Consistent classification of non-stationary time series using stochastic wavelet representations

Piotr Fryzlewicz[1] and Hernando Ombao[2]

December 23, 2007

Abstract

A method is proposed for classifying an observed non-stationary time series using a bias-corrected non-decimated wavelet transform. Wavelets are ideal for identifying highly discriminant local time and scale features. We view the observed signals as realizations of locally stationary wavelet (LSW) processes. The LSW model provides a time-scale decomposition of the signals under which we can define and rigorously estimate the evolutionary wavelet spectrum. The evolutionary spectrum, which contains the second-moment information on the signals, is used as the classification signature. For each time series to be classified, we compute the empirical wavelet spectrum and the divergence from the wavelet spectrum of each group. It is then assigned it to the group to which it is the least dissimilar. Under the LSW framework, we rigorously demonstrate that the classification procedure is consistent, i.e., misclassification probability goes to zero at the rate that is proportional to divergence between the true spectra. The method is illustrated using seismic signals and is demonstrated to work very well in simulation studies.

Keywords: Non-decimated wavelet transform, discrimination, locally stationary wavelet processes, evolutionary wavelet spectrum, seismic data.

[1] Corresponding author. Department of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, UK. Email: p.z.fryzlewicz@bristol.ac.uk

[2] Center for Statistical Sciences, Brown University, 121 South Main Street, Providence, Rhode Island, 02912, USA. Email: ombao@stat.brown.edu

1 Introduction

In many applied sciences, it is often of interest to be able to classify an observed time series into one of two or more groups. Examples include underwater acoustics [Huynh et al. (1998)], speech recognition [Mesagarani et al. (2006)], the analysis of EEG signals [Pfurtscheller et al. (2006)], or the analysis of geophone recordings to detect and identify passing animals [Wood et al. (2005)]. The problem of discriminating between earthquake and mining explosion recordings is described in Shumway and Stoffer (2006). It is of importance as it serves as a "proxy" for discriminating between earthquakes and nuclear explosions: a task which is critical in monitoring nuclear non-proliferation. Plots of seismic signals are displayed in Figure 1.

We highlight the primary contributions of our paper. We develop a new procedure for classification of non-stationary time series. Our method uses the (bias-corrected) non-decimated wavelet transform and thus is ideal for identifying localized time-scale features. We view the observed signals as realizations of locally stationary wavelet (LSW) processes. The LSW model provides a time-scale decomposition of the signals under which we can define and rigorously estimate the evolutionary wavelet spectrum. The evolutionary spectrum, which contains the second-moment information on the signals, is used as the classification signature. The procedure is described as follows. For each time series that is to be classified, we compute the empirical wavelet spectrum and the divergence from the wavelet spectrum of each group. It is then assigned it to the group to which it is the least dissimilar. Under the LSW framework, we rigorously demonstrate that the classification procedure is consistent, i.e., the misclassification probability goes to zero at the rate that is proportional to divergence between the true spectra. The method is illustrated using seismic signals and is demonstrated to work well in simulation studies. The remainder of this section motivates our approach and places it in the context of previous work.

A large number of heuristic methods for discriminating and classifying time series data have been proposed in the engineering and applied sciences literature. To quote but one example, an interesting attempt at collecting articles and datasets related to time series data mining is the "University of California Time Series Data Mining Archive", available at `http://www.cs.ucr.edu/~eamonn/TSDMA/`.

In this work, we are concerned with methodology based on more rigorous statistical modelling, existing instances of which fall into two broad categories: those based on stationary, and non-stationary time series modelling. Literature on discrimination and classification derived from stationary time series modelling is substantial, includes both time- and frequency-domain approaches, and is reviewed in detail in Shumway and Stoffer (2006). In practice, however, many observed time series cannot be modelled accurately as stationary. The important class of non-stationary time series includes, for example, time series measured in a changing environment or those describing an evolving phenomenon, such as a speech signal or some seismic activity.

The majority of statistically rigorous approaches to modelling non-stationary time series are derived from the following Cramér-Fourier representation of stationary processes: all zero-mean discrete-time second-order stationary processes X_t can be represented as

$$X_t = \int_{(-\pi,\pi]} A(\omega) \exp(i\omega t) dZ(\omega), \quad t \in \mathbb{Z}, \tag{1.1}$$

where $A(\omega)$ is the transfer function, and $Z(\omega)$ is a random process with zero mean and orthonormal increments. Priestley (1965) introduced an analogous representation for non-stationary processes, where the transfer function $A(\omega)$ was permitted to vary over time. However, this approach did not offer a rigorous framework for asymptotic inference. Dahlhaus (1997) introduced the *locally stationary* modelling philosophy whereby the time-dependent transfer function was defined on a compact interval representing "rescaled time", to enable asymptotic considerations of consistency and inference. Other recent approaches stemming from (1.1) include Mallat, Papanicolau and Zhang (1998) and Swift (2000). Ombao et al. (2001, 2002) proposed the SLEX methodology, in which a locally stationary time series was segmented into dyadic, approximately stationary pieces, each of which was modelled similarly to (1.1), the main difference being that the Fourier harmonics $\exp(i\omega t)$ were replaced by their more localized modifications.

The idea of local stationarity permits asymptotic considerations also in discrimination and classification problems. Recognizing this, Sakiyama and Taniguchi (2004) proposed a classification technique for Dahlhaus' multivariate locally stationary processes, based on an approximation to the Gaussian Kullback-Leibler divergence. A discussion of some practical aspects of this procedure appeared in Shumway (2003). Huang, Ombao and Stoffer (2004) proposed a discrimination and classification technique for locally stationary time series in the SLEX model. Their method consisted of a best dyadic basis selection step followed by computing a discriminant statistic related to the Kullback-Leibler distance in the chosen basis. Chandler and Polonik (2006) proposed a discrimination procedure for locally stationary AR processes (with a time-varying variance function) which was not based on a distance measure but instead used the shape of the variance function as the discriminant criterion. A Fourier-based procedure, in the context of functional classification where the functions were assumed to lie in a Hilbert space, was proposed by Biau, Bunea and Wegkamp (2005).

Being localized both in time and in frequency, wavelets provide a natural alternative to the Fourier-based approach for modelling non-stationary phenomena whose spectral characteristics evolve over time. See Vidakovic (1999) for an introduction to wavelets and their use in statistics. With this in mind, Nason, von Sachs and Kroisandt (2000) proposed the Locally Stationary Wavelet (LSW) time series model, which uses non-decimated wavelets, rather than Fourier exponentials, as building blocks. The LSW model provides a time-scale decomposition of the process and permits rigorous estimation of the *evolutionary wavelet spectrum* and the *local autocovariance*, offering the user freedom in choosing the underlying wavelet family. Wavelet-based estimators of the second-order structure of LSW processes are naturally localized and can typically be computed more efficiently than the corresponding statistics based on the local periodogram in the Dahlhaus model. Also, unlike the SLEX model, the LSW model does not suffer from the constraint of dyadic segmentation.

Wavelets and wavelet packets have been utilized in supervised learning (discrimination and classification) and unsupervised learning (clustering). Using a Kullback-Leibler criterion, Meyer and Chinrungrueng (2003) developed a procedure for clustering signals based on their wavelet packet coefficients. In a related work, Vannucci, Sha and Brown (2005) proposed a Bayesian method for selecting wavelet packet features for clustering. The emphasis of our work is different in that we combine the use of wavelets with rigorous stochastic non-stationary time series modelling. Our LSW-based approach permits us to analyze the asymptotic behavior and establish consistency of our classifiers and also leads to algorithms which are different to those proposed by the above authors.

The article is organized as follows. In Section 2, we discuss the locally stationary wavelet (LSW) model and estimation of the wavelet spectrum. The LSW classification method is discussed in Section 3 along with the main result on consistency. In Section 4, we analyze the seismic data set and demonstrate that our method works very well via simulation studies.

2 The LSW model

2.1 Definition

We start by defining the LSW model for locally stationary time series. The defnition is as in Fryzlewicz and Nason (2006).

Definition 2.1. *A triangular stochastic array $\{X_{t,T}\}_{t=0}^{T-1}$, for $T = 1, 2, \ldots$, is in the class of LSW processes if there exists a mean-square representation*

$$X_{t,T} = \sum_{j=-\infty}^{-1} \sum_{k=-\infty}^{\infty} W_j(k/T)\psi_{j,t-k}\xi_{j,k}, \tag{2.1}$$

where $j \in \{-1, -2, \ldots\}$ and $k \in \mathbb{Z}$ are, respectively, scale and location parameters, $\psi_j = (\psi_{j,0}, \ldots, \psi_{j,\mathcal{L}_j})$ are discrete, real-valued, compactly supported, non-decimated wavelet vectors normalised to one in the l_2 norm, and $\xi_{j,k}$ are zero-mean orthonormal identically distributed random variables. Also, for each $j \leq -1$, $W_j(z) : [0, 1] \to \mathbb{R}$ is a real-valued, piecewise constant function with a finite (but unknown) number of jumps. Let L_j denote the total magnitude of jumps in $W_j^2(z)$. The functions $W_j(z)$ satisfy

- $\sum_{j=-\infty}^{-1} W_j^2(z) < \infty$ *uniformly in z,*
- $\sum_{j=-\infty}^{-1} 2^{-j} L_j < \infty$.

In formula (2.1), the parameters $W_j(k/T)$ can be thought of as a scale- and location-dependent transfer function, while the non-decimated wavelet vectors ψ_j can be thought of as building blocks analogous to the Fourier exponentials in (1.1). Throughout the paper, we work with Gaussian LSW processes, i.e. our $\xi_{j,k}$ are distributed as $N(0,1)$. This is merely for technical convenience and, in principle, our approach could be extended to other distributions.

Haar wavelets are the simplest example of a wavelet system which can be used in formula (2.1). Denote $\mathbb{I}_A(k) = 1$ when k is in A and zero otherwise. Haar wavelets are defined by

$$\psi_{j,k} = 2^{j/2}\mathbb{I}_{\{0,\ldots,2^{-j-1}-1\}}(k) - 2^{j/2}\mathbb{I}_{\{2^{-j-1},\ldots,2^{-j}-1\}}(k),$$

for $j \in \{-1,-2,\ldots\}$ and $k \in \mathbb{Z}$, where $j = -1$ corresponds to the finest scale. Other Daubechies' compactly supported wavelets [Daubechies (1992)] can also be used.

The main quantity of interest in the LSW framework is the evolutionary wavelet spectrum $S_j(z) := W_j^2(z)$, $j = -1, -2, \ldots$, defined on the rescaled-time interval $z \in [0,1]$. Indeed, the (empirical) evolutionary wavelet spectrum is the core concept in the classification rule proposed in Section 3. The rescaled-time formulation is as in nonparametric regression and is done to enable rigorous asymptotic considerations.

From Definition 2.1, it is immediate that $\mathbb{E}X_{t,T} = 0$ and indeed, throughout the paper, we work with zero-mean processes. Such processes arise, for example, when the trend has been removed from the data, see e.g. von Sachs and MacGibbon(2000) for a wavelet-based technique for detrending locally stationary processes. The primary interest is in the second order structure of the process (i.e., covariance, correlation, spectrum) and the goal is to identify specific time-frequency or time-scale components that most effectively discriminate the groups.

The piecewise constant constraint on $W_j(z)$ enables the modelling of processes whose second-order structure evolves over time in a discontinuous (piecewise constant) manner, but is also convenient for processes which can be well approximated as piecewise stationary. We note that unlike the SLEX model, we do not require that breaks in the spectrum occur at dyadic locations.

For an extensive discussion of the philosophy and several aspects of LSW modelling the reader is referred to Nason, von Sachs and Kroisandt (2000). Estimation in the LSW framework is also considered in Fryzlewicz and Nason (2006), who use the Haar-Fisz methodology, and by Van Bellegem and von Sachs (2007), who consider pointwise estimation based on the idea of adaptive intervals of near-homogeneity.

2.2 Empirical wavelet spectrum

In this section, we construct a "pre-estimator" of the evolutionary wavelet spectrum $S_j(z)$, which will form the basis of of classification rule of Section 3. A starting point for this discussion is the definition of the *wavelet periodogram*, which follows.

Definition 2.2. *Let $X_{t,T}$ be an LSW process constructed using the wavelet system ψ. The triangular stochastic array*

$$I_{t,T}^{(j)} = \left| \sum_s X_{s,T} \psi_{j,s-t} \right|^2$$

is called the wavelet periodogram of $X_{t,T}$ at scale j.

Throughout the paper, we assume that the reader is familiar with the fast Discrete Wavelet Transform [DWT; Mallat (1989)], as well as with the fast Non-decimated DWT [NDWT; see Nason and Silverman (1995)]. In practice, we only observe a single row of the triangular array $X_{t,T}$. The wavelet periodogram is not computed separately for each scale j but instead, we compute the full NDWT transform of the observed row of $X_{t,T}$ (e.g. with periodic boundary conditions), and then square the wavelet coefficients to obtain $I_{t,T}^{(j)}$ for $t = 0, \ldots, T-1$ and $j = -1, \ldots, -J(T)$, where $J(T) \leq \log_2 T$.

It is convenient to recall two further definitions from Nason, von Sachs and Kroisandt (2000) at this point: the *autocorrelation wavelets* $\Psi_j(\tau) = \sum_{k=-\infty}^{\infty} \psi_{j,k}\psi_{j,k+\tau}$ and the invertible *autocorrelation wavelet inner product matrix* $A_{i,j} = \sum_\tau \Psi_i(\tau)\Psi_j(\tau)$, whose entries are all positive.

Without the need to make this statement more precise at the moment, we mention that the wavelet periodogram is, in a certain sense, an asymptotically unbiased estimator of the quantity

$$\beta_j(z) := \sum_{i=-\infty}^{-1} S_i(z) A_{i,j}. \tag{2.2}$$

For the above result to hold in the sense of Proposition 2.1 from Fryzlewicz and Nason (2006), we require the following assumption.

Assumption 2.1. *The set of those locations z where (possibly infinitely many) functions $S_j(z)$ contain a jump, is finite. In other words, let $\mathcal{B} := \{z : \exists j \quad \lim_{u \to z^-} S_j(u) \neq \lim_{u \to z^+} S_j(u)\}$. We assume $B := \#\mathcal{B} < \infty$.*

In addition to Assumption 2.1, the following assumption guarantees the uniqueness of the evolutionary wavelet spectrum $S_j(z)$ in the L_2 sense, as discussed in Fryzlewicz and Nason (2006).

Assumption 2.2. *There exists a positive constant C_1 such that for all j, $S_j(z) \leq C_1 2^j$.*

We note that, in particular, Assumption 2.2 is satisfied if $X_{t,T}$ is the standard white noise process, for which $S_j(z) = S_j = 2^j$ [see Fryzlewicz, Van Bellegem and von Sachs (2003)].

Formula (2.2) suggests that the natural pre-estimator of $S_j(k/T)$ is the *empirical wavelet spectrum*, given by

$$L_{k,T}^{(j)} = \sum_{i=-1}^{-J^*} (A^{-1})_{i,j} I_{k,T}^{(i)}, \tag{2.3}$$

where the clip-off scale $-J^*$ will be specified later. Indeed, the theory of Section 3 demonstrates that our classification rule, based on $L_{k,T}^{(j)}$, is asymptotically consistent.

3 Classification method

In this section, we propose a classification rule which assigns an LSW process $X_{t,T}$ to one of $C \geq 2$ groups $\Pi_1, \ldots, \Pi_C$ with evolutionary wavelet spectra $S_j^{(1)}(z), \ldots, S_j^{(C)}(z)$. For simplicity of exposition, as in Huang, Ombao and Stoffer (2004), we only consider the case $C = 2$, although our method is applicable to the general case $C \geq 2$.

We first outline the generic algorithm, and then discuss its various important aspects. The classification algorithm proceeds as follows.

1. For a (suitably selected) subset $\mathcal{M}$ of scale and location indices (j, k), where $\mathcal{M} \subseteq \{-1, \ldots, -\log_2 T\} \times \{0, \ldots, T-1\}$, compute the empirical wavelet spectrum $L_{k,T}^{(j)}$ of $X_{t,T}$.

2. Compute the squared quadratic distances between the empirical wavelet spectrum and the evolutionary wavelet spectra of each of the groups

$$D_g = \sum_{(j,k)\in\mathcal{M}} \{L_{k,T}^{(j)} - S_j^{(g)}(k/T)\}^2,$$

for group $g = 1, 2$.

3. If $D_1 < D_2$, then classify $X_{t,T}$ to Π_1. Otherwise, classify it to Π_2.

In this section, we discuss some remarks on the implementation and state our main result on the consistency of our classification procedure.

3.1 Remarks on the methodology and implementation

Case $C > 2$. The theory and practice of our classification methodology extends naturally to the case where there are more than two groups. In that case, we classify the time series $X_{t,T}$ under consideration to the group for which the squared quadratic distance between the wavelet spectrum of the group and the empirical wavelet spectrum of $X_{t,T}$ is the smallest. Extending the result of Theorem 3.1 to this case is straightforward but notationally burdensome, so we omit it.

Obtaining $S_j^{(g)}(z)$ in practice. Obviously, in practice the true wavelet spectra $S_j^{(g)}(z)$ are unknown, and are replaced by the empirical wavelet spectra, averaged across time series replicates. Suppose that there are N_g independent time series from class Π_g for $g = 1, 2$, and denote the empirical wavelet periodogram for n-th series in this group by $L_{k,T}^{(j),g,n}$. Then we replace $S_j^{(g)}(k/T)$ by its estimate

$$\widehat{S}_j^{(g)}(k/T) = \frac{1}{N_g}\sum_{n=1}^{N_g} L_{k,T}^{(j),g,n}.$$

In cases where N_g is small, extra smoothing over k, e.g. via kernels, may be required. In our implementation, kernel smoothing with the Gaussian kernel is applied.

Choice of the discriminating set $\mathcal{M}$. In Theorem 3.1, we assume that the set $\mathcal{M}$ is “arbitrary but fixed” and that it satisfies certain technical conditions. Each time-scale coefficient is a potential candidate for membership in the set $\mathcal{M}$. For each time-scale index (j,k), we compute a divergence index

$$\Delta(j,k) = \left[S_j^{(1)}(k/T) - S_j^{(2)}(k/T)\right]^2$$

which measures their ability to separate the groups. Then we order these divergence values and choose only the top pre-specified proportion of the coefficients. Variants of our approach to choosing this proportion are discussed at the end of this section.

3.2 Main result

Theorem 3.1. *Suppose that Assumptions 2.1 and 2.2 hold, and that the constants L_j from Definition 2.1 decay as $O(a^j)$ for $a > 2$. Let $S_j^{(1)}(z)$ and $S_j^{(2)}(z)$ be two non-identical wavelet spectra from processes satisfying the short-memory assumption $\sup_t \sum_\rho |corr(X_t, X_{t+\rho})| < \infty$. Let $I_{k,T}^{(j)}$ be the wavelet periodogram constructed from a process with spectrum $S^{(1)}(z)$, and let $L_{k,T}^{(j)}$ be the corresponding bias-corrected periodogram (see formula (2.3)), with $J^* = \log_2 T$. Let the set $\mathcal{M}$ satisfy*

$$\sum_{(j,k)\in\mathcal{M}} \{S_j^{(1)}(k/T) - S_j^{(2)}(k/T)\}^2 = O(T). \tag{3.1}$$

The probability of misclassifying $L_{k,T}^{(j)}$ as coming from a process with spectrum $S_j^{(2)}(z)$ can be bounded as follows.

$$P(D_1 > D_2) = O(T^{-1}\log_2^3 T + T^{1/\{2\log_2(a)-1\}-1}\log_2^2 T).$$

3.3 Other variations of the method

Choice of a divergence index. The discrepancy measures of choice for classification in Dahlhaus' locally stationary model [Shumway (2003)] and in the SLEX model [Huang, Ombao and Stoffer (2004)] are certain approximations to the likelihood ratios, which both sets of authors show to be linked to the Kullback-Leibler divergence. However, there are good reasons for us to depart from this convention and use the simple squared quadratic distance instead. We briefly summarize them below.

- Unlike, for example, the SLEX model, in which the SLEX periodogram is asymptotically distributed as a collection of independent scaled exponentials, the empirical wavelet spectrum $L_{k,T}^{(j)}$ is a collection of random variables which neither are independent, nor is their (joint or marginal) distribution easy to determine. Thus following the likelihood ratio route is not an attractive option for us.

- $L_{k,T}^{(j)}$ is a sum of a (typically logarithmic) number of terms, see formula (2.3), which, by a CLT-type mechanism, brings it closer to Gaussianity than the plain wavelet periodogram $I_{k,T}^{(j)}$, which is distributed as a scaled χ_1^2. The proximity of $L_{k,T}^{(j)}$ to Gaussian suggests that the use of the quadratic distance might be an appealing option.

- As a matter of fact, the quadratic distance is closely linked to the likelihood ratio for an exponential random variable (the latter occurring in SLEX classification, as described above). We briefly illustrate it now. For the purpose of this paragraph, let X denote an exponential variable which we wish to classify as having mean λ_1 or λ_2. The log-likelihood ratio is

  $$\log(\lambda_2/\lambda_1) + X(1/\lambda_2 - 1/\lambda_1), \tag{3.2}$$

  where we classify to λ_1 if (3.2) is positive. However, Taylor-expanding (3.2) around $(\lambda_1, \lambda_2) = (X, X)$ up to the quadratic term, we obtain

  $$\frac{1}{2X^2}\left\{(\lambda_2 - X)^2 - (\lambda_1 - X)^2\right\},$$

  which is a scaled version of the squared quadratic distance which we use, where the scaling is by the factor of $1/(2X^2)$. Our classification procedure also uses the idea of scaling, albeit from a slightly different perspective: namely, prior to the classification, we always rescale all processes to have the sample variance of one. This corresponds to the theoretical assumption that $\int_0^1 \sum_j S_j^{(g)}(z)dz$ is constant for all groups g. The approach of classifying the time across series according to the relative decomposition of its variance, rather than absolute variance, is taken in Shumway (2003) and Huang, Ombao and Stoffer (2004).

- Last but not least, the squared quadratic distance offers very good practical performance (see Section 4), and is relatively easy to analyze theoretically.

Objective criteria for obtaining $\mathcal{M}$. The non-decimated wavelet transform provides a family of time-scale features that could be used for classification. In particular, the wavelet spectra is defined on a time-scale grid which consists of $T \log_2 T$ coefficients. However, some of these features are unable to separate groups. For example, if we were to consider the distribution of the empirical wavelet spectra at a particular (j^*, k^*), the histogram plots of the two groups may share a large overlap and thus fail to distinguish between groups. Such features are typically not good for classification. The goal of this paper is not to propose a feature extraction procedure. Rather, our primary contribution is a classification procedure which is rigorously demonstrated to be consistent under well-separated spectra of groups. However, to complete our discussion, we point out that this particular problem of feature selection can be cast under the general framework of variable selection for a large number of potential variables. This is an extremely active area of statistical research. One objective approach to choosing the features is to perform multiple tests of hypothesis (with a suitable correction procedure). We may test the null hypothesis that $S_{j^*}^{(1)}(k^*/T) = S_{j^*}^{(2)}(k^*/T)$ for each index (j^*, k^*) using non-parametric procedures such as the Wilcoxon test and the permutation test if we are unwilling to make strong parametric assumptions. Recent discussions on the subject of variable selection and multiple hypothesis testing are discussed in Bunea (2007), Bunea, Wegkamp and Auguste (2006), Genovese, Roeder and Wasserman (2006), Jin and Cai (2006), Sun and Cai (2007) and Wasserman and Roeder (2007), among many others. Another objective approach is to choose, from a pre-selected grid on the interval $(0, 1)$, the optimal proportion which minimizes the classification error. This may entail performing a leave-one-out cross validation or some $k-$fold cross validation which is popular in the machine learning literature. This approach is computationally intensive but is attractive because it is tied directly to the goal of minimizing prediction or classification error. We note that in the seismic data example of Section 4.2, the discriminatory coefficients were selected automatically from the data via leave-one-out cross-validation.

4 Applications of the method

4.1 Simulation studies

The goals of the simulation studies were twofold. First, we examined how well the procedure selected time-scale features that were good discriminators. Second, we investigated how well the classification scheme worked for finite samples. Two types of processes were considered. The first consisted of piecewise auto-regressive processes while the second consisted of auto-regressive processes with slowly evolving parameters.

We generated 100 training data sets, each of which consisted of $N = 8$ time series for each of the two groups. To evaluate the classification error, a testing data set (one time series for each of the two groups) is generated for each training data set. The testing data sets were independent of the training data. Time-scale coefficients were selected according to how well they separated the two training groups according to the L_2 criterion. Test time series were classified according to the top $p = 10\%$, 25% and 50% of the time-scale coefficients. We note that in the seismic data example of Section 4.2, the optimal proportion p was selected from the data via cross-validation.

The performance of the proposed LSW method is compared with the SLEX method in Huang, Ombao and Stoffer (2004), which is the current state-of-the-art among statistically rigorous procedures. Though there are other classification procedures for non-stationary time series, the SLEX, which has been demonstrated to do very well in practice, is its most comparable competitor in terms of over-all structure and philosophy.

Piecewise AR processes. The n-th time series from group g, denoted $X_{n,t}^{(g)}$, is generated from the piecewise first order auto-regressive process defined by

$$X_{n,t}^{(g)} \;=\; \begin{cases} \phi_1^{(g)}\, X_{n,t-1}^{(g)} \;+\; \epsilon_{n,t}^{(g)}, & t = 1, \ldots, T_1 \\ \phi_2^{(g)} X_{n,t-1}^{(g)} \;+\; \epsilon_{n,t}^{(g)}, & t = T_1 + 1, \ldots, T_1 + T_2 \end{cases}$$

where $T = T_1 + T_2$; $\epsilon_{n,t}^{(g)}$ is iid Gaussian with mean 0 and variance σ^2.

In the ensuing illustration, we used the following simulation parameters. $G = 2$ is the total number of groups; $N = 8$ is the total number of subjects in each group; $T_1 = 100$ is the length of the first block of time series; $T_2 = 156$ is the length of the second block of time series. The noise variance is $\sigma^2 = 1$. The AR model parameters are as follows. For group 1, $\phi_1^{(1)} = -0.05$ and $\phi_2^{(1)} = 0.10$. For group 2, $\phi_1^{(2)} = -0.05$ and $\phi_2^{(2)} = -0.10$. In the remainder of this section, we use the Daubechies' Extremal Phase wavelet no. 10 from the `wavethresh` software package for R, although any wavelet system could in principle be used: see Section 5 for a discussion of this issue. Representative time series plots from each group and the estimated wavelet spectra are shown in Figure 2. The squared divergence between the wavelet spectra and the best discriminant features selected are displayed in Figure 3. The SLEX method when applied with either finest level $J = 3$ or $J = 4$ gave a correct classification rate of $54 - 58\%$. The proposed LSW method gave the following correct classification rates: 75%, 83% and 87% when the top $p = 0.10, 0.25$ and $p = 0.50$, respectively, of the coefficients were selected.

Slowly-varying AR processes. We considered AR(2) processes with slowly-varying coefficients defined by

$$Y_{n,t}^{(g)} = \phi_1^{(g)}(t) Y_{n,t-1}^{(g)} + \phi_2^{(g)}(t) Y_{n,t-2}^{(g)} + \epsilon_{n,t}^{(g)}$$

where the AR parameters are

$$\begin{aligned} \phi_1^{(g)}(t) &= -0.8 \left[1 - \alpha^{(g)} \cos(\pi t/T)\right] \\ \phi_2^{(g)}(t) &= -0.81. \end{aligned}$$

The α parameters are $\alpha^{(1)} = 0.7$ for group 1 and $\alpha^{(2)} = 0.1$ for group 2. Representative signals from each group and an estimate of the wavelet spectra are plotted in Figure 4. The squared difference in the average wavelet spectra and the most highly discriminant time-scale features are given in Figure 5. The SLEX method gave the correct classification rates of 62% and 71% for $J = 3$ and $J = 4$ respectively. The LSW method gave correct rates of 94%, 88% and 90% when using the top proportions $p = 0.10, 0.25, 0.50$ respectively.

The simulation studies provide empirical evidence that the LSW method works very well even for time series data of length $T = 256$ and a relatively small training dataset of $N = 8$. Moreover, it overcomes the inherent weaknesses of the SLEX approach. It does not suffer from the dyadic restrictions and it is superior both for processes with slowly-varying, and for processes with piecewise-constant, parameters.

4.2 Seismic signals

In the monitoring a comprehensive test ban treaty, it is critical to develop methods for discriminating between nuclear explosions and earthquakes. We applied the proposed LSW methodology for classifying a time series as either an explosion or earthquake. The proliferation of nuclear explosions are monitored in regional distances of $100 - 2000$ km and thus the data of mining explosions can serve as a reasonable proxy. A data set constructed by Blandford (1993) which are regional ($100 - 2000$ km) recordings of several typical Scandinavian earthquakes and mining explosions measured by stations in Scandinavia are used in this study. The data set, consisting of eight earthquakes and eight explosions and an event of uncertain origin located in the Novaya Zemlya region of Russia (called the NZ event), are given in Kakizawa, Shumway and Taniuguchi (1998). The problem is discussed in detail in Shumway and Stoffer (2006, Chapter 7). and the data are available online from `http://lib.stat.cmu.edu/general/tsa2/`. Using a leave-one-out cross validation procedure to choose the proportion p (by examining different values of p over a pre-specified grid, in this case $p = 0.1\,k$, $k = 1, \ldots, 10$), 14 out of the 16 seismic signals were correctly classified. Moreover, the NZ event is classified as an explosion which is consistent with the findings in Shumway (2003) and Huang, Ombao and Stoffer (2004). Figure 6 further illustrates our analysis of this data set.

5 Discussion

In summary, we developed a procedure for classifying non-stationary time series using the locally stationary wavelet (LSW) processes. The LSW model provides a time-scale decomposition of the signals under which we can define and rigorously estimate the evolutionary wavelet spectrum. Under the LSW framework, we rigorously demonstrated that the classification procedure is consistent, i.e., misclassification probability goes to zero at the rate that is proportional to divergence between the true spectra. The method was demonstrated to work well in simulation studies and in the seismic data example. In this section, we discuss those aspects of our work which we believe open up interesting avenues for further research.

1. In the selection of the subset $\mathcal{M}$, we select those coefficients which differ the most in terms of their quadratic distance from their counterparts in the other group. We chose the quadratic selection distance for "consistency" with our L_2 classification distance but we note that any other L_p selection distance would obviously lead to the same $\mathcal{M}$ as functions $f(x) = x^p$ are order-preserving. A more interesting option would be to try other procedures for selecting $\mathcal{M}$ such as that based on maximising the Fisz distance $|S_j^{(1)}(k/T) - S_j^{(2)}(k/T)|/\{|S_j^{(1)}(k/T)| + |S_j^{(2)}(k/T)|\}$ (which measures the relative discrepancy between the spectra; see Fryzlewicz and Nason (2006) for a discussion of the Fisz transform) or that based on choosing those (j, k) for which, for example, as many as possible empirical $S_j^{(1)}(k/T)$'s for the first group lie above (or below) all $S_j^{(2)}(k/T)$'s for the second group. We mention again that regardless of how $\mathcal{M}$ was selected, the consistency theory still holds as long as $\mathcal{M}$ satisfies (3.1), or even a weaker version of (3.1) where we simply require that the two spectra are identifiably different (however, this weakening would then affect the consistency rate).

2. Although $S_j(z)$ typically decays exponentially as $j \to -\infty$ (e.g. $S_j(z) = 2^j$ for standard white noise), it was our deliberate choice not to assign higher weights to the selection distance for coarser scales. This was because, typically, empirical wavelet spectra are less accurate (as estimators of $S_j(z)$) at coarser scales, so downweighting those less accurate quantities would not necessarily be detrimental to the quality of the classification procedure. However, some way of up-weighting those coefficients for processes with non-negligible low-frequency components would be desirable: we note that none of our examples throughout the paper had prominent low-frequency features. We leave this interesting issue for future research.

3. The wavelet system used is a parameter of our procedure which can be optimized over: ideally, we would envisage the use of the wavelet which, given a data set, provides

the best classification results for those series which are known to lie in either group. However, in simulations, we found that our method was fairly robust to the choice of the wavelet, as long as the wavelet was non-adaptive. We have experimented with the use of some adaptive wavelet transforms, such as the Unbalanced Haar transform [Fryzlewicz (2007)], and obtained encouraging preliminary results which will be reported elsewhere.

Appendix. Proof of Theorem 3.1

We write

$$\begin{aligned} D_1 - D_2 &= -2 \sum_{(j,k)\in\mathcal{M}} \{L^{(j)}_{k,T} - S^{(1)}_j(k/T)\}\{S^{(1)}_j(k/T) - S^{(2)}_j(k/T)\} \\ &- \sum_{(j,k)\in\mathcal{M}} \{S^{(1)}_j(k/T) - S^{(2)}_j(k/T)\}^2 =: X - t \end{aligned}$$

and use the Chebyshev inequality to bound

$$P(X - t > 0) \leq E(X^2)/t^2. \tag{5.1}$$

Instead of bounding $E(X^2)$, we bound $E(\tilde{X}^2)$, where

$$\tilde{X} = -2 \sum_{j=-1}^{-J_0} \sum_{k=1}^{T} \{L^{(j)}_{k,T} - S^{(1)}_j(k/T)\}\{S^{(1)}_j(k/T) - S^{(2)}_j(k/T)\}.$$

Since $\mathcal{M} \subset \{-1, \ldots, -J_0\} \times \{1, \ldots, T\}$, the bound for $E(X^2)$ is found similarly and does not exceed the bound for $E(\tilde{X}^2)$. We have

$$E(\tilde{X}^2) = 4E\left\{\sum_{j=-1}^{-J_0}\sum_{k=1}^{T}\left\{\sum_{i=-1}^{-J^*} A^{-1}_{i,j} I^{(i)}_{k,T} - \sum_{i=-1}^{-\infty} A^{-1}_{i,j}\beta^{(1)}_i(k/T)\right\}\{S^{(1)}_j(k/T) - S^{(2)}_j(k/T)\}\right\}^2 =$$

$$4E\left\{\sum_{j=-1}^{-J_0}\sum_{k=1}^{T}\left\{\sum_{i=-1}^{-J^*} A^{-1}_{i,j}[I^{(i)}_{k,T} - \beta^{(1)}_i(k/T)] - \sum_{i=-J^*-1}^{-\infty} A^{-1}_{i,j}\beta^{(1)}_i(k/T)\right\}\{S^{(1)}_j(k/T) - S^{(2)}_j(k/T)\}\right\}^2$$

$$\leq 8E\left\{\sum_{j=-1}^{-J_0}\sum_{k=1}^{T}\sum_{i=-1}^{-J^*} A^{-1}_{i,j}\{I^{(i)}_{k,T} - \beta^{(1)}_i(k/T)\}\{S^{(1)}_j(k/T) - S^{(2)}_j(k/T)\}\right\}^2 +$$

$$+8E\left\{\sum_{j=-1}^{-J_0}\sum_{k=1}^{T}\sum_{i=-J^*-1}^{-\infty} A^{-1}_{i,j}\beta^{(1)}_i(k/T)\{S^{(1)}_j(k/T) - S^{(2)}_j(k/T)\}\right\}^2 =: I + II.$$

As in the proof of Lemma B.3 from Fryzlewicz, Van Bellegem and von Sachs (2003), we have $|A^{-1}_{i,j}| \leq C 2^{i/2} 2^{j/2}$, where C is a generic constant which will also appear later and which is not necessarily the same at each occurrence. Further, by Assumption 2.2, we have $S_j(z) \leq C2^j$ and, by Proposition 2 from Fryzlewicz and Nason (2006), we have $\beta^{(1)}_i(k/T) \leq C$ under Assumptions 2.1 and 2.2. Given the above, we have

$$II/8 \leq C \left(\sum_{j=-1}^{-J_0} \sum_{k=1}^{T} \sum_{i=-J^*-1}^{-\infty} 2^{i/2} 2^{j/2} 2^j \right)^2 \leq CT^2 2^{-J^*}. \tag{5.2}$$

We now turn to I. Denote

$$b_{i,j} = \sum_{k=1}^{T} \{I^{(i)}_{k,T} - \beta^{(1)}_i(k/T)\}\{S^{(1)}_j(k/T) - S^{(2)}_j(k/T)\}.$$

With this notation, and using the Cauchy inequality, we have

$$I/8 \leq E\left\{ \sum_{j=-1}^{-J_0} \sum_{i=-1}^{-J^*} A^{-1}_{i,j} b_{i,j} \right\}^2 \leq J_0 J^* \sum_{j=-1}^{-J_0} \sum_{i=-1}^{-J^*} \{A^{-1}_{i,j}\}^2 E\{b^2_{i,j}\} \leq C J_0 J^* \sum_{j=-1}^{-J_0} \sum_{i=-1}^{-J^*} 2^{i+j} E\{b^2_{i,j}\}. \tag{5.3}$$

We now evaluate $E\{b^2_{i,j}\}$.

$$\begin{aligned} E\{b^2_{i,j}\} &= E\left\{ \sum_{k=1}^{T} \{I^{(i)}_{k,T} - E(I^{(i)}_{k,T}) + E(I^{(i)}_{k,T}) - \beta^{(1)}_i(k/T)\}\{S^{(1)}_j(k/T) - S^{(2)}_j(k/T)\} \right\}^2 \\ &\leq 2E\left\{ \sum_{k=1}^{T} \{I^{(i)}_{k,T} - E(I^{(i)}_{k,T})\}\{S^{(1)}_j(k/T) - S^{(2)}_j(k/T)\} \right\}^2 \\ &+ 2\left\{ \sum_{k=1}^{T} \{E(I^{(i)}_{k,T}) - \beta^{(1)}_i(k/T)\}\{S^{(1)}_j(k/T) - S^{(2)}_j(k/T)\} \right\}^2 =: 2A + 2B. \end{aligned}$$

Using Cauchy inequality, we have

$$B \leq \sum_{k=1}^{T} \{E(I^{(i)}_{k,T}) - \beta^{(1)}_i(k/T)\}^2 \sum_{k=1}^{T} \{S^{(1)}_j(k/T) - S^{(2)}_j(k/T)\}^2. \tag{5.4}$$

Using the result of Proposition 1 from Fryzlewicz and Nason (2006) , we have

$$\sum_{k=1}^{T} \{E(I^{(i)}_{k,T}) - \beta^{(1)}_i(k/T)\}^2 \leq C(2^{-i} + T^{1/\{2\log_2(a)-1\}}).$$

Substituting it back into (5.4), we have $B \leq C(T2^{2j-i} + T^{1+1/\{2\log_2(a)-1\}}2^{2j})$. Now turn to A. Denote $c_{j,k} = \{S_j^{(1)}(k/T) - S_j^{(2)}(k/T)\}$. As before, $|c_{j,k}| \leq C2^j$. We have

$$A = E\left\{\sum_{k=1}^{T}\{I_{k,T}^{(i)} - E(I_{k,T}^{(i)})\}c_{j,k}\right\}^2 \leq 2^{2j}\sum_{k,k'=1}^{T}|\text{cov}(I_{k,T}^{(i)}, I_{k',T}^{(i)})|.$$

Let $d_{k,T}^{(i)}$ be the wavelet coefficient corresponding to $I_{k,T}^{(i)}$, i.e. $I_{k,T}^{(i)} = (d_{k,T}^{(i)})^2$. By Isserlis' theorem, we have

$$\begin{aligned} |\text{cov}(I_{k,T}^{(i)}, I_{k',T}^{(i)})| &= 2\text{cov}^2(d_{k,T}^{(i)}, d_{k',T}^{(i)}) \leq 2\sup_k \text{Var}^2(d_{k,T}^{(i)})\text{corr}^2(d_{k,T}^{(i)}, d_{k',T}^{(i)}) \\ &\leq 2\{E(I_{k,T}^{(i)})\}^2|\text{corr}(d_{k,T}^{(i)}, d_{k',T}^{(i)})|. \end{aligned} \tag{5.5}$$

Using again Proposition 1 from Fryzlewicz and Nason (2006) , we have

$$E(I_{k,T}^{(i)}) \leq C\sum_{j=-1}^{-\infty}2^j\sum_{t=-\infty}^{\infty}\Psi_{j,i}^2(t-k) = C\sum_{j=-1}^{-\infty}2^jA_{j,i} = C.$$

Substituting it back into (5.5), we get $|\text{cov}(I_{k,T}^{(i)}, I_{k',T}^{(i)})| \leq C|\text{corr}(d_{k,T}^{(i)}, d_{k',T}^{(i)})|$, so that the bound for A is

$$\begin{aligned} A &\leq C2^{2j}\sum_{k,k'=1}^{T}|\text{corr}(d_{k,T}^{(i)}, d_{k',T}^{(i)})| \\ &\leq C2^{2j}\sum_{k=1}^{T}\sum_{\tau}\left|\text{corr}\left(\sum_t X_t\psi_{i,k-t}, \sum_s X_s\psi_{i,k+\tau-s}\right)\right| \\ &\leq C2^{2j}\sum_t\sum_s|\text{corr}(X_t, X_s)|\sum_{\tau=s-t-M2^{-i}}^{s-t+M2^{-i}}\left\{\sum_k|\psi_{i,k-t}||\psi_{i,k-s+\tau}|\right\} \\ &\leq C2^{2j-i}\sum_t\sum_\rho|\text{corr}(X_t, X_{t+\rho})| \leq C2^{2j-i}T, \end{aligned}$$

where the last inequality is implied by the short-memory assumption. Substituting A and B back into (5.3), the bound for I becomes

$$\begin{aligned} I &\leq CJ_0J^*\sum_{j=-1}^{-J_0}\sum_{i=-i}^{-J^*}2^{i+j}\{2^{2j-i}T + T^{1+1/\{2\log_2(a)-1\}}2^{2j}\} \\ &\leq C\log_2^2 T\sum_{j=-1}^{-J_0}\sum_{i=-i}^{-J^*}2^{3j}T + 2^{i+3j}T^{1+1/\{2\log_2(a)-1\}} \\ &\leq C\log_2^2 T\{T\log_2 T + T^{1+1/\{2\log_2(a)-1\}}\}. \end{aligned}$$

With this result, the overall bound for $E(X^2)$ is

$$E(X^2) \leq C\{T^2 2^{-J^*} + T\log_2^3 T + T^{1+1/\{2\log_2(a)-1\}} \log_2^2 T\}.$$

Choosing $J^* = \log_2 T$ makes the above bound $E(X^2) \leq C\{T\log_2^3 T + T^{1+\varepsilon} \log_2^2 T\}$. By the assumption on $\mathcal{M}$, t^2 is of order T^2. Considering again (5.1), we can see that the probability of misclassification is of order

$$P(X > t) = O(T^{-1}\log_2^3 T + T^{1/\{2\log_2(a)-1\}-1} \log_2^2 T),$$

which completes the proof. $\square$

References

Biau, G., Bunea, F. and Wegkamp, M. (2005). Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory*, **51**, 2163–2172.

Blandford, R. (1993). Discrimination of Earthquakes and Explosions. *AFTAC-TR-93-044 HQ, Air Force Technical Applications Center, Patrick Air Force Base, FL.*

Bunea, F., Wegkamp, M. and Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. *Journal of Statistical Planning and Inference*, **136**, 4394-4364.

Bunea, F. (2007). Consistent selection via the Lasso for high dimensional approximating regression models. In *IMS Lecture Notes Monograph Series*, in press.

Chandler, G. and Polonik, W. (2006). Discrimination of locally stationary time series based on the excess mass functional. *Journal of the American Statistical Association*, **101**, 240-253.

Dahlhaus, R. (1996). On the Kullback-Leibler information divergence of locally stationary processes. *Stochastic Processes and Applications*, **62**, 139-168.

Daubechies, I. (1992). *Ten Lectures on Wavelets.* Philadelphia, PA: SIAM.

Fryzlewicz, P. (2007). Unbalanced Haar technique for nonparametric function estimation. *Journal of the American Statistical Association*, in press.

Fryzlewicz. P. and Nason, G. (2006). Haar-Fisz estimation of evolutionary wavelet spectra. *Journal of the Royal Statistical Society, Series B*, **68**, 611-634.

Fryzlewicz, P., Van Bellegem, S. and von Sachs, R. (2003). Forecasting non-stationary time series by wavelet process modelling. *Annals of the Institute of Statistical Mathematics*, **55**, 737-764.

Genovese, C., Roeder, K. and Wasserman, L. (2007). False discovery control with p-value weighting. *Biometrika*, in press.

Huynh, Q., Cooper, L., Intrator, N., Shouval, H. (1998). Classification of underwater mammals using feature extraction basedon time-frequency analysis and BCM theory. *IEEE Transactions on Signal Processing*, **46**, 1202-1207.

Huang, H-S., Ombao, H. and Stoffer, D. (2004). Discrimination and classification of non-stationary time series using the SLEX model. *Journal of the American Statistical Association*, **99**, 763-774.

Jin, J. and Cai, T. (2006). Estimating the null and the proportion non-null effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, **102**, 495-506.

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

Mallat, S. (1998). *A Wavelet Tour of Signal Processing.* New York: Academic Pres.

Mallat, S., Papanicolaou, G. and Zhang, Z. (1998). Adaptive covariance estimation of locally stationary processes. *Annals of Statistics*, **26**, 1-47.

Mesgarani, N., Slaney, M. and Shamma, S. (2006). Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Transactions on Audio, Speech and Language Processing*, **14**, 920-930.

Meyer, F. and Chinrungreung, J. (2003). Analysis of Event-Related fMRI Data using Local Clustering Bases. *IEEE Transactions on Medical Imaging*, **22**, 933-939.

Nason, G. and Silverman, B. (1995). The stationary wavelet transform and some statistical applications. In *Lecture Notes in Statistics, vol. 103, Eds. Antoniadis and Oppenheim*, 281-300. New York: Springer-Verlag.

Nason, G. and von Sachs, R. (1999). Wavelets in time series analysis. *Philosophical Transactions of the Royal Society of London, Series A*, **357**, 2511-2526.

Ombao, H., Raz, J., von Sachs, R. and Malow, B. (2001). Automatic statistical analysis of bivariate nonstationary time series. *Journal of the American Statistical Association*, **96**, 543–560.

Ombao, H., Raz, J., von Sachs, R. and Guo, W. (2002). The SLEX Model of a Non-Stationary Random Process. *Annals of the Institute of Statistical Mathematics*, **54**, 171-200.

Pfurtscheller, G., Brunner, C., Schlögl, A. and Lopes da Silva, F. (2006). Mu rhythm de-synchronization and EEG single-trial classification of different motor imagery tasks. *Neuroimage*, **31**, 153-159.

Priestley, M. (1965). Evolutionary spectra and non-stationary processes. *Journal of the Royal Statistical Society, Series B*, **27**, 204-237.

Sakiyama, K. and Taniguchi, M. (2004). Discriminant analysis for locally stationary processes. *J. Multivariate Analysis*, **90**, 282-300.

Shumway, R. (2003). Time-frequency clustering and discriminant analysis. *Statistics and Probability Letters*, **63**, 307-314.

Shumway, R. and Stoffer, D. (2006). *Time Series Analysis and Its Applications With R Examples.* New York: Springer.

Sun, W. and Cai, T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, **102**, 901-912.

Van Bellegem, S. and von Sachs, R. (2007). Locally adaptive estimation of evolutionary wavelet spectra. *Annals of Statistics*, in press.

Vannucci. M., Sha, N. and Brown, P. (2005). NIR and mass spectra classification: Bayesian methods for wavelet-based feature selection. *Chemometrics and Intelligent Laboratory Systems*, **77**, 139-148.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets.* New York: Wiley.

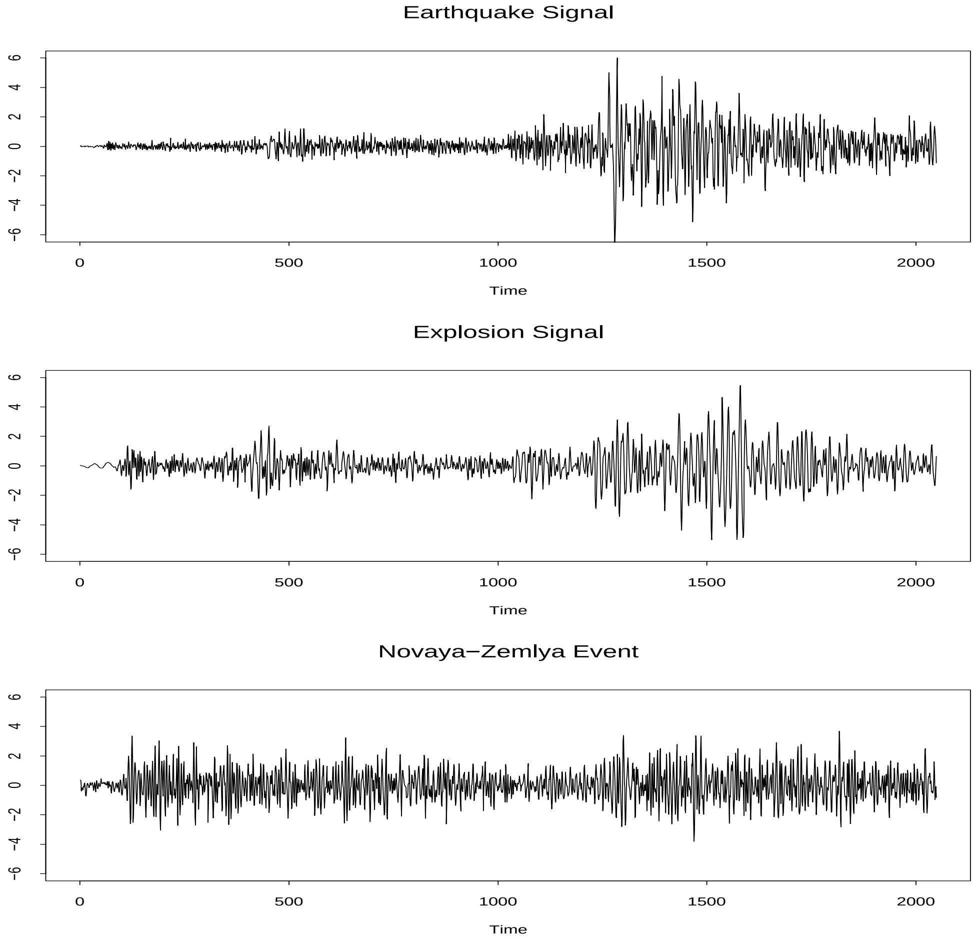
von Sachs, R. and MacGibbon, B. (2000). Non-parametric curve estimation by wavelet thresholding with locally stationary errors. *Scandinavian Journal of Statististics*, **27**, 475-499.

Wasserman, L. and Roeder, K. (2006). Weighted Hypothesis Testing. *Department of Statistics Technical Report, Carnegie Mellon University.*

Wasserman, L. and Roeder, K. (2007). High dimensional variable selection. *Department of Statistics Technical Report, Carnegie Mellon University.*

Wood, J., O'Connell-Rodwell, C. and Klemperer, S. (2005). Using seismic sensors to detect elephants and other large mammals: a potential census technique. *Journal of Applied Ecology*, **42**, 587-594.

FIGURES



Earthquake Signal

Explosion Signal

Novaya–Zemlya Event

Figure 1: Seismic Signals. $(1, 1)$: Earthquake Signal (no. 1, out of 8). $(2, 1)$: Explosion Signal (no. 1, out of 8). $(3, 1)$: Event at Novaya Zemlya (unknown origin). Source of the data: the file `eq+exp.dat` from Shumway and Stoffer (2006) available from `http://lib.stat.cmu.edu/general/tsa2/`.

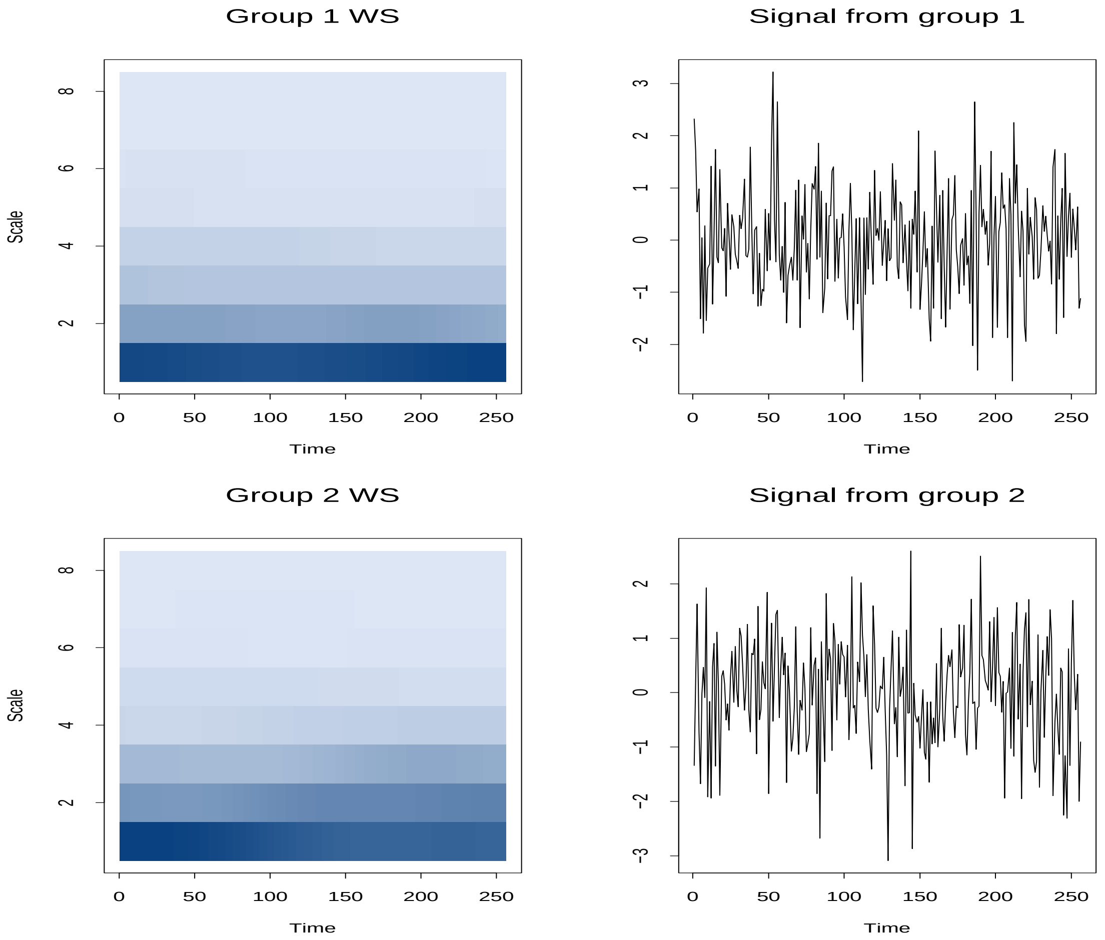


Figure 2: Piecewise-stationary AR(1) processes. $(1,1)$: Estimated wavelet spectrum of group 1 (vertical scale denotes the scale of the wavelet transform; the darker the color the larger the amplitude of the corresponding spectrum estimate). $(1,2)$: A time series from group 1. $(2,1)$: Estimated wavelet spectrum of group 2. $(1,2)$: A time series from group 2. Length of the time series is $T = 256$. Maximal scale is $J = \log_2(256) = 8$. Total number of signals in each group is $N = 8$. In wavelet spectrum estimation, kernel smoothing over time was used with the Gaussian kernel and bandwidth $= 100$.

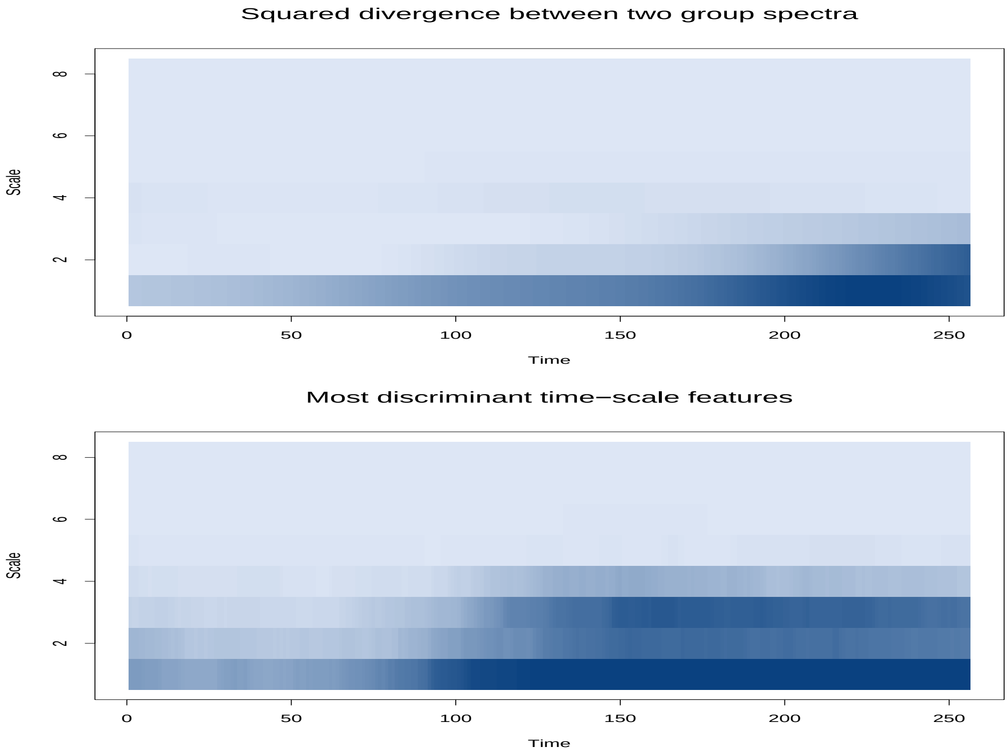


Figure 3: Piecewise-stationary AR(1) processes. $(1,1)$: Squared difference between wavelet spectra of the two groups (the darker the color, the greater the magnitude of the difference). $(2,1)$: Discriminant time-scale features selected (the darker the color, the larger the proportion of times the given feature was selected; $p = 0.25$ was used). Length of the time series is $T = 256$. Maximal scale is $J = \log_2(256) = 8$. Total number of signals in each group is $N = 8$. In wavelet spectrum estimation, kernel smoothing over time was used with the Gaussian kernel and bandwidth $= 100$.

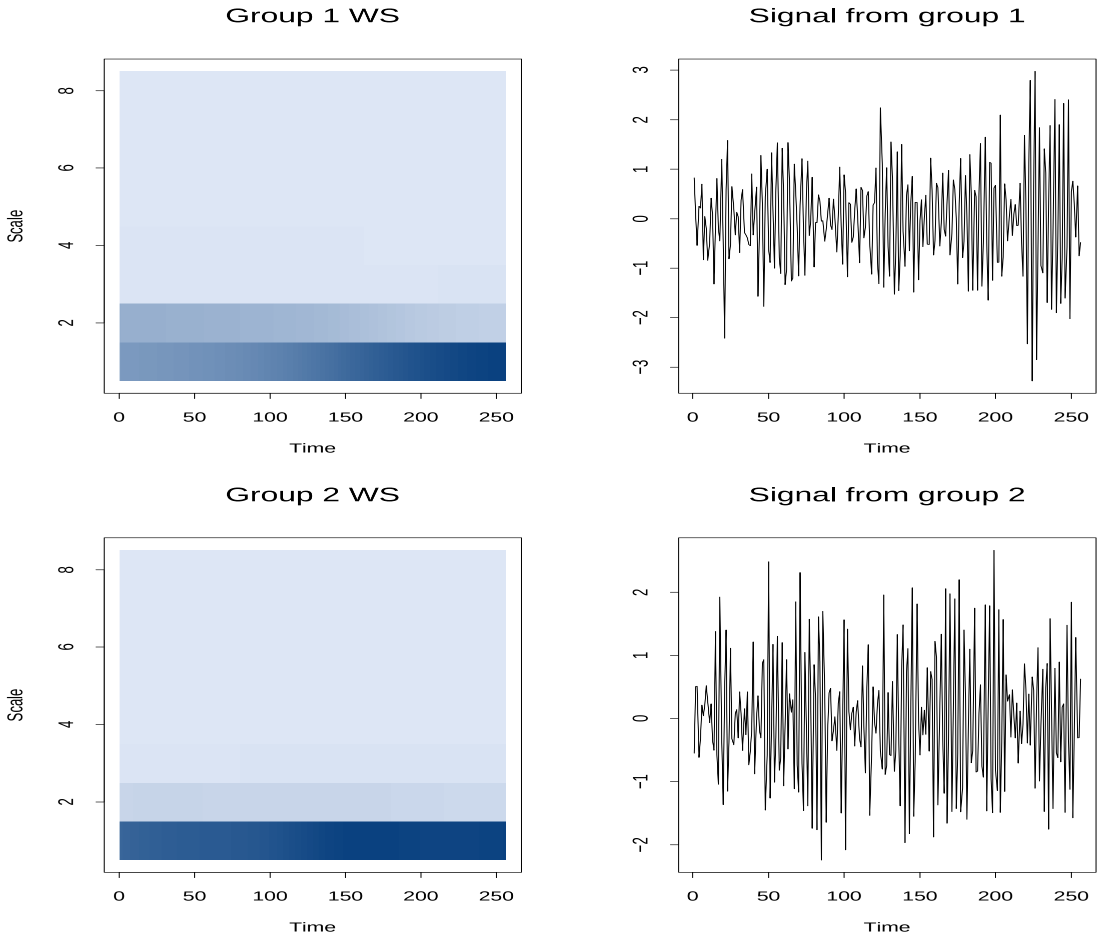


Figure 4: AR(2) processes with slowly time-varying coefficients. $(1,1)$: Estimated wavelet spectrum of group 1 (vertical scale denotes the scale of the wavelet transform; the darker the color the larger the amplitude of the corresponding spectrum estimate). $(1,2)$: A time series from group 1. $(2,1)$: Estimated wavelet spectrum of group 2. $(1,2)$: A time series from group 2. Length of the time series is $T = 256$. Maximal scale is $J = \log_2(256) = 8$. Total number of signals in each group is $N = 8$. In wavelet spectrum estimation, kernel smoothing over time was used with the Gaussian kernel and bandwidth = 100.

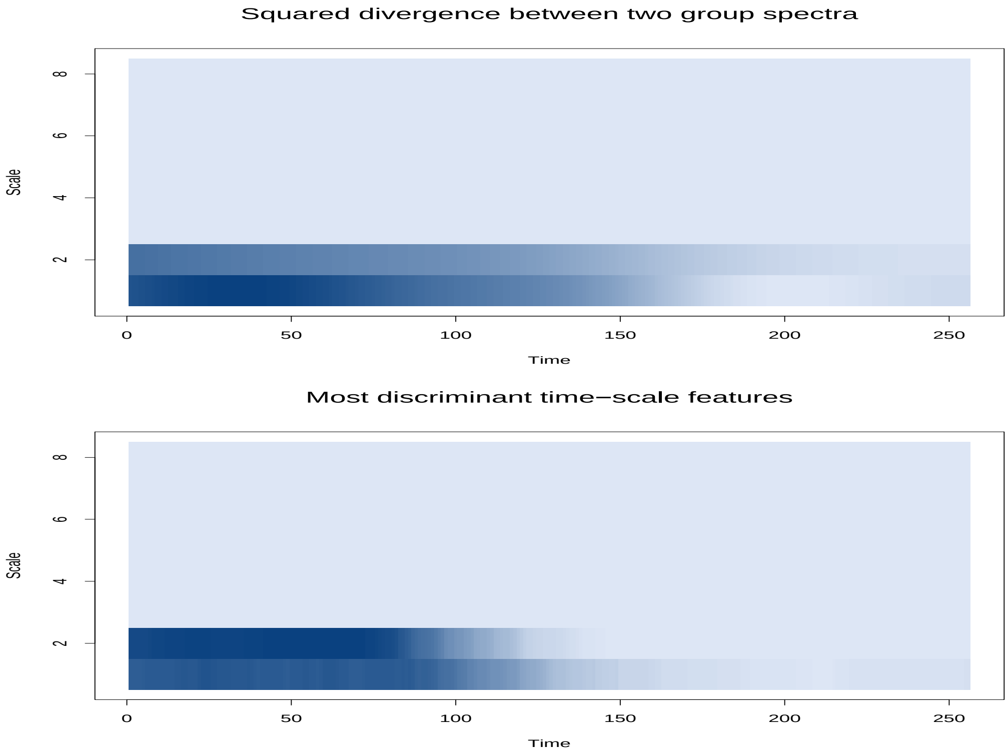


Figure 5: AR(2) processes with slowly time-varying coefficients. $(1, 1)$: Squared difference between wavelet spectra of the two groups (the darker the color, the greater the magnitude of the difference). $(2, 1)$: Discriminant time-scale features selected (the darker the color, the larger the proportion of times the given feature was selected; $p = 0.25$ was used). Length of the time series is $T = 256$. Maximal scale is $J = \log_2(256) = 8$. Total number of signals in each group is $N = 8$. In wavelet spectrum estimation, kernel smoothing over time was used with the Gaussian kernel and bandwidth $= 100$.

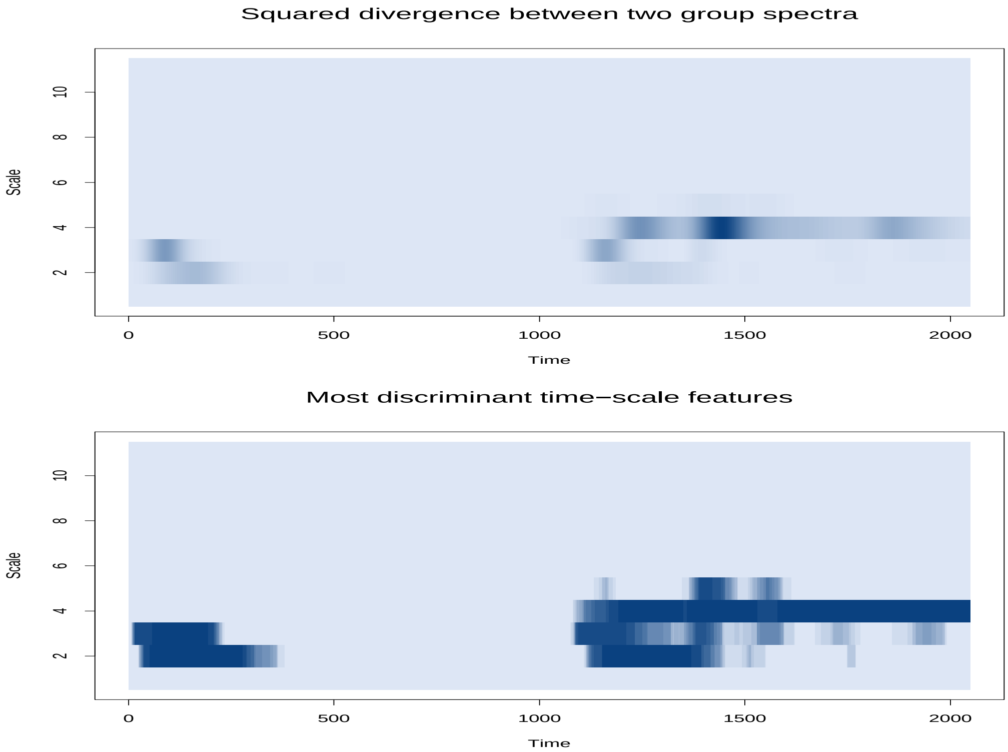


Figure 6: Earthquake vs Explosion Signals. $(1,1)$: Squared difference between wavelet spectra of the two groups (the darker the color, the greater the magnitude of the difference). $(2,1)$: Discriminant time-scale features selected (the darker the color, the larger the proportion of times the given feature was selected; $p = 0.25$ was used). Length of the time series is $T = 2048$. Maximal scale is $J = \log_2(2048) = 11$. Total number of signals in each group is $N = 8$. In wavelet spectrum estimation, kernel smoothing over time was used with the Gaussian kernel and bandwidth $= 100$.