Discrimination of Locally Stationary Time Series Based on the Excess Mass Functional

Gabriel CHANDLER and Wolfgang POLONIK

Discrimination of time series is an important practical problem with applications in various scientific fields. We propose and study a novel approach to this problem. Our approach is applicable to cases where time series in different categories have a different "shape." Although based on the idea of feature extraction, our method is not distance-based, and as such does not require aligning the time series. Instead, features are measured for each time series, and discrimination is based on these individual measures. An AR process with a time-varying variance is used as an underlying model. Our method then uses shape measures or, better, measures of concentration of the variance function, as a criterion for discrimination. It is this concentration aspect or shape aspect that makes the approach intuitively appealing. We provide some mathematical justification for our proposed methodology, as well as a simulation study and an application to the problem of discriminating earthquakes and explosions.

KEY WORDS: Modulated AR process; Nonparametric estimation; Pool-adjacent-violators algorithm; Shape restrictions.

1. INTRODUCTION

There exists a large amount of literature dealing with the problem of discriminating nonstationary time series. One class of proposed discrimination methods finds certain features in time series, such as structural breaks of certain kinds, and then bases the discrimination on the presence or absence of those features. Usually, however, time series under consideration must be aligned in a certain way (e.g., using time warping) for the procedures to work. Moreover, the resulting methodologies, although having an intuitive appeal, also often have an ad hoc flavor because the mathematical underpinning is missing. This kind of work can be found in the area of data mining (e.g., time series data mining archive at University of California, Riverside). The statistical literature includes several approaches for discriminating nonstationary time series that are based on nonparametric spectral density estimation. These methods are distance-based, and hence also require alignment of the time series.

The novel methodology proposed and studied in this article is also based on feature extraction. However, in contrast to many other methods, it is not distance-based, and hence no alignment of time series is necessary. Moreover, the method is computationally feasible, and we provide a more rigorous mathematical underpinning of our procedure. Our approach is semiparametric in nature by using a time-varying AR model. More specifically, given a training set of time series, such as recordings of a special event by a seismograph at a given specific geographic location, say, we construct a discrimination rule that automatically assigns a new time series (the event) into one of k groups, π_i, $i = 1, \ldots, k$. In a time-varying AR model, X_t has mean ν_t, and the centered time series, $X_t^c = X_t - \nu_t$, satisfies the equations $X_t^c = \sum_{k=1}^{p} \psi_k(t) X_{t-k}^c + \epsilon_t \tau(t)$, where the errors ϵ_t are assumed to be iid with mean 0. Here, however, we consider a rescaled version of this model. As introduced by Dahlhaus (1997), we rescale time to the unit interval; that is, instead of time running from 1 to T, it now runs from $1/T$ to 1. In this rescaled time, we denote by the mean function by $\mu(u)$, the AR coefficient functions by $\phi_k(u)$, $k = 1, \ldots, p$, and the variance function by $\sigma^2(u)$, $u \in [0, 1]$. With $X_{t,T}^c = X_t - \mu(t/T)$, our model is

$$X_{t,T}^c = \sum_{k=1}^{p} \phi_k(t/T) X_{t-k,T}^c + \epsilon_t \sigma(t/T). \tag{1}$$

Our procedures are based on $X_{1,T}, \ldots, X_{T,T}$. The rescaling will enable us to derive asymptotic results for our procedures. In this article we consider the class of problems in which all of the discrimination information is contained in $\sigma^2(\cdot)$. Thus $\mu(\cdot)$ and $\phi_k(\cdot)$, $k = 1, \ldots, p$, are treated as nuisance parameters. The order p assumed to be known. Our method for discriminating between classes of time series is then based on characteristics of the variance function $\sigma^2(\cdot)$ only. The motivation for this approach comes from the problem of discriminating earthquakes and mining explosions. We would like to point out, however, that the idea underlying our approach is more general. Other parameter functions can, of course, be included in the discrimination procedure. What is needed is some information on the shape of these functions to construct an appropriate criterion. Our theoretical results allow estimation of the nuisance parameters as a function, and open the door to derive joint distributions of our criterion with additional criteria based on other parameter functions.

In our motivating example of discriminating earthquakes and mining explosions, we actually assume that the AR parameters are constant over time, that is, $\phi_k(u) = \phi_k$, $k = 1, \ldots, p$, where the ϕ_k are real constants. We also assume that the mean is constant at 0, a standard assumption in this context. Intuitively, the purpose of the (now) stationary AR part in our model is to remove the underlying "noise," leaving us with the "signal," which is represented by the variance function. Seismic events of the type under consideration actually consist of two different signals, the p-wave and the s-wave (see Fig. 1). It is apparent from Figure 1 that the underlying variance function of both the p-wave and the s-wave has a unimodal shape; this applies to earthquakes as well as to explosions. The concentration

Gabriel Chandler is Assistant Professor, Department of Mathematics, Connecticut College, New London, CT 06320 (E-mail: *gabriel.chandler@conncoll.edu*). Wolfgang Polonik is Associate Professor, Department of Statistics, University of California, Davis, CA 95616 (E-mail: *polonik@wald.ucdavis.edu*). Most of this work was completed while Chandler was a graduate student at University of California, Davis. This work has been supported by National Science Foundation grants 0103606 and 0406431. The authors would like to thank three referees and the associate editor for constructive criticisms that led to a significant improvement of the original manuscript. The authors also thank Robert Shumway for several valuable discussions, and for his willingness to share computer code and datasets.

Journal of the American Statistical Association
March 2006, Vol. 101, No. 473, Theory and Methods
DOI 10.1198/016214505000000899

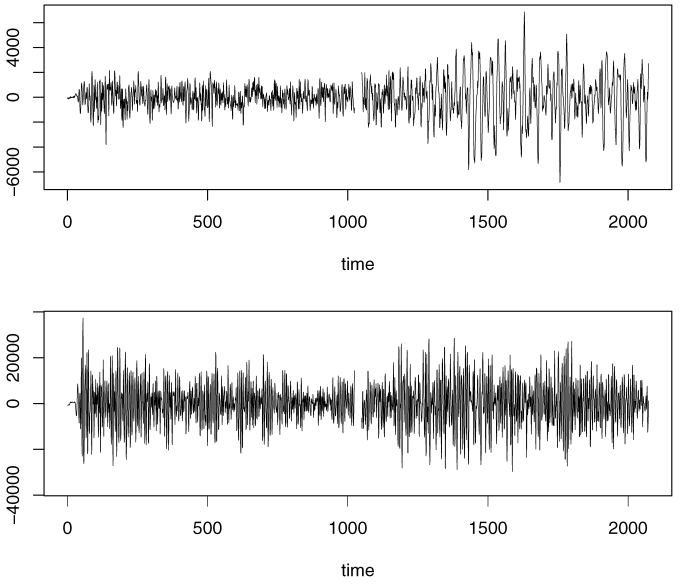

Figure 1. A Typical Earthquake and Explosion. The two parts in each series (indicated by a small gap) are called the p-wave (left) and the s-wave (right).

of this (unimodal) variance function then serves as a characteristic used for discrimination of the earthquake/explosion data. Using the concentration of the variance function for discrimination is motivated by the known fact that decay of volatility is different for earthquakes and explosions. In other words, the concentration of the variance function is the essential feature that allows us to discriminate between different classes. Different measures of concentration on which to base the discrimination will be proposed, all relating to the excess mass functional introduced independently by Müller and Sawitzki (1987) and Hartigan (1987).

Other potential applications of our method include the discrimination of types of animals based on geophone recordings of their footfalls. In one particular study (J. Wood, personal communication, 2004), the focus was on discriminating elephant footfalls from those of other animals. Yet another possible application is the analysis of electroencephalography recordings of seizures. In all of these applications, the signals of interest show different shapes compared with either the background or other signals not of main interest. This is the situation where the basic idea of our method applies.

The discrimination of earthquakes and explosions used as a guiding example is an interesting problem. It can be used to monitor nuclear proliferation, where we would want to discriminate between earthquakes, mining explosions, and small nuclear testing. This problem has been approached in several different ways. Early procedures dealt with considering relative amplitudes of the two main phases present in each seismographic reading or considering ratios of power components in particular frequency bands of the two phases (e.g., Bennett and Murphy 1986). A more recent approach was to consider the discrepancy between the spectral structure of multivariate processes (Kakizawa, Shumway, and Taniguchi 1998) while treating such processes as stationary and later considering dynamic or time-varying spectral densities (Sakiyama and Taniguchi 2001). Kakizawa et al. considered different discrepancy measures that formed quasi-distances, allowing for classification. Sakiyama and Taniguchi essentially considered a ratio of Gaussian log-likelihoods to perform discriminant analysis. Yet another approach based on the so-called SLEX model was considered by Huang, Ombao, and Stoffer (2004). There the time-varying spectral density is estimated using a SLEX basis, and again a certain generalized likelihood ratio statistic is used for classification. Although these approaches all deal with the spectrum itself or an amplitude that is proportional to the integrated spectrum, we propose a technique that bases discrimination on feature extraction in the time domain.

Our model (1) is a special case of a nonstationary time series whose distribution varies (smoothly) in time. Under appropriate assumptions, the model is a special case of a locally stationary time series as defined by Dahlhaus (1997). This definition was extended by Dahlhaus and Polonik (2004), who required only bounded variation of all the parameter functions. In particular, this includes the possibility of jumps in the variance function, our target. Note also that although this model has a parametric form, the variance is a function of rescaled time, and we use a nonparametric method to estimate this variance function. In fact, as motivated earlier, we estimate the variance function under the assumption of unimodality.

We would like to point out, however, that the purpose of this article is not only to study the specific method considered here, but also to promote the way of thinking underlying our approach.

The article is organized as follows. Section 2, assuming that we have access to an estimate of the variance function, introduces two discrimination measures based on the excess mass functional of the variance function. Section 3 deals with estimation of both the finite- and infinite-dimensional parameters. Section 4 presents theorems dealing with the asymptotic behavior of the concentration measures. They serve as justification of our discrimination methods introduced earlier. Section 5 contains a numeric study consisting of earthquakes and explosions recorded in Scandinavia, as well as a comparison of our methods with some of the spectral-based methods mentioned earlier. We defer all of the proofs to Section 6.

2. THE DISCRIMINATION METHOD

2.1 Measures of Concentration

Our discrimination method is based on our model (1) and on measures of concentration of the variance function $\sigma^2(\cdot)$. As indicated earlier, this is motivated by the well-known fact that variation in earthquakes and explosions decay differently. Two different types of measures of concentration are considered. As stated earlier, both are based on the excess mass idea. The excess mass functional of a distribution F with pdf f is defined as

$$E_F(\lambda) = \sup_C \left(\int_C dF(x) - \lambda |C| \right) = \int_{-\infty}^{\infty} (f(x) - \lambda)^+ \, dx, \tag{2}$$

where $a^+ = \max(a, 0)$, the sup is extended over all (measurable) sets C, and $|C|$ denotes the Lebesgue measure of C. In

our application, the role of the density f is taken over by the (normalized) variance function.

The excess mass measures the concentration of the underlying distribution. In fact, the excess mass functional is a convex function (as a supremum of linear functions) that is linear for a uniform distribution, and a higher degree of convexity indicates a higher concentration. Hence different rates of decay can be expected to lead to different behavior of the excess mass, or of functionals thereof.

The excess mass approach has been used to, for instance, investigate the modality of a distribution and to estimate level sets of densities and regression functions (Müller and Sawitzki 1987; Hartigan 1987; Nolan 1991; Polonik 1995; Cavalier 1997; Cheng and Hall 1998a,b; Polonik and Wang 2005).

For our purposes, we propose to consider the excess mass functional of the *normalized* variance function

$$\bar{\sigma}^2(\alpha) = \frac{\sigma^2(\alpha)}{\int_0^1 \sigma^2(u)\,du}. \tag{3}$$

$\bar{\sigma}^2(\cdot)$ does not depend on the magnitude of the series. This is important, because in the case of discrimination between earthquakes and explosions, the magnitude of the signals can vary greatly. The excess mass functional of the normalized variance function is then

$$E(\lambda) = E_{\bar{\sigma}^2}(\lambda) = \int_0^1 \bigl(\bar{\sigma}^2(u) - \lambda\bigr)^+ du, \qquad \lambda \in \mathbb{R}. \tag{4}$$

This excess mass functional is used as a basis to define two different types of measures of concentration that eventually will be used for discrimination.

Integrated Excess Mass. Our first discrimination measure is based on an integral functional of the excess mass. Less concentrated functions will have an excess mass with a higher degree of convexity compared to concentrated functions. Hence we might suspect that the tail behavior of the excess mass functional contains a lot of the discriminatory power. Because the tail would have little weight if we only considered the integral of the excess mass, we include a weight to allow for a greater contribution from the tail to our measure. Our discrimination measure then is

$$IE(\beta) = \int_0^\infty \lambda^\beta E(\lambda)\,d\lambda, \qquad \beta > 0. \tag{5}$$

Larger values of β will result in a larger emphasis on the tail of the excess mass functional, which corresponds to the peak of the variance function. For applications we propose to choose β from the data. For the theory presented here, however, we consider a fixed β.

We would like to make additional comments on $IE(\beta)$. Using $IE(\beta)$ is equivalent to using nonlinear functionals of $\bar{\sigma}^2(u)$. In fact, using Fubini's theorem, it is straightforward to see that $\int_0^\infty \lambda^\beta E(\lambda)\,d\lambda = 1/[(\beta+1)(\beta+2)]\int_0^1 (\bar{\sigma}^2(u))^{\beta+2}\,du$. Hence, using $IE(0)$ for discrimination is equivalent to base discrimination on the L_2-norm of $\bar{\sigma}^2(\cdot)$.

"Quantile" of the Excess Mass Functional. The second type of discrimination measure is a "quantile" of the excess mass functional. In other words, we define

$$\lambda(q) = E^{-1}(q) = \sup\{\lambda : E(\lambda) \geq q\}, \qquad 0 < q < 1. \tag{6}$$

More concentrated variance functions have larger excess mass quantiles than less concentrated variance functions. Similar to the tuning parameter β, the parameter q will be considered fixed for the theory that we provide later in this article. However, for the applications herein we propose a method to automatically select a value for q.

2.2 Empirical Versions of the Discrimination Measures

Empirical versions, or estimates, of the two discrimination measure are constructed via a plug-in method by using an estimator $\hat{\bar{\sigma}}_s^2(\cdot)$ of $\bar{\sigma}^2(\cdot)$ which we define in Section 3. Using this estimate, we define

$$\widehat{E}(\lambda) = E_{\hat{\bar{\sigma}}_s^2}(\lambda) = \int_0^1 \bigl(\hat{\bar{\sigma}}_s^2(u) - \lambda\bigr)^+ du, \qquad \lambda \in \mathbb{R}, \tag{7}$$

and define the empirical measures of concentration $\widehat{IE}(\beta)$ and $\widehat{\lambda}(q)$ as

$$\widehat{IE}(\beta) = \int_0^\infty \lambda^\beta \widehat{E}(\lambda)\,d\lambda, \qquad \beta > 0, \tag{8}$$

and

$$\widehat{\lambda}(q) = \widehat{E}^{-1}(q) = \sup\{\lambda : \widehat{E}(\lambda) \geq q\}, \qquad 0 < q < 1. \tag{9}$$

3. ESTIMATION OF THE VARIANCE FUNCTION

In this section we describe the construction of our estimator $\widehat{\sigma}^2(\cdot)$, and we also introduce a smoothed version $\widehat{\sigma}_s^2(\cdot)$. The reason for considering a smoothed version is that smoothing helps in some special cases. This will become clear later.

Our estimator $\widehat{\sigma}^2(u)$ can be considered a minimum distance estimator that minimizes a criterion function W_T over the parameter space. Motivated by the application of our method to seismic data, we define the parameter space of our model as the class of all unimodal functions on $[0, 1]$. To introduce some notation, let $\mathcal{U}(m)$ denote the class of all positive unimodal function on $[0, 1]$ with mode at m, that is, positive functions that are increasing to the left of m and decreasing to the right. Then $\mathcal{U} = \bigcup_{m\in[0,1]} \mathcal{U}(m)$ denotes the class of all unimodal functions on $[0, 1]$.

As a criterion function, we consider the negative conditional Gaussian likelihood with estimated nuisance parameters. Let

$$W_T(\sigma^2) = \sum_{t=1}^{T} \left\{ \log \sigma^2\left(\frac{t}{T}\right) + \frac{\bigl(X_{t,T}^{\widehat{c}} - \widehat{\phi}_1(\frac{t}{T})X_{t-1,T}^{\widehat{c}} - \cdots - \widehat{\phi}_p(\frac{t}{T})X_{t-p,T}^{\widehat{c}}\bigr)^2}{\sigma^2(\frac{t}{T})} \right\},$$

where $\widehat{\boldsymbol{\phi}} = (\widehat{\phi}_1, \ldots, \widehat{\phi}_p)$ denotes an estimator of $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$, $X_{t;T}^{\widehat{c}} = X_{t,T} - \widehat{\mu}(t/T)$, and $\widehat{\mu}(\cdot)$ is an estimator of $\mu(\cdot)$. We define

$$\widehat{\sigma}^2 = \operatorname*{argmin}_{\sigma^2 \in \mathcal{U}} W_T(\sigma^2). \tag{10}$$

To speed up the calculation, the mode might be estimated in a preliminary step. Let $\widehat{m}$ denote an estimate of the unknown mode m of the true variance function. Then we define

$$\widehat{\sigma}^2_{\widehat{m}} = \operatorname*{argmin}_{\sigma^2 \in \mathcal{U}(\widehat{m})} W_T(\sigma^2). \tag{11}$$

As our final estimate of $\sigma^2(\cdot)$, we propose using "smoothed" versions of the estimates just defined.

Note that a good approximation of the negative Gaussian log-likelihood is given by the Whittle likelihood (Whittle 1962). Thus the estimation approach can be considered a maximum Whittle likelihood approach. However, whereas the Whittle likelihood is considered a function in the frequency domain, our approach considers the time domain.

Finding the estimator $\widehat{\sigma}^2$ or $\widehat{\sigma}^2_{\widehat{m}}$ means solving a constrained optimization problem.

Basic Algorithm for Finding the Minimizers in (10) *and* (11). First, we discuss the case where the mode is estimated. Note that the solution $\widehat{\sigma}^2_{\widehat{m}}(\cdot)$ is monotonically increasing to the left of the mode $\widehat{m}$ and decreasing to the right, and estimation of both parts can be done separately, using essentially the same techniques. The key observation is that the foregoing likelihood $W_T(\cdot)$ is of exactly the form needed to apply the theory of generalized isotonic regression (e.g., Robertson, Wright, and Dykstra 1988). It follows that both the decreasing and the increasing parts of $\sigma^2(\cdot)$ can be found by a generalized isotonic regression on the squared residuals based on given values for the AR parameters. For both parts, the pool-adjacent-violators algorithm can be applied. Let e_t^2 denote the squared residual, that is,

$$e_t^2 = \left(X^{\widehat{c}}_{t,T} - \widehat{\phi}_1\left(\frac{t}{T}\right) X^{\widehat{c}}_{t-1,T} - \cdots - \widehat{\phi}_p\left(\frac{t}{T}\right) X^{\widehat{c}}_{t-p,T} \right)^2, \tag{12}$$

To the left of the estimated mode $\widehat{m}$ it is the (right-continuous) slope of the greatest convex minorant to the cumulative sum diagram given by the points $(k/T, (1/T)\sum_{t=1}^{k} e_t^2)$, $k = 1, \ldots, T$, and to the left of $\widehat{m}$ it is the (left-continuous) slope of the least concave majorant to the same cumulative sum diagram (cf. Robertson et al. 1988, thm. 1.2.1).

In the event the mode is not estimated, then the algorithm just described must be performed T times, because any time point t/T, $t = 1, \ldots, T$, is a potential mode. (The estimator is known to have jumps only at t/T for some $t = 1, \ldots, T$.) Among the resulting T different estimators, the one with the overall largest value of W_T is chosen.

There is, however, a small problem with the solution of the algorithm just described. It is well known in density estimation that the maximum likelihood estimator of a unimodal density does not behave well at the mode (i.e., is not consistent at the mode, and behaves irregularly close to the mode). This is sometimes called the spiking problem. The spiking problem also applies when estimating the variance function under the unimodality restriction. In some cases, however, estimation of the maximum value of $\sigma^2(\cdot)$ is of importance for our method, because it may be needed to estimate the (asymptotic) variance of one of our discrimination measures (cf. Thm. 1). The spiking problem can be avoided by introducing some smoothing.

Smoothing to Avoid Irregular Behavior at the Mode. Intuitively, the spiking problem can be understood by observing that the isotonic regression is given by the slope of the least concave majorant to the cumulative sum diagram based on the squared residuals and hence the estimate of the mode depends on only a few observations, regardless of sample size. One large squared residual in the neighborhood of the mode will therefore cause problems.

Sun and Woodroofe (1993) proposed a penalized least squares approach to make the estimate of a *density* consistent at the mode. Although this does not readily generalize beyond the density case, another approach to stabilize the estimate is to first smooth the squared residuals, via a kernel smoother, and then perform isotonic regression on the smoothed data. This is the approach that we use. That is, we form the smoothed residuals

$$e_t^{*2} = \frac{\sum_{s=-\infty}^{\infty} K\left(\frac{s-t}{[bT]}\right) e_s^2}{\sum_{s=-\infty}^{\infty} K\left(\frac{s-t}{[bT]}\right)}, \tag{13}$$

and find

$$\widehat{\sigma}^2_s(\cdot) = \operatorname*{argmin}_{\sigma^2 \in \mathcal{U}} \sum_{t=1}^{T} \left\{ \log \sigma^2\left(\frac{t}{T}\right) + \frac{e_t^{*2}}{\sigma^2(\frac{t}{T})} \right\} \tag{14}$$

or

$$\widehat{\sigma}^2_{s,\widehat{m}}(\cdot) = \operatorname*{argmin}_{\sigma^2 \in \mathcal{U}(\widehat{m})} \sum_{t=1}^{T} \left\{ \log \sigma^2\left(\frac{t}{T}\right) + \frac{e_t^{*2}}{\sigma^2(\frac{t}{T})} \right\}. \tag{15}$$

Using these estimates, we define

$$\widehat{E}(\lambda) = \int_0^1 \left(\widehat{\sigma}^2_s(u) - \lambda\right)^+ du \tag{16}$$

and similarly $\widehat{E}_{\widehat{m}}(\lambda)$, where $\widehat{\sigma}^2_s$ is replaced by $\widehat{\sigma}^2_{s,\widehat{m}}$. These two empirical excess mass functions have corresponding empirical discrimination measures $\widehat{IE}(\beta)$ and $\widehat{\lambda}(q)$, as defined in (8) and (9), as well as the corresponding quantities $\widehat{IE}_{\widehat{m}}(\beta)$ and $\widehat{\lambda}_{\widehat{m}}(q)$, which are defined using $\widehat{E}_{\widehat{m}}(\lambda)$ instead of $\widehat{E}(\lambda)$. We study the large-sample behavior of all these quantities in the next section.

We discuss later that using e_t^{*2} instead of the regular squared residuals e_t^2 makes no difference asymptotically. We also mention that isotonizing the variance function followed by a smoothing step would also alleviate the spiking problem. Mammen (1991) showed that the resulting estimator is asymptotically first-order equivalent to the estimator we use.

4. ASYMPTOTIC NORMALITY OF THE EMPIRICAL CONCENTRATION MEASURES

Results on asymptotic normality of the empirical concentration measures formulated in this section motivate the use of our discrimination procedures, which are known to be optimal in the normal case. We need the following assumptions.

Assumption 1. (a) $\epsilon_1, \epsilon_2, \ldots$ are iid with $\mathrm{E}\epsilon_t^2 = 1$ and $\mathrm{E}(\epsilon_1^4 \log|\epsilon_1|) < \infty$.

(b) $\sup_{1 \le t \le T} \mathrm{E} X^4_{t,T} < \infty$.

(c) $\sigma^2(\cdot) \in \mathcal{U}(m)$, $\sup_{\alpha \in [0,1]} \sigma^2(\alpha) < M < \infty$, and $0 < \int_0^1 \sigma^2(\alpha)\, d\alpha$.

Further, we assume that $\sigma^2(\cdot)$ has no "flat parts," that is,

$$\sup_{\lambda>0}\left|\left\{u\in[0,1]:|\sigma^2(u)-\lambda|<\epsilon\right\}\right|\to 0 \quad \text{as } \epsilon\to 0. \tag{17}$$

Recall that for a (measurable) set C, we denote its Lebesgue measure by $|C|$. Assumption 1(b) follows from Assumption 1(a) if, for instance, we assume $X_{t,T}$ to be a "locally stationary" process in the sense discussed by Dahlhaus and Polonik (2004). Note that this definition of local stationarity does not require that the parameter functions and the variance functions be continuous; only a finite bounded variation is required. In particular, our variance function is allowed to have jumps. We would also like to point out that assumption (17) in particular implies that the excess mass functional is continuously differentiable for unimodal $\sigma^2(\cdot)$. This fact is needed in the proofs. It is important to note, however, that unimodality of $\sigma^2(\cdot)$ is not a necessary assumption for our discrimination procedure. It could be dropped for the cost of additional, more complex assumptions.

Our next assumption uses bracketing numbers of a class of functions $\mathcal{G}$. In this context, a set $[g_*, g^*] := \{h\in\mathcal{G}: g_*\le h\le g^*\}$ is called a "bracket." Given a metric ρ on $\mathcal{G}$ and $\delta>0$, the bracketing number $N(\delta,\mathcal{G},\rho)$ is the smallest number of brackets $[g_*, g^*]$ with $\rho(g_*, g^*)\le\delta$ needed to cover $\mathcal{G}$. If there is no such class of brackets, then $N(\delta,\mathcal{G},\rho)=\infty$.

Assumption 2. (a) The kernel function K is symmetric around 0, is bounded, and has support $[-1,1]$. The smoothing parameter b satisfies $bT\to\infty$ and $b\sqrt{T}\to 0$ as $T\to\infty$.

(b) The estimators $\widehat{\phi}_j$ of ϕ_j, $j=1,\dots,p$, satisfy the following conditions:

1. $\widehat{\phi}_j\in\mathcal{G}$, where $\mathcal{G}$ is a class of functions with $\|\phi\|_\infty<a$ for some $0<a<\infty$, and $\int_0^1\sqrt{\log N(\delta,\mathcal{G},\|\cdot\|_\infty)}\,d\delta<\infty$, where $\|\cdot\|_\infty$ denotes sup-norm.
2. $\|\widehat{\mu}-\mu\|_\infty=o_P(T^{-1/4})$, $\|\widehat{\phi}_j-\phi_j\|_\infty=o_P(T^{-1/4})$ for all $j=1,\dots,p$.

Examples of function classes satisfying the finite integral assumption in (b.1) include Hölder smoothness classes (with appropriate smoothness parameters), the class of monotonic functions, and of course constant functions. (For more details and more examples, as well as for a definition of the bracketing covering numbers, see van der Vaart and Wellner 1996.) In our application presented later we use a model with constant AR parameters; thus the rate of convergence of these estimators is $\sqrt{T}$. Note, however, that we do not require a $\sqrt{T}$-consistent estimator in (b.2). [For estimators satisfying (b.2), see Dahlhaus and Neumann 2001.]

When using a global estimate of the mode $\widehat{m}$, we also require the following assumption.

Assumption 3.

$$\int_0^1\left(\sigma^2(u)-\sigma^2(\widehat{m})\right)^+du=o_P(1/\sqrt{T}).$$

Assumption 3 is satisfied if, for instance $T^{1/6}(\widehat{m}-m)=o_P(1)$ and $\sigma^2(\cdot)$ behaves like a quadratic around the mode m.

Our concentration measures are smooth functionals of the empirical excess mass $\widehat{E}$, which itself is asymptotic normal. This explains their asymptotic normality. Asymptotic normality of the excess mass is also of independent interest.

Lemma 1 (Asymptotic normality of the excess mass). Under model (1) and Assumptions 1 and 2 the excess mass process $\sqrt{T}(\widehat{E}-E)(\lambda)$ converges in distribution to a mean-0 Gaussian process in $C[0,\bar{\sigma}^2(m)]$ with covariance function $c(\lambda_1,\lambda_2)$ given by

$$\begin{aligned}(\mu_4-1)\Bigg(&\Sigma(\Gamma(0))\int_{\Gamma(\lambda_1)\cap\Gamma(\lambda_2)}\bar{\sigma}^4(u)\,du\\&+\Sigma(\Gamma(\lambda_1))\Sigma(\Gamma(\lambda_2))\int_0^1\bar{\sigma}^4(u)\,du\\&-\Sigma(\Gamma(\lambda_1))\Sigma(\Gamma(0))\int_{\Gamma(\lambda_2)}\bar{\sigma}^4(u)\,du\\&-\Sigma(\Gamma(\lambda_2))\Sigma(\Gamma(0))\int_{\Gamma(\lambda_1)}\bar{\sigma}^4(u)\,du\Bigg),\end{aligned}$$

where $\mu_4=\mathrm{E}\epsilon_1^4$, $\Gamma(\lambda)=\{u\in[0,1]:\bar{\sigma}^2(u)\ge\lambda\}$ and $\Sigma(C)=\int_C\bar{\sigma}^2(u)\,du$. If in addition Assumption 3 is assumed, then the same results holds for $\sqrt{T}(\widehat{E}_{\widehat{m}}-E)(\lambda)$.

Theorem 1 (Asymptotic normality of the $\widehat{IE}(\beta)$). Under the corresponding assumptions of Lemma 1, both $\sqrt{T}(\widehat{IE}(\beta)-IE(\beta))$ and $\sqrt{T}(\widehat{IE}_{\widehat{m}}(\beta)-IE(\beta))$ are asymptotically normal with mean 0 and variance

$$\int_0^{\bar{\sigma}^2(m)}\lambda_1^\beta\int_0^{\bar{\sigma}^2(m)}\lambda_2^\beta c(\lambda_1,\lambda_2)\,d\lambda_1\,d\lambda_2,$$

where $c(\lambda_1,\lambda_2)$ is the covariance function of the excess mass process given in Lemma 1.

Theorem 2 (Asymptotic normality of the $\widehat{\lambda}(q)$). Let $q\in(0,1]$ be fixed. Under the corresponding assumptions of Lemma 1, both $\sqrt{T}(\widehat{\lambda}(q)-\lambda(q))$ and $\sqrt{T}(\widehat{\lambda}_{\widehat{m}}(q)-\lambda(q))$ are asymptotically normal with mean 0 and variance

$$\left(\frac{1}{|\Gamma(\lambda(q))|}\right)^2 c\big(\lambda(q),\lambda(q)\big), \tag{18}$$

where $c(\cdot,\cdot)$ is the asymptotic variance of the empirical excess mass given in Lemma 1.

An estimate of the variance can be found by plugging in the empirical estimates for the corresponding theoretical values.

5. THE DISCRIMINATION PROCEDURE

In this section we describe our discrimination procedure in some detail. The common setup is to assume the availability of a training set with known types or classes. This training data are used to estimate unknown quantities and to derive the actual classification rule. A new, unknown event is then assigned to a category based on the value of the classification rule. Our discrimination rule is based on the (empirical) discrimination measures described earlier.

The actual format of the rule is motivated by asymptotic normality of the empirical concentration measures. Under the assumption that the observations within each class are iid Gaussian random variables, optimal discrimination techniques are well studied. Motivated by asymptotic normality of our discrimination measures, we use such optimal normality-based discrimination rules.

A Two-Dimensional Variant. In the application study presented herein, where we apply our procedure to the discrimination of seismic times series, we follow a well-known practice (see, e.g., Kakizawa et al. 1998) by treating each seismic time series as bivariate by considering the p-wave and the s-wave individually, and also assume the two waves are independent. Figure 1 shows two examples of seismic time series (one earthquake and one explosion) split into p-waves and s-waves.

Applying our discrimination measures to the p-waves and s-waves separately, we end up with a two-dimensional measure of discrimination. We then apply a normality-based quadratic classification rule to assign future observations into one of the two classes, $\pi_1 =$ "earthquake" or $\pi_2 =$ "explosion." That is, if we let $T(\mathbf{X})$ denote one of the two discrimination measures $\widehat{IE}(\beta)$ or $\widehat{\lambda}(q)$, and we let $\mathbf{X} = (X_p, X_s)$ denote the decomposition of the seismic time series into p-wave X_p and s-wave X_s, then $\mathbf{T}(\mathbf{X}) = (T(X_p), T(X_s))$ denotes the two-dimensional discrimination measure. Then we allocate a newly observed time series, $\mathbf{X}$, to π_1 based on $\mathbf{T}(\mathbf{X})$ if

$$-\frac{1}{2}\mathbf{T}(\mathbf{X})'(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})\mathbf{T}(\mathbf{X}) + (\overline{\mathbf{T}}_1'\mathbf{S}_1^{-1} - \overline{\mathbf{T}}_2'\mathbf{S}_2^{-1})\mathbf{T}(\mathbf{X}) - k \leq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right], \quad (19)$$

where

$$k = \frac{1}{2}\ln\left(\frac{|\mathbf{S}_1|}{|\mathbf{S}_2|}\right) + \frac{1}{2}(\overline{\mathbf{T}}_1'\mathbf{S}_1^{-1}\overline{\mathbf{T}}_1 - \overline{\mathbf{T}}_2'\mathbf{S}_2^{-1}\overline{\mathbf{T}}_2), \quad (20)$$

$\overline{\mathbf{T}}_i$ denotes the sample average classification measure for group i, $\mathbf{S}_i$ is the estimated covariance matrix of the empirical concentration measures for group i, $|\mathbf{S}_i|$ is the determinant of $\mathbf{S}_i$, $c(i|j)$ is the cost associated with allocating an observation from group j into group i, and p_i is the proportion of observations from group i in the population (see, e.g., Johnson and Wichern 1998). Because of the assumed independence of the p-wave and s-wave $\mathbf{S}_i$, $i = 1, 2$, will be diagonal matrices. Next, we study this discrimination procedure numerically.

5.1 Classifying Seismic Time Series

The dataset that we use here was first published by Kakizawa et al. (1998), who also provided a description of the events. (For a detailed discussion of the problem, see also Shumway and Stoffer 2000.) The dataset consists of eight known earthquakes and 8 mining explosions, all measured by stations in Scandinavia at regional distances. It also contains an event of unknown origin. One could envision a scenario where this procedure would be used to monitor nuclear treaties, where classifying a nuclear test as an earthquake would carry with it a different cost than the opposite type of error. Nevertheless, for what is done in the following, we set the cost associated with allocating a series to the wrong class as equal for both cases. Likewise, we assume that the proportions of each type are equal, that is, $p_1 = p_2$ in (19). For this example, we use a second-order autoregressive process, because it seems to be a good fit for the data available.

In this application we use the special case of our general methodology described earlier, where we assume mean 0, constant AR parameters, and $p = 2$. Further evidence that this assumption is justified for classification purposes is provided by the fact that our classification results presented herein did not change when we applied the methodology with the mean function estimated (nonparametrically via a kernel estimator).

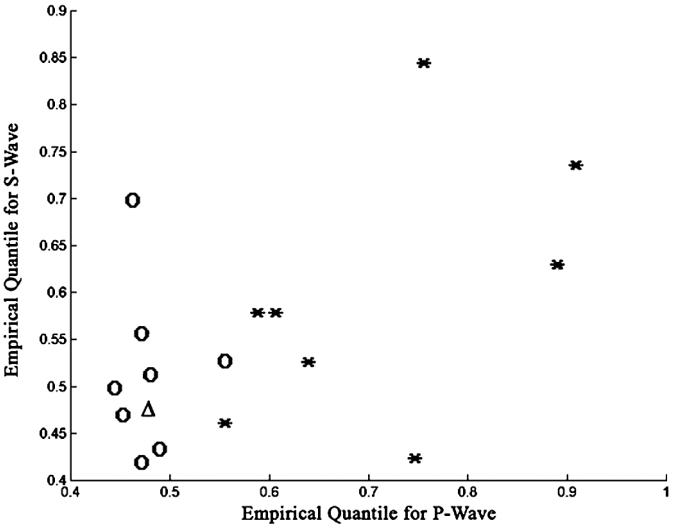

Figure 2. Discrimination Scatterplot for the Quantile Method (o, earthquake; ∗, explosion; △, unknown event).

We use the excess mass quantile as our discrimination measure, and first discuss the selection of the parameter q determining the value $\lambda(q) = E^{-1}(q)$ on which to base our discrimination. Clearly, selection of this quantity is crucial, because $q = 1$ would provide no discriminatory power, since $\lambda(1) = 0$ for all series. Because we want to discriminate between series belonging to π_1 and π_2, we search for a quantile q that makes observations within π_i as homogenous as possible, while simultaneously making series in different categories as different as possible. Therefore, we search for the value q that maximizes the ratio of between sums to squares to within sums of squares. Assuming each wave to be independent, we choose a different q independently for each type of wave (see Fig. 3). As may be apparent from Figure 2, much of the discriminatory power lies in the p-waves. The data select a value of $q = .01$ for the p-wave. The scatterplot of the discrimination measures is shown in Figure 2. It is of interest to note that the unknown event is clearly classified as an earthquake. Whereas the discrimination measure results in complete separation of the two classes, the discrimination rule results in two misclassifications as a result of the identical distribution assumption having been apparently violated. Inherent in using the quadratic discrimination rule is the assumption that the discrimination measures for each explosion or earthquake are identically distributed with respect to others in the same class. It is not clear from Figure 2 whether this assumption is justified; nevertheless, we present the example.

In the foregoing analysis we estimate AR parameters and the variance function simultaneously by minimizing the Whittle likelihood. More precisely, we define our parameter space as

$$\Theta \times \mathcal{U} = \left\{\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p) \in \Phi, \sigma^2(\cdot) \text{ unimodal on } [0, 1]\right\},$$

where Φ denotes the set

$$\Phi = \left\{\boldsymbol{\phi} = (\phi_1, \phi_2) \in \mathbb{R}^2 : \sum_{i=1}^{2} \phi_i z^i \neq 0 \; \forall z \in \mathbb{C}, 0 < |z| \leq 1\right\}.$$

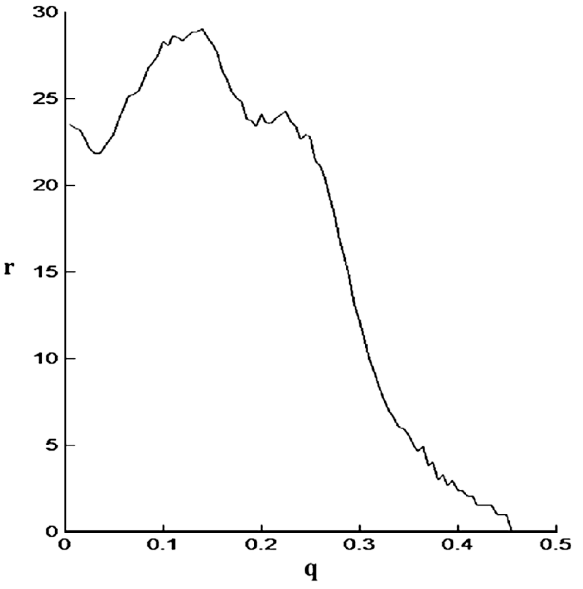

Figure 3. The Ratio of Between Sums of Squares versus Within Sums of Squares as a Function of q for Simulated Data Using the Quantile Method.

Note that if in addition $\sigma^2(\cdot)$ is bounded (which we assume; see Assumption 1), then this choice of Φ corresponds to our model being locally stationary in the sense of Dahlhaus and Polonik (2004). This in turn implies that the estimator $\widehat{\boldsymbol{\phi}}_{\mathcal{L}}$, defined in the following, is $\sqrt{T}$-consistent, and hence our assumptions apply to this case.

Let

$$W_T(\boldsymbol{\phi}, \sigma^2) = \sum_{t=1}^{T} \left\{ \log \sigma^2\left(\frac{t}{T}\right) + \frac{(x_{t,T} - \phi_1 x_{t-1,T} - \cdots - \phi_p x_{t-p,T})^2}{\sigma^2(\frac{t}{T})} \right\},$$

and define

$$(\widehat{\phi}_{\mathcal{L},\widehat{m}}, \widehat{\sigma}^2_{\mathcal{L},\widehat{m}}) = \underset{(\boldsymbol{\phi},\sigma^2)\in\Theta\times\mathcal{U}(\widehat{m})}{\operatorname{argmin}} W_T(\boldsymbol{\phi}, \sigma^2). \tag{21}$$

Note that this can also be viewed as a profile likelihood method. This is, we have $\widehat{\sigma}^2_{\mathcal{L},\widehat{m}} = \operatorname{argmin}_{\sigma^2\in\mathcal{U}(\widehat{m})} W_T(\sigma^2, \widehat{\boldsymbol{\phi}}(\sigma^2))$, where $W_T(\sigma^2, \widehat{\boldsymbol{\phi}}(\sigma^2)) = \operatorname{argmin}_{\boldsymbol{\phi}\in\Theta} W_T(\boldsymbol{\phi}, \sigma^2)$. Thus the foregoing fits into the framework described in Section 3. The estimator $\widehat{\boldsymbol{\phi}}$ has the desired convergence rate. In fact, as shown by Dahlhaus and Polonik (2005), it is of $1/\sqrt{T}$-rate.

The final estimate of $\sigma^2(\cdot)$ that we use in the numerical work is the "smoothed" version of $\widehat{\sigma}^2_{\mathcal{L},\widehat{m}}$ as described earlier. Here the estimate $\widehat{m}$ is the global mode of a kernel-smoothed squared observations based on a Gaussian kernel with bandwidth of 50 time points. This bandwidth was chosen based on a visual inspection of the data. This (rough) estimate was chosen by computational convenience, and other estimates that we tried worked similarly well.

The Algorithm. To find the minimizers in (21), we use an iterative procedure described by Dahlhaus and Polonik (2005). Note that finding the AR parameters for a given $\sigma^2(t)$ is just a weighted least squares problem. And given the values for the AR parameters we use the algorithm for estimating the variance function described earlier. These two steps are iterated to find the global minimizers.

5.2 Simulations

In this section we present simulation results comparing our method with the spectral, distance-based method of Sakiyama and Taniguchi (2001). This spectral-based method essentially assigns an observation to the category with spectra closest to an estimate of the spectral density of the observed processes with respect to an approximation of the Gaussian likelihood. The simulated time series detailed herein were generated so that alignment of the observations in time is not necessary, thus aiding the distance-based procedure.

We emulate the original dataset by simulating observations from each of two categories [meaning two different AR(2) models]. We then try to reclassify each observation using the remainder as a training set. The goal of this simulation study is to show how the two proposed measures compare with the spectral-based method under the proposed model. We mimic the data of the numerical study in that we generate eight observations from each of two categories, and show that when the assumptions regarding identically distributed random variables is satisfied, the quadratic discrimination rule performs quite well, in many cases outperforming the spectral-based classification rule.

For an observation from group i, we simulate from a second-order AR process of the form:

$$x_{t,T} = 1.58x_{t-1,T} - .64x_{t-2,T} + \epsilon_t \sigma_i(t/T). \tag{22}$$

The values for the AR parameter are similar to the estimated parameters for the data in the numeric example and can be shown to satisfy the requirement for causality.

The two different variance functions that we use for the two categories in the following tables are of the form

$$\begin{aligned} \sigma_1(u) &= 300\big(u^a \mathbb{1}(u < .5) + (1-u)^b \mathbb{1}(u > .5)\big), \\ \sigma_2(u) &= 300\big(u^c \mathbb{1}(u < .5) + (1-u)^d \mathbb{1}(u > .5)\big), \end{aligned} \tag{23}$$

where a and b are adjusted to illustrate how the discrimination measures behave at different levels of similarity between the two categories. Let X_i denote the number of misclassifications in one run of the simulation; thus $X_i \in (0, 1, \ldots, 16)$. In Tables 1 and 2, $\bar{X}$ denotes the misclassification rate and s_X estimates the standard deviation of the number of misclassifications. "Spec" denotes the spectral-based methods of Sakiyama

Table 1. Misclassification Rate, $\bar{X}$, and Estimated Standard Deviation, s_X, of the Number of Misclassifications Based on 100 Simulation Runs Under Model (23)

	$\widehat{IE}$		$\widehat{\lambda}$		Spec	
a	$\bar{X}$	s_X	$\bar{X}$	s_X	$\bar{X}$	s_X
2.6	.07	.26	.01	.11	.59	.77
2.5	.2	.42	.03	.17	.7	.67
2.4	.49	.70	.06	.24	.84	.84
2.25	2.52	1.61	.4	.62	1.45	1.14
2.2	3.2	1.22	.86	.98	1.3	1.25

Table 2. Misclassification Rate, $\bar{X}$, and Estimated Standard Deviation, s_X, of the Number of Misclassifications Based on 100 Simulation Runs Under Model (24)

		$\widehat{IE}$		$\hat{\lambda}$		Spec	
a	b	$\bar{X}$	s_X	$\bar{X}$	s_X	$\bar{X}$	s_X
3.6	3.0	.07	.25	.08	.31	1.44	1.12
3.5	2.9	.29	.55	.14	.34	1.72	1.16
3.4	2.75	.62	.84	.38	.60	1.84	1.24
3.3	2.6	1.50	1.23	.91	1.00	1.94	1.27

and Taniguchi (2001). The tables are based on 100 simulation runs. In all cases examined, the quantile method outperforms the spectral-based methods, and in several cases the integrated excess mass approach is better as well. Table 1 examines the case in which the varaince function is continuous, setting $c = d = 2$ and $a = b$. Table 2 considers a variance function that contains jumps, thus not satisfying local stationarity in the sense of Dahlhaus (1997), but still covered by the definition given by Dahlhaus and Polonik (2004). This is accomplished by choosing $c = 2$ and $d = 3$, so that the jump occurs at $u = .5$ and is as large as half of the maximum of the variance function:

$$\begin{aligned}\sigma_1(u) &= 300\big(u^3\mathbb{1}(u < .5) + (1-u)^2\mathbb{1}(u > .5)\big),\\ \sigma_2(u) &= 300\big(u^a\mathbb{1}(u < .5) + (1-u)^b\mathbb{1}(u > .5)\big).\end{aligned} \tag{24}$$

It is interesting to note that even when both methods make misclassifications for a particular dataset, the misclassified observations often are not the same across methods. It seems that discrimination is based on different information in the data. Thus it seems possible to define a discrimination rule, that is a combination of spectral-based and time-based methods that would outperform each method individually.

6. PROOFS

To ease notation, we write X_t instead of $X_{t,T}$ throughout this section. For the proofs, we assume that $X_{-[bT]+1} = X_{-[bT]+2} = \cdots$, $X_0 = 0 = X_{T+1} = X_{T+2} = \cdots = X_{T+[bT]}$. Let $\widehat{\Sigma}(\alpha)$ and $\widehat{\Sigma}^*(\alpha)$ denote the partial sum processes for the ordinary and the smoothed residuals, that is,

$$\widehat{\Sigma}(\alpha) = \frac{1}{T}\sum_{t=1}^{[\alpha T]} e_t^2 \qquad \text{and} \qquad \widehat{\Sigma}^*(\alpha) = \frac{1}{T}\sum_{t=1}^{[\alpha T]} e_t^{*2}. \tag{25}$$

Recall that $e_t^{*2} = \sum_{s=t-[bT]}^{t+[bT]} (K(\frac{t-s}{bT})/A)e_s^2$, where $A = \sum_{t=s-[bT]}^{s+[bT]} K(\frac{s-t}{[bT]})$. Hence, by rearranging sums, we can write

$$\widehat{\Sigma}^*(\alpha) = \frac{1}{T}\sum_{t=1}^{[\alpha T]+[bT]} \omega_t(\alpha)e_t^2, \tag{26}$$

where the nonrandom weights $\omega_t(\alpha)$ are normalized sums of kernel weights. The specific form of the weights plays no role in our proofs, and hence no precise formula is provided. However, what is important is that most of the weights equal 1—namely, if $\alpha > 2b$, then $\omega_t(\alpha) = 1$ for all $[bT] \le t \le [\alpha T] - [bT]$; otherwise, $0 < \omega_t(\alpha) < 1$.

First, we prove a lemma that allows us to disregard estimation of the nuisance parameters.

Lemma 2. Suppose that Assumptions 1 and 2 hold. Then

$$\sqrt{T}\sup_{0\le\alpha\le1}\left|\widehat{\Sigma}(\alpha) - \frac{1}{T}\sum_{t=1}^{[\alpha T]}\sigma^2(t/T)\epsilon_t^2\right| = o_P(1)$$

and

$$\sqrt{T}\sup_{0\le\alpha\le1}\left|\widehat{\Sigma}^*(\alpha) - \frac{1}{T}\sum_{t=1}^{[\alpha T]+[bT]}\omega_t(\alpha)\sigma^2(t/T)\epsilon_t^2\right| = o_P(1).$$

Proof. For ease of notation, we present the proof for the case where $p = 2$. The case of a general order $p \ge 1$ follows similarly. To simplify the proof, we also assume that the parameter functions ϕ_1, ϕ_2, and σ are defined on the whole real line with $\phi_1(u) = \phi_2(u) = \sigma(u) = 0$ for $u \notin [0,1]$. We first prove the assertion for the ordinary residual partial sum process $\widehat{\Sigma}(\alpha) = 1/T\sum_{t=1}^{[\alpha T]} e_t^2$.

First, we consider the effect of estimating $\mu(\cdot)$. With $\tilde{e}_t = X_t^c - \widehat{\phi}_1(t/T)X_{t-1}^c - \widehat{\phi}_2(t/T)X_{t-2}^c$ and $\Delta_T(t) = \mu(t/T) - \widehat{\mu}(t/T)$, we can write the difference $\frac{1}{T}\sum_{t=1}^{[\alpha T]} e_t^2 - \frac{1}{T}\sum_{t=1}^{[\alpha T]}\tilde{e}_t^2$ as

$$\begin{aligned}&\frac{2}{T}\sum_{t=1}^{[\alpha T]}\tilde{e}_t\bigg(\Delta_T(t) - \widehat{\phi}_1\Big(\frac{t}{T}\Big)\Delta_T(t-1)\\ &\qquad - \widehat{\phi}_2\Big(\frac{t}{T}\Big)\Delta_T(t-2)\bigg) \qquad (27)\\ &+ \frac{1}{T}\sum_{t=1}^{[\alpha T]}\bigg(\Delta_T(t) - \widehat{\phi}_1\Big(\frac{t}{T}\Big)\Delta_T(t-1)\\ &\qquad - \widehat{\phi}_2\Big(\frac{t}{T}\Big)\Delta_T(t-2)\bigg)^2. \qquad (28)\end{aligned}$$

Using the Cauchy–Schwarz inequality, (27) can be bounded from above by

$$\begin{aligned}&2\alpha\left[\frac{1}{\alpha T}\sum_{t=1}^{[\alpha T]}\tilde{e}_t^2\right]\left[\frac{1}{\alpha T}\sum_{t=1}^{[\alpha T]}\bigg(\Delta_T(t) - \widehat{\phi}_1\Big(\frac{t}{T}\Big)\Delta_T(t-1)\right.\\ &\qquad\left. - \widehat{\phi}_2\Big(\frac{t}{T}\Big)\Delta_T(t-2)\bigg)^2\right]\\ &\le 2\left[1 + \sup_t\left|\widehat{\phi}_1\Big(\frac{t}{T}\Big)\right| + \sup_t\left|\widehat{\phi}_2\Big(\frac{t}{T}\Big)\right|\right]^2\\ &\quad\times \sup_t|\Delta_T(t)|^2\frac{1}{T}\sum_{t=1}^{[\alpha T]}\tilde{e}_t^2\\ &= o_P(1/\sqrt{T})\frac{1}{T}\sum_{t=1}^{[\alpha T]}\tilde{e}_t^2. \qquad (29)\end{aligned}$$

The last equality follows from the fact that by assumption, the $\widehat{\phi}_i$ are uniformly consistent estimates of ϕ_i, the ϕ_i are bounded, and $\sup_t|\Delta_T(t)|^2 = o_P(1/\sqrt{T})$. Later we show that $\sup_{0\le\alpha\le1}\frac{1}{T}\sum_{t=1}^{[\alpha T]}\tilde{e}_t^2 \le \frac{1}{T}\sum_{t=1}^{T}\tilde{e}_t^2 = O_P(1)$. This, together with the foregoing, then implies that $\sup_{0<\alpha\le1}|\frac{1}{T}\sum_{t=1}^{[\alpha T]}e_t^2 - \frac{1}{T}\sum_{t=1}^{[\alpha T]}\tilde{e}_t^2| = o_P(1/\sqrt{T})$. The fact that (28) is also $o_P(1/\sqrt{T})$ follows similarly, but is somewhat easier. Details are omitted.

We now show that $\sup_{0<\alpha\le 1}|\frac{1}{T}\sum_{t=1}^{[\alpha T]}\tilde{e}_t^2 - \frac{1}{T}\sum_{t=1}^{[\alpha T]}\sigma^2(\frac{t}{T})\times\epsilon_t^2| = o_P(1/\sqrt{T})$. This then completes the proof, because $\frac{1}{T}\sum_{t=1}^{[\alpha T]}\sigma^2(\frac{t}{T})\epsilon_t^2 \le \frac{1}{T}\sum_{t=1}^{T}\sigma^2(\frac{t}{T})\epsilon_t^2 = O_P(1)$. We can write

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{[\alpha T]}\tilde{e}_t^2 - \frac{1}{T}\sum_{t=1}^{[\alpha T]}\sigma^2\left(\frac{t}{T}\right)\epsilon_t^2\\
&\quad= \frac{1}{T}\sum_{t=1}^{[\alpha T]}\left[-2\left(X_t^c - \phi_1\left(\frac{t}{T}\right)X_{t-1}^c - \phi_2\left(\frac{t}{T}\right)X_{t-2}^c\right)X_{t-1}^c\right]\\
&\qquad\qquad\times\left(\widehat{\phi}_1\left(\frac{t}{T}\right) - \phi_1\left(\frac{t}{T}\right)\right)\\
&\qquad+\frac{1}{T}\sum_{t=1}^{[\alpha T]}\left[-2\left(X_t^c - \phi_1\left(\frac{t}{T}\right)X_{t-1}^c - \phi_2\left(\frac{t}{T}\right)X_{t-2}^c\right)X_{t-2}^c\right]\\
&\qquad\qquad\times\left(\widehat{\phi}_2\left(\frac{t}{T}\right) - \phi_2\left(\frac{t}{T}\right)\right)\\
&\qquad+\frac{1}{T}\sum_{t=1}^{[\alpha T]}\left(\widehat{\boldsymbol{\phi}}\left(\frac{t}{T}\right) - \boldsymbol{\phi}\left(\frac{t}{T}\right)\right)'\mathbf{C}_t\left(\widehat{\phi}\left(\frac{t}{T}\right) - \phi\left(\frac{t}{T}\right)\right)\\
&\quad= -\frac{2}{T}\sum_{j=1}^{2}\sum_{t=1}^{[\alpha T]}\sigma\left(\frac{t}{T}\right)\epsilon_t X_{t-j}^c\left(\widehat{\phi}_j\left(\frac{t}{T}\right) - \phi_j\left(\frac{t}{T}\right)\right)\\
&\qquad+\frac{1}{T}\sum_{t=1}^{[\alpha T]}\left(\widehat{\boldsymbol{\phi}}\left(\frac{t}{T}\right) - \boldsymbol{\phi}\left(\frac{t}{T}\right)\right)'\mathbf{C}_t\left(\widehat{\boldsymbol{\phi}}\left(\frac{t}{T}\right) - \boldsymbol{\phi}\left(\frac{t}{T}\right)\right),
\end{aligned}\tag{30}$$

where $\mathbf{C}_t = ((X_{t-i}^c X_{t-j}^c)_{i,j})$, $i,j = 1,2$ and $\boldsymbol{\phi}(t/T) = (\phi_1(t/T), \phi_2(t/T))'$. We now show that both sums in (30) are $o_P(1/\sqrt{T})$ uniformly in $\alpha \in [0,1]$. As for the second of these two sums, note that

$$P\left(\sup_{\alpha\in[0,1]}\left|\frac{1}{T}\sum_{t=1}^{[\alpha T]}X_{t-i}^c X_{t-j}^c\right| > C\right) \le P\left(\frac{1}{T}\sum_{t=1}^{T}|X_{t-i}^c X_{t-j}^c| > C\right) \le \frac{\sup_t \operatorname{var}(X_t)}{C} = O\left(\frac{1}{C}\right).$$

Hence $\sup_{0<\alpha\le 1}\frac{1}{T}\sum_{t=1}^{[\alpha T]}|X_{t-i}^c X_{t-j}^c| = O_P(1)$, $i,j = 1,2$, and because, by assumption, $\|\widehat{\phi}_j - \phi_j\|_\infty = o_P(T^{-1/4})$ for $j = 1,2$, the second sum in (30) is of order $o_P(1/\sqrt{T})$ (uniformly in α).

Now we concentrate on the first sum in (30). For a function h on $[0,1]$, let

$$M_{T,j}(h) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\sigma\left(\frac{t}{T}\right)\epsilon_t X_{t-j}^c h\left(\frac{t}{T}\right), \qquad j = 1,2. \tag{31}$$

Note that the first sum in (30) can be written as $-\frac{2}{\sqrt{T}}\times\sum_{j=1}^{2}M_{T,j}((\widehat{\phi}_j - \phi_j)\mathbb{1}_{[0,\alpha]})$. Hence defining $\mathcal{H}_{T,j} = \{h : h = (g - \phi_j)\mathbb{1}_{[0,\alpha]};\ g \in \mathcal{G}, \alpha \in [0,1], \|g - \phi_j\|_\infty \le T^{-1/4}\}$, and recalling that, by assumption, $\widehat{\phi}_j \in \mathcal{G}$, $j = 1,2$, and $\|T^{1/4}(\widehat{\phi}_j - \phi_j)\|_\infty = o_P(1)$, we see that $T^{1/4}(\widehat{\phi}_j - \phi_j) \in \mathcal{H}_{T,j}$ with probability tending to 1 as $T \to \infty$. Hence the first sum in (30) being $o_P(1/\sqrt{T})$ uniformly in α follows from

$$\sup_{h\in\mathcal{H}_{T,j}}|M_{T,j}(h)| = o_P(1) \quad \text{for } j = 1,2. \tag{32}$$

For ease of notation, let $\mathcal{H} = \mathcal{H}_{T,j}$. Suppose that for each $\delta > 0$, there exists a finite partition of $\mathcal{H}$ into sets $\mathcal{H}_k$, $k = 1,\dots,N(\delta)$ (and we construct an appropriate partition later). Then we have, for $\eta, \delta > 0$,

$$\begin{aligned}
&P\Big(\sup_{h\in\mathcal{H}}|M_{T,j}(h)| > \eta\Big)\\
&\quad\le P\Big(\max_{k=1,\dots,N(\delta)}|M_{T,j}(h_{k,*})| > \eta/2\Big)\\
&\qquad+P\Big(\max_{k=1,\dots,N(\delta)}\sup_{g,h\in\mathcal{H}_k}|M_{T,j}(g-h)| > \eta/2\Big).
\end{aligned}\tag{33}$$

Note that $M_{T,j}(h)$ for each fixed h is a sum of martingale differences with respect to the filtration $\{\mathcal{F}_t\}$, where $\mathcal{F}_t$ is the σ-algebra generated by $\epsilon_t, \epsilon_{t-1}, \dots$. Our assumptions imply that $\operatorname{var}(M_{T,j}(h)) = O(\|h\|_\infty) = o(1)$ for any $h \in \mathcal{H}$. This implies that the first term on the right side of (33) tends to 0 as $T \to \infty$. The fact that the second term on the right side of (33) also becomes small follows from lemma 3.3 of Nishiyama (2000) as applied to a martingale process in discrete time as considered here (cf. Nishiyama 2000, sec. 4). To apply this lemma, we need to verify two conditions, called $[PE']$ and $[L1']$ in Nishiyama's article. We do this next.

First, we construct the needed (nested) partition of $\mathcal{H}$. Let $\delta > 0$. Recall that, by assumption, $\mathcal{G}$ has a finite bracketing integral. Without loss of generality, the corresponding partitions can be assumed to be nested (see, e.g., van der Vaart and Wellner 1996; Nishiyama 1996). Let $\mathcal{G}_k$, $1 \le k \le N_B(\delta)$, be the corresponding brackets, that is, $\mathcal{G}_k = \{g \in \mathcal{G} : g_{k,*} \le g \le g_k^*\}$ with functions g_k^* and $g_{*,k}$ satisfying $\rho(g_k^*, g_{*,k}) < \delta$. Divide $[0,1]$ into small intervals $[b_{k-1}, b_k]$ with $0 = b_0 < b_1 < \cdots < b_{M(\delta^2)-1} < b_{M(\delta^2)} = 1$ and $|b_k - b_{k-1}| < \delta^2$ for $1 \le k \le M(\delta^2)$ with $M(\delta^2) = O(1/\delta^2)$. Then let the partition of $\mathcal{H}$ consist of the sets $\mathcal{H}^{k,\ell} = \{h = (g - \phi_j)\mathbb{1}_{[0,\alpha]} \in \mathcal{H} : g \in \mathcal{G}_k, \alpha \in [b_\ell - b_{\ell-1}]\}$, $k = 1,\dots,N_B(\delta)$, $\ell = 1,\dots,M(\delta^2)$. By construction $\int_0^1\sqrt{\log N(\eta)}\,d\eta \le \int_0^1\sqrt{\log N_B(\eta)}\,d\eta + \int_0^1\sqrt{\log M(\eta^2)}\,d\eta < \infty$, and the partition can without loss of generality be chosen to be decreasing in δ.

We need some more notation. Let E_t denote conditional expectation given $\mathcal{F}_t$. Write $M_T(h) = \sum_{t=1}^{T}\xi_t(h)$, where $\xi_t(h) = 1/\sqrt{T}\sigma(t/T)\epsilon_t X_{t-j}^c h(t/T)$. With $\bar{\xi}_t = \sup_{h\in\mathcal{H}}|\xi_t(h)|$, let $V_T(\eta) = \sum_{t=1}^{T}E_{t-1}(\bar{\xi}_t I(\bar{\xi}_t > \eta))$. In this notation, Nishiyama's condition $[L1']$ reads as $V_T(\eta) = o_P(1)$ for every $\eta > 0$. Note that there exists a constant $C > 0$ such that $\bar{\xi}_t \le C\frac{1}{\sqrt{T}}|\epsilon_t X_{t-j}^c|$ and hence $V_T(\eta) \le \frac{C^2}{T}\sum_{t=1}^{T}(X_{t-j}^c)^2 P(|\epsilon_t X_{t-j}^c| > \eta\sqrt{T}/C|X_{t-j}^c)$. Now, note that on the set $A_T = \{\max_{t=1,\dots,T}|X_t^c| < \sqrt{T}/\log T\}$, we have $P(|\epsilon_t X_{t-j}^c| > \eta\sqrt{T}/C|X_{t-j}^c) \le P(|\epsilon_t| > \eta\log T/C) \to 0$ for each $\eta > 0$. Hence on the set A_T, we have $V_T(\eta) = o_P(1)$ for every $\eta > 0$. Because, by assumption, X_t, $t = 1,\dots,T$, have uniformly bounded fourth moments, we also have $P(A_T) \to 1$ as $T \to \infty$.

Next, we turn our attention to Nishiyama's condition $[PE']$. For this, we need an estimate of $E_{t-1}(\mathcal{H}^{k,\ell}) := E_{t-1}|\sup_{\phi,\psi\in\mathcal{H}^{k,\ell}}(\xi_t(\phi) - \xi_t(\psi))^2|$. Note that for $\phi, \psi \in \mathcal{H}^{k,\ell}$, we have $\phi = (h - \phi_j)I_{[0,\alpha]}$ and $\psi = (g - \phi_j)I_{[0,\beta]}$, with $\|g - h\|_\infty \le \eta$ and $|\alpha - \beta| \le \eta^2$. Hence, writing $\phi - \psi =$

$(h-g)I_{[0,\alpha]} + (g-\phi_j)(I_{[0,\alpha]} - I_{[0,\beta]})$ and using the fact that, by assumption, $\sigma^2(\cdot) < M < \infty$, we get

$$E_{t-1}(\mathcal{H}^{k,l}) \le \frac{2M\big(\eta^2 \operatorname{var}(\epsilon_t)(X^c_{t-1})^2 + \frac{1}{\sqrt{T}} \operatorname{var}(\epsilon_t)(X^c_{t-1})^2 I_{[\alpha,\beta]}(\frac{t}{T})\big)}{T},$$

and hence

$$\sup_{\eta\in(0,1)} \max_{1\le k\le N_B(\eta), 1\le \ell\le M(\eta^2)} \frac{\sqrt{\sum_{t=1}^T E_{t-1}(\mathcal{H}^{k,l})}}{\eta} \le M\sqrt{\operatorname{var}(\epsilon_t)}\sqrt{\frac{1}{T}\sum_{t=1}^T (X^c_{t-1})^2} + M\frac{1}{\eta}\sqrt{\frac{\frac{1}{\sqrt{T}}\operatorname{var}(\epsilon_t)\sum_{t=1}^T (X^c_t)^2 I_{[\alpha,\beta]}(\frac{t}{T})}{T}}.$$

Condition [PE'] requires that the expression on the left side of the last inequality be stochastically bounded. We have $|\alpha-\beta| \le \eta^2$, $1/T\sum_{t=1}^T (X^c_{t-1})^2 = O_P(1)$ and $\max_{1\le t\le T}(X^c_t)^2 = O_P(\sqrt{T})$, where for the latter we use Assumption 1(b). Hence we see that the right side is stochastically bounded, and [PE$'$] follows.

This completes the proof for $\widehat{\Sigma}(\alpha)$ the partial sum process based on the regular residuals. The proof for $\widehat{\Sigma}^*(\alpha)$ is similar by using the representation (26) and observing that $|\omega_t(\alpha)| \le 1$.

We now show the fact that $\widehat{\Sigma}(\alpha)$ and $\widehat{\Sigma}^*(\alpha)$ are very close uniformly in α. This implies that the empirical excess mass functionals based on the smoothed and ordinary squared residuals have the same limiting behavior.

Lemma 3. Under Assumptions 1 and 2, we have

$$\sqrt{T}\sup_{\alpha\in[0,1]} |\widehat{\Sigma}(\alpha) - \widehat{\Sigma}^*(\alpha)| = o_P(1).$$

Proof. Using Lemma 2, we can write

$$\widehat{\Sigma}(\alpha) = \frac{1}{T}\sum_{t=1}^{[\alpha T]} \sigma^2(t/T)\epsilon_t^2 + o_P(1/\sqrt{T}), \tag{34}$$

and similarly, for the smoothed version we have

$$\widehat{\Sigma}^*(\alpha) = \frac{1}{T}\sum_{t=1}^{[\alpha T]+[bT]} \sigma^2(t/T)\epsilon_t^2\omega_t(\alpha) + o_P(1/\sqrt{T}), \tag{35}$$

where the $o_P(1/\sqrt{T})$-terms are uniform in $\alpha \in [0,1]$. Because $\omega_t(\alpha) \ne 1$ only for values t close to 1 and $[\alpha T]$, we can write

$$\sqrt{T}\big(\widehat{\Sigma}(\alpha) - \widehat{\Sigma}^*(\alpha)\big) = \begin{cases} Q_T(\alpha) + B_T(\alpha) + C_T(0) + o_P(1) & \text{for } \alpha > 2b \\ C_T(\alpha) + o_P(1) & \text{for } \alpha \le 2b, \end{cases} \tag{36}$$

where

$$Q_T(\alpha) = \frac{1}{\sqrt{T}}\sum_{t=[\alpha T]-[bT]+1}^{[\alpha T]+[bT]} \sigma^2(t/T)\big(\mathbb{1}\{1\le t\le [\alpha T]\} - \omega_t(\alpha)\big) \times (\epsilon_t^2 - 1),$$

$$B_T(\alpha) = \frac{1}{\sqrt{T}}\sum_{t=[\alpha T]-[bT]+1}^{[\alpha T]+[bT]} \sigma^2(t/T)\big(\mathbb{1}\{1\le t\le [\alpha T]\} - \omega_t(\alpha)\big),$$

and

$$C_T(\alpha) = \frac{1}{\sqrt{T}}\sum_{t=1}^{[\alpha T]+[bT]} \sigma^2(t/T)\big(\mathbb{1}\{1\le t\le [\alpha T]\} - \omega_t(\alpha)\big)\epsilon_t^2.$$

The $o_P(1)$ terms in (36) (which are uniform in $\alpha \in [0,1]$) are the sum of the two o_P terms from (34) and (35). First, we consider $C_T(\alpha)$. With $W_t = \sigma^2(\frac{t}{T})(\mathbb{1}\{1\le t\le [\alpha T]\} - \omega_t(\alpha))\epsilon_t^2$, we have

$$\sup_{[\alpha T]\le 2[bT]} |C_T(\alpha)| \le \frac{1}{\sqrt{T}}\sum_{t=1}^{3[bt]} |W_t| = \sqrt{T}b\cdot\frac{1}{bT}\sum_{t=1}^{3[bt]} |W_t| = o(1)O_P(1).$$

The last equality follows because by assumption, $\sqrt{T}b = o(1)$ and $\frac{1}{bT}\sum_{t=1}^{3[bt]} |W_t| = O_P(1)$, because the random variables $|W_t|$ have (uniformly) finite expected values.

Next, we show $\sup_{\alpha\in[0,1]} |Q_T(\alpha)| = o_P(1)$, where, to simplify notation, we have extended the definition of $Q_T(\alpha)$ to all $\alpha \in [0,1]$ to

$$Q_T(\alpha) = \frac{1}{\sqrt{T}}\sum_{t=\max([\alpha T]-[bT]+1,1)}^{[\alpha T]+[bT]} v_t(\alpha)Z_t,$$

where $v_t(\alpha) = \sigma^2(t/T)(\mathbb{1}\{1\le t\le [\alpha T]\} - \omega_t(\alpha))$ and $Z_t = \epsilon_t^2 - 1$. Note that $Q_T(\alpha)$ is a sum of at most $2[bT]$ independent random variables that are not necessarily bounded. Therefore, we use a truncation argument. By assumption, $\mathrm{E}(Z_t^2 \log|Z_t|) < \infty$, we can truncate the variables $|Z_t|$ at $\sqrt{KT/\log T}$ for some appropriate $K > 0$, to be determined later. For any $K > 0$, the difference is negligible, as the following argument shows:

$$\begin{aligned} D_n(\alpha) &:= \left| \frac{1}{\sqrt{T}}\sum_{t=\max([\alpha T]-[bT]+1,1)}^{[\alpha T]+[bT]} v_t(\alpha)Z_t \right. \\ &\qquad \left. - \frac{1}{\sqrt{T}}\sum_{t=\max([\alpha T]-[bT]+1,1)}^{[\alpha T]+[bT]} v_t(\alpha)Z_t\mathbb{1}\left(|Z_t| \le \sqrt{\frac{KT}{\log T}}\right)\right| \\ &\le \frac{1}{\sqrt{T}}\sum_{t=\max([\alpha T]-[bT]+1,1)}^{[\alpha T]+[bT]} M|Z_t|\mathbb{1}\left(|Z_t| > \sqrt{\frac{KT}{\log T}}\right) \\ &\le \frac{1}{\sqrt{T}}\sum_{t=1}^{T+[bT]} M|Z_t|\mathbb{1}\left(|Z_t| > \sqrt{\frac{KT}{\log T}}\right), \end{aligned}$$

and hence for any $\eta > 0$,

$$\begin{aligned} &P\Big(\sup_{\alpha\in[0,1]} D_n(\alpha) > \eta\Big) \\ &\quad \le P\left(\exists\, t \in \{1,\dots,T+[bT]\} : |Z_t| > \sqrt{\frac{KT}{\log T}}\right) \\ &\quad \le (T+[bT])\cdot P\left(|Z_t| > \sqrt{\frac{KT}{\log T}}\right) \to 0 \quad \text{as } T\to\infty, \end{aligned}$$

where the convergence to 0 follows from the moment assumption on ϵ_t. Hence we need only show that the sum of the truncated random variables tends to 0 uniformly over $\alpha \in [0,1]$. This is done by exploiting Bernstein's inequality. Toward this end, we center the truncated variables. Let $\mu_T = EZ_t\mathbb{1}(|Z_t| \le \sqrt{T}/(K\log T))$. First, we show that centering is negligible. Because $EZ_t = 0$, we have $\mu_T \to 0$ as $T \to \infty$. Because, by assumption $b\sqrt{T} \to 0$, it follows that

$$\sup_{\alpha\in[0,1]}\left|\frac{1}{\sqrt{T}}\sum_{t=[\alpha T]-[bT]+1}^{[\alpha T]+[bT]} v_t(\alpha)\mu_T\right| \le M\frac{2[bT]+1}{\sqrt{T}}\mu_T = o(1).$$

Further, with $\mu_4 = \operatorname{var}(Z_t)$, we have $\operatorname{var}(v_t(\alpha)Z_t\mathbb{1}(|Z_t| \le \frac{\sqrt{T}}{K\log T})) \le M^2\mu_4$. Hence, for every fixed $\epsilon > 0$, we have

$$\begin{aligned}
&P\left(\sup_{\alpha\in[0,1]}\left|\frac{1}{\sqrt{T}}\sum_{t=[\alpha T]-[bT]+1}^{[\alpha T]+[bT]} v_t(\alpha)\right.\right.\\
&\qquad\left.\left.\times\left(Z_t\mathbb{1}\left(|Z_t| \le \frac{\sqrt{T}}{K\log T}\right) - \mu_T\right)\right| > \epsilon\right)\\
&\le 2T\cdot P\left(\left|\frac{1}{\sqrt{bT}}\sum_{t=[\alpha T]-[bT]+1}^{[\alpha T]+[bT]} v_t(\alpha)\right.\right.\\
&\qquad\left.\left.\times\left(Z_t\mathbb{1}\left(|Z_t| \le \frac{\sqrt{T}}{K\log T}\right) - \mu_T\right)\right| > \frac{\epsilon}{\sqrt{b}}\right)\\
&\le 2T\cdot\exp\left\{-\frac{1}{2}\frac{\epsilon^2/b}{2M^2\mu_4 + \frac{1}{3}\frac{\epsilon}{\sqrt{b}}\frac{2M\sqrt{T}}{K\log T}/\sqrt{2[bT]}}\right\}\\
&\le 2T\cdot\exp\left\{-\frac{1}{2}\frac{\epsilon^2/b}{2M^2\mu_4 + \frac{\sqrt{2}M\epsilon}{3bK\log T}}\right\}\\
&\le 2T\cdot\exp\{-c_1\epsilon K\log T\},
\end{aligned}$$

for an appropriate constant $c_1 > 0$ and T large enough. For $K > 0$ large enough, the last expression in the foregoing series of inequalities tends to 0 as $T \to \infty$.

It remains to show that $\sup_{\alpha\in[2b,1]}|B_T(\alpha)| = o(1)$. Writing

$$B_T(\alpha) = \sqrt{T}b\left(\frac{1}{bT}\sum_{t=[\alpha T]-[bT]+1}^{[\alpha T]+[bT]}\sigma^2\left(\frac{t}{T}\right)\right.$$
$$\left.\times\left(\mathbb{1}\{1 \le t \le [\alpha T]\} - w_t(\alpha)\right)\right),$$

we observe that the summands are positive and bounded. Because there are (at most) $2[bT]+1$ many such summands, the term in parentheses is bounded uniformly in α. Because $\sqrt{T}b \to 0$, by assumption, it follows that $\sup_{\alpha\in[2b,1]}|B_T(\alpha)| = o(1)$.

Proof of Lemma 1

We mention that the some elements of the proof are similar to those of Polonik (1995). We consider only the case $\widehat{E}_{\widehat{m}}$. The case of the regular empirical excess mass $\widehat{E}$ follows similarly but is simpler, and hence we omit the proof.

Our target here is the empirical excess mass of the (standardized) variance function $\tilde{\sigma}^2_{s,\widehat{m}}(\cdot)$, based on the smoothed residuals and the estimated mode. We present the proof for the excess mass process based on the regular residuals. However, a closer inspection of the proof reveals that everything depends only on properties of the process $\{\sqrt{T}(\widehat{\Sigma} - \Sigma)(\alpha), \alpha \in [0,1]\}$, where $\Sigma(\alpha) = \int_0^\alpha \sigma^2(u)\,du$. In case of the smoothed residuals, this process would be replaced by $\{\sqrt{T}(\widehat{\Sigma}^* - \Sigma)(\alpha), \alpha \in [0,1]\}$, and Lemma 3 shows that these two processes have the same limiting behavior.

We first prove the assertion for the excess mass of $\widehat{\sigma}^2_{\widehat{m}}(\cdot)$, instead of the standardized function $\tilde{\sigma}^2_{\widehat{m}}(\cdot)$. The behavior of the latter follows from the former by an easy argument.

Recall the characterization of $\widehat{\sigma}^2_{\widehat{m}}(\cdot)$ as consisting of two isotonic regressions. To the left of the estimated mode $\widehat{m}$, it is the (right continuous) slope of the greatest convex minorant to the cumulative sum diagram given by the points $(k/T, \sum_{t=1}^k e_t^2)$, $k = 1,\dots,T$ (cf. Robertson et al. 1988), and to the left of $\widehat{m}$, it is the (left-continuous) slope of the least concave majorant to the same cumulative sum diagram. It follows that $\widehat{\sigma}^2_{\widehat{m}}(\cdot)$ is a piecewise constant function with level sets $\{\alpha \in [0,1] : \widehat{\sigma}^2_{\widehat{m}}(\alpha)\} = [\widehat{a}_\lambda, \widehat{b}_\lambda]$ being an interval and $\int_0^{\widehat{b}_\lambda}\widehat{\sigma}^2_{\widehat{m}}(\alpha)\,d\alpha = \sum_{t=1}^{[\widehat{b}_\lambda T]} e_t^2$. Hence, letting

$$\widehat{\Sigma}(a,b) = \frac{1}{T}\sum_{t=[aT]+1}^{[bT]} e_t^2,$$

for the excess mass $E_{\widehat{\sigma}^2,\widehat{m}}(\lambda)$ of $\widehat{\sigma}^2_{\widehat{m}}(\cdot)$, we have that

$$\begin{aligned}
E_{\widehat{\sigma}^2,\widehat{m}}(\lambda) &= \int_0^1 \left(\widehat{\sigma}^2_{\widehat{m}}(u) - \lambda\right)^+ du = \int_{\widehat{a}_\lambda}^{\widehat{b}_\lambda}\left(\widehat{\sigma}^2_{\widehat{m}}(u) - \lambda\right)du\\
&=: \widehat{H}_\lambda(\widehat{a}_\lambda, \widehat{b}_\lambda),
\end{aligned}$$

where $\widehat{H}_\lambda(a,b) = \widehat{\Sigma}(a,b) - \lambda(b-a)$. We also have, by definition of $\widehat{H}_\lambda$, that $\widehat{H}_\lambda(\widehat{a}_\lambda, \widehat{b}_\lambda) = \sup_{0\le a<b\le 1}\widehat{H}_\lambda(a,b)$. We can expect this empirical excess mass to be close to

$$\begin{aligned}
E_{\sigma^2,\widehat{m}}(\lambda) &= \sup_{0\le a\le\widehat{m}\le b\le 1}\left(\Sigma(a,b) - \lambda(b-a)\right)\\
&= \sup_{0\le a\le\widehat{m}\le b\le 1} H_\lambda(a,b) = H(\tilde{a}_\lambda, \tilde{b}_\lambda), \qquad (37)
\end{aligned}$$

where $\tilde{a}_\lambda$ and $\tilde{b}_\lambda$ are defined through the last equality, $\Sigma(a,b) = \int_a^b \sigma^2(u)\,du$, and $H_\lambda(a,b) = \Sigma(a,b) - \lambda(b-a)$. It is straightforward to see that $E_{\sigma^2,\widehat{m}}(\lambda) = \Sigma(a_\lambda, b_\lambda) - \lambda(b_\lambda - a_\lambda)$, where for $\lambda \le \sigma^2(\widehat{m})$, we have $[a_\lambda, b_\lambda] = cl\{u \in [0,1] : \sigma^2(u) \ge \lambda\}$. [Here $cl(A)$ denotes the closure of a set A.] For $\lambda > \sigma^2(\widehat{m})$, the mode $\widehat{m}$ becomes one of the limits of the interval $[\tilde{a}_\lambda, \tilde{b}_\lambda]$ from (37). If $\widehat{m} \le m$, then $\tilde{a}_\lambda = \widehat{m}$, and without loss of generality, we assume this to be the case. It is also not difficult to see that $\tilde{b}_\lambda = b_\lambda$ as long as $E_{\sigma^2,\widehat{m}}(\lambda) > 0$, and there is level $\tilde{\lambda}_{\max} < \sigma^2(m)$ with $E_{\sigma^2,\widehat{m}}(\lambda) = 0$ for all $\lambda > \tilde{\lambda}_{\max}$. Note also that the supremum in the definition of $E_{\sigma^2,\widehat{m}}$ is extended over a smaller set than in the definition of E_{σ^2}. Hence we have $E_{\sigma^2,\widehat{m}}(\lambda) \le E_{\sigma^2}(\lambda)$, with equality for $\lambda \le \sigma^2(\widehat{m})$. Because excess mass functionals are positive and monotonically decreasing, we obtain the "approximation error," $\sup_{\lambda\ge 0}|(E_{\sigma^2,\widehat{m}}(\lambda) - E_{\sigma^2}(\lambda))| \le E_{\sigma^2}(\sigma^2(\widehat{m}))$. The latter is, in fact, the quantity from Assumption 3, and hence is of the order $o_P(1/\sqrt{T})$.

We obtain

$$\begin{aligned}\sqrt{T}\big(E_{\widehat{\sigma}^2,\widehat{m}}(\lambda) - E_{\sigma^2}(\lambda)\big) &= \sqrt{T}\big(E_{\widehat{\sigma}^2\widehat{m}}(\lambda) - E_{\sigma^2,\widehat{m}}(\lambda)\big) + \sqrt{T}\big(E_{\sigma^2,\widehat{m}}(\lambda) - E_{\sigma^2}(\lambda)\big) \\ &= \sqrt{T}\big(E_{\widehat{\sigma}^2\widehat{m}}(\lambda) - E_{\sigma^2,\widehat{m}}(\lambda)\big) + o_P(1).\end{aligned}$$

Hence it remains to consider the process $\sqrt{T}(E_{\widehat{\sigma}^2\widehat{m}}(\lambda) - E_{\sigma^2,\widehat{m}}(\lambda))$. We write

$$\begin{aligned}\sqrt{T}\big(E_{\widehat{\sigma}^2}(\lambda) - E_{\sigma^2,\widehat{m}}(\lambda)\big) &= \sqrt{T}\big(\widehat{H}_\lambda(\widehat{a}_\lambda,\widehat{b}_\lambda) - H_\lambda(a_\lambda,b_\lambda)\big) \\ &= \sqrt{T}(\widehat{H}_\lambda - H_\lambda)(\widehat{a}_\lambda,\widehat{b}_\lambda) + R_1(\lambda) \\ &= \sqrt{T}(\widehat{H}_\lambda - H_\lambda)(a_\lambda,b_\lambda) + R_1(\lambda) + R_2(\lambda), \qquad (38)\end{aligned}$$

where $R_1(\lambda) = \sqrt{T}[H_\lambda(\widehat{a}_\lambda,\widehat{b}_\lambda) - H_\lambda(a_\lambda,b_\lambda)]$ and $R_2(\lambda) = \sqrt{T}[(\widehat{H}_\lambda - H_\lambda)(\widehat{a}_\lambda,\widehat{b}_\lambda) - (\widehat{H}_\lambda - H_\lambda)(a_\lambda,b_\lambda)]$.

Observe that by using the definitions of a_λ, b_λ and $\widehat{a}_\lambda, \widehat{b}_\lambda$ as maximizers of H_λ and $\widehat{H}_\lambda$, we have

$$\begin{aligned}0 &\le H_\lambda(a_\lambda,b_\lambda) - H_\lambda(\widehat{a}_\lambda,\widehat{b}_\lambda) \\ &= (H_\lambda - \widehat{H}_\lambda)(a_\lambda,b_\lambda) - (H_\lambda - \widehat{H}_\lambda)(\widehat{a}_\lambda,\widehat{b}_\lambda) \\ &\quad + \big(\widehat{H}_\lambda(a_\lambda,b_\lambda) - \widehat{H}_\lambda(\widehat{a}_\lambda,\widehat{b}_\lambda)\big) \\ &\le (H_\lambda - \widehat{H}_\lambda)(a_\lambda,b_\lambda) - (H_\lambda - \widehat{H}_\lambda)(\widehat{a}_\lambda,\widehat{b}_\lambda). \qquad (39)\end{aligned}$$

Hence we obtain that

$$\begin{aligned}&|R_1(\lambda) + R_2(\lambda)| \\ &\quad \le 2\big|\sqrt{T}\big((\widehat{H}_\lambda - H_\lambda)(\widehat{a}_\lambda,\widehat{b}_\lambda) - (\widehat{H}_\lambda - H_\lambda)(a_\lambda,b_\lambda)\big)\big|. \qquad (40)\end{aligned}$$

The key observation now is that (40), and also the main term in (38), can be controlled if we have a hand on the process $\sqrt{T}(\widehat{H}_\lambda - H_\lambda)(a,b)$. We show that this process converges weakly to a tight Gaussian process. This, together with (39), implies that $\widehat{a}_\lambda$ and $\widehat{b}_\lambda$ are consistent estimators for a_λ and b_λ (see Lemma 4). A combination of these two results immediately implies [because of (40)] that $|R_1(\lambda) + R_2(\lambda)| = o_P(1)$, and this in turn entails asymptotic normality of our target quantity $\sqrt{T}(E_{\widehat{\sigma}^2}(\lambda) - E_{\sigma^2,\widehat{m}}(\lambda))$ [cf. (38)].

Lemma 4. If $\sup_{0\le a\le b\le 1}|(\widehat{H}_\lambda - H_\lambda)(a,b)| = o_P(1)$ then $\sup_{\lambda>0}|\widehat{a}_\lambda - a_\lambda| = o_P(1)$ and $\sup_{\lambda>0}|\widehat{b}_\lambda - b_\lambda| = o_P(1)$.

The proof of this lemma follows the proof of theorem 3.5 of Polonik (1995). Details are omitted here. Note also that the results that follow imply the assumed uniform consistency of $\widehat{H}_\lambda$.

It remains to prove weak convergence of the process $\sqrt{T}(\widehat{H}_\lambda - H_\lambda)(a,b)$, $a,b\in[0,1]$, to a tight limit. Observe that because $\sigma^2(\cdot)$ is assumed to be totally bounded, we can write $\Sigma(a,b) = \frac{1}{T}\sum_{t=[aT]+1}^{[bT]}\sigma^2(t/T) + O(1/T)$, and hence we have

$$\begin{aligned}&\sqrt{T}(\widehat{H}_\lambda - H_\lambda)(a,b) \\ &\quad = \sqrt{T}\left(\frac{1}{T}\sum_{t=[aT]+1}^{[bT]}\sigma^2(t/T)(\epsilon_t^2 - 1)\right) + O(1/\sqrt{T}), \\ &\quad = Z_T(b) - Z_T(a) + O(1/\sqrt{T}),\end{aligned}$$

where $Z_T(\alpha) = 1/\sqrt{T}\sum_{t=1}^{[\alpha T]}\sigma^2(t/T)(\epsilon_t^2 - 1)$. We now prove that the partial sum process $\{Z_T(\alpha), \alpha\in[0,1]\}$ as a process in $D[0,1]$ converges in distribution to a Gaussian process. This then completes the proof, because it entails asymptotic stochastic equicontinuity, as well as asymptotically normality of the process $Z_T(b) - Z_T(a)$. These are the two properties that we need to prove. The asserted covariance function of the excess mass process then follows straightforwardly from Lemma 5.

Lemma 5. Under the assumptions of Lemma 1, we have as $T\to\infty$ that

$$Z_T(\alpha) \to G(\alpha) \quad \text{in distribution in } D[0,1],$$

where $G(\alpha)$ is a mean-0 Gaussian process with $\operatorname{cov}(G(\alpha), G(\beta)) = (\mu_4 - 1)\int_0^{\min(\alpha,\beta)}\sigma^4(u)\,du$.

Proof. We first prove convergence of the finite-dimensional distributions. Let $Y_{t,T} = \sigma^2(t/T)(\epsilon_t^2 - 1)$, and define

$$B_T^2(\alpha) = \operatorname{var}(\sqrt{T}Z_T(\alpha)) = (\mu_4 - 1)\sum_{t=1}^{[\alpha T]}\sigma^4(t/T), \qquad (41)$$

where μ_4 denotes the fourth moment of ϵ_t.

We use the Lindeberg–Feller central limit theorem. Hence we need to show that

$$\frac{1}{B_T^2(\alpha)}\sum_{t=1}^{[\alpha T]}E\big\{Y_{t,T}^2 I\big(|Y_{t,T}| \ge \epsilon B_T(\alpha)\big)\big\} \to 0 \qquad (42)$$

as $T\to\infty$. We have

$$\begin{aligned}&\frac{1}{B_T^2(\alpha)}\sum_{t=1}^{[\alpha T]}E\big[Y_{t,T}^2 I\big(|Y_{t,T}| \ge \epsilon B_T(\alpha)\big)\big] \\ &\quad = \frac{1}{B_T^2(\alpha)}\sum_{t=1}^{[\alpha T]}E\left[\sigma^4\left(\frac{t}{T}\right)(\epsilon_t^2 - 1)^2 I\left(|\epsilon_t^2 - 1| \ge \frac{\epsilon B_T(\alpha)}{\sigma^2(\frac{t}{T})}\right)\right] \\ &\quad \le \frac{1}{B_T^2(\alpha)}\sum_{t=1}^{[\alpha T]}\sigma^4\left(\frac{t}{T}\right)E\left[(\epsilon_t^2 - 1)^2 I\left(|\epsilon_t^2 - 1| \ge \frac{\epsilon B_T(\alpha)}{M}\right)\right] \\ &\qquad [\text{since } \sigma^2(u) \le M] \\ &\quad = \frac{1}{\mu_4 - 1}E\left[(\epsilon_t^2 - 1)^2 I\left(|\epsilon_t^2 - 1| \ge \frac{\epsilon B_T(\alpha)}{M}\right)\right] \\ &\qquad (\text{because the } \epsilon_t \text{ are iid}),\end{aligned}$$

which converges to 0 because $B_T(\alpha)\to\infty$. Hence we have that for every $\alpha\in(0,1)$, $Z_T(\alpha)/B_T(\alpha) \to^d \mathcal{N}(0,1)$, so that by the Cramer–Wold device and the univariate central limit theorem, for every $(\alpha_1,\ldots,\alpha_m)\in[0,1]^m$, $Z_T(\alpha) \to^d \mathcal{N}(\mathbf{0},\mathbf{\Lambda})$, where $\mathbf{\Lambda}$ is the $m\times m$ variance–covariance matrix $(\Lambda(i,j))$ with

$$\Lambda(i,j) = (\mu_4 - 1)\int_0^{\min(\alpha_i,\alpha_j)}\sigma^4(u)\,du. \qquad (43)$$

It remains to show asymptotic equicontinuity of $Z_T(\alpha)$; that is, we have to show that for every $\epsilon > 0$,

$$\lim_{\delta\downarrow 0}\limsup_{T\to\infty}P\Big(\sup_{|\alpha-\beta|<\delta}|Z_T(\alpha - \beta)| > \epsilon\Big) = 0. \qquad (44)$$

To prove this, we show that

$$E\big[|Z_T(s) - Z_T(r)|^2|Z_T(t) - Z_T(s)|^2\big] \le [F(t) - F(r)]^2 \qquad (45)$$

for $r \le s \le t$, where F is a nondecreasing continuous function on $[0,1]$. Asymptotic equicontinuity then follows (see Billingsley 1999). We have

$$\begin{aligned}
&E\bigl[|Z_T(s) - Z_T(r)|^2 |Z_T(t) - Z_T(s)|^2\bigr] \\
&\quad = E\left(\sqrt{T} \sum_{j=[rT]+1}^{[sT]} \frac{\sigma^2(\frac{j}{T})}{T}(\epsilon_j^2 - 1)\right)^2 \\
&\qquad \times E\left(\sqrt{T} \sum_{j=[sT]+1}^{[tT]} \frac{\sigma^2(\frac{j}{T})}{T}(\epsilon_j^2 - 1)\right)^2 \\
&\quad = \left(\frac{1}{T} \sum_{j=[rT]+1}^{[sT]} \sigma^4\left(\frac{j}{T}\right) \operatorname{var}(\epsilon_j^2)\right) \\
&\qquad \times \left(\frac{1}{T} \sum_{j=[sT]+1}^{[tT]} \sigma^4\left(\frac{j}{T}\right) \operatorname{var}(\epsilon_j^2)\right) \\
&\quad = (\mu_4 - 1)\left(\frac{1}{T} \sum_{j=[rT]+1}^{[sT]} \sigma^4\left(\frac{j}{T}\right)\right)\left(\frac{1}{T} \sum_{j=[sT]+1}^{[tT]} \sigma^4\left(\frac{j}{T}\right)\right) \\
&\quad \le \left((\mu_4 - 1)\int_r^t \sigma^4(u)\,du\right)^2 + O(1/T) \\
&\quad = C\bigl(F(t) - F(r)\bigr)^2,
\end{aligned}$$

where $F(\alpha) = (\mu_4 - 1)\int_0^\alpha \sigma^4(u)\,du$. It is obvious that the last equality holds (for an appropriate C) for all $|t - r| > 1/T$. It also holds for $|t - r| \le 1/T$, because in this case the left side equals 0.

Proof of Lemma 4, Continued. We have studied the excess mass of the estimated variance function without standardization. What we are really interested in is the behavior of $\widehat{E}(\lambda) = E_{\hat{\sigma}}(\lambda)$, which is the excess mass functional of the *standardized* variance function. Using the foregoing notation, we can write the normalizing factor $\int_0^1 \widehat{\sigma}^2(u)\,du = \widehat{\Sigma}(0,1)$, and, similarly, we have $\int_0^1 \sigma^2(u)\,du = \Sigma(0,1)$. It follows that

$$\widehat{E}(\lambda) = \sup_{0 \le a \le b \le 1}\left(\frac{\widehat{\Sigma}(a,b)}{\widehat{\Sigma}(0,1)} - \lambda(b-a)\right).$$

Compared with $E_{\widehat{\sigma}^2}$, which was treated earlier, the quantity $\widehat{\Sigma}(a,b)$ is now replaced by $\widehat{\Sigma}(a,b)/\widehat{\Sigma}(0,1)$. We have seen that $\sqrt{T}(E_{\widehat{\sigma}^2} - E_{\sigma^2})(\lambda) \approx \sqrt{T}(\widehat{\Sigma} - \Sigma)(a_\lambda, b_\lambda)$, where a_λ and b_λ are the maximizers of $\Sigma(a,b) - \lambda(b-a)$. Very similar arguments show that, uniformly in $\lambda > 0$,

$$\left|\sqrt{T}(\widehat{E} - E_{\tilde{\sigma}^2})(\lambda) - \sqrt{T}\left(\frac{\widehat{\Sigma}(a_\lambda^*, b_\lambda^*)}{\widehat{\Sigma}(0,1)} - \frac{\Sigma(a_\lambda^*, b_\lambda^*)}{\Sigma(0,1)}\right)\right| = o_P(1), \tag{46}$$

where a_λ^* and b_λ^* are the maximizers of the standardized excess mass $\Sigma(a,b)/\Sigma(0,1) - \lambda(b-a)$, which means that $(a_\lambda^*, b_\lambda^*) = (a_{\lambda\Sigma(0,1)}, b_{\lambda\Sigma(0,1)})$. All of this holds provided that the process

$$\sqrt{T}\left(\frac{\widehat{\Sigma}(a,b)}{\widehat{\Sigma}(0,1)} - \frac{\Sigma(a,b)}{\Sigma(0,1)}\right) \tag{47}$$

converges to a tight Gaussian limit process. But this is an immediate consequence of the weak convergence of the process $\sqrt{T}(\widehat{\Sigma}(a,b) - \Sigma(a,b))$, as can be seen from rewriting (47) as

$$\begin{aligned}
&\sqrt{T}\left(\frac{1}{\widehat{\Sigma}(0,1)}\bigl(\widehat{\Sigma}(a,b) - \Sigma(a,b)\bigr)\right. \\
&\qquad \left. - \frac{\Sigma(a,b)}{\widehat{\Sigma}(0,1)\Sigma(0,1)}\bigl(\widehat{\Sigma}(0,1) - \Sigma(0,1)\bigr)\right) \\
&\quad = \frac{1}{\widehat{\Sigma}(0,1)}\bigl(Z_T(b) - Z_T(a)\bigr) \\
&\qquad - \frac{\Sigma(a,b)}{\widehat{\Sigma}(0,1)\Sigma(0,1)}\bigl(Z_T(1) - Z_T(0)\bigr) + o_P(1),
\end{aligned}$$

where the last equality follows using Lemma 2. [Note that the $o_P(1)$-term is uniform in λ.] The covariance function of the limit follows through straightforward computation by using the fact that for any two intervals $(a_1, b_1), (a_2, b_2) \subset [0,1]$, we have

$$\begin{aligned}
&\operatorname{cov}\bigl(Z_T(a_1) - Z_T(b_1), Z_T(a_2) - Z_T(b_2)\bigr) \\
&\qquad = (\mu_4 - 1)\int_{(a_1,b_1)\cap(a_2,b_2)} \sigma^4(u)\,du,
\end{aligned}$$

which follows from Lemma 5.

Proof of Theorem 1

Lemma 1, in conjunction with the continuous mapping theorem, yields the assertion.

Proof of Theorem 2

It is sufficient to present only the proof for $\widehat{\lambda}(q)$; the proof for $\widehat{\lambda}_{\widehat{m}}(q)$ follows similarly. We first prove consistency of $\widehat{\lambda}(q)$. In fact, we prove a stronger result; namely, we show that for each $\epsilon > 0$, we have

$$\sqrt{T} \sup_{q \in [\epsilon, 1]} |\widehat{\lambda}(q) - \lambda(q)| = O_P(1). \tag{48}$$

Let $q_T^- = \max(0, q - \sup_{\lambda>0} |(\widehat{E} - E)(\lambda)|)$. Then we can write

$$\begin{aligned}
\widehat{\lambda}(q) &= \inf\{\lambda \ge 0 : \widehat{E}(\lambda) > q\} \\
&= \inf\bigl\{\lambda \ge 0 : E(\lambda) > q + [E(\lambda) - \widehat{E}(\lambda)]\bigr\} \\
&\le \inf\Bigl\{\lambda \ge 0 : E(\lambda) > \max\Bigl(0, q - \sup_{\lambda>0} |(\widehat{E} - E)(\lambda)|\Bigr)\Bigr\} \\
&= \lambda(q_T^-).
\end{aligned}$$

Similarly, we have, with $q_T^+ = \min(1, q + \sup_{\lambda>0} |(\widehat{E} - E)(\lambda)|)$, that $\widehat{\lambda}(q) \ge \lambda(q_T^+)$ such that

$$\lambda(q_T^+) \le \widehat{\lambda}(q) \le \lambda(q_T^-). \tag{49}$$

Our assumptions ensure that $\lambda(\cdot)$ is differentiable, and it is straightforward to see that $\lambda'(q) = -1/(b_{\lambda(q)} - a_{\lambda(q)})$. Because in addition $\sigma^2(\cdot)$ is unimodal and has no flat parts (Assumption 1), it follows that $\lambda'(q)$ is continuous. Lemma 1 together with (46) shows that $\sqrt{T}\sup_{\lambda>0} |\widehat{E}(\lambda) - E(\lambda)| = O_P(1)$. Hence a one-term Taylor expansion applied to $\lambda(q_T^+) - \lambda(q)$ and to $\lambda(q_T^-) - \lambda(q)$ implies (48).

Asymptotic normality follows by a refinement of the foregoing arguments. Let $q_0 \in [0,1)$ be fixed. We show that $\sqrt{T}(\widehat{\lambda}(q_0) - \lambda(q_0))$ has the asserted asymptotic normal distribution. Further let $\delta_T = o(1)$ such that $\sqrt{T}\delta_T \to \infty$, and define $A_T = \{\sup_{q\in[q_0/2,1]} |\widehat{\lambda}(q) - \lambda(q)| < \delta_T\}$. The foregoing shows that $P(A_T) \to 1$ as $T \to \infty$. On A_T, we have, with $q^{\pm}_{T,*} = q_0 + (E - \widehat{E})(\lambda(q_0)) \pm \sup_{|\lambda-\mu|<\delta_T} |(\widehat{E} - E)(\lambda) - (\widehat{E} - E)(\mu)|$, that for T large enough,

$$\begin{aligned}
\widehat{\lambda}(q_0) &= \inf\{\lambda : \widehat{E}(\lambda) > q_0\} \\
&= \inf\big\{\lambda : E(\lambda) > q_0 + [E(\lambda) - \widehat{E}(\lambda)]\big\} \\
&\geq \inf\Big\{\lambda : E(\lambda) > \max\Big[0, q_0 + (E - \widehat{E})(\lambda(q_0)) \\
&\qquad + \sup_{|\lambda-\mu|<\delta_T} |(\widehat{E} - E)(\lambda) - (\widehat{E} - E)(\mu)|\Big]\Big\} \\
&= \lambda(q^{+}_{T,*}).
\end{aligned}$$

Thus on A_T, we can write

$$\begin{aligned}
&\sqrt{T}\big(\widehat{\lambda}(q_0) - \lambda(q_0)\big) \\
&\geq \sqrt{T}\big(\lambda(q^{+}_{T,*}) - \lambda(q_0)\big) = \sqrt{T}\lambda'(\xi^{+}_T)(q^{+}_{T,*} - q_0) \\
&= \lambda'(\xi^{+}_T)\sqrt{T}(E - \widehat{E})(\lambda(q_0)) \\
&\quad + \sqrt{T}\lambda'(\xi^{+}_T) \sup_{|\lambda-\mu|<\delta_T} |(\widehat{E} - E)(\lambda) - (\widehat{E} - E)(\mu)| \\
&= \lambda'(\xi^{+}_T)\sqrt{T}(E - \widehat{E})(\lambda(q_0)) + o_P(1), \qquad (50)
\end{aligned}$$

where the last equality follows from asymptotic stochastic equicontinuity of the process $\sqrt{T}(\widehat{E} - E)(\lambda)$ (see Lemma 1). A similar argument using $\lambda(q^{-}_{T,*})$ instead of $\lambda(q^{+}_{T,*})$ shows that also

$$\sqrt{T}\big(\widehat{\lambda}(q_0) - \lambda(q_0)\big) \leq \lambda'(\xi^{-}_T)\sqrt{T}(E - \widehat{E})(\lambda(q_0)) + o_P(1). \qquad (51)$$

Because the $\xi^{\pm}_T$ stochastically converge to q, and λ' is continuous, the asserted asymptotic normality follows from the asymptotic normality of $\sqrt{T}(E - \widehat{E})(\lambda(q))$ and (50) and (51).

[Received May 2004. Revised June 2005.]

REFERENCES

Bennett, T. J., and Murphy, J. R. (1986), "Analysis of Seismic Discrimination Capabilities Using Regional Data From Western U.S. Events," *Bulletin of the Seismological Society of America*, 76, 1069–1086.

Billingsley, P. (1999), *Convergence of Probability Measures* (2nd ed.), New York: Wiley.

Cavalier, L. (1997), "Nonparametric Estimation of Regression Level Sets," *Statistics*, 29, 131–160.

Cheng, M.-Y., and Hall, P. (1998a), "On Mode Testing and Empirical Approximations to Distributions," *Statistics and Probability Letters*, 39, 245–254.

——— (1998b), "Calibrating the Excess Mass and Dip Tests of Modality," *Journal of the Royal Statistical Society*, Ser. B, 60, 579–589.

Dahlhaus, R. (1997), "Fitting Time Series Models to Nonstationary Processes," *The Annals of Statistics*, 25, 1–37.

Dahlhaus, R., and Neumann, M. (2001), "Locally Adaptive Fitting of Semiparametric Models to Nonstationary Time Series," *Stochastic Processes and Their Applications*, 91, 277–308.

Dahlhaus, R., and Polonik, W. (2004), "Nonparametric Quasi-Maximum Likelihood Estimation for Gaussian Locally Stationary Processes," unpublished manuscript.

——— (2005), "Inference Under Shape Restrictions for Time-Varying Autoregressive Models," unpublished manuscript.

Hartigan, J. A. (1987), "Estimation of a Convex Density Contour in Two Dimensions," *Journal of the American Statistical Association*, 82, 267–270.

Huang, H.-Y., Ombao, H., and Stoffer, D. S. (2004), "Discrimination and Classification of Nonstationary Time Series Using the SLEX Model," *Journal of the American Statistical Association*, 99, 763–774.

Johnson, R. A., and Wichern, D. W. (1998), *Applied Multivariate Statistical Analysis* (4th ed.), Engelwood Cliffs, NJ: Prentice-Hall.

Kakizawa, Y., Shumway, R. H., and Taniguchi, M. (1998), "Discrimination and Clustering for Multivariate Time Series," *Journal of the American Statistical Association*, 93, 328–340.

Mammen, E. (1991), "Estimating a Smooth Monotone Regression Function," *The Annals of Statistics*, 19, 724–740.

Müller, D. W., and Sawitzki, G. (1987), "Using Excess Mass Estimates to Investigate the Modality of a Distribution," Preprint 398, SFB 123, Universität Heidelberg.

——— (1991), "Excess Mass Estimates and Tests for Multimodality," *Journal of the American Statistical Association*, 86, 738–746.

Nishiyama, Y. (1996), "A Central Limit Theorem for ℓ^{∞}-Valued Martingale Difference Arrays and Its Application," Preprint 971, Utrecht University, Dept. of Mathematics.

——— (2000), "Weak Convergence of Some Classes of Martingales With Jumps," *The Annals of Probability*, 28, 685–712.

Nolan, D. (1991), "The Excess Mass Ellipsoid," *Journal of Multivariate Analysis*, 39, 348–371.

Polonik, W. (1995), "Measuring Mass Concentration and Estimating Density Contour Clusters: An Excess Mass Approach," *The Annals of Statistics*, 23, 855–881.

Polonik, W., and Wang, Z. (2005), "Estimation of Regression Contour Clusters: An Application of the Excess Mass Approach to Regression," *Journal of Multivariate Analysis*, 94, 227–249.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988), *Order-Restricted Statistical Inference*, New York: Wiley.

Sakiyama, K., and Taniguchi, M. (2001), "Discriminant Analysis for Locally Stationary Processes," Report S-58, Research Reports in Statistics, Osaka University.

Shumway, R. H., and Stoffer, D. S. (2000), *Time Series Analysis and Its Applications*, New York: Springer-Verlag.

Sun, J., and Woodroofe, M. (1993), "A Penalized Likelihood Estimate of $f(0+)$ When f Is Nonincreasing," *Statistica Sinica*, 3, 501–515.

Whittle, P. (1962), "Gaussian Estimation in Stationary Time Series," *Bulletin of the International Statistical Institute*, 39, 105–129.

Wood, J. (2004), personal communication.

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York: Springer-Verlag.