Journal of the Royal Statistical Society
SERIES B Statistical Methodology
J. R. Statist. Soc. B (2014)

Multiple-change-point detection for auto-regressive conditional heteroscedastic processes

P. Fryzlewicz

London School of Economics and Political Science, UK

and S. Subba Rao

Texas A&M University, College Station, USA

[Received March 2010. Final revision August 2013]

Summary. The emergence of the recent financial crisis, during which markets frequently underwent changes in their statistical structure over a short period of time, illustrates the importance of non-stationary modelling in financial time series. Motivated by this observation, we propose a fast, well performing and theoretically tractable method for detecting multiple change points in the structure of an auto-regressive conditional heteroscedastic model for financial returns with piecewise constant parameter values. Our method, termed BASTA (binary segmentation for transformed auto-regressive conditional heteroscedasticity), proceeds in two stages: process transformation and binary segmentation. The process transformation decorrelates the original process and lightens its tails; the binary segmentation consistently estimates the change points. We propose and justify two particular transformations and use simulation to fine-tune their parameters as well as the threshold parameter for the binary segmentation stage. A comparative simulation study illustrates good performance in comparison with the state of the art, and the analysis of the Financial Times Stock Exchange FTSE 100 index reveals an interesting correspondence between the estimated change points and major events of the recent financial crisis. Although the method is easy to implement, ready-made R software is provided.

Keywords: Binary segmentation; Cumulative sum; Mixing; Non-stationary time series; Process transformation; Unbalanced Haar wavelets

1. Introduction

Log-returns on speculative prices, such as stock indices, currency exchange rates and share prices, often exhibit the following well-known properties (see for example Rydberg (2000)): the sample mean of the observed series is close to 0; the marginal distribution is roughly symmetric or slightly skewed, has a peak at zero and is heavy tailed; the sample auto-correlations are 'small' at almost all lags, although the sample auto-correlations of the absolute values and squares are significant for a large number of lags; volatility is 'clustered', in that days of either large or small movements are likely to be followed by days with similar characteristics.

To capture these properties, we need to look beyond the stationary linear time series framework, and to preserve stationarity a large number of non-linear models have been proposed. Among them, two branches are by far the most popular: the families of auto-regressive conditional heteroscedastic (ARCH) (Engle, 1982) and generalized auto-regressive conditional heteroscedastic (GARCH) (Bollerslev, 1986; Taylor, 1986) models, as well as the

Address for correspondence: P. Fryzlewicz, Department of Statistics, Columbia House, London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, UK.
E-mail: p.fryzlewicz@lse.ac.uk

© 2013 Royal Statistical Society 1369–7412/14/76000

family of 'stochastic volatility' models (Taylor, 1986). For a review of recent advances in ARCH, GARCH and stochastic volatility modelling, we refer the reader to Fan and Yao (2003) and Giraitis *et al.* (2005).

Although stationarity is an attractive assumption from the estimation point of view, some researchers point out that the above properties can be better explained by resorting to non-stationary models. Dahlhaus and Subba Rao (2006) proposed a time varying ARCH model, where the model parameters evolve over time in a continuous fashion. Mikosch and Stărică (2004) considered probabilistic properties of a piecewise stationary GARCH model and showed that it explains well the 'long memory' effect in squared log-returns. Underlying these approaches is the observation that, given the changing pace of the world economy, it is unlikely that log-return series should stay stationary over long time intervals. For example, given the 'explosion' of market volatility during the recent financial crisis, it is hardly plausible that the volatility dynamics before and during the crisis could be well described by the same stationary time series model. Indeed, Janeway (2009) went further and argued that financial theorists' belief that

> 'statistical attributes of financial time series—such as variance, correlation and liquidity—are stable observables generated by stationary processes'

might have been a contributing factor in the crisis.

In this paper, we focus on processes with piecewise constant parameter values as the simplest form of departure from stationarity. The appeal of this kind of modelling is in that it is easily interpretable as it provides segmentation of the data into time intervals where the parameters of the process remain constant. Also, the piecewise constant parameter approach can be of use in forecasting, where it is often of interest to obtain the 'last' stationary segment of the data which can then be used to forecast the future behaviour. The model we consider is that of an ARCH process with piecewise constant parameters. We note that Fryzlewicz *et al.* (2008) demonstrated that time varying ARCH processes capture well the empirical features of log-return series that were listed in the first paragraph. Since estimation for time varying GARCH processes is a much more challenging task (because likelihood functions are typically 'flat'), and time varying ARCH processes often describe typical log-returns sufficiently well, we do not consider time varying GARCH processes in this paper.

In any model with piecewise constant parameters, one task of interest is to detect, *a posteriori*, the change points, i.e. time instants when the parameters of the process changed. The problem of detecting a single change point was studied, for example, by Chu (1995) and Kulperger and Yu (2005) for the GARCH model and by Kokoszka and Leipus (2000) for the ARCH model. The problem of multiple-change-point detection (or segmentation) of linear time series has been studied by, among others, Adak (1998), Stoffer *et al.* (2002), Davis *et al.* (2006), Last and Shumway (2008), Paparoditis (2010) and Cho and Fryzlewicz (2012). This task for ARCH-type processes is more difficult and has not been studied rigorously by many researchers. The heuristic procedure of Andreou and Ghysels (2002) for the GARCH model was based on the work of Lavielle and Moulines (2000) for detecting multiple breaks in the mean of an otherwise stationary time series. We also mention the computationally intensive procedure of Davis *et al.* (2008) for non-linear time series, based on the minimum description length principle, and the method of Lavielle and Teyssiere (2005), based on penalized Gaussian log-likelihood where the penalization parameter is chosen automatically.

Our aim in this paper is to devise a statistically rigorous, well performing and fast technique for multiple-change-point detection in the ARCH model with piecewise constant parameters, where neither the number nor the amplitudes of the changes are assumed to be known. Our

method, termed BASTA (binary segmentation for transformed auto-regressive conditional heteroscedasticity), proceeds in two stages: the *process transformation* and the *binary segmentation* stage, which we now briefly describe in turn.

1.1. Process transformation

Given a stretch of data from an ARCH process X_t, the initial step is to form a transformation of the data, $U_t = g(X_t, \ldots, X_{t-\tau})$ (for a certain fixed τ which will be specified later), whose aim is twofold: to ensure that the marginal distribution of U_t is bounded, and to ensure that the degree of auto-correlation in U_t is less than that in X_t^2. Formally speaking, the aim of BASTA will be to detect change points in the mean value of U_t. We discuss two suitable choices of g, leading to two algorithms: BASTA-res and BASTA-avg. In the former, more complex construction, we initially choose g in such a way that it corresponds to the sequence of empirical residuals of X_t under the null hypothesis of no change points present. This then leads us to consider an entire family of transformations, $g_{\mathbf{C}}$, indexed by a vector constant $\mathbf{C}$, whose suitable default choice is discussed. In the latter, simpler, construction, g corresponds to local averages of X_t^2, suitably subsampled to reduce auto-correlation and logged to stabilize variance.

1.2. Binary segmentation

In the second stage of the BASTA algorithm, we perform a binary segmentation procedure on the sequence U_t, with the aim of detecting the change points in $\mathbb{E}(U_t)$. Algorithmically, our binary segmentation procedure is performed similarly to Venkatraman's (1992) method for detecting mean shifts in Gaussian white noise sequences, except that we use a more general form of the threshold. We demonstrate that BASTA leads to a consistent estimator of the number and location of change points in $\mathbb{E}(U_t)$. We note that modifications in proof techniques are needed because U_t is a highly structured time series rather than a Gaussian white noise sequence. On the basis of an extensive simulation study, we propose a default choice of the threshold constants, which (reassuringly) works well for both proposed transformations g.

The paper is organized as follows. Section 2 describes the model and the problem. Section 3 introduces our generic algorithm and shows its consistency. Section 4 discusses two particular choices of the function g and the threshold constants. Section 5 describes the outcome of a comparative simulation study where we compare our method with the state of the art. Section 6 describes the application of our methodology to the Financial Times Stock Exchange FTSE 100 index and reveals a (possible) fascinating correspondence between the estimated change points and some major events of the recent financial crisis. The proof of our consistency result appears in Appendix A. Appendix B provides extra technical material.

R software implementing BASTA can be obtained from `http://stats.lse.ac.uk/fryzlewicz/basta/basta.html`.

2. Model and problem set-up

The piecewise constant parameter ARCH(p) model X_t that we consider in this paper is defined as follows:

$$\begin{aligned} X_t &= Z_t \sigma_t, \\ \sigma_t^2 &= a_0(t) + \sum_{j=1}^{p} a_j(t) X_{t-j}^2, \qquad t = 0, \ldots, T-1, \end{aligned} \tag{1}$$

where the independent and identically distributed innovations Z_t are such that $\mathbb{E}(Z_t) = 0$ and $\mathbb{E}(Z_t^2) = 1$, and the non-negative (with $a_0(t) > 0$ and $a_p(t) \not\equiv 0$) piecewise constant parameter vectors $\{a_j(t)\}_t$ have N change points $0 < \eta_1 < \ldots < \eta_N < T - 1$ ($\eta_0 = 0, \eta_{N+1} = T - 1$), i.e., for each η_i, $i = 1, \ldots, N$, there is at least one parameter vector $\{a_j(t)\}_t$ such that $a_j(\eta_i) \neq a_j(\eta_i - 1)$. In assumption 1, in Section 3.2, we state assumptions on $\{a_j(t)\}$ such that X_t admits almost surely a well-defined solution and specify the degree to which we require the parameters to differ between each segment of constancy. For completeness, we assume that X_t for $t = -1, -2, \ldots$ comes from a stationary ARCH(p) process with parameters $\{a_j(0)\}_{j=0}^p$.

Neither the number N nor the locations η_i of the change points are assumed to be known, and our goal is to estimate them. Naturally, we also do not assume that the parameter values $a_j(t)$ are known. N is permitted to increase slowly with the sample size T, but in such a way that a minimum spacing between the η_is is preserved (see assumption 1 for precise rates). We assume that we observe $\{X_t; t = 1, \ldots, T\}$, where formally, $a_j(t)$, N and η_i all depend on the sample size T, although for simplicity this is not reflected in our notation. We do not study the issue of order selection for piecewise constant parameter ARCH processes: if the order p is not known, we note that, in our setting, both underfitting and overfitting the model is permitted in the sense that choosing p to be different from the true order does not affect the validity of either our theory or our algorithm but may reduce the quality of the estimators.

3. Generic algorithm and consistency result

3.1. General approach and motivation

Our method for multiple-change-point detection in the framework that was described in Section 2 is termed BASTA. Its main ingredient is the binary segmentation procedure, suitably modified for use in the ARCH model with piecewise constant parameters. The binary segmentation procedure for detecting a change in the mean of normal random variables was first introduced by Sen and Srivastava (1975). Consistency of binary segmentation for a larger class of processes was shown by Vostrikova (1981); however, conditions for consistency were formulated under more restrictive assumptions on change points than ours, and the procedure itself was not easily implementable owing to the difficulty in computing the null distribution of the change point detection statistic. Venkatraman (1992) proved consistency of the binary segmentation procedure in the Gaussian function plus noise model, using a particularly simple form of the test statistic.

We note that Fryzlewicz's (2007) unbalanced Haar technique for function estimation in the Gaussian function plus noise model is related to binary segmentation in that both proceed in a recursive fashion by iteratively fitting best step functions to the data (see Fryzlewicz (2007) for a discussion of similarities and differences between the two approaches). Indeed, our choice of binary segmentation as a suitable methodology for change point detection in the ARCH model with piecewise constant parameters is motivated by the good practical performance of the unbalanced Haar estimation technique for the Gaussian function plus noise model.

Since BASTA proceeds in a recursive fashion by acting on subsamples determined by previously detected change points, it can be viewed as a 'multiscale' procedure. The next section provides a more precise description of BASTA and formulates a consistency result.

3.2. Algorithm and consistency result

The BASTA algorithm consists of two stages.

Stage I: in the first stage, a process $U_t = g(X_t, X_{t-1}, \ldots, X_{t-\tau})$ is formed. Suitable choices of $g(\cdot)$ and τ will be discussed in Section 4. The process U_t is designed in such a way that its time varying expectation carries information about the changing parameters of X_t and the corresponding change points.

Stage II has the following three steps.

Step 1: begin with $(j,l) = (1,1)$. Let $s_{j,l} = 0$ and $u_{j,l} = T - 1$.
Step 2: denoting $n = u_{j,l} - s_{j,l} + 1$, compute

$$\tilde{U}^b_{s_{j,l},u_{j,l}} = \frac{\sqrt{(u_{j,l}-b)}}{\sqrt{\{n(b-s_{j,l}+1)\}}} \sum_{t=s_{j,l}}^{b} U_t - \frac{\sqrt{(b-s_{j,l}+1)}}{\sqrt{\{n(u_{j,l}-b)\}}} \sum_{t=b+1}^{u_{j,l}} U_t$$

for all $b \in (s_{j,l}, u_{j,l})$. Denote $b_{j,l} = \arg\max_b |\tilde{U}^b_{s_{j,l},u_{j,l}}|$.
Step 3: for a given threshold $\tilde{b}_T$, if $|\tilde{U}^{b_{j,l}}_{s_{j,l},u_{j,l}}| < \tilde{b}_T$, then stop the algorithm on the interval $[s_{j,l}, u_{j,l}]$. Otherwise, add $b_{j,l}$ to the set of estimated change points, and

(a) store $(j_0, l_0) = (j,l)$, let $(s_{j+1,2l-1}, u_{j+1,2l-1}) := (s_{j,l}, b_{j,l})$, update $j := j+1$, $l := 2l-1$, and go to step 2;
(b) recall $(j,l) = (j_0, l_0)$ stored in step (a), let $(s_{j+1,2l}, u_{j+1,2l}) := (b_{j,l}+1, u_{j,l})$, update $j := j+1$, $l := 2l$, and go to step 2.

The maximization of the statistic $|\tilde{U}^b_{s_{j,l},u_{j,l}}|$ in step 2 of the above algorithm is a version of the well-known cumulative sum test and is described in more detail in Brodsky and Darkhovsky (1993), section 3.5. If U_t were a serially independent Gaussian sequence with one change point in the means of otherwise identically distributed variables, $b_{1,1}$ would be the maximum likelihood statistic for detecting the likely location of the change point and would be optimal in the sense of theorem 3.5.3 in Brodsky and Darkhovsky (1993). In our setting, it simply furnishes a least-squares-type estimator; note that, since our U_t is a highly structured time series, exact maximum likelihood estimators of its change points are not easy to obtain and, even if they were, their optimality (or otherwise) would not be easy to investigate. Steps 3(a) and 3(b) describe the binary recursion to the left and to the right of each detected change point; hence the name 'binary segmentation'. We denote the number of the thus-obtained change point estimates by $\hat{N}$ and their locations $b_{j,l}$, sorted in the increasing order, by $\hat{\eta}_1, \ldots, \hat{\eta}_{\hat{N}}$. We note that the threshold $\tilde{b}_T$ depends on the length T of the initial sample, and not on the changing length of each subsegment $[s_{j,l}, u_{j,l}]$.

The following notation prepares the ground for the main result of this paper: a consistency result for BASTA. Let $\{\tilde{X}^i_t\}_t$ denote a *stationary* ARCH(p) process with parameters $a_0(\eta_i), \ldots, a_p(\eta_i)$ $(i = 0, \ldots, N)$, constructed by using the same sequence of innovations Z_t as the original process (1). For each i, we form the process $\tilde{U}^i_t = g(\tilde{X}^i_t, \ldots, \tilde{X}^i_{t-\tau})$, with any fixed τ. Let $\upsilon(t)$ be the index i of the largest change point η_i less than or equal to t. We define

$$g_t = \mathbb{E}(\tilde{U}^{\upsilon(t)}_t).$$

We note that, unlike $\mathbb{E}\{g(X_t, \ldots, X_{t-\tau})\}$, g_t is exactly constant between each pair of change points (η_i, η_{i+1}). The proof of our consistency result below will rely on g_t being, in a certain sense, a limiting value for $\mathbb{E}\{g(X_t, \ldots, X_{t-\tau})\}$.

Before we formulate a consistency result for BASTA, we specify the following technical assumption. C denotes a generic positive constant, not necessarily the same in value each time it is used.

Assumption 1.

(a) For all T, we have $\min_{\{i=0,\ldots,N\}}\{\eta_{i+1}-\eta_i\} \geqslant \delta_T$, where the minimum spacing δ_T satisfies $\delta_T = CT^{\Theta}$ with $\Theta \in (\frac{3}{4}, 1]$.
(b) The number N of change points is bounded from above by the function of the sample size T specified in Appendix B.
(c) The function $g: \mathbb{R}^{\tau+1} \to \mathbb{R}$ satisfies $|g(\cdot)| \leqslant \bar{g} < \infty$ and is Lipschitz continuous in its squared arguments (i.e. satisfies $|g(x_0,\ldots,x_\tau) - g(y_0,\ldots,y_\tau)| \leqslant C\Sigma_{i=0}^{\tau}|x_i^2 - y_i^2|$).
(d) For some $m > 0$ and all T, the sequence g_t satisfies $\min_{\{i=1,\ldots,N\}} |g_{\eta_i} - g_{\eta_i - 1}| \geqslant m$.
(e) The threshold $\tilde{b}_T$ satisfies $\tilde{b}_T = cT^{\theta}$ with $\theta \in (\frac{1}{4}, \Theta - \frac{1}{2})$ and $c > 0$.
(f) For some $\delta_1 > 0$ and all T, we have $\max_{1\leqslant t\leqslant T} \Sigma_{i=1}^{p} a_i(t) \leqslant 1 - \delta_1$.
(g) For some $\delta_2 > 0$ and $C < \infty$, and all T, we have $\min_{1\leqslant t\leqslant T} a_0(t) > \delta_2$ and $\max_{1\leqslant t\leqslant T} a_0(t) \leqslant C < \infty$.
(h) Let f_{Z^2} denote the density of Z_t^2 in expression (1). For all $a > 0$ we have $\int |f_{Z^2}(u) - f_{Z^2}\{u(1+a)\}|\,\mathrm{d}u \leqslant Ka$ for some K independent of a.

Assumption 1, part (a), specifies the minimum permitted distance between consecutive change points; part (b) determines how fast the number of change points is allowed to increase with the sample size. In part (c), the boundedness and Lipschitz continuity of g are technical assumptions which not only facilitate our proofs but also mean that we can avoid placing bounded or normality assumptions on Z_t. Assumption 1, part (d), requires that the consecutive levels of the asymptotic mean function g_t should be sufficiently well separated from their neighbours. Assumption 1, part (e), determines the magnitude of the threshold. Part (f) means that almost surely X_t has a unique causal solution. In addition, parts (f)–(h) are required to guarantee that X_t is strongly mixing at a geometric rate; see assumption 3.1 (and its discussion) as well as theorem 3.1 in Fryzlewicz and Subba Rao (2011). Assumption 1, part (h), is a mild assumption and is satisfied by many well-known distributions, as explained below assumption 3.1 in Fryzlewicz and Subba Rao (2011). The following theorem specifies a consistency result for BASTA.

Theorem 1. Suppose that assumption 1 holds. Let N and $\eta_1,\ldots,\eta_N$ denote respectively the number and the locations of change points. Let $\hat{N}$ denote the number, and $\hat{\eta}_1,\ldots,\hat{\eta}_{\hat{N}}$ the locations, sorted in increasing order, of the change point estimates obtained by BASTA. There exist positive constants C and α such that $P(A_T) \to 1$, where

$$A_T = \{\hat{N} = N; \quad |\hat{\eta}_j - \eta_j| \leqslant C\varepsilon_T \text{ for } 1 \leqslant j \leqslant N\},$$

with $\varepsilon_T = T^{1/2}\log^{\alpha}(T)$.

We note that the factor of $T^{1/2}\log^{\alpha}(T)$ appearing in the event A_T is due to the fact that the change points η_j are measured in the 'real' time $t \in \{0,\ldots,T-1\}$, as opposed to the rescaled time $t/T \in [0,1]$. Another way to interpret the above result is $|\hat{\eta}_j/T - \eta_j/T| \leqslant CT^{-1/2}\log^{\alpha}(T)$. The proof of theorem 1 appears in Appendix A.

Finally, we remark that although the most part of the proof of theorem 1 relies on the mixing properties of X_t and its mixing rates, rather than its coming from a particular time series model, the crucial lemma 1 is specific to the ARCH(p) model with piecewise constant parameters. Its generalization to, for example, the ARCH(∞) model with piecewise constant parameters is possible, but technically challenging (see section 4.2 in Fryzlewicz and Subba Rao (2011)) and cannot proceed without extra assumptions on the ARCH(∞) parameters. We do not pursue this extension in the current work.

4. Two particular choices of the $g(\cdot)$ function and selection of threshold constant c

4.1. General requirements

In this section, we discuss our recommended choices of the transformation function g. We start by recalling the desired properties of the transformed process $U_t = g(X_t, X_{t-1}, \ldots, X_{t-\tau})$.

(a) The time varying expectation of U_t should carry information about the change points, i.e. should change at the change point locations.
(b) A high degree of auto-correlation in U_t would not be desirable as it would have the potential to affect the statistic $\tilde{U}^b_{s_{j,l},u_{j,l}}$, thus giving a false picture of the locations of the change points. Thus, we aim at processes U_t with as little degree of auto-correlation as possible.
(c) In addition, assumption 1, part (c), requires that the function g should be bounded and Lipschitz continuous in its squared arguments.

Intuitively, requirement (a) implies that the process U_t should be a function of *even* powers of X_t. This is because, if Z_t has a symmetric distribution, then so does X_t, which means that, for q odd, if $\mathbb{E}(X_t^q)$ exists, then it equals 0. Thus, the expectation of odd powers of X_t is 'uninteresting' from the point of view of change point detection.

Requirement (b) suggests that any 'diagonal' transformation, where $g(X_t, X_{t-1}, \ldots, X_{t-\tau})$ is a function of X_t only, should not be used. (Examples of such transformations include $U_t = g(X_t) = X_t^2$ or $U_t = g(X_t) = \log(X_t^2)$.) This is because, by the definition of the ARCH process X_t, the squared process X_t^2 has a high degree of auto-correlation, which would typically be preserved in a diagonal transformation of the type $g(X_t)$.

We also remark that requirement (c) guards against transformations which are for example linear in X_t^2, such as the transformation $g(X_t) = X_t^2$. Even for Gaussian innovations Z_t, X_t^2 does not typically have all finite moments, which we refer to as 'heavy-tailedness' throughout the paper. However, heavy tails in $g(\cdot)$ could distort the performance of binary segmentation in the sense of reducing, perhaps to an empty set in the most extreme cases, the permitted range of thresholds for which the procedure would yield consistent results.

4.2. BASTA-res: the residual-based BASTA

Our first proposed transformation U_t, leading to the *BASTA-res* algorithm (the residual-based BASTA), is constructed as follows. Under the null hypothesis of stationarity, the process

$$U_t^{(1)} = \frac{X_t^2}{a_0 + \sum_{i=1}^{p} a_i X_{t-i}^2} = Z_t^2$$

is stationary, and perfectly decorrelated as it is simply an independent and identically distributed sequence of squared innovations Z_t^2. Obviously, in practice, this transformation is impossible to effect as it involves the unknown parameter values a_i. Instead, we 'approximate' it with a transformation

$$U_t^{(2)} = \frac{X_t^2}{C_0 + \sum_{i=1}^{p} C_i X_{t-i}^2},$$

which, under the null hypothesis, also results in a process which is stationary, and hopefully almost decorrelated because of its closeness to $U_t^{(1)}$. The parameter $\mathbf{C} = (C_0, \ldots, C_p)$ will need to be estimated from the data and we describe later how.

To ensure the boundedness of U_t, we add an extra term εX_t^2 in the denominator, which results in the transformation

$$U_t^{(3)} = \frac{X_t^2}{C_0 + \sum_{i=1}^{p} C_i X_{t-i}^2 + \varepsilon X_t^2}. \tag{2}$$

In this paper, for simplicity, we do not dwell on the choice of ε: in fact, in the numerical experiments that are described later, we always assume that ε has the default value of 10^{-3}, with X_t being normalized in such a way that the sample variance of the data vector X_t equals 1. Constructed with a different purpose, a transformation related to $U_t^{(3)}$ has also appeared in Politis (2007).

As discussed above, the hope is that the transformation $U_t^{(3)}$ will produce, with a suitable choice of $\mathbf{C}$, a sequence approximating the squared empirical residuals of the process under the null hypothesis of stationarity. Under the alternative hypothesis, $U_t^{(3)}$ is still (by construction) a sequence of non-negative random variables whose changing expectation from one (approximately) stationary segment to another reflects the different parameter regimes. In practice, $U_t^{(3)}$ tends to have a distribution which is highly skewed to the right. This is unsurprising as $U_t^{(3)}$ is of the form $\tilde{\sigma}_t^2 Z_t^2$, where Z_t^2 are the true squared residuals, and $\tilde{\sigma}_t^2$ is a non-negative random variable (that resembles the conditional variance).

To alleviate the above rightward skew, and to bring the model closer to additive, we consider our final transformation

$$U_t^{(4)} = \log(\varepsilon + U_t^{(3)}), \tag{3}$$

where, for simplicity, the default value of ε is as in $U_t^{(3)}$ above. Note that we cannot simply use $\log(U_t^{(3)})$ as one of the requirements on the function $g(\cdot)$ is that it should be bounded (since $U_t^{(3)}$ is bounded and non-negative and $\varepsilon > 0$, it follows that $U_t^{(4)}$ is bounded).

We note that the transformation $U_t^{(4)}$ is invertible, i.e. X_t^2 can be recovered from it by applying the inverse transformation. This implies that any changes in the joint distribution of X_t^2 (i.e. changes in the time varying ARCH parameters) must be recoverable by examining the joint distribution of $U_t^{(4)}$. BASTA-res proceeds by searching for changes in the mean of $U_t^{(4)}$, rather than in the complete (joint) distribution, and there are good reasons for this simplification. Firstly, $U_t^{(4)}$ is specifically constructed to have less auto-correlation than the original process X_t^2. Secondly, the logarithmic transformation in $U_t^{(4)}$ is designed to stabilize the variance in this sequence, i.e. to make it more homogeneous. Altogether, the hope is that this brings $U_t^{(4)}$ close to a 'function plus independent and identically distributed noise' set-up, in which any changes in the joint distribution would have to be reflected in changes in the mean of $U_t^{(4)}$, which is what BASTA-res looks for. Although this argument for considering the mean only is merely heuristic, we feel that it is vindicated by the good empirical performance of BASTA-res. In our simulations described later, both $U_t^{(3)}$ and $U_t^{(4)}$ are used.

4.2.1. Default choice of $\mathbf{C}$

We propose the following default choice of the vector constant $\mathbf{C}$ in our transformations (2) and (3). In the first stage, we (not necessarily correctly) act as if $\{X_t\}_{t=1}^n$ were a realization of a stationary ARCH process with parameters $a_0, \ldots, a_p$ and follow a normalized least squares procedure (Fryzlewicz *et al.*, 2008) to estimate the values of $a_0, \ldots, a_p$ as $\hat{a}_0, \ldots, \hat{a}_p$. If $\{X_t\}_{t=1}^n$ indeed happened to be stationary, i.e. contained no change points, the computed values $\hat{a}_0, \ldots, \hat{a}_p$ would then form meaningful estimates of the true parameters $a_0, \ldots, a_p$.

Thus, in the null hypothesis of no change points present, if we were to set $C_i := \hat{a}_i$ for $i = 0, \ldots, p$, the corresponding transformed sequences $U_t^{(3)}$ and $U_t^{(4)}$ would indeed be close to the (squared, and squared and logged respectively) empirical residuals from the model, as explained above. The hope is that our change point detection procedure would then correctly react to this construction by determining that no change points were present in the model. However, rather than directly setting $C_i := \hat{a}_i$ for $i = 0, \ldots, p$, we add extra flexibility to our construction by introducing a positive factor $F \geqslant 1$, which we use to 'dampen' the values of the constants $C_1, \ldots, C_p$ as follows:

$$C_0 := \hat{a}_0,$$
$$C_i := \hat{a}_i / F, \qquad i = 1, \ldots, p.$$

The effect of this above dampening of the values of $C_1, \ldots, C_p$ is that, as F increases, $U_t^{(3)}$ is, up to a multiplicative constant, increasingly closer to X_t^2 itself. Indeed, in the limit as $F \to \infty$, we have

$$U_t^{(3)} \approx \frac{X_t^2}{C_0 + \varepsilon X_t^2}$$

(bear in mind that the default value of ε is small). Empirical evidence suggests that larger values of F can lead to better exposure of change points in the alternative hypothesis of change points being present, at the expense of introducing a higher degree of auto-correlation and thicker tails in the empirical distribution of $U_t^{(3)}$. Naturally, this also applies to $U_t^{(4)}$, but to a less extent.

Since, typically, higher values of F will lead to better exposure of change points but will also introduce higher auto-correlation, it is desirable to choose F to obtain a trade-off between these two trends. Section 4.2.2 will discuss the proposed default choice of F based on an extensive simulation study.

4.2.2. Default choice of F and c through simulation

A simulation study was performed in which we assessed the empirical performance of our procedure for a variety of ARCH(1) and ARCH(2) models with piecewise constant parameters and various sample sizes. We mention that our empirical experience suggests that time varying ARCH processes of order up to 2 are typically sufficient to model and forecast a wide range of low frequency returns well; see for example Fryzlewicz *et al.* (2008). The dampening constant F (see Section 4.2.1) ranged from 1 to 10, and the threshold constant c (see Section 3.2) ranged from 0.1 to 1. The number of change points ranged from 0 to 2, and, if present, they were located two-thirds and a third the way through the series. Sample sizes varied from $n = 750$ to $n = 3000$.

It was found that the algorithm based on the sequence $U_t^{(4)}$ performed better than that based on $U_t^{(3)}$: this was because the 'noise' in $U_t^{(4)}$ has a more homogeneous structure due to the use of the log-transform. This is unsurprising: recall that $U_t^{(3)}$ is of the form $\tilde{\sigma}_t^2 Z_t^2$, where Z_t^2 are the true squared residuals, and $\tilde{\sigma}_t^2$ is a non-negative random variable. Therefore, the logarithmic transformation in $U_t^{(4)}$ brings the model close to the additive model $\log(\tilde{\sigma}_t^2) + \log(Z_t^2)$, in which the noise $\log(Z_t^2)$ has a constant variance. Hence, our threshold $\tilde{b}_T$, whose magnitude does not vary locally with t, can be expected to offer better performance for the more homogeneous model $U_t^{(4)}$, although we emphasize that our method is consistent for both $U_t^{(3)}$ and $U_t^{(4)}$.

Performance was surprisingly robust across all models tested with respect to the choice of F. We found that the value of c ranging in the interval $[0.4, 0.6]$ was the best choice in terms of the probability of correctly detecting the true number of change points. The obvious exceptions were 'null hypothesis' models not containing change points, for which, as expected, higher values of

c resulted in better performance than lower values. Below, we provide details of the models that were used:

(a) an ARCH(1) model with one change point, two-thirds the way (a_1 is constant and equal to 0.7, and a_0 changes from 1 to 1.5, 2.0 or 2.5);
(b) an ARCH(1) model with one change point, two-thirds the way (a_0 is constant and equal to 1, and a_1 changes from 0.7 to 0, 0.3 or 0.9);
(c) an ARCH(1) model with one change point, two-thirds the way (a_1 changes from 0.7 to 0, 0.3 or 0.9, and a_0 also changes at the same time point in such a way that the unconditional variance of the process remains constant throughout);
(d) an ARCH(1) model with no change points (a_0 equals 1, and a_1 is set equal to 0, 0.5 or 0.9);
(e) an ARCH(2) model, with two change points, occurring a third the way (in a_2 only) and two-thirds the way (in a_1 only) ($a_0 = 1$ throughout; if a_1 changes from α to β, then a_2 changes from β to α; the values of (α, β) are $(0, 0.7)$, $(0.2, 0.6)$ or $(0.4, 0.1)$).

In the colour 'maps' of Figs 1 and 2, the lighter the colour, the higher the proportion (over 100 simulations) of correctly detected numbers of change points for the various models, with $U_t^{(4)}$. Figs 1(a)–1(c) show maps averaged over models (a), (b), (c) and (e) (and the various changes in a_0 and a_1) for sample sizes 750 (Fig. 1(a)), 1500 (Fig. 1(b)) and 3000 (Fig. 1(c)). Figs 1(d)–1(f) show maps averaged over model (d) (and the three values of a_1); the sample sizes correspond. The difference in the patterns between the columns is explained by the fact that, for the no-change-point model (d), the higher the threshold, the better.

Fig. 2(a) shows results averaged over models (a), (b), (c) and (e) (and the various changes in a_0 and a_1) and averaged over sample sizes 750, 1500 and 3000. Fig. 2(b) shows a similar result averaged over model (d) (and the three values of a_1) and sample sizes 750, 1500 and 3000.

From the results, it appears that the configuration $(c, F) = (0.5, 8)$ is a sensible default choice. However, in practice it may be beneficial to apply an extra 'correction' and to use a slightly lower threshold for higher sample sizes and a slightly higher threshold for lower sample sizes. This is because, as indicated in Fig. 1, the constant $c = 0.4$ yields the best results for sample size 3000, $c = 0.5$ is the best for sample size 1500 and $c = 0.6$ is the best for sample size 750. This is not surprising as it must be borne in mind that our simulations use a threshold $\tilde{b}_T$ of order $T^{3/8}$, whereas the maximum permitted range of the exponent θ in $\tilde{b}_T$ is $\theta \in (\frac{1}{4}, \frac{1}{2})$. Applying the extra correction would correspond to choosing a slightly lower exponent θ in the threshold. To summarize, we issue the following practical recommendation: use $(c, F) = (0.6, 8)$ for sample sizes of up to 1000, $(c, F) = (0.5, 8)$ for sample sizes of between 1000 and 2000, and $(c, F) = (0.4, 8)$ for sample sizes of between 2000 and 3000. For longer series, we recommend applying the procedure to segments of length up to 3000, rather than to the entire series at once. We emphasize that BASTA-res is a completely specified procedure in the sense that we provide default values for all of its parameters.

4.3. BASTA-avg: BASTA based on subsampled local averages

We now describe the construction of the BASTA-avg algorithm, which is a simpler alternative to BASTA-res, that requires the choice of fewer parameters than the latter. We earlier argued that diagonal transformations in which $g(\cdot)$ was a function of X_t^2 only would not be suitable for our purpose as they preserved the high degree of auto-correlation that is normally present in the process X_t^2. However, non-overlapping local averages of the process X_t^2 are an interesting candidate for our transformation $g(\cdot)$ as they exhibit less auto-correlation and have lighter tails than X_t^2. More formally, we take

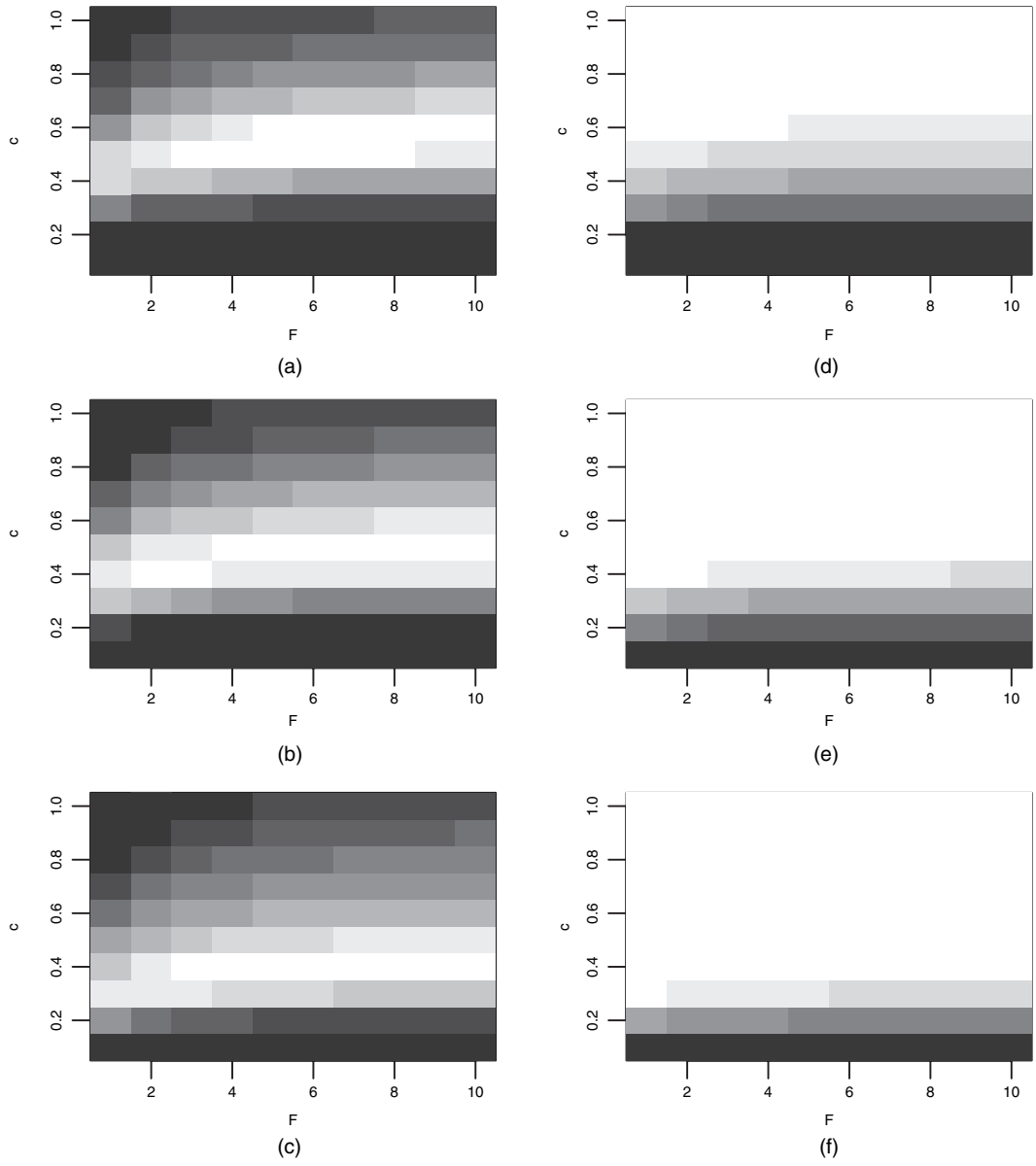

Fig. 1. Maps of correctly detected numbers of change points in BASTA-res, depending on F and c: see Section 4.2.2 for a description

$$U_t^{(5)} = \log\left\{\min\left(\frac{1}{s}\sum_{j=ts}^{(t+1)s-1} X_j^2 + \varepsilon, M\right)\right\}$$

(which is simply a bounded version of $\log(s^{-1}\Sigma_{j=ts}^{(t+1)s-1} X_j^2)$). The effective sample size for $U_t^{(5)}$ is T/s but, since s is a constant, this is still of order $O(T)$ and the rates in theorem 1 are unaffected. We now investigate the performance of our binary segmentation procedure on the sequence $U_t^{(5)}$. Always normalizing our process X_t so that its sample variance is 1, we set ε equal to 10^{-3}

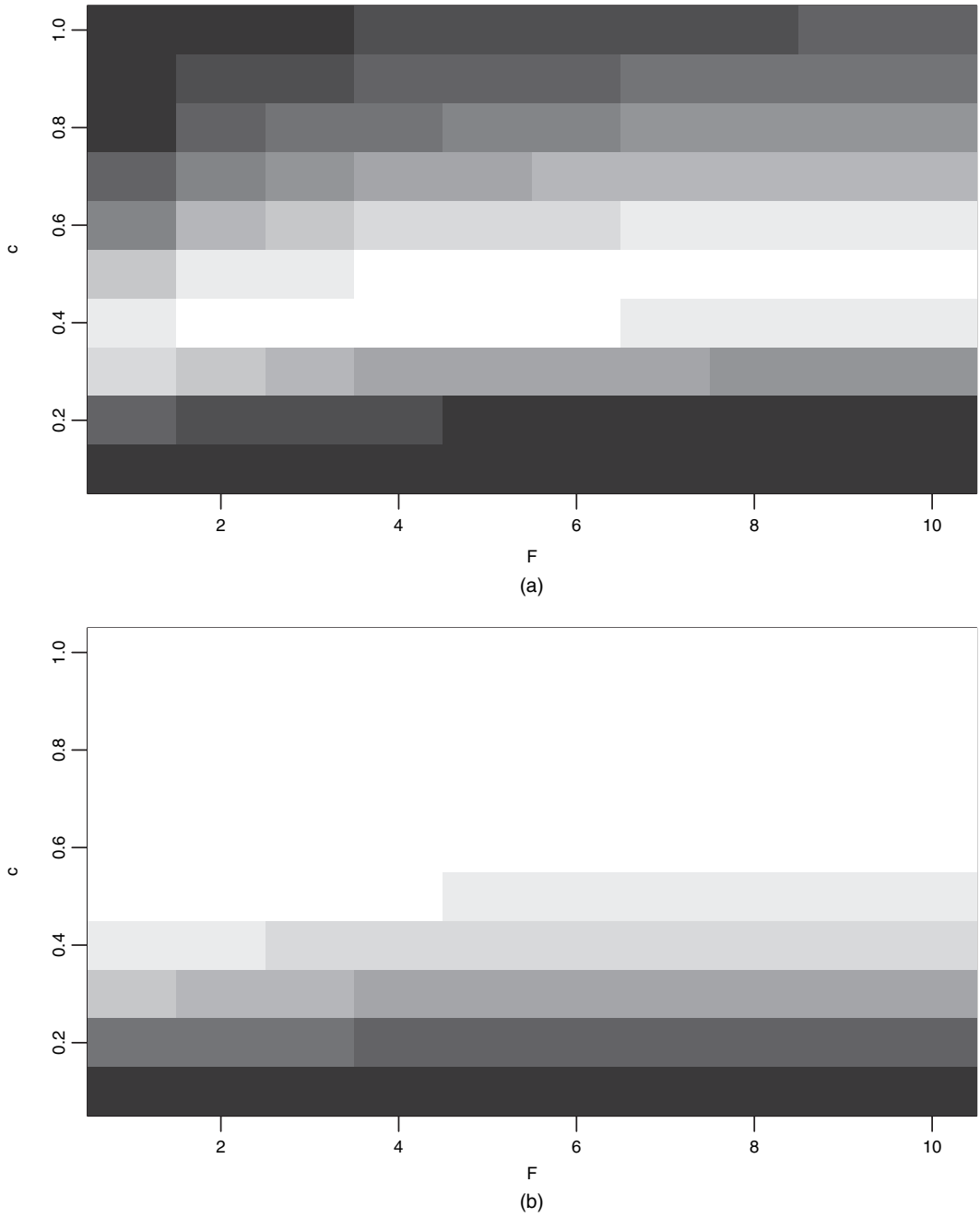

Fig. 2. Maps of correctly detected numbers of change points in BASTA-res, depending on F and c; see Section 4.2.2 for a description

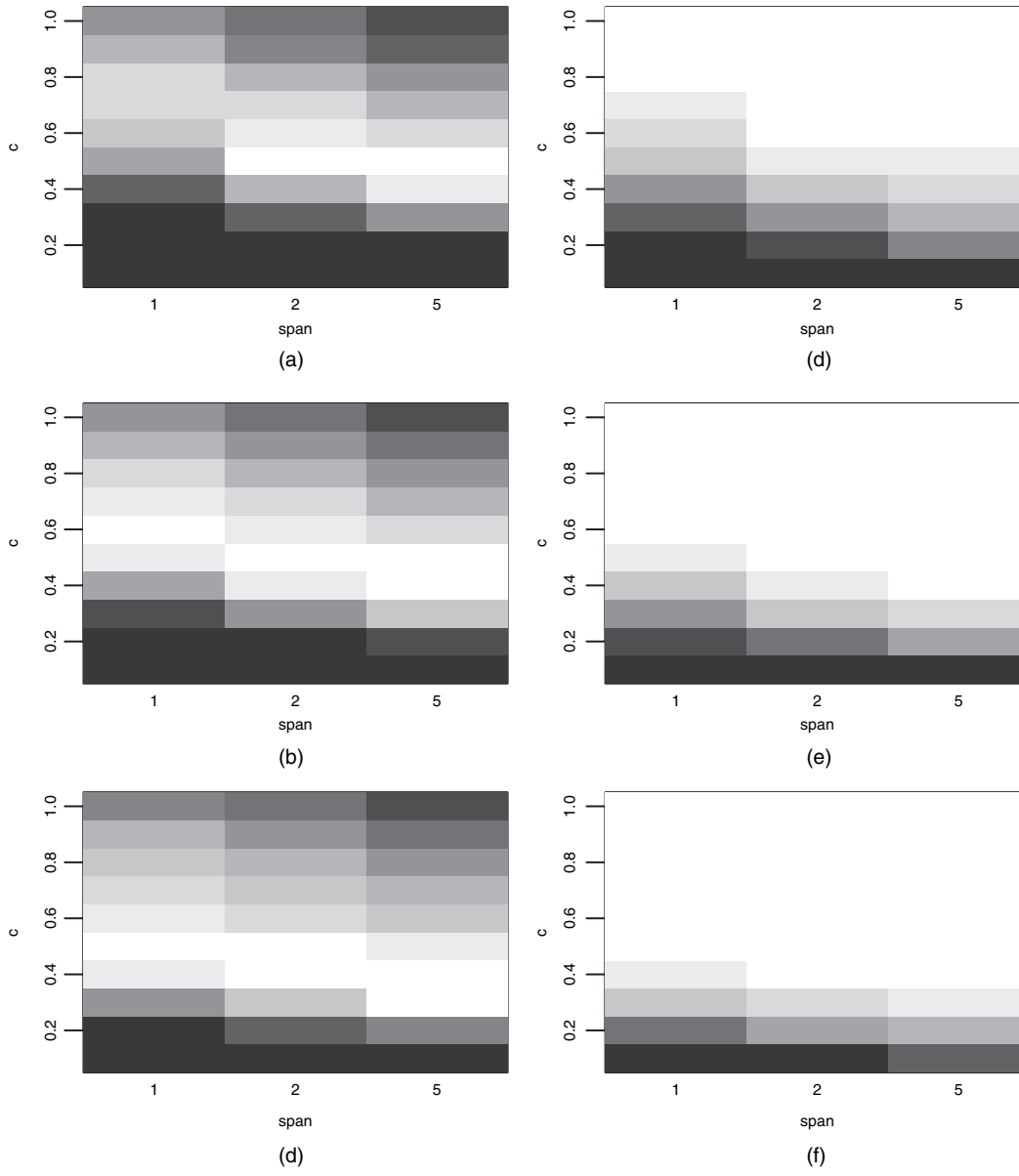

Fig. 3. Maps of correctly detected numbers of change points in BASTA-avg, depending on span s and c; see Section 4.3 for details

as in BASTA-res; the ceiling M is set equal to 10. These two parameters do not seem to have much effect on the practical performance of the procedure and we do not dwell on their choice in this work. As before, the threshold exponent constant θ is taken to be $\frac{3}{8}$: this is in the middle of the maximum permitted range $(\frac{1}{4}, \frac{1}{2})$ and, not unexpectedly, was found to perform the best in our numerical experiments. There remains the issue of choosing the span constant s and the threshold scaling constant c. We first examine the performance of the new procedure for a range of these two parameters on the training models (a)–(e) from Section 4.2.2.

Fig. 3 is analogous to Fig. 1 for BASTA-res: it visualizes the performance of BASTA-avg averaged over the non-stationary models (a), (b), (c) and (e) (Figs 3(a)–3(c)) and for the stationary model (d) (Figs 3(d)–3(f)), for sample sizes 750 (Figs 3(a) and 3(d)), 1500 (Figs 3(b) and 3(e)) and 3000 (Figs 3(c) and 3(f)). Lighter colour in each image means that the correct number of change points was detected a large proportion of times (over 100 runs).

As expected, span $s = 1$ (equivalent to no averaging at all) does not yield good performance, compared with spans 2 or 5. For the latter spans, we can observe that the best performance occurs for values of c roughly around 0.5, although the 'best' value seems to be slightly lower for $s = 5$ and for longer data sets.

Finally, we re-emphasize that our theory does not permit a data-dependent choice of the constants c and θ, so it is important to have reliable default constant values at our disposal. We have found it reassuring that, although the data transformations in BASTA-res and BASTA-avg are constructed in two completely different ways, values of (c, θ) close to $(0.5, \frac{3}{8})$ have been found to perform the best for both of these algorithms.

5. Performance evaluation

In this comparative simulation study, we use our algorithms, BASTA-res and BASTA-avg, to re-examine the examples of GARCH processes that were reported in Davis *et al.* (2008), which appears to be the state of the art procedure for change point detection in GARCH models. We recall that a process Y_t follows a GARCH(p,q) model if it is defined as in expression (1) except that σ_t^2 is defined as

$$\sigma_t^2 = a_0 + \sum_{i=1}^{p} a_i Y_{t-i}^2 + \sum_{j=1}^{q} b_j \sigma_{t-j}^2.$$

Among other models, Davis *et al.* (2008) considered 10 GARCH(1,1) models with sample size $n = 1000$, and with at most one change point occurring in the triple (a_0, a_1, b_1) at time $t = 501$ as follows:

(a) $(0.4, 0.1, 0.5) \to (0.4, 0.1, 0.5)$ (note that this model is stationary);
(b) $(0.1, 0.1, 0.8) \to (0.1, 0.1, 0.8)$ (note that this model is stationary);
(c) $(0.4, 0.1, 0.5) \to (0.4, 0.1, 0.6)$;
(d) $(0.4, 0.1, 0.5) \to (0.4, 0.1, 0.8)$;
(e) $(0.1, 0.1, 0.8) \to (0.1, 0.1, 0.7)$;
(f) $(0.1, 0.1, 0.8) \to (0.1, 0.1, 0.4)$;
(g) $(0.4, 0.1, 0.5) \to (0.5, 0.1, 0.5)$;
(h) $(0.4, 0.1, 0.5) \to (0.8, 0.1, 0.5)$;
(i) $(0.1, 0.1, 0.8) \to (0.3, 0.1, 0.8)$;
(j) $(0.1, 0.1, 0.8) \to (0.5, 0.1, 0.8)$.

Table 1 shows the proportion of simulation runs for which the correct number of change points (0 for models (a) and (b); 1 for the rest) has been detected, for three competing methods: that of Andreou and Ghysels (2002), that of Davis *et al.* (2008) and ours (BASTA-res and BASTA-avg). (The results for the Andreou–Ghysels method have been taken from Davis *et al.* (2008).) BASTA-res used the default values $c = 0.6$ and $F = 8$ (as recommended in Section 4.2.2), was based on the sequence $U_t^{(4)}$ and used order $p = 1$. BASTA-avg used two pairs of values for s and c: $(s, c) = (2, 0.5)$ (BASTA-avg1 in Table 1) and $(5, 0.4)$ (BASTA-avg2). 100 simulation runs were performed. We also tried the method from Lavielle and Teyssiere (2005) (using the MATLAB implementation DCPC) and, although we found that its performance was

Table 1. Proportion of times that the correct number of change points was detected in models (a)–(j), as well as on average across all models for the three competing methods

Method	*Results for the following models:*										*Average*
	(a)	*(b)*	*(c)*	*(d)*	*(e)*	*(f)*	*(g)*	*(h)*	*(i)*	*(j)*	
Davis *et al.* (2008)	0.96	0.96	0.19	0.96	0.63	0.98	0.12	0.91	0.91	0.95	0.757
Andreou and Ghysels (2002)	0.96	0.88	0.24	0.95	0.75	0.72	0.14	0.94	0.94	0.86	0.738
BASTA-res	0.98	0.93	0.25	0.94	0.75	0.95	0.18	0.90	0.96	0.93	0.777
BASTA-avg1	0.98	0.97	0.17	0.91	0.88	0.91	0.07	0.96	0.86	0.92	0.763
BASTA-avg2	0.98	0.86	0.29	0.92	0.91	0.89	0.11	0.99	0.9	0.85	0.77

good, it was overall substantially inferior to all the above methods, so we do not report details here.

Although BASTA was not always the best of the three methods, we note that it was always either the best or came very close to the best, including in models where there was a large difference in performance between the best and the worst. Indeed, BASTA-res achieved the highest average correct proportion across all models tested. Despite its simplicity, BASTA-avg also performed very well for both parameter sets, with the overall results placing it just behind BASTA-res.

6. FTSE 100 index analysis

In this section, we apply our BASTA-res technique to the series of differenced closing values of the FTSE 100 index: the share index of the 100 most highly capitalized UK companies listed on the London Stock Exchange, with the aim of investigating whether and how any detected change points correspond to the milestones of the recent financial crisis. The series has 1000 observations ranging from July 27th, 2005, to July 13th, 2009, i.e. roughly 4 trading years. As before, our method used the default values as recommended in Section 4.2.2, was based on the sequence $U_t^{(4)}$ and used order $p = 1$.

It is fascinating to observe that the estimated change points, which are shown in Fig. 4, do indeed correspond to important events in the recent financial crisis. More precisely, the estimated change points are as follows.

(a) $t = 467$, corresponding to June 5th, 2007: the summer of 2007 is widely regarded as the start of the subprime mortgage hedge fund crisis, with the major investment bank Bear Stearns revealing, in July 2007, that their two subprime hedge funds had lost nearly all of their value.

(b) $t = 773$, corresponding to August 18th, 2008: it is probably safe to attribute this estimated change point to the collapse of Lehman Brothers, a major financial services firm.

(c) $t = 850$, corresponding to December 4th, 2008: although it is difficult to attribute this date to a specific event, we point out that the end of the year 2008 was the time when governments, national banks and international institutions such as the International Monetary Fund announced and began to implement a range of financial measures to help the ailing world economy.

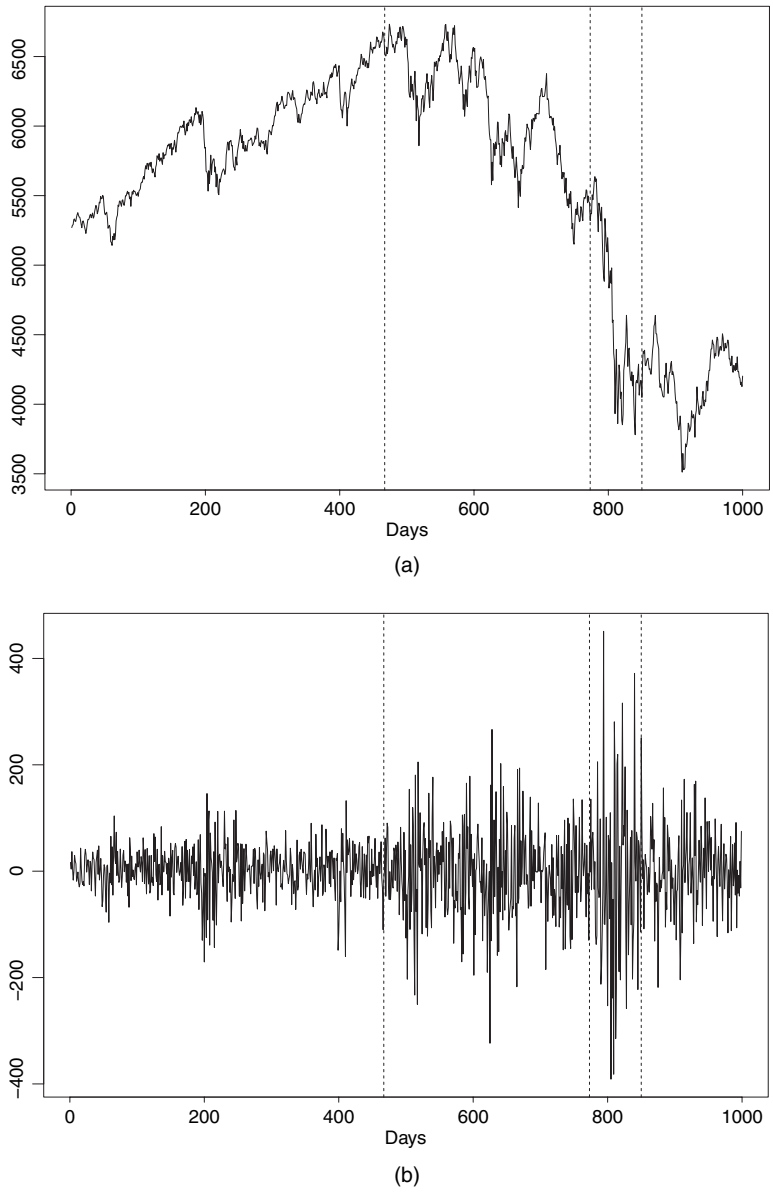

Fig. 4. (a) Closing values of the FTSE 100 index from July 27th, 2005, to July 13th, 2009 (1000 observations: roughly 4 trading years) and (b) differenced values of the index over this period (see Section 6 for comment): ⋮, change points detected by BASTA-res on the bottom series as input

Appendix A: Proof of theorem 1

The lemmas below are proved under assumption 1.

We first collect some of the definitions given in the main section. Let $\tilde{X}_t^i$ satisfy $\tilde{X}_t^i = \tilde{\sigma}_t^i Z_t$ where

$$(\tilde{\sigma}_t^i)^2 = a_0(\eta_i) + \sum_{j=1}^{p} a_j(\eta_i)(\tilde{X}_{t-j}^i)^2, \qquad t \leqslant T-1. \tag{4}$$

Let $\upsilon(t)$ be the index i of the closest change point η_i less than or equal to t and $\eta(t)$ be the location of the largest change point less than or equal to t ($\eta_i \leqslant t$). In the following lemma we show that the piecewise constant squared ARCH process X_t^2 is 'close' to $(\tilde{X}_t^{\upsilon(t)})^2$.

Lemma 1. Let X_t and $\tilde{X}_t^i$ be defined as in expression (1) and (4) respectively; then we have

$$|X_t^2 - (\tilde{X}_t^{\upsilon(t)})^2| \leqslant V_t,$$

where $\mathbb{E}(V_t) \leqslant C\rho^{t-\eta(t)}$, with $0 < \rho < 1$ and C being some constants independent of t.

Proof. As the proof involves only the squared ARCH processes, to reduce cumbersome notation we let $\xi_t = X_t^2$ and $\tilde{\xi}_t^{\upsilon(t)} = (\tilde{X}_t^{\upsilon(t)})^2$. Let $[\cdot]_i$ denote the ith element of a vector. For a generic squared ARCH(p) process $Y_t = Z_t^2\{\alpha_0(t) + \Sigma_{j=1}^p \alpha_j(t) Y_{t-j}\}$ (be it time varying or not) iterating k steps backwards gives

$$Y_t = Z_t^2\{P_{k,t}^Y(\mathbf{Z}_{t,t-k}) + Q_{k,t}^Y(\mathbf{Y}_{t-k})\},$$

where $\mathbf{Z}_{t,t-k} = (Z_t^2, \ldots, Z_{t-k+1}^2)$, $\mathbf{Y}_{t-k} = (Y_{t-k}, \ldots, Y_{t-k-p+1})$,

$$P_{k,t}^Y(\mathbf{Z}_{t,t-k}) = \alpha_0(t) + \left[A_t \sum_{r=0}^{t-k} \prod_{j=1}^{r} \tilde{A}_{t-j} \mathbf{b}_{t-r-1}\right]_1,$$

$$Q_{k,t}^Y(\mathbf{Y}_{t-k}) = \left[A_t \prod_{j=1}^{k-1} \tilde{A}_{t-j} \mathbf{Y}_{t-k}\right]_1$$

and

$$\tilde{A}_t = \begin{pmatrix} \alpha_1(t)Z_t^2 & \alpha_2(t)Z_t^2 & \ldots & \alpha_p(t)Z_t^2 \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ & & \ddots & \vdots \\ \ldots & \ldots & & \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$$\mathbf{b}_t = (\alpha_0(t)Z_t^2, 0, \ldots, 0)'$$

and $A_t = \mathbb{E}(\tilde{A}_t)$. We now consider the above expansion for both ξ_t and $\tilde{\xi}_t^{\upsilon(t)}$. Using the above notation and iterating ξ_t backwards $k = t - \eta(t)$ steps (i.e. to its nearest change point from below) gives

$$\xi_t = Z_t^2\{P_{t-\eta(t),t}^{\xi}(\mathbf{Z}_{t,\eta(t)}) + Q_{t-\eta(t),t}^{\xi}(\boldsymbol{\xi}_{\eta(t)})\}$$

and a similar expansion of $\tilde{\xi}_t^{\upsilon(t)}$ yields

$$\tilde{\xi}_t^{\upsilon(t)} = Z_t^2\{P_{t-\eta(t),t}^{\tilde{\xi}^{\upsilon(t)}}(\mathbf{Z}_{t,\eta(t)}) + Q_{t-\eta(t),t}^{\tilde{\xi}^{\upsilon(t)}}(\tilde{\boldsymbol{\xi}}_{\eta(t)}^{\upsilon(t)})\}.$$

Recalling that both ξ_t and $\tilde{\xi}_t^{\upsilon(t)}$ share the same time varying coefficients on $\eta(t), \ldots, t$ and the same innovation sequence $\{Z_t\}_t$ we have $P_{t-\eta(t),t}^{\xi}(\mathbf{Z}_{t,\eta(t)}) = P_{t-\eta(t),t}^{\tilde{\xi}^{\upsilon(t)}}(\mathbf{Z}_{t,\eta(t)})$. Thus, taking differences, we have

$$\xi_t - \tilde{\xi}_t^{\upsilon(t)} = Z_t^2\{Q_{t-\eta(t),t}(\boldsymbol{\xi}_{\eta(t)}) - Q_{t-\eta(t),t}(\tilde{\boldsymbol{\xi}}_{\eta(t)}^{\upsilon(t)})\}.$$

Define the positive random variable $V_t = Z_t^2\{Q_{t-\eta(t),t}(\boldsymbol{\xi}_{\eta(t)}) + Q_{t-\eta(t),t}(\tilde{\boldsymbol{\xi}}_{\eta(t)}^{\upsilon(t)})\}$; then it is clear that $|\xi_t - \tilde{\xi}_t^{\upsilon(t)}| \leqslant V_t$. Therefore, since $A_t, \ldots, A_{t-\eta(t)}$ share the same ARCH coefficients, we have $\mathbb{E}(V_t) = [A_t^{t-\eta(t)}\mathbb{E}(\boldsymbol{\xi}_{t-k})]_1 + [A_t^{t-\eta(t)}\mathbb{E}(\tilde{\boldsymbol{\xi}}_{t-k}^{\upsilon(t)})]_1$ and thus $\mathbb{E}(V_t) \leqslant \|A_t^{t-\eta(t)}\mathbb{E}(\boldsymbol{\xi}_{t-k})\|_1 + \|A_t^{t-\eta(t)}\mathbb{E}(\tilde{\boldsymbol{\xi}}_{t-k}^{\upsilon(t)})\|_1$. Since the entries of A_t are non-negative, if x and y are column vectors of length p such that $\mathbf{0} \leqslant x \leqslant y$ componentwise (where $\mathbf{0}$ is a column vector of 0s, of length p), then $A_t^{t-\eta(t)}x \leqslant A_t^{t-\eta(t)}y$ componentwise. From assumptions (f) and (g) of theorem 1, both $\mathbb{E}(\boldsymbol{\xi}_{t-k})$ and $\mathbb{E}(\tilde{\boldsymbol{\xi}}_{t-k}^{\upsilon(t)})$ are bounded from above, componentwise, by the vector $C/\delta_1\mathbf{1}$, where $\mathbf{1}$ is a column vector of 1s, of length p. Therefore $\mathbb{E}(V_t) \leqslant C\|A_t^{t-\eta(t)}\mathbf{1}\|_1$ for some constant C. But from the form of the matrix A_t, and using assumption (f) again, it is clear that $A_t^p\mathbf{1} \leqslant (1-\delta_1)\mathbf{1}$ componentwise. This eventually leads to $\mathbb{E}(V_t) \leqslant C\{(1-\delta_1)^{1/p}\}^{t-\eta(t)}$, as required, which completes the proof of 1emma 1.

In the remainder of the proof of theorem 1, we use the decomposition

$$U_t = g_t + \varepsilon_t, \qquad t = 0, \ldots, T-1,$$

noting that $|\varepsilon_t| \leqslant 2\bar{g}$, where $\bar{g} = \sup_t |U_t|$. In general, $\mathbb{E}(\varepsilon_t) \neq 0$, but it is close to 0 in the sense that is explained in lemma 4 below. Since X_t is a strongly α-mixing process with a geometric rate ρ, where $1 - \delta_1 < \rho < 1$ (see

theorem 3.1 in Fryzlewicz and Subba Rao (2011)), so are U_t and ε_t, with the same rates (see for example theorem 14.1 of Davidson (1994)).

Let s and u satisfy $\eta_{p_0} \leqslant s < \eta_{p_0+1} < \ldots < \eta_{p_0+q} < u \leqslant \eta_{p_0+q+1}$ for $0 \leqslant p_0 \leqslant N-q$, which will always be so at all stages of the algorithm. Denoting $n = u-s+1$, we define

$$\tilde{U}^b_{s,u} = \frac{\sqrt{(u-b)}}{\sqrt{\{n(b-s+1)\}}} \sum_{t=s}^{b} U_t - \frac{\sqrt{(b-s+1)}}{\sqrt{\{n(u-b)\}}} \sum_{t=b+1}^{u} U_t,$$

$$\tilde{g}^b_{s,u} = \frac{\sqrt{(u-b)}}{\sqrt{\{n(b-s+1)\}}} \sum_{t=s}^{b} g_t - \frac{\sqrt{(b-s+1)}}{\sqrt{\{n(u-b)\}}} \sum_{t=b+1}^{u} g_t,$$

where b satisfies $s \leqslant b < u$.

In lemmas 2–6 below, we impose at least one of the following conditions:

$$s < \eta_{p_0+r} - C\delta_T < \eta_{p_0+r} + C\delta_T < u \qquad \text{for some } 1 \leqslant r \leqslant q; \tag{5}$$

$$\max\{\min(\eta_{p_0+1} - s, s - \eta_{p_0}),\ \min(\eta_{p_0+q+1} - u, u - \eta_{p_0+q})\} \leqslant C\varepsilon_T. \tag{6}$$

Both condition (5) and condition (6) hold throughout the algorithm for all those segments starting at s and ending at u which contain previously undetected change points. As lemma 7 concerns the case when all change points have been detected, it does not use either of these conditions.

The structure of the proof is as follows: lemma 2 is used in lemma 3; lemma 1 in 4; lemmas 3 and 4 in 5; lemma 5 in 6; lemmas 6 and 7 prove theorem 1.

Lemma 2. Let s and u satisfy condition (5); then there exists $1 \leqslant r^* \leqslant q$ such that

$$|\tilde{g}^{\eta_{p_0+r^*}}_{s,u}| = \max_{s<t<u} |\tilde{g}^t_{s,u}| \geqslant C\delta_T T^{-1/2}. \tag{7}$$

Proof. The equality in expression (7) is the exact statement of lemma 2.3 of Venkatraman (1992). For the inequality part, we note that, in the case of a single change point in g_t, r in condition (5) coincides with r^* and we can use the constancy of g_t to the left and to the right of the change point to show that

$$|\tilde{g}^{\eta_{p_0+r}}_{s,u}| = \left| \frac{\sqrt{(\eta_{p_0+r} - s + 1)}\sqrt{(u - \eta_{p_0+r})}}{\sqrt{n}} (g_{\eta_{p_0+r}} - g_{\eta_{p_0+r}+1}) \right|,$$

which is bounded from below by $C\delta_T T^{-1/2}$. In the case of multiple change points, we remark that, for any r satisfying condition (5), the above order remains the same and thus result (7) follows.

Lemma 3. Suppose that condition (5) holds, and further assume that $\tilde{g}^{\eta_{p_0+r}}_{s,u} > 0$ for some $1 \leqslant r \leqslant q$. Then for any k satisfying $|\eta_{p_0+r} - k| = C_0\varepsilon_T$ and $\tilde{g}^{\eta_{p_0+r}}_{s,u} > \tilde{g}^k_{s,u}$ we have, for sufficiently large T, $\tilde{g}^{\eta_{p_0+r}}_{s,u} \geqslant \tilde{g}^k_{s,u} + CC_0\varepsilon_T T^{-1/2}$.

Proof. Without loss of generality, assume that $\eta_{p_0+r} < k$. As in lemma 2, we first derive the result in the case of a single change point in g_t. The following equations hold:

$$\tilde{g}^k_{s,u} = \frac{\sqrt{(\eta_{p_0+r} - s + 1)}\sqrt{(u-k)}}{\sqrt{(u - \eta_{p_0+r})}\sqrt{(k-s+1)}} \tilde{g}^{\eta_{p_0+r}}_{s,u},$$

and

$$\begin{aligned}
\tilde{g}^{\eta_{p_0+r}}_{s,u} - \tilde{g}^k_{s,u} &= \left\{ 1 - \frac{\sqrt{(\eta_{p_0+r} - s + 1)}\sqrt{(u-k)}}{\sqrt{(u - \eta_{p_0+r})}\sqrt{(k-s+1)}} \right\} \tilde{g}^{\eta_{p_0+r}}_{s,u} \\
&= \frac{\sqrt{\left(1 + \dfrac{k - \eta_{p_0+r}}{\eta_{p_0+r} - s + 1}\right)} - \sqrt{\left(1 - \dfrac{k - \eta_{p_0+r}}{u - \eta_{p_0+r}}\right)}}{\sqrt{\left(1 + \dfrac{k - \eta_{p_0+r}}{\eta_{p_0+r} - s + 1}\right)}} \tilde{g}^{\eta_{p_0+r}}_{s,u} \\
&\geqslant \frac{(1 + c_1 C_0 \varepsilon_T/2\delta_T) - (1 + c_2 C_0 \varepsilon_T/2\delta_T) + o(\varepsilon_T/\delta_T)}{\sqrt{2}} \tilde{g}^{\eta_{p_0+r}}_{s,u} \geqslant CC_0 \frac{\varepsilon_T}{\delta_T} \frac{\delta_T}{T^{1/2}} = CC_0 \varepsilon_T T^{-1/2}.
\end{aligned}$$

Lemma 4. Define

$$B_T = \left\{ \max_{s,b,u} |\tilde{U}^b_{s,u} - \tilde{g}^b_{s,u}| \leqslant C \log^{\alpha}(T) \right\}.$$

We have $P(B_T) \to 1$ for some positive α and C.

Proof. We denote

$$\tilde{U}_{s,b} = \frac{\sqrt{(u-b)}}{\sqrt{\{n(b-s+1)\}}} \sum_{t=s}^{b} U_t, \qquad \tilde{U}_{b,u} = \frac{\sqrt{(b-s+1)}}{\sqrt{\{n(u-b)\}}} \sum_{t=b+1}^{u} U_t,$$

$$\tilde{g}_{s,b} = \frac{\sqrt{(u-b)}}{\sqrt{\{n(b-s+1)\}}} \sum_{t=s}^{b} g_t, \qquad \tilde{g}_{b,u} = \frac{\sqrt{(b-s+1)}}{\sqrt{\{n(u-b)\}}} \sum_{t=b+1}^{u} g_t,$$

so that $\tilde{U}^b_{s,u} = \tilde{U}_{s,b} - \tilde{U}_{b,u}$ and $\tilde{g}^b_{s,u} = \tilde{g}_{s,b} - \tilde{g}_{b,u}$. We have

$$\begin{aligned} P\{\max_{s,b,u} |\tilde{U}^b_{s,u} - \tilde{g}^b_{s,u}| \geqslant \lambda\} &\leqslant P\{\max_{s,b,u} |\tilde{U}_{s,b} - \tilde{g}_{s,b}| + |\tilde{U}_{b,u} - \tilde{g}_{b,u}| \geqslant \lambda\} \\ &\leqslant P\{\max_{s,b,u} |\tilde{U}_{s,b} - \tilde{g}_{s,b}| \geqslant \lambda/2\} + P\{\max_{s,b,u} |\tilde{U}_{b,u} - \tilde{g}_{b,u}| \geqslant \lambda/2\} \\ &\leqslant 2P\{\max_{s,b,u} |\tilde{U}_{s,b} - \tilde{g}_{s,b}| \geqslant \lambda/2\}. \end{aligned}$$

We now bound the above probability in two different ways depending on the difference $b-s$.

(a) $b-s$ 'small': we have

$$\tilde{U}_{s,b} - \tilde{g}_{s,b} = \frac{\sqrt{(u-b)}}{\sqrt{\{n(b-s+1)\}}} \sum_{t=s}^{b} \varepsilon_t;$$

note that

$$\frac{\sqrt{(u-b)}}{\sqrt{\{n(b-s+1)\}}} = \frac{\sqrt{(u-b)}}{\sqrt{(u-s+1)}} \nu^{-1/2} \leqslant \nu^{-1/2},$$

where $\nu = b-s+1$. We bound $|\tilde{U}_{s,b} - \tilde{g}_{s,b}| \leqslant 2\bar{g}\nu^{1/2}$, which does not exceed $\lambda/2$ as long as $\nu \leqslant \lambda^2/(16\bar{g}^2)$ (where λ is logarithmic, which will be established below), which defines what we mean by a small $b-s$.

(b) $b-s$ 'large': in this case $\nu > \lambda^2/(16\bar{g}^2)$, where we have freedom in choosing λ as long as it is $O\{\log^{\alpha}(T)\}$. The main tool is theorem 1.3, part (i), in Bosq (1998). We first observe that $\mathbb{E}(U_t) \neq g_t$; however, by using lemma 1, we show that, for t far from the change point $\eta(t)$, they are very close. Taking differences and using lemma 1, we have

$$\begin{aligned} |\mathbb{E}(U_t) - g_t| &= |\mathbb{E}(\{g(X_t, \ldots, X_{t-\tau})\} - \mathbb{E}\{g(\tilde{X}_t^{\upsilon(t)}, \ldots, \tilde{X}_{t-\tau}^{\upsilon(t)})\}| \\ &\leqslant C \sum_{i=0}^{\tau} \mathbb{E}|X_{t-i}^2 - (\tilde{X}_{t-i}^{\upsilon(t)})^2| \leqslant C \sum_{i=0}^{\tau} \mathbb{E}(V_{t-i}) \leqslant C(\tau)\rho^{t-\eta(t)}, \end{aligned} \tag{8}$$

where $C(\tau)$ is a generic constant (that varies according to the equation both here and below and depends on τ) and the above is due to the Lipschitz continuity of $g(\cdot)$ in its squared arguments. Therefore we have

$$U_t - g_t = U_t - \mathbb{E}(U_t) + \{\mathbb{E}(U_t) - g_t\} := \varepsilon_t' + d_t \tag{9}$$

where $\varepsilon_t' = U_t - \mathbb{E}(U_t)$ and $|d_t| \leqslant C(\tau)\rho^{t-\eta(t)}$ (by expression (8)). Therefore for all $s \leqslant b$ we have

$$\sum_{t=s}^{b} |d_t| \leqslant \sum_{t=\eta_1}^{T-1} |d_t| \leqslant C(\tau) \sum_{t=\eta_1}^{T-1} \rho^{t-\eta(t)} = C(\tau) \sum_{i=1}^{N} \sum_{t=\eta_i}^{\eta_{i+1}} \rho^{t-\eta_i} \leqslant C(\tau)N. \tag{10}$$

We use this result below. We bound

$$\begin{aligned} P\{\max_{s,b,u} |\tilde{U}_{s,b} - \tilde{g}_{s,b}| \geqslant \lambda/2\} &\leqslant \sum_{s,b} P\{\max_{u} |\tilde{U}_{s,b} - \tilde{g}_{s,b}| \geqslant \lambda/2\} \\ &\leqslant T^2 \max_{s,b} P\{\max_{u} |\tilde{U}_{s,b} - \tilde{g}_{s,b}| \geqslant \lambda/2\}. \end{aligned}$$

Further, by using expressions (9) and (10),

$$\begin{aligned} P\{\max_u |\tilde{U}_{s,b} - \tilde{g}_{s,b}| \geqslant \lambda/2\} &\leqslant P\left\{\left|\nu^{-1/2}\sum_{t=s}^{b}(\varepsilon'_t + d_t)\right| \geqslant \lambda/2\right\} \qquad \text{(by equation (9))} \\ &\leqslant P\left\{\left|\nu^{-1/2}\sum_{t=s}^{b}\varepsilon'_t\right| + \nu^{-1/2}\sum_{t=s}^{b}|d_t| \geqslant \lambda/2\right\} \\ &\leqslant P\left\{\left|\nu^{-1/2}\sum_{t=s}^{b}\varepsilon'_t\right| \geqslant \lambda/2 - C(\tau)\nu^{-1/2}N\right\} \qquad \text{(by expression (10)).} \end{aligned}$$

Denote $\tilde{\lambda} = \lambda/2 - C(\tau)\nu^{-1/2}N$. Using formula (1.25) of Bosq (1998), we have

$$P\left\{\left|\nu^{-1/2}\sum_{t=s}^{b}\varepsilon'_t\right| \geqslant \tilde{\lambda}\right\} \leqslant 4\exp\left\{-\frac{\tilde{\lambda}^2}{\tilde{C}_1\nu}q(\nu,T)\right\} + 22\left(1 + \frac{\tilde{C}_2\nu^{1/2}}{\tilde{\lambda}}\right)^{1/2} q(\nu,T)\,\alpha\left\{\left[\frac{\nu}{2q(\nu,T)}\right]\right\}, \quad (11)$$

where $\tilde{C}_1$ and $\tilde{C}_2$ are positive constants, $q(\nu,T)$ is an arbitrary integer in $[1,\ldots,\nu/2]$, $[a]$ is the integer part of a and $\alpha(k)$ are the α-mixing coefficients of X_t which are of order ρ^k. Suitable choice of $q(\nu,T)$ is crucial. We set it to be $q(\nu,T) = \nu/h(T)$, where $h(T)$ is of the same order as $\tilde{\lambda}$. Clearly $q(\nu,T) \leqslant \nu/2$ as $h(T) \to \infty$ and also $q(\nu,T) \geqslant 1$ as ν is at least of order $O(\tilde{\lambda}^2)$. With this choice of $q(\nu,T)$, the bound in inequality (11) becomes at most $4\exp(-\tilde{\lambda}/\tilde{C}_3) + \tilde{C}_4 T^{5/4}\rho^{\tilde{\lambda}/2}$, which converges to 0 exponentially fast for a suitable logarithmic choice of $\tilde{\lambda}$ (see Appendix B for details of this rate). This completes the proof as the resulting λ is also logarithmic, as required in the statement of the proof.

Lemma 5. Assume expressions (5) and (6). On the event B_T from lemma 4, for $b = \arg\max_{s<t<u}|\tilde{U}^t_{s,u}|$, there exists $1 \leqslant r \leqslant q$ such that, for large T, $|b - \eta_{p_0+r}| \leqslant C\varepsilon_T$.

Proof. Let $b_1 = \arg\max_{s<t<u}|\tilde{g}^t_{s,u}|$. From lemma 4, $|\tilde{g}^{b_1}_{s,u}| \leqslant |\tilde{U}^{b_1}_{s,u}| + C\log^\alpha(T)$ and $|\tilde{U}^b_{s,u}| \leqslant |\tilde{g}^b_{s,u}| + C\log^\alpha(T)$. By the definition of b, we have $|\tilde{U}^{b_1}_{s,u}| \leqslant |\tilde{U}^b_{s,u}|$. Putting these together, we obtain

$$|\tilde{g}^{b_1}_{s,u}| \leqslant |\tilde{U}^{b_1}_{s,u}| + C\log^\alpha(T) \leqslant |\tilde{U}^b_{s,u}| + C\log^\alpha(T) \leqslant |\tilde{g}^b_{s,u}| + 2C\log^\alpha(T). \quad (12)$$

Assume that $b \in (\eta_{p_0+r} + C\varepsilon_T, \eta_{p_0+r+1} - C\varepsilon_T)$ for some r and without loss of generality $\tilde{g}^b_{s,u} > 0$. From lemma 2.2 in Venkatraman (1992), we have

(a) $\tilde{g}^t_{s,u}$ is either monotonic or decreasing and then increasing on $[\eta_{p_0+r}, \eta_{p_0+r+1}]$ and
(b) $\max(\tilde{g}^{\eta_{p_0+r}}_{s,u}, \tilde{g}^{\eta_{p_0+r+1}}_{s,u}) > \tilde{g}^b_{s,u}$.

If $\tilde{g}^b_{s,u}$ locally decreases at b, then $\tilde{g}^{\eta_{p_0+r}}_{s,u} > \tilde{g}^b_{s,u}$ and, from lemma 3, for C sufficiently large, there exists $b' \in (\eta_{p_0+r}, \eta_{p_0+r} + C\varepsilon_T]$ such that $\tilde{g}^{\eta_{p_0+r}}_{s,u} \geqslant \tilde{g}^{b'}_{s,u} + 2C\log^\alpha(T)$. Since $\tilde{g}^{b'}_{s,u} > \tilde{g}^b_{s,u}$, this would in turn lead to $|\tilde{g}^{b_1}_{s,u}| \geqslant |\tilde{g}^{\eta_{p_0+r}}_{s,u}| > |\tilde{g}^b_{s,u}| + 2C\log^\alpha(T)$, which would contradict result (12). Similar arguments (but involving η_{p_0+r+1} rather than η_{p_0+r}) apply if $\tilde{g}^b_{s,u}$ locally increases at b.

Lemma 6. On the event B_T from lemma 4, and under expressions (5) and (6), $|\tilde{U}^b_{s,u}| > CT^\theta$, where $b = \arg\max_{s<t<u}|\tilde{U}^t_{s,u}|$.

Proof. $|\tilde{U}^b_{s,u}| \geqslant |\tilde{U}^{\eta_{p_0+r^*}}_{s,u}| \geqslant |\tilde{g}^{\eta_{p_0+r^*}}_{s,u}| - C\log^\alpha(T) \geqslant C\{\delta_T T^{-1/2} - \log^\alpha(T)\} > CT^\theta$.

Lemma 7. For some positive constants C and C', let s and u satisfy either

(a) $\exists 1 \leqslant p \leqslant N$ such that $s \leqslant \eta_p \leqslant u$ and $(\eta_p - s + 1) \wedge (u - \eta_p) \leqslant C\varepsilon_T$ or
(b) $\exists 1 \leqslant p \leqslant N$ such that $s \leqslant \eta_p \leqslant \eta_{p+1} \leqslant u$ and $(\eta_p - s + 1) \vee (u - \eta_{p+1}) \leqslant C'\varepsilon_T$.

On the event B_T from lemma 4, $|\tilde{U}^b_{s,u}| < CT^\theta$, where $b = \arg\max_{s<t<u}|\tilde{U}^t_{s,u}|$.

Proof.

$$|\tilde{U}^b_{s,u}| \leqslant |\tilde{g}^b_{s,u}| + C\log^\alpha(T) \leqslant C\{\varepsilon_T^{1/2} + \log^\alpha(T)\},$$

where the last inequality uses the definition of $\tilde{g}^t_{s,u}$ and condition (a) or (b). This is, for large T, of a lower magnitude than CT^θ as $\theta > \frac{1}{4}$.

With the use of lemmas 1–7, the proof of theorem 1 is simple; the following occurs on the event B_T. At the start of the algorithm, as $s = 0$ and $u = T - 1$, all conditions for lemma 6 are met and it finds a change point within the distance of $C\varepsilon_T$ from the true change point, by lemma 5. Under the assumption of theorem 1, both condition (5) and condition (6) are satisfied within each segment until every change point in g_t has been identified. Then, either of the two conditions (a) and (b) of lemma 7 is met and no further change points are detected.

Appendix B: Clarification of *N* as a function of *T*

The maximum permitted N can be inferred from lemma 4. The quantity $4\exp(-\tilde{\lambda}/C_3) + C_4 T^{5/4}\rho^{\tilde{\lambda}/2}$ needs to converge to 0. The rate is arbitrary but, to set it to T^Δ ($\Delta < 0$) or faster, we require $-\tilde{\lambda}/C_3 \leqslant \Delta \log(T)$ and $\frac{5}{4}\log(T) + (\tilde{\lambda}/2)\log(\rho) \leqslant \Delta \log(T)$, which give

$$\tilde{\lambda} \geqslant \max\left\{\frac{2(\Delta - \frac{5}{4})}{\log(\rho)}, -C_3\Delta\right\}\log(T) =: \bar{C}\log(T).$$

Recalling that $\tilde{\lambda} = \lambda/2 - C(\tau)\nu^{-1/2}N$ and choosing $\lambda = C\log^\alpha(T)$ as in lemma 4, we obtain $C(\tau)\nu^{-1/2}N \leqslant (C/2)\log^\alpha(T) - \bar{C}\log(T)$, which, recalling that $\nu > \lambda^2/(16\bar{g}^2)$, is guaranteed by

$$N \leqslant \frac{C\log^\alpha(T)\{(C/2)\log^\alpha(T) - \bar{C}\log(T)\}}{4C(\tau)\bar{g}}.$$

This determines the largest permitted number of change points N. As can be seen from the above formula, it is permitted to increase slowly to ∞ with the sample size.

References

Adak, S. (1998) Time-dependent spectral analysis of nonstationary time series. *J. Am. Statist. Ass.*, **93**, 1488–1501.

Andreou, E. and Ghysels, E. (2002) Detecting multiple breaks in financial market volatility dynamics. *J. Appl. Econmetr.*, **17**, 579–600.

Bollerslev, T. (1986) Generalized autoregressive conditional heteroscedasticity. *J. Econmetr.*, **31**, 307–327.

Bosq, D. (1998) *Nonparametric Statistics for Stochastic Processes*. New York: Springer.

Brodsky, B. and Darkhovsky, B. (1993) *Nonparametric Methods in Change-point Problems*. Dordrecht: Kluwer.

Cho, H. and Fryzlewicz, P. (2012) Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statist. Sin.*, **22**, 207–229.

Chu, C.-S. J. (1995) Detecting parameter shift in GARCH models. *Econmetr. Rev.*, **14**, 241–266.

Dahlhaus, R. and Subba Rao, S. (2006) Statistical inference for time-varying ARCH processes. *Ann. Statist.*, **34**, 1075–1114.

Davidson, J. (1994) *Stochastic Limit Theory*. Oxford: Oxford University Press.

Davis, R., Lee, T. and Rodriguez-Yam, G. (2006) Structural break estimation for nonstationary time series models. *J. Am. Statist. Ass.*, **101**, 223–239.

Davis, R., Lee, T. and Rodriguez-Yam, G. (2008) Break detection for a class of nonlinear time series models. *J. Time Ser. Anal.*, **29**, 834–867.

Engle, R. F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, **50**, 987–1008.

Fan, J. and Yao, Q. (2003) *Nonlinear Time Series*. New York: Springer.

Fryzlewicz, P. (2007) Unbalanced Haar technique for nonparametric function estimation. *J. Am. Statist. Ass.*, **102**, 1318–1327.

Fryzlewicz, P., Sapatinas, T. and Subba Rao, S. (2008) Normalized least-squares estimation in time-varying ARCH models. *Ann. Statist.*, **36**, 742–786.

Fryzlewicz, P. and Subba Rao, S. (2011) On some mixing properties of ARCH and time-varying ARCH processes. *Bernoulli*, **17**, 320–346.

Giraitis, L., Leipus, R. and Surgailis, D. (2005) Recent advances in ARCH modelling. In *Long Memory in Economics* (eds A. Kirman and G. Teyssiere), pp. 3–39. Berlin: Springer.

Janeway, W. (2009) Six impossible things before breakfast: lessons from the crisis. *Significance*, **6**, 28–31.

Kokoszka, P. and Leipus, R. (2000) Change-point estimation in ARCH models. *Bernoulli*, **6**, 513–539.

Kulperger, R. and Yu, H. (2005) High moment partial sum processes of residuals in GARCH models and their applications. *Ann. Statist.*, **33**, 2395–2422.

Last, M. and Shumway, R. (2008) Detecting abrupt changes in a piecewise locally stationary time series. *J. Multiv. Anal.*, **99**, 191–214.
Lavielle, M. and Moulines, E. (2000) Least-squares estimation of an unknown number of shifts in a time series. *J. Time Ser. Anal.*, **21**, 33–59.
Lavielle, M. and Teyssiere, G. (2005) Adaptive detection of multiple change-points in asset price volatility. In *Long Memory in Economics* (eds A. Kirman and G. Teyssiere), pp. 129–157. Berlin: Springer.
Mikosch, T. and Stărică, C. (2004) Non-stationarities in financial time series, the long-range dependence, and the IGARCH effects. *Rev. Econ. Statist.*, **86**, 378–390.
Paparoditis, E. (2010) Validating stationarity assumptions in time series analysis by rolling local periodograms. *J. Am. Statist. Ass.*, **105**, 839–851.
Politis, D. (2007) Model-free versus model-based volatility prediction. *J. Finan. Econmetr.*, **5**, 358–389.
Rydberg, T. (2000) Realistic statistical modelling of financial data. *Int. Statist. Rev.*, **68**, 233–258.
Sen, A. and Srivastava, M. S. (1975) Some one-sided tests for changes in level. *Technometrics*, **17**, 61–65.
Stoffer, D., Ombao, H. and Tyler, D. (2002) Local spectral envelope: an approach using dyadic tree-based adaptive segmentation. *Ann. Inst. Statist. Math.*, **54**, 201–223.
Taylor, S. J. (1986) *Modelling Financial Time Series*. Chichester: Wiley.
Venkatraman, E. S. (1992) Consistency results in multiple change-point problems. *Technical Report 24*. Department of Statistics, Stanford University, Stanford. (Available from `http://statistics.stanford.edu/~ckirby/techreports/NSA/SIE%20NSA%2024.pdf`.)
Vostrikova, L. J. (1981) Detecting "disorder" in multidimensional random processes. *Sov. Math. Dokl.*, **24**, 55–59.