INSTITUT DE STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION PAPER

0208

FORECASTING NON-STATIONARY TIME SERIES BY WAVELET PROCESS MODELLING

P. FRYŹLEWICZ, S. VAN BELLEGEM and R. VON SACHS

http://www.stat.ucl.ac.be

Forecasting non-stationary time series by wavelet process modelling

P. Fryźlewicz[1] S. Van Bellegem[2, 4, *] R. von Sachs[3, 4]

February 25, 2002

Abstract

Many time series in the applied sciences display a time-varying second order structure. In this article, we address the problem of how to forecast these non-stationary time series by means of non-decimated wavelets. We first consider a model in which only the variance evolves with time. We define a predictor for this model and show an application to the Dow Jones index. Then, we generalise the definition to the case of time-varying covariance using the class of Locally Stationary Wavelet processes. We introduce a new predictor based on wavelets and derive the prediction equations as a generalisation of the Yule-Walker equations. We propose an automatic computational procedure for choosing the parameters of the forecasting algorithm. Finally, we apply the general prediction algorithm to a meteorological time series.

Keywords: Local stationarity, non-decimated wavelets, prediction, time-modulated processes, volatility, Yule-Walker equations.

Running head: Forecasting non-stationary processes by wavelets

---

[1]University of Bristol, Department of Mathematics, Bristol, UK. E-mail: P.Z.Fryzlewicz@bristol.ac.uk

[2]Research Fellow of the National Fund for Scientific Research (F.N.R.S.). Université catholique de Louvain, Institut de statistique, Louvain-la-Neuve, Belgium. E-mail: vanbellegem@stat.ucl.ac.be

[3]Université catholique de Louvain, Institut de statistique, Louvain-la-Neuve, Belgium. E-mail: rvs@stat.ucl.ac.be

[4]Financial support from the contract 'Projet d'Actions de Recherche Concertées' nr. 98/03-217 of the Belgian Government is gratefully acknowledged.

* *Corresponding author. Address for correspondance: Université catholique de Louvain, Institut de statistique, Voie du Roman Pays, 20, B-1348 Louvain-la-Neuve. Fax: +32 10 47.30.32*

1 Introduction

In a growing number of fields, such as biomedical time series analysis, geophysics, telecommunications, or financial data analysis, to name but a few, explaining and inferring from observed serially correlated data calls for non-stationary models of their second order structure. That is, variance and covariance, or equivalently the spectral structure, are likely to change over time.

In this article, we address the problem of whether and how wavelet methods can help in forecasting non-stationary time series. The use of wavelets has proved successful in capturing *local* features of observed data. Recently, Antoniadis and Sapatinas (2001) used wavelets for forecasting time-continuous stationary processes. There arises a natural question of whether they can also be useful for prediction in situations where too little homogeneous structure at the end of the observed data set prevents the use of classical prediction methods based on stationarity. Obviously, in order to develop a meaningful approach, one needs to control this deviation from stationarity, and hence one first needs to think about what kind of non-stationary *models* to fit to the observed data. Let us give a brief overview of the existing possibilities.

Certainly the simplest approach consists in assuming piecewise stationarity, or approximate piecewise stationarity, where the challenge is to find the stretches of homogeneity optimally in an adaptive, i.e. data-driven way (Ombao *et al.*, 2001). The resulting estimate of the time-varying second order structure is, necessarily, rather blocky over time, so some further thoughts on how to cope with these potentially artificially introduced discontinuities are needed. To name a few out of the many models which allow a smoother change over time, we cite the following approaches to the idea of "local stationarity": the work of Mallat *et al.* (1998), who impose bounds on the derivative of the Fourier spectrum as a function of time; the approaches which allow the coefficients of a parametric model (such as AR) to vary slowly with time (e.g. Mélard and Herteleer-De Schutter (1989), Dahlhaus *et al.* (1999) or Grillenzoni (2000)); and, more generally and in a more non-parametric fashion, the approaches which assume a *slow* change over time of the generalised Cramér spectral representation of the process (Priestley (1965), Dahlhaus (1997), Ombao *et al.* (2002)). All the above models are of the "time-frequency" type as they use, directly or indirectly, the concept of a time-varying spectrum, being the Fourier transform of a time-varying autocovariance.

The work of Nason *et al.* (2000) adopts the concept of local stationarity but replaces the aforementioned spectral representation with respect to the Fourier basis by a representation with respect to non-decimated (or translation-invariant) wavelets. With their model of "Locally Stationary Wavelet" (LSW) processes, the authors introduce a time-scale representation of a stochastic process. The representation allows for a rigorous theory of how to estimate the *wavelet spectrum*, i.e. the coefficients of the resulting representation of the local autocovariance function with respect to *autocorrelation wavelets*. This theory parallels the one developed by Dahlhaus (1996a), where rescaling the time argument of the autocovariance and the Fourier spectrum makes it possible to embed the estimation in the non-parametric regression setting, including asymptotic considerations of consistency and inference.

Therefore, it seems appropriate to us to use the (linear) LSW model to generalise the stationary approach of forecasting X_t by means of a predictor based on the previous observations up to time $t-1$. While the classical linear predictor can be viewed as based on a non-local Fourier-type representation, our generalisation uses a local wavelet-based approach. The use of the LSW model allows us to use the observed data $X_{t-1}, X_{t-2}, \ldots$ to estimate the evolving covariance structure locally, and to construct a predictor of X_t. The particular way in which the estimator is constructed (by means of the *wavelet periodogram*, mimicking the given structure of the LSW model) proves useful in a variety of situations: for instance, when dealing with LSW processes which have a sparse time-scale representation, or so-called *time-modulated* processes.

Time-modulated processes are stationary processes modulated by a time-varying variance function. As we argue later in the article, this particularly simple class of locally stationary processes seems to be extremely useful in modelling and forecasting certain financial time series, such as the log-returns of currency exchange rates or stock indices. Mikosch and Stărică (1999) have already indicated that some of the stylized facts concerning the covariance structure of these series can be explained by fitting a non-stationary model (see also Härdle *et al.* (2000) in the particular case of the time-inhomogeneous stochastic volatility model). Indeed, one of the secondary aims of the article is to compare the prediction ability of this non-stationary but linear model with that of the classical (stationary but non-linear) GARCH model (for an overview of GARCH modelling, see, for example, Bera and Higgins (1993)).

In Section 2, we present the main ideas behind our approach using the special case of time-modulated processes, concentrating on the case where only variance changes over time. Section 3 familiarises the reader with the general LSW model, and shows some interesting examples related to the particular subclass of time-modulated processes of the previous section. In the central Section 4, we treat the general prediction case, where the construction of our linear predictor is motivated by the approach in the stationary case, i.e. the objective is to minimise the mean-square prediction error (MSPE). This leads to a generalisation of the Yule-Walker equations. These can be solved numerically by matrix inversion or standard iterative algorithms such as the innovations algorithm (Brockwell and Davis, 1991), provided that the non-stationary covariance structure is known. However, estimation of a non-stationary covariance structure is the main challenge in this context. In this and the following sections, we address this problem by defining and estimating the localised autocovariance function in an appropriate way. In the remainder of Section 4, we derive an analogue of the classical Kolmogorov formula for the theoretical prediction error, and we generalise the one-step-ahead to h-step-ahead prediction.

Section 5 deals with estimation of the time-varying covariance structure. We discuss some asymptotic properties of our estimators based on the properties of the wavelet periodogram, which is an asymptotically unbiased, but not consistent, estimator of the wavelet spectrum. In the second part of Section 5, we propose a new algorithm for forecasting non-stationary time series, which implements the idea of adaptive forecasting in the LSW model. The principle of "learning from the past", exploited by our algorithm, is also known in the Bayesian forecasting framework (West and Harrison, 1997). In Section 6 we apply our general algorithm to a meteorological time series.

We close with a conclusions section, where we also briefly discuss another promising aspect of our prediction approach: studying the prediction error as a useful tool for model selection. This includes the question of how to choose the wavelet system in the underlying LSW model, a question that was left open in Nason *et al.* (2000). We believe that the use of the MSPE is an appropriate way of solving this problem, and in this respect, our paper closes an important gap in the modelling approach of Nason *et al.* (2000).

Finally, we present our proofs in two appendices. Appendix A contains all the results related to approximating the finite-sample covariance structure of the non-stationary time series by the locally stationary limit. In Appendix B, we show some relevant basic properties of the system of autocorrelation wavelets, and provide the remaining proofs of the statements made in Section 4 and 5.

2 Predicting time-modulated stationary processes

We will motivate our general construction of a predictor of locally stationary time series by the treatment of a particular subclass, the so-called time-modulated stationary processes, for two reasons. On the one hand, the constructed predictor is particularly intuitive, conceptually simple and works well in practice. On the other hand, time-modulated processes constitute an interesting and useful class of models for various applications, in particular in modelling financial data, such as log-returns of currency exchange rates and stock market indices.

2.1 Time-modulated processes as an example of locally stationary processes

A straightforward generalisation from the class of (zero mean) stationary processes is to modulate a given second-order stationary process in discrete time $\{Y_t\}_{t\in\mathbb{Z}}$ by a deterministic time-varying variance σ_t^2. The following question arises: under what conditions can the second order structure of the new process $X_t = \sigma_t\ Y_t,\ t = 0, \dots, T-1,$ be estimated by some non-parametric method? This obviously reduces to the problem of how to model and estimate the variation σ_t^2: if this is a sequence which varies with time in an arbitrary way, there is certainly no hope of estimating it. Hence, some conditions which control the evolution of this sequence over time, are needed. A convenient possibility is to assume a "slow" variation over time, i.e. $\sigma_t^2 \approx \sigma_{t+1}^2$, meaning that one can estimate these quantities by any common non-parametric variance estimator. A formal approach to embed this idea into a rigorous estimation theory is the one of "local stationarity" defined by Dahlhaus (1996a, 1996b, 1997), who assumes the existence of a smooth function defined on the unit time interval, $\sigma_0(z),\ z \in [0, 1]$, such that $\sigma_t \approx \sigma_0(t/T)$, where T again denotes the sample size of the observed time series X_t. We observe that this formalism allows us to embed the estimation of this time-varying variance in the classical setting of non-parametric curve estimation, provided that we impose some regularity assumption on the "true unknown" curve $\sigma_0(z)$, e.g. Lipschitz continuity or differentiability in z. The statistical meaning of this "rescaling in time" is as follows: increasing the sample size T does not mean passing to the future observations of the time series X_t, but allows us to gather more local statistical information about the time-varying behaviour of the variance of this process. In fact, process $\{X_t\}$ becomes a doubly-indexed sequence of processes $\{X_{t,T}\}_{0\leqslant t\leqslant T-1, T\geqslant 1}$ where asymptotically, with $T \to \infty$, we can gather "perfect information" about the curve $\sigma_0(z)$.

Later, we will identify σ_t with $\sigma_0(t/T)$, and, for simplicity, the latter will be denoted by

$\sigma(t/T)$. The class of time-modulated processes is then defined as

$$X_{t,T} = \sigma\left(\frac{t}{T}\right) Y_t \tag{2.1}$$

where Y_t is a zero-mean stationary process, and the local variance function $\sigma^2(z)$ is defined on $(0, 1)$. A possible regularity condition on $\sigma^2(z)$ is, for example, Lipschitz continuity: $|\sigma^2(z)-\sigma^2(z_0)| \leqslant \kappa\cdot|z-z_0|$, where κ is a constant. The idea behind this regularity assumption is that the observed data have a slowly varying variance structure. The usefulness of this class of processes will be demonstrated on a real data set further below.

For the sake of simplicity, Y_t is restricted to have a zero mean, but it is possible to introduce a trend function which may then be estimated separately (see, e.g., von Sachs and MacGibbon (2000)).

2.2 Estimation and prediction

With the rescaled-time model (2.1), the estimation of the variance $\sigma^2(z)$, $z \in (0, 1)$, becomes the common statistical problem of non-parametric curve estimation. A naive non-parametric estimator of the local variance $\sigma^2(t/T)$ is given by the unbiased estimator $\tilde{\sigma}^2_{\mathrm{na}}(t/T) = X^2_{t,T}$. Obviously, this estimator suffers from large variability, and the challenge is to find a more stable alternative estimator. As a less variable estimator will introduce a bias, we will use an estimator which has a reduced variance but still has a "small" bias. This is the *multiscale variance estimator* which we will now define.

2.2.1 A multiscale estimator of the variance

Our estimator is constructed using a *non-decimated wavelet system* $\{\psi_{jk}, j = -1, -2, \ldots, k \in \mathbb{Z}\}$ (see Nason and Silverman (1995) for a detailed introduction). In contrast to Mallat's (1989) discrete wavelet transform, the discrete non-decimated wavelets at all scales can be shifted to any location defined by the finest resolution scale, determined by the observed data. As a consequence, this construction leads to an overcomplete system of vectors. For ease of presentation, recall the simplest discrete non-decimated system, called the *Haar system*, given by

$$\psi_{j0}(t) = 2^{j/2}\mathbb{I}_{\{0,1,\ldots,2^{-j-1}-1\}}(t) - 2^{j/2}\mathbb{I}_{\{2^{-j-1},\ldots,2^{-j}-1\}}(t) \qquad \text{for } j = -1, -2, \ldots \text{ and } t \in \mathbb{Z}\ ,$$

and $\psi_{jk}(t) = \psi_{j0}(t-k)$ for all $k \in \mathbb{Z}$, where $\mathbb{I}_{\mathcal{A}}(t)$ is 1 if $t \in \mathcal{A}$ and 0 otherwise.

We now construct the multiscale estimator as follows. Define the empirical wavelet coefficients by

$$d_{jt;T} = \sum_{s=0}^{T-1} X_{s,T}\psi_{jt}(s)\ , \qquad j = -1, -2, \ldots, -J = -\log_2 T \text{ and } k \in \mathbb{Z}\ ,$$

as the projection of the data on the non-decimated wavelets ψ_{jk}. Note that we adopt the convention that $j = 0$ is the finest possible resolution scale given by the data themselves, such that $j=-1$ is the first wavelet scale, $j=-2$ the next coarser one and so forth, ending with $j=-J=-\log_2 T$. The multiscale estimator is given by

$$\tilde{\sigma}^2_{\text{MS}}\left(\frac{t}{T}\right) = \sum_{j=-J}^{-1} 2^j |d_{jt;T}|^2. \tag{2.2}$$

This estimator is simply a weighted sum of difference-based estimators. As reviewed in Antoniadis and Lavergne (1995), the concept of difference-based estimators of σ^2 has a long history. Here, however, $d^2_{jt;T}$ can be viewed as the contribution to the variance of the process at time t/T and scale j. For this reason, the sequence $|d_{jt;T}|^2$ is called the *wavelet periodogram* of the process. By straightforward calculations, it can be verified that the multiscale estimator has a reduced asymptotic variance and a small bias which tends to zero (see Remark 3 in Section 5). Observe that this estimator is not consistent and needs to be smoothed over time $z=t/T$. A classical kernel estimator may be used at this stage to obtain a consistent estimator $\tilde{\sigma}^2(z)$ from the multiscale estimator (see, for instance, Simonoff (1996)).Other more elaborate methods may also be considered for this purpose, such as the adaptive procedure of Mercurio and Spokoiny (2000). In Section 5.2, we will describe a new automatic data-driven procedure for choosing the bandwidth of a kernel-based smoothing procedure and other parameters of the forecasting algorithm.

2.2.2 Prediction of time-modulated processes

Suppose that we have observed one realisation of a time-modulated process $X_{s,T}=\sigma(s/T)\,Y_s$ up to time $t-1$, and want to predict the value of $X_{t,T}$ by $\hat{X}_{t,T}$ ($t \leqslant T-1$). In this subsection, we propose a method for constructing the predicted value of $\hat{X}_{t,T}$ and the corresponding prediction interval.

Having obtained a consistent estimator $\tilde{\sigma}(z)$ of the local variance of $\{X_{s,T}\}_{0\leqslant s\leqslant t-1}$, we construct the standardised process $\tilde{Y}_{s,T}=X_{s,T}/\tilde{\sigma}(s/T)$. Under the model of time-modulated stationarity, $\{\tilde{Y}_{s,T}\}_{0\leqslant s\leqslant t-1}$ can be treated as a *stationary* process for which we can estimate the autocovariance function $\tilde{\rho}_Y(\tau)$ and apply a classical prediction method to forecast $\hat{Y}_{t,T}$. To compute the predicted value $\hat{X}_{t,T}$ we can use relation (2.1) in which we now have to predict $\sigma(t/T)$ by $\hat{\sigma}(t/T)$. Due to the slow evolution of the variance, $\hat{\sigma}(t/T)$ is constructed by the local left-sided average of length μ of the (unsmoothed) multiscale estimators

$$\tilde{\sigma}_{\text{MS}}\left(\frac{t-\mu}{T}\right), \ldots, \tilde{\sigma}_{\text{MS}}\left(\frac{t-1}{T}\right).$$

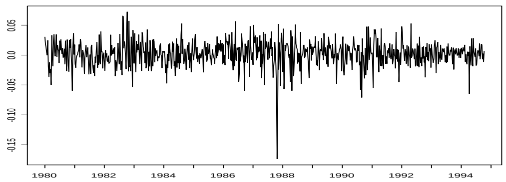
Then we define $\hat{X}_{t,T}=\hat{\sigma}(t/T)\hat{Y}_{t,T}$ and the prediction interval for $\hat{X}_{t,T}$ is obtained by multiplying the classical prediction interval of $\hat{Y}_{t,T}$ by $\hat{\sigma}(t/T)$.

This one-step-ahead procedure can be used as the initial step of a generalisation for the h-step-ahead prediction. Using a classical predictor for $\hat{Y}_{t-1+h}$ from $\{\tilde{Y}_{s,T}\}_{0\leqslant s\leqslant t-1}$, we construct $\hat{\sigma}((t-1+h)/T)$ by the local left-sided average of length μ of

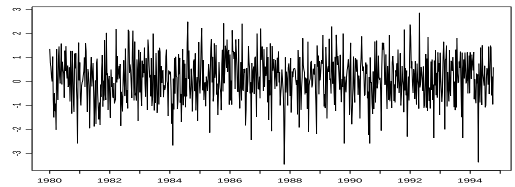
$$\tilde{\sigma}_{\text{MS}}\left(\frac{t-1+h-\mu}{T}\right), \ldots, \tilde{\sigma}_{\text{MS}}\left(\frac{t-1}{T}\right), \hat{\sigma}\left(\frac{t}{T}\right), \ldots, \hat{\sigma}\left(\frac{t-2+h}{T}\right)$$

for $h>1$.

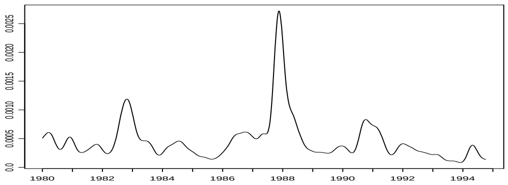
2.3 Real-data example: the Dow Jones Industrial Average index



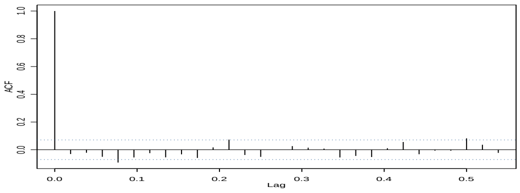
(a) Log-returns of the weekly DJIA index (769 observations).



(b) Standardised data.



(c) Estimated local variance function.



(d) Acf of the squared standardised data.

Figure 1: This example provides evidence that the log-returns of the DJIA index may be modelled as a time-modulated white noise process. The estimation of the local variance uses the multiscale estimator with Haar wavelets, and the Gaussian kernel with bandwidth 0.03.

The weekly returns of the DJIA index from week 1 in 1980 to week 39 in 1994 are analysed by Franses (1998), who fits conditionally heteroscedastic models to the data. Figure 1(a) shows the log-returns of DJIA: $X_t = \log D_t - \log D_{t-1}$, where D_t is the index. A preliminary analysis of the autocorrelation function of X_t suggests that it might be reasonable to attempt to model X_t as a time-modulated white noise process. Figure 1(c) shows an estimate of the local variance function, obtained by means of the multiscale estimator with Haar wavelets, and the Gaussian kernel with bandwidth 0.03. Indeed, after dividing X_t by the square root of the estimated variance function, the non-negligible autocorrelation in the squares, present in the original series X_t, now disappears (see Figure 1(d)). Furthermore, the examination of the cumulant-based kurtosis $\gamma = (\kappa_4/\kappa_2^2) = -0.27$ of the standardised data suggests using normal distribution for the standardised series. Nevertheless, the original data X_t show a large non-normal kurtosis of $\gamma = 6.50$, which is a commonly observed phenomenon in financial log-return series.

Having decided upon the model of time-modulated Gaussian white noise, we compare the predictive power of our method with a predictor based on the GARCH(1,1) model. As the predicted value (i.e. the conditional mean) for both GARCH and the time-modulated white noise is equal to zero, we compare the two methods by looking at the length of the prediction intervals, as well as the proportion of observations that fall within them. The prediction intervals are constructed as in Subsection 2.2.2 with $\mu = 20$, and are based on normality.

Figure 2 shows the one-step-ahead prediction intervals for $t = 100, 101, \ldots, 769$. The proportion of observations which fall within the corresponding prediction intervals using

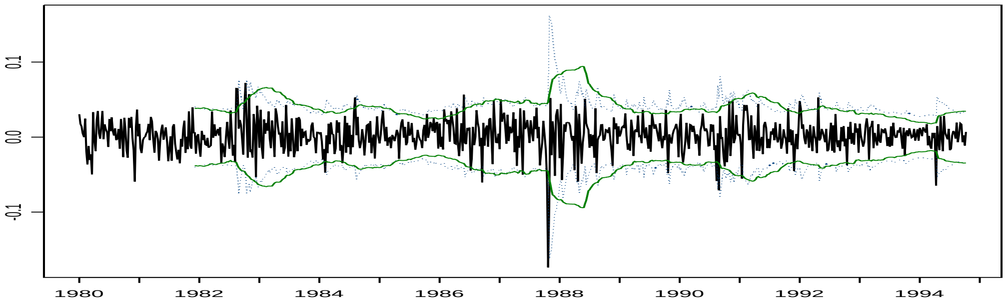


Figure 2: The actual series and prediction intervals for one-step-ahead prediction using GARCH(1,1) (dotted lines) and time-modulated white noise (solid lines).

the GARCH(1,1) model and the time-modulated white noise model, is 94.0% and 94.8%, respectively. The gain obtained using our method is not simply an effect of increasing the length of the prediction intervals. Indeed, in this example, 60.8% of the prediction intervals in the time-modulated white noise model are narrower than the corresponding prediction intervals in the GARCH(1,1) model.

It is interesting to observe that the length of the prediction intervals, as a function of time, evolves "more smoothly" in our model. Intuitively, this is in accord with the assumption of local stationarity. It is thanks to this particular property that the long period of increased volatility which follows the shock of the "Black Monday" week (October 1987), is handled better in our model, even though the shock itself is not captured well by either scheme.

3 Locally stationary wavelet processes

In the previous section, we demonstrated the usefulness of a new predictor for a certain class of locally stationary processes, the time-modulated stationary processes. The key to constructing an appropriate predictor was the construction of a multiscale estimator of the *time-varying variance* of the process, by means of the wavelet periodogram. In this section, we recall the definition of a locally stationary process with a more general *time-varying covariance* structure. We are looking for a class of processes which will allow us to use wavelet-based estimators similar to the one we did in the previous setting. It turns out that the following class of locally stationary wavelet (LSW) processes, introduced in Nason *et al.* (2000), is an appropriate candidate for our purpose. It is based upon combining two ingredients:

1. the idea of sparse modelling of autocovariance sequences which live on certain predominant scales by a superposition of particularly constructed *autocorrelation wavelets*;

2. the idea of local stationarity, expressed in a time-varying spectrum called the *wavelet spectrum*. This spectrum may be written as a decomposition of the autocovariance of the process with respect to the autocorrelation wavelet basis.

The formal definition of a locally stationary wavelet process is quoted from Nason, von Sachs and Kroisandt (2000).

Definition 1. *A sequence of doubly-indexed stochastic processes* $X_{t,T}$ $(t = 0, \dots, T-1)$ *with mean zero is in the class of* locally stationary wavelet processes (LSW processes) *if there exists a representation in the mean-square sense*

$$X_{t,T} = \sum_{j=-J}^{-1} \sum_{k=-\infty}^{\infty} w_{j,k;T}\ \psi_{jk}(t)\ \xi_{jk}, \tag{3.1}$$

where $T = 2^J$ *and* $\{\psi_{jk}(t)\}_{jk}$ *is a discrete non-decimated family of wavelets for* $j = -1, -2, \dots, -J = -\log_2 T$ *based on a mother wavelet* $\psi(t)$ *of compact support, and such that:*

1. ξ_{jk} *is a random orthonormal increment sequence with* $\mathrm{E}\xi_{jk} = 0$ *and* $\mathrm{Cov}\,(\xi_{jk}, \xi_{\ell m}) = \delta_{j\ell}\ \delta_{km}$ *for all* j, ℓ, k, m*; where* $\delta_{j\ell} = 1$ *if* $j = \ell$ *and 0 elsewhere;*
2. *For each* $j \leqslant -1$*, there exists a Lipschitz-continuous function* $W_j(z)$ *on* $(0,1)$ *possessing the following properties:*
   - $\sum_{j=-\infty}^{-1} |W_j(z)|^2 < \infty$ *uniformly in* $z \in (0,1)$ *;*
   - *the Lipschitz constants* L_j *are such that* $\sum_{j=-\infty}^{-1} 2^{-j} L_j < \infty$*;*
   - *there exists a sequence of constants* C_j *such that for each* T

     $$\sup_{k=1,\dots,T} \left| w_{j,k;T} - W_j\left(\frac{k}{T}\right) \right| \leqslant \frac{C_j}{T}\ ,$$

     where the constants C_j *are such that* $\sum_{j=-\infty}^{-1} C_j < \infty$.

LSW processes are not uniquely determined by the sequence $\{w_{jk;T}\}$. However, Nason *et al.* (2000) develop a theory which ensures the existence of a unique wavelet spectrum. This spectrum measures the power of the process at a particular scale and location. Formally, the *evolutionary wavelet spectrum* of an LSW process $\{X_{t,T}\}_{t=0,\dots,T-1}$, with respect to ψ, is given by

$$S_j(z) = |W_j(z)|^2\ ,\ z \in (0,1) \tag{3.2}$$

and is such that, by definition of the process, $S_j(z) = \lim_{T\to\infty} |w_{j,[zT];T}|^2$ for all $z \in (0,1)$, and thus $\sum_{j=-\infty}^{-1} S_j(z) < \infty$ uniformly in $z \in (0,1)$.

In the classical theory of stationary processes, the spectrum and autocovariance function are Fourier transforms of each other. To establish an analogous relationship for the wavelet spectrum, observe that the autocovariance function of an LSW process can be written as

$$c_T\,(z, \tau) = \mathrm{Cov}\left(X_{[zT],T}, X_{[zT]+\tau,T}\right)$$

for $z \in (0,1)$ and τ in $\mathbb{Z}$, and where $[\ \cdot\ ]$ denotes the integer part of a real number. The next result shows that this covariance tends to a local covariance as T tends to infinity. Let us introduce the *autocorrelation wavelets* as

$$\Psi_j(\tau) = \sum_{k=-\infty}^{\infty} \psi_{jk}(0)\ \psi_{jk}(\tau)\ , \qquad j < 0,\ \tau \in \mathbb{Z}.$$

Some useful properties of the system $\{\Psi_j\}_{j<0}$ can be found in Appendix B. By definition, the *local autocovariance function* of an LSW process with evolutionary spectrum (3.2) is given by

$$c(z, \tau) = \sum_{j=-\infty}^{-1} S_j(z) \Psi_j(\tau) \tag{3.3}$$

for all $\tau \in \mathbb{Z}$ and z in $(0, 1)$.

Proposition 1 (Nason *et al.* **(2000)).** *Under the assumptions of Definition 1, if $T \to \infty$, $|c_T(z, \tau) - c(z, \tau)| = O(T^{-1})$ uniformly in $\tau \in \mathbb{Z}$ and $z \in (0, 1)$.*

We observe that equation (3.3) is a decomposition of the autocovariance structure of the process over scales and time locations: the larger the wavelet spectrum $S_j(z)$ at a particular scale j and a point z in the rescaled time, the more dominant the contribution of scale j at time z.

It is worth mentioning that a stationary process with an absolutely summable autocovariance function is an LSW process (Nason *et al.*, 2000, Proposition 3). Stationarity is characterised by a wavelet spectrum which is constant over time: $S_j(z) = S_j$ for all $z \in (0, 1)$. Note also that the time-modulated processes defined in the previous section are LSW. The local variance of these processes is given by the multiscale decomposition

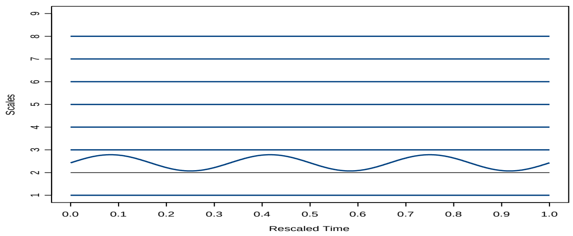
$$\sigma^2(z) = c(z, 0) = \sum_{j=-\infty}^{-1} S_j(z)$$

since $\Psi_j(0) = 1$ for all scales j.

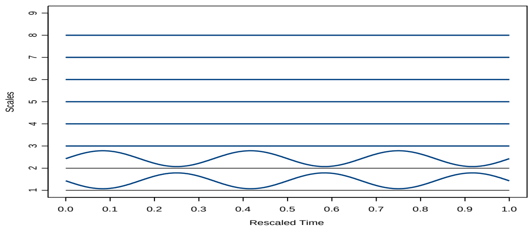
However, the real advantage of introducing general LSW processes lies in their ability to model time-varying covariance functions. Decomposition (3.3) is useful for processes with a sparse time-scale representation, which are "attractive" since only a few coefficients in their wavelet representation have to be estimated. Figure 3 shows simulated examples where the spectrum "lives" only at a limited number of scales. In particular, Figure 3(d) shows a realisation of a process which cannot be modelled as a time-modulated stationary series.

The possibility of having a sparse representation of the wavelet spectrum is a major advantage of LSW processes, in comparison with other locally stationary models. Note that if the local autocovariance function $c(z, \tau)$ is sparsely represented in one particular autocorrelation wavelet system $\{\Psi_j\}$, it is generally not the case for another basis $\{\tilde{\Psi}_j\}$. Then, there arises a question of how to choose the most appropriate autocorrelation wavelet system $\{\Psi_j\}$ for a given local autocovariance function. Even in the case where such a sparse representation does not exist, the LSW framework offers the flexibility of choosing the system $\{\Psi_j\}$ in such a way that (3.3) is the sparsest decomposition possible.

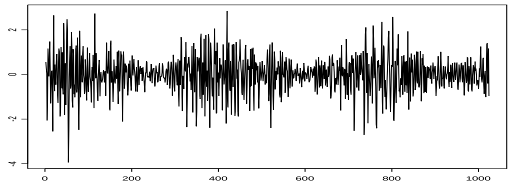
The next section develops the methodology of how to construct a predictor for an LSW process with a given system $\{\Psi_j\}$. Comparing the mean square prediction errors for different choices of this system is one possibility of solving the basis selection problem for LSW models.



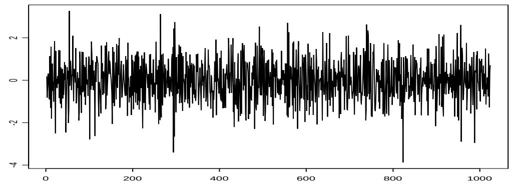
(a) Theoretical wavelet spectrum equal to zero everywhere except scale -2 where $S_{-2}(z) = 0.1 + \cos^2(3\pi z + 0.25\pi)$.



(b) Theoretical wavelet spectrum $S_{-2}(z) = 0.1 + \cos^2(3\pi z + 0.25\pi)$, $S_{-1}(z) = 0.1 + \sin^2(3\pi z + 0.25\pi)$ and $S_j(z) = 0$ for $j \neq -1, -2$.



(c) A sample path of length 1024 simulated from the wavelet spectrum defined in (a).



(d) A sample path of length 1024 simulated from the wavelet spectrum defined in (b).

Figure 3: These simulated examples demonstrate the idea of a sparse representation of the local (co)variance. The left-hand-side column shows an example of a smooth time-varying variance function. The example on the right hand side is constructed in such a way that the local variance function $c(z, 0)$ is constant over time. In this example, the only deviation from stationarity is in the covariance structure. The simulations, like all througout the article, use Gaussian innovations ξ_{jk} and Haar wavelets.

4 The general predictor and its theoretical properties

In this section, we define and analyse the general linear predictor for non-stationary data that are modelled to follow the LSW process representation given in Definition 1. Note that the particular construction of Section 2 for time-modulated processes will arise as a special case.

4.1 Definition of the linear predictor

Given t observations $X_{0,T}, X_{1,T}, \ldots, X_{t-1,T}$ of an LSW process, we define the *one-step-ahead predictor* of $X_{t,T}$ by

$$\hat{X}_{t,T} = \sum_{s=0}^{t-1} b^{(t)}_{t-1-s;T} \, X_{s,T} \ , \tag{4.1}$$

where the coefficients $b^{(t)}_{t-1-s}$ are such that they minimise the Mean Square Prediction Error (MSPE). The MSPE is defined by

$$\mathrm{MSPE}(\hat{X}_{t,T}, X_{t,T}) = \mathrm{E}\left(\hat{X}_{t,T} - X_{t,T}\right)^2 \ .$$

The predictor (4.1) is a linear combination of doubly-indexed observations where the weights need to follow the same doubly-indexed framework. This means that as the number of observation increases, we augment our knowledge about the local structure of the process, which allows us to fit coefficients $b^{(t)}_{t-1-s;T}$ more and more accurately. The double indexing of the weights is necessary due to the non-stationary nature of the data. This scheme is different to the traditional filtering of the data $X_{s,T}$ by a linear filter $\{\boldsymbol{b}_t\}$. In particular, we do not assume the (square) summability of the sequence $\boldsymbol{b}_t$ because (4.1) is a relation which is written in rescaled time. This is in accordance with the following assumption which holds throughout the paper.

Assumption 1. t depends on T and $t/T = c$ is a constant $c \in (0, 1)$ for all T. Moreover, we suppose that $[\log_2(t-1)] = [\log_2(T-1)]$.

Note that we can only forecast the future up to a certain horizon $T-1$. Asymptotically, $T \to \infty$, but also $t \to \infty$ in such a way that $t/T = c$, which ensures that we remain at the same point in the rescaled-time interval $[0, 1]$ as the number of observation increases. The second part of the assumption ensures that the coarsest scale in the wavelet decomposition of the observed data is equal to the coarsest scale for the whole "horizon" $[0, T-1]$.

4.2 Prediction in the wavelet domain

There is an interesting link between the above definition of the linear predictor (4.1) and another, "intuitive" definition of a predictor in the LSW model. Given observations up to time $t-1$, a natural way of defining a predictor of $X_{t,T}$ is to mimic the structure of the LSW model itself by passing to the wavelet domain. The empirical wavelet coefficients are defined by

$$d_{jk;T} = \sum_{s=0}^{t-1} X_{s,T}\ \psi_{jk}(s)$$

for all $j = -1, \dots, -J = -\log_2 T$ and $k \in \mathbb{Z}$. Then, the one-step-ahead predictor is constructed as

$$\hat{X}_{t,T} = \sum_{j=-J}^{-1} \sum_{k \in \mathbb{Z}} d_{jk;T}\ a^{(t)}_{jk;T}\ \psi_{jk}(t)\ , \tag{4.2}$$

where the coefficients $a^{(t)}_{jk}$ have to be estimated and are such that they minimise the MSPE. This predictor (4.2) may be viewed as a projection of $X_{t,T}$ on the space of random variables spanned by $\{d_{j,k;T} | j = -1, \dots, -J = -\log_2 T$ and $k = 0, \dots, T-1\}$.

It turns out that due to the redundancy of the non-orthogonal wavelet system $\{\psi_{jk}(t)\}$, the predictor (4.2) does not have a unique representation: there exists more than one solution $\{a^{(t)}_{jk}\}$ minimising the MSPE, but each solution gives the same predictor (expressed as a different linear combination of the redundant functions $\{\psi_{jk}(t)\}$). One can easily verify this observation by considering, for example, the stationary process $X_s = \sum_{k=-\infty}^{\infty} \psi_{-1k}(s)\zeta_k$, where ψ_{-1} is the non-decimated discrete Haar wavelet at scale -1 and ζ_k is an orthonormal increment sequence.

It is not surprising that the wavelet predictor (4.2) is related to the linear predictor (4.1) by

$$b^{(t)}_{t-s;T} = \sum_{j=-J}^{-1} \sum_{k\in\mathbb{Z}} a^{(t)}_{jk;T}\ \psi_{jk}(t)\ \psi_{jk}(s).$$

Because of the redundancy of the non-decimated wavelet system, for a fixed sequence $b^{(t)}_{t-s;T}$, there exists more than one sequence $a^{(t+1)}_{jk;T}$ such that this relation holds. For this reason, we prefer to work directly with the general linear predictor (4.1), bearing in mind that it can also be expressed as a (non-unique) projection onto the wavelet domain.

4.3 The general linear prediction equations

The goal of this subsection is to minimise the mean square prediction error $\mathrm{MSPE}(\hat{X}_{t;T}, X_{t;T})$ with respect to $b^{(t)}_{t-s;T}$. This quadratic function may be written as

$$\mathrm{MSPE}(\hat{X}_{t;T}, X_{t;T}) = \boldsymbol{b}_t' \boldsymbol{\Sigma}_{t;T} \boldsymbol{b}_t\ ,$$

where $\boldsymbol{b}_t$ is the vector $(b^{(t)}_{t-1;T}, \ldots, b^{(t)}_{0;T}, -1)$ and $\boldsymbol{\Sigma}_{t;T}$ is the covariance matrix of $X_{0;T}, \ldots, X_{t;T}$. However, the matrix $\boldsymbol{\Sigma}_{t;T}$ depends on $w^2_{jk;T}$ which cannot be estimated, as they are not identifiable (recall that the representation (3.1) is not unique due to the redundancy of the system $\{\psi_{jk}\}$). The next proposition shows that the MSPE may be approximated by $\boldsymbol{b}_t' \boldsymbol{B}_{t;T} \boldsymbol{b}_t$, where $\boldsymbol{B}_{t;T}$ is a $(t+1)\times(t+1)$ matrix whose $(m,n)-$th element is given by

$$\sum_{j=-J}^{-1} S_j\left(\frac{n+m}{2T}\right)\Psi_j(n-m)\ ,$$

and can be estimated by estimating the (uniquely defined) wavelet spectrum S_j.

Proposition 2. *Denote by $\mathcal{L}_j$ the length of the support of ψ_j and consider the following assumptions:*

$$\sum_{\tau=0}^{\infty} \sup_z |c(z,\tau)| < \infty \tag{4.3}$$

$$\sum_{j<0} (C_j + \mathcal{L}_j L_j)\mathcal{L}_j < \infty \tag{4.4}$$

$$C_1 := \operatorname*{ess\,sup}_{z,\omega} \sum_{j<0} S_j(z)|\hat{\psi}_j(\omega)|^2 < \infty \tag{4.5}$$

$$C_2 := \operatorname*{ess\,inf}_{z,\omega} \sum_{j<0} S_j(z)|\hat{\psi}_j(\omega)|^2 > 0, \tag{4.6}$$

where

$$\hat{\psi}_j(\omega) = \sum_{s=-\infty}^{\infty} \psi_{j0}(s)\exp(i\omega s).$$

Under Assumptions (4.3), (4.4) and (4.6), the mean square prediction error may be written as

$$\text{MSPE}(\hat{X}_{t;T}, X_{t;T}) = \boldsymbol{b}_t' \boldsymbol{B}_{t;T} \boldsymbol{b}_t \ (1 + o_T(1)) \ . \tag{4.7}$$

Moreover, if $\{b_{s;T}^{(t)}\}$ *are the coefficients which minimise* $\boldsymbol{b}_t' \boldsymbol{B}_{t;T} \boldsymbol{b}_t$*, then* $\{b_{s;T}^{(t)}\}$ *solve the following linear system*

$$\sum_{m=0}^{t-1} b_{t-1-m;T}^{(t)} \left\{ \sum_{j=-J}^{-1} S_j \left(\frac{n+m}{2T} \right) \Psi_j(m-n) \right\} = \sum_{j=-J}^{-1} S_j \left(\frac{t+n}{2T} \right) \Psi_j(t-n) \tag{4.8}$$

for all $n = 0, \ldots, t-1$.

The proof of the first result can be found in Appendix A (see Lemma 4) and uses standard approximations of covariance matrices of locally stationary processes. The second result is simply the minimisation of the quadratic form (4.7) and formula (4.8) is called the *prediction equations*. The key observation here is that minimising $\boldsymbol{b}_t' \boldsymbol{\Sigma}_{t;T} \boldsymbol{b}_t$ is asymptotically equivalent to minimising $\boldsymbol{b}_t' \boldsymbol{B}_{t;T} \boldsymbol{b}_t$. Bearing in mind relation (3.3) between the wavelet spectrum and the local autocovariance function, the prediction equations can also be written as

$$\sum_{m=0}^{t-1} b_{t-1-m;T}^{(t)} \, c\left(\frac{n+m}{2T}, m-n \right) = c\left(\frac{n+t}{2T}, t-n \right). \tag{4.9}$$

The following two remarks demonstrate how the prediction equations simplify in the case of two important subclasses of locally stationary wavelet processes.

Remark 1 (Stationary processes). If the underlying process is stationary, then the local autocovariance function $c(z, \tau)$ is no longer a function of two variables, but only a function of τ. In this context, the prediction equations (4.9) become

$$\sum_{m=0}^{t-1} b_{t-1-m}^{(t)} \, c(m-n) = c(t-n)$$

for all $n = 0, \ldots, t-1$, which are the standard Yule-Walker equations used to forecast stationary processes.

Remark 2 (Time-modulated case). For time-modulated stationary processes, considered in Section 2 (equation (2.1)), the local autocovariance function has a multiplicative structure: $c(z, \tau) = \sigma^2(z) \rho_Y(\tau)$. Therefore, for these processes, the prediction equations (4.9) become

$$\sum_{m=0}^{t-1} b_{t-1-m;T}^{(t)} \sigma^2 \left(\frac{n+m}{2T} \right) \rho_Y(m-n) = \sigma^2 \left(\frac{n+t}{2T} \right) \rho_Y(t-n).$$

We will now study the inversion of the system (4.8) in the general case, and the stability of the inversion. Denote by $\boldsymbol{P}_t$ the matrix of this linear system, i.e.

$$(\boldsymbol{P}_t)_{nm} = \sum_{j=-J}^{-1} S_j \left(\frac{n+m}{2T} \right) \Psi_j(m-n)$$

for $n, m = 0, \ldots, t-1$. Using classical results of numerical analysis (see for instance Kress (1991, Theorem 5.3)) the measure of this stability is given by the so-called *condition number*, which is defined by $\text{cond}(\boldsymbol{P}_t) = \|\boldsymbol{P}_t\| \, \|\boldsymbol{P}_t^{-1}\|$. It can be proved along the lines of Lemma 2 (Appendix A) that, under Assumptions (4.5) and (4.6), $\text{cond}(\boldsymbol{P}_t) \leqslant C_1 C_2$.

4.4 The prediction error

The next result generalises the classical Kolmogorov formula for the theoretical prediction error (Brockwell and Davis, 1991, Theorem 5.8.1). It is a direct modification of a similar result stated by Dahlhaus (1996b, Theorem 3.2(i)) for locally stationary Fourier processes.

Proposition 3. *Suppose that Assumptions (4.3), (4.4) and (4.6) hold. Given t observations $X_{0,T}, \ldots, X_{t-1,T}$ of the LSW process $\{X_{t,T}\}$, the one-step ahead mean square prediction error σ^2_{OSPE} in forecasting $\hat{X}_{t,T}$ is given by*

$$\sigma^2_{\mathrm{OSPE}} = \exp\left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} d\omega \ \ln\Big(\sum_{j=-\infty}^{-1} S_j\Big(\frac{t-1}{T}\Big) |\hat{\psi}_j(\omega)|^2 \Big) \right\} (1 + o_T(1)) \ .$$

Note that due to Assumption (4.6), the sum $\sum_j S_j((t-1)/T)|\hat{\psi}_j(\omega)|^2$ is strictly positive, except possibly on a set of measure zero.

4.5 h-step-ahead prediction

Consider now the h-step-ahead linear prediction of an LSW process $X_{t,T}$, observed up to time $t-1$. In what follows, h may be a function of T. The predictor is defined by

$$\hat{X}_{t+h-1,T} = \sum_{s=0}^{t-1} b^{(t+h-1)}_{t-1-s;T} \ X_{s,T}$$

where the coefficients $b^{(t+h-1)}_{t-1-s}$ have to be estimated and are such that they minimise the mean square prediction error

$$\mathrm{MSPE}(\hat{X}_{t+h-1,T}, X_{t+h-1,T}) = \mathrm{E}\left(\hat{X}_{t+h-1,T} - X_{t+h-1,T}\right)^2 = \boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1},$$

where $\boldsymbol{\Sigma}_{t+h-1;T}$ is the covariance matrix of $X_{0,T}, \ldots, X_{t+h-1,T}$ and $\boldsymbol{b}_{t+h-1}$ is the vector $(b^{(t+h-1)}_{t-1}, \ldots, b^{(t+h-1)}_{0}, b^{(t+h-1)}_{-1}, \ldots, b^{(t+h-1)}_{-h})$, with $b^{(t+h-1)}_{-1}, \ldots, b^{(t+h-1)}_{-h+1} = 0$ and $b^{(t+h-1)}_{-h} = -1$. Like before, we approximate the mean square error by $\boldsymbol{b}'_{t+h-1}\boldsymbol{B}_{t+h-1;T}\boldsymbol{b}_{t+h-1}$, where $\boldsymbol{B}_{t+h-1;T}$ is a $(t+h)\times(t+h)$ matrix whose (m,n)-th element is given by

$$\sum_{j=-J}^{-1} \sum_{k=-\infty}^{\infty} S_j \ \left(\frac{n+m}{2T}\right) \ \Psi_j(n-m) \ .$$

Proposition 4. *Assume that h may depend on T but is such that $h = h_T = o(T)$. Under the assumptions of Proposition 2, the mean square prediction error may be written as*

$$\mathrm{MSPE}(\hat{X}_{t;T}, X_{t;T}) = \boldsymbol{b}'_{t+h-1}\boldsymbol{B}_{t+h-1;T}\boldsymbol{b}_{t+h-1} \ (1 + o_T(1)) \ .$$

The condition $h/T \to 0$ means that the prediction is performed in a local neighbourhood of stationarity around t, in accordance with our model of local stationarity.

5 Prediction based on data

Having treated the prediction problem from a theoretical point of view, we now address the question of how to estimate the unknown time-varying second order structure in the system of equations (4.8). In the second part of this section, we propose a computational procedure for choosing the smoothing parameter, as well as the optimal size of the system (4.8), in such a way as to optimise the performance of prediction.

5.1 Estimation of the time-varying second-order structure

In the general case, our estimator of the local autocovariance function $c(z,\tau)$, with $0 < z < t/T$, is based on replacing the unknown wavelet spectrum $S_j(z)$ by an asymptotically unbiased and consistent estimator. Following Nason *et al.* (2000) we will use a bias-corrected version of the *wavelet periodogram*. The wavelet periodogram has already been used in equation (2.2). Below, we recall its definition as the sequence of squared wavelet coefficients $d_{jk;T}$, where j and k are scale and location parameters, respectively.

$$I_j(k/T) = d^2_{jk;T} = \left(\sum_{s=0}^{t-1} X_{s,T}\ \psi_{jk}(s) \right)^2 , \quad -J = -\log_2(t) \leqslant j \leqslant -1,\ k = 0, \dots, t-1 \ .$$

Recall that, due to the redundancy of the non-decimated wavelet basis $\{\psi_{jk}(t)\}$, a bias correction of this estimator has to be applied: Proposition 4 of Nason *et al.* (2000) states that

$$\mathrm{E}\left[\sum_{\ell=-J}^{-1} A^{-1}_{j\ell} I_\ell(z) \right] - S_j(z) \ = \ O(T^{-1}) \ ,$$

uniformly in $z \in (0,1)$, where the matrix $\boldsymbol{A} = (A_{j\ell})_{j,\ell<0}$ is defined by

$$A_{j\ell} \ := \ < \Psi_j, \Psi_\ell > \ = \ \sum_\tau \Psi_j(\tau)\ \Psi_\ell(\tau) \ . \tag{5.1}$$

Note that the matrix $A_{j\ell}$ is not simply diagonal due to the redundancy in the system of autocorrelation wavelets $\{\Psi_j\}$. Nason *et al.* (2000) proved the invertibility of $\boldsymbol{A}$ if $\{\Psi_j\}$ is constructed using Haar wavelets. If other compactly supported wavelets are used, numerical results suggest that the invertibility of $\boldsymbol{A}$ still holds, but a complete proof of this result has not been established yet. As we need the invertibility of $\boldsymbol{A}$ in our following theoretical results, from now on we restrict ourselves to Haar wavelets, but we conjecture that all results remain valid for more general Daubechies wavelets.

We will now study the following asymptotically unbiased estimator of the local autocovariance function:

$$\tilde{c}\left(\frac{k}{T}, \tau\right) \ = \ \sum_{j=-J}^{-1} \left(\sum_{\ell=-J}^{-1} A^{-1}_{j\ell} \Psi_\ell(\tau) \right) I_j\left(\frac{k}{T}\right) \ , \quad k = 0, \dots, t-1, \tau \neq 0, \tag{5.2}$$

and, in the case of estimating the time-varying variance $c(z,0)$, the estimator (2.2), i.e.,

$$\tilde{c}\left(\frac{k}{T},0\right)=\sum_{j=-J}^{-1}2^{j}\ I_j\left(\frac{k}{T}\right). \tag{5.3}$$

Note that asymptotically, for $\tau=0$, the general estimator (5.2) reduces to the estimator (5.3), by use of Lemma 6 (Appendix B) and recalling that $\Psi_j(0)=1$ for all j.

The results below concern the asymptotic behaviour of the first two moments of the estimators (5.2) and (5.3).

Proposition 5. *Assume that (4.4) holds. Then*

$$\begin{aligned}
\mathrm{E}\,\tilde{c}\left(\frac{k}{T},0\right)&=c\left(\frac{k}{T},0\right)+O\left(T^{-1}\log_2 T\right)\\
\mathrm{E}\,\tilde{c}\left(\frac{k}{T},\tau\right)&=c\left(\frac{k}{T},\tau\right)+O\left(T^{-1/2}\log_2 T\right)\qquad \text{for}\quad \tau\neq 0.
\end{aligned}$$

If, in addition, Assumption (4.3) holds and the increment process $\{\xi_{jk}\}$ in Definition 1 is Gaussian, then

$$\begin{aligned}
\mathrm{Var}\ \tilde{c}\left(\frac{k}{T},0\right)&=2\sum_{j,\ell=-J}^{-1}2^{j}2^{\ell}\left\{\sum_{s,t\in\mathbb{Z}}c\left(\frac{s+t}{2T},t-s\right)\psi_{jk}(s)\psi_{\ell k}(t)\right\}^2+O\left(T^{-1}\log_2 T\right)\\
\mathrm{Var}\ \tilde{c}\left(\frac{k}{T},\tau\right)&=2\sum_{j,\ell=-J}^{-1}h_j(\tau)h_\ell(\tau)\left\{\sum_{s,t\in\mathbb{Z}}c\left(\frac{s+t}{2T},t-s\right)\psi_{jk}(s)\psi_{\ell k}(t)\right\}^2+O\left(T^{-1/2}\log_2 T\right).
\end{aligned}$$

for $\tau\neq 0$, where $h_j(\tau)=\sum_{\ell=-J}^{-1}A_{j\ell}^{-1}\Psi_\ell(\tau)$.

Remark 3 (Time-modulated case). For time-modulated stationary processes considered in Section 2 (equation (2.1)), the variance of the multiscale estimator reduces to

$$\mathrm{Var}\,\tilde{\sigma}^2_{\mathrm{MS}}\left(\frac{t}{T}\right)=2\sum_{j,\ell=-J}^{-1}2^{j+\ell}\left(\sum_{s,u=0}^{T-1}\sigma^2\left(\frac{s+u}{2T}\right)\rho_Y(u-s)\psi_{jt}(s)\psi_{\ell t}(u)\right)^2+O\left(T^{-1}\log_2 T\right),$$

where $\rho_Y(\tau)$ is the autocorrelation function of Y_t (see equation (2.1)). If, in addition, Y_t is a Gaussian white noise, then we can use the Lipschitz continuity of $\sigma^2(z)$ to show that

$$\mathrm{Var}\,\tilde{\sigma}^2_{\mathrm{MS}}\left(\frac{t}{T}\right)=\frac{2}{3}\,\sigma^4\left(\frac{t}{T}\right)+O\left(T^{-1}\log_2 T\right)$$

holds for all compactly supported wavelets ψ. Observing that the variance of the naive estimator $\tilde{\sigma}_{\mathrm{na}}(t/T)$, defined in Section 2, is equal to $2\sigma^4(t/T)$, we conclude that the asymptotic variance of the multiscale estimator is smaller, while, at the same time, the multiscale estimator is still asymptotically unbiased.

Proposition 5 shows the inconsistency of the estimator of the local (co)variance. As already mentioned in the case of time-modulated stationary processes, it needs to be smoothed over time $z = k/T$. We use standard kernel smoothing where the problem of the choice of the bandwidth parameter g arises. The goal of Subsection 5.2 is to provide a fully automatic procedure for choosing g.

To compute the linear predictor in practice, we invert the generalised Yule-Walker equations (4.9) in which the theoretical local autocovariance function is replaced by our estimate $\tilde{c}(k/T, \tau)$. However, in equations (5.2) and (5.3), our estimator is only defined for $k = 0, \ldots, t-1$ while the prediction equations (4.9) require the local autocovariance up to $k = t$. This problem is inherent to our non-stationary framework, and has already been mentioned in Subsection 2.2.2. Like in that subsection, we are motivated by the slow evolution of the local autocovariance function, and we propose to compute $\hat{c}(t/T, \tau)$ by the local smoothing of the (unsmoothed) estimators $\{\tilde{c}(k/T, \tau), k = t-1, \ldots, t-\mu\}$. In practice, the smoothing parameter μ for prediction is set to be equal to the smoothing parameter g for estimation. They can be obtained by the data-driven procedure described in the next subsection.

5.2 Data-driven choice of parameters

In theory, the best linear predictor of X_t is given by (4.1), where $\boldsymbol{b}_t = (b^{(t)}_{t-1-s;T})_{s=0,\ldots,t-1}$ solves the prediction equations (4.8). In practice, each of the t components of vector $\boldsymbol{b}_t$ is estimated using our estimator of the local autocovariance function based on observations $X_{0,T}, \ldots, X_{t-1,T}$. Hence, we have to find a balance between the estimation error, potentially increasing with t, and the prediction error which is a decreasing function of t.

As a natural balancing rule which works well in practice, we suggest to choose a number p such that the "clipped" predictor

$$\hat{X}^{(p)}_{t,T} = \sum_{s=t-p}^{t-1} b^{(t)}_{t-1-s;T} X_{s,T} \tag{5.4}$$

gives a good compromise between the theoretical prediction error and the estimation error. The construction (5.4) is reminiscent of the classical idea of AR(p) approximation for stationary processes.

We propose an automatic procedure for selecting the two parameters: the order p in (5.4) and the bandwidth g, necessary to smooth the inconsistent estimator $\tilde{c}(z, \tau)$ by a kernel method. The idea of this procedure is to start with initial values of p and g and to gradually update these parameters using a criterion which measures how well the series gets predicted using a given pair of parameters. This type of approach is in the spirit of *adaptive forecasting* (Ledolter, 1980) and is well known in the Bayesian forecasting framework (West and Harrison, 1997).

Suppose that we observe the series up to X_{t-1} and we want to predict X_t, using an appropriate pair (p, g). The idea of our method is as follows. First, we move backwards by s observations and choose initial parameters (p_0, g_0) for predicting X_{t-s} from the observed series up to X_{t-s-1}. Next, we compute the prediction of X_{t-s} using the pairs of parameters around our preselected pairs (i.e. $(p_0 - 1, g_0 - \delta)$, $(p_0, g_0 - \delta)$, $\ldots$, $(p_0 + 1, g_0 + \delta)$ for a fixed constant δ). Since the true value of $X_{t-s,T}$ is known, we are able to use a preset criterion to compare the 9 obtained prediction results, and we choose the pair corresponding to the best predictor, according to this preset criterion. This step is called the *update of the parameters* by predicting X_{t-s}. In the next step, this updated pair is used as initial parameters and itself updated by predicting X_{t-s+1} from $X_{0,t}, \ldots, X_{t-s}$. By applying this procedure to predict $X_{t-s+2}, X_{t-s+3}, \ldots, X_{t-1}$, we finally obtain an updated pair (p_1, g_1) which is selected to do the actual prediction.

Many different criteria can be used to compare the quality of the pairs of parameters at each step. Denote by $\hat{X}_{t-i,T}(p, g)$ the predictor of $X_{t-i,T}$ computed using pair (p, g), and by $\mathrm{P}_{t-i,T}(p, g)$ the length of the corresponding prediction interval. The criterion which we used in the simulations reported in the next section is to compute

$$\frac{\left|X_{t-i,T} - \hat{X}_{t-i,T}(p, g)\right|}{\mathrm{P}_{t-i,T}(p, g)}$$

for each of the 9 pairs at each step of the procedure and select the updated pair as the one that minimises this ratio.

We also need to choose the initial parameters (p_0, g_0) and the number s of data points at the end of the series which are used in the procedure. We suggest that s should be set to the length of the largest segment at the end of the series which does not contain any apparent breakpoints observed after a visual inspection. To avoid dependence on the initial values (p_0, g_0), we suggest to iterate the algorithm a few times, using (p_1, g_1) as the initial value for each iteration. We propose to stop when the parameters (p_1, g_1) are such that at least 95% of the observations fall into the prediction intervals (as in Figure 2).

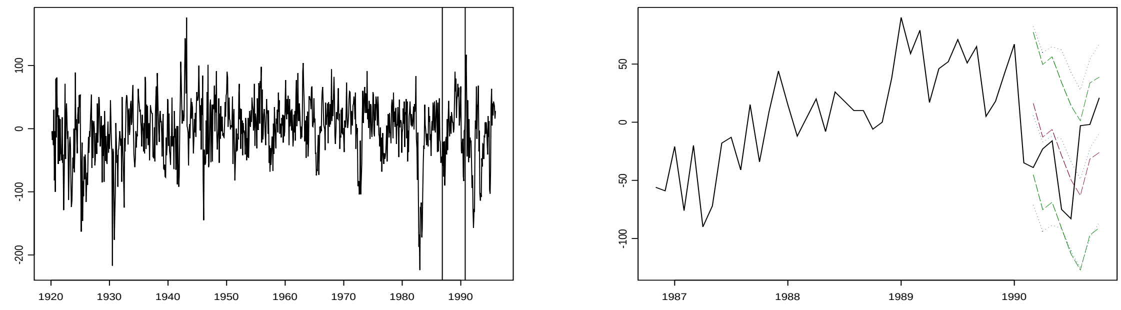
The procedure of choosing (p_1, g_1) to perform the actual prediction $\hat{X}_{t,T}$ is in accordance with the principle of local stationarity: if the pair (p_1, g_1) was good at predicting $X_{t-1,T}$, then we can expect that the same pair of parameters, or possibly one of its neighbours, will also perform well in forecasting $X_{t,T}$. In order to be able to use our procedure completely online, we do not have to repeat the whole algorithm. Indeed, when observation $X_{t,T}$ becomes available, we only have to update the pair (p_1, g_1) by predicting $X_{t,T}$, and we directly obtain the optimal pair for predicting $X_{t+1,T}$.

There are, obviously, many possible variants of our algorithm. Possible modifications include, for example, using a different criterion, restricting the allowed parameter space for (p, g), penalising certain regions of the parameter space, or allowing more than one parameter update at each time point.

We have tested our algorithm on numerous examples, and the following Section presents an application to a real data set. A more theoretical study of this algorithm is left for future work.

6 Application of the general predictor to real data

El Niño is a disruption of the ocean atmosphere system in the tropical Pacific which has important consequences for the weather around the globe. Even though the effect of El Niño is not avoidable, research on its forecast and its impacts allows specialists to attenuate or prevent its harmful consequences (see Philander (1990) for a detailed overview). The effect of the equatorial Pacific meridional reheating may be measured by the deviation of the wind speed on the ocean surface from its average. It is worth mentioning that this effect is produced by conduction, and thus we expect the wind speed variation to be smooth. This legitimates the use of LSW processes to model the speed. In this section, we will study the wind speed anomaly index, i.e. its standardised deviation from the mean, in a specific region of the Pacific (12-2N, 160E-70W). Modelling this anomaly helps to understand the effect of El Niño effect in that region. The time series composed of $T = 910$ monthly observations is available free of charge at `http://tao.atmos.washington.edu/data_sets/eqpacmeridwindts`. Figure 4(a) shows the plot of the series.



(a) The wind anomaly index (in cm/s). The two vertical lines indicate the segment shown in Figure 4(b).

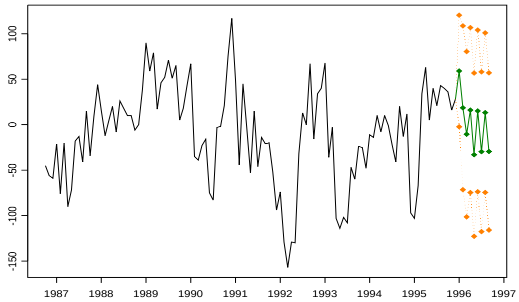
(b) Comparison between the one-step-ahead prediction in our model (dashed lines) and AR (dotted lines).

Figure 4: The wind anomaly data (910 observations from March 1920 to December 1995).

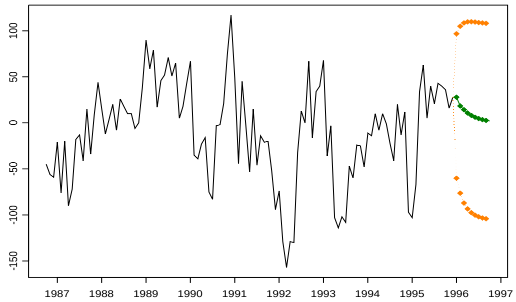
Having provisionally made a safe assumption of the possible non-stationarity of the data, we first attempt to find a suitable pair of parameters (p, g) which will be used for forecasting the series. By inspecting the acf of the series, and by trying different values of the bandwidth, we have found that the pair $(7, 70/T)$ works well for many segments of the data; indeed, the segment of 100 observations from June 1928 to October 1936 gets predicted very accurately in one-step-ahead prediction: 96% of the actual observations are contained in the corresponding 95% prediction intervals (computed assuming normality).

However, the pair $(7, 70/T)$ does not appear to be uniformly well suited for forecasting the whole series. For example, in the segment of 40 observations between November 1986 and February 1990, only 5% of the observations fall into the corresponding one-step-ahead prediction intervals computed using the above pair of parameters. This provides strong evidence that the series is non-stationary (indeed, if it was stationary, we could expect to obtain a similar percentage of accurately predicted values in both segments) This further justifies our approach of modelling and forecasting the series as an LSW process.

Motivated by the above observation, we now apply our algorithm, described in the previous section, to the segment of 40 observation mentioned above, setting the initial parameters to $(7, 70/T)$. After the first iteration along the segment, the parameters drift up to $(14, 90/T)$, and 85% of the observations fall within the prediction intervals, which is indeed a dramatic improvement over the 5% obtained without applying our adaptive algorithm. In the second pass, we set the initial values to $(14, 90/T)$, and obtain a 92.5% coverage by the one-step-ahead prediction intervals, with the parameters drifting up to $(14, 104/T)$. In the last iteration, we finally obtain a 95% coverage, and the parameters get updated to $(14, 114/T)$. We now have every reason to believe that this pair of parameters is well suited for one-step-ahead prediction within a short distance of February 1990. Without performing any further updates, we apply the one-step-ahead forecasting procedure to predict, one by one, the eight observations which follow February 1990, the prediction parameters being fixed at $(14, 114/T)$. The results are plotted in Figure 4(b), which also compares our results to those obtained by means of AR modelling. The results are presented in the same way as in Figure 2, which compares our methodology to GARCH modelling. At each time point, the order of the AR process is chosen as the one that minimises the AIC criterion, and then the parameters are estimated by means of the standard S-Plus routine. We observe that for both models, all of the true observed values fall within the corresponding one-step-ahead prediction intervals. However, the main gain obtained using our procedure is that the prediction intervals are on average 17.45% narrower in the case of our algorithm. This result is not peculiar to AR modelling as this percentage is similar in the case comparison with other stationary models, like ARMA(2,10), believed to accurately fit the series. A similar phenomenon has been observed at several other points of the series. Also, we remind the reader that the same advantage of our procedure has already been observed in subsection 2.3.



(a) 9-step-ahead prediction using LSW modelling



(b) 9-step-ahead prediction using AR modelling

Figure 5: The end of the wind anomaly series and its 1- up to 9-step-ahead forecasts (in cm/s).

We end this section by applying our general prediction method to compute multi-step-ahead forecasts. Figure 5 shows the 1- up to 9-step-ahead forecasts of the series, along with the corresponding prediction intervals, computed at the end of the series (December 1995). In Figure 5(a), the LSW model is used to construct the forecast values, with parameters $(10, 2.18)$ chosen automatically by our adaptive algorithm described above. Figure 5(b) shows the 9-step-ahead prediction based on AR modelling (here, AR(2)). The prediction in Figure 5(a) looks "smoother" because it uses the information from the whole series. This information is averaged out, whereas in the LSW forecast, local information is picked up at the end of the series, and the prediction looks more "jagged".

It is worth mentioning here that our approach is inherently different from the one that attempts to find (almost) stationary segments at the end of the series to perform the prediction. Instead, our procedure is adapting the prediction coefficients to the slow evolution of the covariance.

7 Conclusion

In this paper, we have given an answer to the pertinent question, asked by time series analysts over the past few years, of whether and how wavelet methods can help in forecasting non-stationary time series. To develop the forecasting methodology, we have considered the Locally Stationary Wavelet (LSW) model, which is based on the idea of a localised time-scale representation of a time-changing autocovariance function. This model includes the class of second-order stationary processes and has several attractive features, not only for modelling, but also for estimation and prediction purposes. Its linearity and the fact that the time-varying second order quantities are modelled as smooth functions, have enabled us to formally extend the classical theory of linear prediction to the whole class of LSW processes. This results in a generalisation of the Yule-Walker equations and, in particular, of Kolmogorov's formula for the one-step-ahead prediction error.

In the empirical prediction equations the second-order quantities have to be estimated, and this is where the LSW model proves most useful. The rescaled time, one of the main ingredients of the model, makes it possible to develop a rigorous estimation theory: our multiscale estimators of the time-varying variance and covariance are asymptotically unbiased, yet their variation is substantially reduced compared to the traditional unbiased estimator of local variance. More generally, by using well-localised non-decimated wavelets instead of a Fourier based approach, our estimators are able to capture the local time-scale features of the observed non-stationary data very well (Nason and von Sachs, 1999).

In practice, our new prediction methodology depends on two nuisance parameters which arise in the estimation of the local covariance and the mean-square prediction error. More specifically, we need to smooth our inconsistent estimators over time, and to do so, we have to choose the bandwidth of the smoothing kernel. Moreover, we need to reduce the dimension of the prediction equations to avoid too much inaccuracy of the resulting prediction coefficients due to estimation errors. We have proposed an automatic computational procedure for selecting these two parameters. Our algorithm is in the spirit of adaptive forecasting — it gradually updates the two parameters basing on the success of prediction. This new method is not only essential for the success of our whole prediction methodology, it also seems to be promising in a much wider context of choosing nuisance parameters in non-parametric methods in general.

We have applied our new algorithm to two different data sets: weekly log-returns of the Dow Jones index, and a meteorological time series. We have modelled the former one as a time-modulated process, and the latter one as a general LSW process. (Time-modulated processes constitute a simple but important subclass of LSW processes, and, according to recent literature, can be successfully applied to the analysis of certain financial time series.) In both cases, our non-parametric forecasting algorithm shows very interessing advantages over the classical parametric alternatives (GARCH and AR forecasting, respectively). Moreover, we believe that one of the biggest advantages of our new algorithm is that it can be successfully applied to a variety of data sets, ranging from financial log-returns to series traditionally modelled as ARMA processes, including in particular data sets which are not, or do not appear to be, second-order stationary. The software implementing our algorithm, as well as the both data sets, can be downloaded from the associated web page

`http://www.stats.bris.ac.uk/~mapzf/flsw/flsw.html`

In the future, we intend to derive the theoretical properties of our automatic algorithm for choosing the nuisance parameters of the adaptive predictor. Last but not least, let us emphasize that the approach which we have adopted offers the attractive possibility of using prediction error for model selection purposes. LSW processes are constructed using a fixed wavelet system, e.g. Haar or another Daubechies' system. It is clear that we can compare the fitting quality of each such model by comparing its prediction performance on the observed data. In the future, we intend to investigate this in more detail in order to answer the question, left open by Nason *et al.* (2000), of which wavelet basis to use to model a given series.

8 Acknowledgements

The authors would like to thank Christian Hafner and Guy Nason as well as Peter Brockwell and Brandon Whitcher – the first two for stimulating discussions and suggestions which helped to improve the presentation of this article, and the other two for their expertise and advice in analysing the wind data of Section 6. Sébastien Van Bellegem and Rainer von Sachs would like to express their gratitude to the Department of Mathematics, University of Bristol, and Piotr Fryźlewicz – to the Institut de statistique, Université catholique de Louvain, and SVB, for their hospitality during mutual visits in 2001. RvS and SVB were funded from the National Fund for Scientific Research – Wallonie, Belgium (F.N.R.S.) and by the contract "Projet d'Actions de Recherche Concertées" no. 98/03-217 of the Belgian Government. PF was funded by the Department of Mathematics at the University of Bristol, Universities UK, Unilever Research, and Guy Nason.

A Theoretical properties of the predictor

Let $\boldsymbol{X}_{t;T} = (X_{0;T}, \dots, X_{t-1;T})'$ be a realisation of an LSW process. In this appendix, we study the theoretical properties of the covariance matrix $\boldsymbol{\Sigma}_{t;T} = \mathrm{E}(\boldsymbol{X}'_{t;T} \cdot \boldsymbol{X}_{t;T})$. As we need upper bounds for the spectral norms $\|\boldsymbol{\Sigma}_{t;T}\|$ and $\|\boldsymbol{\Sigma}^{-1}_{t;T}\|$, we base the following results and their proofs on methods developed in Dahlhaus (1996b, Section 4) for approximating covariance matrices of locally stationary Fourier processes. However, in our setting these methods need important modifications. The idea is to approximate $\boldsymbol{\Sigma}_{t;T}$ by overlapping block Toeplitz matrices along the diagonal.

The approximating matrix is constructed as follows. First, we construct a coverage of the time axis $[0, T[$. Let L be a divisor of T such that $L/T \to 0$, and consider the following partition of the time axis:

$$\mathcal{P}_0 = \Big\{[0, L[, [L, 2L[, \dots, [T-L, T[\Big\}.$$

Then, consider another partition of the time axis, which is a shift of $\mathcal{P}_0$ by $\delta < L$:

$$\mathcal{P}_1 = \Big\{[0, \delta[, [\delta, L + \delta[, [L + \delta, 2L + \delta[, \dots , [T - L + \delta, T[\Big\}.$$

In what follows, assume that L is a multiple of δ and that $\delta/L \to 0$ as T tends to infinity. Also, consider the partition of the time axis which is a shift of $\mathcal{P}_1$ by δ:

$$\mathcal{P}_2 = \Big\{[0, 2\delta[, [2\delta, L + 2\delta[, [L + 2\delta, 2L + 2\delta[, \dots , [T - L + 2\delta, T[\Big\}$$

and, analogously, define $\mathcal{P}_3, \mathcal{P}_4, \dots$ up to $\mathcal{P}_M$ where $M = (L/\delta) - 1$. Consider the union of all these partitions $\mathcal{P} = \{\mathcal{P}_0, \mathcal{P}_1, \dots , \mathcal{P}_M\}$, which is a highly redundant coverage of the time axis. Denote by P the number of intervals in $\mathcal{P}$, and denote the elements of $\mathcal{P}$ by L_p, $p = 1, \dots , P$.

For each p, we fix a point ν_p in L_p and consider matrix $\boldsymbol{D}^{(p)}$ defined by:

$$D^{(p)}_{nm} = \sum_{j<0} S_j \left(\frac{\nu_p}{T}\right) \Psi_j(n - m)\mathbb{I}_{n,m\in L_p}$$

where $\mathbb{I}_{n,m\in L_p}$ means that we only include those n, m that are in L_p. The following lemma concerns the approximation of $\boldsymbol{\Sigma}_{t;T}$ by matrix $\boldsymbol{D}$ defined by

$$D_{nm} = \frac{\delta}{L} \sum_{p=1}^{P} D^{(p)}_{nm}.$$

Lemma 1. *Assume that (4.3) and (4.4) hold. If $L \to \infty$, $\delta/L \to 0$ and $L^2/T \to 0$ as $T \to \infty$, then*

$$\boldsymbol{x}' (\boldsymbol{\Sigma}_{t;T} - \boldsymbol{D}) \boldsymbol{x} = \boldsymbol{x}'\boldsymbol{x} o_T(1).$$

Proof. Define matrix $\boldsymbol{\Sigma}^{(p)}_{t;T}$ by $\Sigma^{(p)}_{nm} = \Sigma_{nm}\mathbb{I}_{n,m\in L_p}$. Straightforward calculations yield

$$\boldsymbol{x}' (\boldsymbol{\Sigma}_{t;T} - \boldsymbol{D}) \boldsymbol{x} = \boldsymbol{x}' \left(\frac{\delta}{L} \sum_{p=1}^{P} \boldsymbol{\Sigma}^{(p)}_{t;T} - \boldsymbol{D}^{(p)} \right) \boldsymbol{x} + \mathrm{Rest}_T \tag{A.1}$$

where

$$\mathrm{Rest}_T = \sum_{n,m=0}^{\frac{T}{\delta}-1} \min \left(|n - m| \frac{\delta}{L}, 1 \right) \sum_{u,s=0}^{\delta-1} x_{n\delta+u}\Sigma_{n\delta+u,m\delta+s}x_{m\delta+s}.$$

Let us first bound this remainder. Replace Σ_{nm} by $\sum_j S_j((n + m)/2T, n - m)$ and denote $b(k) = \sum_z | \sum_j S_j(z)\Psi_j(k)|$. We have

$$|\,\mathrm{Rest}_T\,| \leqslant 2\boldsymbol{x}'\boldsymbol{x} \sum_{d=1}^{\frac{T}{\delta}-1} \min \left(d\frac{\delta}{L}, 1 \right) \sum_{k=(d-1)\delta+1}^{d\delta} b(k) + \mathrm{Rest}'_T$$

where Rest'_T will be considered later. Splitting the sum over d at point $1+\sqrt{L}/\delta$, the above can be bounded by

$$\leqslant 2\boldsymbol{x}'\boldsymbol{x}\left(\frac{\delta+\sqrt{L}}{L}\sum_{k=1}^{\infty} b(k) + \sum_{k>\sqrt{L}} b(k)\right) + \mathrm{Rest}'_T$$

and the main term in the above is $o_T(1)$ since $L \to \infty$ and $\delta/L \to 0$ as $T \to \infty$, and by assumption (4.3). Let us now turn to the remainder Rest'_T. We have

$$\mathrm{Rest}'_T \leqslant \sum_{n,m=0}^{T-1} \left| x_n x_m \sum_{j,k} \left(w^2_{jk;T} - S_j\left(\frac{n+m}{2T}\right)\right) \psi_{j,k}(m)\psi_{j,k}(n) \right|$$

which may be bounded as follows using the definition of an LSW process, and the Lipschitz property of S_j:

$$\begin{aligned}
&\leqslant O(T^{-1}) \sum_j (C_j + \mathcal{L}_j L_j) \sum_k \left(\sum_{n=k-\mathcal{L}_j+1}^{k} |x_n \psi_{j,k}(n)| \right)^2 \\
&\leqslant O(T^{-1})\boldsymbol{x}'\boldsymbol{x} \sum_j (C_j + \mathcal{L}_j L_j)\mathcal{L}_j \leqslant O(T^{-1})\boldsymbol{x}'\boldsymbol{x}
\end{aligned}$$

by Assumption (4.4).

Let us finally consider the main term in (A.1). We have

$$\begin{aligned}
\boldsymbol{x}'\left(\frac{\delta}{L}\sum_{p=1}^{P} \boldsymbol{\Sigma}^{(p)}_{t;T} - \boldsymbol{D}^{(p)}\right)\boldsymbol{x} &\leqslant \frac{\delta}{L}\sum_{p=1}^{P}\sum_{jk}\left| w^2_{jk;T} - S_j\left(\frac{\nu_p}{T}\right)\right| \left(\sum_n \psi_{j,k}(u) x_u \mathbb{I}_{u\in L_p}\right)^2 \\
&\leqslant O(T^{-1})\frac{\delta}{L}\sum_{p=1}^{P}\sum_{jk}\left(\sum_n x_n^2 \mathbb{I}_{n\in L_p}\right)\sum_j (\mathcal{L}_j + L)(C_j + (L_j + L)) \\
&= O(T^{-1})\boldsymbol{x}'\boldsymbol{x}\sum_j (\mathcal{L}_j + L)(C_j + (L_j + L))
\end{aligned}$$

where the last equality holds because, by construction, each x_n is contained in exactly L/δ segments of the coverage. Since we assumed that $L^2/T \to 0$ as $T \to \infty$, we obtain the result. □

In what follows, the matrix norm $\|\boldsymbol{M}\|$ denotes the spectral norm of the matrix $\boldsymbol{M}$, i.e. $\max\{\sqrt{\lambda} : \lambda \text{ is the eigenvalue of } \boldsymbol{M}'\boldsymbol{M}\}$. If $\boldsymbol{M}$ is symmetric and nonnegative definite, by standard theory we have

$$\text{(A.2)} \qquad \|\boldsymbol{M}\| = \sup_{\|\boldsymbol{x}\|_2^2=1} \boldsymbol{x}'\boldsymbol{M}\boldsymbol{x} \qquad \|\boldsymbol{M}^{-1}\| = \left(\inf_{\|\boldsymbol{x}\|_2^2=1} \boldsymbol{x}'\boldsymbol{M}\boldsymbol{x}\right)^{-1}.$$

Lemma 2. *Under Assumptions (4.3) and (4.4), condition (4.5) implies that the spectral norm $\|\boldsymbol{\Sigma}_{t;T}\|$ is bounded in t, while condition (4.6) implies that the spectral norm $\|\boldsymbol{\Sigma}^{-1}_{t;T}\|$ is bounded in t.*

Proof. Lemma 1 implies

$$\|\boldsymbol{\Sigma}_{t;T}\| = \sup_{\|\boldsymbol{x}\|_2^2=1} \frac{\delta}{L} \sum_{p=1}^{P} \sum_{j<0} S_j\left(\frac{\nu_p}{T}\right) \sum_k \left(\sum_n x_n \psi_{j,k-n} \mathbb{I}_{n\in L_p}\right)^2 + o_T(1)$$

using Parseval formula, we have

$$\begin{aligned} &= \sup_{\|\boldsymbol{x}\|_2^2=1} \frac{\delta}{2\pi L} \sum_{p=1}^{P} \int_{-\pi}^{\pi} d\omega \sum_{j<0} S_j\left(\frac{\nu_p}{T}\right) |\hat{\psi}_j(\omega)|^2 \left|\sum_n x_n \exp(-i\omega n)\mathbb{I}_{n\in L_p}\right|^2 + o_T(1) \\ &\leqslant \operatorname*{ess\,sup}_{z,\omega} \sum_j S_j(z)|\hat{\psi}_j(\omega)|^2 \sup_{\|\boldsymbol{x}\|_2^2=1} \|\boldsymbol{x}\|_2^2 + o_T(1) = \operatorname*{ess\,sup}_{z,\omega} \sum_j S_j(z)|\hat{\psi}_j(\omega)|^2 + o_T(1) \end{aligned}$$

which is bounded by (4.5). Using (A.2) with $\boldsymbol{M} = \boldsymbol{\Sigma}_{t;T}$, the boundedness of $\|\boldsymbol{\Sigma}_{t;T}^{-1}\|$ is shown in exactly the same way. □

We will now study the approximation of $\boldsymbol{\Sigma}_{t;T}$ by $\boldsymbol{B}_{t;T}$. We allow the prediction lag h to be arbitrary, i.e. $1 \leq h \leq T-t$ and we consider the approximation of $\boldsymbol{\Sigma}_{t+h-1;T}$ by $\boldsymbol{B}_{t+h-1;T}$.

Lemma 3. *Under the assumptions of Proposition 2 and 4,*

$$\mathrm{MSPE}(\hat{X}_{t+h-1;T}, X_{t+h-1;T}) = \boldsymbol{b}'_{t+h-1}\boldsymbol{B}_{t+h-1;T}\boldsymbol{b}_{t+h-1} + \boldsymbol{b}'_{t+h-1}\boldsymbol{b}_{t+h-1}\; o_T(1)$$

and, in particular,

$$\mathrm{MSPE}(\hat{X}_{t;T}, X_{t;T}) = \boldsymbol{b}'_t\boldsymbol{B}_{t;T}\boldsymbol{b}_t + \boldsymbol{b}'_t\boldsymbol{b}_t\; o_T(1)$$

Proof. By the definition of an LSW process, we have $|w_{jk;T}|^2 = S_j(n/T) + (C_j + L_j|k - n|)O(T^{-1})$. Therefore,

$$\begin{aligned} \boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1} &= \sum_{jk} \sum_{n,m=0}^{t+h-1} b_n b_m \psi_{jk}(n)\psi_{jk}(m)|w_{jk;T}|^2 \\ &= \sum_{jk} \sum_{n,m=0}^{t+h-1} b_n b_m \psi_{jk}(n)\psi_{jk}(m) S_j\left(\frac{n+m}{2T}\right) + \mathrm{Rest}_1 \end{aligned} \tag{A.3}$$

We bound Rest_1 as follows:

$$|\,\mathrm{Rest}_1\,| \leqslant O(T^{-1}) \sum_{jk} \sum_{n,m=0}^{t+h-1} \left(\left|k - \left(\frac{n+m}{2}\right)\right| L_j + C_j\right) |b_n b_m \psi_{jk}(n)\psi_{jk}(m)|.$$

If $\mathcal{L}_j$ denotes the length of support of ψ_j, we have $0 \leqslant k-n, k-m \leqslant \mathcal{L}_j$ and so $k-(n+m)/2 \leqslant \mathcal{L}_j$ such that

$$\begin{aligned} |\,\mathrm{Rest}_1\,| &\leqslant O(T^{-1}) \sum_{jk} \sum_{n,m=0}^{t+h-1} (\mathcal{L}_j L_j + C_j)\, |b_n b_m \psi_{jk}(n)\psi_{jk}(m)| \\ &\leqslant O(T^{-1}) \boldsymbol{b}'_{t+h-1}\boldsymbol{b}_{t+h-1} \sum_j \mathcal{L}_j\,(\mathcal{L}_j L_j + C_j) = \boldsymbol{b}'_{t+h-1}\boldsymbol{b}_{t+h-1} o_T(1) \qquad \text{by assumption.} \end{aligned}$$

Finally, by Assumption (4.3), (A.3) yields the result. □

Proof of Proposition 3. The proof uses Lemmas 1 to 3 and is along the lines of Dahlhaus (1996b, Theorem 3.2(i)). The idea is to reduce the problem to the stationary situation by fixing the local time at ν_p. Then, the key point is to use the following relation between the wavelet spectrum of a *stationary* process and its classical Fourier spectrum. If X_t is a stationary process with absolutely summable autocovariance and with Fourier spectrum $f(\cdot)$, then its wavelet spectrum is given by

$$S_j = \sum_l A_{jl}^{-1} \int d\lambda\ f(\lambda)|\hat{\psi}_l(\lambda)|^2 \tag{A.4}$$

for any fixed non-decimated system of compactly supported wavelets $\{\psi_{jk}\}$. We refer to Dahlhaus (1996b, Theorem 3.2(i)) for details. □

Lemma 4. *Under the assumptions of Proposition 4, we have*

$$\boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1} = \boldsymbol{b}'_{t+h-1}\boldsymbol{B}_{t+h-1;T}\boldsymbol{b}_{t+h-1}\,(1+o_T(1))$$

Proof of Lemma 4. By Lemma 3, we have $\boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1} = \boldsymbol{b}'_{t+h-1}\boldsymbol{B}_{t+h-1;T}\boldsymbol{b}_{t+h-1} + \boldsymbol{b}'_{t+h-1}\boldsymbol{b}_{t+h-1}\ \ o_T(1)$ By Lemma 2, the inverse of $\boldsymbol{\Sigma}_{t;T}$ is bounded in T and, by standard properties of the spectral norm, we have

$$\boldsymbol{b}'_{t+h-1}\boldsymbol{b}_{t+h-1} \leqslant \boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1}\ \|\boldsymbol{\Sigma}^{-1}_{t+h-1;T}\|$$

for all sequences $\boldsymbol{b}_{t+h-1}$. The above gives

$$\boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1} \leqslant \boldsymbol{b}'_{t+h-1}\boldsymbol{B}_{t+h-1;T}\boldsymbol{b}_{t+h-1} + \boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1}\ \|\boldsymbol{\Sigma}^{-1}_{t+h-1;T}\|\ o_T(1)$$

which is equivalent to

$$\boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1} \leqslant \boldsymbol{b}'_{t+h-1}\boldsymbol{B}_{t+h-1;T}\boldsymbol{b}_{t+h-1}\Big(1 - \|\boldsymbol{\Sigma}^{-1}_{t+h-1;T}\|\ o_T(1)\Big)^{-1}$$

for large T. On the other hand, we have

$$\boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1} \geqslant \boldsymbol{b}'_{t+h-1}\boldsymbol{B}_{t+h-1;T}\boldsymbol{b}_{t+h-1}\Big(1 + \|\boldsymbol{\Sigma}^{-1}_{t;T}\|\ o_T(1)\Big)^{-1}$$

which implies the result. □

B Estimation of the local autocovariance function

In this section, we study the properties of the estimator of the local autocovariance. We first show some relevant properties of the autocorrelation function $\Psi_j(\tau)$ and the matrix $\boldsymbol{A}$ defined in (5.1).

Lemma 5. *1. The system $\{\Psi_j(\tau), j = -1, -2, \ldots\}$ is linearly independent.*

2. Denote by $\Psi(\tau)$ the wavelet autocorrelation function of a continuous *wavelet ψ, i.e.*

$$\Psi(\tau) = \int du \psi(u)\psi(u-\tau), \qquad \tau \in \mathbb{Z}.$$

We have

$$\Psi_j(\tau) = \Psi\left(2^j|\tau|\right)$$

for all $j = -1, -2, \ldots$ and $\tau \in \mathbb{Z}$.

The proof of the first result can be found in Nason *et al.* (2000, Theorem 2.13). For the proof of the second result, see, for example, Berkner and Wells (2002, Lemma 4.2).

Lemma 6. *$\sum_{j=-\infty}^{-1} 2^j \ \Psi_j(\tau) = \delta_0(\tau)$.*

Proof. Using Lemma 5 and Parseval's formula,

$$\sum_{j=-\infty}^{-1} 2^j \Psi_j(\tau) = \sum_{j=-\infty}^{-1} 2^j \Psi\left(2^j\,|\tau|\right) = \sum_{j=-\infty}^{-1} \int_{-\infty}^{\infty} d\omega\, |\hat{\psi}(2^{-j}\omega)|^2 \exp(i\omega\tau)$$

$$= \sum_{j=-\infty}^{-1} \int_0^{2\pi} d\omega \sum_{k\in\mathbb{Z}} \left|\hat{\psi}\left(2^{-j}(\omega+2k\pi)\right)\right|^2 \exp(i\omega\tau). \tag{B.1}$$

Denote by $m_0(\xi)$ the trigonometric polynomial which corresponds to the construction of wavelet ψ and its corresponding scaling function ϕ (Daubechies, 1992, Theorem 6.3.6). We may write

$$\sum_{k\in\mathbb{Z}} \left|\widehat{\psi}\left(2^{-j}\left(\omega+2k\pi\right)\right)\right|^2 = \sum_{k\in\mathbb{Z}} \left|m_0\left(2^{-j-1}\omega + 2^{-j-1}k2\pi + \pi\right)\right|^2 \left|\widehat{\phi}\left(2^{-j-1}\omega + 2^{-j-1}k2\pi\right)\right|^2$$

and, using the $2\pi k$-periodicity of m_0,

$$= \left|m_0\left(2^{-j-1}\omega + \pi\right)\right|^2 \sum_{k\in\mathbb{Z}} \left|\widehat{\phi}\left(2^{-j-1}\omega + 2^{-j-1}k2\pi\right)\right|^2$$

$$= \left|m_0\left(2^{-j-1}\omega + \pi\right)\right|^2 \sum_{k\in\mathbb{Z}} \left|m_0\left(2^{-j-2}\omega + 2^{-j-2}k2\pi\right)\right|^2 \left|\widehat{\phi}\left(2^{-j-2}\omega + 2^{-j-2}k2\pi\right)\right|^2$$

$$= \left|m_0\left(2^{-j-1}\omega + \pi\right)\right|^2 \left|m_0\left(2^{-j-2}\omega\right)\right|^2 \sum_{k\in\mathbb{Z}} \left|\widehat{\phi}\left(2^{-j-2}\omega + 2^{-j-2}k2\pi\right)\right|^2.$$

By similar transformations, we finally arrive at

$$= \left|m_0\left(2^{-j-1}\omega + \pi\right)\right|^2 \prod_{n=2}^{-j} \left|m_0\left(2^{-j-n}\omega\right)\right|^2 \sum_{k\in\mathbb{Z}} \left|\widehat{\phi}\left(\omega + k2\pi\right)\right|^2$$

$$= (2\pi)^{-1} \left|m_0\left(2^{-j-1}\omega + \pi\right)\right|^2 \prod_{n=2}^{-j} \left|m_0\left(2^{-j-n}\omega\right)\right|^2$$

$$= (2\pi)^{-1} \left|1 - m_0\left(2^{-j-1}\omega\right)\right|^2 \prod_{\ell=0}^{-j-2} \left|m_0\left(2^{\ell}\omega\right)\right|^2.$$

Using (B.1), we obtain

$$\sum_{j=-\infty}^{-1} 2^j \Psi_j(\tau) = (2\pi)^{-1} \int_0^{2\pi} \sum_{j=-\infty}^{-1} d\omega \; \exp(i\tau\omega) \left|1 - m_0\left(2^{-j-1}\omega\right)\right|^2 \prod_{\ell=0}^{-j-2} \left|m_0\left(2^\ell \omega\right)\right|^2.$$

Expanding the telescopic sum over j, we get

$$\sum_{j=-\infty}^{-1} \left|1 - m_0\left(2^{-j-1}\omega\right)\right|^2 \prod_{l=0}^{-j-2} \left|m_0\left(2^l\omega\right)\right|^2 = 1 - \lim_{j\to-\infty} \prod_{l=0}^{-j-1} \left|m_0\left(2^l\omega\right)\right|^2 = 1 - \prod_{l=0}^{+\infty} \left|m_0\left(2^l\omega\right)\right|^2.$$

Thus, we obtain

$$\begin{aligned}\sum_{j=-\infty}^{-1} 2^j \Psi_j(\tau) &= \frac{1}{2\pi} \int_0^{2\pi} d\omega \; \exp(i\tau\omega) \left\{1 - \prod_{\ell=0}^{+\infty} \left|m_0(2^\ell\omega)\right|^2\right\} \\ &= \delta_0(\tau) - \frac{1}{2\pi} \int_0^{2\pi} d\omega \; \exp(i\tau\omega) \prod_{\ell=0}^{+\infty} \left|m_0(2^\ell\omega)\right|^2. \qquad \text{(B.2)}\end{aligned}$$

Now, it remains to prove that the second term in (B.2) is equal to zero. By definition, $m_0(\omega) = 2^{-1/2} \sum_{n=0}^{2N-1} h_n e^{-in\omega}$, where $\{h_k\}_{k\in\mathbb{Z}}$ is the low pass quadrature mirror filter used in the construction of Daubechies' compactly supported continuous time wavelet ψ (Daubechies, 1992, Section 6.4). We have

$$\frac{1}{2\pi} \int_0^{2\pi} d\omega \; \exp(i\tau\omega) \prod_{l=0}^{L} \left|m_0(2^l\omega)\right|^2 = \prod_{l=0}^{L} 2^{-l} \sum_{n,m=0}^{2N-1} h_n \overline{h_m} \delta_0(n-m)$$

which clearly tends to 0 as L tends to infinity. □

Lemma 7. *Matrix $\boldsymbol{A}$ defined in (5.1) has the following property:*

$$\sum_{j,\ell=-\infty}^{-1} 2^j A_{j\ell} = 1. \qquad \text{(B.3)}$$

If, in addition, $\boldsymbol{A}$ is constructed using Haar wavelets, then

$$\sum_{\ell=-\infty}^{-1} |A^{-1}_{j\ell}| \leqslant C \cdot 2^{j/2} \qquad \text{(B.4)}$$

$$\sum_{\ell=-\infty}^{-1} A^{-1}_{j\ell} = 2^j \qquad \text{(B.5)}$$

for all $j < 0$, where C is a constant.

Proof. (B.3) is a straightforward corollary of Lemma 6. To prove (B.4), we introduce the auxiliary matrix $\boldsymbol{\Gamma} = \boldsymbol{D}'\boldsymbol{A}\boldsymbol{D}$, where $\boldsymbol{D} = \mathrm{diag}(2^{j/2})_{j<0}$ is diagonal, i.e. $\Gamma_{j\ell} = 2^{j/2} A_{j\ell} 2^{\ell/2}$.

Nason *et al.* (2000, Theorem 2.15) show that the spectral norm of $\mathbf{\Gamma}^{-1}$ is bounded for Haar wavelets. Therefore, we obtain (B.4) as

$$\sum_{\ell=-\infty}^{-1} |A_{j\ell}^{-1}| = \sum_{\ell=-\infty}^{-1} 2^{j/2} 2^{\ell/2} |\Gamma_{j\ell}^{-1}| \leqslant C \cdot 2^{j/2}$$

To prove (B.5), observe that if $X_{t,T}$ is a white noise, then its classical Fourier spectrum is $f(\lambda) = (2\pi)^{-1}$. On the other hand, white noise is an LSW process such that $\sum_j S_j \Psi_j(\tau) = \delta_0(\tau)$ which implies that $S_j = 2^j$ (Lemma 6). Due to relation (A.4), (B.5) holds for $j < 0$ because $\int d\lambda |\hat{\psi}_\ell(\lambda)|^2 = 2\pi$. $\square$

Proof of Proposition 5. By Definition 1, we have $|w_{\ell,n+k;T}|^2 = S_\ell(k/T) + (C_\ell + |n| L_\ell) T^{-1}$ and we can write

$$\begin{aligned}
\mathrm{E}\tilde{c}\left(\frac{k}{T}, 0\right) &= \sum_{j=-J}^{-1} 2^j \mathrm{E} I_j\left(\frac{k}{T}\right) = \sum_{j=-J}^{-1} 2^j \sum_{\ell=-J}^{-1} S_\ell\left(\frac{k}{T}\right) \sum_n \left(\sum_t \psi_{\ell,n+k}(t)\psi_{j,k}(t)\right)^2 + \mathrm{Rest}_T \\
&= \sum_{j=-J}^{-1} 2^j \left(\sum_{\ell=-J}^{-1} A_{j\ell} S_\ell\left(\frac{k}{T}\right)\right) + \mathrm{Rest}_T = \sum_{\ell=-J}^{-1} S_\ell\left(\frac{k}{T}\right) + \mathrm{Rest}_T
\end{aligned}$$

using (B.3). We now bound $|\,\mathrm{Rest}_T\,|$:

$$|\,\mathrm{Rest}_T\,| \leqslant T^{-1} \sum_{j=-J}^{-1} 2^j \sum_{\ell=-J}^{-1} \sum_{n=-\infty}^{\infty} (C_\ell + |n| L_\ell) \left(\sum_{t=-\infty}^{\infty} \psi_{\ell,n+k}(t)\psi_{j,k}(t)\right)^2$$

and, since the sum over t ranges from $-\infty$ to ∞, we have

$$= T^{-1} \sum_{j=-J}^{-1} 2^j \sum_{\ell=-J}^{-1} \sum_{n=-\infty}^{\infty} (C_\ell + |n| L_\ell) \left(\sum_{t=-\infty}^{\infty} \psi_{\ell,n}(t)\psi_{j,0}(t)\right)^2.$$

Now, given that the length of support of ψ_j is $\mathcal{L}_j = O(2^{-j})$, we have that $|n| \leqslant \max(\mathcal{L}_j, \mathcal{L}_l) \leqslant \mathcal{L}_j \cdot \mathcal{L}_\ell$, and we obtain

$$\leqslant T^{-1} \sum_{j=-J}^{-1} 2^j \sum_{\ell=-J}^{-1} C_\ell\, A_{j\ell}\, O(T^{-1}) + \sum_{j=-J}^{-1} 2^j\, \mathcal{L}_j \sum_{\ell=-J}^{-1} \mathcal{L}_\ell\, L_\ell\, A_{j\ell} = O(JT^{-1})$$

using (4.4) and $\sum_{j=-J}^{-1} A_{j\ell} \leqslant 2J\mathcal{L}_\ell$. To compute the expectation of $\tilde{c}\left(\frac{k}{T}, \tau\right)$, we use the same technique to obtain

$$\mathrm{E}\tilde{c}\left(\frac{k}{T}, \tau\right) = \sum_{j=-J}^{-1} \left(\sum_{\ell=-J}^{-1} A_{j\ell}^{-1} \Psi_\ell(\tau)\right) \mathrm{E} I_j\left(\frac{k}{T}\right) = \sum_{\ell=-J}^{-1} \Psi_\ell(\tau) S_\ell\left(\frac{k}{T}\right) + \mathrm{Rest}_T$$

and now $|\,\mathrm{Rest}_T\,|$ is bounded as follows:

$$|\,\mathrm{Rest}_T\,| \leqslant T^{-1} \sum_{j=-J}^{-1} \Big|\sum_{\ell=-J}^{-1} A_{j\ell}^{-1} \Psi_\ell(\tau)\Big| \sum_{m=-J}^{-1} \sum_{n=-\infty}^{\infty} (C_m + |n| L_m) \left(\sum_{t=-\infty}^{\infty} \psi_{mn}(t)\psi_{j0}(t)\right)^2$$

using $|n| \leqslant \mathcal{L}_j \cdot \mathcal{L}_m$ and $\Psi_m(\tau) \leqslant 1$ uniformly in m and τ, (B.4) implies

$$\leqslant T^{-1} \sum_{j=-J}^{-1} 2^{j/2} \sum_{m=-J}^{-1} (C_m + \mathcal{L}_j \cdot \mathcal{L}_m L_m) A_{mj} = O(T^{-1/2} J)$$

by Assumption (4.4) and since $\sum_{j<0} 2^{j/2} \mathcal{L}_j A_{jm} = O(\sum_{j<0} 2^{-j/2} A_{jm}) = O(\mathcal{L}_m \sqrt{T} \log_2 T)$. To compute the variances, we use a corollary of the Isserlis theorem for normal random variables: if $X \sim N(0, \sigma_X^2)$ and $Y \sim N(0, \sigma_Y^2)$, then $\mathrm{Cov}(X^2, Y^2) = 2\{\mathrm{Cov}(X, Y)\}^2$. We have

$$\begin{aligned} \mathrm{Var}\, \tilde{c}\left(\frac{k}{T}, 0\right) &= \sum_{j=-J}^{-1} 2^{2j} \,\mathrm{Var}\left(\sum_s X_{s,T} \psi_{jk}(s)\right)^2 \\ &\quad + 2 \sum_{j,\ell=-J, j<\ell}^{-1} 2^{j+\ell} \,\mathrm{Cov}\left((\sum_s X_{s,T} \psi_{jk}(s))^2, (\sum_t X_{t,T} \psi_{\ell k}(t))^2 \right). \end{aligned}$$

Each of the two terms can be evaluated following similar reasoning as for the computation of the expectation. Using Assumptions (4.3) and (4.4), the first term of $\mathrm{Var}\, \tilde{c}\left(\frac{k}{T}, 0\right)$ is equal to

$$\begin{aligned} 2 \sum_{j=-J}^{-1} 2^{2j} \left(\mathrm{Var} \sum_s X_{s,T} \psi_{jk}(s)\right)^2 &= 2 \sum_{j=-J}^{-1} 2^{2j} \left(\sum_{s,t} \psi_{jk}(s) \psi_{jk}(t) \,\mathrm{Cov}(X_s, X_t)\right)^2 \\ &= 2 \sum_j 2^{2j} \left(\sum_{s,t} \sum_{n,m} S_n\left(\frac{s+t}{2T}\right) \psi_{jk}(s) \psi_{jk}(t) \psi_{nm}(s) \psi_{nm}(t)\right)^2 + O(T^{-1}) \\ &= 2 \sum_j 2^{2j} \left(\sum_{s,t} c\left(\frac{s+t}{2T}, t-s\right) \psi_{jk}(s) \psi_{jk}(t)\right)^2 + O(T^{-1}) \end{aligned}$$

by (3.3). Similarly, the second term of $\mathrm{Var}\, \tilde{c}\left(\frac{k}{T}, 0\right)$ is equal to

$$4 \sum_{j,\ell=-J, j<\ell}^{-1} 2^j 2^\ell \left\{\sum_{s,t} c\left(\frac{s+t}{2T}, t-s\right) \psi_{jk}(s) \psi_{\ell k}(t)\right\}^2 + O(JT^{-1}),$$

and thus we obtain the result. The variance of $\tilde{c}\left(\frac{k}{T}, \tau\right)$ is obtained using the same scheme.□

References

Antoniadis, A. and Lavergne, C. (1995). Variance Function Estimation in Regression by Wavelet Methods. In A. Antoniadis and G. Oppenheim (Eds.), *Wavelets and Statistics* (Vol. 103, pp. 31–42). New York: Springer-Verlag.

Antoniadis, A. and Sapatinas, T. (2001). *Wavelet methods for continuous-time prediction using representations of autoregressive processes in Hilbert spaces* (Rapport De Recherche No. RR-1042). Laboratoire IMAG-LMC, University Joseph Fourier.

Bera, A. and Higgins, M. (1993). ARCH models: properties, estimation and testing. *J. Econ. Surveys*, *7*(4), 305–366.

Berkner, K. and Wells, R. (2002). Smoothness Estimates for Soft-Threshold Denoising via Translation-Invariant Wavelet Transforms. *Appl. Comput. Harmon. Anal.*, *12*, 1–24.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods* (Second ed.). Springer, New York.

Dahlhaus, R. (1996a). Asymptotic statistical inference for nonstationary processes with evolutionary spectra. In P. Robinson and M. Rosenblatt (Eds.), *Athens conference on applied probability and time series analysis* (Vol. 2). Springer, New York.

Dahlhaus, R. (1996b). On the Kullback-Leibner information divergence of locally stationary processes. *Stochastic Process. Appl.*, *62*, 139–168.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.*, *25*(1), 1–37.

Dahlhaus, R., Neumann, M. and von Sachs, R. (1999). Non-linear wavelet estimation of time-varying autoregressive processes. *Bernoulli*, *5*, 873–906.

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

Franses, P. H. (1998). *Time series models for business and economic forecasting.* Cambridge University Press.

Grillenzoni, C. (2000). Time-varying parameters prediction. *Ann. Inst. Statist. Math.*, *52*, 108–122.

Härdle, W., Spokoiny, V. and Teyssière, G. (2000). *Adaptive estimation for a time inhomogeneous stochastic-volatility model* (SFB Discussion Paper No. 6/00). Berlin: Humboldt University.

Kress, R. (1991). *Numerical Analysis.* New York: Springer.

Ledolter, J. (1980). Recursive estimation and adaptive forecasting in ARIMA models with time varying coefficients. In *Applied Time Series Analysis, II (Tulsa, Okla.)* (pp. 449–471). New York-London: Academic Press.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, *11*, 674–693.

Mallat, S., Papanicolaou, G. and Zhang, Z. (1998). Adaptive covariance estimation of locally stationary processes. *Ann. Statist.*, *26*, 1–47.

Mélard, G. and Herteleer-De Schutter, A. (1989). Contributions to the evolutionary spectral theory. *J. Time Ser. Anal.*, *10*, 41–63.

Mercurio, D. and Spokoiny, V. (2000). *Statistical inference for time-inhomogeneous volatility models* (Preprint No. 583). Berlin: WIAS.

Mikosch, T. and Stărică, C. (1999). *Change of structure in financial data, long-range dependence and GARCH modelling* (Technical Report). University of Groningen.

Nason, G. and Silverman, B. (1995). The Stationary Wavelet Transform and some Statistical Applications. In A. Antoniadis and G. Oppenheim (Eds.), *Wavelets and Statistics* (Vol. 103, pp. 281–299). New York: Springer-Verlag.

Nason, G. and von Sachs, R. (1999). Wavelets in time series analysis. *Phil. Trans. R. Soc. Lond.* A, *357*, 2511–2526.

Nason, G., von Sachs, R. and Kroisandt, G. (2000). Wavelet processes and adaptive estimation of evolutionary wavelet spectra. *J. R. Stat. Soc. Ser.* B, *62*, 271–292.

Ombao, H., Raz, J., Guo, W. and von Sachs, R. (2002). The SLEX Model of a non-stationary random process. *Ann. Inst. Statist. Math.*, *54*. (To appear)

Ombao, H., Raz, J., von Sachs, R. and Malow, B. (2001). Automatic Statistical Analysis of Bivariate Nonstationary Time Series. *J. Amer. Statist. Assoc.*, *96*, 543–560.

Philander, S. (1990). *El Niño, La Niña and the Southern Oscillation.* San Diego: Academic Press.

Priestley, M. (1965). Evolutionary spectra and non-stationary processes. *J. R. Stat. Soc. Ser.* B, *27*, 204–237.

Simonoff, J. (1996). *Smoothing Methods in Statistics.* New York: Springer-Verlag.

von Sachs, R. and MacGibbon, B. (2000). Nonparametric Curve Estimation by Wavelet Thresholding with Locally Stationary Errors. *Scand. J. Statist.*, *27*, 475–499.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models* (Second ed.). Springer Verlag.