

Mode Identification of Volatility in Time-varying Autoregression

Gabriel Chandler and Wolfgang Polonik

Abstract

In many applications, time series exhibit non-stationary behavior that might reasonably be modeled as a time-varying autoregressive process. In the context of such a model, we discuss the problem of testing for modality of the variance function. We propose a test of modality which is local, and which when used iteratively, can be used to identify the total number of modes in a given series. This problem is closely related to peak detection and identification, which has applications in many fields. We propose a test that, under appropriate assumptions, is asymptotically distribution free under the null-hypothesis, even though non-parametric estimation of the AR parameter-functions is involved. Simulation studies and applications to real data sets illustrate the behavior of the test.

Author's Footnote: Gabriel Chandler is Associate Professor, Department of Mathematics, Pomona College, Claremont, CA 91711 (E-mail: *gabriel.chandler@pomona.edu*); and Wolfgang Polonik is Professor, Department of Statistics, University of California, Davis, CA 95616-8705 (E-mail: *wpolonik@ucdavis.edu*). This work has been partially supported by the NSF grant# 0406431. The authors would like to thank the referees and the AE for constructive criticisms which lead to a significant improvement of the original manuscript.

1 INTRODUCTION

Underlying the methodology of this paper is the following nonstationary autoregressive model which allows the volatility to evolve unconditionally on the past as a function of time:

$$Y_t - \sum_{k=1}^p \theta_k\left(\frac{t}{T}\right) Y_{t-k} = \sigma\left(\frac{t}{T}\right) \epsilon_t, \quad t = 1, \dots, T, \quad (1)$$

where θ_k are the autoregressive parameter functions, p is the order of the model, σ is a function controlling the volatility and $\epsilon_t \sim (0, 1)$ i.i.d. Following Dahlhaus (1997), time is rescaled to the interval $[0, 1]$ in order to make a large sample analysis feasible. Observe that this in particular means that the process $Y_t = Y_{t,T}$ satisfying (1) is in fact a triangular array of processes. Under certain additional assumptions, (1) constitutes a locally stationary processes in the sense of Dahlhaus (1997). Such additional assumptions are not used here, however.

Time-varying AR-processes have been in use for quite some while, in particular in the area of signal processing (e.g. Subba Rao, 1970, Grenier, 1983, Hall et al. 1983, Rjan and Rayner, 1996, Girault et al. 1998, Eom, 1999). In this area of work, however, the variance is often assumed to be constant, and only the AR-parameters are modeled as being time-dependent. Orbe et al. (2005) also used such a model in the context of econometric time series. Another line of work, such as Drees and Stărică (2002) and Fryzlewicz et al. (2006), used a model with constant AR-parameters, but time-varying variance function as a simple yet effective model for stock returns, while Chandler and Polonik (2006) used the model as the basis for discriminating different types of seismic events. The focus of this paper is also the inference for the variance function $\sigma^2(\cdot)$. While we allow for the AR-parameters $\theta_k(u)$, $u \in [0, 1]$ to vary with time, they are regarded as nuisance parameters.

There are situations in which the shape of these parameter functions are of particular interest. In an econometric context, seasonality might be a causing factor, and Orbe et al. (2005) incorporate such shape information into the estimation method for the time-varying AR-parameters (the variance is kept constant in this paper). However, the shape of the variance function is of interest in various applications. For instance, a measure of concentration of the variance function was used in the context of discrimination between earthquakes and explosions in Chandler and Polonik (2006). Other applications regard the question of monotonicity or the modality of the variance function. In particular, this question is closely

related to the problem of peak detection and estimation, which shows up in a variety of contexts. For example, peak picking is important for onset detection in musical signals (Bello et al. 2005). In order to estimate the location of a peak, we must first be sure that one exists. This motivates a test for the existence of peaks. For simplicity of presentation, we present a test that only considers whether or not there is a mode in a given interval of time. However, our test is performed locally. Via an iterative procedure, we can thus count the number of modes in a given time series, in addition to an estimate of their locations. The location of a mode is only discussed in passing, as it is not the focus of this paper.

As a motivating example, suppose that we are interested in locating volatility clusters in financial data, often modeled by the celebrated ARCH (Engle, 1982) and GARCH (Bollerslev, 1986) models. From an unconditional heteroscedastic perspective, volatility clusters correspond to modes in the variance function. For instance, we might consider a model in which volatility clusters result from exogenic shocks. Identifying individual volatility clusters (modes) allows us to identify the presence of shocks, which we can then try to identify. Figure 1.1 shows the log returns, r_t , from the Dow Jones Industrial Average (DJI), which we model simply as $r_t = \theta r_{t-1} + \sigma(t)\epsilon_t$.

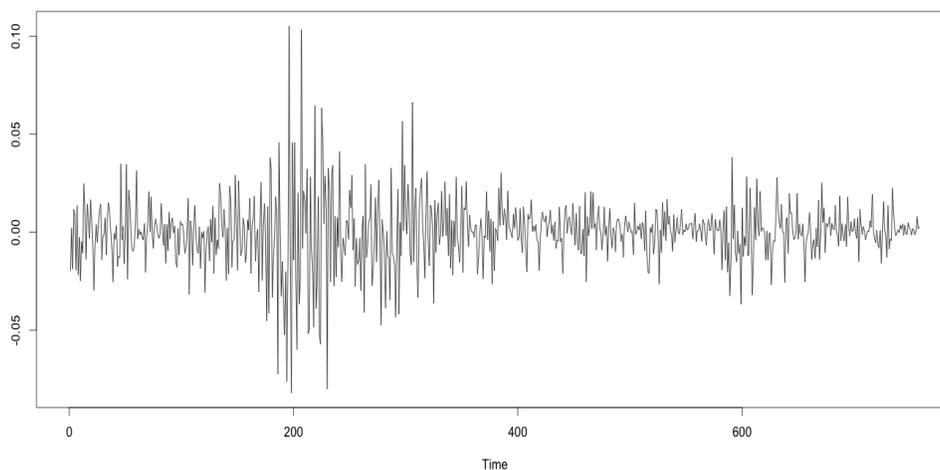


Figure 1.1: Log return on the Dow Jones Industrial Average closing prices from January 2008 to December 2010.

Using a graphical device motivated by the mode tree of Minnotte and Scott (1993), we present the results of using our test of modality (to be discussed in detail below) iteratively on the DJI data set. Whereas the mode tree plots mode locations against the bandwidth at which they first appear, we plot the locations against their p-value. For the DJI data set,

the mode corresponding to the initial crash of October, 2008 (in the plot at observation 300) is readily apparent and highly significant for our test of homoscedasticity. The second most significant mode corresponds to the period following the “flash crash” of May, 2010 (in the plot just before observation 600), with a p-value of .125. Though not significant, studying the stock market on a daily basis obscures much of the variability in the market that would be apparent on a smaller time scale, especially obvious with respect to volatility of the ‘crash’ in question. What appears to be a mode in early 2009 seemingly is only two days of high log returns, and our method finds that mode to be highly non-significant, though due to the outlying values, a kernel smoother (of the squared observations or residuals resulting from fitting some model, say) may very well suggest a mode exists.

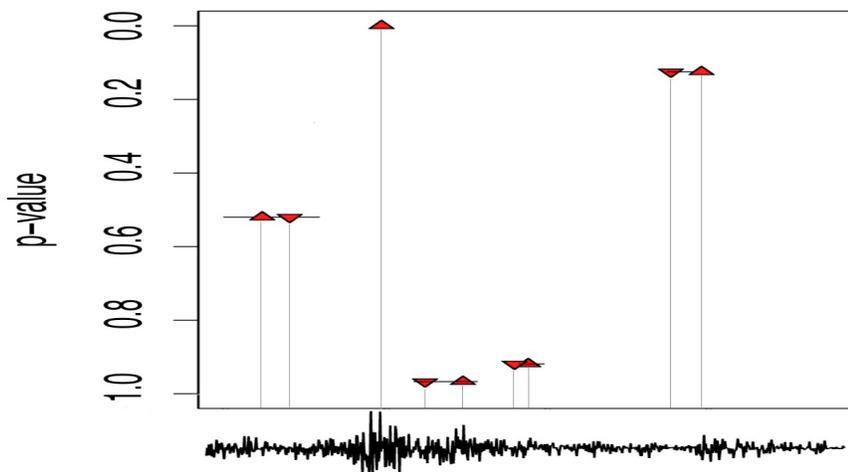


Figure 1.2: A mode tree for the DJI data set. Up arrows denote estimates of locations of modes, down arrows denote estimates of anti-modes. By design, the initial estimate of the mode does not come with an antimode attached to it. Lines are the estimated sets on which our test of modality was run.

In section 2 we present and discuss the testing problem, our proposed test statistic and its asymptotic behavior. Section 3 discusses the application of the testing procedure to mode counting. In section 4 we present simulation studies and applications to real data. Proofs are deferred to section 5.

2 TESTING FOR MODALITY OF THE VOLATILITY

2.1 THE PROBLEM

Tests of modality of a probability density function have been studied in detail. For instance, the excess mass test (Müller and Sawitzki, 1991), the dip test (Hartigan and Hartigan, 1985),

and the bandwidth test (Silverman, 1981) have been proposed. For higher dimensions, Burman and Polonik (2009) propose an approach based on the concept of an antimode. In the regression setting, tests of monotonicity of functions has been explored by a number of authors. See, for instance, Bowman, et al. (1998), Ghosal, et al. (2000) and Hall and Heckman (2000), and references therein.

In contrast to this work, we consider a non-stationary time series context, and our interest is to test for modality of the variance function $\sigma^2(\cdot)$. Compared to other approaches for investigating modality from the literature, the approach studied here is conceptually closest to that of Burman and Polonik (2009), as it also uses the relationship that if a function is unimodal then it contains no anti-mode. Here we say that an anti-mode is present if there exists an interval I composed of a union of two disjoint intervals, i.e. $I = I_0 \cup I_1$ such that $\sigma(\cdot)$ is decreasing on I_0 and increasing on I_1 . On each stretch (I_0 or I_1), the function $\sigma(\cdot)$ does not need to be strictly monotonic, but is not allowed to be fully constant on either I_0 or I_1 . Using this definition of an anti-mode, we consider the hypotheses:

H_0 : $\sigma(\cdot)$ is unimodal (i.e. has no antimode);

H_1 : $\sigma(\cdot)$ has an anti-mode somewhere.

Testing for unimodality and testing for monotonicity are closely related. A unimodal function is simply one that is monotone on either side of the mode (first increasing then decreasing). As such, testing for unimodality is equivalent to testing for the appropriate monotonicity on either side of the mode. In order to detect deviations from monotonicity, one has to consider the ‘right’ region that contains an anti-mode under H_1 .

For now, in order to simplify the discussion, assume that the ‘right’ region is known to be the interval $[a, b] \subset [0, 1]$ to the left of the mode, and consider the hypothesis that $\sigma^2(\cdot)$ is monotonically increasing in this interval. For $k = 1, \dots, p$, let $\hat{\theta}_k(\cdot)$ be an estimator for $\theta_k(\cdot)$, and define the corresponding squared residuals as $\hat{\eta}_t^2 = (Y_t - \sum_{k=1}^p \hat{\theta}_k(\frac{t}{T}) Y_{t-k})^2$. Under our model (1) we can expect the squared residuals to behave like $\sigma^2(\frac{t}{T}) \epsilon_t^2$. Thus, under the hypothesis of $\sigma^2(\cdot)$ being increasing on $[a, b]$, the squared residuals should tend to be increasing in that interval. In fact, we expect the large squared residuals to be concentrated towards b if the function is strictly monotone, or perhaps uniformly placed in the case that the function $\sigma^2(\cdot)$ is constant on $[a, b]$. On the other hand, if the function contains an anti-mode on $[a, b]$, we might expect to see different patterns in the occurrences of large squared residuals, also depending on the choice of $[a, b]$. For instance, consider the function

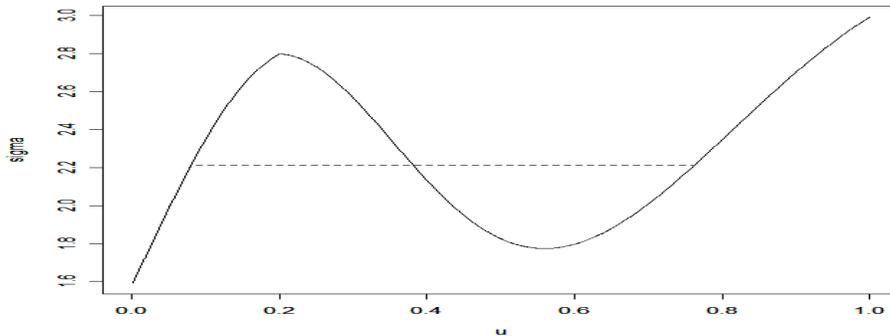


Figure 2.1: A bimodal variance function (solid line) and a similar unimodal function (dashed line) that is the L_2 -projection of the bimodal $\sigma(u)$ onto the class of monotone functions. The two functions agree everywhere the dashed line is not visible.

in figure 2.1. If we choose the set $[a, b]$ to be the set indicated by the dashed line, one can expect that many of the largest residuals will tend to fall within the first half of the interval, as this is where the variance function is largest on the set. The construction of our test statistic exploits these different patterns of occurrences of large residuals. The most difficult case to discriminate against the alternative, an anti-mode, should be a unimodal $\sigma^2(\cdot)$ with $\sigma(u) = c$ for all u on $[a, b]$, which corresponds to an (approximate) uniform distribution of the locations of large squared residuals across the set $[a, b]$.

Our proposed testing procedure also takes into account the possibility of observed outliers in the residuals, as it only uses their ranks and not the values themselves. Our procedure shares the benefits of Hall and Heckman's procedure in that it does not require selecting a bandwidth.

2.2 THE TEST STATISTIC

Let $\mathbf{Y}_{t-1} = (Y_{t-1}, \dots, Y_{t-p})'$, and let $\hat{\boldsymbol{\theta}}_n(\frac{t}{T}) = (\hat{\theta}_{n,1}(\frac{t}{T}), \dots, \hat{\theta}_{n,p}(\frac{t}{T}))'$ be an estimator for the AR-parameters. Consider the γ -proportion largest squared residuals $\hat{\eta}_t^2$ on the set $[a, b]$, i.e. the squared residuals that exceed the empirical upper γ -quantile of the squared residuals on $[a, b]$ given by

$$\hat{q}_\gamma^2 = \min\{q^2 \geq 0 : \frac{1}{n} \sum_{t \in [aT, bT]} \mathbf{1}(\hat{\eta}_t^2 > q^2) \leq \gamma\},$$

where $n = \lfloor bT \rfloor - \lfloor aT \rfloor + 1$ is the number of observations in $[aT, bT]$. The quantity \hat{q}_γ^2 serves as an estimate of the γ -quantile of the squared residual if the variance function is constant

on this set (see section 2.3). For given $[a, b]$ and $0 < \gamma < 1$ define the partial sum process of the ‘large’ squared residuals on $[a, b]$ as

$$\widehat{G}_{n,\gamma}(\alpha) = \frac{1}{n} \sum_{t=\lceil aT \rceil}^{\lceil aT \rceil + \lceil \alpha n \rceil - 1} \mathbf{1}(\widehat{\eta}_t^2 \geq \widehat{q}_\gamma^2), \quad \alpha \in [0, 1]. \quad (2)$$

Notice that $\widehat{G}_{n,\gamma}(\cdot)$ is defined in $[0, 1]$ rather than $[a, b]$. This reparametrization is done for ease of notation.

To simplify the discussion, assume for the moment that all the residuals are different and that γn is an integer. Then by definition $\widehat{G}_{n,\gamma}(1) = \gamma$. The jumps in $\widehat{G}_{n,\gamma}(\cdot)$ occur at the γn ‘large’ squared residuals, and under the worst case null-hypothesis, the location of the large residuals can be expected to be approximately uniform on the interval $[a, b]$. We can then expect $\widehat{G}_{n,\gamma}(\alpha)$ to behave like $\gamma\alpha$ in this case. When testing for $\sigma^2(u)$ monotonically increasing, we expect more of the large squared residuals to occur in the earlier portions of the interval $[a, b]$ than under a uniform distribution. Thus, we reject the null ($\sigma^2(\cdot)$ monotonically increasing on $[a, b]$) for large values of $\sup_{\alpha \in [0,1]} (\widehat{G}_{n,\gamma}(\alpha) - \gamma\alpha)$. When testing for a monotonically decreasing $\sigma^2(\cdot)$ (on $[a, b]$) we modify T_n by taking the supremum over $\gamma\alpha - \widehat{G}_{n,\gamma}(\alpha)$. Letting $m \in [0, 1]$ denote the mode of $\sigma^2(\cdot)$, which might be known or estimated, our test statistic thus is

$$T_n = \begin{cases} \sup_{\alpha \in [0,1]} (\widehat{G}_{n,\gamma}(\alpha) - \gamma\alpha), & [a, b] \subset [0, m] \\ \sup_{\alpha \in [0,1]} (\gamma\alpha - \widehat{G}_{n,\gamma}(\alpha)), & [a, b] \subset [m, 1]. \end{cases} \quad (3)$$

For the simple test of homoscedasticity one can use $[a, b] = [0, 1]$ and $T_n^{(1)} = \sup_{\alpha \in [0,1]} |\widehat{G}_{n,\gamma}(\alpha) - \gamma\alpha|$. In practice we will have to estimate a “good” interval $[a, b]$ and also the mode m . In section 2.4 we will discuss the estimation procedure for this interval. For more details an consistent mode estimation see section 2.3.1.

2.3 ASYMPTOTIC DISTRIBUTION

Under appropriate assumptions, the asymptotic distribution of the appropriately normalized test statistics T_n under the worst case null hypothesis ($\sigma(u) \equiv \sigma_0 > 0$ for all $u \in [a, b]$) is the supremum of a standard Brownian Bridge, and thus is distribution free. See Theorem 1 below. This theorem also shows that the large sample behavior of our test statistic is not influenced by the (in general) non-parametric estimation of the parameter functions, as long

as the rate of convergence of these estimators is sufficiently fast, essentially meaning that the model class for the AR-parameter function is not too rich.

THEOREM 1 *Let $\gamma \in [0, 1]$ and suppose that $0 \leq a < b \leq 1$ are non-random with $m \notin [a, b]$. Assume further that $E|\epsilon|^k < C^k$ for some $C > 0$ and all $k = 1, 2, \dots$. Then, under assumptions (i) - (iv) (see appendix), we have under the worst case null-hypothesis $\sigma(u) \equiv \sigma_0$ for $u \in [a, b]$ that*

$$\sqrt{\frac{n}{\gamma(1-\gamma)}} T_n \rightarrow \sup_{\alpha \in [0, 1]} B(\alpha), \quad \text{in distribution as } n \rightarrow \infty.$$

Moreover, in case the AR-parameter in model (1) are constant, and \sqrt{n} -consistent estimators are used, then the moment assumptions on the innovations can be relaxed to $E\epsilon_t^2 < \infty$. The result continues to hold if the mode m of $\sigma(\cdot)$ is replaced by an estimate \hat{m} with $|\hat{m} - m| = o_P(1)$ as $n \rightarrow \infty$, and if $[a, b]$ is replaced by an estimated set $[\hat{a}, \hat{b}]$ with $P([\hat{a}, \hat{b}] \subset [a, b]) \rightarrow 1$ and $\max\{|c - \hat{a}|, |d - \hat{b}|\} = o_P(1)$ for some $a \leq c < d \leq b$ as $n \rightarrow \infty$,

Consistent estimation of the mode is discussed in subsection 2.3.1. The construction of an appropriate estimated set $[\hat{a}, \hat{b}]$ is presented in section 2.4.

We also note that while our testing procedure is calibrated to have the correct asymptotic level α when the variance function is constant on our interval, it is the case that the asymptotic level under any other form satisfying the null hypothesis, the level is at most α . This is due to the fact that for any variance function satisfying the null hypothesis $E(G_{n,\gamma}(\alpha)) \leq \gamma\alpha + O(n^{-1})$ for all $\alpha \in [0, m]$ and $E(G_{n,\gamma}(\alpha)) \geq \gamma\alpha + O(n^{-1})$ for $\alpha \in [m, 1]$. In fact, with

$$\Psi_{a,b}(\alpha, z) = \int_0^\alpha F\left(\frac{z}{\sigma(a+\beta(b-a))}\right) d\beta - \int_0^\alpha F\left(\frac{-z}{\sigma(a+\beta(b-a))}\right) d\beta,$$

the expected value $E(G_{n,\gamma}(\alpha))$ can be approximated uniformly in α by $\alpha - \Psi_{a,b}(\alpha, q_\gamma)$ (see proof of Theorem 2). This motivates the fact that the deviation $\Psi_{a,b}(\alpha, q_\gamma) - \alpha(1-\gamma)$ influences the large sample behavior of our test statistics (cf. Theorem 2), where q_γ is such that

$$\Psi_{a,b}(1, q_\gamma) = 1 - \gamma. \quad (4)$$

The quantity q_γ is well defined if the distribution function F of ϵ_t is strictly monotonic and $\min_{u \in [a, b]} \sigma(u) > 0$. Under the worst case null-hypothesis of $\sigma^2(\cdot) \equiv \sigma_0^2$, $u \in [a, b]$, this

definition gives us back the $(1 - \gamma)$ -quantile of the distribution of $\sigma_0 \epsilon_t$ (cf. definition of $G_{n,\gamma}(\alpha)$).

THEOREM 2 *If the assumptions of Theorem 1 hold and we have $\sup_{\alpha \in [0,1]} (\alpha(1 - \gamma) - \Psi_{a,b}(\alpha, q_\gamma)) > 0$, then the power of our test for the variance function being monotonically increasing on $[a, b]$ tends to one as $n \rightarrow \infty$. Similarly, the power of the test for a monotonically decreasing variance function tends to one if $\sup_{\alpha \in [0,1]} (\Psi_{a,b}(\alpha, q_\gamma) - \alpha(1 - \gamma)) > 0$.*

2.3.1 CONSISTENT ESTIMATION OF MAXIMA AND MINIMA. In practice we need to estimate maxima and minima of $\sigma^2(\cdot)$ (restricted to a certain region), and Theorem 1 requires the estimation of the mode to be consistent. Under appropriate smoothness conditions, one such consistent estimator is given by the maximum local likelihood estimate discussed in Dahlhaus and Polonik (2009). Consistency of this estimator can be shown by using the uniform consistency of the local likelihood estimates, which is shown in Dahlhaus and Polonik, in conjunction with standard techniques from mode estimation (e.g. see Appendix, section 5.3). The same arguments apply to the estimation of a minimum.

2.4 ESTIMATION OF THE SET OF INTEREST.

In practice, the set $[a, b]$ on which our test statistic is based will often be unknown and a data driven method to select $[a, b]$ will be needed. We propose a method based on a constrained optimization estimate of the variance function, where the constraint is motivated by the null hypothesis of unimodality. We search for the best fitting (in the sense of least squares) unimodal function, which amounts to performing isotonic regression on the squared residuals $\hat{\eta}_t^2$ to each side of the mode m . More precisely, let $\hat{\Sigma}(u) := \frac{1}{T} \sum_{t=1}^{\lfloor uT \rfloor} \hat{\eta}_t^2$, $u \in [0, 1]$, and for $m \in [0, 1]$ let $\hat{\Sigma}_m^*(u)$ denote the greatest convex minorant / least concave majorant (GCM/LCM) to $\hat{\Sigma}(u)$ on $[0, m]$ and $[m, 1]$, respectively. The function $\hat{\sigma}_m^{*2}(u)$ further denotes the slope (right-hand derivative) of $\hat{\Sigma}_m^*(u)$. The mode m may either be fixed or itself be selected by the data. Efficient algorithms to calculate $\hat{\sigma}_m^{*2}(u)$, based on the pooled-adjacent-violators algorithm, exist (see e.g. Robertson et al. 1988). The resulting estimate is a piecewise constant function.

THE ESTIMATION PROCEDURE. $\hat{\Sigma}_m^*(u)$ estimates $\Sigma_m^*(u)$, which denotes the GCM/LCM to

$$\Sigma(u) := \int_0^u \sigma^2(v) dv, \tag{5}$$

with m being the mode of $\sigma^2(u)$. We also denote by $\sigma_m^{*2}(u)$ the derivative of $\Sigma_m^*(u)$. For simplicity we first consider the region left of the mode. Under H_0 where $\sigma^2(u)$ is increasing to the left of m , we have $\sigma^2(u) = \sigma_m^{*2}(u)$ for $u \in [0, m]$, because $\Sigma_m^*(u) = \Sigma(u)$ for $u \in [0, m]$. If $\sigma_m^{*2}(u)$ is constant on a certain stretch $I^0 = [a, b] \subset [0, m]$, then $\Sigma_m^*(u)$ is linear on I^0 , and thus the longest of these linear stretches is our set of interest. If $\sigma^2(u)$ has an antimode to the left of m , then $\Sigma(u)$ is concave on an interval containing the antimode, and this implies the $\Sigma_m^*(u)$ will lie strictly below $\Sigma(u)$ on an interval containing the antimode. Let I^0 denote the closure of the longest such interval. The simple, but crucial, observation is that σ_m^{*2} is constant on this set I^0 . The set I^0 for the function shown in figure 2.1 is exactly the closure of the set for which the solid and dashed lines do not agree.

The intuitive idea now is to use the longest constant stretch, \widehat{I}_0 , of the piecewise constant function $\widehat{\sigma}_m^{*2}$ as an estimator of the longest constant stretch of σ_m^{*2} . However, this will in general not result in a consistent estimator for I^0 (even if I^0 is the only constant stretch of σ_m^{*2}). The reason for this is that the greatest convex minorant will tend to have a few jumps along I^0 , so that \widehat{I}_0 will tend to fall somewhere inside the true flat set, and thus will be shorter. More seriously, considered as a random set \widehat{I}_0 , it does not stabilize, i.e. it does not converge to a fixed subset of I^0 . This ‘instability’ affects the limiting distribution. Our remedy to overcome this inconsistency is to instead estimate a slight enlargement of I^0 . This will result in a consistent estimator (see below). In a second step, we prune back this estimated set somewhat. This will result in a subset of I_0 for the test, but the estimator will stabilize. Besides not affecting the limiting distribution, this estimate also has the important effect of a bias reduction, as discussed below.

The formal definition of our estimator to the left of the mode is as follows. For a given m write

$$\widehat{H}_{m,\lambda}(c, d) = \widehat{\Sigma}_m^*(c, d) - \lambda|c - d|, \quad 0 \leq c \leq d \leq 1,$$

where in slight abuse of notation $\widehat{\Sigma}_m^*(c, d) = \widehat{\Sigma}_m^*(d) - \widehat{\Sigma}_m^*(c)$. In other words, $\widehat{\Sigma}_m^*(c, d)$ denotes the mass content of the interval $[c, d]$ under the positive measure given by $\widehat{\Sigma}_m^*(u)$. Now, for given $\lambda, \delta > 0$ define

$$\widehat{\Delta}_m^\ell(\delta; \lambda) = \sup_{[c,d] \subset [0,m]} \{ |c - d| : -\widehat{H}_{m,\lambda}(c, t) \leq \delta \text{ and } \widehat{H}_{m,\lambda}(t, d) \leq \delta \text{ for all } t \in [c, d] \}$$

Let $[\widehat{c}_\lambda(\delta), \widehat{d}_\lambda(\delta)]$ denote the ‘optimal’ interval at which the supremum in the definition of $\widehat{\Delta}_m^\ell(\delta; \lambda)$ is attained. (For ease of notation we drop the dependence on m of $[\widehat{c}_\lambda(\delta), \widehat{d}_\lambda(\delta)]$.)

In a second step we maximize over $\lambda > 0$:

$$\widehat{\Delta}_m^\ell(\delta) = \sup_{\lambda \geq 0} \widehat{\Delta}_m^\ell(\delta; \lambda) = \sup_{\lambda \geq 0} |\widehat{d}_\lambda(\delta) - \widehat{c}_\lambda(\delta)|. \quad (6)$$

Let $[\widehat{a}_\delta^{*\ell}, \widehat{b}_\delta^{*\ell}] \subset [0, m]$ denote the corresponding maximizing set (again the dependence on m is not indicated). In case there is more than one interval $[\widehat{a}_\delta^{*\ell}, \widehat{b}_\delta^{*\ell}]$ at which the maximum in $\widehat{\Delta}_m^\ell(\delta)$ is assumed, then pick one of these maximizers. In a similar fashion we define an estimated set to the right of the mode, where the definition of $\widehat{\Delta}_m^\ell(\delta; \lambda)$ is modified to

$$\widehat{\Delta}_m^r(\delta; \lambda) = \sup_{[c,d] \subset [m,1]} \{|c - d| : \widehat{H}_{m,\lambda}(c, t) \leq \delta \text{ and } -\widehat{H}_{m,\lambda}(t, d) \leq \delta \text{ for all } t \in [c, d]\}.$$

This then results in an estimated set $[\widehat{a}_\delta^{*r}, \widehat{b}_\delta^{*r}]$ for testing a decreasing function. More precisely, instead of $[\widehat{a}_\delta, \widehat{b}_\delta]$, we propose to use a slightly reduced set

$$[\widehat{a}_{\delta,\beta}, \widehat{b}_{\delta,\beta}]$$

which is defined by cutting off $\frac{\beta}{2} \times 100\%$ from each side of the interval $[\widehat{a}_\delta, \widehat{b}_\delta]$. For the motivation of this trimming procedure, see discussion below.

GEOMETRIC INTERPRETATION. For this discussion we only consider intervals in $[0, m]$, and to ease the notation we omit the superscript ℓ that indicates that we are to the left of the mode. Now, to the left of the mode, the function $\widehat{\sigma}_m^{*2}(t)$ is piecewise constant, increasing, and with a few jumps in the area of the unknown interval of interest $[a, b]$. This is due to the behavior of the LCM. Therefore the idea is to seek an interval that contains the longest constant piece of $\widehat{\sigma}_m^{*2}(t)$ such that on this interval the function $\widehat{\sigma}_m^{*2}(t)$ does not differ ‘by much’ from its value on the longest constant piece. Our measure of distance used here (defined via $\widehat{H}_{m,\lambda}$) is motivated by the fact that there tend to be only a few jumps of $\widehat{\sigma}_m^{*2}(t)$ along the interval of interest. The following makes this precise.

Consider the set $\widehat{T}(\lambda) = \{t \in [0, m] : \widehat{\sigma}_m^{*2}(t-) \leq \lambda \leq \widehat{\sigma}_m^{*2}(t+)\}$ where $\widehat{\sigma}_m^{*2}(t\pm)$ denote right hand and left hand limits, respectively, of $\widehat{\sigma}_m^{*2}$ at t . Since $\widehat{\sigma}_m^{*2}(\cdot)$ is increasing, we have $\widehat{T}(\lambda) = [\widehat{t}_0(\lambda), \widehat{t}_1(\lambda)]$ with $\widehat{t}_0(\lambda) \leq \widehat{t}_1(\lambda)$. If $\lambda \notin \{y : \widehat{\sigma}_m^{*2}(t) = y \text{ for some } t \in [0, m]\}$ then $\widehat{t}_0(\lambda) = \widehat{t}_1(\lambda)$ is a jump point, otherwise $\widehat{T}(\lambda)$ is the closure of the interval on which $\widehat{\sigma}_m^{*2}$ is constant equal to λ . Further, $\widehat{H}_{m,\lambda}(c, d)$ equals the integral of the difference $\widehat{\sigma}_m^{*2}(t) - \lambda$ over $[c, d]$, and since $\widehat{\sigma}_m^{*2}$ is monotonically increasing, we have $\widehat{H}_{m,\lambda}(c, t) \leq 0$ for $c \leq t \leq \widehat{t}_0(\lambda)$ and

$\widehat{H}_{m,\lambda}(t, d) \geq 0$ for $d \geq t \geq \widehat{t}_1(\lambda)$. It follows that we can write

$$\widehat{\Delta}_m(\delta; \lambda) = \sup_{\substack{c \leq \widehat{t}_0(\lambda) \\ d \geq \widehat{t}_1(\lambda)}} \{d - c : -\widehat{H}_{m,\lambda}(c, \widehat{t}_0(\lambda)) \leq \delta \text{ and } \widehat{H}_{m,\lambda}(\widehat{t}_1(\lambda), d) \leq \delta\}. \quad (7)$$

Thus, we can envision constructing $[\widehat{c}_\lambda(\delta), \widehat{d}_\lambda(\delta)]$ by extending $\widehat{T}(\lambda)$ to the left and to the right, such that the areas between the resulting horizontal line and $\widehat{\sigma}_m^{*2}$ to the left and to the right of $\widehat{T}(\lambda)$ both either are equal to δ or we hit one of the boundaries of $[0, m]$, whichever comes first. The largest of the intervals $[\widehat{c}_\lambda(\delta), \widehat{d}_\lambda(\delta)]$, $\lambda > 0$, then gives $[\widehat{a}_\delta^*, \widehat{b}_\delta^*]$. Figure 2.2 schematically shows, for a continuous function, two such sets for two different values of λ .

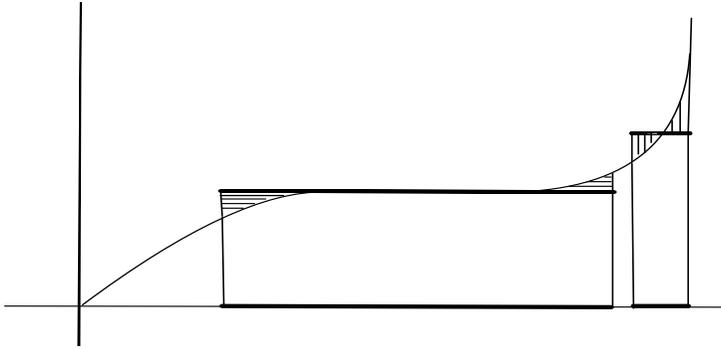


Figure 2.2: Indicated are two optimal sets $[c_\lambda(\delta), d_\lambda(\delta)]$ for two different values of λ and a continuous function $\sigma^{*2}(\cdot)$. The four individual shaded regions have areas equal to δ .

2.5 CONSISTENCY OF THE ESTIMATION METHOD.

In analogy to the above, define the quantities $\Delta_m(\delta; \lambda)$, $\Delta_m(\delta)$, $T(\lambda) = [t_0(\lambda), t_1(\lambda)]$, $[c_\lambda, d_\lambda]$ and the optimal set $[a_\delta^*, b_\delta^*] = [c_{\lambda_\delta^*}, d_{\lambda_\delta^*}]$ by replacing $\widehat{\Sigma}_m^*$ and $\widehat{\sigma}^{*2}$ in the above definitions by Σ_m^* and σ^{*2} , respectively. We assume that

- (A0) the function $\sigma^2(\cdot)$ is bounded;
- (A1) $\sup_{u \in [0, 1]} |\widehat{\Sigma}(u) - \Sigma(u)| = o_P(1)$;
- (A2) the estimator \widehat{m} is weakly consistent for m ;
- (A3) $(\delta, m) \rightarrow \sup_{\lambda \geq 0} \Delta_m(\delta; \lambda)$ is continuous in $(0, \delta_0] \times (0, 1]$ for some $\delta_0 > 0$;
- (A4) the parameter δ is such that the maximizing set $[a_\delta^*, b_\delta^*]$ is unique, i.e. there exists a unique level λ_δ^* and a unique corresponding optimal set $[c_{\lambda_\delta^*}, d_{\lambda_\delta^*}] = [a_\delta^*, b_\delta^*]$ such that $\sup_{\lambda \geq 0} \Delta_m(\delta; \lambda) = \Delta_m(\delta; \lambda_\delta^*) = |c_{\lambda_\delta^*} - d_{\lambda_\delta^*}|$. We also have $|b_\delta^* - a_\delta^*| > 0$.

LEMMA 1 *Let the estimator $[\widehat{a}_\delta^*, \widehat{b}_\delta^*]$ be defined by using $\widehat{\Sigma}_m^*$ with an estimator \widehat{m} , and suppose that assumptions (A1) - (A4) hold. Then, as $n \rightarrow \infty$, we have*

$$|[\widehat{a}_\delta^*, \widehat{b}_\delta^*] \Delta [a_\delta^*, b_\delta^*]| = o_P(1).$$

FACT 1. *Assumption (A1) holds under the assumptions of Theorem 1.*

LEMMA 2 *For $\delta > 0$ sufficiently small we have*

$$T(\lambda^*) \subset [a_\delta^*, b_\delta^*]$$

This result says that the theoretically optimal set contains the longest constant stretch. This together with the consistency of our estimate implies that an appropriate pruning of the initial estimated set provides an estimated set that stabilizes inside $T(\lambda^*)$ as desired.

BIAS CORRECTION. The proposed pruning step, i.e. to cut off $\frac{\beta}{2} \times 100\%$ from each side of the interval $[\widehat{a}_\delta, \widehat{b}_\delta]$, can be understood as a bias reduction. The heuristic behind this is the following. Suppose we are to the left of the mode and the null hypothesis is true (σ^2 increasing), but a flat set exists. Then the distribution of large residuals will not tend to be uniform over the entire estimated flat set due to the way we are estimating $\sigma^2(u)$. To understand this, notice that $\widehat{\sigma}^2(u)$ being flat over an extended period (to the left of the mode, say) means that $\widehat{\Sigma}(u)$ lies strictly above $\widehat{\Sigma}^*(u)$ in the interior of this extended period. For this to happen, the jumps of $\widehat{\Sigma}(u)$, given by the squared residuals at the corresponding locations, have to be relatively small in the interior of the flat set. In other words, it is likely that there exist relative extremes of the squared residuals near the endpoints of the flat set. As a remedy we thus propose to run the test on a slightly smaller set, removing a small percentage of the data at each endpoint. Such a procedure will not harm the test under the null, as the data is still homoscedastic on the smaller interval (should the original interval be homoscedastic). The pruning of the interval $[\widehat{a}_\delta^*, \widehat{b}_\delta^*]$ leads to the set

$$[\widehat{a}_{\delta,\beta}^*, \widehat{b}_{\delta,\beta}^*].$$

While any (strict) subset of the interval on which σ is constant serves the purpose of performing the test, the smaller the interval that is used for testing, the less is the available information, and thus less power can be expected.

2.6 CHOOSING δ AND β .

The choice of δ depends on the behavior of the variance function just beyond the boundaries of the set on which the function is constant, which is generally unknown. (It also depends on the sample size, and we might think to let $\delta \rightarrow 0$ as $T \rightarrow \infty$). Under the null hypothesis, for a fixed value of β , choosing δ too large will cause the set on which the test is run to tend to be too large as well. Thus, the variance function will only be monotone on the interval, not constant. Theoretically, we don't provide a result for the behavior of the test statistic in this case. Empirically, however, the actual level of the set will be smaller than the designated value of α as the small residuals will tend to be near the left hand boundary, with large residuals tending to be near the right, the opposite of what the test is looking for. While a conservative test is not a major problem, the loss of power under the alternative causes some concern. On the other hand, if δ is chosen too small, relative to the sample size, the level of the test is actually larger, as the estimated set tends to vary too much. We propose choosing δ via a semi-parametric bootstrap procedure, searching over a grid of δ with fixed β until the level in the bootstrapped tests is approximately α . We simulate from the model

$$Y_t = \sum_{k=1}^p \hat{\theta}_k Y_{t-k} + \hat{\sigma}\left(\frac{t}{T}\right) e_t^*$$

where both the AR parameters and the variance function are estimated from the data and the e_t^* are bootstrapped from the centered standardized residuals $\hat{\eta}_t/\hat{\sigma}^*(\frac{t}{T})$ where $\hat{\sigma}^*$ is any consistent estimator of σ (notice that the unimodal estimator is only consistent everywhere should H_0 be true, and thus should not be used). The test is then run on these simulated data sets and the best value of δ is chosen (i.e. the one that results in the level being closest to the nominal level). Simulation results show that the selected value of δ from the bootstrap procedure is often close to the best value of δ coming from sampling from the true underlying model.

3 FROM TESTING FOR MODALITY TO ANTIMODE COUNTING.

Given an initial estimate of the mode in the data, we can perform tests of monotonicity on either side of it. Our test of monotonicity by design comes with an estimated set $[\hat{a}_{\delta,\beta}^*, \hat{b}_{\delta,\beta}^*]$, which by construction can be expected to contain a second mode and the corresponding anti-mode, should they exist. We estimate the maximum and minimum of $\hat{\sigma}^2(\cdot)$ on this

region. This partitions the interval from the boundary to the mode into three new intervals, and we can run the test on each of these. We can do this repeatedly until no estimated set has length exceeding some lower bound $\eta > 0$. This method, which exhibits the flavor of a multi-scale procedure, leads to an estimate of the number of sufficiently pronounced antimodes in the data, given by the number of rejections in this procedure. Here we call an antimode of $\sigma^2(u)$ sufficiently pronounced if the corresponding critical set $[a_{\delta,\beta}^*, b_{\delta,\beta}^*]$ is long enough, meaning that it satisfies $b_{\delta,\beta}^* - a_{\delta,\beta}^* > \eta$.

The requirement to only consider sufficiently pronounced antimodes is due to the fact that this lower bound keeps the maximum number of tests we will run finite, which keeps multiple testing corrections under control.

Discussion of consistency. Let m_η denote the number of sufficiently pronounced antimodes of $\sigma^2(\cdot)$, and \hat{m}_η the number of antimodes estimated by applying our test repeatedly with critical values k_n . To see consistency of the antimode counting procedure, we need to show (i) that with probability tending to 1 our algorithm will find all the critical sets $[a_{\delta,\beta}^*, b_{\delta,\beta}^*]$ with $b_{\delta,\beta}^* - a_{\delta,\beta}^* > \eta$; (ii) that we can consistently estimate the local maxima and minima of $\sigma^2(u)$ for $u \in [a, b]$, where $[a, b]$ is an arbitrary given interval with $b - a > \eta$; and (iii) that the power of each individual test tends to one for each individual antimode tested, and the level tends to zero to avoid wrongly rejecting the null. Thus, we need the critical value to tend to infinity, but it cannot exceed $\sqrt{n} \sup_{\alpha \in [0,1]} (\alpha(1-\gamma) - \Psi_{a,b}(\alpha, q_\gamma))$ in order for the power to tend to one (see proof of Theorem 2). In other words, if we choose a critical value k_n for our test statistic with $k_n \rightarrow \infty$ and $k_n = o(\sqrt{n})$ then for each of the finitely many tests that we will have to perform the level tends to 0 and the power tends to 1. For more details of this see Appendix, section 5.4.

4 SIMULATIONS AND APPLICATIONS

4.1 SIMULATIONS.

Data was simulated under the null hypothesis using an AR(1) model with $\phi = .9$ and three different variance functions, the two pictured in figure 2.1 as well as a strictly monotone function where the modes of the bimodal function shown in figure 2.1 are connected linearly. Errors were generated from a standard normal distribution. We used $\beta = 0.2$, and the bootstrap procedure generally suggested a value in the neighborhood of $\delta = .3$. Under the

null hypothesis with a flat set, the empirical level based on 1000 simulations was .049 using an $\alpha = .05$. The p-value was calculated using the limiting Brownian Bridge distribution. Under the strictly monotone null, the level was approximately .02. Under the alternative hypothesis, the power of the test was .809 in 1000 simulations. Using errors that are t-distributed with 5 degrees of freedom decreased the power of the test to .646. It is not surprising that the power decreased some amount with a heavier tailed distribution, though the results are still promising.

We also considered the full tvAR model, using a first order model with $\theta(u)$ being quadratic. The parameter functions and an example of a simulated example are pictured in figure 4.2. Using normal errors, the power of the test was estimated to be .832 in 1000 simulations, insignificantly larger than under the constant AR parameter simulations above. The AR parameter function was estimated using a local least squares estimator (see Dahlhaus and Giraitis, 1998) with $\hat{\theta}(u)$ coming from a quadratic fit of the local estimates.

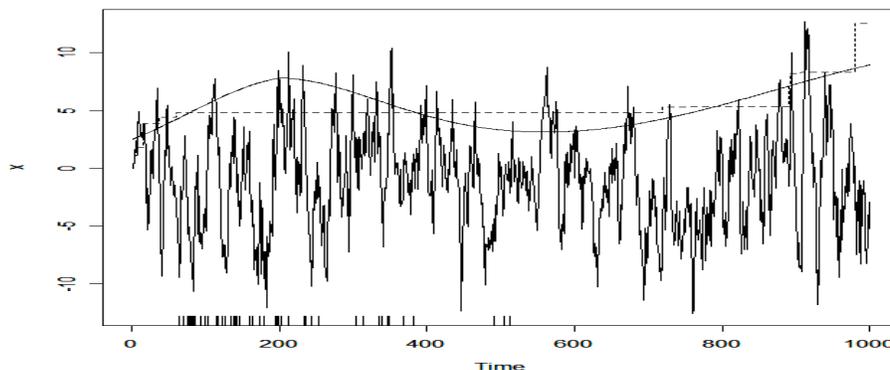


Figure 4.1: A simulated data set from AR(1) with normal errors. The variance function used (solid) and its estimate (dashed) are overlaid. A rug plot of the location of the largest residuals is also shown.

As our squared residuals are estimates of the quantity $\sigma^2(\frac{t}{T})\epsilon_t^2$, we might consider looking at $\log(\hat{\eta}^2)$ which is approximating $\log(\sigma^2(\frac{t}{T})) + \log(\epsilon_t^2)$, i.e. a regression function plus (mean zero) white noise. Thus, we might use a test for modality designed for the regression case after computing the log of the squared residuals. In particular, we consider the method of Davies and Kovac (2001). For residuals close to zero, taking the log results in outliers in the data, and thus the multi-resolution criterion struggles. We replace this with a bandwidth, using their R package *ftnonpar* (Davies, 2010). We sample under the null hypothesis, using

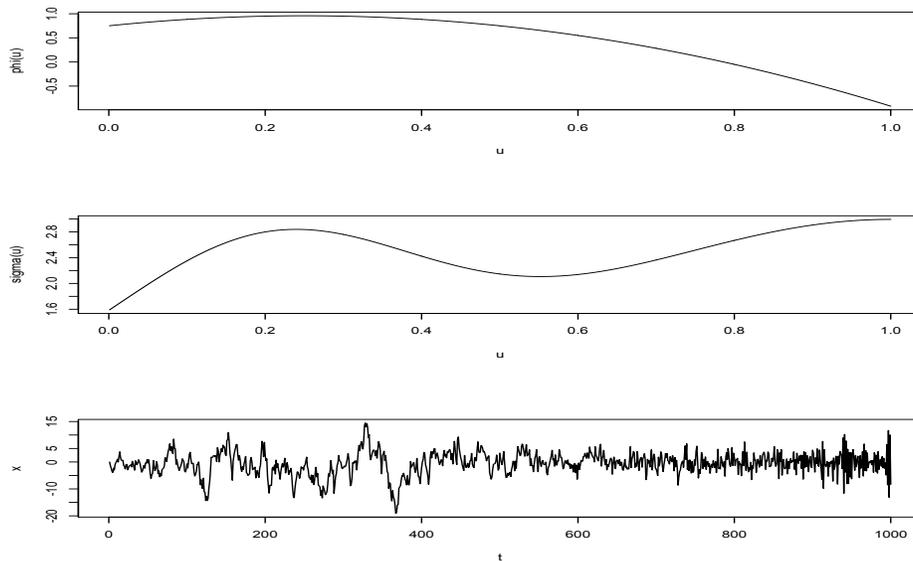


Figure 4.2: The AR parameter function (top), the variance function (middle) and a realization of this model (bottom).

the same process as in the first simulation above, and perform a grid search over bandwidths to identify the bandwidth that achieves a type I error rate of .05 (so that the method recognizes that there are no local extrema approximately 95 percent of the time). We then sample under the alternative, recording the power as the proportion of times the method recognizes there are (at least) two local extrema, which we calculate to be .386, well below the results using the proposed method.

Our final simulations explore the behavior of the test under model misspecification. We use a constant variance function, and an AR(1) model that is a white noise sequence ($\theta(u) = 0$) except in two intervals where $\theta(u)$ smoothly increases up to a height h and back to 0 over about 10 percent of the total time interval ($T=1000$). This is followed by identical behavior

h	.1	.2	.3	.4	.5	.6	.7	.8	.9
β	0.05	0.05	0.06	0.07	0.10	0.13	0.20	0.30	0.46

Table 1: The type II error probability under a misspecified model of constant AR parameter vs the maximum of $\theta(u)$ on the two intervals where it is not 0.

a short time later (allowing θ to briefly return to 0) to give the impression of an anti-mode should one naïvely fit a constant AR parameter model. The behavior of the test is given in table 4.1. Unsurprisingly, the more pronounced these peaks are, the more likely the test

will conclude the presence of an anti-mode. However, the effect is relatively small even when $h = .5$. This is partly due to $\hat{\theta}$ being wrong essentially everywhere, and thus the residuals are larger than they should be on the white noise segments of the data. The solution is of course to fit the appropriate model, having time varying autoregressive parameters.

4.2 APPLICATIONS.

4.2.1 SEISMIC DATA AND REFLECTION WAVES We consider testing for modality of a seismic time series. Seismograms generated by earthquakes are known to include a reflection (pP-)wave that follows the primary (P-)wave, arriving at a delay due to the depth of the event. Shumway et al. (2004) proposed estimation for the time delay which can be used to estimate the depth of the event, as well as possibly give some information towards the type of event, as shallow explosions should not exhibit such a reflection wave. Consider the filtered seismogram shown in figure 4.3, generated by an earthquake in northern Chile. For more information about the time series, see Shumway et al. (2004). We are interested in testing if the estimated pP-wave is in fact a second mode from the P-wave.

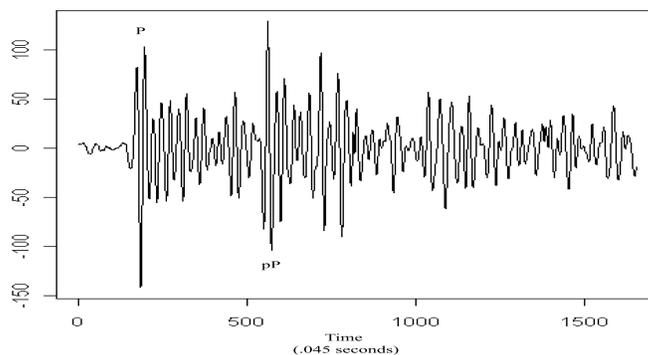


Figure 4.3: Seismic reading of an earthquake centered in northern Chile.

Selecting observation 184 as the primary mode (the maximum of the absolute series), a decreasing estimate of the variance was fit to the squared residuals from the estimated mode on using isotonic regression. The bootstrap procedure suggested $\delta = .125$. The resulting set before trimming extended from time points 204 to 795, i.e. from just after the first mode to beyond the mode caused by the arrival of the pP-wave. Again using $\beta = .2$, the resulting set was the interval $[293, 706]$, which contains the antinode as hoped. The test statistic was 3.33, with a corresponding p-value of $2 * 10^{-10}$ based on the limiting distribution. Thus,

the method correctly identifies the second mode resulting from the reflection of the p-wave. As such, we would suspect the origin of the event was an earthquake, rather than a shallow explosion, nuclear or otherwise.

4.2.2 FOOTSTEP COUNTING Our second application regards the problem of tracking elephants via geo-phones. Wood et. al. (2005) were interested in tracking the size of elephant populations in forested areas of their habitat where aerial surveillance is not possible.

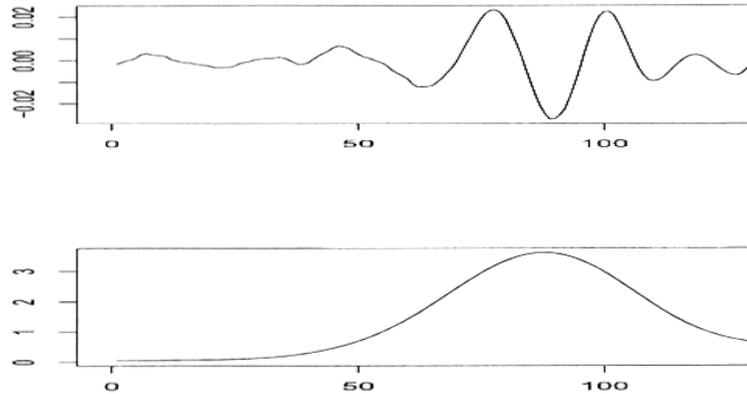


Figure 4.4: A geo-phone recording of an elephant’s footfall (top) and the estimated time-dependent variance function (bottom) assuming unimodality.

In order to differentiate between single elephants and small groups of elephants, one possible solution is to attempt to count the number of footsteps in a certain time frame. We can model footsteps using an autoregression with unimodal variance function. An example of a single footfall is provided in figure 4.4. Figure 4.5 is a recording taken from a geo-phone of

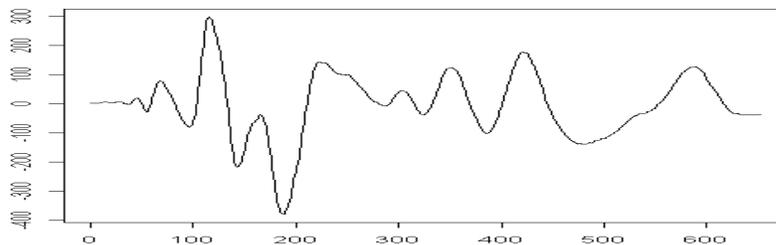


Figure 4.5: A geo-phone recording of footsteps from two elephant bulls.

seismic vibrations created by the footsteps of two elephant bulls ($n=758$). An AR(1) model

is used, with $\hat{\phi} = .9975$, and the mode is chosen to be at observation 186. The bootstrap procedure suggests $\delta = 1$. The set on which the test was run, using $\beta = .2$, was [255, 437], which graphically matches the apparent anti-mode in the data. The resulting test statistic is 2.65, with corresponding p-value $7 * 10^{-7}$, thus confirming the presence of two footsteps in the given window, corresponding to the fact that there were two elephant bulls passing the geo-phone array.

4.2.3 PEAK PICKING AND BEAT DETECTION We consider the problem of finding individual drum hit events in an audio recording, a problem that is related to both tempo estimation and onset detection. Here, we consider the problem of automatically transcribing a single measure of isolated drums, commonly referred to as a ‘drum break’. The repetitive looping of these drum breaks are ubiquitous in modern popular music, especially the genre of rap. In the example below, we consider the first measure of the drum break of the song ‘Amen Brother’ by the band The Winstons. The notated drum pattern is shown in Figure 4.6, with the sound recording pictured in Figure 4.7 (T=4796). Grid lines are added according to where the events occur as stated in the notated drum pattern, though there is some variability as to when the actual event occurs. In this particular example, we have additional information regarding the possible location of modes. Should they exist, they should fall roughly on 16^{th} notes. Exploiting this information, we apply our method repeatedly over dyadic subsets of the data, half notes, quarter notes, eighth notes, and sixteenth notes. We assume a model of the form

$$Y_{t;T} = \theta\left(\frac{t}{T}\right)Y_{t-1;T} + \sigma\left(\frac{t}{T}\right)\epsilon_{t;T}$$

and θ is estimated by a windowed estimate that is then smoothed via a normal kernel (we might also assume that θ is piecewise constant for such data). The squared residuals are then calculated based on this estimate.



Figure 4.6: Notation for measure one of the ‘Amen Brother’ break. The top notes represent the ride cymbal, the middle notes correspond to the snare drum, and the bottom notes to the bass drum.

We mention only a selection of the results of the iterative procedure for brevity. On the first

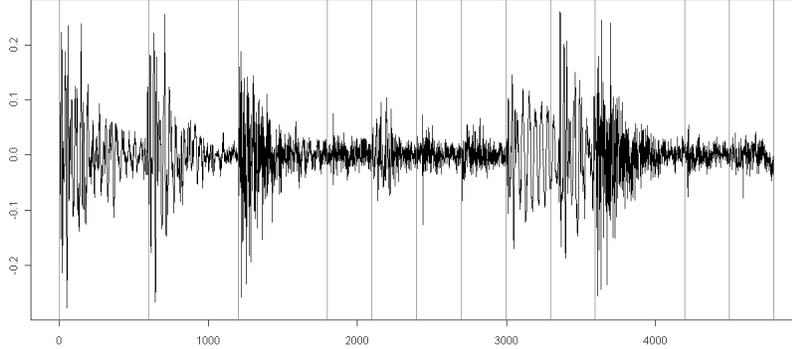


Figure 4.7: Data for measure one of the ‘Amen’ break. Grid lines correspond to the notated pattern.

eighth note segment of the measure, we calculate a p-value of .14, leading us to conclude there is no drum hit between the first and second eighth notes, consistent with the notation. Similarly, with the second eighth note, a p-value of .433, and the third, a p-value of .299 is found, again suggesting the lack of a drum being played between these eighth notes. Testing whether there is in fact a hit on the 4th eighth note, which may be questionable based on a visual inspection of the series, the resulting p-value is .019. This was based on testing the interval that ran from the 3rd eighth note to a sixteenth note past the 4th eighth note. Similarly, the test detects the presence of the snare drum being played on the sixteenth note following the 4th eighth note.

5 APPENDIX

5.1 ASSUMPTIONS

First we formulate the assumptions underlying our results. For a function $h : [a, b] \rightarrow \mathbf{R}$ we denote

$$\|h\|_{\infty} := \sup_{u \in [a, b]} |h(u)| \quad \text{and} \quad \|h\|_n^2 := \frac{1}{n} \sum_{t \in [aT, bT]} h^2\left(\frac{t}{T}\right).$$

Let \mathcal{H} denote a class of functions defined on $[a, b]$. For a given $\delta > 0$, a δ -bracket in \mathcal{H} with respect to $\|\cdot\|_n$ is of the form $\mathcal{B}_k = [\underline{g}, \bar{g}] := \{h \in \mathcal{H} : \underline{g} \leq h \leq \bar{g}\}$ for pairs of functions $\underline{g} \leq \bar{g}$

with $\|\bar{g} - g\|_n \leq \delta$. Then the minimal number $N = N_B(\delta, \mathcal{H})$ of brackets $\mathcal{B}_1, \dots, \mathcal{B}_N$ with $\mathcal{H} \subset \bigcup_{k=1}^N \mathcal{B}_k$ is called a *bracketing covering number*, and $\log N_B(\delta, \mathcal{H})$ is called the *metric entropy with bracketing* of \mathcal{H} with respect to $\|\cdot\|_n$.

ASSUMPTIONS. (i) The process $Y_t = Y_{t,T}$ has an MA-type representation

$$Y_{t,T} = \sum_{j=0}^{\infty} a_{t,T}(j) \epsilon_{t-j}$$

where $\epsilon_t \sim_{i.i.d.} (0, 1)$. The distribution F of ϵ_t has a strictly positive Lipschitz continuous Lebesgue density f . The function $\sigma(u)$ in (1) is Lipschitz continuous with $0 < m_* < \sigma(u) \leq \sigma(m) = m^* < \infty$ for all $u \in [0, 1]$ and a unique global maximum $m \in [0, 1]$.

(ii) The coefficients $a_{t,T}(\cdot)$ of the MA-type representation of $Y_{t,T}$ given in (i) satisfy

$$\sup_{1 \leq t \leq T} |a_{t,T}(j)| \leq \frac{K}{\ell(j)}$$

for some $K > 0$, and where for some $\kappa > 0$ we have $\ell(j) = j(\log j)^{1+\kappa}$ for $j > 1$ and $\ell(j) = 1$ for $j = 0, 1$.

(iii) There exists a class \mathcal{G} with

$$\widehat{\theta}_{n,k}(\cdot) - \theta_k(\cdot) \in \mathcal{G}, \quad k = 1, \dots, p,$$

such that $\sup_{g \in \mathcal{G}} \|g\|_{\infty} < \infty$ and for some $\gamma \in [0, 1)$ and $C, c > 0$ and for all $\delta > 0$

$$\log N_B(\delta, \mathcal{G}) \leq \begin{cases} C \delta^{-\gamma}, & \text{for } \frac{c}{n} < \gamma < 1 \\ C \log(\delta^{-1}), & \text{for } \gamma = 0. \end{cases}$$

(iv) For $k = 1, \dots, p$ we have $\frac{1}{T} \sum_{t=1}^T (\widehat{\theta}_{n,k}(\frac{t}{T}) - \theta_k(\frac{t}{T}))^2 = O_P(m_n^{-1})$ as $T \rightarrow \infty$ with $\frac{n^{1/2}}{m_n^2 \log n} = o(1)$.

We refer to Chandler and Polonik (2012) for a discussion of these assumptions. Here we only mention that the wavelet estimators $\widehat{\theta}_n$ of Dahlhaus et al. (1999) satisfy conditions (iii) and (iv). (See Chandler and Polonik, 2010 for more details.)

A crucial ingredient into the proofs of our results is the following theorem that is proven (for $[a, b] = [0, 1]$) in Chandler and Polonik (2012). Their proof directly translates to the case $[a, b] \subset [0, 1]$.

Theorem A (Chandler and Polonik, 2010). *Let $\gamma \in [0, 1]$ and suppose that $0 \leq a < b \leq 1$ are non-random. Assume further that $E|\epsilon|^k < C^k$ for some $C > 0$ and all $k = 1, 2, \dots$. Then, under assumptions (i) - (iv), with $\frac{n^{1/2}}{m_n^2 \log n} = o(1)$ we have*

$$\sqrt{n} \sup_{\alpha \in [0, 1]} \left| \widehat{G}_{n,\gamma}(\alpha) - G_{n,\gamma}(\alpha) + c(\alpha)(G_{n,\gamma}(1) - EG_{n,\gamma}(1)) \right| = o_p(1), \quad (8)$$

where

$$c(\alpha) = \frac{\int_a^{a+\alpha(b-a)} [f_u(q_\gamma) + f_u(-q_\gamma)] du}{\int_a^b [f_u(q_\gamma) + f_u(-q_\gamma)] du},$$

and $f_u(x) = \frac{1}{\sigma(u)} f\left(\frac{x}{\sigma(u)}\right)$. Under the null-hypothesis $\sigma(u) \equiv \sigma_0 > 0$ for $u \in [a, b]$ we have $c(\alpha) = \alpha$. Moreover, in case the AR-parameter in model (1) are constant, and \sqrt{n} -consistent estimators are used, then the moment assumptions on the innovations can be significantly relaxed to $E\epsilon_t^2 < \infty$.

5.2 PROOFS

5.2.1 SKETCH OF THE PROOF OF THEOREM 1 First we consider that case of $[a, b]$ and m being known. The above theorem from Chandler and Polonik (2012) says that under certain assumptions, $\widehat{G}_{n,\gamma}(\alpha)$ can be approximated well by a sum of independent random variables, $G_{n,\gamma}(\alpha)$ plus a correction term $-c(\alpha)(G_{n,\gamma}(1) - EG_{n,\gamma}(1))$ that in fact accounts for the estimation of q_γ . Thus, under the null hypothesis of a constant variance function (on $[a, b]$), we obtain that the process $\sqrt{\frac{n}{\gamma(1-\gamma)}}(\widehat{G}_{n,\gamma}(\alpha) - \gamma\alpha)$ has the same asymptotic behavior as $\sqrt{\frac{n}{\gamma(1-\gamma)}}(G_{n,\gamma}(\alpha) - \gamma\alpha - \alpha(G_{n,\gamma}(1) - \gamma))$. And under H_0 the latter obviously converges to a standard Brownian Bridge. This implies the asserted asymptotic distribution.

First notice that the estimation of the set is not influenced by the consistent estimation of m since by assumption the mode is unique and thus the mode does not lie in the theoretical optimal set $[a, b]$, and since by assumption $P([\widehat{a}, \widehat{b}] \subset [a, b]) \rightarrow 1$, the estimation of the mode does not have an effect asymptotically.

To see that we can use a consistent estimator $[\widehat{a}, \widehat{b}]$ instead of $[a, b]$, first notice the obvious fact that n , q_γ , \widehat{q}_γ and the parametrization itself, used in the definition of $\widehat{G}_{n,\gamma}(\alpha)$, all depend on the interval used in the definition. In order to make this more clear, we denote by $\widehat{G}_{n,\gamma}(\alpha; [c, d])$ the quantity defined as $\widehat{G}_{n,\gamma}(\alpha)$ but with $[a, b]$ replaced by $[c, d] \subset [0, 1]$, and let $\overline{\Psi}_{c,d}(\alpha, z) = \alpha - \Psi_{c,d}(\alpha, z)$. Define

$$Z_{n_{c,d},\gamma}(\alpha; [c, d]) = \sqrt{n} \left[G_{n,\gamma}(\alpha; [c, d]) - \overline{\Psi}_{c,d}(\alpha, q_{\gamma;c,d}) - c_{c,d}(\alpha) (G_{n_{c,d},\gamma}(1; [c, d]) - \overline{\Psi}_{c,d}(\mathbf{1}, q_{\gamma;c,d})) \right],$$

where $n_{c,d}$ denotes the number of observations in $[c, d]$, $q_{\gamma;c,d}$ and $c_{c,d}(\alpha)$ are defined as q_γ and $c(\alpha)$ (cf. (4) and in Theorem A, respectively), but with $[a, b]$ replaced by $[c, d]$. To ease the notation we write $\hat{n} = n_{\hat{a},\hat{b}}$ and $\hat{c}(\alpha) = c_{\hat{a},\hat{b}}(\alpha)$. We now indicate that $\max\{|\hat{a} - a|, |\hat{b} - b|\} = o_P(1)$ for some estimate $[\hat{a}, \hat{b}]$ with $P([\hat{a}, \hat{b}] \subset [a, b]) \rightarrow 1$ implies that we have

$$\sup_{\alpha \in [0,1]} \left| \sqrt{\hat{n}} \left(\hat{G}_{\hat{n},\gamma}(\alpha; [\hat{a}, \hat{b}]) - \bar{\Psi}_{\hat{a},\hat{b}}(\alpha, q_{\gamma;\hat{a},\hat{b}}) \right) - Z_{\hat{n},\gamma}(\alpha; [\hat{a}, \hat{b}]) \right| = o_P(1), \quad (9)$$

and that

$$\sup_{\alpha \in [0,1]} |Z_{\hat{n},\gamma}(\alpha; [\hat{a}, \hat{b}]) - Z_{n,\gamma}(\alpha; [a, b])| = o_P(1). \quad (10)$$

These two approximations implies the desired result by noticing that for any set $[c, d] \subset [a, b]$ (whether random or not) we have under the null hypothesis $\sigma^2(u) \equiv c$ for $u \in [a, b]$ that $\bar{\Psi}_{c,d}(\alpha, z) = \alpha - \Psi_{c,d}(\alpha, z) = \alpha(1 - \gamma)$ and $c_{c,d}(\alpha) = \alpha$. The proof of (9) follows similar ideas as the proof of Theorem A (see Chandler and Polonik, 2012), and (10) follows from asymptotic equicontinuity of the Kiefer-Müller-type process

$$\nu_n(\alpha, z) = \frac{1}{\sqrt{n}} \sum_{s=1}^{\lfloor \alpha n \rfloor} \left[\mathbf{1}\left\{ \sigma\left(a + \frac{s}{T}\right) \epsilon_t \leq z \right\} - P\left\{ \sigma\left(a + \frac{s}{T}\right) \epsilon_t \leq z \right\} \right], \quad (11)$$

by using the fact that $|q_{\gamma;\hat{a},\hat{b}} - q_{\gamma;a,b}| = o_P(1)$. The latter follows easily by observing that under our assumptions $\Psi_{a,b}(1, z)$ is continuous in a and b , uniformly over z , and that $\gamma \rightarrow \Psi_{a,b}^{-1}(1, \gamma)$ is continuous. The fact that under our assumption the equicontinuity of (11) holds follows from empirical process theorem. (For instance, it is a special case of Lemma 1 in Chandler and Polonik (2012).)

5.2.2 PROOF OF THEOREM 2 Let $n = n_{c,d}$ and $\hat{n} = \hat{n}_{\hat{a},\hat{b}}$ denote the numbers of observations in $[c, d]$ and $[\hat{a}, \hat{b}]$, respectively. Let further $\Psi(\alpha, q_\gamma) = \Psi_{c,d}(\alpha, q_\gamma)$.

Properties (9) and (10) say that $\sqrt{\hat{n}} \left(\hat{G}_{\hat{n},\gamma}(\alpha; [\hat{a}, \hat{b}]) - \gamma\alpha \right)$ has the same asymptotic behavior as $\sqrt{n} \left(G_{n,\gamma}(\alpha; [c, d]) - \gamma\alpha - \alpha \left(G_{n,\gamma}(1; [c, d]) - \gamma \right) \right)$. We have

$$\mathbb{E}G_{n,\gamma}(\alpha; [c, d]) = \frac{1}{n} \sum_{t=\lceil cT \rceil}^{\lceil dT \rceil + \lfloor \alpha n \rfloor - 1} P\left(\epsilon_t^2 \geq \frac{q_\gamma^2}{\sigma^2(\frac{t}{T})}\right)$$

$$\begin{aligned}
&= \alpha - \frac{1}{n} \sum_{t=\lceil cT \rceil}^{\lceil dT \rceil + \lfloor \alpha n \rfloor - 1} \left[F\left(\frac{q_\gamma}{\sigma(\frac{t}{T})}\right) - F\left(\frac{-q_\gamma}{\sigma(\frac{t}{T})}\right) \right] + O\left(\frac{1}{n}\right) \\
&= \alpha - \Psi(\alpha, q_\gamma) + O\left(\frac{1}{n}\right),
\end{aligned}$$

where the $O(\frac{1}{n})$ -term comes from the fact that by our assumptions $u \rightarrow F\left(\frac{q_\gamma}{\sigma(u)}\right) - F\left(\frac{-q_\gamma}{\sigma(u)}\right)$ is of bounded variation. Thus we can write

$$\begin{aligned}
&\sqrt{n}(G_{n,\gamma}(\alpha) - \gamma\alpha - \alpha(G_{n,\gamma}(1) - \gamma)) \\
&= \sqrt{n}(G_{n,\gamma}(\alpha) - \mathbb{E}G_{n,\gamma}(\alpha) - \alpha(G_{n,\gamma}(1) - \gamma)) - \sqrt{n}(\Psi(\alpha, q_\gamma) - \alpha(1 - \gamma)) + O\left(\frac{1}{n}\right).
\end{aligned}$$

The fact that under model (1) we have

$$\sup_{\alpha \in [0,1]} \left| \sqrt{n}(G_{n,\gamma}(\alpha) - \mathbb{E}G_{n,\gamma}(\alpha) - \alpha(G_{n,\gamma}(1) - \gamma)) \right| = O_P(1)$$

follows from the fact that the process $\sqrt{n}(G_{n,\gamma}(\alpha) - \mathbb{E}G_{n,\gamma}(\alpha) - \alpha(G_{n,\gamma}(1) - \gamma))$ converges weakly (in the sense of Hoffmann-Jørgensen) to a Gaussian process $D(\alpha)$ with mean zero and covariance $\text{Cov}(D(\alpha), D(\beta)) = (\alpha \wedge \beta - \Psi(\alpha \wedge \beta, q_\gamma)) - (\alpha - \Psi(\alpha, q_\gamma))(\beta - \Psi(\beta, q_\gamma))$. This weak convergence follows from asymptotic stochastic equicontinuity of the process (which in turn follows from Lemma 1 of Chandler and Polonik, 2010) together with the asymptotic normality of the finite dimensional distributions which follows from a straightforward application of Lindeberg-Feller CLT coupled with the Cramer-Wold-device. Thus, if $\sup_{\alpha \in [0,1]} |\Psi(\alpha, q_\gamma) - \alpha(1 - \gamma)| \neq 0$ then the above implies that $\sup_{\alpha \in [0,1]} |\sqrt{n}(G_{n,\gamma}(\alpha) - \gamma\alpha - \alpha(G_{n,\gamma}(1) - \gamma))|$ is not stochastically bounded, and the assertion follows by utilizing (9).

5.2.3 PROOF OF LEMMA 1. We first replace **(A1)** and **(A2)** by the stronger assumptions

$$\mathbf{(A1)'} \sup_{u \in [0,1]} |\widehat{\Sigma}(u) - \Sigma(u)| = o(1) \text{ a.s.}$$

$$\mathbf{(A2)'} |\widehat{m} - m| = o(1) \text{ a.s.}$$

Under these stronger assumptions we first show strong consistency of the set estimate. Then we utilize this result together with a well-known characterization of convergence in probability via the almost sure convergence of sub-subsequences to obtain the desired result.

We first show that

$$\left| \sup_{\lambda \geq 0} \widehat{\Delta}_{\widehat{m}}(\delta; \lambda) - \sup_{\lambda \geq 0} \Delta_m(\delta; \lambda) \right| = o(1) \text{ a.s.} \quad (12)$$

To this end, observe that

$$\widehat{H}_{\widehat{m}, \lambda}(u) = H_{m, \lambda}(u) + \widehat{\Sigma}_{\widehat{m}}^*(u) - \Sigma_m^*(u).$$

Further, because taking the greatest convex minorant (and also the least concave majorant) is a contraction with respect to the supremum-norm, assumption **(A1)'** implies that $\|\widehat{\Sigma}_{\widehat{m}}^*(\cdot) - \Sigma_m^*(\cdot)\|_\infty \leq \|\widehat{\Sigma}(\cdot) - \Sigma(\cdot)\|_\infty = o(1)$ a.s.. This, together with assumption **(A3)** now implies that $\|\widehat{\Sigma}_{\widehat{m}}^* - \Sigma_m^*\|_\infty = o(1)$ a.s., and thus we have

$$\widehat{D}_n = \sup_{\lambda > 0} \sup_{u \in [0, 1]} |\widehat{H}_{\widehat{m}, \lambda}(u) - H_{m, \lambda}(u)| = o(1) \text{ a.s.} \quad (13)$$

Write

$$\begin{aligned} \widehat{\Delta}_{\widehat{m}}(\delta; \lambda) = & \sup_{[c, d] \subset [0, \widehat{m}]} \left\{ |c - d| : -H_{m, \lambda}(c, t) \leq \delta + \widehat{H}_{\widehat{m}}(c, t) - H_m(c, t) \right. \\ & \left. \text{and } H_{m, \lambda}(t, d) \leq \delta + \widehat{H}_{\widehat{m}}(t, d) - H_m(t, d); \text{ for all } t \in [c, d] \right\} \end{aligned}$$

and a similar rewriting can be done for $\Delta_{\widehat{m}}(\delta; \lambda)$. Consequently, we have for each $\omega \in A$ and for n sufficiently large (such that $\delta - 2\|\widehat{\Sigma}_{\widehat{m}}^* - \Sigma_m^*\|_\infty \geq 0$), that

$$\Delta_{\widehat{m}}(\delta - 2D_n; \lambda) \leq \widehat{\Delta}_{\widehat{m}}(\delta; \lambda) \leq \Delta_{\widehat{m}}(\delta + 2D_n; \lambda). \quad (14)$$

Assumption **(A3)** and the assumed consistency of \widehat{m} now imply (12).

Now let A denote the event of probability 1 on which **(A1)'** and **(A2)'** hold. Let further $\widehat{\lambda}_\delta^*$ be such that $\widehat{\Delta}_{\widehat{m}}(\delta; \widehat{\lambda}_\delta^*) = \sup_{\lambda \geq 0} \widehat{\Delta}_{\widehat{m}}(\delta; \lambda)$ and let $[\widehat{a}_\delta^*, \widehat{b}_\delta^*]$ be the corresponding maximizing set, i.e. by definition of $\widehat{\Delta}_{\widehat{m}}(\delta; \lambda)$ we have

$$\widehat{\Delta}_{\widehat{m}}(\delta; \widehat{\lambda}_\delta^*) = |\widehat{a}_\delta^* - \widehat{b}_\delta^*| \quad (15)$$

and

$$-\widehat{H}_{\widehat{m}, \widehat{\lambda}_\delta^*}(\widehat{a}_\delta^*, t) \leq \delta \quad \text{for all } t \in [\widehat{a}_\delta^*, \widehat{b}_\delta^*]; \quad (16)$$

$$\widehat{H}_{\widehat{m}, \widehat{\lambda}_\delta^*}(t, \widehat{b}_\delta^*) \leq \delta \quad \text{for all } t \in [\widehat{a}_\delta^*, \widehat{b}_\delta^*]. \quad (17)$$

We will show below, that for $\omega \in A$ fixed, there exists $L > 0$ such that $\widehat{\lambda}^* \leq L$ for $n \geq n_0$. Thus, we can assume that $\{(\widehat{\lambda}_\delta^*, \widehat{a}_\delta^*, \widehat{b}_\delta^*)\}_n \in [0, L] \times [0, 1]^2$, and since $[0, L] \times [0, 1]^2$ is compact, we can assume that $\{(\widehat{\lambda}_\delta^*, \widehat{a}_\delta^*, \widehat{b}_\delta^*)\}_n$ converges. (Otherwise we have to work with subsequences.) Let $(\widetilde{\lambda}_\delta, \widetilde{a}_\delta, \widetilde{b}_\delta)$ denote the corresponding limit.

It follows from (13) and (15) - (17) that $|\widetilde{a}_\delta - \widetilde{b}_\delta| = |a_\delta^* - b_\delta^*|$, and that $-H_{m, \widetilde{\lambda}}(\widetilde{a}_\delta, t) \leq \delta$ for all $t \in [\widetilde{a}_\delta, \widetilde{b}_\delta]$, and $H_{m, \widetilde{\lambda}}(t, \widetilde{b}_\delta) \leq \delta$ for all $t \in [\widetilde{a}_\delta, \widetilde{b}_\delta]$. The uniqueness assumption **(A4)** implies that $[\widetilde{a}_\delta, \widetilde{b}_\delta] = [c_{\lambda_\delta^*}, d_{\lambda_\delta^*}]$. This implies almost sure consistency of $[\widehat{a}_\delta^*, \widehat{b}_\delta^*]$ as an estimator for $[a_\delta^*, b_\delta^*]$.

The existence of an upper bound for $\widehat{\lambda}^*$ can be seen as follows. By assumption, $\sigma^2(\cdot)$ is bounded, and thus for $\lambda > M^* = \max_{u \in [0, 1]} \sigma^2(u)$ we have $-H_{m, \lambda}(c, d) \geq (\lambda - M^*)|c - d| \geq 0$. Consequently, we have $-H_{m, \lambda}(c, d) \leq \delta$ only if $|c - d| \leq \delta/(\lambda - M^*)$. This implies that the length of the optimal interval in the definition of $\Delta_m(\delta; \lambda)$ tends to zero as $\lambda \rightarrow \infty$. Utilizing (12) we can find an L such that for n sufficiently large, the largest empirical optimal interval is attained at a level $\lambda < L$. This completes the proof for the strong consistency under the assumption **(A1)**'.

The proof of weak consistency utilizes the well know fact that $X_n = o_P(1)$ iff every subsequence of X_n has a further subsequence that converges a.s. Consider the sequence $(\widehat{\lambda}_\delta^*, \widehat{a}_\delta^*, \widehat{b}_\delta^*)$ and pick a subsequence $(n') \subset (n)$. This provides a corresponding subsequences $Z_{n'} = \sup_{u \in [0, 1]} |\widehat{\Sigma}(u) - \Sigma(u)|$ and $W_{n'} = |\widehat{m} - m|$. Since $Z_n = o_P(1)$ and $W_n = o_P(1)$, there exists a further subsequence (n'') such that both $\{Z_{n''}\}$ and $\{W_{n''}\}$ converge to zero almost surely. Then, utilizing the above proof we can conclude that the sub-subsequence of $\{(\widehat{\lambda}_\delta^*, \widehat{a}_\delta^*, \widehat{b}_\delta^*)\}_{n''}$ converges a.s. to $(\lambda_\delta^*, a_\delta^*, b_\delta^*)$. For any sub-subsequence this is the same limit, and thus we know that every subsequence of $\{(\widehat{\lambda}_\delta^*, \widehat{a}_\delta^*, \widehat{b}_\delta^*)\}$ has a further subsequence that converges a.s. to $(\lambda_\delta^*, a_\delta^*, b_\delta^*)$. By the above fact this means that $(\widehat{\lambda}_\delta^*, \widehat{a}_\delta^*, \widehat{b}_\delta^*)$ converges in probability to $(\lambda_\delta^*, a_\delta^*, b_\delta^*)$.

5.2.4 PROOF OF FACT 1. We have

$$\begin{aligned} \widehat{\Sigma}\left(\frac{k}{T}\right) - \Sigma\left(\frac{k}{T}\right) &= \frac{1}{T} \sum_{t=1}^k \left[(Y_t - \widehat{\boldsymbol{\theta}}_n\left(\frac{t}{T}\right)' \mathbf{Y}_{t-1})^2 - (Y_t - \boldsymbol{\theta}\left(\frac{t}{T}\right)' \mathbf{Y}_{t-1})^2 \right] \\ &= -\frac{2}{T} \sum_{t=1}^k (\widehat{\boldsymbol{\theta}}_n\left(\frac{t}{T}\right) - \boldsymbol{\theta}\left(\frac{t}{T}\right))' \mathbf{Y}_{t-1} \sigma\left(\frac{t}{T}\right) \epsilon_t + \frac{1}{T} \sum_{t=1}^k \left((\widehat{\boldsymbol{\theta}}_n\left(\frac{t}{T}\right) - \boldsymbol{\theta}\left(\frac{t}{T}\right))' \mathbf{Y}_{t-1} \right)^2. \end{aligned}$$

Consequently,

$$\begin{aligned}
& \mathbb{E} \sup_{1 \leq k \leq T} \left| \widehat{\Sigma}\left(\frac{k}{T}\right) - \Sigma\left(\frac{k}{T}\right) \right| \\
& \leq \mathbb{E} \frac{2}{T} \sum_{t=1}^T \left| \left(\widehat{\boldsymbol{\theta}}_n\left(\frac{t}{T}\right) - \boldsymbol{\theta}\left(\frac{t}{T}\right) \right)' \mathbf{Y}_{t-1} \sigma\left(\frac{t}{T}\right) \epsilon_t \right| + \mathbb{E} \frac{1}{T} \sum_{t=1}^T \left(\left(\widehat{\boldsymbol{\theta}}_n\left(\frac{t}{T}\right) - \boldsymbol{\theta}\left(\frac{t}{T}\right) \right)' \mathbf{Y}_{t-1} \right)^2. \quad (18)
\end{aligned}$$

It will become clear from the below that these expectations exist. The proof is complete if we have shown that the two expectations in the last line converge to zero as $n \rightarrow \infty$. By utilizing the Cauchy-Schwarz inequality we have

$$\begin{aligned}
& \mathbb{E} \frac{1}{T} \sum_{t=1}^T \left| \left(\widehat{\boldsymbol{\theta}}_n\left(\frac{t}{T}\right) - \boldsymbol{\theta}\left(\frac{t}{T}\right) \right)' \mathbf{Y}_{t-1} \sigma\left(\frac{t}{T}\right) \epsilon_t \right| \\
& \leq \sum_{j=1}^p \mathbb{E} \left(\sqrt{\frac{1}{T} \sum_{t=1}^T \left(\widehat{\theta}_{n,j}\left(\frac{t-j}{T}\right) - \theta_j\left(\frac{t-j}{T}\right) \right)^2} \sqrt{\frac{1}{T} \sum_{t=1}^T Y_{t-j}^2 \sigma^2\left(\frac{t}{T}\right) \epsilon_t^2} \right) \\
& \leq \sum_{j=1}^p \sqrt{\mathbb{E} \left(\frac{1}{T} \sum_{t=1}^T \left(\widehat{\theta}_{n,j}\left(\frac{t-j}{T}\right) - \theta_j\left(\frac{t-j}{T}\right) \right)^2 \right)} \sqrt{\mathbb{E} \left(\frac{1}{T} \sum_{t=1}^T Y_{t-j}^2 \sigma^2\left(\frac{t}{T}\right) \epsilon_t^2 \right)} \\
& \leq m^* \sum_{j=1}^p \sqrt{\mathbb{E} \left(\frac{1}{T} \sum_{t=1}^T \left(\widehat{\theta}_{n,j}\left(\frac{t-j}{T}\right) - \theta_j\left(\frac{t-j}{T}\right) \right)^2 \right)} \sqrt{\sup_{-p \leq t \leq T} \mathbb{E} Y_t^2}.
\end{aligned}$$

Assumptions (i) and (ii) imply that $\sup_{-p \leq t \leq T} \mathbb{E} Y_t^2 < \infty$ (cf. Chandler and Polonik, 2010, proof of (22)). Since further $\frac{1}{T} \sum_{t=1}^T \left(\widehat{\theta}_{n,j}\left(\frac{t-j}{T}\right) - \theta_j\left(\frac{t-j}{T}\right) \right)^2$ is bounded and converges to zero stochastically (cf. assumptions (iii) and (iv)), it follows that the first expectation in (18) converges to 0 as $n \rightarrow \infty$. The second expectation in (18) can be treated similarly by the following crude estimation. We have

$$\begin{aligned}
& \frac{1}{T} \sum_{t=1}^T \left[\left(\boldsymbol{\theta}\left(\frac{t}{T}\right) - \widehat{\boldsymbol{\theta}}_n\left(\frac{t}{T}\right) \right)' \mathbf{Y}_{t-1} \right]^2 \\
& = \sum_{j=1}^p \sum_{k=1}^p \frac{1}{T} \sum_{t=1}^T \left(\theta_k\left(\frac{t-k}{T}\right) - \widehat{\theta}_{n,k}\left(\frac{t-k}{T}\right) \right) \left(\theta_j\left(\frac{t-j}{T}\right) - \widehat{\theta}_{n,j}\left(\frac{t-j}{T}\right) \right) Y_{t-j} Y_{t-k} \\
& \leq 2 \sum_{j=1}^p \sum_{k=1}^p \sqrt{\frac{1}{T} \sum_{t=1}^T \left(\theta_k\left(\frac{t-k}{T}\right) - \widehat{\theta}_{n,k}\left(\frac{t-k}{T}\right) \right)^2 \left(\theta_j\left(\frac{t-j}{T}\right) - \widehat{\theta}_{n,j}\left(\frac{t-j}{T}\right) \right)^2} \sqrt{\frac{1}{T} \sum_{t=1}^T Y_{t-k}^2 Y_{t-j}^2}
\end{aligned}$$

and thus by another application of the Cauchy-Schwarz inequality we obtain

$$\begin{aligned}
& \mathbb{E} \frac{1}{T} \sum_{t=1}^T [(\boldsymbol{\theta}(\frac{t}{T}) - \widehat{\boldsymbol{\theta}}(\frac{t}{T}))' \mathbf{Y}_{t-1}]^2 \\
& \leq 2 \sum_{j=1}^p \sum_{k=1}^p \sqrt{\mathbb{E} \frac{1}{T} \sum_{t=1}^T (\theta_k(\frac{t-k}{T}) - \widehat{\theta}_{n,k}(\frac{t-k}{T}))^2 (\theta_j(\frac{t-j}{T}) - \widehat{\theta}_{n,j}(\frac{t-j}{T}))^2} \sqrt{\mathbb{E} \frac{1}{T} \sum_{t=1}^T Y_{t-k}^2 Y_{t-j}^2} \\
& \leq 2pC \sum_{k=1}^p \sqrt{\mathbb{E} \|\theta_k - \widehat{\theta}_{n,k}\|_n} \sqrt{\sup_{1 \leq s, t \leq T} \mathbb{E}(Y_t^2 Y_s^2)}.
\end{aligned}$$

Here we again used the fact that by assumption $\theta_k(\cdot) - \widehat{\theta}_k(\cdot)$ is bounded. Also the quantity $\sup_{1 \leq s, t \leq T} \mathbb{E}(Y_t^2 Y_s^2)$ is bounded (cf. proof of $\text{Var}(\frac{1}{n} \sum_{t=1}^n Y_t^2) = o(1)$ given in Chandler and Polonik, 2010). Since $\|\theta_k - \widehat{\theta}_{n,k}\|_n = o_P(1)$ for all $1 \leq k \leq p$ (and since each of these quantities also is bounded), the assertion follows.

5.2.5 PROOF OF LEMMA 2. Consider $\lambda \geq 0$ and let $\widetilde{\lambda}$ be closest to λ among the levels $\mathcal{L} = \{\mu \geq 0 : |T(\mu)| > 0\}$. If $\lambda \in \mathcal{L}$, then obviously $T(\lambda) \subset [c_\lambda(\delta), d_\lambda(\delta)]$. However, λ does not need to be in \mathcal{L} for $[c_\lambda(\delta), d_\lambda(\delta)]$ to contain a flat set. We only need λ to be close enough to \mathcal{L} in the following sense. Let $T(\lambda) = \{t(\lambda)\}$ (since $\lambda \notin \mathcal{L}$ the set $T(\lambda)$ consists only of one point). If $\lambda < \widetilde{\lambda}$ and we have $\int_{t(\lambda)}^{t_1(\widetilde{\lambda})} (\sigma^{*2}(u) - \lambda) du \leq \delta$, then we have $T(\widetilde{\lambda}) \subset [c_\lambda(\delta), d_\lambda(\delta)]$. To see this, we only have to observe that $T(\widetilde{\lambda}) \subset [t(\lambda), t_1(\widetilde{\lambda})]$ and use the definition of $\Delta_m(\delta; \lambda)$. For $\lambda > \widetilde{\lambda}$ a similar condition, where we integrate from $t_0(\widetilde{\lambda})$ to $t(\lambda)$, is sufficient. These conditions hold for λ close enough to $\widetilde{\lambda}$, because $t(\lambda) \rightarrow t_0(\widetilde{\lambda})$ as $\lambda \rightarrow \widetilde{\lambda}$ and $\widetilde{\sigma}^2(u) = \widetilde{\lambda}$ on $T(\widetilde{\lambda}) = [t_0(\widetilde{\lambda}), t_1(\widetilde{\lambda})]$ by definition of $T(\widetilde{\lambda})$. For λ not close enough to \mathcal{L} (in the sense just described), the length of $[c_\lambda(\delta), d_\lambda(\delta)]$ tends to zero as δ tends to zero. The assertion follows.

5.3 CONSISTENCY OF MODE ESTIMATOR DISCUSSED IN SECTION 2.3.1.

Suppose that $\widehat{\sigma}_2(u)$ is such that

$$\sup_{u \in [h, 1-h]} |\widehat{\sigma}_2(u) - \sigma^2(u)| = o_P(1).$$

(See Dahlhaus and Polonik (2009) for an estimator and conditions under which this holds.) Let $m = \arg \max_{u \in [0,1]} \sigma^2(u)$ be the unique maximizer of $\sigma^2(u)$ with $m \in (0, 1)$, and define $\widehat{m} = \arg \max_{u \in [h, 1-h]} \widehat{\sigma}_2(u)$. Since $\sigma^2(u)$ is continuous (in a neighborhood of m) there exists a $\delta > 0$ such that $m \in [\delta, 1 - \delta] \subset [h, 1 - h]$ for n sufficiently large. For such n we have

$$|\widehat{\sigma}^2(\widehat{m}) - \sigma^2(m)| = \left| \sup_{u \in [h, 1-h]} \widehat{\sigma}^2(u) - \sup_{u \in [h, 1-h]} \sigma^2(u) \right| \leq \sup_{u \in [h, 1-h]} |\widehat{\sigma}^2(u) - \sigma^2(u)| = o_P(1).$$

We obtain that

$$\begin{aligned} |\sigma^2(\widehat{m}) - \sigma^2(m)| &\leq |\widehat{\sigma}^2(\widehat{m}) - \sigma^2(m)| + |\widehat{\sigma}^2(\widehat{m}) - \sigma^2(\widehat{m})| \\ &\leq \underset{n \text{ sufficiently large}}{\leq} |\widehat{\sigma}^2(\widehat{m}) - \sigma^2(m)| + \sup_{u \in [h, 1-h]} |\widehat{\sigma}^2(u) - \sigma^2(u)| = o_P(1). \end{aligned}$$

Using the uniqueness of the mode together with the continuity of $\sigma^2(\cdot)$ at m it is straightforward to conclude that this implies $|\widehat{m} - m| = o_P(1)$ as $n \rightarrow \infty$.

5.4 CONSISTENCY OF ANTIMODE COUNTING METHODOLOGY DISCUSSED IN SECTION 3.

Recall that the sets $[a_{\delta, \beta}^*, b_{\delta, \beta}^*]$ correspond to sufficiently pronounced antimodes of $\sigma^2(\cdot)$, i.e. the sets satisfy $b_{\delta, \beta}^* - a_{\delta, \beta}^* > \eta$ for some pre-specified $\eta > 0$. First we indicate that under appropriate conditions all the set estimates of the sets $[a_{\delta, \beta}^*, b_{\delta, \beta}^*]$ used in our algorithm are consistent. To see this we need to generalize Lemma 1 to include the case of the GCM/LCM being applied on an estimated set $[\widehat{a}, \widehat{b}]$ (rather than $[0, 1]$). To this end consider the function $\Sigma_{a,b}(u) = \int_a^u \sigma^2(v) dv$, $u \in [a, b]$ on a set $[a, b]$ and denote by $\Sigma_{m;a,b}^*$ its GCM/LCM with m being the mode of $\sigma^2(u)$ on $[a, b]$. Replace assumption **(A3)** by the stronger assumption

(A3) $(\delta, m, c, d) \rightarrow \sup_{\lambda \geq 0} \Delta_{m;c,d}(\delta; \lambda)$ is continuous in $(0, \delta_0] \times (0, 1] \times \{(c, d) : d - c \geq 0\}$ with respect to sup-norm.

With this stronger assumption the lines of arguments of the proof of Lemma 1 can be followed (and appropriately modified) to result in consistency of each of the (finitely many) sets $[a_{\delta, \beta}^*, b_{\delta, \beta}^*]$ corresponding to sufficiently pronounced antimodes. The fact that this estimation is performed iteratively does not have an impact asymptotically, since there will be only finitely many iterations.

Our algorithm also involves the estimation of maxima and minima of $\sigma^2(u)$, and consistency of these estimates is needed for consistently estimating the critical sets to run our tests for monotonicity. The fact that all these estimates of local extrema are consistent follows from uniform consistency of $\widehat{\sigma}^2(u)$ as an estimator of $\sigma^2(u)$. This is discussed in section 5.3. (One such uniformly consistent estimator is the one from Dahlhaus and Polonik (2009).)

REFERENCES

- ABRAHAM, C., BIAU, G., and CADRE, B. (2003) Simple estimation of the mode of a multivariate density. *Canad. J. Statist.* **3**, 23-34.
- BELLO, J.P., DAUDET, L., ABDALLAH, S., DUXBURY, C., DAVIES, M., and SANDLER, M. B. (2005). A tutorial on onset detection in music signals, *IEEE Trans. Speech Audio Process.*, **13**, 1035-1047.
- BOLLERSLEV, T. (1986). Generalized autoregressive conditional heteroskedasticity, *J. Econometrics* **31**, 307-327.
- BOWMAN A. W., JONES, M. C. and GIJBELS, I. (1998). Testing monotonicity of regression. *J. Comput. Graph. Statist.* **7**, 489-500.
- BURMAN, P. and POLONIK, W. (2009). Mode hunting in multi dimensions: data analytic tools with measures of significance. *J. Multivariate Anal.* **100**, 1198-1218.
- CHANDLER, G. and POLONIK, W. (2006). Discrimination of locally stationary time series based on the excess mass functional. *J. Amer. Statist. Assoc.* **101**, 240-253.
- CHANDLER, G. and POLONIK, W. (2012). Weighted sums and residual empirical processes for time-varying processes. *Submitted*.
- DAHLHAUS, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **25**, 1-37.
- DAHLHAUS, R., and GIRAITIS, L. (1998). On the optimal segment length for parameter estimates for locally stationary time series. *J. Time Ser. Anal.*, **19**, 629-655.
- DAHLHAUS, R., NEUMANN, M. and V. SACHS, R. (1999). Nonlinear wavelet estimation of time-varying autoregressive processes. *Bernoulli* **5**, 873-906.
- DAHLHAUS, R. and POLONIK, W. (2009). Empirical spectral processes for locally stationary time series *Bernoulli* **15**, 1 - 39.
- DAVIES, P.L. and KOVAC, A. (2001). Local Extremes, Runs, Strings and Multiresolution *Ann Statist.* **29**, 1-65.

- DAVIES, P.L. and KOVAC, A. (2010). `ftnonpar`: features and strings for nonparametric regression. R package version 0.1-84. <http://CRAN.R-project.org/package=ftnonpar>
- DREES, H. and STĂRICĂ, C. (2002). A simple non-stationary model for stock returns. *Preprint*.
- FRYZLEWICZ, P., SAPATINAS, T., and SUBBA RAO, S. (2006). A Haar-Fisz technique for locally stationary volatility estimation. *Biometrika* **93**, 687-704.
- ENGLE, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of variance of United Kingdom inflation", *Econometrica* **50**, 987-1008.
- EOM, K.B. (1999). Analysis of acoustic signatures from moving vehicles using time-varying autoregressive models, *Multidimensional Systems and Signal Processing* **10**, 357-378.
- GIRAULT, J-M., OSSANT, F. OUAHABI, A., KOUAMÉ, D. and PATAT, F. (1998). Time-varying autoregressive spectral estimation for ultrasound attenuation in tissue characterization. *IEEE Trans. Ultrasonic, Ferroelectric, and Frequency Control* **45**, 650 - 659.
- GHOSAL, S., SEN, A. and V.D. VAART, A. (2000). Testing monotonicity of regression, *Ann. Statist.* **28**, 1054-1082.
- GRENIER, Y. (1983). Time-dependent ARMA modeling of nonstationary Signals, *IEEE Trans. on Acoustics, Speech, and Signal Processing* **31** , 899911.
- HALL, M.G., OPPENHEIM, A.V. and WILLSKY, A.S. (1983). Time varying parametric modeling of speech, *Signal Processing* **5**, 267285.
- HALL, P. and HECKMAN, N. (2000). Testing for monotonicity of a regression mean without selecting a bandwidth, *Ann. Statist.* **28**, 20-39.
- HARTIGAN, J.A., and HARTIGAN, P.M. (1985). The dip test of unimodality. *Ann. Statist.* **13**, 70-84
- MINNOTTE, M.C. and SCOTT, D.W. (1993). The mode tree: A tool for visualization of nonparametric density features. *J Comput. Graph. Stat.* **2**, 51-68.
- MÜLLER, D.W., and SAWITZKI, G. (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86**, 738-746.

- ORBE, S., FERREIRA, E., and RODRIGUES-POO, J. (2005). Nonparametric estimation of time varying parameters under shape restrictions, *J. Econometrics* **126**, 53 - 77.
- RAJAN, J.J. and RAYNER, P.J. (1996). Generalized feature extraction for time-varying autoregressive models, *IEEE Trans. on Signal Processing* **44**, 24982507.
- ROBERTSON, T., WRIGHT, F., and DYKSTRA, R. (1988). *Order restricted statistical inference*. John Wiley & Sons, New York.
- SHUMWAY, R.H., BONNER, J.L., and REITER, D.T. (2004). Nonparametric deconvolution of seismic depth phases. D.R. Brillinger, E. A. Robinson and F. Schoenberg ed., *Time Series Anal. Applications to Geophysical Systems*, 1-10, IMA Volume 139, Springer, New York.
- SILVERMAN, B. W. (1981). Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. Ser. B* **43**, 9799.
- SUBBA RAO, T. (1970). The fitting of nonstationary time-series models with time-dependent parameters, *J. Royal Stat. Soc. - Ser. B* **32**, 312322.
- WOOD, J.D., O'CONNELL-RODWELL, C.E., and KLEMPERER, S.L. (2005). Using seismic sensors to detect elephants and other large mammals: a potential census technique. *J. Appl. Ecol.* **42**, 587-594.