Robust Risk Management. Accounting for Nonstationarity and Heavy Tails

Chen, Ying
Weierstrass Institute,
Mohrenstr. 39,
10117 Berlin, Germany
chen@wias-berlin.de

Spokoiny, Vladimir
Weierstrass Institute,
Mohrenstr. 39,
10117 Berlin, Germany
spokoiny@wias-berlin.de

Abstract

In the ideal Black-Scholes world, financial time series are assumed 1) stationary (time homogeneous) and 2) having conditionally normal distribution given the past. These two assumptions have been widely-used in many methods such as the RiskMetrics, one risk management method considered as industry standard. However these assumptions are unrealistic. The primary aim of the paper is to account for nonstationarity and heavy tails in time series by presenting a local exponential smoothing approach, by which the smoothing parameter is adaptively selected at every time point and the heavy-tailedness of the process is considered. A complete theory addresses both issues. In our study, we demonstrate the implementation of the proposed method in volatility estimation and risk management given simulated and real data. Numerical results show the proposed method delivers accurate and sensitive estimates.

Keywords: exponential smoothing; spatial aggregation; heavy-tailed distribution.

1 Introduction

In the ideal Black-Scholes world, financial time series are assumed 1) stationary (time homogeneous) and 2) having conditionally normal distribution given the past. These two assumptions have been widely-used in many methods such as the RiskMetrics which has been considered as industry standard in risk management after introduced by J.P. Morgan in 1994. However, these assumptions are very questionable as far as the real life data is concerned. The time homogeneous assumption does not allow to model structure shifts or breaks on the market and to account for e.g. macroeconomic, political or climate changes. The assumption of conditionally Gaussian innovations leads to underestimation of the market risk. Recent studies show that the Gaussian and sub-Gaussian distributions are too light to model the market risk associated with sudden shocks and crashes and heavy-tailed distributions like Student-t or Generalized Hyperbolic are more appropriate. A realistic risk management system has to account for the both stylized facts of the financial data, which is a rather complicated task. The reason is that these two issues are somehow contradictory. A robust risk management which is stable against extremes and large shocks in financial data is automatically less sensitive to structural changes and vice versa. The aim of the present paper is to offer an approach for a flexible modeling of financial time series which is sensitive to structural changes and robust against extremes and shocks on the market.

1.1 Accounting for Non-stationarity

It is rational to surmise that the structure of volatility process shifts through time, possibly due to policy adjustments or economic changes. This non-stationary effect is illustrated in Figure 1, by which the realized variances, the sum of squared returns sampled at 15 minutes tick-by-tick, of Dow Jones Euro StoXX 50 Index futures are presented ranging from December 8, 2004 to May 2, 2005. The realized variance measure has been considered as a robust estimator of the variance of financial asset, see Anderson, Bollerslev, Diebold and Labys (2001) and Zhang, Mykland and Ait-Sahalia (2005). We here use the realized variance

to illustrate the movement of the unobserved variance. In the figure, an evident change of market situation is observed in the last 10 days. It indicates that volatility estimates obtained by averaging over a long historical interval will significantly underestimate the current volatility and lead to a large estimation bias.

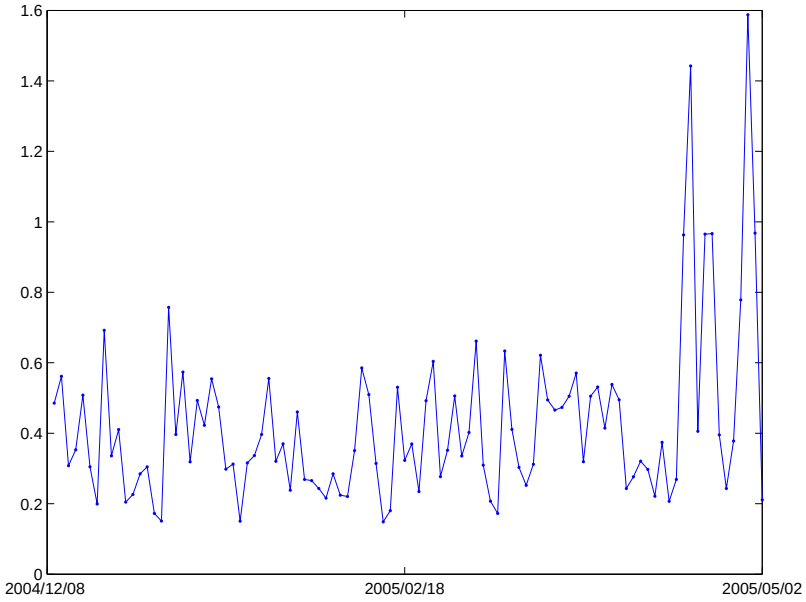

Figure 1: The realized variances, the sum of squared returns sampled at 15 minutes tick-by-tick, of Dow Jones Euro StoXX 50 Index futures ranging from December 8, 2004 to May 2, 2005.

The standard way of accounting for non-stationarity is to recalibrate (reestimate) the model parameters at every time point using the latest available information from a *time varying window*. Alternatively, the *exponential smoothing* approach assigns some weights to historical data which exponentially decrease with the time. The choice of a small window or rapidly decreasing weights results in high variability of the estimated volatility and, as a consequence, of the estimated value of the portfolio risk from day to day. In turns, a large window or a low pass volatility filtering method results in the loss of sensitivity of the risk management system to the significant changes of the market situation.

An *adaptive* approach aims to select large windows or slowly decreasing weights in the time homogeneous situation and it switches to high pass filtering if some structural change

is detected.

Recently a number of local parametric methods has been developed, which investigates the structure shifts, or equivalently to say, adjusts the smoothing parameter to avoid serious estimation errors and achieve the best possible accuracy of estimation. For example, Fan and Gu (2003) introduce several semiparametric techniques of estimating volatility and portfolio risk. Mercurio and Spokoiny (2004) present an approach to specify local homogeneous interval, by which volatility is approximated by a constant. Belomestny and Spokoiny (2006) present the spatial aggregation of the local likelihood estimates (SSA). Among others, we refer to Spokoiny (2006) for a detailed description of the local estimation methods. These works however concern only one issue, namely the nonstationarity of time series, and rely on the unrealistic Gaussian distributional assumption.

1.2 Accounting for Heavy Tails in Innovations

As already mentioned, the evidence of non-Gaussian heavy-tailed distribution for the standardized innovations of the financial time series is well documented. For instance, Student-t or Generalized Hyperbolic distributions are much more accurate in estimating the quantiles of the standardized returns, see e.g. Embrechts, McNeil and Straumann (2002) and Eberlein and Keller (1995), among other. However, the existent methods and approaches to modeling such phenomena are based on one or another kind of parametric assumptions, and hence, are not flexible for modeling structural changes in the financial data.

The primary aim of the paper is to present a realistic approach that accounts for the both features: nonstationarity and heavy tails in financial time series. The whole approach can be decomposed in few steps. First we develop an adaptive procedure for estimation of the time dependent volatility under the assumption of the conditionally Gaussian innovations. Then we show that the procedure continues to apply in the case of sub-Gaussian innovations (under some exponential moment conditions). To make this approach applicable to the heavy-tailed data, we make a power transformation of the underlying process. Box and Cox (1964) stimulated the application of the power transformation to non-Gaussian variables to

obtain another distribution more close to the normal and homoscedastic assumption. Here we follow this way and replace the squared returns by their p-power to provide that the resulting "observations" have exponential moments.

1.3 Volatility Estimation by Exponential Smoothing

Let S_t be an observed asset process in discrete time, $t = 1, 2, \ldots$, while R_t defines the corresponding return process: $R_t = \log(S_t/S_{t-1})$. We model this process via the *conditional heteroskedasticity* assumption:

$$R_t = \sqrt{\theta_t}\varepsilon_t\,, \tag{1.1}$$

where ε_t, $t \geq 1$, is a sequence of standardized innovations satisfying

$$I\!E\big(\varepsilon_t \mid \mathcal{F}_{t-1}\big) = 0, \qquad I\!E\big(\varepsilon_t^2 \mid \mathcal{F}_{t-1}\big) = 1$$

where $\mathcal{F}_{t-1} = \sigma(R_1, \ldots, R_{t-1})$ is the (σ-field generated by the first $t-1$ observations), and θ_t is the *volatility* process which is assumed to be predictable with respect to $\mathcal{F}_{t-1}$.

In this paper we focus on the problem of filtering the parameter θ_t from the past observations $R_1, \ldots, R_{t-1}$. This problem naturally arises as an important building block for many tasks of financial engineering like Value-at-Risk or Portfolio Optimization. Among others, we refer to Christoffersen (2003) for a systematic introduction of risk analysis.

The exponential smoothing (ES) and its variation have been considered as good functional approximations of variance by assigning weights to the past squared returns:

$$\theta_t \;=\; \frac{1}{1-\eta}\sum_{m=0}^{\infty}\eta^m R_{t-m-1}^2, \quad \eta \in [0,1).$$

Many time series models such as the ARCH proposed by Engle (1982) and the GARCH by Bollerslev (1986) can be considered as variation of the ES. For example, the GARCH(1,1)

setup can be reformulated as:

$$\theta_t \;=\; \omega + \alpha R_{t-1}^2 + \beta\theta_{t-1} = \frac{\omega}{1-\beta} + \alpha \sum_{m=0}^{\infty} \beta^m R_{t-m-1}^2.$$

With a proper reparametrization, this is again an exponential smoothing estimate.

It is worth noting that the ES is in fact a local maximum likelihood estimate (MLE) based on the Gaussian distributional assumption of the innovations, see e.g. Section 2. One can expect that this method also does a good job if the innovations are not conditionally Gaussian but their distribution is not far away from normal. Our theoretical and numerical results confirm this hint for the case of a sub-Gaussian distribution of the innovations ε_t, see Section 2 for more details.

To implement the ES approach, one first faces the problem to choose the smoothing parameter η (or β) which can be naturally treated as a memory parameter. The values of η close to one correspond to a slow decay of the coefficients η^m and hence, to a large averaging window, while the small values of η result in a high-pass filtering. The classical ES methods choose one constant smoothing (memory) parameter. For instance, in the RiskMetrics design, $\eta = 0.94$ has been thought of as an optimized value. This, however, raises the question whether the experience-based value is really better than others. Another more reliable but computationally demanding approach is to choose η by optimizing some objective function such as forecasting errors (Cheng, Fan and Spokoiny, 2003) or log-likelihood function (Bollerslev and Woolridge, 1992).

In our study, the smoothing parameter is adaptively selected at every time point. Given a finite set $\eta_1, \ldots, \eta_K$ of the possible values of the memory parameter, we calculate K local MLEs $\{\widetilde{\theta}_t^{(k)}\}$ at every time point t. Then these "weak" estimates are aggregated in one adaptive estimate by using the *Spatial Stagewise Aggregation* (SSA) procedure from Belomestny and Spokoiny (2006). Alternatively, we choose one η_k such that its corresponding MLE $\widetilde{\theta}_t^{(k)}$ has the best performance in the estimation among the considered set of K estimates, referred as LMS. Furthermore, we extend the local exponential smoothing in the

heavy-tailed distributional framework. Chen, Härdle and Jeong (2005) show that the normal inverse Gaussian (NIG) distribution with four distributional parameters is successful in imitating the distributional behavior of real financial data. It is therefore practically interesting to show that the quasi ML estimation is applicable under the NIG distributional assumption. Finally, we demonstrate the implementation of the proposed local exponential smoothing method in volatility estimation and risk management.

The paper is organized as follows. The local exponential smoothing is described, by which the SSA and LMS methods are used to select the smoothing parameter in Section 2. In particular, Section 2.4 investigates the choice of parameters involved in the localization. Sensitivity analysis is reported. Later in this section, an alternative parameter tuning is illustrated by minimizing forecasting errors. The quasi ML estimation under the NIG distributional assumption is discussed in Section 3. Section 4 compares the proposed methods with the stationary ES approach based on simulated data. Moreover, risk exposures of two German assets, one US equity and two exchange rates are examined using the proposed local volatility estimation under the normal and NIG distributional assumption.

Our theoretical study in Section 2.2 claims a kind of "oracle" optimality for the proposed procedure while the numerical results for simulated and real data demonstrates the quite reasonable performance of the method in the situations we focus on.

2 Accounting for Non-Stationarity. Gaussian and Sub-Gaussian Innovations

This section presents the method of adaptive estimation of time inhomogeneous volatility process θ_t based on aggregating the ES estimates with different memory parameters η. For this section the innovations ε_t in the model (1.1) are assumed to be Gaussian or sub-Gaussian. An extension to heavy-tailed innovations will be discussed in Section 3.

We follow the local parametric approach from Spokoiny (2006). First we show that the ES estimate is a particular case of the local parametric volatility estimate and study some

of its properties. Then we introduce the SSA procedure for aggregating a family of "weak" ES estimates into one adaptive volatility estimate and study its properties in the case of sub-Gaussian innovations.

2.1 Local Parametric Modeling

A *time-homogeneous* (*time-homoskedastic*) model means that θ_t is a constant. For the homogeneous model $\theta_t \equiv \theta$ for t from the given time interval I, the parameter θ can be estimated using the (quasi) maximum likelihood method. Suppose first that the innovations ε_t are conditionally on $\mathcal{F}_{t-1}$ standard normal. Then the joint distribution of R_t for $t \in I$ is described by the log-likelihood

$$L_I(\theta) = \sum_{t \in I} \ell(Y_t, \theta)$$

where $\ell(y, \theta) = -(1/2)\log(2\pi\theta) - y/(2\theta)$ is the log-density of the normal distribution $\mathcal{N}(0, \theta)$ and Y_t mean the squared returns, $Y_t = R_t^2$. The corresponding maximum likelihood estimate (MLE) maximizes the likelihood:

$$\widetilde{\theta}_I = \operatorname*{argmax}_{\theta \in \Theta} L_I(\theta) = \operatorname*{argmax}_{\theta \in \Theta} \sum_{t \in I} \ell(Y_t, \theta),$$

where Θ is a given parametric subset in $\mathbb{R}_+$.

If the innovations ε_t are not conditionally standard normal, the estimate $\widetilde{\theta}_I$ is still meaningful and it can be considered as a *quasi MLE*.

The assumption of time homogeneity is usually too restrictive if the time interval I is sufficiently large. The standard approach is to apply the parametric modeling in a vicinity of the point of interest t. The localizing scheme is generally given by the collection of weights $W_t = \{w_{st}\}$ which leads to the *local log-likelihood*

$$L(W_t, \theta) = \sum_s \ell(Y_s, \theta) w_{st}$$

and to the local MLE $\widetilde{\theta}_t$ defined as the maximizer of $L(W_t, \theta)$. In this paper we only consider the localizing scheme with the exponentially decreasing weights $w_{st} = \eta^{t-s}$ for $s \le t$, where η is the given "memory" parameter. We also cut the weights when they become smaller than some prescribed value $c > 0$, e.g. $c = 0.01$. However, the properties of the local estimate $\widetilde{\theta}_t$ are general and apply to any localizing scheme.

We denote by $\widetilde{\theta}_t$ the value maximizing the local log-likelihood $L(W_t, \theta)$:

$$\widetilde{\theta}_t = \operatorname*{argmax}_{\theta \in \Theta} L(W_t, \theta).$$

The volatility model is a particular case of an exponential family, so that a closed form representation for the local MLE $\widetilde{\theta}_t$ and for the corresponding fitted log-likelihood $L(W_t, \widetilde{\theta}_t)$ are available, see Polzehl and Spokoiny (2006) for more details.

Theorem 2.1. *For every localizing scheme W_t*

$$\widetilde{\theta}_t = N_t^{-1} \sum_s Y_s w_{st}$$

where N_t denotes the sum of the weights w_{st}:

$$N_t = \sum_s w_{st}.$$

Moreover, for every $\theta > 0$ the fitted likelihood ratio $L(W_t, \widetilde{\theta}, \theta) = \max_{\theta'} L(W_t, \theta', \theta)$ with $L(W_t, \theta', \theta) = L(W_t, \theta') - L(W_t, \theta)$ satisfies

$$L(W_t, \widetilde{\theta}_t, \theta) = N_t \mathcal{K}(\widetilde{\theta}_t, \theta) \tag{2.1}$$

where

$$\mathcal{K}(\theta, \theta') = -0.5\big\{\log(\theta/\theta') + 1 - \theta/\theta'\big\}$$

is the Kullback-Leibler information for the two normal distributions with variances θ and θ': $\mathcal{K}(\theta, \theta') = I\!\!E_\theta \log\big(I\!\!P_\theta / dI\!\!P_{\theta'}\big)$.

Proof. One can see that

$$L(W_t, \theta) = -\frac{N_t}{2} \log(2\pi\theta) - \frac{1}{2\theta} \sum_s Y_s w_{st} \tag{2.2}$$

This representation yields the both assertions of the theorem by simple algebra. □

Remark 2.1. The results of Theorem 2.1 only rely on the structure of the function $\ell(y, \theta)$ and do not utilize the assumption of conditional normality of the innovations ε_t. Therefore, they apply whatever the distribution of the innovations ε_t is.

2.2 Some Properties of the Estimate $\widetilde{\theta}_t$ in the Homogeneous Situation

This section collects some useful properties of the (quasi) MLE $\widetilde{\theta}_t$ and of the fitted log-likelihood $L(W_t, \widetilde{\theta}_t, \theta^*)$ in the homogeneous situation $\theta_s = \theta^*$ for all s. We assume the following condition on the set Θ of possible values of the volatility parameter.

($\boldsymbol{\Theta}$) The set Θ is a compact interval in $\mathbb{R}_+$ and does not containing $\theta = 0$.

First we discuss the case of Gaussian innovations ε_s.

Theorem 2.2 (Polzehl and Spokoiny, 2006). *Assume* (Θ). *Let* $\theta_s = \theta^* \in \Theta$ *for* s. *If the innovations* ε_s *are i.i.d. standard normal, then for any* $\mathfrak{z} > 0$

$$\mathbb{P}_{\theta^*}\big(L(W_t, \widetilde{\theta}_t, \theta^*) > \mathfrak{z}\big) \equiv \mathbb{P}_{\theta^*}\big(N_t \mathcal{K}(\widetilde{\theta}_t, \theta^*) > \mathfrak{z}\big) \le 2e^{-\mathfrak{z}}.$$

Theorem 2.2 claims that the estimation loss measured by $\mathcal{K}(\widetilde{\theta}_t, \theta^*)$ is with high probability bounded by $\mathfrak{z}/N_t$ provided that $\mathfrak{z}$ is sufficiently large. This result helps to establish a risk bound for a power loss function and to construct the confidence sets for the parameter θ^*.

Theorem 2.3. *Assume* (Θ). *Let* Y_t *be i.i.d. from* $\mathcal{N}(0, \theta^*)$. *Then for any* $r > 0$

$$\mathbb{E}_{\theta^*}\big|L(W_t, \widetilde{\theta}_t, \theta^*)\big|^r \equiv \mathbb{E}_{\theta^*}\big|N_t \mathcal{K}(\widetilde{\theta}_t, \theta^*)\big|^r \le \mathfrak{r}_r\,.$$

where $\mathfrak{r}_r = 2r\int_{\mathfrak{z}\geq 0}\mathfrak{z}^{r-1}e^{-\mathfrak{z}}d\mathfrak{z} = 2r\Gamma(r)$. *Moreover, if* $\mathfrak{z}_\alpha$ *satisfies* $2e^{-\mathfrak{z}_\alpha} \leq \alpha$, *then*

$$\mathcal{E}_{t,\alpha} = \bigl\{\theta : N_t\mathcal{K}\bigl(\widetilde{\theta}_t, \theta\bigr) \leq \mathfrak{z}_\alpha\bigr\} \tag{2.3}$$

is an α *-confidence set for the parameter* θ^* *in the sense that*

$$I\!P_{\theta^*}\bigl(\mathcal{E}_{t,\alpha} \not\ni \theta^*\bigr) \leq \alpha.$$

Proof. By Theorem 2.2

$$\begin{aligned} I\!E_{\theta^*}\bigl|L(W_t, \widetilde{\theta}_t, \theta^*)\bigr|^r &\leq -\int_{\mathfrak{z}\geq 0}\mathfrak{z}^r dI\!P_{\theta^*}(L(W_t, \widetilde{\theta}_t, \theta^*) > \mathfrak{z}) \\ &\leq r\int_{\mathfrak{z}\geq 0}\mathfrak{z}^{r-1}I\!P_{\theta^*}(L(W_t, \widetilde{\theta}_t, \theta^*) > \mathfrak{z})d\mathfrak{z} \leq 2r\int_{\mathfrak{z}\geq 0}\mathfrak{z}^{r-1}e^{-\mathfrak{z}}d\mathfrak{z} \end{aligned}$$

and the first assertion is fulfilled. The last assertion is proved similarly. □

The assumption of normality for the innovations ε_t is often criticized in the financial literature. The basic result of Theorem 2.2 and its corollaries can be extended to the case of non-Gaussian innovations under some exponential moment conditions. We refer to this situation as the *sub-Gaussian* case. Later these results in combination with the power transformation of the data will be used for studying the heavily tailed innovations, see Section 5.

Theorem 2.4. *Assume* (Θ) . *Let the innovations* ε_s *be i.i.d.,* $I\!E\varepsilon_s^2 = 1$, *and*

$$\log I\!E\exp\bigl\{\lambda(\varepsilon_s^2 - 1)\bigr\} \leq \varkappa(\lambda) \tag{2.4}$$

for some $\lambda > 0$ *and some constant* $\varkappa(\lambda)$. *Then there is a constant* $\mu_0 > 0$ *such that for all* $\theta^*, \theta \in \Theta$

$$I\!E_{\theta^*}\exp\bigl\{\mu_0 L(W_t, \theta, \theta^*)\bigr\} \equiv I\!E_{\theta^*}\exp\bigl\{\mu_0 N_t\mathcal{K}(\widetilde{\theta}_t, \theta^*)\bigr\} \leq 1 \tag{2.5}$$

and

$$I\!P_{\theta^*}\bigl(L(W_t,\widetilde{\theta}_t,\theta^*) > \mathfrak{z}\bigr) \equiv I\!P_{\theta^*}\bigl(N_t\mathcal{K}(\widetilde{\theta}_t,\theta^*) > \mathfrak{z}\bigr) \le 2e^{-\mu_0\mathfrak{z}}. \tag{2.6}$$

Proof. For brevity of notation we omit the subscript t. It holds for $L(W,\theta,\theta^*) = L(W,\theta) - L(W,\theta^*)$

$$2L(W,\theta,\theta^*) = -N\log(\theta/\theta^*) - (1/\theta - 1/\theta^*)\sum_s Y_s w_s\,.$$

Under the measure $I\!P_{\theta^*}$, the squared returns Y_t can be represented as $Y_t = \theta^*\varepsilon_t^2$ leading to the formula

$$\begin{aligned} 2L(W,\theta,\theta^*) &= N\log(\theta^*/\theta) - (\theta^*/\theta - 1)\sum_s \varepsilon_s^2 w_s \\ &= N\log(1+u) - u\sum_s \varepsilon_s^2 w_s = N\log(1+u) - Nu - u\sum_s(\varepsilon_s^2 - 1)w_s \end{aligned}$$

with $u = \theta^*/\theta - 1$. For any μ such that $\max_s u\mu w_s \le \lambda$ this yields by independence of the ε_s's

$$\begin{aligned} \log I\!E_{\theta^*}\bigl\{2\mu L(W,\theta,\theta^*)\bigr\} &= \mu N\log(1+u) - \mu N u + \sum_s \log I\!E_{\theta^*}\exp\bigl\{-u\mu w_s(\varepsilon_s^2 - 1)\bigr\} \\ &= \mu N\log(1+u) - \mu N u + \sum_s \varkappa(-u\mu w_s). \end{aligned}$$

It is easy to see that the condition (Θ) implies $\varkappa(-u\mu w_s) \le \varkappa_0 u^2\mu^2 w_s^2 \le \varkappa_0 u^2\mu^2 w_s$ for some $\varkappa_0 > 0$. This yields

$$\begin{aligned} \log I\!E_{\theta^*}\bigl\{2\mu L(W,\theta,\theta^*)\bigr\} &\le \mu N\log(1+u) - \mu N u + \sum_s \varkappa_0 u^2\mu^2 w_s \\ &= \mu N\bigl\{\log(1+u) - u + \varkappa_0\mu u^2\bigr\}. \end{aligned}$$

The condition (Θ) ensures that $u = u(\theta) = \theta^*/\theta - 1$ is bounded by some constant u^* for all $\theta \in \Theta$. The expression $\log(1+u) - u + \varkappa_0\mu u^2$ is negative for all $|u| \le u^*$ and

sufficiently small μ yielding (2.5).

Lemma 6.1 from Polzehl and Spokoiny (2006) implies that

$$\{N_t \mathcal{K}(\widetilde{\theta}_t, \theta^*) > \mathfrak{z}\} \subseteq \{N_t \mathcal{K}(\theta^-, \theta^*) > \mathfrak{z}\} \cup \{N_t \mathcal{K}(\theta^+, \theta^*) > \mathfrak{z}\}$$

for some fixed points θ^+, θ^- depending on $\mathfrak{z}$. This and (2.5) prove (2.6). □

The results of Theorem 2.3 can be similarly extended to the case of sub-Gaussian innovations.

Theorem 2.5. *Assume* (Θ) *and (2.4). Then for any* $r > 0$

$$I\!\!E_{\theta^*} \big| L(W_t, \widetilde{\theta}_t, \theta^*) \big|^r \equiv I\!\!E_{\theta^*} \big| N_t \mathcal{K}(\widetilde{\theta}_t, \theta^*) \big|^r \leq \mathfrak{r}_r \, \mu_0^{-r}.$$

Moreover, if $\mathfrak{z}_\alpha$ *satisfies* $2e^{-\mu_0 \mathfrak{z}_\alpha} \leq \alpha$, *then* $\mathcal{E}_{t,\alpha}$ *from (2.3) is an* α*-confidence set for the parameter* θ^*.

2.3 Spatial Stagewise Aggregation (SSA) Procedure

In this section we focus on the problem of adaptive (data-driven) estimation of the parameter θ_t. We assume that a finite set $\{\eta_k,\, k = 1, \ldots, K\}$ of values of the smoothing parameter is given. Every value η_k leads to the localizing weighting scheme $w_{st}^{(k)} = \eta_k^{t-s}$ for $s \leq t$ and to the local ML estimate $\widetilde{\theta}_t^{(k)}$:

$$\begin{aligned} N_k &= \sum_s w_{st}^{(k)} = \sum_{m=0}^{M_k} \eta_k^m, \\ \widetilde{\theta}_t^{(k)} &= N_k^{-1} \sum_s w_{st}^{(k)} Y_s = N_k^{-1} \sum_{m=0}^{M_k} \eta_k^m y_{t-m-1}. \end{aligned} \tag{2.7}$$

where $M_k = \log c / \log \eta_k - 1$ is the cutting point and guarantees that the weights after M_k are bounded by the prescribed value c, i.e. $\eta_k^{M_k+1} \leq c$. It is easy to see that the sum of weights $N_k = \sum_s w_{st}^{(k)}$ does not depend on t, thus we suppress the index t in the notation.

The corresponding fitted log-likelihood $L(W_t^{(k)}, \widetilde{\theta}_t^{(k)}, \theta)$ reads as

$$L(W_t^{(k)}, \widetilde{\theta}_t^{(k)}, \theta) \quad = \quad N_k \mathcal{K}(\widetilde{\theta}_t^{(k)}, \theta).$$

The local MLEs $\widetilde{\theta}_t^{(k)}$ will be referred to as "weak" estimates. Usually the parameter η_k runs over a wide range from values close to one to rather small values, so that at least one of them is "good" in the sense of estimation risk. However, the proper choice of the parameter η generally depends on the variability of the unknown random process θ_s. We aim to construct a data-driven estimate $\widehat{\theta}_t$ which performs nearly as good as the best one from this family.

In what follow we consider the *spatial stagewise aggregation* (SSA) method which originates from Belomestny and Spokoiny (2006). The underlying idea of the method is to aggregate all the weak estimates in form of a convex combination instead of choosing one of them. The procedure is sequential and starts with the estimate $\widetilde{\theta}_t^{(1)}$ having the largest variability, that is, we set $\widehat{\theta}_t^{(1)} = \widetilde{\theta}_t^{(1)}$. At every step $k \geq 2$ the new estimate $\widehat{\theta}_t^{(k)}$ is constructed by aggregating the next "weak" estimate $\widetilde{\theta}_t^{(k)}$ and the previously constructed estimate $\widehat{\theta}_t^{(k-1)}$. Following to Spokoiny (2006), the aggregation is done in terms of the canonical parameter υ which relates to the natural parameter θ by $\upsilon = -1/(2\theta)$. With $\widetilde{\upsilon}_t^{(k)} = -1/(2\widetilde{\theta}_t^{(k)})$ and $\widehat{\upsilon}_t^{(k-1)} = -1/(2\widehat{\theta}_t^{(k-1)})$

$$\begin{aligned} \widehat{\upsilon}_t^{(k)} &= \gamma_k \widetilde{\upsilon}_t^{(k)} + (1-\gamma_k)\widehat{\upsilon}_t^{(k-1)}, \\ \widehat{\theta}_t^{(k)} &= -1/(2\widehat{\upsilon}_t^{(k)}). \end{aligned}$$

Equivalently one can write

$$\widehat{\theta}_t^{(k)} = \left(\frac{\gamma_k}{\widetilde{\theta}_t^{(k)}} + \frac{1-\gamma_k}{\widehat{\theta}_t^{(k-1)}} \right)^{-1}$$

The mixing weights $\{\gamma_k\}$ are computed on the base of the fitted log-likelihood by checking that the previously aggregated estimate $\widehat{\theta}_t^{(k-1)}$ is in agreement with the next "weak"

estimate $\widetilde{\theta}_t^{(k)}$. The difference between these two estimates is measured by the quantity

$$\gamma_k = K_{\mathrm{ag}}\Big(\frac{1}{\mathfrak{z}_{k-1}} L(W_t^{(k)}, \widetilde{\theta}_t^{(k)}, \widehat{\theta}_t^{(k-1)})\Big) = K_{\mathrm{ag}}\Big(\frac{1}{\mathfrak{z}_{k-1}} N_k \mathcal{K}(\widetilde{\theta}_t^{(k)}, \widehat{\theta}_t^{(k-1)})\Big) \tag{2.8}$$

where $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ are the parameters of the procedure, see Section 2.4 for more details, and $K_{\mathrm{ag}}(\cdot)$ is the *aggregation kernel*. This kernel monotonously decreases on $\mathbb{R}_+$, is equal to one in a neighborhood of zero and vanishes outside the interval $[0, 1]$, so that the mixing coefficient γ_k is one if there is no essential difference between $\widetilde{\theta}_t^{(k)}$ and $\widehat{\theta}_t^{(k-1)}$ and zero, if the difference is significant. The significance level is measured by the "critical value" $\mathfrak{z}_{k-1}$. In the intermediate case, the mixing coefficient γ_k is between zero and one. The procedure terminates after step k if $\gamma_k = 0$ and we define in this case $\widehat{\theta}_t^{(m)} = \widehat{\theta}_t^{(k)} = \widehat{\theta}_t^{(k-1)}$ for all $m > k$. The formal definition reads as

1. Initialization: $\widehat{\theta}_t^{(1)} = \widetilde{\theta}_t^{(1)}$.

2. Loop: for $k \geq 2$

$$\widehat{\theta}_t^{(k)} = \left(\frac{\gamma_k}{\widetilde{\theta}_t^{(k)}} + \frac{1-\gamma_k}{\widehat{\theta}_t^{(k-1)}}\right)^{-1}$$

   where the aggregating parameter γ_k is computed as by (2.8). If $\gamma_k = 0$ then terminate by letting $\widehat{\theta}_t^{(k)} = \ldots = \widehat{\theta}_t^{(K)} = \widehat{\theta}_t^{(k-1)}$.

3. Final estimate: $\widehat{\theta}_t = \widehat{\theta}_t^{(K)}$.

In a special case of the SSA procedure with the binary γ_k equal to zero or one, every estimate $\widehat{\theta}_t^{(k)}$ and hence, the resulting estimate $\widehat{\theta}_t$ coincide with one of the "weak" estimates $\widetilde{\theta}_t^{(k)}$. This fact can easily be seen by induction arguments. Indeed, if $\gamma_k = 1$, then $\widehat{\theta}_t^{(k)} = \widetilde{\theta}_t^{(k)}$ and if $\gamma_k = 0$, then $\widehat{\theta}_t^{(k)} = \widehat{\theta}_t^{(k-1)}$. Therefore, in this situation the SSA method reduces to a kind of *local model selection procedure* (LMS). One limitation of the SSA compared to the alternative approach LMS is that it may magnify the bias through the summation, which will be illustrated in the later simulation study. On the meanwhile, the LMS may suffer

from a high variability since it merely concerns discrete and finite values of the smoothing parameter.

The next section discusses in details the problem of the parameter choice and critical values identification for the SSA procedure.

2.4 Parameter Choice and Implementation Details

To run the procedure, one has to specify the setup and fix the parameters of the procedure.

The considered setup mainly concerns the set of localizing schemes $W_t^{(k)} = \{w_{st}^{(k)}\}$ for $k = 1, \ldots, K$ yielding a set of "weak" estimates $\widetilde{\theta}_t^{(k)}$. Due to Theorem 2.4, variability of every $\widetilde{\theta}_t^{(k)}$ is characterized by the local sample size N_k (the sum of the corresponding weights $w_{st}^{(k)}$ over s) which increases with k. In this paper we focus on the exponentially decreasing localizing schemes, so that every $W_t^{(k)}$ is completely specified by the rate η_k and the cutting level c.

So, the aggregating procedure for a family of the "weak" ES estimates assumes that a growing sequence of values $\eta_1 < \eta_2 < \ldots < \eta_K$ is given in advance. This set leads to the sequence of localizing schemes $W_t^{(k)}$ with $w_{st}^{(k)} = \eta_k^{t-s}$ for $s \le t$ and $\eta_k^{t-s} > c$ otherwise $w_{st}^{(k)} = 0$. The set corresponding "weak" estimates $\widetilde{\theta}_t^{(k)}$ is defined by (2.7). The procedure applies to any such sequence for which the following condition is satisfied:

$(\boldsymbol{MD})$ for some $\mathfrak{u}_0, \mathfrak{u}$ with $0 < \mathfrak{u}_0 \le \mathfrak{u} < 1$, the values $N_1, \ldots, N_K$ satisfy

$$\mathfrak{u}_0 \le N_{k-1}/N_k \le \mathfrak{u}.$$

Here we present one example of constructing such a set $\{\eta_k\}$ which is used in our simulation study and application examples.

Example 2.1. [Set $\{\eta_k\}$] Given values $\eta_1 < 1$ and $a > 1$, define

$$\frac{N_{k+1}}{N_k} \approx \frac{1-\eta_k}{1-\eta_{k+1}} = a > 1. \tag{2.9}$$

The coefficient a controls the decreasing speed of the variations. The starting value η_1 should be sufficiently small to provide a reasonable degree of localization. Our default values are $a = 1.25$, $\eta_1 = 0.6$, and $c = 0.01$. The total number K of the considered localizing schemes is fixed by the condition that η_K does not exceed the prescribed value $\eta^* < 1$. One can expect a very minor influence of the mentioned parameters a, c on the performance of the procedure. This is confirmed by our simulation study in Section 4.

The definition of the mixing coefficients γ_k involves the "aggregation" kernel K_{ag}. Our theoretical study is done under the following assumptions on this kernel:

(K_{ag}) The aggregation kernel K_{ag} is monotonously decreasing for $u \in I\!R_+$, $K_{\mathrm{ag}}(0) = 1$, $K_{\mathrm{ag}}(1) = 0$. Moreover, there exists some $u_0 \in (0,1)$ such that $K_{\mathrm{ag}}(u) = 1$ for $u \le u_0$.

Our default choice is $K_{\mathrm{ag}}(u) = \{1 - (u - 1/6)_+\}_+$ so that $K_{\mathrm{ag}}(u) = 1$ for $u \le 1/6$.

Another choice is the uniform aggregation kernel $K_{\mathrm{ag}}(u) = \mathbf{1}(u \le 1)$. This choice leads the binary mixing coefficients γ_k and hence, to the local model selection procedure.

Next we discuss the most important question of choosing the critical values $\mathfrak{z}_k$.

The idea of selecting the critical values $\mathfrak{z}_k$ is to provide the prescribed performance of the procedure in the simple parametric situation with $\theta_t \equiv \theta^*$. In this situation, all the squared returns Y_t are i.i.d. and follow the equation $Y_t = \theta^* \varepsilon_t^2$. The corresponding joint distribution of all Y_t is denoted by $I\!P_{\theta^*}$. The approach assumes that the distribution of the innovations ε_s is known and it satisfies the condition (2.4). A natural candidate is the Gaussian distribution. However, we consider below in Section 3 the case when the ε_s's are obtained from the normal inverse Gaussian distribution, the heavy-tailed distribution, by some power transformation.

The way of selecting the critical values is based on the so called "propagation" condition and it can be formulated in a quite general setup. Recall that the SSA procedure is sequential and delivers after the step k the estimate $\widehat{\theta}_t^{(k)}$ which depends on the parameters $\mathfrak{z}_1, \ldots, \mathfrak{z}_{k-1}$. We now consider the performance of this procedure in the simple "paramet-

ric" situation of constant volatility $\theta_t \equiv \theta^*$. In this case the "ideal" or optimal choice among the first k estimates $\widetilde{\theta}_t^{(1)}, \ldots, \widetilde{\theta}_t^{(k)}$ is the one with the smallest variability, that is, the latest estimate $\widetilde{\theta}_t^{(k)}$ whose variability is measured by the quantity N_k, see Theorem 2.3. Our approach is similar to the one which is widely used in the hypothesis testing problem: to select the parameters (critical values) by providing the prescribed error under the "null", that is, in the parametric situation. The only difference is that in the estimation problem the risk is measured by another loss function. This consideration leads to the following condition: for all $\theta^* \in \Theta$ and all $k = 2, \ldots, K$

$$I\!E_{\theta^*} \bigl| L(W_t^{(k)}, \widetilde{\theta}_t^{(k)}, \widehat{\theta}_t^{(k)}) \bigr|^r \equiv I\!E_{\theta^*} \bigl| N_k \mathcal{K}\bigl(\widetilde{\theta}_t^{(k)}, \widehat{\theta}_t^{(k)}\bigr) \bigr|^r \le \frac{(k-1)\alpha \mathfrak{r}_r}{K-1} \,. \tag{2.10}$$

Here $\mathfrak{r}_r$ is from Theorem 2.3, and r and α are the fixed global parameters. The meaning of this condition is that the statistical difference between the adaptive estimate $\widehat{\theta}_t^{(k)}$ and the "oracle" estimate $\widetilde{\theta}_t^{(k)}$ after the first k steps measured by the left hand-side of (2.10) is bounded by a prescribed constant which linearly grows with k. As a particular case for $k = K$, the condition (2.10) implies for $\widehat{\theta}_t = \widehat{\theta}_t^{(K)}$

$$I\!E_{\theta^*} \bigl| N_K \mathcal{K}\bigl(\widetilde{\theta}_t^{(K)}, \widehat{\theta}_t\bigr) \bigr|^r \le \alpha \mathfrak{r}_r \,.$$

This means that the final adaptive estimate $\widehat{\theta}_t$ is sufficiently close to its non-adaptive counterpart $\widetilde{\theta}_t^{(K)}$.

The relation (2.10) gives us $K-1$ inequalities to fix $K-1$ parameters $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$. However, these parameters only implicitly enter in (2.10) and it is unclear, how they can be selected in a numerical algorithmic way. The next section describes a sequential procedure for selecting the parameters $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ one after another by Monte Carlo simulations.

The condition (2.10) is stated uniformly over θ^*. However, the following technical result allows to reduce the condition to any one particular θ^*, e.g. for $\theta^* = 1$.

Lemma 2.6. *Let the squared returns Y_t follow the parametric model with the constant*

volatility parameter θ^**, that is,* $Y_t = \theta^* \varepsilon_t^2$*. Then the distribution of the "test statistics"* $L(W_t^{(k)}, \widetilde{\theta}_t^{(k)}, \widehat{\theta}_t^{(k-1)}) = N_k \mathcal{K}(\widetilde{\theta}_t^{(k)}, \widehat{\theta}_t^{(k-1)})$ *under* $I\!P_{\theta^*}$ *is the same for all* $\theta^* > 0$*.*

Proof. Under $I\!P_{\theta^*}$ the squared returns Y_s fulfill $Y_t = \theta^* \varepsilon_t^2$ and for every k, the estimate $\widetilde{\theta}_t^{(k)}$ can be represented as

$$\widetilde{\theta}_t^{(k)} = N_k^{-1} \sum_s Y_s w_{st}^{(k)} = \theta^* N_k^{-1} \sum_s \varepsilon_s^2 w_{st}^{(k)},$$

so that $\widetilde{\theta}_t^{(k)}$ is θ^* times the estimate computed for $\theta^* = 1$. The same applies by simple induction argument to the aggregated estimate $\widehat{\theta}_t^{(k-1)}$. It remains to note that the Kullback-Leibler divergence $\mathcal{K}(\widetilde{\theta}_t^{(k)}, \widehat{\theta}_t^{(k-1)})$ is a function of the ratio $\widetilde{\theta}_t^{(k)} / \widehat{\theta}_t^{(k-1)}$, in which θ^* cancels. □

The condition (2.10) involves two more "hyperparameters" r and α. The parameter r in (2.10) specifies the selected loss function. To provide a stable performance of the method and to minimize the Monte Carlo error we suggest the choice $r = 1/2$. The parameter α is similar to the test level parameter, and, exactly as in the testing setup, its choice depends upon the subjective requirements on the procedure. Small values of α mean that we put more attention to the performance of the methods in the time homogeneous (parametric) situation and such a choice leads to a rather conservative procedure with relatively large critical values. Increasing α would result in a decrease of the critical values and an increase of the sensitivity of the method to the changes in the underlying parameter θ_t at cost of some loss of stability in the time homogeneous situation. For the most of applications, a reasonable range of values α is between 0.2 and 1. Section 4 presents a small simulation study which demonstrates the dependence of the critical values on the parameters r and α.

It is important to note that the "hyperparameters" r and α are *global* and their proper choice depends on the particular application while the estimation procedure is *local* and it constructs the estimate $\widehat{\theta}_t$ separately at each point. The parameters r and α can be selected in a data driven way by fixing some objective function, e.g., by minimizing the

forecasting error, see Section 2.5, however, we prefer to keep this choice free for the user.

Below we present one way of selecting the critical values $\mathfrak{z}_k$ using Monte Carlo simulations from the parametric model successively, starting from $k = 1$. To specify the contribution of $\mathfrak{z}_1$ in the final risk of the method, we set all the remaining values $\mathfrak{z}_2, \ldots, \mathfrak{z}_{K-1}$ equal to infinity: $\mathfrak{z}_2 = \ldots = \mathfrak{z}_{K-1} = \infty$. Now, for every particular $\mathfrak{z}_1$, the whole set of critical values $\mathfrak{z}_k$ is fixed and can run the procedure leading to the estimates $\widehat{\theta}_t^{(k)}(\mathfrak{z}_1)$ for $k = 2, \ldots, K$. The value $\mathfrak{z}_1$ is selected as the minimal one for which

$$I\!\!E_{\theta^*} \big| N_k \mathcal{K}\big(\widetilde{\theta}_t^{(k)}, \widehat{\theta}_t^{(k)}(\mathfrak{z}_1)\big)\big|^r \le \frac{\alpha \mathfrak{r}_r}{K-1}, \qquad k = 2, \ldots, K. \tag{2.11}$$

Such a value exists because the choice $\mathfrak{z}_1 = \infty$ leads to $\widehat{\theta}_t^{(k)}(\mathfrak{z}_1) = \widetilde{\theta}_t^{(k)}$ for all k. Notice that the rule of "early stop" (the procedure terminates and sets $\widehat{\theta}_t^{(k)} = \ldots, \widehat{\theta}_t^{(K)} = \widehat{\theta}_t^{(k-1)}$ if $\gamma_k = 0$) is important here, otherwise $\mathfrak{z}_k = \infty$ leads to $\gamma_k = 1$ and $\widehat{\theta}_t^{(k)} = \widetilde{\theta}_t^{(k)}$ for all $k \ge 2$.

Next, with $\mathfrak{z}_1$ fixed in this way, we select $\mathfrak{z}_2$. The method is similar: set $\mathfrak{z}_3 = \ldots = \mathfrak{z}_{K-1} = \infty$ and play with $\mathfrak{z}_2$. Every particular value of $\mathfrak{z}_2$ determines the whole set of critical values $\mathfrak{z}_1, \mathfrak{z}_2, \infty, \ldots, \infty$. The procedure with such critical values results in the estimates $\widehat{\theta}_t^{(k)}(\mathfrak{z}_1, \mathfrak{z}_2)$ for $k = 3, \ldots, K$. We select $\mathfrak{z}_2$ as the minimal value which fulfills

$$I\!\!E_{\theta^*} \big| N_k \mathcal{K}\big(\widetilde{\theta}_t^{(k)}, \widehat{\theta}_t^{(k)}(\mathfrak{z}_1, \mathfrak{z}_2)\big)\big|^r \le \frac{2\alpha \mathfrak{r}_r}{K-1}, \qquad k = 3, \ldots, K. \tag{2.12}$$

Such a value exists because the choice $\mathfrak{z}_2 = \infty$ provides a stronger inequality (2.11). We continue this way for all $k < K$. Suppose $\mathfrak{z}_1, \ldots, \mathfrak{z}_{k-1}$ have been already fixed. We set $\mathfrak{z}_{k+1} = \ldots = \mathfrak{z}_{K-1} = \infty$ and play with $\mathfrak{z}_k$. Every particular choice of $\mathfrak{z}_k$ leads to the estimates $\widehat{\theta}^{(m)}(\mathfrak{z}_1, \ldots, \mathfrak{z}_k)$ for $m = k+1, \ldots, K$ coming out of the procedure with the parameters $\mathfrak{z}_1, \ldots, \mathfrak{z}_k, \infty, \ldots, \infty$. We select $\mathfrak{z}_k$ as the minimal value which fulfills

$$I\!\!E_{\theta^*} \big| N_l \mathcal{K}\big(\widetilde{\theta}_t^{(l)}, \widehat{\theta}_t^{(l)}(\mathfrak{z}_1, \ldots, \mathfrak{z}_k)\big)\big|^r \le \frac{k\alpha \mathfrak{r}_r}{K-1}, \qquad l = k+1, \ldots, K. \tag{2.13}$$

By simple induction arguments one can see that such a value exists and that the final procedure with the such defined parameters fulfills (2.10).

Note that the proposed Monte Carlo procedure heavily relies on the joint distribution of the estimates $\widetilde{\theta}_t^{(1)}, \ldots, \widetilde{\theta}_t^{(K)}$ under the parametric measure $I\!P_{\theta^*}$. In particular, it automatically accounts for the correlation between the estimates $\widetilde{\theta}_t^{(k)}$.

It is also worth mentioning that the numerical complexity of the proposed procedure is not very high. It suffices to generate once M samples from $I\!P_{\theta^*}$ and compute and store the estimates $\widetilde{\theta}_t^{(k,m)}$ for every realization, $m = 1, \ldots, M$ and $k = 1, \ldots, K$. The SSA procedure operates with the estimates $\widetilde{\theta}_t^{(k)}$ and there is no need to keep the samples themselves. Now, with the fixed set of parameters $\mathfrak{z}_k$, computing the estimates $\widehat{\theta}_t^{(k)}$ requires only the finite number of operations proportional to K. One can roughly bound the total complexity of the Monte Carlo study by CMK^2 for some fixed constant C.

Below we present some numerical results for the proposed procedures for selecting the critical values. We first specify our setup. Then we illustrate how the resulting critical values depend on the other "hyperparameters" like r and α.

The parameters $\{\eta_k\}$ defining the weighting scheme $W_t^{(k)}$ are fixed by setting the values c, a, η_1 .We select $c = 0.01$, $a = 1.25$ and $\eta_1 = 0.6$. We also restrict the largest η_K to be smaller than $\eta^* = 0.985$.

To understand the impact of using a continuous aggregation kernel, we also consider the LMS procedure which comes out of the algorithm for the uniform aggregation kernel $K_{\mathrm{ag}}(u) = \mathbf{1}(u \le 1)$.

For the above defined family of localizing schemes, the critical values $\mathfrak{z}_k$ of the SSA and LMS procedures are fixed by the method from Section 2.4. The coefficients $\{\eta_k\}$, the corresponding local window width M_k and the resulting critical values are reported in Table 1. An interesting observation is that the first critical value $\mathfrak{z}_1$ is relatively small compared with the second and third values. A possible explanation is that the first two localizing schemes $W_t^{(1)}$ and $W_t^{(2)}$ are close to each other leading to a strong correlation between the estimates $\widetilde{\theta}_t^{(1)}$ and $\widetilde{\theta}_t^{(2)}$. The parameter $\mathfrak{z}_1$ is responsible just for the risk associated

k	η_k	M_k	N_k	$\mathfrak{z}_k$ (SSA)	$\mathfrak{z}_k$ (LMS)
1	0.600	9	2.485	0.192	0.192
2	0.680	11	3.095	0.548	0.141
3	0.744	15	3.872	0.587	0.091
4	0.795	20	4.843	0.220	0.065
5	0.836	25	6.045	0.134	0.053
6	0.869	32	7.555	0.145	0.043
7	0.895	41	9.446	0.117	0.035
8	0.916	52	11.806	0.087	0.030
9	0.933	66	14.759	0.076	0.025
10	0.946	83	18.446	0.065	0.020
11	0.957	104	23.051	0.050	0.016
12	0.966	131	28.816	0.037	0.012
13	0.973	165	36.024	0.022	0.007
14	0.978	207	45.029	0.015	0.001
15	0.982	259	56.280		

Table 1: Critical values of the SSA and LMS methods w.r.t. the default choice: $c = 0.01$, $a = 1.25$, $\eta_1 = 0.6$, $r = 0.5$ and $\alpha = 1$.

with the discrepancy $N_2\mathcal{K}(\widetilde{\theta}_t^{(2)}, \widetilde{\theta}_t^{(1)})$ which can be bounded with a high probability by a relatively small value $\mathfrak{z}_1$.

Next few numerical results illustrate the influence of the parameters r, α, a, and c on the critical values $\mathfrak{z}_k$.

The sequences of the critical values $\mathfrak{z}_k$ for the SSA procedure for different combinations of r, α, a, and c are detailed in Table 2. We start with the default choice and then slightly vary one parameter fixing the others to the default.

The numerical results can be summarized as follows:

- r (Default choice: $r = 0.5$): The parameter r is the power of the loss function. Our numerical results confirm that the growth of the power loss results in an increase of the critical values and hence, in a more conservative and less sensitive procedure, see Section 2.4.

- α (Default choice: $\alpha = 1$): As already mentioned, the parameter α has the same meaning as the test level. Correspondingly, a decrease of α results in an increase of

		r			α			c	
k	default	0.3	0.7	1.0	0.5	0.7	1.5	0.005	0.02
1	0.19	0.12	0.29	0.57	0.24	0.22	0.17	0.19	0.34
2	0.54	0.28	0.92	1.54	0.69	0.60	0.43	0.50	0.60
3	0.58	0.23	1.05	1.69	0.93	0.75	0.42	0.56	0.51
4	0.22	0.10	0.41	0.76	0.41	0.28	0.15	0.20	0.19
5	0.13	0.07	0.17	0.19	0.15	0.15	0.11	0.13	0.17
6	0.14	0.07	0.24	0.40	0.21	0.17	0.10	0.14	0.16
7	0.11	0.06	0.20	0.54	0.20	0.15	0.08	0.11	0.11
8	0.08	0.05	0.12	0.20	0.13	0.11	0.06	0.08	0.09
9	0.07	0.04	0.10	0.12	0.11	0.09	0.05	0.07	0.08
10	0.06	0.04	0.09	0.14	0.10	0.08	0.04	0.06	0.06
11	0.05	0.03	0.06	0.10	0.09	0.07	0.03	0.04	0.05
12	0.03	0.02	0.04	0.05	0.06	0.05	0.01	0.03	0.03
13	0.02	0.01	0.02	0.02	0.05	0.03	0.00	0.02	0.02
14	0.01	0.02	0.00	0.00	0.06	0.03	0.00	0.01	0.01
$\mathfrak{r}_r$	0.40	0.54	0.32	0.25	0.40	0.40	0.40	0.40	0.40

Table 2: Sensitivity analysis: comparison of the SSA critical values $\mathfrak{z}_k$.

$\mathfrak{z}_k$ and hence, in a less sensitive procedure.

- a (Default choice: $a = 1.25$): This parameter specifies how dense is the set of possible values η_k. The values of a close to one result in a rather dense set which becomes more and more rare with the increase of a. Therefore, for smaller a-values we have more estimates to select between. This can be helpful for improving the accuracy of approximation and thus, for reducing the bias of estimation. This improvement is however, at cost of some loss of sensitivity, because the growth of K requires more conditions to be checked. Note also that our theoretical upper bound for the critical values $\mathfrak{z}_k$ from Theorem 2.7 presented later linearly increases with K. From the other side, the use of a relatively small a results in a strong correlation between the estimates $\widetilde{\theta}_t^{(k)}$ which leads to a decrease of the critical values $\mathfrak{z}_k$. Figure 2 shows the critical values $\mathfrak{z}_k$ for the default choice ($K = 15$), $a = 1.5$ ($K = 9$) and $a = 1.1$ ($K = 34$).

- c (Default choice: $c = 0.01$): The parameter c specifies the cutting point of the

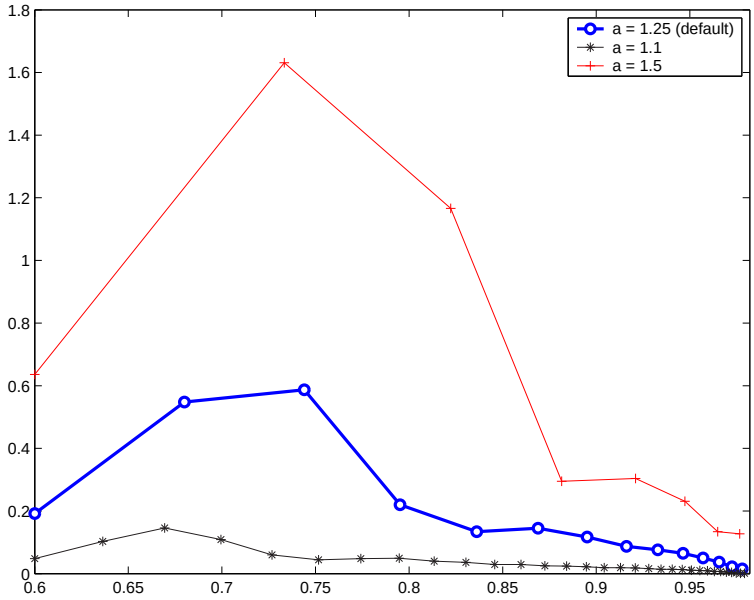

Figure 2: Sequences of critical values $\mathfrak{z}_k$ for the default choice $a = 1.25\,(\,K = 15\,)$, $a = 1.5$ $(\,K = 9\,)$ and $a = 1.1$ $(\,K = 34\,)$ w.r.t. the smoothing parameter η_k for $k = 1, \ldots, K-1$.

exponential smoothing window. As one can expect, this value has only minor influence on the critical values and on the whole procedure. This is in agreement with our numerical results.

2.5 Parameter Tuning by Minimizing the Forecast Errors

The proposed procedure is *local* in the sense that the the adaptation (model selection or aggregation) is performed at every time instant t separately. However, the procedure involves some global parameters like the loss power r or the level α. Their choice can be done in a data-driven way by minimizing the global forecasting error as suggested in Cheng et al. (2003). The estimated value $\widehat{\theta}_t$ can be viewed as a forecast for the volatility for a short forecasting horizon h. So, a good performance of the method means a relatively small forecasting error which is measured as

$$\text{mean } h\text{-step-ahead forecasting errors:} \qquad \sum_{t=t_0}^{T} \frac{1}{h} \sum_{m=0}^{h-1} \bigl| y_{t+m} - \widehat{\theta}_t \bigr|^p$$

for some power $p > 0$.

2.6 Some Theoretical Properties of the SSA Estimate

Belomestny and Spokoiny (2006) claimed some "oracle" property of the SSA estimate $\widehat{\theta}_t$. However, the results presented there only apply to the local maximum likelihood estimates obtained from independent observations. Here we show that the similar results continue to apply in the sub-Gaussian case and in the time series framework.

The first result gives an upper bound for the critical values $\mathfrak{z}_k$.

Theorem 2.7 (Belomestny and Spokoiny (2006, Theorem 5.1)). *Let the innovations ε_t be i.i.d. standard normal. Assume (MD) and (K_{ag}). There are three constants a_0, a_1 and a_2 depending on $\mathfrak{u}_0$, $\mathfrak{u}$ and u_0 only such that the choice*

$$\mathfrak{z}_k = a_0 + a_1 \log \alpha^{-1} + a_2 r \log N_k$$

ensures (2.10) for all $k \le K$.

The result and the proof extend in a straightforward way to the case of the sub-Gaussian innovations using the result of Theorem 2.4. In that case, the constants a_0, a_1, and a_2 also depend on μ_0 shown in Theorem 2.4.

The construction of the procedure ensures some risk bound for the adaptive estimate $\widehat{\theta}$ in the time homogeneous situation, see (2.10). It is natural to expect that a similar behavior is valid in the situation when the time varying parameter θ_t does not significantly deviates from some constant value θ. Here we quantify this property and show how the deviation from the parametric time homogeneous situation can be measured.

Denote by $I_t^{(k)}$ the support of the kth weighting scheme corresponding to the memory parameter η_k: $I_t^{(k)} = [t - M_k, t]$, $k = 1, \ldots, K$. Define for each k and θ

$$\Delta_t^{(k)}(\theta) = \sum_{s \in I_t^{(k)}} I\!K\big(P_{\theta_s}, P_\theta\big), \tag{2.14}$$

where $I\!K\bigl(P_{\theta_s}, P_\theta\bigr)$ means the Kullback-Leibler distance between two distributions of Y_s with the parameter values θ_s and θ. In the case of Gaussian innovations, $I\!K\bigl(P_{\theta_s}, P_\theta\bigr) = \mathcal{K}(\theta_s, \theta)$. The value $\Delta_t^{(k)}(\theta)$ can be considered as a distance from the time varying model at hand to the parametric model with the constant parameter θ on the interval $I_t^{(k)}$.

Note that the volatility θ_s is in general a random process. Thus, the value $\Delta_t^{(k)}(\theta)$ is random as well. Our *small modeling bias* condition means that there is a number k^* such that the modeling bias $\Delta_t^{(k)}(\theta)$ is small with a high probability for some θ and all $k \le k^*$. Consider the corresponding estimate $\widehat{\theta}_t^{(k^*)}$ obtained after the first k^* steps of the algorithm. The next "propagation" result claims that the behavior of the procedure under the small modeling bias condition is essentially the same as in the pure parametric situation.

Theorem 2.8. *Assume* (Θ), (MD), *and (2.4). Let* θ *and* k^* *be such that*

$$\max_{k \le k^*} I\!E \Delta_t^{(k)}(\theta) \le \Delta \tag{2.15}$$

for some $\Delta \ge 0$. *Then for any* $r > 0$

$$\begin{aligned} I\!E \log\biggl(1 + \frac{N_{k^*}^r \mathcal{K}^r\bigl(\widetilde{\theta}_t^{(k^*)}, \widehat{\theta}_t^{(k^*)}\bigr)}{\alpha \mathfrak{R}_r}\biggr) &\le 1 + \Delta, \\ I\!E \log\biggl(1 + \frac{N_{k^*}^r \mathcal{K}^r\bigl(\widetilde{\theta}_t^{(k^*)}, \theta\bigr)}{\mathfrak{R}_r}\biggr) &\le 1 + \Delta \end{aligned}$$

where $\mathfrak{R}_r = \mathfrak{r}_r$ *in the case of Gaussian innovations and* $\mathfrak{R}_r = \mu_0^{-r}\mathfrak{r}_r$ *in the case of sub-Gaussian innovations with the constant* μ_0 *from Theorem 2.4.*

Proof. The proof is based on the following general result.

Lemma 2.9. *Let* $I\!P$ *and* $I\!P_0$ *be two measures such that the Kullback-Leibler divergence* $I\!E \log(dI\!P/dI\!P_0)$, *satisfies*

$$I\!E \log(dI\!P/dI\!P_0) \le \Delta < \infty.$$

Then for any random variable ζ *with* $I\!E_0\zeta < \infty$

$$I\!E \log\bigl(1+\zeta\bigr) \le \Delta + I\!E_0\zeta.$$

Proof. By simple algebra one can check that for any fixed y the maximum of the function $f(x) = xy - x\log x + x$ is attained at $x = e^y$ leading to the inequality $xy \le x\log x - x + e^y$. Using this inequality and the representation $I\!E\log\bigl(1+\zeta\bigr) = I\!E_0\bigl\{Z\log\bigl(1+\zeta\bigr)\bigr\}$ with $Z = dI\!P/dI\!P_0$ we obtain

$$\begin{aligned} I\!E\log\bigl(1+\zeta\bigr) &= I\!E_0\bigl\{Z\log\bigl(1+\zeta\bigr)\bigr\} \\ &\le I\!E_0\bigl(Z\log Z - Z\bigr) + I\!E_0(1+\zeta) \\ &= I\!E_0\bigl(Z\log Z\bigr) + I\!E_0\zeta - I\!E_0 Z + 1. \end{aligned}$$

It remains to note that $I\!E_0 Z = 1$ and $I\!E_0\bigl(Z\log Z\bigr) = I\!E\log Z$. □

The first assertion of the theorem is just a combination of this result and the condition (2.10). The second follows in a similar way from Theorem 2.3 for the case of Gaussian innovations and from Theorem 2.4 in the sub-Gaussian case. □

Due to the "propagation" result, the procedure performs well as long as the "small modeling bias" condition $\Delta_k(\theta) \le \Delta$ is fulfilled. To establish the accurate result for the final estimate $\widehat{\theta}$, we have to check that the aggregated estimate $\widehat{\theta}_k$ does not vary much at the steps "after propagation" when the divergence $\Delta_k(\theta)$ from the parametric model becomes large.

Theorem 2.10 (Belomestny and Spokoiny (2006), Theorem 5.3). *It holds for every* $k \le K$

$$N_k \mathcal{K}\bigl(\widehat{\theta}_t^{(k)}, \widehat{\theta}_t^{(k-1)}\bigr) \le \mathfrak{z}_k. \tag{2.16}$$

Moreover, under (MD), *it holds for every* k' *with* $k < k' \le K$

$$N_k \mathcal{K}\bigl(\widehat{\theta}_t^{(k')}, \widehat{\theta}_t^{(k)}\bigr) \le \mathfrak{a}^2 c_{\mathfrak{u}}^2 \bar{\mathfrak{z}}_k \tag{2.17}$$

where $c_{\mathfrak{u}} = (\mathfrak{u}^{-1/2} - 1)^{-1}$, $\mathfrak{a}$ *is a constant depending on* Θ *only, and* $\bar{\mathfrak{z}}_k = \max_{l \ge k} \mathfrak{z}_l$.

Combination of the "propagation" and "stability" statements implies the main result concerning the properties of the adaptive estimate $\widehat{\theta}_t$.

The result claims again the "oracle" accuracy $N_{k^*}^{-1/2}$ for $\widehat{\theta}$ up to the log factor $\mathfrak{z}_{k^*}$. We state the result for $r = 1/2$ only. An extension to an arbitrary $r > 0$ is obvious.

Theorem 2.11 ("Oracle" property). *Assume* (Θ), (MD), *(2.4), and let* $\mathbb{E}\Delta_t^{(k)} \le \Delta$ *for some* k^*, θ *and* Δ. *Then*

$$\mathbb{E} \log\biggl(1 + \frac{N_{k^*}^{1/2} \mathcal{K}^{1/2}\bigl(\widehat{\theta}_t, \theta\bigr)}{\mathfrak{a}\mathfrak{R}_{1/2}}\biggr) \le \log\Bigl(1 + c_{\mathfrak{u}} \mathfrak{R}_{1/2}^{-1} \sqrt{\bar{\mathfrak{z}}_{k^*}}\Bigr) + \Delta + \alpha + 1$$

where $c_{\mathfrak{u}}$ *is the constant from Theorem 2.10 and* $\mathfrak{R}_{1/2}$ *from Theorem 2.8.*

Remark 2.2. Before proving the theorem, we briefly comment on the result claimed. By Theorem 2.8, the "oracle" estimate $\widetilde{\theta}_t^{(k^*)}$ ensures that the estimation loss $\mathcal{K}^{1/2}\bigl(\widetilde{\theta}_t^{(k^*)}, \theta\bigr)$ is stochastically bounded by Const.$/N_{k^*}^{1/2}$ where Const. is a constant depending on Δ from the condition (2.15). The "oracle" result claims the same property for the adaptive estimate $\widehat{\theta}_t$ but the loss $\mathcal{K}^{1/2}(\widehat{\theta}_t, \theta)$ is now bounded by Const.$\sqrt{\bar{\mathfrak{z}}_{k^*}/N_{k^*}}$. By Theorem 2.7, the parameter $\bar{\mathfrak{z}}_{k^*}$ is at most logarithmic in the sample size. Hence, the accuracy of adaptive estimation is the same in order as for the "oracle" up to a logarithmic factor which can be viewed as "payment for adaptation". Belomestny and Spokoiny (2006) argued that the "oracle" result implies rate optimality of the adaptive estimate $\widehat{\theta}$ and that the log-factor $\bar{\mathfrak{z}}_{k^*}$ cannot be removed or improved.

Proof. Similarly to the proof of Theorem 2.10,

$$\begin{aligned}\mathcal{K}^{1/2}\big(\widehat{\theta}_t,\theta\big) &\leq \mathfrak{a}\mathcal{K}^{1/2}\big(\widetilde{\theta}_t^{(k^*)},\theta\big)+\mathfrak{a}\mathcal{K}\big(\widetilde{\theta}_t^{(k^*)},\widehat{\theta}_t^{(k^*)}\big)+\mathfrak{a}\sum_{l=k^*+1}^{\widehat{k}}\mathcal{K}^{1/2}\big(\widehat{\theta}_t^{(l)},\widehat{\theta}_t^{(l-1)}\big)\\ &\leq \mathfrak{a}\mathcal{K}^{1/2}\big(\widetilde{\theta}_t^{(k^*)},\theta\big)+\mathfrak{a}\mathcal{K}^{1/2}\big(\widetilde{\theta}_t^{(k^*)},\widehat{\theta}_t^{(k^*)}\big)+\mathfrak{a}c_{\mathrm{u}}\sqrt{\bar{\mathfrak{z}}_{k^*}/N_{k^*}}.\end{aligned}$$

This, the elementary inequality $\log(1+a+b)\leq\log(1+a)+\log(1+b)$ for $a,b\geq 0$ implies similarly to Theorem 2.8 that

$$\begin{aligned}&I\!E\log\Big(1+\frac{N_{k^*}^{1/2}\mathcal{K}^{1/2}\big(\widehat{\theta}_t,\theta\big)}{\mathfrak{a}\mathfrak{R}_{1/2}}\Big)\\ &\quad\leq\log\Big(1+\frac{c_{\mathrm{u}}\sqrt{\bar{\mathfrak{z}}_{k^*}}}{\mathfrak{R}_{1/2}}\Big)+I\!E\log\Big(1+\frac{N_{k^*}^{1/2}\mathcal{K}^{1/2}\big(\widetilde{\theta}_t^{(k^*)},\widehat{\theta}_t^{(k^*)}\big)+N_{k^*}^{1/2}\mathcal{K}^{1/2}\big(\widetilde{\theta}_t^{(k^*)},\theta\big)}{\mathfrak{R}_{1/2}}\Big)\\ &\quad\leq\log\Big(1+\frac{c_{\mathrm{u}}\sqrt{\bar{\mathfrak{z}}_{k^*}}}{\mathfrak{R}_{1/2}}\Big)+\Delta+\alpha+1\end{aligned}$$

as required. □

3 Accounting for Heavy Tails

The proposed local exponential smoothing methods and the calculation of the critical values are valid in the Gaussian framework. They can be easily extended to the sub-Gaussian framework considered in Section 2.2. Financial time series however often indicates a heavily tailed behaviour which goes far beyond the sub-Gaussian case. In this section, we extend the methods in the normal inverse Gaussian (NIG) distributional framework which can well describe the heavy-tailed behavior of the real series. The density is of the form:

$$f_{\mathrm{NIG}}(\varepsilon;\phi,\beta,\delta,\mu)=\frac{\phi\delta}{\pi}\frac{K_1\big(\phi\sqrt{\delta^2+(\varepsilon-\mu)^2}\big)}{\sqrt{\delta^2+(\varepsilon-\mu)^2}}\exp\{\delta\sqrt{\phi^2-\beta^2}+\beta(\varepsilon-\mu)\},$$

where the distributional parameters fulfill conditions: $\mu \in \mathbb{R}$, $\delta > 0$ and $|\beta| \leq \phi$, and $K_\lambda(\cdot)$ is the modified Bessel function of the third kind which is of the form:

$$K_\lambda(y) = \frac{1}{2} \int_0^\infty y^{\lambda-1} \exp\{-\frac{y}{2}(y + y^{-1})\}\, dy.$$

We refer to Prause (1999) for a detailed description of the NIG distribution.

One can easily see that the exponential moment $I\!E\{\exp(\lambda\varepsilon_t^2)\}$ of the squared NIG innovations ε_t^2 does not exist. Hence, the results of Section 2.2 do not apply to NIG innovations. Apart the theoretical reasons, the quasi MLE $\widetilde{\theta}_t$ computed from the squared returns Y_t with the heavy-tailed innovations indicates high variability and is very volatile. To ensure a robust and stable risk management, we suggest to replace the squared returns Y_t by their p-power. The choice of $0 \leq p < 1/2$ ensures that the resulting "observations" $y_{t,p} = Y_t^p$ have exponential moments, see Chen et al. (2005). This enables us to apply the proposed SSA procedure to the transformed data $y_{t,p}$ to estimate the parameter ϑ_t. One easily gets

$$I\!E\{y_{t,p} \mid \mathcal{F}_{t-1}\} \quad = \quad I\!E\{Y_t^p \mid \mathcal{F}_{t-1}\} = \theta_t^p I\!E|\varepsilon_t|^{2p} = \theta_t^p C_p = \vartheta_{t,p} \tag{3.1}$$

where $C_p = I\!E|\varepsilon_t|^{2p}$ is a constant and relies on p and the distribution of the innovations ε_t which is assumed to be NIG. Note that the equation (3.1) can be rewritten as

$$y_{t,p} = \vartheta_{t,p}\varepsilon_{t,p}^2$$

where the "new" standardized squared innovations

$$\varepsilon_{t,p}^2 = y_{t,p}/\vartheta_{t,p} = Y_t^p/(C_p\theta_t^p)$$

satisfy $I\!E\{\varepsilon_{t,p}^2 \mid \mathcal{F}_{t-1}\} = 1$.

An important question for this application is the choice of parameters of the method,

especially of the critical values $\mathfrak{z}_k$. The formal application of the approach of Section 2.4 requires to use the underlying NIG distribution of the innovations ε_t for the Monte Carlo simulations. This means that one has to first simulate the NIG data Y_t under the time homogeneous situation $Y_t = \theta^* \varepsilon_t^2$ with NIG ε_t and then compute the transformed data $y_{t,p}$ for the calculation of "weak" estimates $\widetilde{\vartheta}_{t,p}^{(k)}$. This approach would require the exact knowledge of the parameters of the NIG distribution of ε_t which is difficult to expect in real life situation. On the other hand, the use of power transformation with an appropriate choice of p makes the distribution of the "new" innovations $\varepsilon_{t,p}$ close to the Gaussian case. This suggests to apply the critical values $\mathfrak{z}_k$ computed for the Gaussian case. Below in Section 4 we calculate critical values $\mathfrak{z}_k$ given the true distributional parameters of the NIG innovations, which shows that the use of Gaussian $\varepsilon_{t,p}$ in the Monte Carlo simulations and the values of p around $1/2$ works well and delivers almost the same results as if the true NIG distribution for the ε_t's would be utilized.

The adaptive procedure delivers the estimate $\widehat{\vartheta}_{t,p}$ of the "new" variable $\vartheta_{t,p}$. To get the estimate of the original variance θ_t from the relation (3.1), we need to know the constant C_p which depends upon the parameters of the NIG distribution. We suggest two ways to fix this constant. One is based on the fact that the standardized innovations $\varepsilon_t^2 = Y_t/\theta_t$ should satisfy $I\!E\varepsilon_t^2 = 1$. The estimates $\widehat{\theta}_t = \widehat{\vartheta}_{t,p}^{1/p}/C_p^{1/p}$ lead to the estimated squared innovations $\widetilde{\varepsilon}_t^2 = Y_t/\widehat{\theta}_t = C_p^{1/p} Y_t/\widehat{\vartheta}_{t,p}^{1/p}$, so that an estimate of C_p can be obtained from the equation

$$n^{-1} C_p^{1/p} \sum_{t=t_0}^{t_1} \frac{Y_t}{\widehat{\vartheta}_{t,p}^{1/p}} = 1, \tag{3.2}$$

where $n = t_1 - t_0 + 1$ means the number of observations based on which the estimation is done. A small problem with this approach is that the presented sum of $Y_t/\widehat{\vartheta}_{t,p}^{1/p}$ is quite sensitive to extreme values of Y_t and even one or two outliers can dramatically destroy the resulting estimate.

The other method of fixing the constant C_p is based on the proposal of Section 2.5

to minimize the mean of forecasting error. Namely, we define the value C_p in a way to minimize

$$\sum_{t=t_0}^{t_1} \frac{1}{h} \sum_{m=0}^{h-1} \bigl| Y_{t+m} - \widehat{\theta}_t \bigr|^p = \sum_{t=t_0}^{t_1} \frac{1}{h} \sum_{m=0}^{h-1} \bigl| Y_{t+m} - \widehat{\vartheta}_{t,p}^{1/p} / C_p^{1/p} \bigr|^p .$$

After the constant C_p is estimated one can use the estimated returns $\widetilde{\varepsilon}_t$ for fixing the NIG parameters which will be used for our risk evaluation.

The adaptive procedure for the NIG innovations is summarized as:

1. Do power transformation to the squared returns Y_t: $Y_{t,p} = Y_t^p$.
2. Compute the estimate $\widehat{\vartheta}_{t,p}$ of the parameter $\vartheta_{t,p}$ from $Y_{t,p}$ applying the critical values $\mathfrak{z}_k$ obtained for the Gaussian case.
3. Estimate the value C_p from the equation (3.2).
4. Compute the estimates $\widehat{\theta}_t = (\widehat{\vartheta}_{t,p}/C_p)^{1/p}$ and identify the NIG distributional parameters from $\widetilde{\varepsilon}_t = R_t \widehat{\theta}_t^{-1/2}$.
5. (Optional) Calculate critical values $\mathfrak{z}_k$ with the identified NIG parameters using Monte Carlo simulation. Repeat the above procedure to estimate θ_t.

All the theoretical results from Section 2.6 applies to the such constructed estimate $\widehat{\vartheta}_{t,p}$ of the parameter $\vartheta_{t,p}$ if $p < 1/2$ is taken. This automatically yields the "oracle" accuracy for the back transformed estimate $\widehat{\theta}_t$ of the original volatility θ_t. For reference convenience, we present the "oracle" result. Below P_ϑ means the distribution of $Y_{t,p} = \vartheta |\varepsilon_t|^{2p}$ with NIG ε_t. Note that neither the procedure nor the result does not assume that the parameter of the NIG distribution are known.

Theorem 3.1 ("Oracle" property for NIG innovations). *Let the innovations ε_t be NIG and $p < 1/2$. Assume (Θ), (MD), and let, for some k^*, ϑ and Δ,*

$$I\!\!E \sum_{t \in I} I\!\!K\bigl(P_{\vartheta_{t,p}}, P_\vartheta\bigr) \le \Delta .$$

Then

$$\mathbb{E}\log\Big(1+\frac{N_{k^*}^{1/2}\mathcal{K}^{1/2}\big(\widehat{\vartheta}_{t,p},\vartheta\big)}{\mathfrak{a}\mathfrak{R}_{1/2}}\Big)\le\log\Big(1+c_{\mathrm{u}}\mathfrak{R}_{1/2}^{-1}\sqrt{\bar{\mathfrak{z}}_{k^*}}\Big)+\Delta+\alpha+1$$

where c_{u} *is the constant from Theorem 2.10 and* $\mathfrak{R}_{1/2}$ *from Theorem 2.8.*

4 Simulation Study

This section aims to compare the performance of the proposed adaptive procedures and the well established stationary ES estimation with the default parameter $\eta = 0.94$ and with the optimized parameter for the given data by hand. We consider two versions of the SSA procedure: one with the default parameter set and the other one with the uniform kernel K_{ag} which does a model selection and therefore, referred to as LMS.

In the simulation study, we generate 1000 stochastic processes driven by the hidden Markov model: $R_t = \sqrt{\theta_t}\varepsilon_t$ with ε_t either standard normal or NIG with parameters $\phi = 1.340\,,\ \beta = -0.015\,,\ \delta = 1.337\,,\ \mu = 0.010\,.$ These NIG parameters are in fact the maximum likelihood estimates of the devolatilized Deutsche Mark to the US Dollar daily rates (innovations) from 1979/12/01 to 1994/04/01. The data is available at the FEDC (http://sfb649.wiwi.hu-berlin.de/fedc). The designed volatility process has 7 states : $0.2, 0.25, 0.3, 0.4, 0.5, 0.7$ and 1 , see Figure 3. The sample size of the stochastic processes is $T = 1000\,.$ The first 300 observations are reserved as a training set for the very beginning volatility estimations since the largest smoothing parameter η_K in the adaptive procedure corresponds to 259 past observations.

In the simulation study, we apply the power transform with the frontier value $p = 0.5$ as a default choice. We also present a small sensitivity analysis by varying values of p and show the accuracy of estimation based on the critical values driven in the Gaussian and NIG distributional assumptions respectively. Two criteria are used to measure the accuracy of estimation:

1. Sum of the absolute error (AE) of the estimated volatility.

$$\mathrm{AE} = \sum_{t=301}^{T} |\widehat{\theta}_t^{1/2} - \theta_t^{1/2}|.$$

2. Ratio of the AE (RAE) of the adaptive approach to that of the stationary ES.

$$\mathrm{RAE} = \frac{\mathrm{AE_{SSA}}}{\mathrm{AE_{ES}}} \quad \text{or} \quad \frac{\mathrm{AE_{LMS}}}{\mathrm{AE_{ES}}}$$

The volatility estimates of one realization with $\varepsilon_t \sim \mathrm{N}(0,1)$ is displayed in Figure 3, by which the adaptive SSA estimates fast react to jumps of the process. The LMS displays the similar pattern and the difference between these two adaptive approaches is not significant. It shows that the adaptive estimates better illustrate the movement of the generated volatility process than the ES.

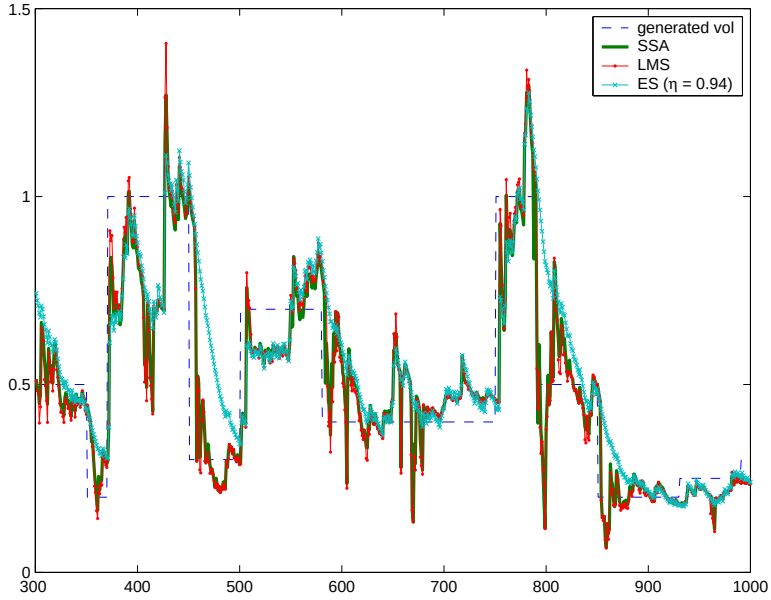

Figure 3: Estimated volatility process based on one realized simulation data with $\varepsilon_t \sim \mathrm{N}(0,1)$. The "optimized" ES ($\eta = 0.94$), LMS and SSA estimates and the generated volatility process are displayed.

Over the 1000 simulations with the Gaussian innovations, the LMS with the average

AE of 68.84 and the SSA with 69.55 are more accurate than the "optimized" stationary ES 82.50 with $\eta = 0.94$. The corresponding average values of RAE of the SSA is 84.42% indicating a roughly 16% improvement over the ES. Moreover, Figure 4 illustrates the boxplot of RAEs w.r.t. not only the adaptive but also the stationary ES approaches with smoothing parameters in the default sequence of $\{\eta_k\}$ for $k = 1, \ldots, 15$, see Table 1. The best performance of the stationary ES is realized for $\eta = 0.895$ that corresponds to $k = 7$. The adaptive ES approaches, namely the SSA and the LMS, show even better performance than the "best" stationary ES approach. The figure also approves that a potential limitation of the SSA compared to the LMS is that it may magnify the bias through the summation as mentioned before.

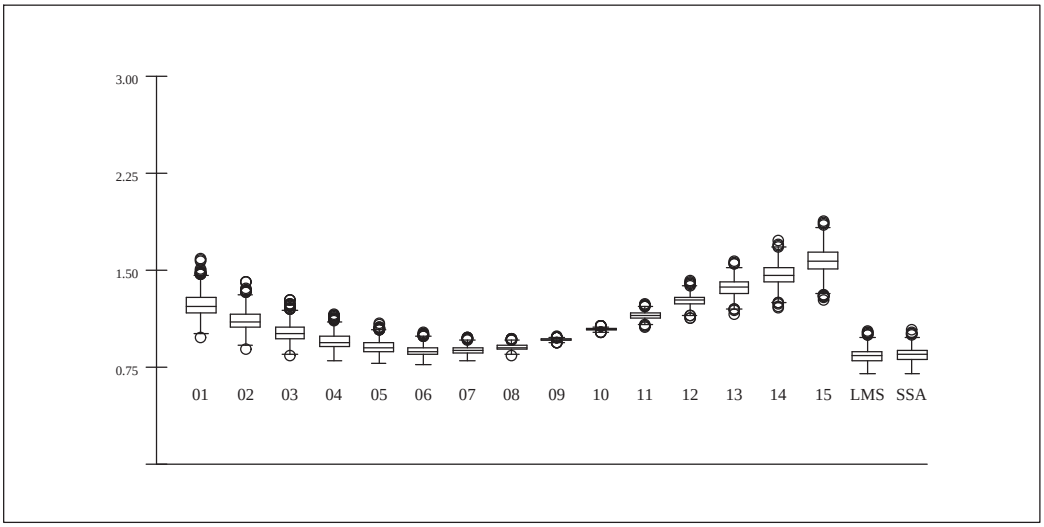

Figure 4: The boxplots of the RAEs of the SSA, LMS and ES with η_k for $k = 1, \ldots, K$.

Table 3 summarizes the estimation errors w.r.t. different values of the four parameters analyzed in Section 2.4. The comparison of the RAEs reasons the default choice in the SSA approach.

Given the simulated heavy-tailed data with the NIG innovations, we follow the procedure explained in Section 3 by first applying the critical values $\mathfrak{z}_k$ computed for the Gaussian

def. SSA	r, def. 0.5			α, def. 1			a, def. 1.25		c, def. 0.01	
	0.3	0.7	1.0	0.5	0.7	1.5	1.1	1.5	0.005	0.02
0.84	0.85	0.87	0.92	0.88	0.86	0.84	0.84	0.86	0.84	0.85

Table 3: Average RAE of the 1000 simulation data sets with $\varepsilon_t \sim \mathrm{N}(0,1)$, by which the SSA method is applied w.r.t. several values of the parameters involved in the adaptive approach. In the stationary ES, $\eta = 0.94$ is applied.

case to the transformed data. Furthermore, we calculate the critical values given the true NIG distributional parameters in the Monte Carlo simulation and reestimate the volatility following the adaptive procedure. Compared to the "optimized" ES, the SSA approach is sensitive to the structure shifts. One realization of the estimated volatility process is displayed in Figure 5. In our study, we also measure the influence of the parameter p over a range from 0.1 to 1 on the estimation, see Table 4. The default choice $p = 0.5$ for example results in an average value of RAE with 90.27% over the 1000 simulations, indicating a better performance of the adaptive method than the "optimized" ES. The RAEs of the SSA estimates based on the critical values under the Gaussian case and the NIG case are reported in the table as well. It is observed that the Gaussian-based critical values works well and the accuracy of estimation is improved as the values of p are close to the default choice.

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
CV N(0, 1)	1.09	1.08	1.06	1.03	0.99	0.94	0.91	0.90	0.90	0.91
CV NIG	1.01	0.96	0.93	0.91	0.90	0.90	0.90	0.90	0.90	0.91

Table 4: Average RAEs over 1000 simulated NIG data sets with different values of p, by which $p = 0.5$ is default choice. Two sequences of critical values calculated in the Gaussian case and given the true NIG parameters are used in the adaptive procedure.

5 Application to Risk Analysis

The aim of this section is to illustrate the performance of the risk management approach based on the adaptive SSA procedure.

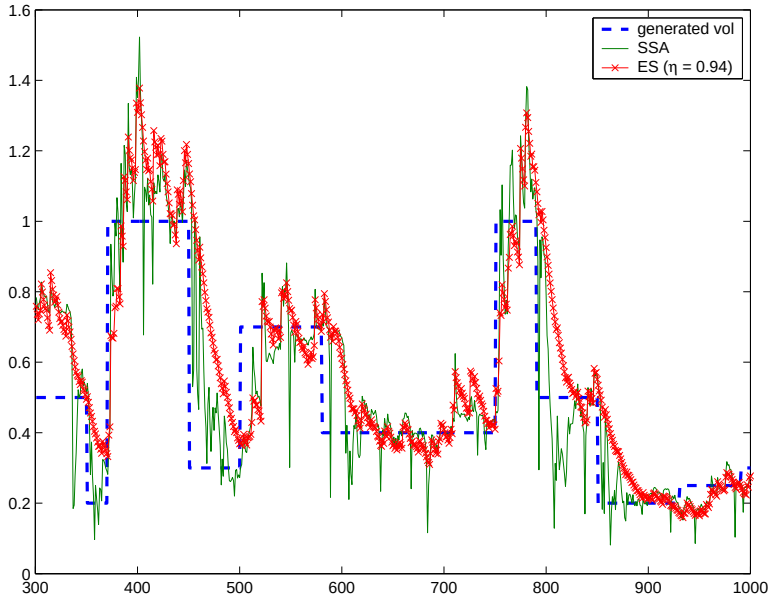

Figure 5: Estimated volatility process based on one realized simulation data with $\varepsilon_t \sim \text{NIG}(1.340, -0.015, 1.337, 0.010)$. The ES ($\eta = 0.94$) and SSA ($p = 0.5$ and critical values given the true NIG parameters) estimates and the generated volatility process are displayed.

A sound risk management system is of great importance, because a large devaluation in the financial market is often followed by economic depression and bankruptcy of credit system. Therefore, it is necessary to measure and control risk exposures using accurate methods. As mentioned before, a realistic risk management method should account for nonstationarity and heavy tailedness of financial time series. In this section, we implement the proposed local exponential smoothing approaches to estimate the time-varying volatility and assume that the innovations are either NIG or Gaussian distributed:

$$R_t = \sqrt{\theta_t}\varepsilon_t, \quad \text{where} \quad \varepsilon_t \sim \mathcal{N}(0,1) \quad \text{or} \quad \varepsilon_t \sim \text{NIG} \tag{5.1}$$

We consider here log-returns of three assets Microsoft (MC), Volkswagen (VW), Deutsche Bank (DB) with daily closed price from 2002/01/01 to 2006/01/05 (972 observations) and of two exchange rates: EUR/USD (EURUSD) and EUR/JPY (EURJPY) ranging from 1997/01/02 to 2006/01/05 (2332 observations). The data sets have been provided by the

data	vola	mean	s.d.	skewness	kurtosis	KPSS
MC	SSA	0.001	1.235	0.261	10.494	0.059
	LMS	-0.004	1.204	0.065	10.173	0.085
	ES	-0.003	1.071	0.545	12.492	0.036
VW	SSA	-0.063	1.150	0.493	9.530	0.065
	LMS	-0.061	1.132	0.477	10.382	0.076
	ES	-0.054	1.050	0.680	10.016	0.056
DB	SSA	-0.097	1.142	-0.661	7.868	0.317
	LMS	-0.100	1.132	-0.631	8.855	0.308
	ES	-0.087	1.025	-0.558	6.561	0.242
EURUSD	SSA	-0.008	1.091	-0.172	4.190	0.317
	LMS	-0.006	1.074	-0.051	4.175	0.258
	ES	-0.014	1.043	-0.278	3.773	0.270
EURJPY	SSA	-0.007	1.121	0.164	4.942	0.313
	LMS	-0.006	1.092	0.186	4.953	0.274
	ES	-0.010	1.051	0.164	4.646	0.292

Table 5: Descriptive statistics of the standardized returns. The critical value of the KPSS test without trend is 0.347 (90%).

financial and economic data center (FEDC) of the Collaborative Research Center 649 on Economic Risk of the Humboldt-Universität zu Berlin. The NIG innovations (standardized returns) are assumed to be stationary. The KPSS tests of stationarity are not rejected at the 90% confidence level, see Table 5.

Two mainly used risk measures at probability pr , Value-at-Risk (VaR) and expected shortfall (ExS), are calculated:

$$\begin{aligned} \mathrm{VaR}_{t,\mathrm{pr}} &= -\mathrm{quantile}(R_t)_{\mathrm{pr}} = -\sqrt{\theta_t} * \mathrm{quantile}(\varepsilon_t)_{\mathrm{pr}} \\ \mathrm{ExS} &= I\!E\{-R_t | -R_t > \mathrm{VaR}_{t,\mathrm{pr}}\}. \end{aligned}$$

The performance of the proposed local exponential smoothing approaches is evaluated from the viewpoints of regulator, investors and internal supervisor.

Minimum regulatory requirement: The main goals of risk regulatory are to ensure the adequacy of capital and restrict the happening of large losses of financial institutions. It

regulates that the financial institutions shall reserve appropriate amount of capital related to 1% risk level of their portfolio, namely the market risk charge (RC), in the central bank:

$$\mathrm{RC}_t = \max\left(M_f \frac{1}{60}\sum_{i=1}^{60} \mathrm{VaR}_{t-i}, \mathrm{VaR}_t\right) \tag{5.2}$$

where the multiplicative factor M_f has a floor value 3. According to the modification of the Basel market risk paper in 1997, financial institutions are allowed to use their internal models to measure the market risks. The internal models are verified in accordance with the "traffic light" rule that counts the number of exceedances over VaR at 1% probability spanning the last 250 days and identifies the multiplicative factor M_f in the form (5.2), see Table 6, cited from Franke, Härdle and Hafner (2004). It is clear that an increase of M_f

Number of exceedances	Increase of M_f	Zone
0 bis 4	0	**green**
5	0.4	**yellow**
6	0.5	**yellow**
7	0.65	**yellow**
8	0.75	**yellow**
9	0.85	**yellow**
More than 9	1	**red**

Table 6: Traffic light as a factor of the exceeding amount.

or concerning an extremal risk level such as 0.5% results in a large amount of risk charge and consequently a low ratio of profit. This observation indicates that the regulatory rule in fact motivates financial institutions to control VaR at $1.6\% = \frac{4}{250}$ level instead of 1%. Therefore an internal model is particularly desirable by generating an empirical probability $\widehat{\mathrm{pr}}$ that is smaller or equal to 1.6%,

$$\widehat{\mathrm{pr}} = \frac{\text{number of exceedances}}{\text{number of total observations}},$$

and simultaneously requiring risk charge as small as possible.

Table 7 gives a detailed report of risk analysis, which shows that all the considered

models locate either in the green or yellow zone. The Gaussian-based adaptive ES models successfully fulfill the minimal requirement of regulatory. To be more specific, the LMS for MC, VW and EURUSD and the SSA for DB generate the favorable results. The EURJPY data is extraordinary by which the models with the Gaussian noise can not fulfill the regulatory requirement. A compensate choice is the ES with the NIG noise.

Investors' review: From the viewpoint of investors, it is important to measure the size of loss instead of the frequency of loss since investors suffer loss at bankruptcy. Even in the "best" case, the loss equals to the difference between the total realized loss and the reserved risk capital. As a consequence, investors care the ExS more than the VaR.

The risk analysis report shows that the Gaussian-based model in general generates larger values of ExS than the NIG-based model. Furthermore, the adaptive ES are desirable for investors concerning extreme risk level. The ExS values of EURJPY at the expected 0.5% level, for example, are 0.231 (SSA), 0.255 (LMS) and 0.263 (ES) with NIG innovations, see Table 7. It is clear that the SSA procedure is superior to the other two.

Internal supervisory review: It is important for internal supervisory to exactly measure the market risk exposures before controlling them. Based on this criterion, it is rational to choose a model that generates the empirical risk level $\widehat{\text{pr}}$ as close as possible to the target one.

In the real data analysis, the models with the NIG innovations and using the local exponential smoothing approaches generate more precise empirical values than other alternative methods at two risk levels 0.5% and 1%.

On summary, the models based on the local volatility estimates and the NIG distributed residuals best suit the requirements of investors and supervisory. The VaR models based on the adaptive approaches and the normal distributional assumption, on the contrary, is successful to fulfill the regulatory requirement.

	1% $\varepsilon_t \sim$ N(0,1)			1% $\varepsilon_t \sim$ NIG			0.5% $\varepsilon_t \sim$ N(0,1)			0.5% $\varepsilon_t \sim$ NIG		
MC	SSA	LMS	ES	SSA	LMS	ES	SSA	LMS	ES	SSA	LMS	ES
except.	12	11	8	7	6	5	10	7	6	4	3	3
prob.	0.018	0.016	0.012	0.010 ^{s}	0.009	0.007	0.015	0.010	0.009	0.006	0.004 ^{s}	0.004
$\sum$ ExS	0.409	0.377	0.325	0.317	0.285	0.265 ^{i}	0.374	0.303	0.286	0.225	0.193 ^{i}	0.202
$\sum$ VaR	17.18	17.43 ^{r}	18.92	22.08	21.94	21.94						
VW	SSA	LMS	ES	SSA	LMS	ES	SSA	LMS	ES	SSA	LMS	ES
except.	12	10	10	7	8	6	8	8	7	2	2	3
prob.	0.018	0.015	0.015	0.010 ^{s}	0.012	0.009	0.012	0.012	0.010	0.003	0.003	0.004 ^{s}
$\sum$ ExS	0.623	0.567	0.567	0.443	0.492	0.360 ^{i}	0.488	0.488	0.439	0.167 ^{i}	0.167	0.215
$\sum$ VaR	27.83	28.37 ^{r}	28.82	32.44	32.58	33.21						
DB	SSA	LMS	ES	SSA	LMS	ES	SSA	LMS	ES	SSA	LMS	ES
except.	10	10	7	5	4	6	8	7	4	3	3	3
prob.	0.015	0.015	0.010	0.007	0.006	0.009 ^{s}	0.012	0.010	0.006	0.004 ^{s}	0.004	0.004
$\sum$ ExS	0.397	0.397	0.285	0.190	0.148 ^{i}	0.259	0.323	0.301	0.168	0.099 ^{i}	0.099	0.126
$\sum$ VaR	28.00 ^{r}	28.35	29.06	31.23	31.84	30.19						
EURUSD	SSA	LMS	ES	SSA	LMS	ES	SSA	LMS	ES	SSA	LMS	ES
except.	34	30	22	15	16	18	20	21	9	11	10	7
prob.	0.017	0.015	0.011	0.008	0.008	0.009 ^{s}	0.010	0.010	0.004	0.005	0.005 ^{s}	0.003
$\sum$ ExS	0.417	0.372	0.309	0.207 ^{i}	0.212	0.248	0.255	0.254	0.134	0.149	0.143	0.105 ^{i}
$\sum$ VaR	28.00	28.35 ^{r}	28.65	29.51	29.95	29.53						
EURJPY	SSA	LMS	ES	SSA	LMS	ES	SSA	LMS	ES	SSA	LMS	ES
except.	52	50	41	21	20	21	34	30	28	10	11	10
prob.	0.026	0.025	0.020	0.010	0.010 ^{s}	0.010	0.017	0.015	0.014	0.005 ^{s}	0.005	0.005
$\sum$ ExS	0.884	0.900	0.797	0.442	0.428 ^{i}	0.463	0.655	0.597	0.572	0.231 ^{i}	0.255	0.263
$\sum$ VaR	32.53	33.09	33.67	40.32	40.21 ^{r}	40.31						

Table 7: Risk analysis of the real data. The exceedances are marked in green, yellow or red according to the traffic light rule. An internal model is accepted if it is in the green zone. The best results to fulfill the regulatory requirement are marked by ^{r}. The recommended method to the investor is marked by ^{i}. For the internal supervisory, we recommend the method marked by ^{s}.

References

Anderson, T., Bollerslev, T., Diebold, F. and Labys, P. (2001). The distribution of realized exchange rate volatility, *Journal of the American Statistical Association* pp. 42–55.

Belomestny, D. and Spokoiny, V. (2006). Spatial aggregation of local likelihood estimates with applications to classification, *WIAS Preprint*.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* pp. 307–327.

Bollerslev, T. and Woolridge, J. (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances, *Econometric Reviews* pp. 143–172.

Box, G. and Cox, D. (1964). An analysis of transformations, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 211–252.

Chen, Y., Härdle, W. and Jeong, S. (2005). Nonparametric risk management with generalized hyperbolic distributions, *SFB 649, discussion paper*.

Cheng, M., Fan, J. and Spokoiny, V. (2003). Dynamic nonparametric filtering with application to volatility estimation, *in* M. Akritas and D. Politis (eds), *Recent advances and trends in nonparametric statistics*, Elsevier, pp. 315–334.

Christoffersen, P. (2003). *Elements of Financial Risk Management*, Adademic Press.

Eberlein, E. and Keller, U. (1995). Hyperbolic distributions in finance, *Bernoulli* **1**: 281–299.

Embrechts, P., McNeil, A. and Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls, *in* M. Dempster (ed.), *Risk Management: Value at Risk and Beyond*, Cambridge University Press.

Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of uk inflation, *Econometrica* pp. 987–1008.

Fan, J. and Gu, J. (2003). Semiparametric estimation of value-at-risk, *Econometrics Journal* **6**: 261–290.

Franke, J., Härdle, W. and Hafner, C. (2004). *Statistics of Financial Markets*, Springer-Verlag Berlin Heidelberg New York.

Mercurio, D. and Spokoiny, V. (2004). Statistical inference for time inhomogeneous volatility models, *The Annals of Statistics* **32**: 577–602.

Polzehl, J. and Spokoiny, V. (2006). Propagation-separation approach for local likelihood estimation, *Probability Theory and Related Fields* pp. 335–362.

Prause, K. (1999). The generalized hyperbolic model: Estimation, financial derivatives and risk measures, *dissertation*.

Spokoiny, V. (2006). *Local parametric methods in nonparametric estimation*, Springer-Verlag Berlin Heidelberg New York.

Zhang, L., Mykland, P. and Ait-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data, *Journal of The American Statistical Association* pp. 1394–1411.