{t-1}^2 + 0.1R_{t-1}^2\\
\text{GARCH(1,3)}: &\quad \sigma_t^2 = 0.00002 + 0.8\sigma_{t-1}^2 + 0.02R_{t-1}^2 + 0.05R_{t-2}^2 + 0.11R_{t-3}^2.\\
\text{ARCH(2)}: &\quad \sigma_t^2 = 0.00085 + 0.1R_{t-1}^2 + 0.05R_{t-2}^2.
\end{aligned}$$

As shown in (10), the problem of volatility estimation is closely related to the filtering problems in Section 3. Therefore, the measure of effectiveness of each method can be gauged by MAFE and MSFE in Section 3.4. Tables 2 and 3 summarize the result for $\gamma = 0.5$ and $\gamma = 2$ in a similar format to Table 1. Table 4 summarizes the results using "rank" as a measure. For example, for the GARCH(1,3) model (second block), using untransformed data transformation (right block), in terms of MAFE, among 500 simulations, the GARCH(1,1), ES and AR methods ranked respectively, 334, 162 and 4 times in the first place, 159, 309 and 32 times in the second place and 7, 29 and 464 times in the third place.

First of all, from Tables 2 and 3, the ES and AR with their parameters chosen from data perform nearly as well as their corresponding estimators using the ideal filtering parameters. This is consistent with our theoretical result, which is the theme of our study. The GARCH(1,1) and ES estimation methods are quite robust. When the true model is GARCH(1,1), the GARCH(1,1) method performs the best, as expected, followed by ES global and then AR global. When the true model is the GARCH(1,3), which can still reasonably be well approximated by a GARCH(1,1) model, the performance of the GARCH(1,1) method and the ES method is nearly the same, though the GARCH(1,1) method performs somewhat better. It is clear that the relative performance of the GARCH(1,1) method gets deteriorated from the GARCH(1,1) model to the GARCH(1,3) model. When the series comes from the ARCH(2) model, the AR filter performs the best, as expected.

Table 2
Relative MAFE performance. ES and AR filtering of $Y_t = |R_t|^{1/2}$. Empirical mean (first row), sample standard deviation (second row), first quartile (third row), median (fourth row), and third quartile (fifth row) of MAFE ratios.

	Relative to GARCH(1,1)		Relative to ideal counterparts	
Model	ES global	AR global	ES global	AR global
	2.898	3.078	1.026	1.095
	1.747	2.045	0.060	0.164
GARCH(1,1)	1.816	1.900	0.998	1.000
	2.464	2.544	1.006	1.060
	3.381	3.564	1.050	1.187
	1.485	1.610	1.034	1.063
	0.246	0.304	0.070	0.122
GARCH(1,3)	1.314	1.401	1.000	1.000
	1.482	1.571	1.000	1.034
	1.639	1.794	1.051	1.120
	2.914	1.330	1.000	1.061
	1.448	0.731	0.000	0.131
ARCH(2)	1.899	0.797	1.000	1.000
	2.575	1.139	1.000	1.000
	3.473	1.626	1.000	1.111

Table 3
Relative MAFE performance. ES and AR filtering of $Y_t = R_t^2$. Empirical mean (first row), sample standard deviation (second row), first quartile (third row), median (fourth row), and third quartile (fifth row) of MAFE ratios.

	Relative to GARCH(1,1)		Relative to ideal counterparts	
Model	ES global	AR global	ES global	AR global
	2.119	2.789	1.115	1.171
	1.413	1.823	0.152	0.249
GARCH(1,1)	1.283	1.655	1.010	1.000
	1.815	2.340	1.055	1.101
	2.449	3.318	1.165	1.308
	1.111	2.147	1.132	1.179
	0.222	2.381	0.181	0.291
GARCH(1,3)	0.971	1.448	1.000	1.000
	1.092	1.778	1.070	1.108
	1.220	1.171	1.181	1.325
	2.484	0.964	1.002	1.152
	1.229	0.562	0.032	0.353
ARCH(2)	1.632	0.565	1.000	1.000
	2.166	0.816	1.000	1.000
	2.939	1.237	1.000	1.213

Table 4
Rank performance of GARCH(1,1), ES global, and AR global.

Model	Filtering $Y_t = \|R_t\|^{1/2}$ GARCH(1,1)	ES	AR	Filtering $Y_t = R_t^2$ GARCH(1,1)	ES	AR
	487	9	4	451	42	7
GARCH(1,1)	11	286	203	40	383	77
	2	205	293	9	75	416
	491	7	2	334	162	4
GARCH(1,3)	8	347	145	159	309	32
	1	146	353	7	29	464
	299	0	201	183	0	317
ARCH(2)	197	4	299	310	8	182
	4	496	0	7	492	1

4.4. Applications

We apply the GARCH(1,1) approach $\widehat{f}_{t,1,1}$, the adaptive global ES smoothing $\widehat{f}_{t,\widehat{\lambda}}$, and the global AR smoothing $\widehat{f}_{t,\widehat{p}}$ to estimate the volatility of the log-returns of the S&P500 index and the three-month Treasury Bills. For the ES and AR approaches, we consider the square root transformation $Y_t = |R_t|^{1/2}$, which yields more stable estimates than the square transformation $Y_t = R_t^2$. The order of the AR filtering was searched among the candidate set $\mathcal{P} = \{1, \cdots, 15\}$ and the collection of grids of ES smoothing parameters was taken to be $\Lambda = \{[5 \times 1.2^k], k = 0, 1, \cdots, 15\}$. For the real data, we don't know the true volatility. Hence, we use the Average of Prediction Errors (APE) as a measure of effectiveness:

$$APE1 = \frac{1}{T - t_0 + 1} \sum_{t=t_0}^{T} \left(|R_t| - C_1 \widehat{\sigma}_t\right)^2 \quad \text{and} \quad APE2 = \frac{1}{T - t_0 + 1} \sum_{t=t_0}^{T} \left|R_t^2 - \widehat{\sigma}_t^2\right|.$$

As noted in [8], the prediction errors consist of stochastic errors and estimation (filtering) errors. The former is independent of estimation methods and dominates the latter. Therefore, a small percentage of improvement in prediction errors implies a large improvement in the filtering error.

Table 5
Relative prediction performance for yields of three-month Treasury Bills over four different periods

Time period	ES $\widehat{f}_{t,\widehat{\lambda}}$ relative to GARCH(1,1) $APE1$	$APE2$	AR $\widehat{f}_{t,\widehat{p}}$ relative to GARCH(1,1) $APE1$	$APE2$
12/09/55–07/02/65	1.012	1.038	1.051	0.979
06/09/67–12/31/76	0.956	0.889	0.983	0.858
12/08/78–07/01/88	0.772	0.696	0.840	0.724
06/08/90–12/31/99	1.004	0.879	0.989	0.948

The three-month Treasury bills data consist of weekly observations (Fridays' closing) of interest rates of the three-month Treasury bills, from August 1954 to December 1999. The rates are based on quotes at the official close of the U.S. government securities market on a discount basis. To attenuate the time effects, we divided the entire series into four sub-series. The gaps between the time periods are the length t_0 used for the subsequent series. The volatility is computed based on the difference of the yields series. The relative performance of global ES and global AR smoothing and GARCH(1,1) is given in Table 5. The values are smaller than one most of the time and are sometimes as small as 0.696. This implies that with the adaptive choice of filtering parameters, the exponential smoothing and the autoregressive model outperform the GARCH(1,1) model for the periods studied. Figure 4 depicts first one hundred lags of the autocorrelation of the absolute returns and the absolute returns divided by the standard deviations estimated by the three methods. The horizontal lines indicate the 95% confidence limits. All of the three estimation methods explain well the volatility: the standardized returns rarely exhibit significant correlations.

Table 6
Relative prediction performance for the Standard and Poor 500 index over two time periods.

	ES $\widehat{f}_{t,\widehat{\lambda}}$ relative to GARCH(1,1)		AR $\widehat{f}_{t,\widehat{p}}$ relative to GARCH(1,1)	
Time period	$APE1$	$APE2$	$APE1$	$APE2$
03/08/90–18/07/94	0.950	0.883	1.002	0.983
08/12/94–20/11/98	0.993	0.952	1.031	0.898

The S&P500 data consist of the daily closing of the Standard and Poor 500 index. The volatility estimation methods are applied to the data in the time periods 03/08/90–18/07/94 and 08/12/94–20/11/98. Again the AR and ES methods with our adaptive choice of filtering parameters provide satisfactory estimate of the underlying volatility. The ACF plots of the standardized log-returns (not shown here, similar to Figure 4) indicate success of the three methods. The relative performance against GARCH(1,1) is shown in Table 6. Again, the ES and AR filters with filtering parameters chosen by data outperform the GARCH(1,1).

The adaptive local ES filter and local AR filter were also applied to the above two data sets. We do not report the details here to save space. They both perform reasonably well. However, the local ES method does not perform as well as global one. The local AR filter performs quite well and is often better than the global AR filter, for the two financial series data that we examined.

REFERENCES

1. Barndoff-Neilsen, O.E. and Shephard, N. (2001). Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics (with discussion). *Jour. Roy. Statist. Soc. B*, **63**, 167-241.

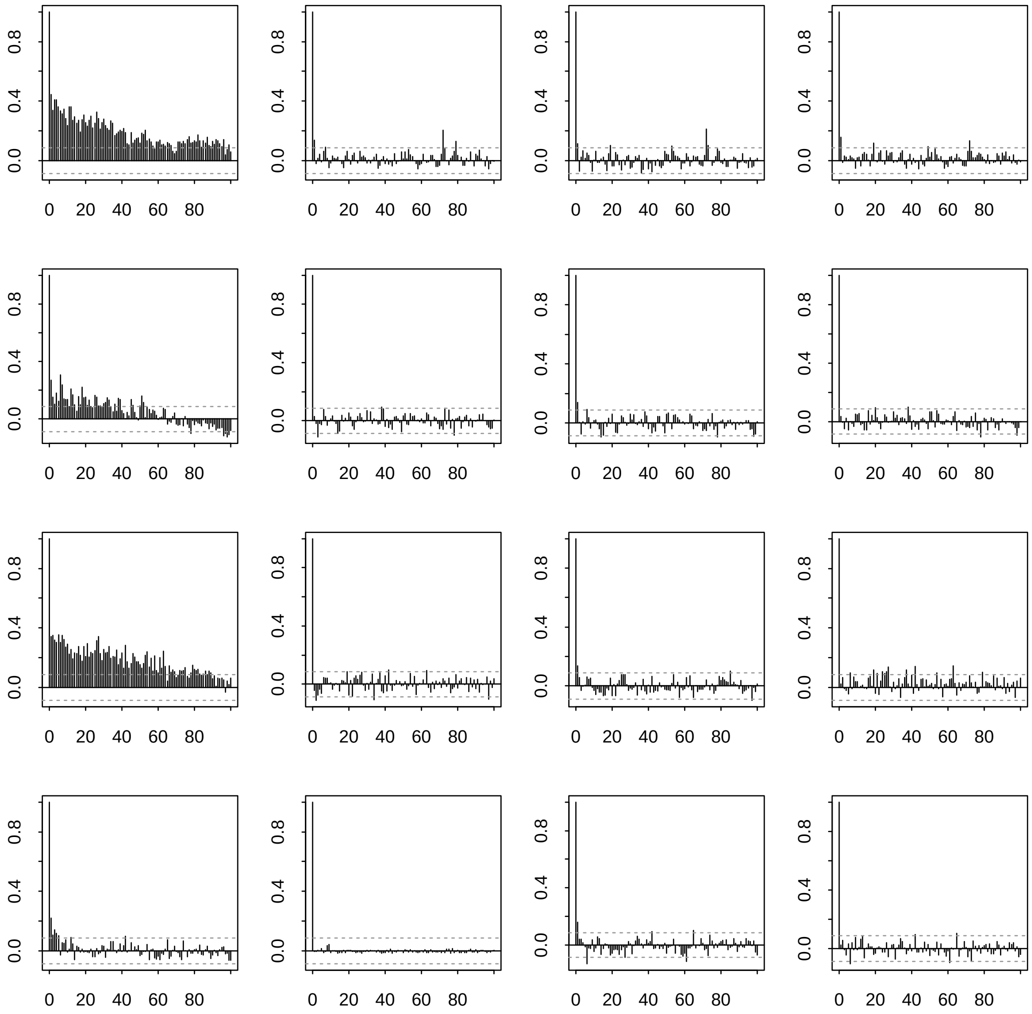

Figure 4. First one hundred lags of the autocorrelation function (ACF). Left to right: ACF of the absolute log-returns, ACF of the absolute log-returns divided by volatility estimated by GARCH(1,1) model, global ES, and global AR. From top to bottom: time periods 12/09/55–07/02/65, 06/09/67–12/31/76, 12/08/78–07/01/88, and 06/08/90–12/31/99.

2. Barndoff-Neilsen, O.E. and Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Jour. Roy. Statist. Soc. B*, **64**, 253-280.
3. Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, **31**, 307-327.
4. Bollerslev, T. and Zhou, H. (2002). Estimating stochastic volatility diffusion using conditional moments of integrated volatility. *Jour. Econometrics*, **109**, 33-65.
5. Brockmann, M., Gasser, T. and Herrmann, E. (1993). Locally adaptive bandwidth choice for kernel regression estimators. *Jour. Amer. Statist. Assoc.*, **88**, 1302–1309.

6. Engle, R.F. (1982). Autoregressive conditional heteroscedasticy with estimates of the variance of U.K. inflation. *Econometrica*, **50**, 987-1008.
7. Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Royal Statist. Soc. B*, **57**, 371–394.
8. Fan, J. and Gu, J. (2003). Data-analytic approaches to the estimation of value-at-risk. *2003 International Conference on Computational Intelligence for Financial Engineering*, to appear.
9. Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, **85**, 645-660.
10. Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer-verlag, New York.
11. Gijbels, I., Pope, A., Wand, M.P. (1999). Understanding exponential smoothing via kernel regression. *Journal of the Royal Statistical Society, Series B*, **61**, 39–50.
12. Gouriéroux, C. (1997) ARCH models and financial applications. Springer-Verlag, New York.
13. Jorion, P. (2000). *Value at Risk: The new benchmark for managing financial risk* (2nd ed.). McGraw-Hill, New York.
14. Liptser, R. and Spokoiny, V. (2000). Deviation probability bound for martingales with applications to statistical estimation. *Statist. Probab. Lett.*, **46**, 347–357.
15. Mercurio, D. and Spokoiny, V. (2003). Statistical inference for time-inhomogeneous volatility models. *Ann. Statist.*, to appear.
16. Morgan, J. P. (1996). *RiskMetrics Technical Document.* Fourth edition, New York.
17. Ruppert, D., Wand, M.P., Holst, U. and Hössjer, O. (1997). Local polynomial variance function estimation. *Technometrics*, **39**, 262-273.
18. Tong, H. (1990). *Non-Linear Time Series: A Dynamical System Approach.* Oxford University Press, Oxford.