ADAPTIVE ESTIMATION FOR A TIME INHOMOGENEOUS STOCHASTIC-VOLATILITY MODEL

HÄRDLE, W., SPOKOINY, V.G., AND TEYSSIÈRE, G.

Humboldt Universität, Institute for Statistics and Econometrics,
Spandauer str. 1, 10178 Berlin
and
Weierstrass Institute for Applied Analysis and Stochastics,
Mohrenstr. 39, 10117 Berlin

Abstract. Let a process $S_1, \ldots, S_T$ obey the conditionally heteroskedastic equation $S_t = v_t \varepsilon_t$ where ε_t is a random noise and v_t is the volatility coefficient which in turn obeys an autoregression type equation $\log v_t^2 = \omega + \alpha S_{t-1}^2 + \eta_t$ with an additional noise η_t. We consider the situation when the parameters ω and α might also depend on the 'time' t, and we study the problem of online estimation of current values of $\omega = \omega(T)$ and $\alpha = \alpha(T)$ from the observations $S_1, \ldots, S_T$. We propose an adaptive method of estimation which does not use any information about time homogeneity of the observed process.

1. Introduction

Let $S_1, \ldots, S_T$ be an asset price process obeying the conditional heteroskedasticity equation

$$S_t = v_t \varepsilon_t \tag{1}$$

where ε_t is a multiplicative random noise and v_t is the volatility coefficient, $t = 1, 2, \ldots$. We assume that the errors ε_t are i.i.d. random variables satisfying the condition

$$\underset{\sim}{E}\, \varepsilon_t^2 = 1$$

and our aim is to make some statistical inference about the volatility process v_t. Below we use an equivalent representation of this model

1991 *Mathematics Subject Classification.* 62G10; Secondary 62G20.
Key words and phrases. adaptive estimation, local homogeneity, non-stationarity .

with notation $Y_t = \log S_t^2$. Then applying the function $\log(x^2)$ to both sides of (1) and using (3) we get

$$Y_t = \log v_t^2 + \log \varepsilon_t^2 = z_t + \xi_t \tag{2}$$

with

$$\begin{aligned} z_t &= \log v_t^2 + \mu, \\ \xi_t &= \log \varepsilon_t^2 - \mu, \\ \mu &= \underline{\mathrm{E}} \ \log \varepsilon_t^2. \end{aligned}$$

A typical approach to study the model (1) or the equivalent model (2), see Gourieroux (1997), is based on a parametric assumption about the underlying process v_t. One specific example studied in this paper is delivered by so called *stochastic-volatility model*:

$$z_t = \omega + \alpha Y_{t-1} + \eta_t, \qquad t = 2, 3, \ldots$$

where η_t is some other noise process, and ω, α are unknown coefficients. If we know the distribution of the noise processes ε_t and η_t, then the properties of the volatility process v_t and the asset price process Y_t are completely determined by the coefficients ω and α. This reduces the original problem to the problem of estimating ω, α, and standard well developed parametric theory applies in this situation: the maximum likelihood estimate (MLE) is root-n consistent and asymptotically efficient, see e.g. Gourieroux (1997).

By practical applications, the main drawback of this approach is that the model assumption (3) is quite restrictive and being applied to real time series, it may lead to a huge modelling bias. One possibility to make such modelling more flexible is to allow for a high order autoregression and moving averaging structure

$$z_t = \omega + \alpha_1 Y_{t-1} + \cdots + \alpha_p Y_{t-p} + \beta_1 z_{t-1} + \cdots + \beta_q z_{t-q} + \eta_t. \tag{3}$$

Then the coefficients $\alpha_1, \ldots, \alpha_p$, $\beta_1, \ldots, \beta_q$ and even the parameters p, q can be estimated from data with a reasonable accuracy, see e.g. ?? Such type of modelling allows to fit real data much better but it still restricted to the class of time homogeneous processes because the structure of the equation (3) does not depend of time index t.

Dahlhaus (1997), Neumann and von Sachs (1997) studied the problem of statistical inference for locally stationary time series. Their approach allows the coefficients ω and α for time dependence of the form $\alpha = \alpha(t/m)$ and $\omega = \omega(t/m)$. Here $\alpha(\cdot)$ and $\omega(\cdot)$ are smooth unknown functions and the parameter m controls the degree of time homogeneity: the processes S and v are almost stationary inside each time interval of length smaller (in order) than m.

In this paper we consider a situation when the coefficients ω and α in (3) are general piecewise smooth functions, $\omega = \omega(t)$ and $\alpha = \alpha(t)$. This means that for every point s, there exists a number $m = m(s)$ such that the parameters $\omega(t)$ and $\alpha(t)$ do not change significantly within the interval $[s-m, s]$. The difference between this approach and that of Dalhaus (1997) is that we allow the parameter m (which controls the degree of local time homogeneity) to vary from point to point. This particularly helps to include the case of spontaneous change in parameters raised by some external (e.g. political) perturbation. Under such an approach, by estimating $\omega(T), \alpha(T)$, the main intention is to determine by data the largest possible interval of local homogeneity $[T-m, T]$ corresponding to the time point T. Then the values $\omega(T)$ $\alpha(T)$ can be estimated from observations from this interval. The parameter m describing the degree of local homogeneity before the time point T, can be viewed as a smoothness parameter which usually appears in nonparametric estimation. In that sense, we aim to adapt our estimate to this unknown parameter.

Our adaptive estimation procedure is presented in Section 3. The proposed method is a modification of the pointwise adaptive estimation procedure from Spokoiny (1998) for the regression model, adapted to the time series case.

The paper is organized as follows. In the next section we formulate the problem. Section 3 contains the description of the estimation procedure and its properties. The proofs are postponed to the Appendix.

2. The model and the problem

Below we discuss the problem of statistical inference for a time heterogeneous stochastic-volatility model described by the following two equations

$$\begin{aligned} S_t &= v_t \varepsilon_t, \\ \log v_t^2 &= \omega(t) + \alpha_1(t) \log S_{t-1}^2 + \alpha_p(t) \log S_{t-p}^2 + \eta_t, \qquad t = p+1, p+2, \ldots . \end{aligned}$$

We observe $S_1, \ldots, S_T$ and the goal is to estimate the current values $\omega(T)$ and $\alpha_1(T), \ldots, \alpha_p(T)$.

In the sequel we use the equivalent representation (2) and we treat $Y_t = \log S_t^2$ as our observations. It is also useful to introduce some vector notation. Let X_t be a vector in $\Re^{p+1}$ with $X_t = (1, Y_{t-1}, \ldots, Y_{t-p})^T$

and $\theta(t) = \big(\theta_0(t), \dots, \theta_p(t)\big)^T$ be the vector of parameters where

$$\begin{aligned} \theta_0(t) &= \omega(t) + \mu, \\ \theta_k(t) &= \alpha_k(t), \qquad k = 1, \dots, p. \end{aligned}$$

Then the model equation (2) reads

$$Y_t = X_t^T \theta(t) + \zeta_t \tag{4}$$

with

$$\zeta_t = \xi_t + \eta_t = \log \varepsilon_t^2 - \mathop{\mathbb{E}} \log \varepsilon_t^2 + \eta_t$$

and we treat Y_t as our observations and ζ_t as a 'noise' process which is i.i.d. with $\mathop{\mathbb{E}} \zeta_t = 0$ and $\sigma^2 = \mathop{\mathbb{E}} \zeta_t^2 = \mathop{\mathbb{E}} (\xi_t + \eta_t)^2$.

To construct a consistent estimation of $\theta(T)$ we need some additional assumptions on $\theta(\cdot)$. Our approach assumes some kind of local homogeneity of the process Y_t. Namely we suppose that there is an interval of homogeneity $[T - m, T]$ in which the vector function θ can be well approximated by a constant vector and hence, the process Y is nearly homogeneous in this interval.

2.1. **Local homogeneity assumption.** Time homogeneity assumption means that the vector function $\theta(t)$ does not essentially differ from a constant vector θ within some interval I of the form $I = [t_1, T]$ with the right end-point T and hence the original model (4) can be approximated within this interval by a parametric model

$$Y_t = X_t^T \theta + \zeta_t, \qquad t \in I. \tag{5}$$

Based on this assumption, we can construct an estimate of θ from the observations Y_t for $t \in I$ and use this value for estimating $\theta(T)$. In the sequel we consider the pseudo maximum likelihood estimate (pseudo MLE)

$$\tilde{\theta}_I = (\tilde{\theta}_{0,I}, \dots, \tilde{\theta}_{p,I})^T = \left(\sum_{t \in I} X_t X_t^T \right)^{-1} \sum_{t \in I} X_t Y_t \tag{6}$$

which is the ML estimate for the case of Gaussian errors ζ_t. The interval I here can be interpreted as the parameter of the method. If no information about degree of homogeneity of the process Y_t is available, then this approach leads to the problem of adaptive estimation: we have a family of estimates $\{\tilde{\theta}_I\}$ indexed by the parameter I (or, equivalently, by the corresponding left end-point t_1) and we have to select in a data-driven way one of them which provides the minimal possible risk.

Before presenting our adaptive procedure, we need in a more detailed study of the properties of the estimate $\tilde{\theta}_I$.

2.2. **Properties of $\tilde{\theta}_I$.** Due to our assumption of local homogeneity, the vector $\theta(t)$ is close to a constant vector θ for all $t \in I$. This means that the value

$$\Delta_I = \sup_{t \in I} |\theta(t) - \theta(T)|$$

is small. (Here $|v|$ means the Euclidean norm of the vector $v \in \Re^{p+1}$.) We also denote by V_I and W_I the random matrices defined as:

$$\begin{aligned} V_I &= \sigma^{-2} \sum_{t \in I} X_t X_t^T, \\ W_I &= V_I^{-1}. \end{aligned}$$

The elements of these matrices will be denoted by $v_{ij,I}$ and $w_{ij,I}$ respectively, $i,j = 0,1,\ldots,p$. We also use the notation $\sigma_{i,I}^2$ for diagonal elements of W_I, $i = 0,\ldots,p$.

For the case of a regression model of the form (5) with the deterministic design $X_1,\ldots,X_T$, the estimate $\tilde{\theta}_I$ is the least squares estimate and W_I is its covariance matrix. In particular, the diagonal element $w_{ii,I} = \sigma_{i,I}^2$ of this matrix is the variance of the estimate $\tilde{\theta}_{i,I}$, $i = 0,1,\ldots,p$. In our situation the design points X_t's are random and correlated to the errors ζ_t's. Nevertheless, we shall see that these values allow to control the stochastic component of the estimate $\tilde{\theta}_{i,I}$. By analogy with the regression case, we call the values $\sigma_{i,I}^2$ the *conditional variances* of $\tilde{\theta}_{i,I}$, $i = 0,\ldots,p$.

To guarantee some properties of the estimate $\tilde{\theta}_I$, the matrix V_I has to fulfill some regularity conditions which roughly mean that this matrix is well defined. Since this matrix is random, we introduce a random set where such conditions are satisfied and in the sequel restrict our consideration to this set. Given positive numbers $b > 0$, $B \ge 1$, $r \ge 1$ and $\delta < 1$, denote for every $i = 0,\ldots,p$ by $A_{i,I}$ the random set where the following conditions are fulfilled:

$$A_{i,I} = \left\{ \begin{array}{c} b \le w_{ii,I}^{-1} \le bB, \\ \\ w_{ii,I} \|V_I\| \le r, \\ \\ \left| \dfrac{w_{ij,I}}{w_{ii,I}} \right| \le \rho, \quad \forall j = 0,\ldots,p \end{array} \right\}.$$

Here $\|A\|$ denotes the sup-norm of the matrix A: $\|A\| = \sup_{\lambda \in \Re^{d+1}: |\lambda| = 1} |A\lambda|$.

Theorem 1. *Let $Y_1, \ldots, Y_T$ obey (4) and let the errors ζ_t be i.i.d. $\mathcal{N}(0, \sigma^2)$ -distributed. Then it holds for the estimate $\tilde{\theta}_I$*

$$\mathbf{P}\left(|\tilde{\theta}_{i,I} - \theta_i(T)| > \Delta_I + \lambda\sigma_{i,I}, A_{i,I}\right) \\ \leq 4e\,(1 + \log B)\,(1 + 2\rho\sqrt{rp}\lambda)^p\,\lambda \exp\left(-\lambda^2/2\right), \qquad i = 0, \ldots, p.$$

Proof. The model equation (4) and the definition of Δ_I imply

$$\begin{aligned} |\tilde{\theta}_I - \theta(T)| &= \left| \left(\sum_{t \in I}^{T} X_t X_t^T \right)^{-1} \sum_{t \in I} \left(X_t Y_t - \theta(T) X_t X_t^T \right) \right| \\ &= \left| \left(\sum_{t \in I} X_t X_t^T \right)^{-1} \sum_{t \in I} \left(X_t \zeta_t + [\theta(t) - \theta(T)]\, X_t X_t^T \right) \right| \\ &\leq \left| \left(\sum_{t \in I} X_t X_t^T \right)^{-1} \sum_{t \in I} X_t \zeta_t \right| + \Delta_I. \end{aligned}$$

This is nothing but the usual decomposition of the estimate $\tilde{\theta}_I$ into the bias and stochastic components. Denote by $\xi_I = (\xi_{0,I}, \ldots, \xi_{p,I})^T$ the stochastic component of $\tilde{\theta}_I$:

$$\xi_I = \left(\sum_{t \in I} X_t X_t^T \right)^{-1} \sum_{t \in I} X_t \zeta_t$$

We end up by using the following general results for martingales from Liptser and Spokoiny (1999):

$$\mathbf{P}\left(|\xi_{i,I}| > \lambda\sigma_{i,I}\right) \leq 4e\,(1 + \log B)\,(1 + 2\rho\sqrt{rp}\lambda)^p\,\lambda \exp\left(-\lambda^2/2\right). \tag{7}$$

□

3. An adaptive choice of the interval of homogeneity

Due to Theorem 1, the losses of the estimate $\tilde{\theta}_{i,I}$ can be bounded by the sum of two terms Δ_I and $\lambda\sigma_{i,I}$, $i = 0, 1, \ldots, p$. The first term Δ_I characterizes the variability of the function $\theta(\cdot)$ and it is typically small if $|I|$ is small, where $|I|$ denotes the number of points X_t within I , i.e., for $I = [t_1, T]$, we have $|I| = T - t_1 + 1$. In opposite, the term $\sigma_{i,I}$ is relatively large for small $|I|$ and it decreases with $|I|$. More precisely, the values $\sigma_{i,I}^2$, $i = 0, \ldots, p$ are the diagonal elements of the matrix $W_I = \sigma^2 \left(\sum_{t \in I} X_t X_t^T \right)^{-1}$ and they, under some stationarity assumptions, behave as Const. $|I|^{-1}$. Moreover, all values $\sigma_{i,I}$ can be exactly calculated from the data, since they depend on the observations $Y_1, \ldots, Y_T$ only.

Let $\mathfrak{I}$ be a set of all considered intervals I. By N we denote the number of elements in $\mathfrak{I}$. Our adaptive method is based on comparison of different estimates corresponding to subintervals of I. Namely, for every interval I from $\mathfrak{I}$, we suppose to be given a set $\mathfrak{J}(I)$ of subintervals J of I (one example of these sets $\mathfrak{I}$ and $\mathfrak{J}(I)$ is given in the next section). The number of elements in $\mathfrak{J}(I)$ is denoted by N_I. For every $J \in \mathfrak{J}(I)$, we construct the corresponding estimate $\tilde{\theta}_J$ from the observations Y_t with $t \in J$, see (6). If there exists an interval J from $\mathfrak{J}(I)$ such that the estimates $\tilde{\theta}_I$ and $\tilde{\theta}_J$ differ significantly then we reject the hypothesis of homogeneity for the interval I. Finally, our adaptive estimate corresponds to the largest interval I such that the hypothesis of homogeneity is not rejected for all $I' \in \mathfrak{I}$ with $I' \subseteq I$.

For the formal description, given $I \in \mathfrak{I}$, define $\mathfrak{I}(I)$ as the set of all intervals $I' \in \mathfrak{I}$ with $I' \subseteq I$ and

$$\sigma_{i,I}^{*} = \max\{\sigma_{i,I'} : I' \in \mathfrak{I}(I)\}.$$

Now, with two constants λ and D, define the adaptive choice of the interval of homogeneity by the following iterative procedure:

Initialization: Select the smallest interval in $\mathfrak{I}$.

Iteration: Select the next interval I in $\mathfrak{I}$ and calculate the corresponding estimate $\tilde{\theta}_I$ and its conditional variance σ_I.

Testing homogeneity: Reject I, if there exists one $J \in \mathfrak{J}(I)$ such that

$$|\tilde{\theta}_{i,I} - \tilde{\theta}_{i,J}| > \lambda(\sigma_{i,I} + \sigma_{i,J}) + 2D\sigma_{i,I}^{*}$$

Loop: If I is not rejected, then continue with the iteration step. Otherwise, set $\hat{I} =$ "the latest non rejected I".

The adaptive estimate $\hat{\theta}(T)$ of $\theta(T)$ is defined by making use of this selected interval $\hat{I}$:

$$\hat{\theta}_i(T) = \tilde{\theta}_{i,\hat{I}}, \qquad i = 0, \dots, p.$$

We also define the "ideal" choice

$$I\!\!I = \max\left\{ I \in \mathfrak{I} : \Delta_I \le D\sigma_{i,I}^{*},\ i = 0, 1, \dots, p \right\}. \tag{8}$$

For this 'ideal' choice $I = I\!\!I$ we have the balance between the accuracy of approximation (which is controlled by Δ_I) and the stochastic error characterized by the stochastic variance $\sigma_{i,I}^{*}$. This allows us to bound with a high probability the losses of the "ideal" estimate $\tilde{\theta}_{i,I\!\!I}$ by $(D + \lambda)\sigma_{i,I\!\!I}^{*}$ provided that λ is sufficiently large, $i = 0, 1, \dots, p$. The next assertion claims that the risk of the adaptive estimate is of the same order $\sigma_{i,I\!\!I}^{*}$.

Let the random events $A_{i,I}$ be introduced before Theorem 1 and

$$A_i = \bigcap_{I \in \mathfrak{I}} A_{i,I}.$$

Theorem 2. *Let errors $\zeta_1, \ldots, \zeta_T$ be i.i.d. Gaussian r.v.'s with zero mean and the variance σ^2. Then it holds for the adaptive estimate $\hat{\theta} = \tilde{\theta}_{\hat{I}}$*

$$\mathbf{P}\left(|\hat{\theta}_i - \theta_i(T)| > 3(\lambda + D)\sigma^*_{i,I\!I}, A_i\right) \tag{9}$$

$$\leq \left(1 + \sum_{I \in \mathfrak{I}(I\!I)} N_I\right) 4e\left(1 + \log B\right)\left(1 + 2\rho\sqrt{rp}\lambda\right)^p \lambda \exp\left(-\lambda^2/2\right).$$

Proof. Let $I\!I$ be the "ideal" interval from (8). Obviously

$$\mathbf{P}\left(\left|\hat{\theta}_i - \theta_i(T)\right| > 3(\lambda + D)\sigma^*_{i,I\!I},\ A_i\right)$$

$$\leq \mathbf{P}\left(\left|\hat{\theta}_i - \theta_i(T)\right| > 3(\lambda + D)\sigma^*_{i,I\!I},\ I\!I \subseteq \hat{I},\ A_i\right) + \mathbf{P}\left(I\!I \text{ is rejected},\ A_i\right).$$

We evaluate separately each summand in the right side of this inequality. It holds on the event $\{I\!I \subseteq \hat{I}\}$ in view of the definition of $\hat{I}$

$$|\tilde{\theta}_{i,\hat{I}} - \tilde{\theta}_{i,I\!I}| \leq \lambda\big(\sigma_{i,\hat{I}} + \sigma_{i,I\!I}\big) + 2D\sigma^*_{i,\hat{I}} \leq 2(\lambda + D)\sigma^*_{i,I\!I}.$$

Further, using the inequality $\Delta_{I\!I} \leq D\sigma_{i,I\!I}$ and applying Theorem 1, we get

$$\begin{aligned}
&\underline{\mathrm{P}}\left(|\tilde{\theta}_{i,I\!I} - \theta_i(T)| > (\lambda + D)\sigma_{i,I\!I}, A_i\right) \\
&\quad \leq \underline{\mathrm{P}}\left(|\tilde{\theta}_{i,I\!I} - \theta_i(T)| > \lambda\sigma_{i,I\!I} + \Delta_{I\!I}\,,\ A_i\right) \\
&\quad \leq C_1\lambda\left(1 + C_2\lambda\right)^p \exp\left(-\frac{\lambda^2}{2}\right)
\end{aligned}$$

with

$$\begin{aligned}
C_1 &= 4e(1 + \log B), \\
C_2 &= 2\rho\sqrt{rp}.
\end{aligned}$$

Coupled with the previous inequality the latter implies

$$\underline{\mathrm{P}}\left(|\hat{\theta}_i - \theta_i(T)| > 3(\lambda + D)\sigma_{i,I\!I}, A_i, I\!I \subseteq \hat{I}\right) \leq C_1\lambda\left(1 + C_2\lambda\right)^p \exp\left(-\frac{\lambda^2}{2}\right). \tag{10}$$

It remains to evaluate $\underline{\mathrm{P}}(I\!I$ is rejected$)$. Due to the definition

$$\{I\!I \text{ is rejected}\} = \bigcup_{I \in \mathfrak{I}(I\!I)} \bigcup_{J \in \mathfrak{J}(I)} \left\{|\hat{\theta}_{i,I} - \hat{\theta}_{i,J}| > \lambda\big(\sigma_{i,I} + \sigma_{i,J}\big) + 2D\sigma^*_{i,I}\right\}. \tag{11}$$

Note also that for every $I \in \mathcal{I}(I\!I)$ and every $J \in \mathcal{J}(I)$, it holds

$$\begin{aligned} \Delta_I &\le \Delta_{I\!I} \le D\sigma_{i,I\!I} \le D\sigma^*_{i,I} \\ \Delta_J &\le \Delta_{I\!I} \le D\sigma_{i,I\!I} \le D\sigma^*_{i,I}. \end{aligned}$$

Hence, by Theorem 1

$$\begin{aligned} &\mathbf{P}\left(|\tilde{\theta}_{i,I} - \tilde{\theta}_{i,J}| > \lambda\bigl(\sigma_{i,I} + \sigma_{i,J}\bigr) + 2D\sigma^*_{i,I}\,,\, A_i\right) \\ &\quad\le \mathbf{P}\left(|\tilde{\theta}_{i,I} - \theta_i(T)| > \lambda\sigma_{i,I} + \Delta_I\,, A_i\right) + \mathbf{P}\left(|\tilde{\theta}_{i,J} - \theta_i(T)| > \lambda\sigma_{i,J} + \Delta_J\,, A_i\right) \\ &\quad\le 2C_1\lambda\,(1 + C_2\lambda)^p \exp\left(-\frac{\lambda^2}{2}\right). \end{aligned}$$

This along with (10) and (11) yields the assertion.

□

3.1. **The choice of the sets $\mathcal{I}$, $\mathcal{J}(I)$ and the parameters λ and D.** The presented procedure involves the sets $\mathcal{I}$ and $\mathcal{J}(I)$ of considered indices and two numeric parameters λ and D. We now discuss how these parameters can be selected.

First we mention that the numbers N, N_I of elements in the sets $\mathcal{I}$ and $\mathcal{J}(I)$ enter in the upper bound of the error probability of the estimate $\hat{\theta}$: if we have too many estimates to select between, then the error probability becomes larger. A reasonable choice of this set which allows for a rate optimal estimation with only a logarithmic number of elements in $\mathcal{I}$ and $\mathcal{J}(I)$ is as follows. Let $\mathcal{S}_k$ denote the regular grid of points with the step 2^k, that is, $\mathcal{S}_k = \{t \le T : t = 1 + j2^k, j = 0, 1, 2, \ldots\}$, $k = k_0, 1, \ldots$. For every point t, we define $S_{k,+}(t)$ (resp. $S_{k,-}(t)$) as the closest to t from the right (resp. from the left) point of $\mathcal{S}_k$ with $|t - S_{k,\pm}(t)| \ge 2^{k-1}$ ($S_{k,-}(t)$ is undefined for $t \le 2^{k-1}$ and $S_{k,+}$ for $t \ge T - 2^{k-1}$). Next, given k_0 satisfying $2^{k_0} > 2p$ and k^* satisfying $2^{k^*} \le T/2$, define for every t

$$\begin{aligned} \mathcal{S}_-(t) &= \{t' = S_{k,-}(t),\, k = k_0, \ldots, k^*\} \cup \{t' \in \mathcal{S}_{k^*} : t' < t\} \\ \mathcal{S}_+(t) &= \{t' = S_{k,+}(t),\, k = k_0, \ldots, k^*\} \cup \{t' \in \mathcal{S}_{k^*} : t' > t\} \end{aligned}$$

so that $\mathcal{S}_-(t)$ contains all points t' from the set $\mathcal{S}_{k^*}$ with $t' < t$ and also one (closest from the left) points from each set $\mathcal{S}_k$ with $k < k^*$ and similarly for $\mathcal{S}_+(t)$.

Now, given T, the set $\mathcal{I}$ is defined in the form

$$\mathcal{I} = \{I = [t, T],\, t \in \mathcal{S}_-(T)\}.$$

Next, for each $I \in \mathcal{I}$ with the left end point t, the set $\mathcal{J}(I)$ is defined as the set of intervals with either the same (as for I) left end point t

and with the right end point in $\mathcal{S}_+(t)$ or the same right end point T and the left end point in $\mathcal{S}_-(T)$:

$$\mathcal{J}(I) = \{J = [t', T],\, t' \in \mathcal{S}_-(T) : t' > t\} \cup \{J = [t, t'],\, t' = \mathcal{S}_+(t)\}.$$

Clearly, for $I = [t, T] \in \mathcal{I}$ with $T - t \le 2^k \le 2^{k^*}$, it holds $N_I \le 2(k - k_0)$. If $T - t > 2^{k^*}$, then $N_I \le 2(k^* - k_0) + 2(T - t)2^{-k^*}$. We now aim to bound the sum $\sum_{I \in \mathcal{I}} N_I$ entering in the result of Theorem 2. We consider separately two different situations: with a large k^* satisfying $2^{k^*} \asymp T/2$ and with a small k^*. If $2^{k^*} \asymp T/2$, then

$$\sum_{I \in \mathcal{I}} N_I \le \sum_{k=k_0}^{k^*} 2(k - k_0 + 1) \le (k^* - k_0 + 1)^2 \le (\log_2(T/2) - k_0 + 1)^2.$$

If 2^{k^*} is small compared with T, then

$$\sum_{I \in \mathcal{I}} N_I \le (k^* - k_0 + 1)^2 + \sum_{\ell=1}^{T2^{-k^*}} 2\ell \le (k^* - k_0 + 1)^2 + T^2 2^{-2k^*}.$$

In view of the result of Theorem 2, this bound motivates the choice of a large k^* when we have only a logarithmic number of estimates to compare between. Such a choice also reduces the computational difficulty of the method. On the other side, the choice of a smaller k^* improves the behaviour of the estimate after a change-point, see the next section. Therefore, the parameter k^* is to be taken as a reasonable compromise between the sensitivity of the procedure and the computational time we agree to spend.

Concerning the parameter λ, it should be taken to provide that the value $\sum_{I \in \mathcal{I}(\mathbb{I})} N_I (1 + 2\rho\sqrt{rp}\,\lambda)^p \lambda e^{-\lambda^2/2}$ is sufficiently small. This leads to the theoretical recommendation $\lambda \approx \sqrt{(2+\delta)\log M}$ with $M = \sum_{I \in \mathcal{I}(\mathbb{I})} N_I$ some $\delta > 0$. If $2^{k^*} \asymp T/2$, then we get $M \approx \log(\log_2(T/2))^2$ and $\lambda \approx \sqrt{2(2+\delta)\log(\log_2(T/2))}$. Note meanwhile that the bound from Theorem 2 is rather conservative since we replace the probability of the union of dependent events by the sum of the underlying probabilities. The same applies for the bound (7) used for evaluating the deviation probabilities of the difference $\hat{\theta}_{i,I} - \theta_i(T)$. The simulations results show that the choice $\lambda = 3$ provides a reasonable quality of estimation.

There is much more degree of freedom in the choice of D. This parameter controls the balance between the accuracy of approximation of the vector-function θ by a constant vector and the stochastic error in the definition of the "ideal" choice $\mathbb{I}$ (see (8)). The results from

Lepski and Spokoiny (1997) and Spokoiny (1998) lead to the choice $D = \mathrm{Const}\,\lambda$. At the same time, Lepski and Levit (1997) argued that for a smooth function $\theta(t)$, a relevant choice would be $D = 0$. Our simulation results are in agreement with the latter proposal.

3.2. **Sensitivity to change-points.** Here we briefly discuss the behaviour of the proposed adaptive estimate in the situation when the underlying parameter θ is a piecewise constant function of time, that is, this vector spontaneously changes at some time moments t_m. In this situation, for each time moment T, the corresponding interval of homogeneity coincides with $[t_m, T]$ where t_m is the latest change-point before T. Due to the result of Theorem 2, the quality of estimating the vector θ is of the same order as for the "ideal" estimate $\tilde{\theta}_{I\!I}$. We now can show a slightly stronger assertion: the adaptively selected interval $\hat{I}$ is essentially as $[t_m, T]$ provided that the amplitude of the change is sufficiently large compared to the noise level.

Note first, that the arguments used in the proof of Theorem 2 yield that the probability to reject an interval $I \in \mathfrak{I}$ which is strictly contained in $[t_m, T]$ is small (provided that λ is large enough to make the right hand side of (9) small). Now it remains to show that, if $I \in \mathfrak{I}$ is larger than $[t_m, T]$ in the sense that there is an interval $J \in \mathfrak{I}(I)$ lying from the left of t_m, then such I will be rejected with a probability close to one.

Let θ denote the parameter vector in the latest interval of homogeneity and let θ' be the similar vector the preceding interval $[t_{m-1}, t_m]$.

Proposition 1. *Let $I\!I \in \mathfrak{I}$ be the largest interval of homogeneity within $[t_m, T]$ and let I be a larger interval from $\mathfrak{I}$ such that there is some $J \in \mathfrak{J}(I)$ with $J \subseteq [t_{m-1}, t_m]$. If, for some $i \le p$,*

$$|\theta_i - \theta_i'| \ge 2\lambda(\sigma_{i,I} + \sigma_{i,J} + \sigma_{i,I\!I}) + 4D\sigma^*_{i,I\!I}, \tag{12}$$

then

$$\mathbf{P}\left(I \textit{ is rejected}, A_i\right) \ge 1 - 8e\,(1 + \log B)\,(1 + 2\rho\sqrt{rp}\lambda)^p\,\lambda \exp\left(-\lambda^2/2\right). \tag{13}$$

Proof. Clearly the event $\{I \text{ is not rejected}\}$ includes

$$\begin{aligned} |\tilde{\theta}_{i,I} - \tilde{\theta}_{i,I\!I}| &\le \lambda\,(\sigma_{i,I} + \sigma_{i,I\!I}) + 2D\sigma^*_{i,I\!I}, \\ |\tilde{\theta}_{i,I} - \tilde{\theta}_{i,J}| &\le \lambda\,(\sigma_{i,I} + \sigma_{i,J}) + 2D\sigma^*_{i,I}. \end{aligned}$$

Next, by Theorem 2, using homogeneity within $I\!I$ and J

$$\begin{aligned} &\mathbf{P}\left(|\tilde{\theta}_{i,I\!I} - \theta_i| \ge \lambda\sigma_{i,I\!I}, A_i\right) + \mathbf{P}\left(|\tilde{\theta}_{i,J} - \theta_i'| \ge \lambda\sigma_{i,J}, A_i\right) \\ &\qquad \le 8e\,(1 + \log B)\,(1 + 2\rho\sqrt{rp}\lambda)^p\,\lambda \exp\left(-\lambda^2/2\right). \end{aligned}$$

Since also $\sigma^*_{i,I} \le \sigma^*_{i,I\!I}$, the inequality (12) implies (13). □

We now specify the previous results for the most important situation when we estimate the parameters immediately after a change point.

Given $T > t_m$, define k by the inequalities $2^{k-1} \le T - t_m < 2^k$ and suppose that $k < k^*$. Assume also for simplicity that $t_m \in \mathcal{S}_k$. Then $S_{k,-}(T) = t_m$, that is, $I\!I = [t_m, T]$. We are interested to describe the minimal k which provides $\hat{I} = I\!I$ with a probability close to one.

Corollary 1. *Let $I\!I = [t_m, T]$ with $t_m = S_{k,-}(T)$ for some $k < k^*$. If the conditional variances σ_I^2 satisfy*

$$\sigma_I^2 \le C|I|^{-1}$$

for all $I \in \mathcal{I}$ and if, for some $i \le p$,

$$|\theta_i - \theta_i'| \ge C 2^{-(k-1)/2}(6\lambda + 4D)$$

then

$$\mathbf{P}\left(\hat{I} = I\!I\right) \ge 1 - \left[2 + (k - k_0)^2\right] 4e\,(1 + \log B)\,(1 + 2\rho\sqrt{rp}\lambda)^p\,\lambda \exp\left(-\lambda^2/2\right).$$

Proof. The arguments from the proof of Theorem 2 lead to the bound

$$\mathbf{P}\,(I\!I \text{ is rejected}) \le (k - k_0)^2 4e\,(1 + \log B)\,(1 + 2\rho\sqrt{rp}\lambda)^p\,\lambda \exp\left(-\lambda^2/2\right).$$

Let $t = S_{k+1,-}(T)$. It easily follows from the definition of $S_{k+1,-}$ that either $t_m - t = 2^{k-1}$ or $t_m - t = 2^k$. Provided that $I\!I$ is not rejected, we have $\{\hat{I} \ne I\!I\} = \{I \text{ is not rejected}\}$. Now the assertion follows directly from Proposition 1. □

Remark 1. *Let $a = \max_i |\theta_i - \theta_i'|$ be the amplitude of the change point. The message of Corollary 1 is that the procedure detects the this change point in the sense that it starts to select $\hat{I} \approx I\!I = [t_m, T]$ when the distance $T - t_m$ between the point of estimation and the change point becomes of order $Const.\lambda^2 a^{-2}$. It follows from Spokoiny (1998, Theorem) that the proposed procedure provides the optimal in rate sensitivity to change points.*

References

[1] Gourieroux (1997). *ARCH Models and Financial applications.* Springer. New York.

[2] Dahlhaus (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **25** , no. 1, 1–37.

[3] Neumann,M. and von-Sachs,R. (1997). Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *Ann. Statist.* **25** no. 1, 38–76.

[4] Lepski, O. and Spokoiny, V. (1997) Optimal Pointwise Adaptive Methods in Nonparametric Estimation. *Annals Statist.*, **25**, no.6, 2512–2546.

[5] Liptser, R. and Spokoiny, V. (1999). Deviation probability bound for martingales with applications to statistical estimation. Preprint **??**. Weierstrass-Institute, Berlin.

[6] Spokoiny, V. (1998). Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *Annals Statist.*, **26**, no. 4, 1356–1378.

HUMBOLDT UNIVERSITÄT BERLIN, SPANDAUERSTR. 1, 10178 BERLIN, GERMANY.

WEIERSTRASS INSTITUTE FOR APPLIED ANALYSIS AND STOCHASTICS, MOHRENSTR. 39, 10117 BERLIN, GERMANY

HUMBOLDT UNIVERSITÄT BERLIN, SFB 373, & GREQAM, SPANDAUERSTR. 1, 10178 BERLIN, GERMANY.