UNIVERSITÉ CATHOLIQUE DE LOUVAIN
INSTITUT DE STATISTIQUE
Voie du Roman Pays, 20
B-1348 Louvain-la-Neuve
Belgique

ADAPTIVE METHODS FOR MODELLING, ESTIMATING AND FORECASTING LOCALLY STATIONARY PROCESSES

Membres du jury:

Prof. Rainer Dahlhaus
Prof. Irène Gijbels
Prof. Guy P. Nason
Prof. Jean-Marie Rolin
Prof. Rainer von Sachs *(Promoteur)*
Prof. Léopold Simar *(Président du Jury)*

Thèse présentée en vue de
l'obtention du grade de
Docteur en Sciences
(orientation statistique) par :

Sébastien Van Bellegem

Louvain-la-Neuve
Défense du 16 décembre 2003

Ackowledgements

I would like to express my deep gratitude to Professor Rainer von Sachs for guiding me in the world of nonparametric statistics and time series, for his willingness to share his ideas in research problems and for the energy he put in advising my thesis work.

I'm also grateful to the readers of the thesis for their helpul comments and remarks. These, I believe, helped me to improve the presentation of my results. I'm especially grateful to Professor Guy P. Nason (Bristol, UK) and to Professor Rainer Dahlhaus (Heidelberg, Germany) for welcoming me in their departments for scientific visits during my PhD. I would like to thank them for their kindness and generosity. These visits were fruitful: Chapter 2 of this thesis was prepared in Heidelberg, and Chapter 6 was initiated in Bristol. This last chapter is a joint work with Piotr Fryźlewicz. I thank him very much for all the interesting discussions that were especially useful in the begining of my PhD.

Finally, thanks to everyone who gave their valuable time, skills and enthusiasm during these last years. In particular to the members of the Institut de statistique (academic, scientific and administrative staff). And, of course, to Marie-Pierre for her continuous support and patience.

*

This research was supported by the National Fund for Scientific Research (FNRS), Belgium, by the "Fonds Spéciaux de Recherche" from the Université catholique de Louvain, by "Projet d'Action et de Recherche Concertes" (No. 98/03-217) from the Belgian government and by the IAP research network No. P5/24 of the Belgian State (Federal Office for Scientific, Technical and Cultural Affairs).

Contents

Notations and abbreviations

List of symbols

Var	Variance	11
$\mathrm{Var}(\,\cdot \mid \cdot\,)$	Conditional variance	11
$\boldsymbol{x}$	Sequence $(x_1, \ldots, x_t)$	190
$\boldsymbol{X}_{t;T}$	Random vector $(X_{1;T}, \ldots, X_{t;T})$	189
$\hat{X}_t$	Prediction of X_t	182
ξ_{jk}	Random orthonormal increment sequence	95
W_t	White noise process	154
$W^{s,p}$	Sobolev space	218
$\wp(\mathcal{R})$	Set of subsets of $\mathcal{R}$	117
$\mathbb{Z}$	Set of integers	42

List of abbreviations

ACW	Autocorrelation wavelet function	90
AIC	Akaike information criterion	25
ARMA	Autoregressive moving average process	12
Cov.	Coverage	27
CUSUM	Cumulative sum	15
CWP	Corrected wavelet periodogram	103
DailySR	Daily stock returns	20
ECG	Electrocardiogram	170
EGARCH	Exponential GARCH process	30
ESD	Evolutionary spectral density	46
EWS	Evolutionary wavelet spectrum	97
Ex-DM-BP	Exchange rate Deutsche Mark/British Pound	20
Ex-US-BP	Exchange rate US dollar/British Pound	20
For. hor.	Forecasting horizon	27
GARCH	Generalized autoregressive heteroskedasticity process	11
iid	Independent, identically distributed	120
Lip	Lipschitz continuous	11
LSW	Locally stationary wavelet process	95
MAD	Mean-absolute deviation	26
MAPE	Mean-absolute percentage error	26
MLPI	Median of the length of some prediction intervals	26
MRA	Multiresolution analysis	82
MSPE	Mean square prediction error	182
Nas-100	Log-returns of the Nasdaq-100 index	20
NasFin-100	Log-returns of the Nasdaq Financial-100 index	20

PC	Piecewise continuous	11
PSP	Post-sample prediction	21
RMSE	Root-mean-square error	26
tm-AR	Time-modulated autoregressive process	33
tm-ARMA	Time-modulated ARMA process	11
tm-GARCH	Time-modulated GARCH process	11
tm-WN	Time-modulated White Noise process	11
tv-ARMA	Time-varying ARMA process	41
WFT	Windowed Fourier transform	172

List of Tables

List of Figures

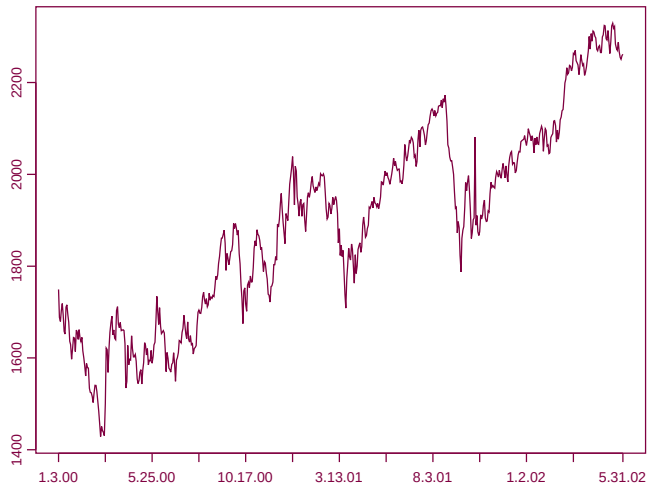
Introduction

This work is concerned with data coming in the form of a univariate, discrete-time stochastic process X_t ($t = 0, 1, 2, \ldots$). We focus on the analysis of its covariance structure, and therefore we assume that the process is zero-mean.

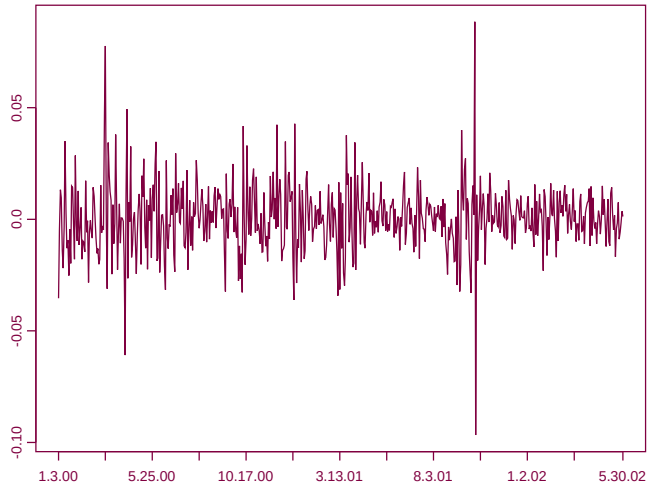
Zero-mean processes arise, for example, when the global trend has been removed from the data. Trend estimation is a well studied problem in the literature, and we refer to [51, 52, 97, 120] for some existing methods. Sometimes, the global trend of a time series can be removed without computing an estimator explicitely. In that case, a first-order difference is applied on the observed process, that is we compute $X_t - X_{t-1}$, and the resulting process is zero-mean. It is well-known that this phenomenon arises for a lot of economic indices [7]. An example is given in Figure 1(a), which represents the values of the Nasdaq-100 index from January 3, 2000 through May 31, 2002. This index includes hundred of the largest non-financial domestic and international companies listed on the Nasdaq National Market. It is clear that this process contains a trend, which can be removed if we compute the first-order difference of the logarithm of the Nasdaq-100 series (the log-return index), see Figure 1(b). The resulting zero-mean process still contains valuable information for the analyst, in terms of the volatility of the time series.

In time series analysis, most existing models assume that the zero-mean process X_t is covariance stationary. This means that the covariance between X_s and X_t depends only on the lag $|s-t|$. This assumption is very useful in order to have some estimators for the autocovariance structure of the process with good statistical properties, such as consistency, efficiency or central limit theorems (see Brockwell and Davis [15] for a review).

However, many time series in applied science are *not* covariance sta-



(a) 604 daily observations of the index



(b) The log-returns of the index.

Figure 1: The Nasdaq-100 index includes 100 of the largest non-financial domestic and international companies listed on the Nasdaq National Market.

tionary and show a *time-varying second-order structure*. That is, variance and covariance can change over time. For instance, the log-returns showed in Figure 1(b) are likely to have an inhomogeneous variance. By this, we mean that the variance of the process is clearly not constant over time. In this thesis, we apply a test of covariance stationarity to these data, and it confirms this observation. Many other examples may be found in economics, as is showed in Chapter 1 of this thesis. But this lack of covariance-stationarity has also been observed in many other fields of the applied science, such as biomedical time series (see Chapter 5 of this thesis, or [83, 87]), seismology [103] or meteorology (see Chapter 6 of this thesis, or [80]), to name but a few.

Gradually, more and more attention has been paid to this challenging problem on how to model such processes with an evolutionary autocovariance structure. Among the pioneers, we can cite the work of Loynes [65], Page [90], Priestley [94], Silverman [104]. For instance, this last author proposed in 1957 the approximation

$$\mathrm{Cov}(X_s, X_t) \approx m\left(\frac{s+t}{2}\right) c(s-t)$$

i.e. the covariance behaves locally as a typical stationary autocovariance but then varies from place to place depending of the midpoint between s and t. As in Silverman's definition, each model on nonstationary covariance has to define explicitely its departure from stationarity. However, from a statistical viewpoint, many questions remain. For instance, with this lack of an invariant second-order structure, how can we estimate the time-varying covariance with a high accuracy? Even if we add some regularity assumptions on the function m, a serious problem here is that we cannot build an asymptotic theory for the estimation of m. Consequently, the standard statistical properties like consistency, efficiency or central limit theorems cannot be used to measure and compare the quality of different estimators.

In the last decades, many authors worked on this problem of modelling and estimating an evolutionary autocovariance structure [44, 46, 54, 55, 69, 71, 73]. A decisive idea was introduced recently by Dahlhaus [23] with his new concept of "local stationarity". This concept allows the modelling of a time-varying autocovariance structure which can be estimated rigorously. By this, we mean that an appropriate asymptotic theory can be developed and the usual statistical properties of estima-

tors may be derived. This simple and elegant idea is reviewed in the beginning of Chapter 1.

The thesis contributes to the development of this approach of nonstationarity. Herewith, by "nonstationarity", we always refer to a zero-mean process with a possibly time-varying autocovariance structure. Our contributions concern both the theoretical and empirical analysis of nonstationary processes. We bring new results in terms of modelling and estimating the autocovariance of these processes. We also address the problem of how to forecast nonstationary data. The proposed estimation and forecasting procedures are adaptive, in the sense that all the parameters needed in the procedure are chosen in a data-driven way. Applications are provided on economic, biomedical and meteorological time series.

The first chapter presents a simple model for nonstationarity, where only the variance is time-varying. The aim of this chapter is twofold. First, it makes the reader familiar with the fundamental concepts of the whole thesis. The notion of "local stationarity" is discussed, and the fundamental problems of the analysis of nonstationary data are presented on a model which is particularly simple to understand. These problems concern the modelling, the estimation and the forecasting of nonstationary data. The second aim of this chapter is to show that this simple model satisfactorily explains the nonstationary behaviour of several economic data sets, among which are the U.S. stock returns and exchange rates. The nonstationary behaviour of the economic processes has been neglected in the past, and very often estimation and forecasting procedures based on the assumption of stationarity are applied without checking the covariance-stationarity of the data. Our major contribution here is to show that very often, the assumption of stationarity is rejected by standard testing procedures, and we provide a simple model for explaining this nonstationary behaviour. This chapter is based on [111].

In the second chapter, we study more complex semiparametric models, where not only the variance is evolutionary. A typical example of these models is given by ARMA(p, q) models with time-varying coefficients. Our aim is to fit such semiparametric models to some nonstationary data. Our data-driven estimator is constructed from a minimisation of a penalised contrast function, where the contrast function is an approximation to the Gaussian likelihood of the model. The theoretical

performance of the estimator is analysed via non asymptotic risk bounds for the quadratic risk. In our results, we do not assume that the observed data follow the semiparamatric structure, that is our results hold in the misspecified case.

The third chapter introduces a fully nonparametric model for local nonstationarity. This model is a wavelet-based model of local stationarity which enlarges the class of models defined by Nason et al. [83]. Our main contribution here is for modelling: we allow the evolutionary autocovariance to change very suddenly in time. This is in contrast with the work of Nason et al. [83], who model a smoothly varying autocovariance. This extension is not only proposed in order to work in a general setting. It is also crucial if one wishes to model time series with intermittent phenomena, such as transients followed by regions of smooth behaviour. A notion of time-varying "wavelet spectrum" is uniquely defined as a wavelet-type transform of the autocovariance function with respect to so-called "autocorrelation wavelets". This leads to a natural representation of the autocovariance which is localised on scales. We also provide some useful mathematical properties of the autocorrelation wavelet system. Chapter 3 is based on [39] and [112].

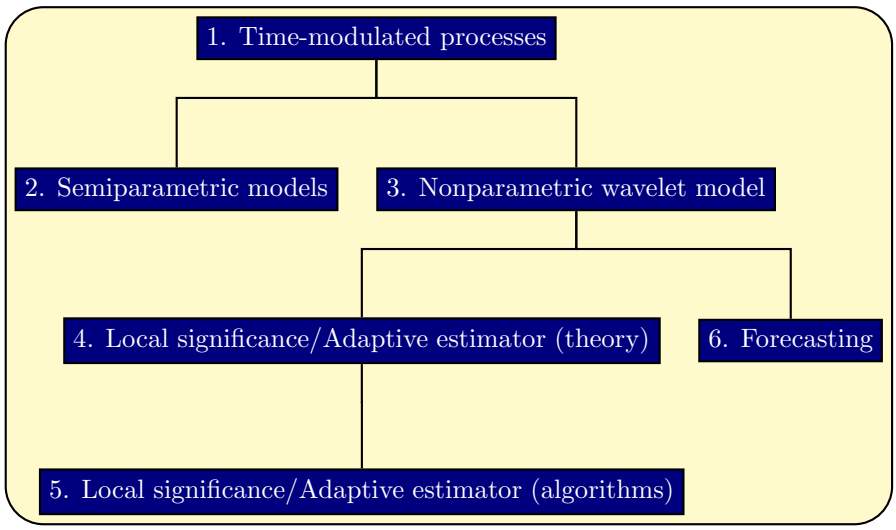
Similarly to the classical theory of stationary time series, a wavelet periodogram can be defined as a preliminary estimator of the wavelet spectrum. One particularly interesting question is to test the significance of the coefficients of the wavelet periodogram. This question has been presented in Nason et al. [83] as a challenging problem, with potentially important practical applications. In Chapters 4 and 5, we address this problem. We derive some theoretical properties of our test of significance, including a discussion on its consistency, its power, and its local alternative. The test rule is based on a non-asymptotic result on the deviations of a functional of the periodogram. This key result also allows to derive a new pointwise adaptive estimator of the wavelet spectrum. Theoretical properties of this new estimator are also presented in Chapter 4. This chapter is based on [112].

However, the use of the test of significance and the pointwise adaptive estimators is not straightforward and necessitates the study of some appropriate algorithmic procedures. This is provided in the next Chapter 5, where we give a description of a full algorithm for each procedure. These algorithms are evaluated on simulated nonstationary processes, and applied on a case study in biostatistics. In this chapter, we also

derive a new test of covariance stationarity. This test is illustrated on another case study in biostatistics, and compared with the wavelet-based test of stationarity of von Sachs and Neumann [99]. This chapter is based on [113]

Finally, Chapter 6 addresses the problem how to forecast the general nonstationary process introduced in Chapter 3. We present a new predictor 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. Then we apply the prediction algorithm to a meteorological data set. This chapter is based on [39] and [110].

Diagram of the thesis



CHAPTER 1

Modelling and forecasting economic time series with unconditional time-varying variance

1.1 Motivation

To forecast economic time series, many analyses are based on the assumption that the probabilistic properties of the underlying process are time-invariant. Even if this assumption is very useful in order to construct simple predictors, it seems not to be the best strategy in practice. Indeed, taking into account structural changes of parameters may lead to better forecasting performance because it is sometimes more adequate to capture some aspects of the real world [18, Section 7.4]. More surprisingly, some stylised facts, as the long range dependence of the absolute returns of financial time series, can even be explained by a parameter change of the model [77]. In terms of forecasting, Swanson and White [108] conclude that models allowing parameter variability can show better accuracy and adaptability than constant models on macroeconomic data. These empirical and theoretical works suggest a model with evolutionary parameters, reflecting the evolution of the economy over time. An important consequence of this observation is the lack of stationarity of the data, which means that the *unconditional* moments of the time

series *can vary over time*.

This evolution of unconditional moments has been statistically tested on various economic processes. For instance, Pagan and Schwert [88, 89] develop statistical tests rejecting the hypothesis that the monthly U.S. stock returns are covariance stationary, as long as the period of Great Depression (1929–1939) is included in the series. Loretan and Phillips [63] confirm this conclusion for stock returns, but also for a set of exchange rate data. Similar results have been found by Los [64] on short series of weekly price indeces of Asian stock markets. Other approaches for testing structural changes in the variance of economic time series are reviewed in Hansen [47].

In this introductory chapter, we go one step further and study how this nonstationary behaviour can be modeled. We focus on a very simple model of nonstationarity with an unconditional variance evolving with time. This model includes the so-called "time-modulated processes" defined in Section 1.2 below. In this section, we also derive some basic properties of these processes. Estimation and forecasting procedures are described in Section 1.3. Section 1.4 is devoted to the practical evaluation of these nonstationary processes on several financial time series among which are the U.S. stock returns and exchange rates. We show that our simple and meaningful model of nonstationarity provides a satisfactory explanation power of the nonstationary behaviour of the observed data.

In Section 1.5, a comparison between our forecasts and standard ARCH-type models is provided. However, the standard comparison tests of forecast accuracy cannot be used in our context, as they usually work under the maintained assumption that the process is variance-stationary. Hence, we show that a generalisation of the Meese and Rogoff's test [72] may be used for comparing the mean forecasts, while the Christoffersen's test [17] may be used to compare the interval forecasts.

Finally, Section 1.6 presents some concluding remarks including the essential points for the understanding of the whole thesis.

1.2 Modelling evolutionary variance

The most simple nonstationary model consists of a second-order stationary process modulated by a deterministic time-varying variance. If $Y_t, t = 0, 1, 2, \ldots$, is a zero mean stationary process, a very simple nonstationary model is given by $X_t = \sigma_t Y_t$, where σ_t is a deterministic

time-varying function which is strictly positive. As we suppose the process Y_t to be zero mean and stationary, the nonstationarity of X_t is only explained by its evolving unconditional variance.

With this model of variance nonstationarity, it may seem contradictory to construct a forecasting theory, since a predictor exploits generally an invariance structure in the unconditional moments of the process. This problem is overcome if we add regularity assumptions on the deterministic function σ_t. For instance, we can impose that σ_t is a piecewise constant function. More generally, we can assume that σ_t is nearly constant along intervals of a certain length τ. Using this regularity assumption, we can estimate and extrapolate the deterministic variance to build a predictor.

However, this approach is not satisfactory since it implicitly imposes that the function σ_t is estimable only using τ observations. In this framework, when the length of the data set increases, no improvement is possible in the estimation of σ_t over this interval of length τ. This implies that asymptotic considerations can not be used in the statistical inference of such process. This is a substantial drawback, because the usual statistical properties of estimators such as consistency, efficiency or central limit theorems cannot be used to measure and to compare the quality of different estimators.

To overcome this (theoretical) problem, Dahlhaus [23] introduced a concept of "local stationarity" in a general context of covariance evolution. In our situation, this concept is as follows: Suppose we observe the series from time 0 up to $T-1$ (T observations). The local stationarity assumption postulates the existence of a deterministic function $\sigma(z)$ defined for $z \in [0,1)$ such that the approximation $\sigma_t \approx \sigma(t/T)$ holds in an appropriate way that we will define below. In this approach, two scales of time are defined: The *observed time*, which is the usual scale of time $0, \ldots, T-1$, and the *rescaled time* defined on the interval $[0,1)$. The deterministic function $\sigma(z)$ is defined on the rescaled time. There exists a mapping between these two scales of time, and since this mapping depends on the sample size T, the resulting nonstationary process is doubly indexed:

$$X_{t,T} = \sigma\left(\frac{t}{T}\right) Y_t, \tag{1.1}$$

where Y_t is a zero-mean stationary process with unit variance. $X_{t,T}$ is

called a *time-modulated process*. The regularity assumptions are now made on the function $\sigma(z)$ defined on $[0,1)$. Due to the mapping between $0,\dots,T-1$ and $[0,1)$, the estimation of $\sigma(z)$ becomes a standard statistical problem: For instance if $\sigma(z)$ is constant on an interval of length $\tau < 1$ in the rescaled time, then it may be estimated using $\tau \cdot T$ observed data in the real time (see Figure 1.1).

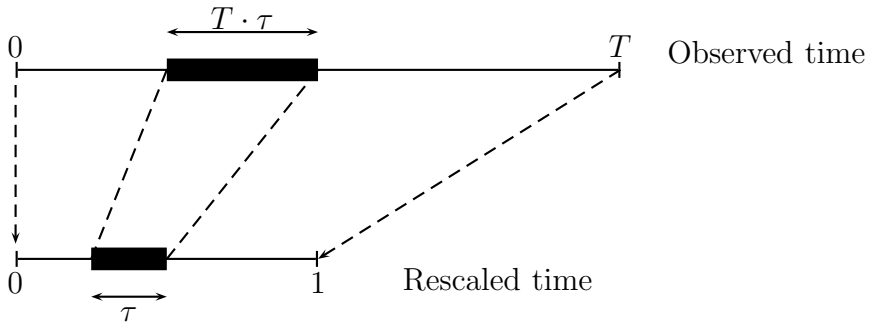


Figure 1.1: The rescaled time principle.

Consequently, if T increases, we get more observations to estimate $\sigma(z)$ on τ. Estimation of $\sigma(z)$ is then parallel to the estimation of a regression function in a nonparametric regression problem.

An important consequence of the rescaled time is the interpretation of "asymptotics". When T tends to infinity, we get more information on the local structure of $\sigma^2(z)$ in the rescaled time, because the mapping defines a finer grid in the rescaled time. However, it does *not* mean that we look into the future, because the rescaled time has a fixed bounded support $[0,1)$. To understand this point, an analogy with the spectral analysis of stationary time series may help. Indeed, the spectrum of a stationary process is defined on the interval $[-\pi,\pi]$. When we observe T data ($T < \infty$), the spectrum is not identifiable on $[-\pi,\pi]$ because we can only estimate the covariance of the process up to lag $T-1$, and then the lowest frequencies of the spectrum are not identifiable. When T increases, we get more information on the spectrum on $[-\pi,\pi]$, and this spectrum is uniquely defined only *asymptotically* [21]. The rescaled time framework is analogous, in the sense that "T tends to infinity" now means that we observe finer details in the rescaled time.

In this chapter, we study model (1.1) in some particular cases, corresponding to the specification of Y_t and $\sigma(z)$. In Section 1.4, we consider

tm-WN. Time-modulated White Noise processes, which are processes (1.1) where Y_t is a White Noise process;

tm-ARMA. Time-modulated ARMA processes, which are processes (1.1) where Y_t is a general ARMA process;

In addition, we also consider the following model in Section 1.5:

tm-GARCH. Time-modulated GARCH processes, which are processes (1.1) where Y_t is a general GARCH process.

Another important specification is the regularity assumption made on the deterministic function $\sigma(z)$. In this chapter, we study two cases of regularity:

PC is the case where $\sigma^2(z)$ is modeled by a piecewise constant function, with a finite number of jumps,

Lip is the case where $\sigma^2(z)$ is a continuous function without jump. In this case, $\sigma^2(z)$ is modeled as a Lipschitz continuous function, i.e. $|\sigma^2(z) - \sigma^2(z_0)| \leqslant C \cdot |z - z_0|$ for all $0 < z_0, z < 1$.

In addition, we assume in the two cases that the variance is bounded away from zero, i.e. there exists $\delta > 0$ such that $\sigma(z) \geqslant \delta$ uniformly in z.

Let us now derive some properties of these processes. We first consider the case of a tm-WN process $\{X_{t,T}\}$. Our definition implies that the expectation and the conditional expectation of $X_{t,T}$ is zero, which we denote by

$$\mathrm{E}\left(X_{t,T}\right) = \mathrm{E}\left(X_{t,T}|\mathcal{F}_{t-1,T}\right) = 0$$

where $\mathcal{F}_{t-1,T}$ stands for the set of the observed values $X_{0,T}, \ldots, X_{t-1,T}$. Concerning the (conditional) variance, we get

$$\mathrm{Var}\left(X_{t,T}\right) = \mathrm{Var}\left(X_{t,T}|\mathcal{F}_{t-1,T}\right) = \sigma^2\left(\frac{t}{T}\right)$$

(recall that the variance of the stationary process Y_t in equation (1.1) is assumed to be one). The covariance and conditional covariance of

a tm-WN are trivially zero. A more interesting property concerns the kurtosis κ_2:

$$\kappa_2 := \frac{\mathrm{E}X_{t,T}^4}{\left(\mathrm{E}X_{t,T}^2\right)^2} = \frac{\mathrm{E}Y_{t,T}^4}{\left(\mathrm{E}Y_{t,T}^2\right)^2} \tag{1.2}$$

and it follows that *the kurtosis of a tm-WN model is the kurtosis of its White Noise*. This simple property has a particular importance in the modeling of economic series, which are known to be leptokurtic. This stylised fact and equation (1.2) indicate that the distribution of Y_t in equation (1.1) is potentially leptokurtic. Consequently, we allow the white-noise process Y_t to be non Gaussian and to follow a possible very general distribution.

These results are easily extended to the case of tm-ARMA processes. Consider equation (1.1) where Y_t is the stationary and invertible ARMA(p, q) process

$$Y_t - \phi_1 Y_{t-1} - \ldots - \phi_p Y_{t-p} = Z_t + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}$$

where $\{Z_t\}$ is a zero-mean white noise process such that $\mathrm{Var}\, Y_t = 1$. Direct considerations yield

$$\mathrm{E}X_{t,T} = 0 \ , \tag{1.3}$$

$$\begin{aligned}&\mathrm{E}\left(X_{t,T}|\mathcal{F}_{t-1,T}\right)\\ &\qquad = \sigma\left(\frac{t}{T}\right)\left\{\phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \theta_1 Z_{t-1} + \ldots + \theta_q Z_{t-q}\right\}\end{aligned}$$

and

$$\begin{aligned}&\mathrm{Var}\, X_{t,T} = \sigma^2\left(\frac{t}{T}\right) \ ,\\ &\mathrm{Var}\left(X_{t,T}|\mathcal{F}_{t-1,T}\right) = \mathrm{Var}\left(Y_t|\mathcal{F}_{t-1}\right)\cdot\sigma^2\left(\frac{t}{T}\right).\end{aligned}$$

In this last expression, $\mathrm{Var}(Y_t|\mathcal{F}_{t-1})$ is replaced by the classical formula for the conditional variance of a stationary ARMA process (again, note that we have supposed the unconditional variance of Y_t to be 1). In the particular case of an MA(q) process, this conditional variance is

$$\mathrm{Var}\left(Y_t|\mathcal{F}_{t-1}\right) = (1 + \theta_1^2 + \ldots + \theta_q^2)^{-1}.$$

In the case where Y_t is an AR(p) process, this conditional variance is

$$\mathrm{Var}\,(Y_t|\mathcal{F}_{t-1}) = 1 - a_1^2 - \ldots - a_p^2.$$

The case where Y_t is a mixed model with both an AR and a MA part is formally more complicated to be written, but follows from the general expression for the autocorrelation function of these processes (see Brockwell and Davis [15] for instance).

A similar property holds for the kurtosis: The kurtosis of a time-modulated ARMA process is the kurtosis of the constituting ARMA process, and then the distribution of the process $\{Z_t\}$ can usefully be non Gaussian. This allows the tm-ARMA process to be leptokurtic, in accordance with one of the stylised facts of economic processes.

We now consider the case of time-modulated GARCH processes. In that case, the model (1.1) is defined with a GARCH process for Y_t. Recall that Y_t is a GARCH(p, q) process if $Y_T|\mathcal{F}_{t-1}$ has a normal distribution $\mathcal{N}(0, \sigma_t^2)$, where the conditional variance σ_t^2 follows the following ARMA-type representation [12, 36]:

$$\sigma_t^2 = \alpha_0 + \alpha_1 Y_{t-1}^2 + \ldots + \alpha_p Y_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \ldots + \beta_q \sigma_{t-q}^2$$

for some parameters $\alpha_0 > 0$, $\alpha_i \geqslant 0$ ($i = 1, \ldots, p$) and $\beta_j \geqslant 0$ ($j = 1, \ldots, q$). The unconditional variance of Y_t is given by

$$\mathrm{Var}(Y_t) = \alpha_0 \left(1 - \alpha_1 - \ldots - \alpha_p - \beta_1 - \ldots - \beta_q\right)^{-1}$$

provided that $\alpha_1 + \ldots + \alpha_p + \beta_1 + \ldots + \beta_q < 1$. This imposes that $\alpha_0 = (1 - \alpha_1 - \ldots - \alpha_p - \beta_1 - \ldots - \beta_q)$ since we are assuming that this variance is equal to 1. Then, the conditional variance of the time-modulated GARCH process is given by

$$\mathrm{Var}\,(X_{t,T}|\mathcal{F}_{t-1,T}) = \sigma^2 \left(\frac{t}{T}\right) \sigma_t^2 \ ,$$

and a similar property holds for the kurtosis: the kurtosis of a time-modulated GARCH process is the kurtosis of the constituting GARCH process, and these processes are known to be leptokurtic [7].

In our results, we also consider an extension of the GARCH model given by the EGARCH model. The motivation of this model is to model a conditional variance σ_t which is not symmetric in the lagged Y_t's.

Indeed, Nelson [84] suggested that a symmetric conditional variance function may be inappropriate for modelling the volatility of returns or stocks because it cannot represent a phenomena known as the "leverage effect, which is a negative correlation between volatility and past returns. More specifically, he proposed the model

$$\log(\sigma_t) = \alpha_0 + \sum_{i=1}^{p} \alpha_i g(Y_{t-i}) + \sum_{i=1}^{q} \beta_i \log(\sigma_{t-i})$$

where

$$g(Y_t) = \theta Y_t + \gamma\{|Y_t| - \mathrm{E}|Y_t|\}.$$

This specification of the conditional variance is known as exponential GARCH (EGARCH). The sequence $g(Y_t)$ is independent with mean zero and constant variance. Therefore, EGARCH represents a linear ARMA model for $\log(Y_t)$ with innovations $g(Y_t)$.

1.3 Estimation and forecasting

In the scope of the present chapter, we want to study how time-modulated processes explain the nonstationary behaviour of some data sets. This section introduces all the tools we need for this goal. First, we recall two tests for stationarity studied in the literature. Then, we propose three different estimators of the local variance function from data. Finally, we address the problem of how to forecast time-modulated processes, and propose a way to construct prediction intervals in practice.

1.3.1 Testing for covariance stationarity

In what follows, we will use tests of covariance stationarity. Let us briefly recall two of the tests presented in Pagan and Schwert [89].

The first test is called *post-sample prediction test*. Suppose we observe the zero-mean process $X_{0,T}, \ldots, X_{T-1,T}$ and split the time axis by $T = T_1 + T_2$ with $T_1 = T_2$. If we want to test the hypothesis that the variance on $X_{0,T}, \ldots, X_{T_1-1,T}$ is equal to the variance on $X_{T_1,T}, \ldots, X_{T-1,T}$, a suitable test statistic is

$$\hat{\tau} = \hat{\sigma}_1^2 - \hat{\sigma}_2^2$$

where $\hat{\sigma}_i^2$ is the sample variance on the ith segment. Under the null hypothesis, the distribution of $T_1^{1/2}\hat{\tau}$ is asymptotically normal if $X_{t,T}^2$ is

a stationary process with autocovariance γ_j [50]:

$$T_1^{1/2}\hat{\tau} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 2\nu) \tag{1.4}$$

as T tends to infinity, where

$$\nu = \gamma_0 + 2\sum_{j=1}^{\infty} \gamma_j \ .$$

ν is estimated using the kernel-based estimate

$$\hat{\nu}_\ell = \hat{\gamma}_0 + 2\sum_{j=1}^{\ell} \left(1 - \frac{j}{\ell+1}\right) \hat{\gamma}_j$$

where $\hat{\gamma}_j$ is the jth serial covariance of X_t^2 and ℓ is a truncation number. A discussion on this estimator can be found in Newey and West [86], where a consistency result is established when $\ell = \ell(T)$ tends to infinity with T and is such that $\ell(T) = O(T^{1/4})$. Discussions about the choice of ℓ may be found in Phillips [92] and White and Domowitz [118], who proposed to first investigate the decay of the sample autocorrelations $\hat{\gamma}_j$. In that case, ℓ is selected equal to the lag where $\hat{\gamma}_j$ becomes non significant.

Note that the post-sample prediction test crucially depends on the time point where we split the series into two parts. As in practice this time point is arbitrary, we recall a second test for covariance stationarity, the *CUSUM test*. This test does not require to split the time series into two parts. Define

$$\psi(r) = \frac{1}{\sqrt{T\nu}} \sum_{t=1}^{[Tr]} \left(X_{t,T}^2 - \hat{\sigma}_T^2\right) \tag{1.5}$$

where $0 < r < 1$ and $\hat{\sigma}_T^2$ is the classical variance estimate over the whole segment of length T. This test compares the global variance estimate with the partial sum of the squared process (recall that we assume the process to be zero-mean). If the $X_{t,T}$ obey the moment and mixing conditions in Phillips [92], then Lo [62] proves that, under the null, $\psi(r)$ converges in distribution to a Brownian bridge. This means that the probability $\Pr(\psi(r) < c)$ is equal to the probability that an $\mathcal{N}(0, r(1-r))$ random variable is less than c.

Remark 1.1. These two tests of stationarity come from the standard literature in econometrics. In this thesis, a new test of stationarity is developed in Chapter 5 below. ◊

1.3.2 Estimation of the unconditional variance

The unconditional variance function $\sigma^2(z)$, defined in the rescaled time $z \in [0,1)$, is a function which is piecewise constant PC or continuous Lip.

In the Lip case, the unconditional variance is estimated as follows. Consider the squared process $X_{t,T}^2$. This process is clearly an unbiased estimate of $\sigma(t/T)$ but it suffers from a large variability (see Proposition 1.1 below for a precise quantification of this phenomenon). Then we smooth it using a kernel estimator.

In the PC case, we proceed as follows. Suppose the variance function has no structural break (jump) in the interval $I = [z_0, z_1]$, $0 < z_0 < z_1 < 1$. A natural estimate of $\sigma(z)$ for $z \in I$ is the empirical variance

$$\hat{s}_T := \frac{1}{|z_1 - z_0|T} \sum_{t=[z_0T]}^{[z_1T]} X_{t;T}^2, \tag{1.6}$$

where $[\,\cdot\,]$ denotes the integer part of a real number. The resulting estimate is an unbiased and consistent estimator of $\sigma^2(z)$ on I (see Proposition 1.2 below). However, we have to detect the structural breaks, and for this we propose to follow one of these two procedures:

1. Consider the random variable

$$s(t) = \left|\hat{\sigma}_{t+1}^2 - \hat{\sigma}_t^2\right| \qquad t = 0, \ldots, T-2, \tag{1.7}$$

where $\hat{\sigma}_t^2$ denotes the sample stationary variance computed on $\{X_{0,T}, \ldots, X_{t,T}\}$. We decide that there is a structural break in the variance at $t = t_0$ if $s(t_0)$ is larger than a certain threshold λ. In practice, we consider a data-dependent threshold: We choose λ such that 3 % of the realisations of $s(t)$ are larger than λ. Thus, on average, 3 % of the data are considered as structural breaks. Other numerical choices for λ could be taken but, as we shall see later, this choice leads to good results in practice.

2. We can also use the CUSUM test to decide if there is a structural break in the unconditional variance. Fixing the level of the CUSUM test at α, we select the first structural break by searching the time point $t_0 = r_0 T$, $0 < r_0 < 1$, such that $\psi(r_0)$ in (1.5) is outside the confidence interval of the test. Then, we reproduce the procedure on segment $X_{[r_0 T],T}, \ldots, X_{T-1,T}$. In practice, we set $\alpha = 0.10$.

1.3.3 Forecasting procedure

In the local stationarity framework, the nonstationary process is doubly indexed, see equation (1.1). If we want to forecast h values of an observed process, we arrange the indices as follows: We denote by $X_{0,T}, \ldots, X_{T-h-1,T}$ the observed values with the convention that the ratio h/T tends to zero as T tends to infinity. This implies that the forecasting horizon h may depend on T, but does not increase faster than T.

The last observed realisation is denoted by $X_{T-h-1;T}$, which implies that, in the rescaled time, the local variance function $\sigma^2(z)$ can be estimated only on the interval

$$\left[0, 1 - \frac{h+1}{T}\right]$$

which asymptotically tends to $[0, 1)$. If we denote $\zeta_h = 1 - (h+1)/T$, then the value of $\sigma^2(z)$ outside $[0, \zeta_h]$ are the *predicted values of the unconditional variance in the rescaled time* (see Figure 1.2).

Remark 1.2 (Future observations and rescaled time). With this remark, we would like to give some explanations about the forecasting mechanics for doubly indexed processes. For clarity of presentation, we restrict ourselves to the case $h = 1$. An important ingredient of the rescaled time is that the data come in the form of a triangular array whose rows correspond to *different* stochastic processes, only linked through the unconditional variance sampled on a finer and finer grid. This mechanism is inherently different to what we observe in practice, where, typically, observations arrive one by one and neither the values of the "old" observations, nor their corresponding second-order structure, change when a new observation arrives. One way to reconcile the practical setup with our theory is to assume that for an observed process

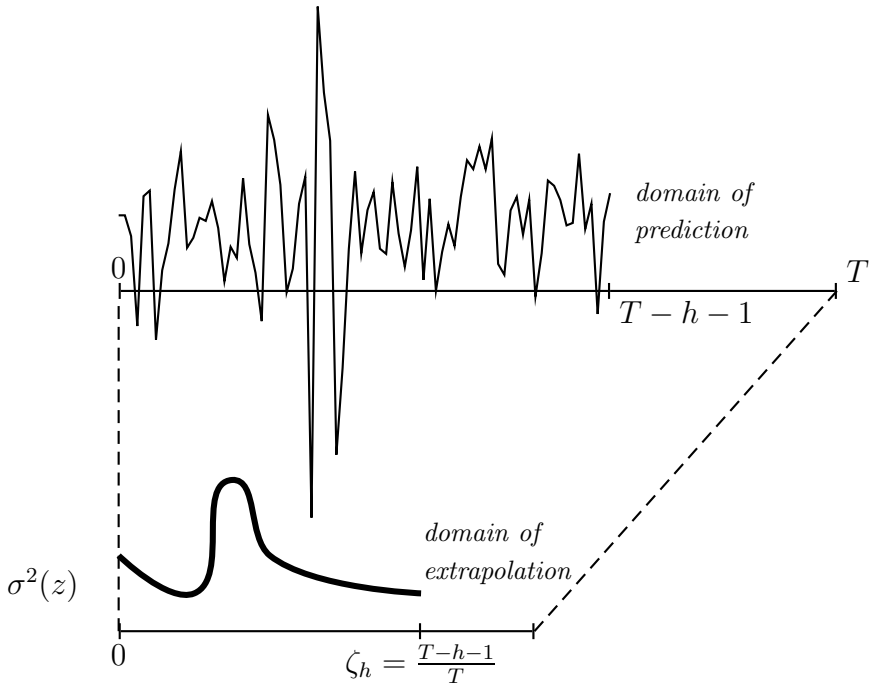


Figure 1.2: This picture illustrates the forecasting procedure explained in Section 1.3.3.

$X_0, \ldots, X_{t-1}$, there exists a doubly-indexed time-modulated process $\mathbf{Y}$ such that $X_k = Y_{k,T}$ for $k = 0, \ldots, t-1$. When a new observation X_t arrives, the underlying time-modulated process changes, i.e. there exists another TM process $\mathbf{Z}$ such that $X_k = Z_{k,T+1}$ for $k = 0, \ldots, t$. $\Diamond$

We can now present our forecasting procedure for time-modulated processes. With our indices convention, the h-step predictor corresponds to the index $T-1$, and then is denoted by $\hat{X}_{T-1,T}$.

1. Predict the local variance. Predicted values of the variance consist in an extrapolation of the estimated variance computed on $[0, \zeta_h]$. In the PC case, the most natural extrapolation is to prolong the last constant piece of the function by the same value. In the Lip case, we use a kernel extrapolator.

2. Define $\tilde{\sigma}^2(z)$ as the estimator of the local variance for $z \leqslant \zeta_h$ and its extrapolation for $\zeta_h < z < 1$.

3. Define the standardised data by

$$\tilde{X}_{t,T} = \frac{X_{t,T}}{\tilde{\sigma}\left(\frac{t}{T}\right)} \qquad t = 0, \ldots, T-h-1. \tag{1.8}$$

4. From equation (1.1), the standardised process $\tilde{X}_{t,T}$ is a zero-mean stationary process. Then, use a standard forecaster for stationary process to forecast $\tilde{X}_{t,T}$.

5. From equation (1.3), we define the *h-step predictor* $\hat{X}_{T-1,T} = \hat{X}_{T-1,T}(h)$ in a natural way as

$$\hat{X}_{T-1,T} = \tilde{\sigma}\left(\frac{T-1}{T}\right) \mathrm{E}\left(Y_{T-1} | \mathcal{F}_{T-h-1}\right) \tag{1.9}$$

   where $\mathrm{E}\left(Y_{T-1} | \mathcal{F}_{T-h-1}\right)$ denotes the classical h-step predictor of a stationary process constructed from the standardised data (1.8).

Note that this predictor is zero for a tm-WN process.

The construction of the prediction intervals is similar. However, recall that the unconditional distribution of Y_t in equation (1.1) is not supposed to be Gaussian due to (1.2). It means that the (zero-mean) random variable $\tilde{X}_{t,T}$ is not Gaussian in general. Then, to construct the prediction interval at the level α, we propose to compute the interval

$$\left[\tilde{q}_{\frac{\alpha}{2}} \quad ; \quad \tilde{q}_{1-\frac{\alpha}{2}}\right]$$

where $\tilde{q}_\alpha$ is the α-quantile of the empirical distribution of the standardised process $\{\tilde{X}_{t,T}\}$, $t = 0, \ldots, n-1$. Finally the prediction interval for the forecasted value of $X_{T-1,T}$ is given by

$$\left[\hat{X}_{T-1,T} - \tilde{\sigma}\left(\frac{T-1}{T}\right)\tilde{q}_{\frac{\alpha}{2}} \quad ; \quad \hat{X}_{T-1,T} + \tilde{\sigma}\left(\frac{T-1}{T}\right)\tilde{q}_{1-\frac{\alpha}{2}}\right]. \tag{1.10}$$

Based on the empirical distribution of $\{\tilde{X}_{t,T}\}$, these prediction intervals are more suitable for the model (1.1), in which the distribution of the stationary process Y_t may be very general.

1.4 Empirical results

In this section, we focus on the practical evaluation of time-modulated processes. Our experiments are presented for five different economic data sets that we present in Section 1.4.1. In Section 1.4.2, we study the question of the covariance stationarity of these time series, and we apply the statistical tests recalled in section 1.3.1. Then, Section 1.4.3 studies the covariance stationarity of the standardised process (1.8), when we estimate the local variance $\tilde{\sigma}(z)$ with the estimators proposed in Section 1.3.2. Finally, Section 1.4.4 is devoted to evaluate the forecasting performance of the procedure described in Section 1.3.3.

1.4.1 The data sets

Nas-100. The log-returns of the Nasdaq-100 Index. This index includes 100 of the largest non-financial domestic and international companies listed on the Nasdaq National Market tier of The Nasdaq Stock Market, Inc. from January 3, 2000 through May 31, 2002 (603 observations);

NasFin-100. The log-returns of the Nasdaq Financial-100 Index. This index includes 100 of the largest financial organizations listed on the Nasdaq National Market tier of The Nasdaq Stock Market, Inc. from January 3, 2000 through May 31, 2002 (603 observations);

Daily-SR. The daily stock returns to the Standard & Poor's composite portfolio from January 3, 1950 through December 29, 1962 (from Schwert [102] – 3100 observations);

Ex-DM-BP. Daily percentage nominal returns for the Deutsche Mark/ British Pound exchange rate from January 3, 1984 to December 31, 1991 (from Bollerslev and Ghysel [13] – 1974 observations);

Ex-US-BP. Daily spot exchange rate of the US dollar to the British pound over a period of 1000 days ending on August 9, 1996 (1000 observations).

1.4.2 Testing covariance stationarity

In a first step, we apply the two tests of covariance stationarity reviewed in Section 1.3.1 on the five data sets. All results are summarized in the column called "Original series" of Table 1.1. In this table, "PSP-test"

	Original series		Stand. (i) — Lip			Stand. (ii) — PC			Stand. (iii) — PC		
	PSP-test	CS-test	PSP-test	CS-test	κ_2	PSP-test	CS-test	κ_2	PSP-test	CS-test	κ_2
Nas-100	0.02	0.535	0.81	*	2.99	0.43	*	3.77	0.86	*	3.48
NasFin-100	0.20	0.053	0.74	*	3.51	0.58	*	9.64	0.71	*	6.83
Daily SR	0.37	0.291	0.78	*	3.62	0.17	0.23	9.13	0.69	0.70	5.30
Ex DM-BP	$< 10^{-2}$	1.338	0.71	*	4.26	0.22	0.15	7.80	0.69	*	8.68
Ex US-BP	$< 10^{-2}$	2.039	0.75	*	4.18	0.05	1.18	5.46	0.68	0.50	7.30

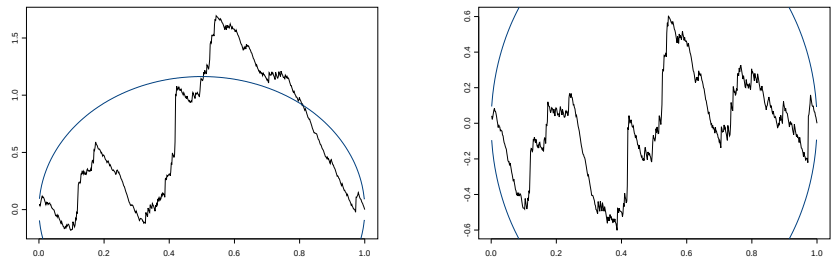
Table 1.1: "PSP-test" denotes the post-sample prediction test (1.4) for which we indicate the two-sided p-value. "CS-test" refers to the CUSUM test of covariance stationarity. This test compares the values of $\psi(r)$ defined in equation (1.5) and the percentiles $c^{\pm}_{0.01}(r)$ of a Brownian bridge. If $\psi(r)$ is fully contained in the confidence intervals, we indicate *, and in the contrary, we compute δ as in equation (1.11) — see text. The table shows results for three different standardisations (1.8) corresponding to the three different estimates (i), (ii) and (iii) of the local variance $\tilde{\sigma}^2(z)$ explained in text. κ_2 is the empirical kurtosis computed for each standardised process.

stands for the post-sample prediction test (1.4) with $\ell = 7$ and "CS-test" refers to the CUSUM-test (1.5). For the first test, we indicate the p-value of the test. For the CUSUM-test, we present the results in a different way: We fix the level of the test at 0.01 and compare the values of $\psi(r)$ defined by equation (1.4) with the percentiles $c^{\pm}_{0.01}$ of a Brownian bridge. If the curve $\psi(r)$ is fully contained in the confidence intervals, the test does not reject the null assumption of covariance stationarity and, in this case, we indicate the symbol *. In the contrary, if $\psi(r)$ intersects the curve $c^{+}_{0.01}$ or $c^{-}_{0.01}$, then the CUSUM test rejects the assumption of covariance stationarity. In this case, we indicate in the table the value of

$$\delta := \max_{0<r<1} \left(|\psi(r)| - c^{+}_{0.01}(r) \right) \tag{1.11}$$

which measures the maximal deviation between $|\psi(r)|$ and $c^{+}_{0.01}(r)$ with respect to r.

To illustrate the CUSUM-test, we plot in Figure 1.3(a) the result of the CUSUM test applied on the Nas-100 series. This test concludes to a lack of homogeneity in the unconditional variance of Nas-100, with $\delta = 0.535$, which motivates the introduction of time-modulated processes.



(a) The CUSUM test applied on the Nasdaq index

(b) The CUSUM test applied on the standardised Nasdaq index

Figure 1.3: In the CUSUM test, the function $\psi(r)$ is ploted for $0 < r < 1$ (see equation (1.5)). The smooth lines which are (partly) plotted are the confidence intervals of the CUSUM test at level 0.01.

From Table 1.1, we can see that the post-sample prediction test (1.4) concludes to the lack of homogeneity in the variance of Nas-100, Ex-DM-BP and Ex-US-BP. On the other hand, the CUSUM test (1.5) shows the

lack of stationarity of all the series. Note that the PSP-test splits the time interval into two segments of equal length and tests the equality of variance on these two segments. This test crucially depends on the time point where we split the series into two parts. Contrary to this test, the CUSUM test controls the variance changes at each time point. Consequently, the reason why the series Daily-SR and NasFin-100 are not considered to be inhomogeneous in variance by the first test is simply due to the fact that the variance is not clearly broken at the middle of the series. Nevertheless, observe that there is only little doubt about the lack of homogeneity for NasFin-100.

1.4.3 Empirical properties of the standardised process

As a second step, we estimate the local variance and, for this, we have three possibilities, given by the three procedures mentioned in Section 1.3:

(i) in the Lip case, we smooth the squared data using a normal kernel and a bandwidth of $40/n$, where n is the total number of observations of the series;

(ii) in the PC case, we use the estimator (1.6) and a segmentation based on $s(t)$, see equation (1.7);

(iii) in the PC case the estimator (1.6) is used with the segmentation based on the CUSUM procedure.

Using our estimation, we standardise the data in the three cases, and run again the two tests of stationarity onto the standardised data. Figure 1.3(b) shows the result for Nas-100 with standardisation (iii), and we observe that the standardised process is variance homogeneous. In other words, the time-varying variance $\sigma^2(z)$ modeled with PC fully explains the lack of stationarity of the original Nas-100 series.

The results for the other series are given in Table 1.1. For each series, we apply the PSP-test and the CUSUM-test on the original data. Then, these two tests are applied on standardised data (1.8) for the three different standardisations (i), (ii) and (iii). If we focus on the CUSUM-test, we see that there is a large doubt about the stationarity of the five original series. This is partly confirmed by the PSP-test, which rejects the null of stationarity at a sensible level of test for the series Nas-100, Ex DM-BP and Ex US-BP. At a first glance, the result

of the two tests seems contradictory for the data NasFin-100 and Daily SR. However, as we have mentioned above, the PSP-test is not a good test when the breakpoint in the variance does not occur at the middle of the time series. Then, the observed difference between the two tests may be explained by this phenomenon.

If we focus on the CUSUM-test and standardisation (i) only, then Table 1 shows that the curve $\psi(r)$ (see (1.5)) is fully contained in the confidence intervals of the test. In other words, this test does not reject the null of stationarity for the standardised data in the Lip case. This result is not so clear for the two other standardisations.

However, even for the PSP-test, some results are still interesting. Consider for instance the series Ex DM-BP. On the original series, the PSP-test rejects the null of stationarity (p-value $< 10^{-2}$). But for each standardised data, this test does not reject the null (p-value > 0.22). This conclusion applies also for Ex US-BP and Nas-100 data. This means that if the original series has a breakpoint in the variance occuring in the middle of the series, then the corresponding standardised process no longer presents this breakpoint.

In conclusion, the first standardisation fully explains the lack of homogeneity of the series. The two other standardisations, corresponding to the PC case, provide less clear results. However, except for the Daily-SR series, the third standardisation leads to better results than the second standardisation in the PC case. Then, we can conclude that, in most of the cases, time-modulated processes satisfactorily explain the nonstationary behaviour observed in Section 1.4.2.

These first results are in favour of a smooth variance modelling. We could also model a time-modulated process (1.1) with a piecewise smooth variance. However, in this case, the estimation of the variance leads to the simultaneous problem of finding a good segmentation and choosing an appropriate smoother. Since the Lip and the PC conditions lead to satisfactory results, we will not consider this general case for the scope of this chapter.

In Table 1.1, we also mention the empirical kurtosis of the standardised process. This is in order to remark that the standardised process generally does not have a Gaussian distribution, and this motivates the construction of prediction intervals based on the empirical quantiles, as explained in Section 1.3.3 for tm-ARMA processes.

1.4.4 Forecasting

In the next step, we want to compare the forecasting accuracy of time-modulated models using the PC and Lip condition. In this section, we will only consider the estimators (i) and (iii). We now explain how we measure this accuracy for the one-step ahead prediction.

First, we clip the series at time t_0 and consider the prediction at time t_0+1. With notations of Figure 1.2, we set $t_0 = T-h-1 = T-2$ because the horizon prediction is $h = 1$. Then, we can proceed to the forecast as described in Section 1.3.3. In our experiments, we work as follows:

1. Fit an autoregressive model to the standardised process obtained in Section 1.4.3. This is done through the minimisation of the AIC criterion [3];

2. Compute the forecasted value at t_0+1 for the stationary process;

3. To compute the forecasted value at time t_0+1, we need to extrapolate the local variance (see equation (1.9)). This extrapolation is constant in the case of the estimator (iii). For the estimator (i), we compute $\tilde{\sigma}\left((t_0+1)/T\right)$ as the average of the local variance estimated at $t_0/T, (t_0-1)/T, \ldots, (t_0-39)/T$. (We also could take a weighted sum on more than 40 data. 40 leads to good results in practice, and, in our experimental results, the results are rather robust with respect to the choice of this parameter.)

This procedure is repeated for different values of t_0. Consequently, the autoregressive process fitted at each t_0 may also change with time, in accordance with Swanson and White [108]. In our experiments, we applied this algorithm to the 100 last values of the observed process.

We follow a similar procedure for h-step ahead prediction. In this case however, the estimator (i) is computed recursively: First, we extrapolate the variance at t_0+1, then we extrapolate the variance at t_0+2 averaging the estimated values of the local variance at $(t_0+1)/T, t_0/T, \ldots, (t_0-38)/T$, and so on until the extrapolation at $(T-1)/T$. From this moving average procedure, we can see that, for a long-range prediction, the extrapolated variance is a weighted sum of the estimated variance at $t_0/T, (t_0-1)/T, \ldots, (t_0-39)/T$.

1.4.5 Evaluating the forecasts

The empirical results are listed in tables 1.2 and 1.3 for a one-step ahead prediction and a four-steps ahead prediction. For each series, the three following criteria are computed [121]:

- The Root-Mean-Square Error (RMSE), given by

$$\sqrt{\frac{1}{100}\sum_{t=T-101}^{T-1}\left(X_{t,T}-\hat{X}_{t,T}\right)^2},$$

  is a widely used measure of forecast accuracy.

- The Mean-Absolute Deviation (MAD), defined by

$$\frac{1}{100}\sum_{t=T-101}^{T-1}\left|X_{t,T}-\hat{X}_{t,T}\right|.$$

- The Mean-Absolute Percentage Error (MAPE)

$$\frac{1}{K_{100}}\sum_{t=T-101}^{T-1}\frac{\left|X_{t,T}-\hat{X}_{t,T}\right|}{|X_{t,T}|}\,\mathbb{I}_{X_{t,T}\neq 0}$$

  where $\mathbb{I}_{X_{t,T}\neq 0}$ is 1 if $X_{t,T}\neq 0$ and 0 if not, and K_{100} is the number of nonzero terms in the sum.

These three criteria concern the accuracy of the predicted mean. We are also interested in the accuracy of the predicted variance of the process. To measure this accuracy, we compute the two following measures:

- the percentage of observations (denoted by "Cov." for "Coverage") which fall within the corresponding prediction intervals. These intervals are constructed using formula (1.10).

- the median of the length of these prediction intervals (MLPI).

Finally, we provide a statistical test of accuracy for the interval forecast. This test is based on the methodology of Christoffersen [17], which

	Series	For. hor.	RMSE	MAD	MAPE	Cov.	MLPI	Christoffersen test (p-values)
Lip	Nas-100	$h = 1$	0.024	0.019	1.362	95 %	0.090	0.39
	NasFin-100		0.009	0.007	1.034	96 %	0.032	0.76
	Daily SR		0.004	0.003	1.022	96 %	0.020	0.76
	Ex DM-BP		0.267	0.188	0.997	93 %	0.928	1
	Ex US-BP		0.005	0.004	1.150	92 %	0.020	0.03
	Nas-100	$h = 4$	0.024	0.019	1.214	95 %	0.092	0.30
	NasFin-100		0.009	0.007	0.997	96 %	0.032	0.75
	Daily SR		0.005	0.003	1.004	95 %	0.020	0.70
	Ex DM-BP		0.266	0.188	1.000	91 %	0.913	0.98
	Ex US-BP		0.005	0.004	1.090	92 %	0.020	0.03

Table 1.2: This table summarizes the prediction accuracy of time-modulated processes with Lip condition, i.e. using the estimator (i).

	Series	For. hor.	RMSE	MAD	MAPE	Cov.	MLPI	Christoffersen test (p-values)
PC	Nas-100	$h = 1$	0.024	0.019	1.558	99 %	0.137	0.09
	NasFin-100		0.008	0.007	1.057	100 %	0.063	1
	Daily SR		0.004	0.003	1.107	100 %	0.026	1
	Ex DM-BP		0.266	0.188	1.000	99 %	2.013	0.09
	Ex US-BP		0.005	0.004	1.074	100 %	0.038	0.98
	Nas-100	$h = 4$	0.025	0.020	1.494	99 %	0.137	0.09
	NasFin-100		0.009	0.007	0.997	100 %	0.026	1
	Daily SR		0.004	0.003	1.160	100 %	0.026	0.87
	Ex DM-BP		0.266	0.188	1.000	99 %	2.014	0.09
	Ex US-BP		0.005	0.004	1.002	100 %	0.038	0.87

Table 1.3: This table summarizes the prediction accuracy of time-modulated processes with PC condition, i.e. using the estimator (iii).

we now recall briefly. For an interval forecast $\mathcal{I}_{t|t-1}(\pi)$ at time t with a coverage probability π, define

$$I_t = \begin{cases} 1 & \text{if } X_{t,T} \in \mathcal{I}_{t|t-1}(\pi) \ , \\ 0 & \text{if } X_{t,T} \notin \mathcal{I}_{t|t-1}(\pi) \ . \end{cases}$$

Christoffersen [17] notes that the quality of the interval forecasts should be tested by a conditional coverage hypothesis, i.e.

$$H_0 : \mathrm{E}(I_t | I_{t-1}, \ldots, I_1) = \pi \text{ for all } t. \tag{1.12}$$

He shows that this test is equivalent to testing that the sequence $\{I_t\}$ consists in identically and independent distributed Bernoulli random variables, with parameter π. The likelihood ratio test of conditional coverage is

$$\mathcal{L}_T(\pi) = -2 \log \frac{\ell(\pi; I_1, \ldots, I_T)}{\ell(\hat{\Pi}_1; I_1, \ldots, I_T)}, \tag{1.13}$$

where

$$\ell(p; I_1, \ldots, I_T) = (1-\pi)^{n_0} \pi^{n_1}$$

and

$$\ell(\hat{\Pi}_1; I_1, \ldots, I_T) = (1-\hat{\pi}_{01})^{n_{00}} \hat{\pi}_{01}^{n_{01}} (1-\hat{\pi}_{11})^{n_{10}} \hat{\pi}_{11}^{n_{11}},$$

where n_{ij} is the number of observations with value i followed by j ($i, j =$ 0 or 1), n_i is the number of observations i (0 or 1), and $\hat{\pi}_{ij}$ is an estimator of $\Pr(I_t = i | I_{t-1} = j)$ given by

$$\begin{pmatrix} \hat{\pi}_{00} & \hat{\pi}_{01} \\ \hat{\pi}_{10} & \hat{\pi}_{11} \end{pmatrix} = \begin{pmatrix} \frac{n_{00}}{n_{00}+n_{01}} & \frac{n_{01}}{n_{00}+n_{01}} \\ \frac{n_{10}}{n_{10}+n_{11}} & \frac{n_{11}}{n_{10}+n_{11}} \end{pmatrix}.$$

Christoffersen [17] proves that, under H_0, the likelihood ratio $\mathcal{L}_T(\pi)$ is asymptotically distributed as a χ^2 random variable with 2 degrees of freedom.

Depending on the application we have in mind, one of these criteria will be the most important for the practitioner. The ideal predictor will have a minimal loss in the mean (RMSE, MAD or MAPE), with the best

coverage, the shortest MLPI and the best accuracy (Christoffersen test). However, it may happen that these quantities are balancing each other. For instance, we can have a predictor with the best loss in the mean, but which has not the best conditional or unconditional coverage, etc. In this case, selecting the most appropriate predictor would be advised by the goals of the practitioner or the physical nature of the series. In other words, for an observed time series, if it is more important to get a precise forecast in the mean, then select the predictor with the best loss in the mean. On the contrary, if a precise value of the mean is less of interest than accurate predicted variability, then consider a predictor with the best MLPI, and having a sensible percentage of coverage and loss in the mean.

1.4.6 Results

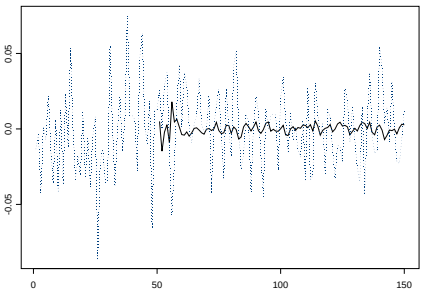
The results listed in tables 1.2 and 1.3 are an illustration of this balancing phenomenon. Indeed, this table indicates that the PC model leads to better results than the Lip condition as long as the accuracy in the predicted mean is considered. However, if we consider the variance prediction, the Lip model shows the best MLPI's. It means that at level 0.05 (i.e. if we accept 5 % of error in our prediction) the forecasting procedure based on the Lip condition leads to smaller prediction intervals.

Another observation is the difficulty to have a good coverage for exchange rate data. This phenomenon is often observed for these data. This coverage is lower than 95 % with the Lip model. It is better with the PC model if we consider the coverage and the Christoffersen's test, but is not satisfactory for the Ex-DM-BP series, which leads to very large prediction intervals.

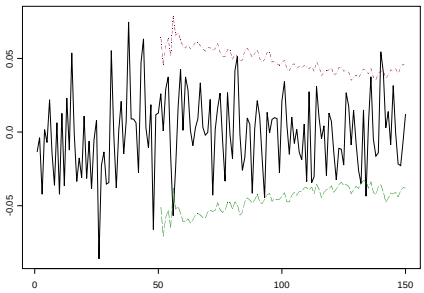
Finally, Figure 1.4 illustrates the forecasting procedure on a segment of the Nas-100 series. This prediction is provided under the Lip assumption, and we can observe in Figure 1.4(b) the smooth variation of the local variance, which automatically follows the smooth evolution of the actual series.

1.5 Comparisons with standard models

Our last empirical study is devoted to the comparison of our forecasting procedure with standard models in econometrics, namely GARCH$(1,1)$ and EGARCH$(1,1)$. Many out-of-sample tests are proposed in the lit-



(a) The dotted line is the actual series, and the solid line is the one-step ahead prediction.



(b) The prediction intervals of the same forecasting procedure as in Figure 1.4(a).

Figure 1.4: One-step ahead prediction at 100 time points of the series Nas-100. The prediction is done between January 3, 2000 and March 14, 2002 with the Lip model.

erature in order to test the equality of forecast accuracy [34, 48, 117]. However, most of these tests are based on strong assumptions and cannot be of a direct use in our situation. More precisely, these tests are based on the maintained assumption that the data are stationary in covariance, and, in our context, this assumption is precisely not acceptable. As a consequence, we need to develop specific tests in order to compare the forecasts between a stationary and a potential nonstationary models. This section proposes two approaches of comparison. A deeper study of this relevant question should be addressed in a future work.

1.5.1 A generalised Meese-Rogoff test

This test of forecast accuracy is introduced in Meese and Rogoff [72] and reviewed in Diebold and Mariano [34]. It is based on the maintained assumptions that the forecasting error $X_{t,T} - \hat{X}_{t,T}$ are zero-mean, covariance stationary and Gaussian. These two last assumptions are certainly too strong in our context, and, in this subsection, we show how to derive this test under weaker conditions.

Suppose we would like to compare the forecasting performance between the two models

$$X_{t,T}^{(1)} = \sigma\left(\frac{t}{T}\right) Y_t^{(1)}$$

and

$$X_{t,T}^{(2)} = \sigma\left(\frac{t}{T}\right) Y_t^{(2)}$$

with a known unconditional variance $\sigma(t/T)$ (possibly time-varying). The standardised errors of prediction for these two models are denoted by

$$e_{t,T}^{(i)} = \frac{X_{t,T} - \hat{X}_{t,T}^{(i)}}{\sigma(t/T)} \qquad i = 1, 2, \tag{1.14}$$

and the null hypothesis of equal forecasting performance (in a quadratic sense) is

$$H_0 : \mathrm{E}\left[T^{-1}\sum_{t=0}^{T-1}\left(e_{t,T}^{(1)}\right)^2\right] = \mathrm{E}\left[T^{-1}\sum_{t=0}^{T-1}\left(e_{t,T}^{(2)}\right)^2\right].$$

Let $x_{t,T} = e_{t,T}^{(1)} + e_{t,T}^{(2)}$ and $z_{t,T} = e_{t,T}^{(1)} - e_{t,T}^{(2)}$. If the processes $Y_t^{(i)}$ are Gaussian, computations similar to Priestley [95, Section 5.3.3] lead to

$$\sqrt{T}\hat{\gamma}_{xz} \xrightarrow{\mathrm{d}} \mathcal{N}(0, V) \qquad \text{under } H_0 \tag{1.15}$$

where

$$\hat{\gamma}_{xz} = \frac{x'z}{T}$$

and

$$V = \frac{1}{T}\sum_s \left[\gamma_{xx}(s)\gamma_{zz}(s) + \gamma_{xz}(s)\gamma_{zx}(s)\right]$$

with

$$\gamma_{xx}(s) = \mathrm{Cov}\left(x_{t,T}, x_{t-s,T}\right), \quad \gamma_{zz}(s) = \mathrm{Cov}\left(z_{t,T}, z_{t-s,T}\right),$$
$$\gamma_{xz}(s) = \gamma_{zx}(s) = \mathrm{Cov}\left(x_{t,T}, z_{t-s,T}\right).$$

In our context, we allow the processes $Y_t^{(i)}$ to be non-Gaussian, but we assume the Central Limit Theorem (1.15) holds with

$$V = \frac{1}{T}\sum_s \left[\gamma_{xx}(s)\gamma_{zz}(s) + \gamma_{xz}(s)\gamma_{zx}(s)\right] + \frac{1}{T^2}\sum_{s,t}\kappa_{4,xz}(s,t)$$

where $\kappa_{4,xz}(s,t)$ is the fourth cumulant of the distribution of $(x_{s,T}, x_{t,T}, z_{s,T}, z_{t,T})$. Estimation of V is done by plugging in the formula consistent estimates for $\gamma_{xx}, \gamma_{xz}, \gamma_{zx}, \gamma_{zz}, \kappa_{4,xz}(s,t)$. There are many standard results in the literature for this estimation of V, and we use the estimator reviewed in Diebold and Mariano [34] in our empirical comparisons.

In practice, an estimation of $\sigma(t/T)$ in (1.14) is needed. If the model $X_{t,T}^{(i)}$ is covariance-stationary (i.e. σ is constant over time), then we compute the root variance of $\{X_{0,T}, \ldots, X_{T-1,T}\}$. If the model $X_{t,T}^{(i)}$ is time-modulated with a time-varying variance, $\sigma(\cdot)$ is estimated like in Subsection 1.3.2, using the whole data $\{X_{0,T}, \ldots, X_{T-1,T}\}$.

1.5.2 Christoffersen tests

The generalised Meese-Rogoff test is for comparing the forecast accuracy between two forecasting procedures focusing on the predicted mean. To compare the forecasting in terms of conditional coverage, the Christoffersen's test may be used, even for time-modulated processes. As in Christoffersen [17], the test (1.12) may be applied for a wide range of π between 0.9 and 0.99.

1.5.3 Results

The two procedures have been applied on the five data sets of Subsection 1.4.1. We compute the generalised Meese-Rogoff test to compare the forecasting procedure of Subsection 1.3.3 (with estimator (i) and (iii)) against the two conditional models GARCH(1,1) and EGARCH(1,1). Again, in our experiments, we compute 100 one-step ahead forecasted values at the end of each time series. In all the cases, the p-value of the test is larger than 0.91. This means that, in terms of the predicted mean, the time-modulated approach does not show any improvement over the two ARCH-type models. This is not a surprising result, as time-modulated processes differ from standard models in terms of the variance only.

More relevant is the comparison in terms of interval forecasts by the Christoffersen test. Figure 1.5 plots the likelihood ratio statistics for some values of π between 0.9 and 0.99. The three figures correspond to the NasFin-100, Daily-SR and Ex-DM-BP data respectively. In each figure, the likelihood ratio statistic (1.13) is computed for the tm-AR model with a variance estimated under the Lip model (solid line) and the tm-AR model with a variance estimated under the PC condition (short

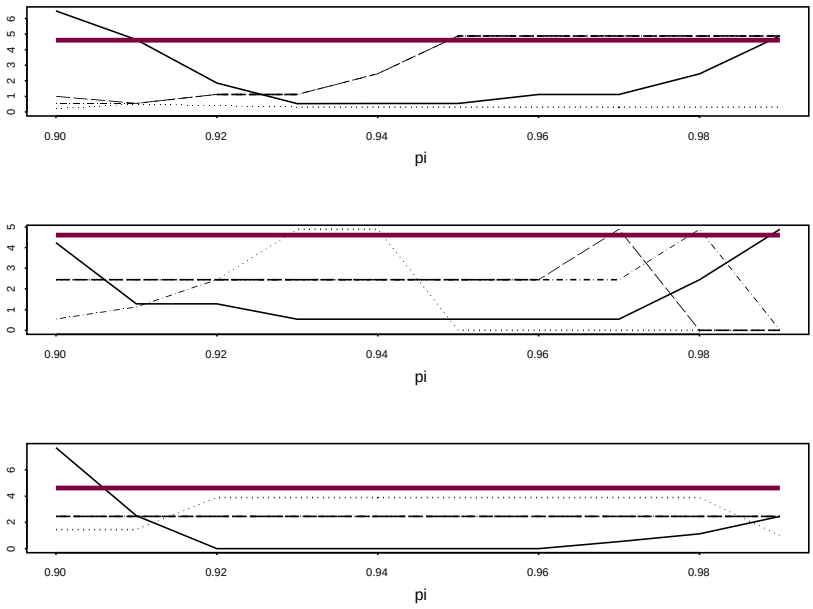


Figure 1.5: NasFin-100, Daily-SR and Ex-DM-BP: Likelihood ratio statistics of conditional coverage with different models. The solid line corresponds to tm-AR model with a variance estimated under the Lip model. The short dash is tm-AR model with a variance estimated under the PC condition. The long dash is GARCH(1,1) and the short-long dash is EGARCH(1,1) (these two lines are very close in the first figure). The bold line corresponds to 10 per cent significance level of the appropriate χ^2 distribution.

dash). They are compared to the GARCH(1,1) model (long dash) and the EGARCH(1,1) model (short-long dash). The bold line corresponds to 10 per cent significance level of the appropriate χ^2 distribution.

The behaviour of the tm-AR model is similar in the three cases. It starts with high values at 0.9, then has good behaviour since its values are below the significance level, and then increases near 0.99. Between 0.92 and 0.98, it performs better than the (E)GARCH models. The tm-AR model with a variance estimated under the PC condition has a more stable behaviour, except for the stock returns.

We have to note that, from a modelling viewpoint, the time-modula-

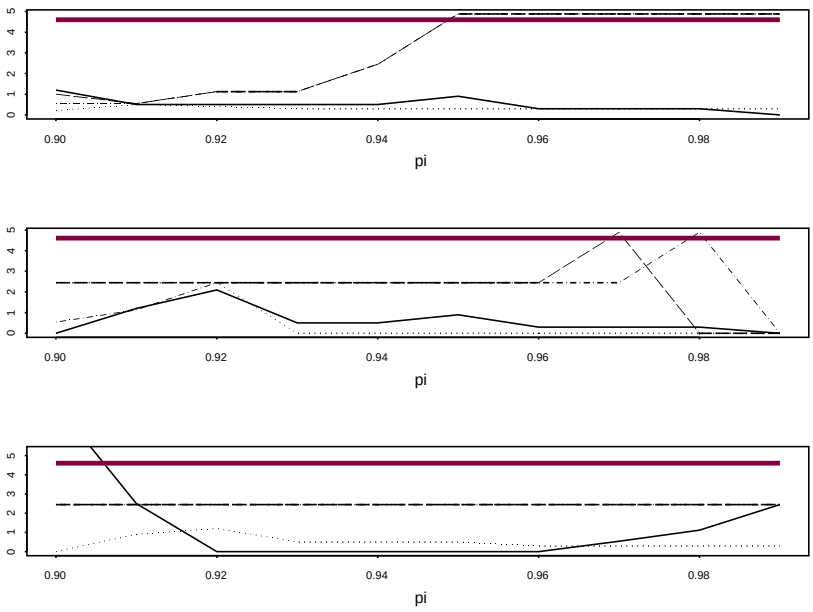


Figure 1.6: NasFin-100, Daily-SR and Ex-DM-BP: Likelihood ratio statistics of conditional coverage with different models. The solid line corresponds to tm-GARCH(1,1) model with a variance estimated under the Lip model. The short dash is tm-GARCH(1,1) model with a variance estimated under the PC condition. The long dash is GARCH(1,1) and the short-long dash is EGARCH(1,1). The bold line corresponds to 10 per cent significance level of the appropriate χ^2 distribution.

ted AR processes is not a competitor to the GARCH models. Indeed, it may happen that the standardised process (1.8) contains a time-varying conditional variance. It then makes sense to fit an ARCH-type model to the standardised process. Figure 1.6 presents the same results as Figure 1.5, except that a GARCH model is fitted to the standardised processes. (Again, the solid line corresponds to the Lip modelling of the unconditional variance, while the dashed line is under the PC model.) Comparing figures 1.5 and 1.6, we see that the behaviour of time-modulated GARCH models performs better than the time-modulated AR models.

1.6 Concluding remarks

In this chapter, we have provided a methodology to forecast economic time series when the assumption of variance stationarity is rejected by statistical tests. This lack of homogeneity of unconditional variance has been observed on daily stock returns and exchange rate data and they confirm preliminary results published in the literature [63, 64, 88, 89]. Using the framework of rescaled time, we have shown that a simple process with a time-varying variance may be defined and explains satisfactorily the nonstationary behaviour of the observed series. This process consists in a stationary process modulated by a time-varying variance defined in the rescaled time. Two kinds of time-varying unconditional variance were considered: A smooth time-varying variance and a piecewise constant variance. The first specification gives better results for forecasting the variance, and the second specification gives better results in order to forecast the mean.

Our results confirm that economic processes can evolve over time. Consequently, it is advisable to do a test of covariance stationarity before fitting an econometric model which assumes the constancy of the unconditional moments. This paper shows a simple and meaningful way to standardise nonstationary data in order to obtain a process which can be assumed to be stationary. It is then possible to apply the standard econometric modelling on the standardized data, and to forecast it. Forecasting the actual series is then straightforward on the non-standardized data.

Our method of standardisation is based on the estimation of an unconditional variance. The proposed estimation procedures require the choice of parameters, like a smoothing parameter. In this chapter, these parameters are not automatically chosen. In the next chapters, more complex models are studied and adaptive data-driven procedures are developped for the estimation of their parameters. However, even if our parameters are chosen by hand in the empirical studies of this chapter, they provide satisfactory results, as the standardised process is stationary and the results presented in this study are robust with respect to these parameters. It means that, taking a slightly different parameter does not significantly change the results.

We also point out the inherent statistical difficulty to compare the forecasting accuracy between models which are possibly not covariance-

stationary. We examine two procedures for comparing the accuracy of the forecasting between time-modulated processess and ARCH-type models. As time-modulated processes model a time-varying unconditional variance, the predicted mean does not show significantly better results than ARCH-type models. However, when considering the accuracy of interval forecasts, we show that time-modulated models with Lipschitz-continuous variance perform better.

As a follow-up, we also compare the results with time-modulated GARCH $(1,1)$ models. In terms of interval forecast, our results show the gain in prediction accuracy, in comparison with stationary GARCH $(1,1)$ and EGARCH $(1,1)$ models. Consequently, the conclusion of our results may be summarized as follows:

1. Very often, economic processes exhibit a time-varying unconditional variance.

2. Standard ARCH-type models assume the stationarity of the data. In this paper, we show how a standardisation of the data may be applied in order to "stationarize" the data. Then ARCH-type models may be applied on the *standardised* data and forecasts may be provided.

3. If this standardisation is not provided, the accuracy of the interval forecast decreases.

Then, one ultimate message of our results for the analysis of economic processes is that it could be useful to perform more empirical studies on modelling and fitting time-modulated GARCH processes to time series. This poses some challenging problems both from a theoretical and practical viewpoint, in particular for identification. Here, the results of Dahlhaus and Subba Rao [29] may be a starting point.

Other approaches of forecasting nonstationary signals are possible (see for instance Grillenzoni [43] or Ramsey and Zhang [96]). However, our approach based on the rescaled time leads to a consistent statistical theory modelling time-varying processes. Moreover, analysis with a time-modulated stationary process is related to classical statistical problems, such as nonparametric regression problem, nonparametric extrapolation and estimation of quantiles for a non-Gaussian distribution.

*

This first chapter aimed to focus on the essential problems treated in the thesis. We have introduced important concepts on a simple model: the local stationarity, the rescaled time, the estimation problem and the need of adaptive methods for selecting the parameters of the estimation, and the forecasting problem with nonstationary data. In the next chapters, we introduce more complex models for nonstationarity, and develop adaptive methods for the estimation of their parameters and for forecasting. In these models, not only the variance is time-varying, but all the second-order quantities of the model (autocovariance function, spectrum).

1.7 Appendix: Some statistical properties of the variance estimator

In this Appendix, we focus on the statistical properties of the variance estimates. We will successively consider the case of the Lip assumption and the PC assumption.

Proposition 1.1. *Assume that the function σ in equation (1.1) obeys the Lip condition. Then,*

$$\begin{aligned} \mathrm{E}X^2_{[zT],T} &= \sigma^2(z) \\ \operatorname{Var} X^2_{[zT],T} &= \sigma^4(z)(\kappa_2 - 1) \end{aligned}$$

for all $0 < z < 1$, where κ_2 is the kurtosis (1.2).

The proof of this proposition is straightforward from equation (1.1), recalling that the variance of the stationary process is equal to one.

Proposition 1.2. *Assume that the function σ in equation (1.1) obeys the PC condition and has no structural break between z_0 and z_1, $0 < z_0 < z_1 < 1$. Then the estimator (1.6) is such that*

$$\mathrm{E}\hat{s}_T = \frac{1}{|z_1 - z_0|} \int_{z_0}^{z_1} dz\, \sigma^2(z)$$

and

$$\operatorname{Var}\hat{s}_T = \frac{1}{|z_1 - z_0|^2\, T}\gamma(0)\int_{z_0}^{z_1} dz\ \sigma^4(z)$$
$$+\frac{1}{|z_1 - z_0|^2\, T}\sum_{u=1}^{[(z_1-z_0)T]-1}\gamma(u)\int_{z_0}^{z_1-\frac{u}{T}} dz\ \left[\sigma^4(z) + \sigma^4\left(z+\frac{u}{T}\right)\right]$$

where $\gamma(s-t) = \operatorname{Cov}\left(Y_s^2, Y_t^2\right)$.

Proof. Taking the expectation of (1.6) we get, with definition (1.1):

$$\mathrm{E}\hat{s}_T = \frac{1}{|z_1 - z_0|\, T}\sum_{t=[z_0T]}^{[z_1T]-1}\sigma^2\left(\frac{t}{T}\right)$$
$$= \frac{1}{|z_1 - z_0|}\sum_{t=[z_0T]}^{[z_1T]-1}\int_0^{\frac{1}{T}} dz\ \sigma^2\left(\frac{t}{T}\right)$$

Because $\sigma^2(\cdot)$ is constant over $[z_0, z_1]$, it follows

$$\mathrm{E}\hat{s}_T = \frac{1}{|z_1 - z_0|}\sum_{t=[z_0T]}^{[z_1T]-1}\int_0^{\frac{1}{T}} dz\ \sigma^2\left(z+\frac{t}{T}\right)$$
$$= \frac{1}{|z_1 - z_0|}\sum_{t=[z_0T]}^{[z_1T]-1}\int_{\frac{t}{T}}^{\frac{t+1}{T}} dz\ \sigma^2(z)$$

which leads to the result for the expectation. The computation of the variance is similar: If $\gamma(u)$ denotes the autocovariance function of the process $\{Y_t^2\}$, then we can write

$$\operatorname{Var}\hat{s}_T = \frac{1}{|z_1 - z_0|^2\, T^2}\sum_{s,t=[z_0T]}^{[z_1T]-1}\gamma(s-t)\,\sigma^2\left(\frac{s}{T}\right)\sigma^2\left(\frac{t}{T}\right)$$
$$= \frac{1}{|z_1 - z_0|^2\, T}\sum_{s,t=[z_0T]}^{[z_1T]-1}\gamma(s-t)\int_{\frac{t}{T}}^{\frac{t+1}{T}} dz\ \sigma^2\left(\frac{s-t}{T}+z\right)\sigma^2(z)$$

Because $\sigma^2(\cdot)$ is constant over $[z_0, z_1]$, it follows

$$\operatorname{Var}\hat{s}_T = \frac{1}{|z_1 - z_0|^2\, T}\sum_{s,t=[z_0T]}^{[z_1T]-1}\gamma(s-t)\int_{\frac{t}{T}}^{\frac{t+1}{T}} dz\ \sigma^4(z)$$

Defining $u := s - t$, direct computations lead to

$$\operatorname{Var} \hat{s}_T = \frac{1}{|z_1 - z_0|^2 \, T} \left\{ \sum_{u=0}^{[(z_1-z_0)T]-1} \gamma(u) \sum_{t=[z_0T]}^{[z_1T]-1-u} \int_{\frac{t}{T}}^{\frac{t+1}{T}} dz \; \sigma^4(z) \right.$$
$$\left. + \sum_{u=[(z_0-z_1)T]+1}^{-1} \gamma(u) \sum_{t=[z_0T]-u}^{[z_1T]-1} \int_{\frac{t}{T}}^{\frac{t+1}{T}} dz \; \sigma^4(z) \right\}$$

and the result follows using the symmetry of $\gamma(u)$. □

CHAPTER 2

Semiparametric estimation by model selection for locally stationary processes

2.1 Introduction

In the previous chapter, we have introduced a simple model of nonstationarity where only the variance is time-varying. A typical model in this class is an ARMA process with an evolutionary variance.

The present chapter studies more complex models, where not only the variance is evolutionary, but all the coefficients of the model. A typical example is given by the time-varying ARMA(p, q) model (tv-ARMA in short) defined by

$$\sum_{j=0}^{p} a_j\left(\frac{t}{T}\right) X_{t-j,T} = \sum_{k=0}^{q} b_j\left(\frac{t}{T}\right) \varepsilon_{t-j,T}. \tag{2.1}$$

where $a_0(\cdot) \equiv b_0(\cdot) \equiv 1$ and $\varepsilon_{t,T}$ are independent normal random variables $\mathcal{N}(0, \sigma(t/T)^2)$. In this example, the parameter of interest is the D-dimensional vector of curves

$$\theta(u) = (a_1(u), \ldots, a_p(u), b_1(u), \ldots, b_q(u), \sigma^2(u))$$

with $D = p + q + 1$.

More generally, this chapter concerns processes characterised by such a D-dimensional time-varying curve θ and we address the problem of

how to estimate this vector. This is what we call a semiparametric estimation problem because, on one hand, the parameter of interest is parametrised by D curves and, on the other hand, each of these curves will be estimated nonparametrically.

The example of tv-ARMA shows that this task is complicated by the fact that the curve θ is not observed "directly". This is in contrast with the situation of the classical nonparametric regression framework, where the curve $\theta(\cdot)$ is observed plus some noise. In this chapter, the characteristics of the process (such as the spectrum) may depend on the parameter curves in a highly nonlinear way.

The goal of this chapter is to develop a data-driven method that automatically selects an estimator $\hat{\theta}_{\hat{m}}$ from a collection of estimators $\hat{\theta}_m$ for varying index m. These estimators are constructed as minimum contrast estimators, and the contrast function is an approximation of the Gaussian likelihood of the model. The estimator $\hat{\theta}_{\hat{m}}$ follows from a method of model selection procedure. We show that the estimator achieves good theoretical properties, in a sense which is described in Section 2.3.

The precise definition of the nonstationary model is given in the next section. This definition starts from the spectral representation of time series. It is worth mentioning that, in this chapter, we do not assume the model to follow the semiparametric structure explained below. In other words, our results include the case of model misspecification. In Section 2.3, we describe our fitting procedure, and present the main results of the chapter. For sake of clarity, all technical tools and proofs are deferred to Section 2.4.

2.2 The model of local stationarity

The model of local stationarity is defined in the spectral domain. We first recall some facts about the spectral representation of time series.

2.2.1 Spectral representation of time series

Spectral analysis of time series is a large field presenting a great interest from both theoretical and practical viewpoints. The fundamental starting point of this analysis is the *Cramér representation*, stating that all second-order zero-mean stationary process X_t, $t \in \mathbb{Z}$ may be written

$$X_t = \int_{[-\pi,\pi)} A(\lambda) \exp(i\lambda t) dZ(\lambda), \qquad t \in \mathbb{Z}\,, \tag{2.2}$$

where $A(\lambda)$ is the *amplitude* or *transfer function* of the process X_t and $dZ(\lambda)$ is an orthonormal increment process, i.e. $\mathrm{E}(dZ(\lambda), \overline{dZ(\mu)}) = \delta_0(\lambda - \mu)$, see Priestley [95]. Correspondingly, under mild conditions, the autocovariance function can be expressed as

$$c_X(\tau) = \int_{[-\pi,\pi)} f_X(\lambda) \exp(i\lambda\tau) d\lambda,$$

where f_X is the *spectral density* of X_t.

There is not a unique way to relax the assumption of stationarity, i.e. to define a second-order process with a time-depending spectrum. However, this modelling is a theoretical challenge which may be helpful in practice, since a lot of studies have shown that models with evolutionary spectra or time-varying parameters are necessary to explain some observed data, even over short periods of time. Examples may be found in numerous fields, such as economics (Chapter 1), biostatistics (Chapter 5 or [87]) or meteorology (Chapter 6 or [80]) to name but a few.

Among the different possibilities for modelling nonstationary second-order processes, we can emphasize the approaches consisting in a modification of the Cramér representation (2.2). Different modifications of (2.2) are possible. First, we can replace the process $dZ(\lambda)$ by a nonorthonormal process, leading for instance to the *harmonizable* processes [61]. A second possibility is to replace the amplitude function $A(\lambda)$ by a time-varying version $A_t(\lambda)$ and assume a slow change of $A_t(\lambda)$ over time. Such approach is followed to define *oscillatory* processes [94].

However, a major statistical drawback of the oscillatory processes is the intrinsic impossibility to construct an asymptotic theory for consistency and inference. To overcome this problem, Dahlhaus [23] introduced the class of *locally stationary processes*, in which the transfer function is rescaled in time. In this approach, a doubly-indexed process is defined as

$$X_{t,T} = \int_{[-\pi,\pi)} A\left(\frac{t}{T}, \lambda\right) \exp(i\lambda t) dZ(\lambda), \qquad t = 0, \ldots, T-1, \quad T > 0\,, \quad (2.3)$$

where the transfer function $A(z, \lambda)$ is defined on $(0, 1) \times [-\pi, \pi)$.

Dahlhaus [23, 24] investigated statistical inference for such processes, with a discussion on maximum likelihood, Whittle and least squares

estimates, and showed that asymptotic results when T tends to infinity can be considered. A precise definition is given in the next subsection.

2.2.2 Locally stationary processes

Assume we observe data $X_1, \ldots, X_T$ from some nonstationary process and we want to fit a semiparametric model to the data. An appropriate framework, which allows for a rigorous asymptotic treatment of nonstationary time series, is the following model for locally stationary processes introduced in Dahlhaus [23].

Definition 2.1 ([23]). *A sequence of stochastic processes* $\{X_{t,T};\ t = 1, \ldots, T\}$ *is called* locally stationary *with transfer function* A° *if there exists a representation*

$$X_{t,T} = \int_{-\pi}^{\pi} A^\circ_{t,T}(\lambda) \exp(i\lambda t) dZ(\lambda), \qquad t = 1, \ldots, T, \qquad T > 0, \tag{2.4}$$

where

1. $Z(\lambda)$ *is a complex valued Gaussian process on* $[-\pi, \pi]$ *with* $\overline{Z}(\lambda) = Z(-\lambda)$, $\mathrm{E}Z(\lambda) = 0$ *and orthonormal increments, i.e.*

$$\mathrm{E}\{dZ(\lambda_1), dZ(\lambda_2)\} = \eta(\lambda_1 + \lambda_2)\, d\lambda_1 d\lambda_2$$

   where $\eta(\lambda) = \sum_{j=-\infty}^{\infty} \delta(\lambda + 2\pi j)$ *is the period* 2π *extension of the Dirac delta function (Dirac comb [76]), and where*

2. *there exists a positive constant* K *and a smooth function* $A(u, \lambda)$ *on* $[0, 1] \times [-\pi, \pi]$ *which is* 2π*-periodic in* λ*, with* $A(u, -\lambda) = \overline{A(u, \lambda)}$*, such that for all* T*,*

$$\sup_{t,\lambda} |A^\circ_{t,T}(\lambda) - A(t/T, \lambda)| \leqslant K/T\ .$$

Moreover, a locally stationary process is said to be Gaussian *if its increment process* $\{Z(\lambda), \lambda \in [-\pi, \pi]\}$ *is Gaussian.*

In this definition two different functions $A^\circ_{t,T}(\lambda)$ and $A(t/T, \lambda)$ are defined. This complicated construction is necessary if we want to model a class of processes which is rich enough to cover interesting applications. In particular, if we do not define these two functions, i.e. if $A^\circ_{t,T}(\lambda) =$

$A(t/T, \lambda)$ in the above definition, then the class does no longer include time-varying AR(p) processes (as showed in Dahlhaus [22]).

Observe that we have used the same convention for the asymptotic concept than in Chapter 1. This implies that the nonstationary process is doubly-indexed. The smoothness of A in u defines the departure from stationarity and ensures the locally stationary behaviour of the process. In order to formulate the smoothness assumptions on A, we first need to recall the definition of the total variation norm.

The *total variation norm* of a univariate function f defined on an interval $[a, b]$ is

$$\begin{aligned}&\mathrm{TV}_{[a,b]}(f)\\&= \sup\left\{\sum_{i=1}^{I}\left|f\left(a_i\right)-f\left(a_{i-1}\right)\right| : a<a_0<\ldots<a_I<b, I\in\mathbb{N}\right\}.\end{aligned}$$

If there is no risk of ambiguity of the domain of f, we sometimes write $\mathrm{TV}(f)$ for the total variation norm of f.

We can now formulate the exact smoothness assumptions on A, following the setting of Neumann and von Sachs [85].

Assumption 2.1 ([85]). The function A in Definition 2.1 is such that

1. $\sup_u \mathrm{TV}_{[-\pi,\pi]}\left(A(u,\cdot)\right) \leqslant C_1 < \infty$
2. $\sup_\lambda \mathrm{TV}_{[0,1]}\left(A(\cdot,\lambda)\right) \leqslant C_2 < \infty$
3. $\sup_{u,\lambda} |A(u,\lambda)| \leqslant \kappa_s < \infty$
4. $\inf_{u,\lambda} |A(u,\lambda)| \geqslant \kappa$ for some $\kappa > 0$
5. Let

$$\hat{A}(u,s) := (2\pi)^{-1}\int_{-\pi}^{\pi} d\lambda\; A(u,\lambda)\exp(i\lambda s)$$

for $s \in \mathbb{Z}$ and $u \in [0,1]$. Then $\sup_u \sum_s |\hat{A}(u,s)| < \infty$. ◊

2.2.3 Evolutionary spectral density

Let $\{X_{s,T}\}$ be a locally stationary process as defined in Definition 2.1. The *Wigner-Ville* spectrum for fixed T and $u \in (0,1)$ is given by

$$f_T(u,\lambda) = \frac{1}{2\pi} \sum_{s=\infty}^{\infty} \mathrm{Cov}\left(X_{[uT-s/2],T}, X_{[uT+s/2],T}\right) \exp(-i\lambda s),$$

where we have used the convention $A^{\circ}_{t;T}(\lambda) = A(0,\lambda)$ for $t < 1$ and $A^{\circ}_{t,T}(\lambda) = A(1,\lambda)$ for $t > T$ (the quantity $A^{\circ}_{t;T}(\lambda)$ is actually only defined for $t = 1,\ldots,T$ and this convention is for sake of simplifications in the proofs). The Wigner-Ville spectrum has been introduced by Martin and Flandrin [70] in order to define a time-varying spectrum of a nonstationary process. The next result shows that $f_T(u,\lambda)$ converges asymptotically (in a sense defined below) to

$$f(u,\lambda) = \left|A(u,\lambda)\right|^2.$$

Therefore, we call $f(u,\lambda)$ the *evolutionary spectral density* (ESD) of the process.

Proposition 2.1 ([22, 85]). *If $X_{t,T}$ is locally stationary (Definition 2.1) and under Assumption 2.1,*

$$\int_0^1 du \int_{-\pi}^{\pi} d\lambda \, |f_T(u,\lambda) - f(u,\lambda)|^2 = o_T(1)$$

Proof. See Theorem 3.1 of Neumann and von Sachs [85]. □

This result is important because it shows the uniqueness of the evolutionary spectral density $f(u,\lambda)$. Even if the spectral representation (2.4) is not unique [22, 95], Proposition 2.1 shows that if there exists a spectral representation with a $A(u,\lambda)$ such that Assumption 2.1 is fulfilled, then $|A(u,\lambda)|^2$ is uniquely determined. As it is the limit of the Wigner-Ville spectrum, we may call this quantity a spectrum with no ambiguity.

To conclude this section, we give some specific examples of ESD. First, we consider the case of time-modulated processes studied in Chapter 1. Recall that these processes are zero-mean stationary processes with unit variance that are multiplied by a time-varying function

$X_{t,T} = \sigma(t/T)Y_t$ (see (1.1)). If $f_Y(\lambda)$ denotes the spectral density of the stationary process Y_t, then it is straightforward to derive the ESD of the process $X_{t,T}$, which is $f_X(u,\lambda) = \sigma^2(u) f_Y(\lambda)$.

The second example is for tv-ARMA(p,q). This case is more involved and leads to the following ESD [22]:

$$f_{\theta(u)} = \frac{\sigma^2(u)}{2\pi} \frac{|\sum_{k=0}^{q} b_k(u)\exp(i\lambda k)|^2}{|\sum_{j=0}^{p} a_j(u)\exp(i\lambda j)|^2}. \tag{2.5}$$

2.3 Semiparametric estimation

The model we like to be fitted is characterized by a D-dimensional parameter function $\theta(u)$, $u \in (0,1)$, which defines the evolutionary spectral density $f_{\theta(u)}(\lambda)$. A typical example is the time-varying ARMA process (2.1) with evolutionary spectral density (2.5).

In that context, Dahlhaus and Neumann [26] suggested to use a minimum distance method for the estimation of $\theta(\cdot)$, which is based on a distance between the evolutionary spectral density and some nonparametric pre-estimate of it. We follow this method, and first need to define a suitable nonparametric pre-estimate in the next subsection.

2.3.1 The preperiodogram

Motivated by the convergence result in Proposition 2.1, Neumann and von Sachs [85] define the *preperiodogram* as

$$J_T(u,\lambda) = \frac{1}{2\pi} \sum_k X_{[uT+\frac{k+1}{2}],T} X_{[uT-\frac{k-1}{2}],T} \exp(-ik\lambda) \tag{2.6}$$

where the sum over k if for $k \in \mathbb{Z}$ such that $1 \leqslant [uT-(k-1)/2], [uT+(k+1)/2] \leqslant T$. (In fact, there is a slight difference with the definition of Neumann and von Sachs [85] due to a time-shift in the indices.)

Dahlhaus [24] has derived a meaningful relation between the preperiodogram and the ordinary periodogram. Recall that the *periodogram* is defined as

$$\begin{aligned} I_T(\lambda) &= \frac{1}{2\pi T}\Big|\sum_{t=1}^{T} X_{t,T}\exp(-i\lambda t)\Big|^2 \qquad (2.7)\\ &= \frac{1}{2\pi}\sum_{k=-(T-1)}^{T-1}\left(\frac{1}{T}\sum_{t=1}^{T-|k|} X_{t,T}X_{t+|k|,T}\right)\exp(-i\lambda k). \end{aligned}$$

The periodogram is a widely used tool in the context of stationary processes, i.e. when the spectral density does not vary in time [14, 95]. For stationary processes, it is the Fourier transform of the covariance estimator of lag k over the whole segment of time. In contrast, the preperiodogram $J_T(t/T, \lambda)$ of a locally stationary process just uses the pair $X_{[uT-(k-1)/2],T} X_{[uT+(k+1)/2],T}$ as a "local estimator" of the covariance of lag k at time point $[uT]$ (because $[uT+(k+1)/2]-[uT-(k-1)/2] = k$). That is the reason why Neumann and von Sachs [85] also call $J_T(t/T, \lambda)$ the *localised periodogram*.

What Dahlhaus [24] pointed out is the following relation between the ordinary periodogram and the preperiodogram:

$$I_T(\lambda) = \frac{1}{T} \sum_{t=1}^{T} J_T \left(\frac{t}{T}, \lambda \right) \tag{2.8}$$

that is, the periodogram is the average of the preperiodogram over time.

The preperiodogram may be regarded as a raw estimate of the ESD at time u and frequency λ. Similarly to the behaviour of the ordinary periodogram for stationary processes, the preperiodogram of locally stationary time series is asymptotically unbiased but has a diverging variance as $T \to \infty$. In the following, it is used as a pre-estimator of the evolutionary spectral density. The advantage of this definition is that it does not contain any implicit smoothing, neither in frequency nor in time. Then, the decision about the degree of smoothing in each of these directions is left to the major smoothing step itself [85]. This is in contrast with other pre-estimators of the evolutionary spectral density proposed in the literature, like the periodogram computed on small segments of time of length N [23, 100]. In that case, an additional parameter, the segment length N, acts like a smoothing parameter in time direction.

2.3.2 The contrast function

Suppose we observe data $\{X_{1,T}, \ldots, X_{T,T}\}$ from a locally stationary process with evolutionary spectral density $f(u, \lambda)$. If the goal of the analysis is the estimation of the evolutionary spectral density $f(u, \lambda)$, then we can use a fully nonparametric estimate (e.g. by smoothing the preperiodogram [85], or using the procedure developed in the next chapters).

In this chapter, our goal is to fit a semiparametric model $f_{\theta(u)}(\lambda)$ to the data. We do not assume that f obeys the structure of the semi-

parametric model to be fitted. In other words, we do not assume that the evolutionary spectral density generating the process takes the form $f_{\theta(u)}(\lambda)$.

The distance between the semiparametric model f_θ and the true evolutionary spectral density generating the process f is measured by a *contrast function*. Here, we use

$$\mathcal{L}(\theta) = \frac{1}{4\pi} \int_0^1 du \int_{-\pi}^{\pi} d\lambda \left\{ \log f_{\theta(u)}(\lambda) + \frac{f(u,\lambda)}{f_{\theta(u)}(\lambda)} \right\},$$

which is up to a constant the asymptotic Kullback-Leibler information divergence of a locally stationary process [22]. Thus, we define the *empirical contrast function*

$$\mathcal{L}_T(\theta) = \frac{1}{4\pi T} \sum_{t=1}^{T} \int_{-\pi}^{\pi} d\lambda \left\{ \log f_{\theta(t/T)}(\lambda) + \frac{J_T(t/T,\lambda)}{f_{\theta(t/T)}(\lambda)} \right\}$$

where $J_n(t/T,\lambda)$ is the preperiodogram. $\mathcal{L}_T(\theta)$ is an approximation to the negative log-likelihood of locally stationary stationary process [24].

Remark 2.1. For stationary processes, i.e. if $f(u,\lambda) = f(\lambda)$, equation (2.8) implies that $\mathcal{L}_T(\theta)$ is the classical Whittle likelihood. For univariate stationary processes with mean zero, Whittle [119] introduced an approximation of the negative Gaussian likelihood. This approximation has been used in many different situations. A general overview may be found in the monograph of Dzhaparidze [35]. ◊

2.3.3 The sieve estimator

Our aim is to develop a nonparametric estimator of the parameter curve $\theta(\cdot) = (\theta^{(1)}(\cdot), \ldots, \theta^{(D)}(\cdot))$. In the following, we include the case of model-misspecification, that is we do not assume that the true spectral density $f(u,\lambda)$ follows the semiparametric structure $f_{\theta(u)}(\lambda)$. Hence, our estimator will not converge to the true parameter curve (which does not exist) but to

$$\theta^\circ = \arg\min_{\theta \in \Theta} \mathcal{L}(\theta).$$

Theoretically, an estimator may be constructed by minimizing the empirical contrast function $\mathcal{L}_T(\theta)$ over the class Θ of parameter curves.

Such estimator is called a *minimum contrast estimator*. However, this minimisation procedure may pose serious numerical (computational) problems, in particular if the class Θ is a complicated infinite dimensional space. Another problem arising when the set of parameters is too large, is that we could get suboptimal rates of convergence (as compared to the minimax risk). This phenomenon has been observed in even simpler contexts, e.g. for the maximum likelihood estimator with iid data [9].

The approach we follow here is based on the *method of sieves*, as named by Grenander [41]. In this chapter, each component $\theta^{(i)}$ of the target curve is approximated in a *finite-dimensional* and *linear* space of approximation $\mathcal{F}$. This means that the empirical contrast function $\mathcal{L}_T(\theta)$ is minimised over the product space $\mathcal{F}^D := \mathcal{F} \otimes \ldots \otimes \mathcal{F}$, where the dimension of $\mathcal{F}^D$ is D times the dimension of one single $\mathcal{F}$. The space $\mathcal{F}^D$ may be considered as an approximation space of Θ. The resulting estimator is denoted by $\hat{\theta}_{\mathcal{F}}$. Figure 2.1 summarizes this estimation procedure.

The first main result of this chapter concerns the convergence of $\hat{\theta}_{\mathcal{F}}$ (Theorem 2.1) to θ°. The distance for measuring this convergence is the L^2-norm: if θ° and γ are two D-dimensional curves, we define

$$\|\theta^\circ - \gamma\|_2^2 := \sum_{i=1}^{D} \int_0^1 du \, \left(\theta^{(i)}(u) - \gamma^{(i)}(u)\right)^2 .$$

Then, our main Theorem 2.1 to be formaly stated below basically says that

$$\mathrm{E}\|\theta^\circ - \hat{\theta}_{\mathcal{F}}\|_2 \lesssim \|\theta^\circ - \theta_{\mathcal{F}}\|_2 + c_\theta \sqrt{\frac{D \cdot \dim(\mathcal{F})}{T}} + O\left(\frac{1}{\sqrt{T}}\right) \qquad (2.9)$$

where the symbol $\lesssim$ means less or equal up to a finite constant independent of the parameters, where

$$\theta_{\mathcal{F}} = \arg\min_{\theta \in \mathcal{F}^D} \mathcal{L}(\theta)$$

and

$$\hat{\theta}_{\mathcal{F}} = \arg\min_{\theta \in \mathcal{F}^D} \mathcal{L}_T(\theta) \ ,$$

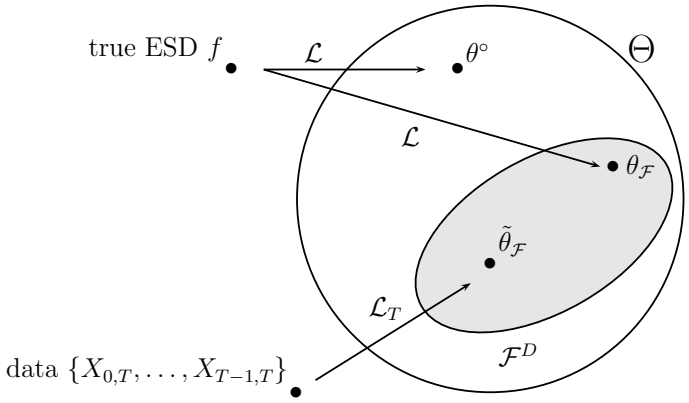


Figure 2.1: This picture illustrates the estimation procedure on one single sieve $\mathcal{F}^D$. The true evolutionary spectral density f is not assumed to follow a semiparametric structure. The space Θ is the space of all ESD that follow the semiparametric structure with fixed D, with a certain regularity on the coefficients θ for instance. The distance between the true ESD f and Θ is given by the Kullback-Leibler information divergence $\mathcal{L}$, and defines the point θ° of Θ as the "nearest" density in Θ from f. $\mathcal{F}^D$ is a finite-dimensional approximation space (a sieve) in Θ and we define $\theta_{\mathcal{F}} \in \mathcal{F}^D$ similarly to the definition of θ°. Empirically, if we observe data $\{X_{0,T}, \ldots, X_{T-1,T}\}$ generated from the ESD f, the estimator $\tilde{\theta}_{\mathcal{F}}$ is defined through the empirical distance $\mathcal{L}_T$, i.e., the Whittle likelihood of locally stationary processes (see text).

and where c_θ is a constant depending on θ° and the $O()$ term does not depend on $\theta_m, \hat{\theta}_m, \theta^\circ, D$ and $\mathcal{F}$. The first term of the right-hand side of (2.9) is known as the bias term, and the second as the variance term of the risk.

In order to be more specific on the choice of the approximation space $\mathcal{F}^D$, let us expand the component $\theta^{(i)}$ in some suitable orthonormal basis $\{\varphi_j\}$. Then, $\theta^{(i)} = \sum_{j=1}^{\infty} \theta_{ij}\varphi_j(u)$ and a typical choice for the space $\mathcal{F}$ is to take the m-dimensional linear space generated by $\varphi_1, \ldots, \varphi_m$. We denote this space by $\mathcal{F}_m$ and the correponding minimum contrast estimator by $\hat{\theta}_m := \hat{\theta}_{\mathcal{F}_m}$.

The problem of choosing the dimension m of the approximation space $\mathcal{F}_m$ occurs. Suppose we have a set $\mathcal{M}_T$ of possible dimensions. The problem is to determine from a data set some $\hat{m}$ in $\mathcal{M}_T$ in such a way that the minimum contrast estimator $\hat{\theta}_{\hat{m}}$ performs as well as the best estimator $\hat{\theta}_m$ among $m \in \mathcal{M}_T$, the criterion for comparing the estimators

being the L_2-risk. The procedure explained below provides a data-driven algorithm for the choice of $\hat{m}$, and the second main result of this chapter (Theorem 2.2) basically says that the inequality

$$\mathrm{E}\|\theta^\circ - \hat{\theta}_{\hat{m}}\|_2^2 \lesssim \inf_{m \in \mathcal{M}_T} \left\{ \|\theta^\circ - \theta_m\|_2^2 + L_m \frac{Dm}{T} \right\} + O\left(\frac{1}{T}\right)$$

holds true for some weights L_m and for $\theta_m := \theta_{\mathcal{F}_m}$. In other words, the procedure leads to a risk which, up to some multiplicative constant, realises the best trade-off between $\|\theta - \theta_m\|_2^2$ and DmL_m/T.

We now give the precise estimation procedure. Consider a collection of nested finite-dimensional linear spaces $\mathcal{F}_m, m \in \mathcal{M}_T$. The estimation procedure on sieves has two steps:

1. On each space $\mathcal{F}_m$, we minimize the empirical contrast function and compute the minimum contrast estimator

   $$\hat{\theta}_m = \arg\min_{\theta \in \mathcal{F}_m} \mathcal{L}_T(\theta)$$

   for each $m \in \mathcal{M}_T$.

2. From the set $\{\hat{\theta}_m : m \in \mathcal{M}_T\}$ of estimators, we choose $\hat{m}$ among the family $\mathcal{M}_T$ such that

   $$\hat{m} = \arg\min_{m \in \mathcal{M}} \left\{ \mathcal{L}_T\left(\hat{\theta}_m\right) + \mathrm{pen}(m) \right\}$$

   where $\mathrm{pen}(m)$ is a penalty function to be specified later.

Finally, the sieve estimator is

$$\hat{\theta} = \hat{\theta}_{\hat{m}}.$$

The precise form of the penalty function is derived in Theorem 2.2 below.

We note that the set $\mathcal{M}_T$ of dimensions is depending on T. The precise quantification of this dependence will be illustrated on some examples in the next subsection.

The above procedure is inspired by the work of Barron et al. [5], Birgé and Massart [10, 11], who studied several types of contrasts and estimates in various contexts but under the assumption of linearity of the contrast function, and under the assumption of independence. An

extension of the procedure to more complex estimation problems, like the estimation of the spectral density for a stationary time series or the estimation of the parameters in a β-mixing regression may be found in the literature [4, 20]. We note that the contrast function used in [4, 20] is the L^2 norm. Our situation is different and complex in the sense that we are dealing with dependent and covariance nonstationary data. Moreover, our contrast function is the Whittle likelihood, which is more natural in the context of spectral density estimation.

The next section presents some remarks on the collection of models $\{\mathcal{F}_m, m \in \mathcal{M}_T\}$ with some typical examples. Then, the two main results are presented in their precise form in Section 2.3.5. All formal considerations and proofs are deferred to Section 2.4.

2.3.4 The collections of models

The choice of a family of models $\{\mathcal{F}_m, m \in \mathcal{M}_T\}$ (i.e. the choice of a sieve) is basically guided by the approximation theory. One would like to use some sieves which are suitable for the approximation of the unknown component $\theta^{(i)}(\cdot)$ of the target curve $\theta^\circ \in \Theta$. Typical examples are trigonometric polynomials, wavelet expansions or piecewise polynomials, because their approximation properties are well studied in the literature (see De Vore and Lorentz [33] for instance).

In this chapter, each space $\mathcal{F}_m$ is a linear finite-dimensional subspace of $L^2([0,1]) \cap L^\infty([0,1])$ spanned by some orthonormal basis $\{\varphi_j; j \in \Lambda_m\}$ with $|\Lambda_m| = d_m$. For a given linear sieve, we need to describe the relationships between its L^2 and L^∞ structures. That is the reason why we introduce the two indices $\overline{r}_m$ and Φ_m, that will be involved in the upper bound for the risk of minimum contrast estimators on this sieve. These indices already play a crucial role in the work of Birgé and Massart [11] and the other work cited above. However, in our context, their definition is slightly different due to the complexity of our situation.

Consider the expansion of $\theta^{(i)}$ in the basis $\mathcal{F}_m$:

$$\theta^{(i)} = \sum_{j \in \Lambda_m} \theta_{ij} \varphi_j(u) \qquad i = 1, \ldots, D \ .$$

Denote by $\|\cdot\|_p$ the L^p-norm on $[0,1]$, and write $|\theta_{i\bullet}|_p$ for the ℓ^p-norm of the sequence θ_{ij} over $j \in \Lambda_m$ and for a fixed component i. Then, set

$$\overline{r}_m^{(i)} = \frac{1}{\sqrt{d_m D}} \sup \frac{\| \sum_{j \in \Lambda_m} \theta_{ij} \varphi_j(u) \|_\infty}{|\theta_{i\bullet}|_\infty}$$

where the supremum is taken over all sequences $(\theta_{ij})_{j\in\Lambda_m}$ (i is fixed) such that $|\theta_{i\bullet}|_\infty \neq 0$. Finally, $\overline{r}_m$ is defined by

$$\overline{r}_m = \sup_{i=1,\ldots,D} \overline{r}_m^{(i)}. \tag{2.10}$$

Next, we define

$$\Phi_m = d_m^{-1/2} \sup_{\theta^{(1)}\in\mathcal{F}_m\setminus\{0\}} \frac{\|\theta^{(1)}\|_\infty}{\|\theta^{(1)}\|_2}. \tag{2.11}$$

where $\theta^{(1)}$ refers to one single component. From Lemma 1 in [11], we derive the following identity

$$\Phi_m^2 d_m = \left\| \sum_{j\in\Lambda_m} \varphi_j^2 \right\|_\infty$$

between the index Φ_m and the system $\{\varphi_j\}$.

There is a link between these two indices, given by

$$\Phi_m \leqslant \overline{r}_m \leqslant \Phi_m \sqrt{d_m}\,,$$

see [5].

For the statement of our results, the general assumptions made on the collection of models is summarized now.

Assumption 2.2. For all $m \in \mathcal{M}_T$, the linear space $\mathcal{F}_m$ is included in $L^2([0,1]) \cap L^\infty([0,1])$ with finite dimension $\dim(\mathcal{F}_m) = |\Lambda_m| = d_m$ such that $\Lambda_T := \max_{m\in\mathcal{M}_T} \Lambda_m \leqslant T$. This space is generated by the orthonormal functions $\{\varphi_j; j \in \Lambda_m\}$ which are such that there exists a finite and positive $\tilde{v}_m$ with $\sup_{j\in\Lambda_m} \mathrm{TV}(\varphi_j) \leqslant d_m \tilde{v}_m$ uniformly in d. Moreover, $\overline{r}_m \leqslant C_{\overline{r}} \sqrt{T/d_m}$ for all $m \in \mathcal{M}_T$ and the collection of models $\{\mathcal{F}_m : m \in \mathcal{M}_T\}$ is nested, that is $\mathcal{F}_m \subset \mathcal{F}_{m'}$ for $m < m'$. ◊

Now we describe some examples of models. These are taken from the standard literature [5, 20, 33].

Example 2.1 (Trigonometric polynomials). Consider spaces $\mathcal{F}_m$ generated from the functions $\varphi_j(u) = \sqrt{2}\cos(2\pi j u)$ for $j = 0, \ldots, m-1$. The dimension of $\mathcal{F}_m$ is $d_m = m$ and any component i of the vector θ can be written $\theta^{(i)} = \sqrt{2}\sum_{j=0}^{m-1} \theta_{ij} \cos(2\pi j u)$. This collection is such that $\mathcal{M}_T = \{1, \ldots, T\}$. It follows from [33] that $C_{\overline{r}} \leqslant \sqrt{2}$ provided that $|\Lambda_T| \leqslant \sqrt{T}$, and $\tilde{v} = 1$. ◊

Example 2.2 (Polynomials). In that case, $\mathcal{F}_m$ is the linear space of polynomials on $[0,1]$ with degree bounded by $m-1$. The next example is a generalisation. $\lozenge$

Example 2.3 (Piecewise polynomials). Consider first dyadic partitions of $[0,1]$ given by $I_m = \{[j2^{-m}, (j+1)2^{-m}], j = 0, \ldots, m-1\}$. Given some integer s, the space $\mathcal{F}_m$ is defined as the space of piecewise polynomials with degree bounded by $s-1$ on the partition I_m. The dimension of $\mathcal{F}_m$ is $r2^m$ and Assumption 2.2 holds with $C_{\overline{r}} = \sqrt{(r+1)(2r+1)}$ independently of m. $\lozenge$

Example 2.4 (Regular compactly supported wavelets). Consider an orthonormal wavelet basis $\{\phi_{j,k} : j \geqslant 0, k \in \mathbb{Z}\}$ of L_2 (see Chapter 3) with the following conventions: $\psi_{0,k}$ are translated of the father wavelet and for $j \geqslant 1$, ϕ_{jk} are affine transforms of the mother wavelet. We consider this construction for compactly supported wavelets, such that the index $k \in \{1, \ldots, 2^j L\}$, $j, L > 0$ [32]. Then, we define $\mathcal{F}_m$ as the space generated by $\{\phi_{j,k}\}$ for (j,k) such that $0 \leqslant j \leqslant m$ and $k \in \{1, \ldots, 2^j L\}$. Then, $\mathcal{F}_m$ has dimension $d_m = L\sum_{j=0}^{m} 2^j = L(2^{m+1}-1)$ for all $m \in \mathcal{M}_T = \{1, \ldots, M_T\}$, where M_T is such that $L(2^{M_T+1}-1) \leqslant T$, i.e. $M_T = O(\ln T)$. From Chapter 2, Lemma 8 of Meyer [76] and from [5] it can be showed that $C_{\overline{r}}$ is bounded by a constant depending on L. $\lozenge$

Finally, we also need the following assumption which controls the number of models in each collection.

Assumption 2.3. There exists some weights L_m and a finite constant Υ such that

$$\sum_{m \in \mathcal{M}_T} \exp(-L_m d_m) \leqslant \Upsilon < \infty \, . \qquad \lozenge$$

For the three examples above, the weights L_m are of order 1. For piecewise polynomials $L_m = 1/r$ can be used [5].

2.3.5 Main results

Before stating the main results, we have to give two more assumptions. We first introduce the following notation. If $g(u,\lambda)$ is a function over $[0,1] \times (-\pi, \pi)$, then we set

$$\tilde{g}(u,j) := \int_{-\pi}^{\pi} d\lambda \, \phi(u,\lambda) \exp(i\lambda j) \, ,$$

and define

$$\rho_2(g) = \left(\int_0^1 \int_{-\pi}^{\pi} d\lambda \, du \, |g(u,\lambda)|^2 \right)^{1/2} ,$$
$$\rho_\infty(g) := \sum_{j=-\infty}^{\infty} \sup_u |\tilde{g}(u,j)| ,$$
$$\tilde{v}(g) := \sup_j \mathrm{TV}\left(\tilde{g}(\cdot, j)\right) .$$

Correspondingly, we set $\rho_2(g_1, g_2) := \rho_2(g_1 - g_2)$, $\rho_\infty(g_1, g_2) := \rho_\infty(g_1 - g_2)$ and $\tilde{v}(g_1, g_2) := \tilde{v}(g_1 - g_2)$.

If θ is a D-dimensional curve, we also need the following definitions:

$$\|\theta\|_2^2 := \sum_{i=1}^{D} \int_0^1 du \, \left(\theta^{(i)}(u)\right)^2 , \tag{2.12}$$
$$\|\theta\|_\infty := \sup_{i=1,\ldots,D} \sup_{u\in[0,1]} |\theta^{(i)}(u)| , \tag{2.13}$$
$$\mathrm{TV}(\theta) := \sum_{i=1}^{D} \mathrm{TV}(\theta^{(i)}) . \tag{2.14}$$

We can now formulate the assumptions. The first assumption is on the total variation norm of the evolutionary spectrum with respect to its time argument.

The second assumption is needed to describe the relationship between, on one side, the distance between the two spectra f_θ and f_γ and, on the other side, the distance between the corresponding curves θ and γ.

Assumption 2.4. The norms ρ_2 and $\|\cdot\|_2$ are equivalent, i.e. there exists two constants K_2, K_2' (which may depend on D) such that

$$K_2'\|\theta - \theta^\star\|_2 \leqslant \rho_2\left(1/f_\theta - 1/f_{\theta^\star}\right) \leqslant K_2\|\theta - \theta^\star\|_2. \tag{2.15}$$

for all θ and $\theta^\star$ in Θ. Moreover, there exists constants K_∞ and K_{tv} depending on D such that

$$\rho_\infty\left(1/f_\theta - 1/f_{\theta^\star}\right) \leqslant K_\infty\|\theta - \theta^\star\|_\infty \tag{2.16}$$
$$\mathrm{TV}\left(1/f_\theta - 1/f_{\theta^\star}\right) \leqslant K_{tv}\,\mathrm{TV}(\theta - \theta^\star) \tag{2.17}$$

for all θ and $\theta^\star$ in Θ. $\Diamond$

To illustrate this last assumption, let us consider a simple example. We consider a time-varying AR(p) model with a known, constant variance σ^2. In this example, the vector θ is the set of p time-varying parameters of the model, $(a_1(u), \dots, a_p(u))$. From (2.5), we see that the ESD of this process is given by

$$f_{\theta(u)} = \frac{\sigma^2}{2\pi} \Big| \sum_{j=0}^{p} a_j(u) \exp(i\lambda j) \Big|^{-2} ,$$

and it is a simple exercice to derive the bounds of Assumption 2.4 in this model. The equivalence between the two quadratic norms holds with constants $K_2 = \|1/f_\theta\|_\infty \sqrt{2\pi}/\sigma$ and $K_2' = \|1/f_\theta\|_\infty^{-1} \sqrt{2\pi}/\sigma$. Here, we used Assumption 2.1, which ensures that the ESD and its inverse are uniformly bounded (points 2 and 3 of Assumption 2.1). The other constants of Assumption 2.4 are $K_\infty = \sqrt{2\pi}/\sigma$ and $K_{tv} = 2(2\pi)^2/\sigma^4$, if we use $|a_j(u)| \leqslant 1$.

The first result is on the mean square error for the estimation of θ by $\hat{\theta}_m$ for a fixed space $\mathcal{F}_m$. In the formulation of the result, we denote by Σ the covariance matrix of the process $\{X_{t,T}\}$, i.e. the entry (s,t) of Σ is $\mathrm{Cov}(X_{s,T}, X_{t,T})$. Moreover, $\|\Sigma\|_{\mathrm{spec}}$ denotes the spectral norm of the matrix Σ (see Appendix A).

Theorem 2.1. *Suppose that we observe data $X_{1,T}, \dots, X_{T,T}$ from a Gaussian locally stationary process (Definition 2.1). Under Assumptions 2.1, 2.2 and 2.4, the minimum contrast estimator $\hat{\theta}_m$ is such that*

$$\mathrm{E}\|\theta^\circ - \hat{\theta}_m\|_2 \leqslant \|\theta^\circ - \theta_m\|_2 + \omega_d + T^{-1/2}(c_1 + c_2\|\Sigma\|_{\mathrm{spec}})$$

where

$$\omega_d = \sqrt{\frac{dD}{c_2 T}} \left\{ 1 \vee \frac{\alpha K_2 \|\Sigma\|_{\mathrm{spec}}}{K_2' K_\infty \overline{r} 2\pi} \sqrt{\frac{20}{e} B + \frac{200 dD}{e^2 c_2 T} A^2} \right\} \tag{2.18}$$

with

$$A = 108\sqrt{2} K_\infty^2 \overline{r}^2 / K_2$$

and

$$B = 1 + 36\sqrt{2} K_\infty \overline{r} K_2^{-1} \sqrt{\pi} (K_2 + d\sqrt{K_{tv} D \overline{r} \tilde{v} K_\infty / T}) .$$

and where c_1, c_2 and α are specific constants depending on $\kappa_1, K_\infty, K_2, \overline{r}$ and derived in the proof below.

The second theorem is about the estimator $\hat{\theta}_{\hat{m}}$ computed from the data-driven procedure described above.

Theorem 2.2. *Suppose that we observe data $X_{1,T}, \ldots, X_{T,T}$ from a Gaussian locally stationary process (Definition 2.1) and suppose that Assumptions 2.1 to 2.4 hold true. Moreover, let us suppose that the space $\Theta = \{\mathcal{F}_m : m \in \mathcal{M}_T\}$. With the same constant α as in the previous theorem, we define*

$$\tau = 8\pi\alpha/K_2' \tag{2.19}$$

$$\Phi = \inf_{m \in \mathcal{M}_T} \sqrt{d_m}\Phi_m \tag{2.20}$$

$$A = K_2 + 2T^{-1}K_{tv}K_\infty\tilde{v}\Phi \tag{2.21}$$

$$B = K_\infty\kappa_s^2 \tag{2.22}$$

$$\zeta = 4\left\{\left(2\pi A \vee \frac{K_2^2}{K_\infty^2\overline{r}^2}\right)\|\Sigma\|_{\text{spec}}^2 + B\|\Sigma\|_{\text{spec}}\tau\right\}/\tau^2\,, \tag{2.23}$$

where the constants are the constants of Assumptions 2.1 and 2.4. If the penalty function $\text{pen}(\cdot)$ *defined on $\mathcal{M}_T$ satisfies*

$$\text{pen}(m) \geqslant \frac{4\pi}{\tau}\left(\omega_{d_m}^2 \vee \frac{\zeta L_m d_m}{T}\right) + \frac{\kappa_1}{8\pi T} + \frac{\overline{C}}{2T\tau}\left(\rho_\infty + \tilde{v}\right),$$

where the constant $\overline{C}$ is derived in the proof, then

$$\mathrm{E}\|\theta^\circ - \hat{\theta}_{\hat{m}}\|_2^2 \leqslant \inf_{m \in \mathcal{M}_T}\left\{\|\theta^\circ - \theta_m\|_2^2 + \frac{8\pi}{\tau}\,\text{pen}(m)\right\} + \frac{3.6\Upsilon\zeta + 1}{T}.$$

In this result, we assumed that the parameter space Θ coincides with the largest sieve $\mathcal{F}_m$ ($m \in \mathcal{M}_T$). It is also possible to state a similar result if Θ is larger than this largest sieve. In that case, an additional bias term appears in the right hand side of the inequality. This bias measures the distance in Θ between θ° and

$$\theta^T := \arg\min \mathcal{L}(\theta)\,,$$

where the minimum is over θ in the largest sieve $\mathcal{F}_m$. This bias is proportional to $\mathcal{L}(\theta^\circ) - \mathcal{L}(\theta^T)$.

A comparison between Theorem 2.1 and Theorem 2.2 shows that the automatic selection of the parameter m does not increase the estimation

error significantly. More precisely, the upper bound derived in Theorem 2.1 is still valid for $\hat{\theta}_{\hat{m}}$ and is the best upper bound among all the classes $\{\mathcal{F}_m : m \in \mathcal{M}_T\}$. Of course, this comment is of asymptotic nature and the constants involved in the $O(T^{-1})$ term are different in the two theorems.

From Theorem 2.2, it is easy to derive adaptation results with respect to the unknown smoothness of f. Let us suppose for instance that each component θ_i° of the target vector θ° belongs to a Besov space $B_\infty^{\beta,2}$ (see Appendix B for a review on functional spaces). If we consider the trigonometric polynomials, the piecewise polynomials or the regular compactly supported wavelets described above, it is known from classical approximation theory (see De Vore and Lorentz [33]) that if $r \geqslant \beta$ and $\theta_i^\circ \in B_\infty^{\beta,2}$, then

$$\|\theta^\circ - \theta_m\|_2 \leqslant C(\beta) \sum_{i=1}^{D} \|\theta_i^\circ\|_{B_\infty^{\beta,2}} m^{-\beta}$$

where r is the regularity. For these models, $L_m = 1$ and the term in brackets in the upper bound of Theorem 2.2 becomes

$$\left(\sum_{i=1}^{D} \|\theta_i^\circ\|_{B_\infty^{\beta,2}} \right)^2 m^{-2\beta} + \frac{Cm}{T}$$

where C is a constant independent of T. Minimising this bound with respect to m, we derive immediately

$$\mathrm{E}\|\theta^\circ - \hat{\theta}_{\hat{m}}\|_2^2 \leqslant C \left(\sum_{i=1}^{D} \|\theta_i^\circ\|_{B_\infty^{\beta,2}} \right)^{2/(1+2\beta)} T^{-2\beta/(2\beta+1)}$$

where the constant C only depends on $\beta, \|\Sigma\|_\infty, \tau, A, B$ and ζ. In other words, the proposed estimator converges to the corresponding target θ° with a rate which is the usual rate of convergence in Besov spaces. This rate is the optimal minimax rate of convergence for a lot of problems (regression, density estimation). However, this optimality has not yet been proved in the framework of semiparametric locally stationary models, but we conjecture that this result also holds in that framework.

2.4 Formal complements and proofs

In this section, we develop the mathematical tools for proving the results of Section 2.3.5.

2.4.1 The main tool: The empirical spectral process

As usual in the context of minimum contrast estimation on sieves, the key point is to establish exponential bounds for the fluctuation of the empirical process. The *empirical spectral process* [27, 28] is defined by

$$E_T(\phi) = \sqrt{T}\,(F_T - F)\,(\phi)$$

where

$$F(\phi) = \int_0^1 du \int_{-\pi}^{\pi} d\lambda\ \phi(u,\lambda) f(u,\lambda)$$

and

$$F_T(\phi) = \frac{1}{T}\sum_{t=1}^{T} \int_{-\pi}^{\pi} d\lambda\ \phi\left(\frac{t}{T},\lambda\right) J_T\left(\frac{t}{T},\lambda\right).$$

In order to explain the need of the empirical spectral process, a useful connection with the contrast functions $\mathcal{L}_T(\cdot)$ and $\mathcal{L}(\cdot)$ can be derived [28]. By definition of θ_m, the inequality $\mathcal{L}(\theta_m) \leqslant \mathcal{L}(\theta)$ holds for all $\theta \in \mathcal{F}_m$. Similarly, by definition of $\hat{\theta}_m$, it holds $\mathcal{L}_T(\hat{\theta}_m) \leqslant \mathcal{L}_T(\theta)$ for all $\theta \in \mathcal{F}_m$. Combining these two inequalities, we get

$$\begin{aligned} 0 &\leqslant \mathcal{L}(\hat{\theta}_m) - \mathcal{L}(\theta_m) \\ &\leqslant (\mathcal{L}_T - \mathcal{L})\,(\theta_m) - (\mathcal{L}_T - \mathcal{L})\,(\hat{\theta}_m) \\ &\leqslant \frac{1}{4\pi\sqrt{T}} E_T\left(\frac{1}{f_{\theta_m}} - \frac{1}{f_{\hat{\theta}_m}}\right) + R(\theta_m) - R(\hat{\theta}_m) \end{aligned} \tag{2.24}$$

where

$$\begin{aligned} R(\theta) &:= \frac{1}{4\pi}\int_{-\pi}^{\pi} d\lambda\ \left\{\frac{1}{T}\sum_{t=1}^{T} \log f_{\theta(t/T)}\,(\lambda) - \int_0^1 du\ \log f_{\theta(u)}(\lambda)\right\} \\ &= \frac{1}{4\pi}\int_{-\pi}^{\pi} d\lambda\ \sum_{t=1}^{T}\int_0^{1/T} du\,\big\{\log f_{\theta(t/T)}\,(\lambda) \\ &\qquad\qquad - \log f_{\theta(u+(t-1)/T)}\,(\lambda)\big\}. \end{aligned}$$

Thus,

$$|R(\theta)| \leqslant \frac{1}{4\pi T} \int_{-\pi}^{\pi} d\lambda \, \mathrm{TV}\left(\log f_{\theta(\cdot)}(\lambda)\right) \leqslant \kappa_1/(4\pi T) = O\left(T^{-1}\right) \tag{2.25}$$

by Assumption 2.1(2) that implies $\sup_{-\pi \leqslant \lambda < \pi} \mathrm{TV}_{[0,1]} \log f_{\theta(\cdot)}(\lambda) \leqslant \kappa_1 < \infty$, where κ_1 is a constant depending only on the constant κ_s. Thus, (2.24) indicates that the convergence of $\mathcal{L}(\hat{\theta}_m) - \mathcal{L}(\theta_m)$ depends only on the behaviour of the empirical spectral process on $(1/f_{\theta_m}) - (1/f_{\hat{\theta}_m})$.

In Section 2.4.2 we derive an exponential inequality on the empirical spectral process. This point is the most technical part of the proof. Then, we combine this element with (2.24) and show that the convergence of $\|\theta - \theta_m\|_2^2$ depends on the behaviour of the empirical spectral process on $(1/f_{\theta_m}) - (1/f_{\hat{\theta}_m})$. Finally, in Section 2.4.3, we can derive the bound given in Theorem 2.1. The exponential bound is also used in Section 2.4.4 for proving Theorem 2.2.

2.4.2 Maximal exponential inequality

An *exponential inequality*, or *Bernstein inequality*, is an exponential bound for the probability of deviation of an empirical process. This inequality is a *maximal inequality* if it holds on the deviation of the *supremum* of the empirical process, over a certain class of functions denoted by $\mathcal{G}$ in this section. We also denote by d the dimension of $\mathcal{G}$. In our applications, in particular in the proof of the first theorem, $\mathcal{G}$ (resp. d) will be replaced by a fixed sieve $\mathcal{F}_m$ (resp. the dimension d_m).

Preliminaries

For the sake of clarity, we first decompose the empirical spectral process as

$$E_T = \tilde{E}_T + \overline{E}_T$$

where

$$\tilde{E}_T = \sqrt{T}\,(F_T - \mathrm{E}F_T) \tag{2.26}$$

is a stochastic term and

$$\overline{E}_T = \sqrt{T}\,(\mathrm{E}F_T - F)\,.$$

is a deterministic term. In the following, a maximal exponential inequality is proved for the process $\tilde{E}_T$.

The usual way for proving maximal inequalities is to start with a Bernstein inequality and then to use the chaining technique, provided that the complexity (entropy) of $\mathcal{G}$ is well controlled [93, 114]. We follow this scheme in our proof, and start by presenting a lemma with the Bernstein inequality.

Lemma 2.1 ([28]). *Suppose that $\{X_{t,T}\}$ is a locally stationary process (Definition 2.1) and suppose that the function $\phi : [0,1] \times [-\pi, \pi] \to \mathbb{R}$ is such that $\rho_\infty(\phi) < \infty$, $\rho_2(\phi) < \infty$ and $\tilde{v}(\phi) < \infty$ (these norms are defined in Section 2.3.5). Set*

$$\rho_{2,T}(\phi) = \left\{ \frac{1}{T} \sum_{t=1}^{T} \int_{-\pi}^{\pi} d\lambda \; \phi\left(\frac{t}{T}, \lambda\right) \right\}^{1/2}$$

and define the process $\tilde{E}_T$ as in (2.26). Then the Bernstein inequality

$$\Pr\left[|\tilde{E}_T(\phi)| \geqslant 2\|\Sigma^{1/2}\|_{\text{spec}}^2 \sqrt{T} \left(2\xi \; \rho_\infty(\phi) + \sqrt{2\pi\xi} \; \rho_{2,T}(\phi) \right) \right] \leqslant \exp\left(-T\xi\right)$$

holds true for all $\xi > 0$.

Proof of the lemma. Straightforward application of Theorem 3.4 in Dahlhaus and Polonik [28]. □

Note that

$$\rho_{2,T}(\phi) \leqslant \rho_2(\phi) + \sqrt{\frac{\rho_\infty(\phi)\tilde{v}(\phi)}{T}}, \tag{2.27}$$

then we can replace $\rho_{2,T}(\phi)$ by this upper bound in the Bernstein inequality. In the following, we also use the following alternative formulation of Lemma 2.1:

$$\Pr\left(|\tilde{E}_T(\phi)| \geqslant \eta \right) \leqslant \exp\left(-\frac{1}{4} \cdot \frac{\eta^2}{2\pi\|\Sigma^{1/2}\|_{\text{spec}}^4 \rho_{2,T}^2(\phi) + \|\Sigma^{1/2}\|_{\text{spec}}^2 \frac{\rho_\infty(\phi)\eta}{\sqrt{T}}} \right) \tag{2.28}$$

Above, we have mentioned that we also need a control on the complexity of the approximation space. The next lemma shows that this is precisely the technical advantage of using the method of sieves.

Lemma 2.2 ([5]). *Suppose that $\mathcal{G}$ is a finite-dimensional linear space of dimension d such that Assumption 2.2 holds, and define the product space $\mathcal{G}^D = \mathcal{G}\otimes\cdots\otimes\mathcal{G}$. Then, for any positive δ one can find a countable set $\mathcal{E}(\delta) \subset \mathcal{G}^D$ and a mapping $\mu : \mathcal{G}^D \to \mathcal{E}(\delta)$ such that*

- *For each ball $\mathcal{B}$ in $\mathbb{R}^D$ with radius $\omega \geqslant 5\delta$, $|\mathcal{E}(\delta) \cap \mathcal{B}| \leqslant (5\omega/\delta)^{dD}$,*
- *$\|\theta - \mu(\theta)\|_2 \leqslant \delta$ for all $\theta \in \mathcal{G}^D$,*
- *$\sup_{t\in\mathcal{E}(\delta)} \|t - \mu^{-1}(t)\|_\infty \leqslant \overline{r}\delta$ for all $t \in \mathcal{N}(\delta)$,*

where $\overline{r}$ is defined in (2.10). (The norms are defined in (2.13) and (2.12).)

Proof of the lemma. Straightforward generalisation to the multidimensional case of Lemma 9 in Barron et al. [5]. □

Observe that, with Assumption 2.2, we may bound the total variation norm between $\theta \in \mathcal{G}$ and $\mu(\theta) \in T(\delta)$. With $\theta^{(i)} = \sum_j \theta_{ij}\varphi_j$ and $\mu(\theta^{(i)}) = \sum_j \bar{\theta}_{ij}\varphi_j$, we get

$$\begin{aligned}
&\sum_{i=1}^{D} \mathrm{TV}\left(\theta^{(i)} - \mu(\theta^{(i)})\right) \\
&\qquad \leqslant \sum_{i=1}^{D} \sup_{\substack{0\leqslant u_0<\ldots<u_N\leqslant 1 \\ N\in\mathbb{N}_0}} \sum_{j\in\Lambda} |\theta_{ij} - \bar{\theta}_{ij}| \sum_{k=0}^{N-1} \left|\varphi_{ij}(u_{k+1}) - \varphi_{ij}(u_k)\right| \\
&\qquad \leqslant D\delta\overline{r}\tilde{v}d^2 \,. \qquad (2.29)
\end{aligned}$$

Chaining on a ball

Fix γ in $\mathcal{G}$. We shall first prove a maximal inequality on a ball $\mathcal{B}(\gamma,\omega)$ centered in γ with radius $\omega > 0$, included in $\mathcal{G}$. More precisely, our goal now is to derive an exponential bound for

$$\mathcal{P}_1 := \Pr\left\{ \sup_{\theta\in\mathcal{B}(\gamma,\omega)} \left|\tilde{E}_T\left(\frac{1}{f_\theta} - \frac{1}{f_\gamma}\right)\right| > \sqrt{T}\frac{K_2}{K_\infty\overline{r}}\|\Sigma\|_{\mathrm{spec}}\xi\omega^2 \right\} \,.$$

To this end, we start with the chaining argument. We fix the sequence $\delta_k = 2^{-k}\delta_0$, $k = 0, 1, \ldots$ (δ_0 will be fixed later on). A straightforward application of Lemma 2.2 shows that there exists a sequence of subsets $\mathcal{E}(\delta_k) \subset \mathcal{G}$ such that $5\delta_k \leqslant \omega$ and

- $|\mathcal{E}(\delta_k) \cap \mathcal{B}| \leqslant (5\omega/\delta_k)^{dD}$,
- Given $\theta \in \mathcal{B}$, there exists a sequence (θ_k) with $\theta_k \in \mathcal{E}(\delta_k)$ and such that $\|\theta - \theta_k\|_2 \leqslant \delta_k$ and $\|\theta - \theta_k\|_\infty \leqslant \overline{r}\delta_k$ hold.

Given some point $\theta \in \mathcal{B}(\gamma, \omega)$, we select an element θ_k in $\mathcal{N}(\delta_k)$ for each k. Then, $\theta_k \to \theta$ in the L^2 and the L^∞ norms, and

$$\theta = \theta_0 + \sum_{k=1}^{\infty} (\theta_k - \theta_{k-1}) .$$

Then, if we choose the sequence (ξ_k) such that

$$\sum_k \xi_k \leqslant (K_2/K_\infty \overline{r}) \|\Sigma\|_{\text{spec}} \xi \omega^2, \tag{2.30}$$

we can write

$$\begin{aligned}
\mathcal{P}_1 &\leqslant \sum_{\theta_0 \in \mathcal{E}(\delta_0)} \Pr\left\{ \left| \tilde{E}_T \left(\frac{1}{f_{\theta_0}} - \frac{1}{f_\gamma} \right) \right| > \xi_0 \sqrt{T} \right\} \\
&\qquad + \sum_{k=1}^{\infty} \sum_{\substack{\theta_k \in \mathcal{E}(\delta_k) \\ \theta_{k-1} \in \mathcal{E}(\delta_{k-1})}} \Pr\left\{ \left| \tilde{E}_T \left(\frac{1}{f_{\theta_k}} - \frac{1}{f_{\theta_{k-1}}} \right) \right| > \xi_k \sqrt{T} \right\} \\
&=: P_0 + \sum_{k=1}^{\infty} P_k .
\end{aligned} \tag{2.31}$$

We now bound $P_0, P_k, k \geqslant 1$. Set $H_k = \ln |\mathcal{E}(\delta_k)|$. Using the Bernstein inequality (Lemma 2.1), we get $P_0 \leqslant \exp(H_0 - c_2 T \eta_0)$ provided that

$$\begin{aligned}
\xi_0 = 2\|\Sigma^{1/2}\|_{\text{spec}}^2 \Bigg\{ & 2\eta_0 \, \rho_\infty \left(\frac{1}{f_{\theta_0}} - \frac{1}{f_\gamma} \right) \\
& + \sqrt{2\pi\eta_0} \, \rho_{2,T} \left(\frac{1}{f_{\theta_0}} - \frac{1}{f_\gamma} \right) \Bigg\} ,
\end{aligned}$$

where η_0 will be chosed later. Using (2.27) with assumptions (2.16), (2.15) and (2.17), ξ_0 is bounded by

$$2\|\Sigma^{1/2}\|_{\text{spec}}^2\left\{2\eta_0 K_\infty\|\theta_0-\gamma\|_\infty\right.$$
$$\left.+\sqrt{2\pi\eta_0}\left(K_2\|\theta_0-\gamma\|_2+\sqrt{\frac{K_{tv}\,\text{TV}(\theta_0-\gamma)K_\infty\|\theta_0-\gamma\|_\infty}{T}}\right)\right\}.$$

Similarly, $P_k \leqslant \exp(H_k + H_{k-1} - c_2 T\eta_k)$ with

$$\xi_k \leqslant 2\|\Sigma^{1/2}\|_{\text{spec}}^2\left\{2\eta_k K_\infty\|\theta_k-\theta_{k-1}\|_\infty+\sqrt{2\pi\eta_k}\times\right.$$
$$\left.\times\left(K_2\|\theta_k-\theta_{k-1}\|_2+\sqrt{\frac{K_{tv}\,\text{TV}(\theta_0-\gamma)K_\infty\|\theta_k-\theta_{k-1}\|_\infty}{T}}\right)\right\}.$$

Now, we set L such that the inequality $L \geqslant \xi^2 \vee 2\ln(5\alpha)$ holds with

$$\alpha := 1+36\sqrt{2}\frac{K_\infty\overline{r}}{K_2}\left\{3K_\infty\overline{r}\sqrt{\frac{dDL}{c_2T}}\right.$$
$$\left.+\sqrt{\pi}\left(K_2+\sqrt{\frac{K_{tv}dD\overline{r}\tilde{v}K_\infty}{T}}\right)\right\}. \quad (2.32)$$

We also choose $\delta_0 = \omega/\alpha$ and assume that the radius of the ball is such that

$$\xi\omega = \sqrt{dDL/c_2T}. \quad (2.33)$$

Now, we choose η_0, η_k such that $c_2T\eta_0 = H_0 + dDL$ and $c_2T\eta_k = H_k + H_{k-1} + (k+1)dDL$ for $k \geqslant 1$. From (2.31), this leads to

$$\begin{aligned}
&\Pr\left\{\sup_{\theta\in\mathcal{B}(\gamma,\omega)}\left|\tilde{E}_T\left(\frac{1}{f_\theta}-\frac{1}{f_\gamma}\right)\right| > \sqrt{T}\frac{K_2}{K_\infty\overline{r}}\|\Sigma\|_{\text{spec}}\xi\omega^2\right\}\\
&\quad\leqslant \exp(-dDL)\left\{1+\sum_{k=1}^{\infty}\exp(-kdDL)\right\}\\
&\quad\leqslant \exp(-dDL)\left\{1-\exp(-dDL)\right\}^{-1}\\
&\quad\leqslant e(e-1)^{-1}\exp(-dDL)\\
&\quad= e(e-1)^{-1}\exp(-c_2\omega^2\xi^2T) \quad (2.34)
\end{aligned}$$

which is the maximal exponential inequality on the ball $\mathcal{B}(\gamma, \omega)$, for a radius ω such that (2.33) holds, and provided that (2.30) and $dDL/2 \geqslant 1$ hold true. As $dD \geqslant 1$, the last constraint holds since $\alpha \geqslant 1$ and then $L \geqslant 2$. Moreover, a very long computation derived at the end of this section shows that (2.30) indeed holds true. Then, the maximal inequality on the ball $\mathcal{B}(\gamma, \omega)$ is proved provided that (2.33) holds true, i.e. $\omega^2 \geqslant dD(c_2 T)^{-1}\{1 \vee 2\xi^{-2} \ln(5\alpha)\}$. In what follows, we show that a sufficient condition for this inequality is

$$\omega^2 \geqslant \frac{dD}{c_2 T} \left\{ 1 \vee \frac{1}{\xi^2} \left(\frac{20}{e} B + \frac{200 dD}{e^2 c_2 T} A^2 \right) \right\} \tag{2.35}$$

where $A := 108\sqrt{2} K_\infty^2 \overline{r}^2 / K_2$ and

$$B := 1 + 36\sqrt{2} K_\infty \overline{r} K_2^{-1} \sqrt{\pi} (K_2 + d\sqrt{K_{tv} D \overline{r} \tilde{v} K_\infty / T}).$$

Indeed, with $\ln |x| \leqslant |x|/e$,

$$\begin{aligned} \frac{2dD}{c_2 T} \ln(5\alpha) \leqslant \frac{10 dD}{e c_2 T} \alpha &= \frac{10 dD}{e c_2 T} \left(A \sqrt{\frac{dDL}{c_2 T}} + B \right) \\ &= \frac{10 dD}{e c_2 T} (A\xi\omega + B) \end{aligned}$$

and with (2.35),

$$\begin{aligned} &\leqslant \frac{10 dDA}{e c_2 T} \omega\xi + \frac{1}{2}\xi^2\omega^2 - \frac{100 d^2 D^2 A^2}{e^2 c_2^2 T^2} \\ &= \frac{3}{4}\xi^2\omega^2 - \left(\frac{10 dDA}{e c_2 T} - \frac{1}{2}\omega\xi \right)^2 \\ &\leqslant \xi^2\omega^2 \end{aligned}$$

Then the exponential inequality (2.34) on a ball $\mathcal{B}(\gamma, \omega)$ holds provided that ω obeys (2.35).

Extension to $\mathcal{G}$

In order to proof the maximal inequality on the whole space $\mathcal{G}$, we define $\omega_0 = 0$ and $\omega_j = 2^j\omega$, $j > 0$. Then

$$\begin{aligned}
&\Pr\left\{\sup_{\theta\in\mathcal{G}}\frac{\left|\tilde{E}_T\left(\frac{1}{f_\theta}-\frac{1}{f_\gamma}\right)\right|}{\omega^2\vee\|\theta-\gamma\|_2^2} > \tau\sqrt{T}\right\}\\
&\quad\leqslant \sum_{j=0}^{\infty}\Pr\left\{\sup_{\theta\in\mathcal{G};\omega_j^2\leqslant\|\theta-\gamma\|_2^2<\omega_{j+1}^2}\frac{\left|\tilde{E}_T\left(\frac{1}{f_\theta}-\frac{1}{f_\gamma}\right)\right|}{\omega_j^2} > \tau\sqrt{T}\right\}\\
&\quad\leqslant \sum_{j=0}^{\infty}\Pr\left\{\sup_{\theta\in\mathcal{B}(\gamma,\omega_{j+1})}\left|\tilde{E}_T\left(\frac{1}{f_\theta}-\frac{1}{f_\gamma}\right)\right| > \omega_j^2\tau\sqrt{T}\right\}. \quad (2.36)
\end{aligned}$$

We can now use the Bernstein inequality on the balls $\mathcal{B}(\gamma,\omega_{j+1})$, with $\tau = K_2\|\Sigma\|_{\text{spec}}\xi/(K_\infty\overline{r})$. From (2.35), with condition

$$\omega^2 \geqslant \frac{dD}{c_2T}\left\{1\vee\frac{K_2^2\|\Sigma\|_{\text{spec}}^2}{K_\infty^2\overline{r}^2\tau^2}\left(\frac{20}{e}B+\frac{200dD}{c_2e^2T}A^2\right)\right\} \quad (2.37)$$

we can bound (2.36) from above by:

$$\begin{aligned}
&\frac{e}{e-1}\sum_{j=0}^{\infty}\exp\left(-\frac{c_2T\tau^2K_\infty^2\overline{r}^2\omega_j^2}{K_2^2\|\Sigma\|_{\text{spec}}^2}\right)\\
&\quad\leqslant \frac{e}{e-1}\exp\left(-\frac{c_2T\tau^2K_\infty^2\overline{r}^2\omega^2}{K_2^2\|\Sigma\|_{\text{spec}}^2}\right)\times\\
&\qquad\times\left\{1+\sum_{j=1}^{\infty}\exp\left(-\frac{c_2T\tau^2K_\infty^2\overline{r}^2\omega^2(2^{2j}-1)}{K_2^2\|\Sigma\|_{\text{spec}}^2}\right)\right\}\\
&\quad\leqslant \frac{e^2}{(e-1)^2}\exp\left(-\frac{c_2T\tau^2K_\infty^2\overline{r}^2\omega^2}{K_2^2\|\Sigma\|_{\text{spec}}^2}\right).
\end{aligned}$$

since (2.37) with $dD > 1$ implies that $K_\infty^2\overline{r}^2\tau^2\omega^2T \geqslant K_2^2\|\Sigma\|_{\text{spec}}^2$.

We sumarize the result in the following proposition.

Proposition 2.2 (Maximal Inequality). *Under assumption 2.4 and 2.2, for all* $\gamma \in \mathcal{G}$,

$$\Pr\left\{\sup_{\theta\in\mathcal{G}} \frac{\left|\tilde{E}_T\left(\frac{1}{f_\theta}-\frac{1}{f_\gamma}\right)\right|}{\omega^2 \vee \|\theta-\gamma\|_2^2} > \tau\sqrt{T}\right\} \leqslant \frac{e^2}{(e-1)^2}\exp\left(-\frac{T\tau^2\omega^2 K_\infty^2 \overline{r}^2}{4K_2^2\|\Sigma\|_{\text{spec}}^2}\right)$$

provided that $\omega^2 \geqslant \omega_d^2(\tau)$ *with*

$$\omega_d^2(\tau) = \frac{dD}{c_2 T}\left\{1 \vee \frac{K_2^2\|\Sigma\|_{\text{spec}}^2}{K_\infty^2\overline{r}^2\tau^2}\left(\frac{20}{e}B + \frac{200dD}{e^2T}A^2\right)\right\} \tag{2.38}$$

with

$$A = 108\sqrt{2}K_\infty^2\overline{r}^2/K_2$$

and

$$B := 1 + 36\sqrt{2}K_\infty\overline{r}K_2^{-1}\sqrt{\pi}(K_2 + \sqrt{K_{tv}dD\overline{r}\tilde{v}K_\infty/T})\ .$$

This key result helps for controling the fluctuations of the empirical spectral process E_T. It is a generalisation of Theorem 5 of Birgé and Massart [11], who proved a similar result for the empirical process of an iid sequence.

Derivation of (2.30)

For ease of presentation, we set $s = 2\|\Sigma\|_{\text{spec}}^2$. Direct considerations yield:

$$\|\theta_0-\gamma\|_2 \leqslant \delta_0, \quad \|\theta_0-\gamma\|_\infty \leqslant \overline{r}\delta_0 \tag{2.39}$$

$$\|\theta_k-\theta_{k-1}\|_\infty \leqslant \overline{r}\,(\delta_k+\delta_{k-1}) \leqslant 3\overline{r}\delta_k \tag{2.40}$$

$$\|\theta_k-\theta_{k-1}\|_2 \leqslant \delta_k+\delta_{k-1} \leqslant 3\delta_k \tag{2.41}$$

Moreover, from (2.29), we can write $\mathrm{TV}(\theta_0-\gamma) \leqslant dD\overline{r}\tilde{v}\delta_0$. Considering property (A.4) on the spectral norm of matrices, we can write

$$\begin{aligned}\xi_0 &\leqslant s\left\{2K_\infty\overline{r}\eta_0\delta_0 + \sqrt{2\pi\eta_0}\left(K_2\delta_0 + \delta_0 d\sqrt{T^{-1}K_{tv}D\overline{r}\tilde{v}K_\infty}\right)\right\}\\ &= s\left\{\delta_0(x_0+\sqrt{y_0})\right\}\end{aligned}$$

with

$$x_0 := 2K_\infty \overline{r}\eta_0$$
$$y_0 := 2\pi\eta_0 \left(K_2 + d\sqrt{T^{-1}K_{tv}D\overline{r}\tilde{v}K_\infty}\right)^2$$

Similarly,

$$\xi_k \leqslant s\left\{\delta_k(x_k + \sqrt{y_k})\right\}$$

for $k \geqslant 1$ with

$$x_k := 6K_\infty \overline{r}\eta_k$$
$$y_k := 18\pi\eta_k \left(K_2 + d\sqrt{T^{-1}K_{tv}D\overline{r}\tilde{v}K_\infty}\right)^2$$

Then, using $(a+b)^2 \leqslant 2a^2 + 2b^2$,

$$\begin{aligned}\left(\sum_{k=0}^{\infty}\xi_k\right)^2 &\leqslant 2s^2\left(\delta_0 x_0 + \sum_{k\geqslant 1}\delta_k x_k\right)^2 \\ &\qquad + 4s^2\left(\delta_0\sqrt{y_0} + \sum_{k\geqslant 1}\delta_k\sqrt{y_k}\right)^2 \\ &=: 2s^2A + 2s^2B.\end{aligned}$$

We first evaluate A. By definition of δ_k, x_0 and x_k, we get

$$A = 4K_\infty^2\overline{r}^2\delta_0^2\left(\eta_0 + 3\sum_{k\geqslant 1}2^{-k}\eta_k\right)^2 \leqslant 36K_\infty^2\overline{r}^2\delta_0^2\left(\sum_{k\geqslant 0}2^{-k}\eta_k\right)^2.$$

We now evaluate B. With repeated use of the inequality $(\sqrt{a} + \sqrt{b})^2 \leqslant 2a + 2b$ we can write

$$\begin{aligned}B &= \left(\delta_0\sqrt{y_0} + \sum_{k\geqslant 1}\delta_k\sqrt{y_k}\right)^2 \leqslant 2\delta_0^2 y_0 + 2^2\delta_1^2 y_1 + \sum_{k\geqslant 2}2^{k+1}\delta_k^2 y_k \\ &= 2\delta_0\sum_{k\geqslant 0}\delta_k y_k \leqslant 36\pi\delta_0^2\left(K_2 + d\sqrt{T^{-1}K_{tv}D\overline{r}\tilde{v}K_\infty}\right)^2\sum_{k\geqslant 0}2^{-k}\eta_k.\end{aligned}$$

We now compute $\sum_{k\geqslant 0} 2^{-k}\eta_k$. First, we note that, by Lemma 2.2, $H_k \leqslant dD\ln(5\omega/\delta_k)$ and this last bound is bounded by$dD\ln(5\omega/\delta_0) + dDk\ln 2 \leqslant dDL + dDk\ln 2$. Then we get

$$\eta_0 \leqslant \frac{2dDL}{c_2T} \qquad \text{and } \eta_k \leqslant (k+3)\frac{dDL}{c_2T} + (2k-1)\frac{dD}{c_2T}\ln 2. \quad (2.42)$$

Using $\sum_{k\geqslant 1} 2^{-k} = 1$ and $\sum_{k\geqslant 1} k2^{-k} = 2$, a direct calculation leads to

$$\sum_{k\geqslant 0} 2^{-k}\eta_k \leqslant \frac{2dDL}{c_2T} + \frac{5dDL}{c_2T} + \frac{3dD}{c_2T}\ln 2$$

and, with (2.33) and $7 + 1.5\ln 2 \leqslant 9$,

$$\sum_{k\geqslant 0} 2^{-k}\eta_k \leqslant 9\frac{dDL}{c_2T}.$$

Finally,

$$\begin{aligned}\sum_{k=0}^{\infty}\xi_k &\leqslant \sqrt{2}s\left(\sqrt{A}+\sqrt{B}\right)\\ &\leqslant 18\sqrt{2}s\delta_0\sqrt{\frac{dDL}{c_2T}}\left\{3K_\infty\overline{r}\sqrt{\frac{dDL}{c_2T}}\right.\\ &\qquad\left. +d\sqrt{\pi}\left(K_2+\sqrt{\frac{K_{tv}D\overline{r}\tilde{v}K_\infty}{T}}\right)\right\}\end{aligned}$$

with (2.33), $\delta_0 = \omega/\alpha$ and (2.32), we get (2.30).

2.4.3 Proof of Theorem 2.1

Recall that the sieve m is fixed in this proof. We first need to establish a link between the error $\|\theta^\circ - \hat{\theta}_m\|_2$ and the empirical process $\tilde{E}_n$. The following lemma will be useful for this task.

Lemma 2.3 ([28]). *If the class $\mathcal{F}_m$ is such that θ_m exists and is unique, and if $\rho_\infty(1/f_\theta)$ and $\rho_2(1/f_\theta)$ are uniformly bounded under $\theta \in \mathcal{F}_m$, then there exists an $\alpha > 0$ such that*

$$\rho_2\left(\frac{1}{f_\theta},\frac{1}{f_{\theta_m}}\right)^2 \leqslant \alpha\left\{\mathcal{L}(\theta) - \mathcal{L}(\theta_m)\right\}$$

for all $\theta \in \mathcal{F}_m$.

Proof. We refer to Dahlhaus and Polonik [28]. □

We start the proof of the theorem with the decomposition $\|\theta^\circ - \hat{\theta}_m\|_2 \leqslant \|\theta^\circ - \theta_m\|_2 + \|\theta_m - \hat{\theta}_m\|_2$ and the goal now is to bound the second term of the right hand side in this inequality. Let $1 - p(\omega)$ be the probability of the event $\mathcal{A}$ defined by

$$\left\{\forall \nu \in \mathcal{F}_m : \frac{1}{\sqrt{T}} \tilde{E}_T \left(\frac{1}{f_\nu} - \frac{1}{f_{\theta_m}}\right) \leqslant k \left(\omega^2 \vee \|\nu - \theta_m\|_2^2\right)\right\}$$

where k is a positive constant that will be specified later on. On $\mathcal{A}$, using Assumption 2.4, we may write, with $C = K_2'^{-1}$,

$$\begin{aligned}
\|\theta_m - \hat{\theta}_m\|_2^2 &\leqslant C\rho_2^2 \left(1/f_{\theta_m}, 1/f_{\hat{\theta}_m}\right) \\
&\leqslant \alpha C \left[\mathcal{L}(\hat{\theta}_m) - \mathcal{L}(\theta_m)\right] \qquad \text{by Lemma 2.3} \\
&\leqslant \frac{\alpha C}{4\pi\sqrt{T}} E_T \left(\frac{1}{f_{\theta_m}} - \frac{1}{f_{\hat{\theta}_m}}\right) + \frac{\kappa_1}{4\pi T}
\end{aligned}$$

where the last term comes from (2.25). Straightforward calculations lead to

$$\begin{aligned}
|E_T(\phi) - \tilde{E}_T(\phi)| = |\overline{E}_T(\phi)| &\leqslant \frac{\overline{C}}{\sqrt{T}} \{\rho_\infty(\phi) + \tilde{v}(\phi)\} \\
&\leqslant \frac{\overline{C}}{\sqrt{T}} (\rho_\infty + \tilde{v})
\end{aligned}$$

Then we can write

$$\|\theta_m - \hat{\theta}_m\|_2^2 \leqslant \alpha C k (4\pi)^{-1} \left(\omega^2 \vee \|\theta_m - \hat{\theta}_m\|_2^2\right) + R_T$$

where

$$R_T = \frac{\alpha C \kappa_1}{4\pi T} + \frac{\alpha \overline{C} C}{4\pi T} (\rho_\infty + \tilde{v}) \ .$$

and thus

$$\begin{aligned}
&\|\theta_m - \hat{\theta}_m\|_2^2 \\
&\qquad \leqslant \alpha C k (4\pi)^{-1} \left(\omega^2 + \|\theta_m - \hat{\theta}_m\|_2^2\right) + \alpha C (\kappa_1 + \overline{C})/(4\pi T) \ .
\end{aligned}$$

We choose $k = 2\pi/(\alpha C)$ and rearrange the inequality to get

$$\|\theta_m - \hat{\theta}_m\|_2^2 \leqslant (2\pi)^{-1} \left(2\pi\omega^2 + \alpha C T^{-1}(\kappa_1 + \overline{C})\right)$$

a.s. on $\mathcal{A}$. Denote $V = \|\theta_m - \hat{\theta}_m\|_2^2 - \alpha C(2\pi T)^{-1}(\kappa_1 + \overline{C})$. Then, $V \leqslant \omega^2$ a.s. on $\mathcal{A}$ with $\Pr(\mathcal{A}) = 1 - p(\omega)$. Then, $\Pr(V > \omega^2) \leqslant 1 - \Pr(\mathcal{A}) = p(\omega)$ and we get

$$\begin{aligned}
\mathrm{E}(V) &= \int_0^\infty dx \ \Pr(V > x) \\
&\leqslant \int_0^\infty dx \ p(\sqrt{x}) \\
&= \int_0^\infty dy \ 2yp(y) \\
&= \int_0^{\omega_d\left(\frac{2\pi}{\alpha C}\right)} dy \ 2yp(y) + \int_{\omega_d\left(\frac{2\pi}{\alpha C}\right)}^\infty dy \ 2yp(y)
\end{aligned}$$

(ω_d is defined in (2.38))

$$\begin{aligned}
&\leqslant \omega_d \left(\frac{2\pi}{\alpha C}\right) + \frac{2e^2}{(e-1)^2} \int_{\omega_d\left(\frac{2\pi}{\alpha C}\right)}^\infty dy \ y \exp\left(-\frac{Tk^2 y^2 K_\infty^2 \overline{r}^2}{K_2^2 \|\Sigma\|_{\text{spec}}^2}\right) \\
&\leqslant \omega_d^2 \left(\frac{2\pi}{\alpha C}\right) + \frac{e^2}{(e-1)^2} \ \frac{K_2^2 \|\Sigma\|_{\text{spec}}^2}{Tk^2 K_\infty^2 \overline{r}^2}
\end{aligned}$$

With the Jensen's inequality and using $e/(e-1) \leqslant 2$, we get

$$\begin{aligned}
\mathrm{E}\|\theta^\circ - \hat{\theta}_m\|_2 \leqslant \|\theta^\circ - \theta_m\|_2 + \omega_d \left(\frac{2\pi}{\alpha C}\right) + \sqrt{\frac{\alpha C(\kappa_1 + \overline{C})}{2\pi T}} \\
+ \frac{2\alpha C K_2 \|\Sigma\|_{\text{spec}}}{K_\infty \overline{r} \sqrt{T}}
\end{aligned}$$

and the result follows. □

2.4.4 Proof of Theorem 2.2

Fix (ν, m') such that the two conditions

- $\nu \in \mathcal{F}_{m'}$, and

- $\mathcal{L}_T(\nu) + \text{pen}(m') \leqslant \mathcal{L}_T(\theta_m) + \text{pen}(m)$ for each $m \in \mathcal{M}_T$

are fulfilled. For all $m \in \mathcal{M}_T$, we can write, with $C = K_2'^{-1}$,

$$\begin{aligned}
\|\theta^\circ - \nu\|_2^2 &\leqslant C\rho_2^2\left(1/f_{\theta^\circ}, 1/f_\nu\right) \qquad \text{by Assumption 2.4} \\
&\leqslant \alpha C\left\{\mathcal{L}(\nu) - \mathcal{L}(\theta^\circ)\right\} \qquad \text{by Lemma 2.3} \\
&\leqslant \alpha C\left\{\mathcal{L}(\nu) - \mathcal{L}_T(\nu) + \mathcal{L}_T(\theta_m) - \mathcal{L}(\theta^\circ) + \text{pen}(m) - \text{pen}(m')\right\} \\
&= \alpha C\Bigg\{\frac{1}{4\pi\sqrt{T}}\tilde{E}_T\left(\frac{1}{f_{\theta_m}} - \frac{1}{f_\nu}\right) + R_T + U_m \\
&\qquad\qquad + \text{pen}(m) - \text{pen}(m')\Bigg\} \qquad (2.43)
\end{aligned}$$

for all $m \in \mathcal{M}_T$, where $U_m = \mathcal{L}(\theta_m) - \mathcal{L}(\theta^\circ)$ is positive and such that $\inf_{m\in\mathcal{M}_T} U_m = 0$ (by definition of θ°), and where

$$R_T = \frac{\kappa_1}{4\pi T} + \frac{\overline{C}}{4\pi T}\left(\rho_\infty + \tilde{v}\right)$$

as in the proof of Theorem 2.1.

Now, we fix $m \in \mathcal{M}_T$. For all $m' \in \mathcal{M}_T$, define

$$\omega_{m'}^2(y) = \omega_{d_m}^2(16\pi^2/\tau) \vee \omega_{d_{m'}}^2(16\pi^2/\tau) \vee \left(\frac{\zeta(L_m d_m \vee L_{m'} d_{m'})}{T}\right) + \frac{y}{T}$$

for $y \geqslant 1$, where $\omega_{d_m}(\cdot)$ is defined in (2.38), and let $p(y)$ the probability of the set

$$\mathcal{A}_y = \left\{\sup_{m'\in\mathcal{M}_T}\sup_{\nu\in\mathcal{F}_{m'}} \frac{\left|\tilde{E}_T\left(\frac{1}{f_{\theta_m}} - \frac{1}{f_\nu}\right)\right|}{\|\theta^\circ - \theta_m\|_2^2 \vee \|\theta^\circ - \nu\|_2^2 \vee \omega_{m'}^2(y)} > \frac{2\pi\sqrt{T}}{\alpha C}\right\}.$$

Bound for $\|\theta^\circ - \hat{\theta}_{\hat{m}}\|_2$ on $\mathcal{A}_y^c$

On $\mathcal{A}_y^c$, we can bound $\|\theta^\circ - \nu\|_2^2$ from (2.43) as follows:

$$\begin{aligned}
\|\theta^\circ - \nu\|_2^2 \leqslant \frac{1}{2}\left(\|\theta^\circ - \theta_m\|_2^2 + \|\theta^\circ - \nu\|_2^2 + \omega_{m'}^2(y)\right) \\
+ \alpha C\left(R_T + U_m + \text{pen}(m) - \text{pen}(m')\right).
\end{aligned}$$

If we rearrange the sum, then the minimum penalized likelihood estimator $\hat{\theta}_{\hat{m}} \in \mathcal{F}_{\hat{m}}$ satisfies

$$\|\theta^\circ - \hat{\theta}_{\hat{m}}\|_2^2 \leqslant \|\theta^\circ - \theta_m\|_2^2 + \omega_{\hat{m}}^2(y) + 2\alpha C \left(R_T + U_m + \mathrm{pen}(m) - \mathrm{pen}(\hat{m})\right). \quad (2.44)$$

As the penalty function is such that

$$\mathrm{pen}(m) \geqslant (2\alpha C)^{-1} \left(\omega_{d_m}^2 \left(\frac{16\pi^2}{\tau}\right) \vee \frac{\zeta L_m d_m}{T}\right) + \frac{R_T}{2},$$

then $(2\alpha C)^{-1}\omega_{\hat{m}}^2(y) + R_T \leqslant \mathrm{pen}(\hat{m}) + \mathrm{pen}(m) + (2\alpha C)^{-1} y/T$ and we can write

$$\|\theta^\circ - \hat{\theta}_{\hat{m}}\|_2^2 \leqslant \|\theta^\circ - \theta_m\|_2^2 + 2\alpha C \left(U_m + 2\,\mathrm{pen}(m)\right) + \frac{y}{T} \quad (2.45)$$

almost surely on $\mathcal{A}_y^c$ and for all fixed $m \in \mathcal{M}_T$.

Bound for $\Pr(\mathcal{A}_y)$

We now compute an upper bound of $p(y) = \Pr(\mathcal{A}_y)$. First, we show that, for all fixed $m' \in \mathcal{M}_T$

$$\begin{aligned}
&\Pr\left\{\sup_{\nu \in \mathcal{F}_{m'}} \frac{\left|\tilde{E}_T\left(\frac{1}{f_{\theta_m}} - \frac{1}{f_\nu}\right)\right|}{\|\theta^\circ - \theta_m\|_2^2 \vee \|\theta^\circ - \nu\|_2^2 \vee \omega^2} > \tau' \sqrt{T}\right\} \\
&\leqslant \left(1 + \frac{e^2}{(e-1)^2}\right) \times \\
&\qquad \times \exp\left(-\frac{T\tau'^2\omega^2}{\left(\frac{K_2^2}{K_\infty^2 \overline{r}^2} \vee 2\pi A_{m,m'}\right) \|\Sigma^{1/2}\|_{\mathrm{spec}}^4 + B\|\Sigma^{1/2}\|_{\mathrm{spec}}^2 \tau}\right)
\end{aligned} \quad (2.46)$$

holds for all fixed $\omega > \omega_{d_m}(\tau') \vee \omega_{d_{m'}}(\tau')$, where $A_{m,m'}$ and B will be derived now. In order to check this inequality, we note that, from Proposition 2.2, for all $\gamma \in \mathcal{F}_{m'}$,

$$\Pr\left\{\sup_{\nu \in \mathcal{F}_{m'}} \frac{\left|\tilde{E}_T\left(\frac{1}{f_\nu} - \frac{1}{f_\gamma}\right)\right|}{\omega^2 \vee \|\nu - \gamma\|_2^2} > \tau\sqrt{T}\right\} \leqslant \frac{e^2}{(e-1)^2} \exp\left(-\frac{T\tau^2\omega^2 K_2^2 \overline{r}^2}{4K_2^2 \|\Sigma\|_{\mathrm{spec}}^2}\right)$$

provided that $\omega^2 \geqslant \omega^2_{d_{m'}}(\tau)$. Moreover, the Bernstein-type inequality (2.28) allows to write a bound, for all $\gamma \in \mathcal{F}_{m'}$,

$$\Pr\left\{\frac{\left|\tilde{E}_n\left(\frac{1}{f_{\theta_m}}-\frac{1}{f_\gamma}\right)\right|}{\|\theta_m-\gamma\|_2^2\vee\omega^2} > \tau\sqrt{T}\right\} \leqslant \exp\left(-\frac{1}{4}\cdot\frac{T\tau^2(\|\theta_m-\gamma\|_2^2\vee\omega^2)}{2A^\circ_{m,m'}\pi\|\Sigma^{1/2}\|^4_{\mathrm{spec}}+B\|\Sigma^{1/2}\|^2_{\mathrm{spec}}\tau}\right)$$

where, using Assumption 2.4, (2.16) and (2.17)

$$\begin{aligned} A^\circ_{m,m'} &:= \frac{\rho^2_{2,T}\left(\frac{1}{f_{\theta_m}}-\frac{1}{f_\gamma}\right)}{\|\theta_m-\gamma\|_2^2\vee\omega^2} \\ &\leqslant \frac{K_2\|\theta_m-\gamma\|_2^2+T^{-1}K_{tv}K_\infty\|\theta_m-\gamma\|_\infty \mathrm{TV}\,(\theta_m-\gamma)}{\|\theta_m-\gamma\|_2^2\vee\omega^2} \end{aligned}$$

with Assumption 2.2 and equations (2.49), (2.50) derived below,

$$\begin{aligned} &\leqslant \frac{K_2\|\theta_m-\gamma\|_2^2+\frac{K_{tv}K_\infty\|\theta_m-\gamma\|_\infty\|\theta_m-\gamma\|_2\tilde{v}D(d_m^2\vee d_{m'}^2)}{T}}{\|\theta_m-\gamma\|_2^2\vee\omega^2} && (2.47)\\ &\leqslant K_2+T^{-1}K_{tv}K_\infty(\Phi_m^2 d_m+\Phi_{m'}^2 d_{m'})^{1/2}\tilde{v}D^{3/2}(d_m^2\vee d_{m'}^2) && (2.48)\\ &=: A_{m,m'} \end{aligned}$$

and with $B = K_\infty\rho_\infty(1/f_\gamma - 1/f_{\theta_m}) \leqslant K_\infty\kappa_s^2$ by Assumption 2.1. Then, we can write

$$\Pr\left\{\frac{\left|\tilde{E}_T\left(\frac{1}{f_{\theta_m}}-\frac{1}{f_\gamma}\right)\right|}{\|\theta_m-\gamma\|_2^2\vee\omega^2} > \tau\sqrt{T}\right\} \leqslant \exp\left(-\frac{1}{4}\cdot\frac{T\tau^2\omega^2}{2\pi A_{m,m'}\|\Sigma^{1/2}\|^4_{\mathrm{spec}}+B\|\Sigma^{1/2}\|^2_{\mathrm{spec}}\tau}\right).$$

With the equality $\|\Sigma\|_{\text{spec}} = \|\Sigma^{1/2}\|^2_{\text{spec}}$, we finally get, for all $\gamma \in \mathcal{F}_{m'}$,

$$\Pr\left\{\sup_{\nu\in\mathcal{F}_{m'}} \frac{\left|\tilde{E}_T\left(\frac{1}{f_{\theta_m}}-\frac{1}{f_\nu}\right)\right|}{\|\gamma-\theta_m\|_2^2 \vee \|\gamma-\nu\|_2^2 \vee \omega^2} > \tau\sqrt{T}\right\}$$
$$\leqslant \Pr\left\{\sup_{\nu\in\mathcal{F}_{m'}} \frac{\left|\tilde{E}_n\left(\frac{1}{f_\nu}-\frac{1}{f_\gamma}\right)\right| + \left|\tilde{E}_n\left(\frac{1}{f_{\theta_m}}-\frac{1}{f_\gamma}\right)\right|}{\|\gamma-\theta_m\|_2^2 \vee \|\gamma-\nu\|_2^2 \vee \omega^2} > \tau\sqrt{T}\right\}$$

with $a^2 \vee b^2 \vee c^2 \geqslant a^2 \vee b^2$,

$$\leqslant \left(1+\frac{e^2}{(e-1)^2}\right) \times \exp\left(-\frac{1}{4}\cdot\frac{T\tau^2\omega^2}{\left(\frac{K_2^2}{K_\infty^2\tau^2}\vee 2\pi A_{m,m'}\right)\|\Sigma^{1/2}\|^4_{\text{spec}} + B\|\Sigma^{1/2}\|^2_{\text{spec}}\tau}\right)$$

(2.46) follows since, with $\omega > 0$ and for any $\varepsilon > 0$, there exists $\gamma \in \mathcal{F}_{m'}$ such that

$$\|\gamma-\theta^\circ\|^2 \leqslant \left[(1+\varepsilon)\inf_{\nu\in\mathcal{F}_{m'}}\|\theta^\circ-\nu\|^2\right] \vee \omega^2$$

and this implies

$$\begin{aligned}\|\gamma-\theta_m\|_2^2 \vee \|\gamma-\nu\|_2^2 &\leqslant \|\theta^\circ-\gamma\|_2^2 + \left(\|\theta^\circ-\nu\|_2^2 \vee \|\theta^\circ-\theta_m\|_2^2\right)\\ &\leqslant \left\{(1+\varepsilon)\|\theta^\circ-\nu\|_2^2 \vee \omega^2\right\} + \left\{\|\theta^\circ-\nu\|_2^2 \vee \|\theta^\circ-\theta_m\|_2^2\right\}\\ &\leqslant (2+\varepsilon)\left\{\omega^2 \vee \|\theta^\circ-\nu\|_2^2 \vee \|\theta^\circ-\theta_m\|_2^2\right\}\end{aligned}$$

where we applied the inequality $a+b \leqslant 2(a\vee b)$, and this argument holds for an arbitrary $\varepsilon > 0$.

(2.46) allows to bound $p(y) = \Pr(\mathcal{A}_y)$ as follows: Using (2.19)–(2.23), $D_m \geqslant 1$, $A_{m,m'} \geqslant A$ (with (2.50) derived below) and $(1+(e/(e-1))^2) \leqslant 3.6$, if $y \geqslant 1$, then we may write

$$\begin{aligned}p(y) &\leqslant 3.6\sum_{m'\in\mathcal{M}_T}\exp\left(-\frac{(\zeta L_{m'}d_{m'}+y)}{\zeta}\right)\\ &\leqslant 3.6\exp\left(-\frac{y}{\zeta}\right)\sum_{m'\in\mathcal{M}_T}\exp\left(-L_{m'}d_{m'}\right)\\ &\leqslant 3.6\Upsilon\exp\left(-\frac{y}{\zeta}\right).\end{aligned}$$

Derivation of (2.47)

We need to quantify the total variation distance between two curves in two different models $\mathcal{F}_m$ and $\mathcal{F}_{m'}$. Assume $\gamma \in \mathcal{F}_m$ and $\eta \in \mathcal{F}_{m'}$. By assumption on our models, $\mathcal{F}_m \subseteq \mathcal{F}_{m'}$ or $\mathcal{F}_{m'} \subseteq \mathcal{F}_m$. Then, if $\gamma^{(i)} = \sum_{j\in\Lambda_m} \gamma_{ij}\varphi_j$ and $\eta^{(i)} = \sum_{j\in\Lambda_{m'}} \eta_{ij}\varphi_j$, we get

$$\begin{aligned}
&\mathrm{TV}(\gamma - \eta) \\
&\quad \leqslant \sum_{i=1}^{D} \sup_{\substack{0\leqslant u_0<\ldots<u_N\leqslant 1 \\ N\in\mathbb{N}_0}} \sum_j |\gamma_{ij} - \eta_{ij}| \sum_{k=0}^{N-1} \left|\varphi_{ij}(u_{k+1}) - \varphi_{ij}(u_k)\right| \\
&\quad \leqslant \|\gamma - \eta\|_2 \tilde{v} D(d_m^2 \vee d_{m'}^2) \qquad (2.49)
\end{aligned}$$

Derivation of (2.48)

We have to compute the supremum norm between two curves in two different models. Assume $\gamma \in \mathcal{F}_{m_1}$ and $\eta \in \mathcal{F}_{m_2}$. Assume: $\mathcal{F}_m = \mathcal{F}_{m_1} + \mathcal{F}_{m_2}$ has a dimension $d_m \leqslant d_{m_1} + d_{m_2}$. Then

$$\begin{aligned}
\|\gamma - \eta\|_\infty &= \sup_i \sup_u \left| \sum_j (\gamma_{ij} - \eta_{ij})\varphi_j(u) \right| \\
&\leqslant \sqrt{D}\|\gamma - \eta\|_2 \sup_u \sqrt{\sum_{j=1}^{d_m} \varphi_j^2(u)} \\
&\leqslant \sqrt{D}\|\gamma - \eta\|_2 \Phi_m \sqrt{d_m}
\end{aligned}$$

where Φ_m is defined in (2.11). Barron et al. [5, equations (3.2)–(3.3)] prove that

$$\Phi_m^2 d_m \leqslant \Phi_{m_1}^2 d_{m_1} + \Phi_{m_2}^2 d_{m_2} \qquad (2.50)$$

and that Φ_m is such that

$$\Phi_m d_m \geqslant \Phi \qquad (2.51)$$

where Φ is defined in (2.20).

Coda

We now define the random variable

$$V = \left[\|\theta - \hat{\theta}_{\hat{m}}\|_2^2 - \|\theta - \theta_m\|_2^2 - 2\alpha C \left(U_m - 2\,\mathrm{pen}(m)\right) \right] \vee 0.$$

From (2.45), we observe that $V \leqslant y/T$ a.s. on $\mathcal{A}_y^c$ with $\Pr(\mathcal{A}_y) = p(y)$. Thus, if $y > 1$, $\Pr(V > y/T) \leqslant p(y)$ and we may write, for any $m \in \mathcal{M}$,

$$\begin{aligned} \mathrm{E}(V) &= \int_0^\infty dy \ \Pr(V > y) \\ &= T^{-1} \left(1 + \int_1^\infty dy \, p(y)\right) \\ &\leqslant T^{-1} \left(1 + 3.6 \Upsilon \zeta\right) \end{aligned}$$

and then we conclude. □

2.5 Conclusions and future research

In this chapter, we have addressed the problem of fitting an evolutionary semiparametric model to nonstationary time series using a data-driven procedure. This procedure has two steps. First, a contrast function is minimised over several linear and finite-dimensional models (the sieve). Then, we select the model for the final estimation by minimising the penalised contrast function. In our results, the contrast function is an approximation of the log-likelihood of the nonstationary model, and we have derived the form of the penalty function which is necessary to perform the model selection procedure.

Very often one is interested in time-varying models that are purely defined in the time domain, such as time-varying ARMA models. In this case the proposed estimation procedure via the spectrum may just be regarded as a technical tool for estimation.

To end this chapter, we would like to mention some possible directions for future research.

First of all, the quality of our estimation is measured with the global norm L^2 defined in (2.12). This means that each component of the vector θ is estimated with the same regularity. It could be interesting to derive the results of this chapter *componentwise*. The main difficulty to establish such results is to provide a link between the empirical spectral process and the behaviour of the estimator component by component. A possible approach is to differentiate the contrast function, with some regularity assumptions on the contrast $\mathcal{L}(\theta)$ as a function of θ. This approach has been followed in the work of Dahlhaus and Giraitis [25] for another task, and it should be also fruitful here. Nevertheless, in that case, the minimisation of the contrast function would lead to serious

computational problems. Moreover, we note that the approach with the product space $\mathcal{F}^D$ considered in this chapter has a connection with many previous works on tv-AR processes. For instance, this viewpoint is also taken in Subba Rao [107]. This approach can also be found in numerous work in the applied science [37, 42, 53, 109, 115].

A limitation of the preceeding method is the fact that the penalty function depends on the spectral norm $s := \|\Sigma\|_{\text{spec}}$ which is obviously unknown from a practical viewpoint. The question here is to find a preliminary estimator $\tilde{s}$ of s such that $\|\tilde{s} - s\| = o_P(1)$. With this property, we think that the estimation error due to the preliminary estimation is negligible with respect to the estimation error due to the model selection procedure itself (i.e. the error derived in our two theorems). In other words, the main error term is due to the main estimation procedure, and the preliminary estimation procedure leads to higher order error terms. Of course, a complete proof has not yet been given and would be a useful result for the future. Moreover, the preliminary estimator needs to be defined properly. In Chapter 4, we derive an $o_P(1)$ preliminary estimator of this quantity in another context, and we think that this estimator could be of use here.

A natural question arises if it could be possible to choose the parameter D by a data-driven procedure. (Recall that D is the number of components of the curve θ.) In the context of tv-ARMA processes, a similar question is how to choose p and q in practice. This problem is not solved by our procedure, and could be a challenging question for future research. The selection of D does not follow immediately from our estimation procedure because our proofs are strongly using the true value of this parameter. However, note that some rules for the selection of D have already been proposed in the literature, among which is a modified AIC criterion [23]. A detailed theoretical and empirical study of such selection procedures is still to be done, and would be of a great interest for future research.

CHAPTER 3

A wavelet-based model for locally stationary processes

3.1 Introduction

The theory of wavelets offers an interesting alternative to the Fourier analysis. It permits the construction of orthogonal bases of $L^2(\mathbb{R})$ which, in contrast to Fourier basis, are local both in time and frequency. As a consequence, wavelets appear more appropriate to decompose signals with high frequencies, like signals with jumps or peaks.

The origin of this theory is not easy to draw. A common starting point is the article of Grossmann and Morlet for the decomposition of seismic signals [45]. The term "wavelet" was used for the first time at this occasion. Then, this theory has been developed and has now numerous applications in many different fields, including statistics.

In the next chapters, we focus on a class of doubly-indexed locally stationary processes defined by replacing the harmonic system $\{\exp(i\omega t)\}$ in Definition (2.1) by a wavelet basis. By this way, we move from a *time-frequency* representation to a *time-scale* representation of the nonstationary process. Because wavelet systems are well localized in time and frequency, they appear more natural to model the time-varying spectra of nonstationary processes. Indeed, by the uncertainty principle, allowing the spectrum to be time-varying implies that we loose resolution in the frequency domain. As wavelets decompose the frequency domain into discrete scales, they offer a well-adapted system to achieve the trade-off resolution between time and frequency [116].

The class of locally stationary wavelet processes studied in the next chapters was initially introduced by Nason, von Sachs, and Kroisandt [83]. Their definition of wavelet processes involves a time-varying amplitude which is smoothly varying and continuous as a function of time. One first goal of this chapter is to extend this definition to the case of time-varying amplitudes with possibly *discontinuous* behaviour in time. This adds some technical difficulties in the proof of our results but we believe the gain due to this extension to be crucial. Our new definition now includes more important examples of nonstationary processes. For instance, this extension of the definition is needed if we wish to model a nonstationary process built as a concatenation of different stationary processes. Moreover, wavelet processes can now be used for the analysis of intermittent phenomena, such as transients followed by regions of smooth behaviour.

Our definition of wavelet processes is presented in Section 3.4, where we also define their *evolutionary spectrum*. This spectrum is a function of time and scales, and measures the power of the process at a particular time and scale.

Before introducing our definition of this process, we first recall some basic results on wavelets (Section 3.2). We only recall the minimal notions needed for the later developments. In the next Section 3.3, we have a particular focus on the wavelet system that will be used in the definition of the random process.

3.2 Standard wavelet systems

In this section, we summarize some standard results on wavelets. These will be useful later in our work. An exhaustive introduction to this theory may be found in [32], [49] or [68].

3.2.1 Multiresolution analysis of L^2

The starting point for the construction of an orthonormal wavelet basis is the notion of *multiresolution analysis* (MRA) *of* $L^2(\mathbb{R})$, that is a sequence of closed vector subspaces V_j of $L^2(\mathbb{R})$

$$\ldots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset \ldots \tag{3.1}$$

such that

1. $\bigcap_{j\in\mathbb{Z}} V_j = \{0\}, \qquad \bigcup_{j\in\mathbb{Z}} V_j$ is dense in $L^2(\mathbb{R})$;

2. for all f in $L^2(\mathbb{R})$ and for all j in $\mathbb{Z}$, we have $f(x) \in V_j \Leftrightarrow f(2x) \in V_{j+1}$;

3. for all f in $L^2(\mathbb{R})$ and for all k in $\mathbb{Z}$, we have $f(x) \in V_0 \Leftrightarrow f(x-k) \in V_0$;

4. there exists a function ϕ in V_0 such that the system $\{\phi_{0k}(x) \equiv \phi(x-k),\ k \in \mathbb{Z}\}$ of shifted functions is an orthonormal basis of V_0.

The function ϕ is called *scaling function* and is such that

$$\phi_{jk}(x) \equiv 2^{j/2}\phi\left(2^j x - k\right), \qquad k \in \mathbb{Z} \tag{3.2}$$

constitutes an orthonormal basis of V_j. The functions ϕ_{jk} are the *dilated-translated functions* of the scaling function. (We sometimes write $\phi_{j,k}$ if an ambiguity may appear in the indices between the dilatation and the time factor.)

Denote by W_j the orthogonal complement of the subspace V_j in V_{j+1}. In the next section, we recall the Meyer-Mallat Theorem, which says that W_j is generated by some dilated-translated function ψ, that is

$$\psi_{jk}(x) = 2^{j/2}\psi\left(2^j x - k\right), \qquad k \in \mathbb{Z}, \tag{3.3}$$

constitutes a basis of W_j. By definition of a multiresolution analysis, the collection $\{\psi_{jk}(x), j, k \in \mathbb{Z}\}$ is an orthonormal basis of $L^2(\mathbb{R})$. This system is a *wavelet basis* generated by the function ψ called the *mother wavelet*. Therefore, the wavelets are constructed from a scaling function ϕ given in an MRA of $L^2(\mathbb{R})$.

In the following, we shall recall how the functions ϕ and ψ are constructed from an MRA. Moreover, we shall see how the MRA structure induces very useful properties on these two functions.

3.2.2 Construction and examples

The following explains how to derive a mother wavelet ψ from an MRA. Since ϕ is in V_0, which is included in V_1, we can decompose ϕ in the basis $\{\phi_{1,k}\}$ of V_1:

$$\phi = \sum_k h_k \phi_{1,k} \tag{3.4}$$

with $h_k = \langle \phi, \phi_{1,k} \rangle_{L^2}$, $k \in \mathbb{Z}$. We expand this decomposition and it comes

$$\phi(x) = \sqrt{2} \sum_k h_k \phi(2x - k) \qquad x \in \mathbb{R}$$

which we can rewrite in the Fourier domain

$$\hat{\phi}(\lambda) = \frac{1}{\sqrt{2}} \sum_k h_k e^{-ik\lambda/2} \hat{\phi}(\lambda/2) \qquad \lambda \in [-\pi, \pi]$$

or, equivalently,

$$\hat{\phi}(\lambda) = m_0(\lambda/2)\, \hat{\phi}(\lambda/2) \qquad \lambda \in [-\pi, \pi] \tag{3.5}$$

where

$$m_0(\lambda) = \frac{1}{\sqrt{2}} \sum_k h_k e^{-ik\lambda}. \qquad \lambda \in [-\pi, \pi] \tag{3.6}$$

The following theorem gives the mother wavelet associated to ϕ. A proof of this result may be found in Mallat [68].

Theorem 3.1 (Meyer [75], Mallat [66]). *If $(V_j)_{j \in \mathbb{Z}}$ is a multiresolution analysis of $L^2(\mathbb{R})$ with scaling function ϕ such that (3.4) and (3.5) hold, then the function ψ defined by*

$$\hat{\psi}(\lambda) = \frac{1}{\sqrt{2}} e^{i\lambda/2} \overline{m_0\left(\frac{\lambda}{2} + \pi\right)} \hat{\phi}(\lambda/2) \qquad \lambda \in [-\pi, \pi]$$

or, equivalently, with $g_k = (-1)^{k-1} \overline{h_{-k-1}}$,

$$\psi(x) = \sum_k g_k \phi_{1,k}(x) \tag{3.7}$$

is a mother wavelet, i.e. the system (3.3) generates W_j.

In addition, if we start with an *r-regular multiresolution analysis* of $L^2(\mathbb{R})$, that is if we start with a multiresolution analysis whose scaling function fulfils

$$\left| \left(\frac{d}{dx} \right)^q \phi(x) \right| \leqslant C_m (1 + |x|)^{-m}$$

for all $q \leqslant r$ and all integer $m \geqslant 1$, then one can show that the resulting function ψ is such that:

1. $\{\psi_{jk}\}_{j,k\in\mathbb{Z}}$ is an orthonormal basis of $L^2(\mathbb{R})$;
2. $\psi(x)$ and all its derivatives until order r are in $L^\infty(\mathbb{R})$;
3. $\left|\left(\frac{d}{dx}\right)^q \psi(x)\right| \leqslant C_m (1+|x|)^{-m}$ with $m \geqslant 1$ and $0 \leqslant q \leqslant r$;
4. $\int_{\mathbb{R}} x^q \psi(x) dx = 0$ for $0 \leqslant q \leqslant r$.

The conditions (2), (3) and (4) show respectively the *regularity*, the *localization* and the *oscillatory behaviour* of the function ψ, sometimes called a *mother wavelet of class* r. The functions ψ_{jk} inherit these three properties from the mother wavelet and are called *wavelets*. The translation parameter k represents the "time" and the dilatation parameter j is the "scale".

We now give some examples of such bases. For all examples, it suffices to define a scaling function (or, by (3.4), a sequence $\{h_k\}$). The regularity of the corresponding wavelet system will depend on the regularity conditions imposed on this scaling function.

Example 3.1 (Haar wavelet system). The simplest example of a multiresolution analysis starts with the *Haar scaling function* $\phi(x) = 1_{[0,1[}(x)$. It is easily verified that $\{\phi(x-k); k \in \mathbb{Z}\}$ forms an orthonormal set of functions. Moreover, the Haar scaling function leads to $h_k = \frac{1}{\sqrt{2}}\delta_0(k) + \frac{1}{\sqrt{2}}\delta_0(k-1)$ and

$$\begin{aligned}\psi(x) &= -\phi(2x) + \phi(2x-1)\\ &= -1_{[0,1/2[}(x) + 1_{[1/2,1[}(x).\end{aligned}$$

ψ is the *Haar mother wavelet* and corresponds to a multiresolution analysis of regularity $r = 0$. Consequently, the regularity of the Haar wavelet basis is limited. ◊

Example 3.2 (Daubechies wavelet system). After the example of the Haar wavelet system, the natural question arises if it is possible to define a more regular function which still has a compact support in time. The answer is positive and was first given by Daubechies [30], who showed that, for all integers $r \geqslant 1$, there exists an r-regular multiresolution analysis of $L^2(\mathbb{R})$ such that the support of ϕ and ψ are compact. She also obtained the following precision: There exists a constant C

such that, for all $r \geqslant 1$, the support of ϕ and ψ are included in the interval $[-Cr, Cr]$, and there exists a constant c such that the length of the supports of ϕ and ψ is larger than cr. Then, the supports increases linearly with the regularity of the multiresolution analysis. Details about Daubechies wavelets may be found in [30, 31, 116]. $\Diamond$

Example 3.3 (Shannon wavelet system). Also called Littlewood-Paley wavelets, the Shannon wavelet system has time-frequency properties which are complementary to those of the Haar basis as it is compactly supported in the Fourier domain. We start with a scaling function ϕ having a Fourier transform $\hat{\phi}(\lambda) = 1_{[-\pi,\pi]}(\lambda)$ for $\lambda \in [-\pi, \pi]$. This scaling function leads to $m_0(\lambda) = 1_{[-\pi/2,\pi/2]}(\lambda)$ and

$$\hat{\psi}(\lambda) = e^{-i\lambda/2}\left\{\hat{\phi}(\lambda/2) - \hat{\phi}(\lambda)\right\}$$

for $\lambda \in [-\pi, \pi]$. The inverse Fourier transform, ψ, is the Shannon mother wavelet:

$$\psi(x) = \mathrm{sinc}\{2\pi(t - 1/2)\} - \mathrm{sinc}\{\pi(t - 1/2)\},$$

where $\mathrm{sinc}(\alpha) := (\sin\alpha)/\alpha$. ψ is a very smooth function (in C^∞), but has obviously weak localisation property in time. $\Diamond$

3.2.3 The decimated wavelet transform for discrete data

Suppose we are given a deterministic sequence $s_0, \ldots, s_{T-1}$ where $T = 2^J$ for some integer J. In this section, we recall the algorithm of Mallat [67] which provides an efficient scheme for performing a wavelet transformation of such sequence.

First, we set

$$c_{0,k} = s_k, \qquad k = 0, \ldots, T-1$$

and define the continuous function

$$s(x) = \sum_{k=-\infty}^{\infty} c_{0,k}\phi_{0,k}(x) \qquad x \in \mathbb{R}.$$

The function s is in the space V_0 of the multiresolution analysis (3.1). Recall that W_j is the orthogonal complement of V_j in V_{j+1}. Then, we

may write the decomposition

$$s(x) = \sum_{k=-\infty}^{\infty} c_{-1,k}\phi_{-1,k}(x) + \sum_{k=-\infty}^{\infty} d_{-1,k}\psi_{-1,k}(x)$$

where

$$c_{-1,k} = \langle s, \phi_{-1,k}\rangle_{L^2}, \qquad d_{-1,k} = \langle s, \psi_{-1,k}\rangle_{L^2}.$$

Using (3.2) and (3.7), we get the recursions

$$\phi_{j-1,k}(x) = \sum_m h_{m-2k}\phi_{jm}(x), \quad \psi_{j-1,k} = \sum_m g_{m-2k}\phi_{jm}(x) \tag{3.8}$$

and therefore, after direct calculations, the coefficients above are such that

$$c_{-1,k} = \sum_m h_{m-2k}c_{0,m}, \qquad d_{-1,k} = \sum_m g_{m-2k}c_{0,m}. \tag{3.9}$$

These relations show how the coefficients at scale $j = -1$ may be computed from the coefficients at scale $j = 0$ (the scale on which the given sequence s "lives"). They involve the sequences $h = (h_k)_k$ and $g = (g_k)_k$ in a linear way. Note that for compactly supported wavelets, such as Haar or Daubechies wavelets, these two sequences have a finite number of non-zero coefficients, and then the summations in formula (3.9) are finite. Observe also that there are half as many coefficients at scale -1 than scale 0. This crucial phenomenon is called the *decimation*. Formulas (3.9) may be written in terms of operators [81]. Let H and G represent the convolution operator with the sequences $\{h_k\}$ and $\{g_k\}$ repectively, i.e.

$$(Hs)_k = \sum_m h_{m-k}s_m \qquad (Gs)_k = \sum_m g_{m-k}s_m$$

and denote by D_0 the decimation operator, i.e.

$$(D_0 s)_k = s_{2k}$$

for all $k \in \mathbb{Z}$. Then the relations (3.9) may be written

$$c^{(-1)} = D_0 H c^{(0)}, \qquad d^{(-1)} = D_0 G c^{(0)}. \tag{3.10}$$

where $c^{(j)}$ and $d^{(j)}$ denote the sequences $(c_{j,k})_k$ and $(d_{j,k})_k$ respectively.

With these conventions, a recursion of (3.10) leads to the following algorithm, called the *discrete wavelet transform* or the *pyramid algorithm* [67]: Given a sequence $s = (s_0, \ldots, s_{T-1})$ as above, set $c_0 = s$. The wavelet coefficients of the sequence are obtained recursively using the relations

$$c^{(j-1)} = D_0 H c^{(j)}, \qquad d^{(j-1)} = D_0 G c^{(j)}$$

for $j = -1, -2, \ldots$ until a level $J = -\log_2 T$. The wavelet coefficients are given by

$$(c^{(-J)}, d^{(-J+1)}, d^{(-J+2)}, \ldots, d^{(-2)}, d^{(-1)}).$$

Provided the sequence $\{h_k\}$ has a finite number of non-zero elements, the step at level j in the transform requires $O(2^{-j})$ operations. Then, the total number of arithmetic operations in the algorithm is $O(2^{-J})$, i.e. $O(T)$. Due to the orthogonality of the wavelet system, this transform is invertible, and it may be showed that the inverse transform is again an $O(T)$ algorithm. We refer to Mallat [67, 68] for further considerations about this algorithm.

3.2.4 Discrete wavelet system

In time series analysis, data are recorded at a finite number of discrete points. As we would like to use wavelets in order to decompose such processes, we need to define a discrete-time analog to the wavelet system (3.3).

At the jth step of the pyramid algorithm, the vector of wavelet coefficients is

$$d^{(j)} = \mathcal{W}_j c^{(0)}$$

where we know from the previous subsection that, at scale $j = -1$, the matrix $\mathcal{W}_{-1}$ is $D_0 G$ and, for scales $j < -1$, the matrix $\mathcal{W}_j$ is $D_0 G (D_0 H)^{-j-1}$. For fixed j, we note that, due to the MRA structure, the rows of matrix $\mathcal{W}_j$ are the same, except that they are simply shifted by 2^{-j}. The *discrete-time wavelet* ψ_j is defined as the first row of the matrix $\mathcal{W}_j$, and this matrix is orthonormal [98].

If we start the pyramid algorithm with finite sequences $\{h_k\}$ and $\{g_k\}$ of length $N_{[h]}$ corresponding to Daubechies wavelets, then we get

the associated discrete wavelet $\psi_j = (\psi_j(0), \ldots, \psi_j(N_j - 1))$ with a compact support of length N_j for scale $j < 0$ such that

$$\begin{aligned}
\psi_{-1}(n) &= \sum_k g_{n-2k}\delta_{0k} = g_n && \text{for } n = 0, \ldots, N_{-1} - 1, \\
\psi_{j-1}(n) &= \sum_k h_{n-2k}\psi_j(k) && \text{for } n = 0, \ldots, N_{j-1} - 1, \\
N_j &= (2^{-j} - 1)(N_{[h]} - 1) + 1,
\end{aligned}$$

where $\delta_{0k} = \delta_0(k)$ is the Kronecker delta [83].

For example, for Haar wavelets we get

$$\begin{aligned}
\psi_{-1} &= \frac{1}{\sqrt{2}}\,(1, -1) \\
\psi_{-2} &= \frac{1}{2}\,(1, 1, -1, -1) \\
\psi_{-3} &= \frac{1}{2\sqrt{2}}\,(1, 1, 1, 1, -1, -1, -1, -1)
\end{aligned}$$

and so on.

Except for Haar wavelets, the discrete wavelets are not just the sampled version of the associated continuous time wavelet $\psi(x)$. They are, however, precisely the vectors $d^{(j)}$ constructed in the discrete wavelet transform described above [83].

3.3 Nondecimated wavelets

The wavelet system used to decompose locally stationary processes is a discrete non-decimated system of compactly supported wavelets. In this section, we give a definition for these wavelets and collect some mathematical results about this system. These results will be useful later in our work.

3.3.1 Nondecimated discrete wavelet system

The key point for *nondecimated* (or *stationary)* discrete wavelets is that the wavelets can be shifted to any location and not just by shifts by 2^{-j} as in the discrete wavelet transform [19, 81]. If $\{\psi_j, j < 0\}$ is a discrete wavelet system (Section 3.2.4), we set $\psi_j = \psi_{j,0}$ and we redefine ψ_{jk} for $k \in \mathbb{Z}$ as:

$$\psi_{jk}(\tau) = \psi_{j,0}(\tau - k)\,, \qquad \tau \in \mathbb{Z}\,,$$

that is the $(\tau - k)$th element of the vector ψ_j. (Note the difference with the standard wavelet system, equation (3.3).)

In the following, we consider discrete mother wavelets with a compact support, that is the nondecimated discrete wavelet system is constructed from a finite sequence $\{h_k\}$ of length $N_{[h]}$, corresponding to a Daubechies wavelet. Then, the length of the support of the discrete wavelets ψ_j is given by $N_j = (2^{-j} - 1)(N_{[h]} - 1) + 1$ for all $j < 0$ (see Section 3.2.4).

3.3.2 The autocorrelation wavelet function

The measure of the local autocovariance structure of the wavelet-based process defined in the next section requires the study of the convolution

$$\Psi_j(\tau) = \sum_{k=-\infty}^{\infty} \psi_{jk}(0)\psi_{jk}(\tau)$$

where $\tau \in \mathbb{Z}$ and $j = -1, -2, -3, \ldots$. The function Ψ_j is called the *discrete autocorrelation wavelet function at scale* j (ACW in short). They inherit localisation properties from wavelets. However, they are symmetric about $\tau = 0$, that is $\Psi_j(\tau) = \Psi_j(-\tau)$ for all scales j and for all τ. In this section, we derive some useful results of these functions.

First of all, one easily derives that Ψ_j is compactly supported for all $j < 0$ and the length of its support is bounded by $2N_j - 1$. Furthermore, the following reasoning shows that $\Psi_j(0) = 1$:

$$\Psi_j(0) = \sum_{k=-\infty}^{\infty} \psi_{jk}(0)^2 = \sum_{k=-\infty}^{\infty} \psi_{jk}^2 = 1 \tag{3.11}$$

where the last equality comes from the orthonormality of *decimated* wavelets.

The next result is not straightforward and is due to Nason et al. [83].

Lemma 3.1. *The autocorrelation wavelet system* $\{\Psi_j; j = -1, -2, \ldots\}$ *is linearly independent.*

Proof. We refer to the proof of Theorem 1 of Nason et al. [83]. □

The autocorrelation wavelet system is related to the associated continuous-time autocorrelation function of wavelets studied by Saito and

Beylkin [101]. This *continuous* autocorrelation wavelet is defined as

$$\Psi^{\circ}(x) = \int_{-\infty}^{\infty} du\ \psi^{\circ}(u)\psi^{\circ}(u-x)$$

for the *continuous-time* compactly supported wavelet ψ° associated to our *discrete-time* wavelet ψ (i.e. defined from the same sequence $\{h_k\}$). A link between these two functions is given in the following lemma.

Lemma 3.2. *If $\Psi^{\circ}(u)$ denotes the continuous-time autocorrelation wavelet associated to the discrete-time autocorrelation wavelet $\Psi(\tau)$, then the formula*

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

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

Proof. See Lemma 4.2 of Berkner and Wells [8]. □

The connection between the continuous-time and the discrete-time ACW allows to prove the next resolution of the identity.

Lemma 3.3. *The discrete-time autocorrelation wavelet system*

$$\{\Psi_j(\tau); j < 0\}$$

is such that the identity

$$\sum_{j=-\infty}^{-1} 2^j \Psi_j(\tau) = \delta_0(\tau) \tag{3.12}$$

holds for all $\tau \in \mathbb{Z}$.

Proof. Using Lemma 3.2 and Parseval's identity,

$$\begin{aligned}\sum_{j=-\infty}^{-1} 2^j \Psi_j(\tau) &= \sum_{j=-\infty}^{-1} 2^j \Psi\left(2^j\,|\tau|\right) \\ &= \sum_{j=-\infty}^{-1} \int_{-\infty}^{\infty} d\omega\ |\hat{\psi}(2^{-j}\omega)|^2 \exp(i\omega\tau) \\ &= \sum_{j=-\infty}^{-1} \int_{0}^{2\pi} d\omega\ \sum_{k\in\mathbb{Z}} \left|\hat{\psi}\left(2^{-j}(\omega+2k\pi)\right)\right|^2 \exp(i\omega\tau).\end{aligned} \tag{3.13}$$

By the Mallat-Meyer Theorem, we may write

$$\sum_{k\in\mathbb{Z}}\left|\widehat{\psi}\left(2^{-j}\left(\omega+2k\pi\right)\right)\right|^2=\sum_{k\in\mathbb{Z}}\left|m_0\left(2^{-j-1}\omega+2^{-j-1}k2\pi+\pi\right)\right|^2\times \\ \times\left|\widehat{\phi}\left(2^{-j-1}\omega+2^{-j-1}k2\pi\right)\right|^2 \quad (3.14)$$

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

$$\begin{aligned}&\sum_{k\in\mathbb{Z}}\left|\widehat{\psi}\left(2^{-j}\left(\omega+2k\pi\right)\right)\right|^2 \quad (3.15)\\ &\quad=\left|m_0\left(2^{-j-1}\omega+\pi\right)\right|^2\sum_{k\in\mathbb{Z}}\left|\widehat{\phi}\left(2^{-j-1}\omega+2^{-j-1}k2\pi\right)\right|^2\\ &\quad=\left|m_0\left(2^{-j-1}\omega+\pi\right)\right|^2\sum_{k\in\mathbb{Z}}\left|m_0\left(2^{-j-2}\omega+2^{-j-2}k2\pi\right)\right|^2\\ &\qquad\qquad\left|\widehat{\phi}\left(2^{-j-2}\omega+2^{-j-2}k2\pi\right)\right|^2\\ &\quad=\left|m_0\left(2^{-j-1}\omega+\pi\right)\right|^2\left|m_0\left(2^{-j-2}\omega\right)\right|^2\\ &\qquad\qquad\sum_{k\in\mathbb{Z}}\left|\widehat{\phi}\left(2^{-j-2}\omega+2^{-j-2}k2\pi\right)\right|^2.\end{aligned}$$

By similar transformations, we finally arrive at

$$\begin{aligned}&=\left|m_0\left(2^{-j-1}\omega+\pi\right)\right|^2\prod_{n=2}^{-j}\left|m_0\left(2^{-j-n}\omega\right)\right|^2\sum_{k\in\mathbb{Z}}\left|\widehat{\phi}\left(\omega+k2\pi\right)\right|^2\\ &=(2\pi)^{-1}\left|m_0\left(2^{-j-1}\omega+\pi\right)\right|^2\prod_{n=2}^{-j}\left|m_0\left(2^{-j-n}\omega\right)\right|^2\\ &=(2\pi)^{-1}\left|1-m_0\left(2^{-j-1}\omega\right)\right|^2\prod_{\ell=0}^{-j-2}\left|m_0\left(2^{\ell}\omega\right)\right|^2.\end{aligned}$$

Using (3.13), 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

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

Thus, we obtain

$$\sum_{j=-\infty}^{-1} 2^j \Psi_j(\tau) = \frac{1}{2\pi}\int_0^{2\pi} d\omega\ \exp(i\tau\omega)\left\{1 - \prod_{\ell=0}^{+\infty}\left|m_0(2^{\ell}\omega)\right|^2\right\}$$
$$= \delta_0(\tau) - \frac{1}{2\pi}\int_0^{2\pi} d\omega\ \exp(i\tau\omega)\prod_{\ell=0}^{+\infty}\left|m_0(2^{\ell}\omega)\right|^2. \tag{3.16}$$

Now, it remains to prove that the second term in (3.16) is equal to zero. By definition, $m_0(\omega) = 2^{-1/2}\sum_{n=0}^{2N_{[h]}-1} h_n e^{-in\omega}$. We have

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

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

3.3.3 The Gram matrix A

As the autocorrelation wavelet system is not orthogonal, we introduce the Gram matrix A defined by

$$A_{j\ell} = \sum_{\tau} \Psi_j(\tau)\Psi_\ell(\tau) \qquad j,\ell = -1,-2,\ldots. \tag{3.17}$$

The invertibility of A has been established when $\{\Psi_j\}$ is constructed using Haar or Shannon wavelets [83]. If other compactly supported wavelets are used, numerical results suggest that the invertibility of A still holds, but a complete proof of this result has not been established yet. As we need the invertibility of A in our following results, from now

on we restrict ourselves to Haar wavelets, but we conjecture that all results remain valid for more general Daubechies wavelets.

We collect the following properties of A, which will be used thereafter.

Lemma 3.4. *For Haar and Shannon wavelets, there exists a finite positive constant ν such that the matrix A fulfills the following properties for all $j = -1, \dots, -[\log_2 T]$:*

$$\sum_{\ell=-[\log_2 T]}^{-1} A_{j\ell}^{-1} = 2^j + O\left(2^{j/2} T^{-1/2}\right) \tag{3.18}$$

$$\sum_{\ell=-[\log_2 T]}^{-1} |A_{j\ell}^{-1}| \leqslant \nu(1+\sqrt{2}) 2^{j/2} \tag{3.19}$$

$$\sum_{\ell=-[\log_2 T]}^{-1} 2^{-\ell/2} |A_{j\ell}^{-1}| \leqslant \nu \cdot 2^{j/2} \log_2 T \tag{3.20a}$$

$$\sum_{\ell=-[\log_2 T]}^{-1} 2^{-\ell} |A_{j\ell}^{-1}| \leqslant \nu(2+\sqrt{2}) 2^{j/2} T^{1/2}. \tag{3.20b}$$

For all compactly supported wavelets, the matrix A fulfills the following property:

$$A_{j\ell} \leqslant (2N_j - 1) \wedge (2N_\ell - 1) \wedge \sqrt{N_\ell N_j} \tag{3.21}$$

where $x \wedge y = \min(x, y)$.

Proof. The following argument shows that the main term in (3.18) is 2^j: Using that $\Psi_\ell(0) = 1$ for all $\ell < 0$ and the identity (3.12), we may write

$$\begin{aligned} \sum_{\ell=-\infty}^{-1} A_{j\ell}^{-1} &= \sum_{\ell=-\infty}^{-1} A_{j\ell}^{-1} \sum_{m,u=-\infty}^{\infty} 2^m \Psi_m(u) \Psi_\ell(u) \\ &= \sum_{m=-\infty}^{-1} 2^m \delta_0(j-m) = 2^j \end{aligned}$$

from the definition of A. Observe that this argument holds for all compactly supported wavelets. To compute the remainder of (3.18), we

introduce the auxiliary matrix $\Gamma = D' \cdot A \cdot D$ with diagonal matrix $D = \mathrm{diag}(2^{\ell/2})_{\ell<0}$, i.e. $\Gamma_{j\ell} = 2^{j/2} A_{j\ell} 2^{\ell/2}$. Nason et al. [83, Theorem 2] have proved that the spectral norm of Γ^{-1} is bounded for Haar and Shannon wavelets. Then, we get

$$\sum_{\ell=-\infty}^{-\log_2(T)-1} A^{-1}_{j\ell} = 2^{j/2} \sum_{\ell=-\infty}^{-\log_2(T)-1} 2^{\ell/2} \Gamma^{-1}_{j\ell} = O\left(2^{j/2} T^{-1/2}\right)$$

To prove (3.19),

$$\sum_{\ell=-\log_2 T}^{-1} |A^{-1}_{j\ell}| = \sum_{\ell=-\log_2 T}^{-1} 2^{j/2} 2^{\ell/2} |\Gamma^{-1}_{j\ell}| \leqslant 2^{j/2}(1+\sqrt{2})\nu$$

using $\sup_{j\ell} |\Gamma^{-1}_{j\ell}| \leqslant \nu$. (3.20) is obtained similarly, using the approximation $\sum_{j=-\log_2 T}^{-1} 2^{-j/2} \leqslant (2+\sqrt{2})\sqrt{T}$. (3.21) follows from the definition of $A_{j\ell}$ and the support of the autocorrelation wavelets, using $|\Psi_j(\tau)| \leqslant 1$ uniformly in j and τ. □

3.4 The process and its evolutionary spectrum

As we will note below, our definition of locally stationary wavelet processes differs from the original definition of Nason et al. [83] as we only impose a total variation condition on the amplitudes instead of a Lipschitz condition.

Definition 3.1. *A sequence of doubly-indexed stochastic processes* $X_{t,T}$ $(t = 0, \ldots, T-1, T > 0)$ *with mean zero is in the class of* locally stationary wavelet processes (LSW processes) *if there exists a representation in the mean-square sense*

$$X_{t,T} = \sum_{j=-\infty}^{-1} \sum_{k=0}^{T-1} w_{jk;T}\ \psi_{jk}(t)\ \xi_{jk}, \tag{3.22}$$

where $\{\psi_{jk}(t) = \psi_{j0}(t-k)\}_{jk}$ *with* $j < 0$ *is a discrete non-decimated family of wavelets 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 if not;*

2. *For each* $j = -1, -2, -3, \ldots$, *there exists a function* $W_j(z)$ *on* $(0,1)$ *possessing the following properties:*

   (a) $\sum_{j=-\infty}^{-1} |W_j(z)|^2 < \infty$ *uniformly in* $z \in (0,1)$,

   (b) *There exists a sequence of constants* C_j *such that for each* T

$$\sup_{k=0,\ldots,T-1} \left| w_{jk;T} - W_j\left(\frac{k}{T}\right) \right| \leqslant \frac{C_j}{T} \ , \tag{3.23}$$

   (c) $W_j^2(z)$ *is bounded by* L_j *in the total variation norm, i.e.*

$$\mathrm{TV}_{[0,1]}\left(W_j^2\right) \leqslant L_j, \tag{3.24}$$

   (d) *The constants* C_j *and* L_j *are such that*

$$\sum_{j=-\infty}^{-1} N_j(N_j L_j + C_j) \leqslant \rho < \infty \tag{3.25}$$

   where $N_j = |\operatorname{supp} \psi_{j0}| = (2^{-j} - 1)(N_{-1} - 1) + 1$.

In this definition, the smoothness of W_j is determined by the constant L_j for all $j < 0$. In view of (3.25), the sequence of constants $(L_j)_j$ is constrained to decrease rapidly when j tends to $-\infty$ (because $\sum_{j<0} 2^{-2j} L_j$ must be uniformly bounded). Intuitively, this contraint means that the variation of the EWS is lower for the coarsest scales, which is logical since the coarsest scales correspond to the lowest frequencies.

LSW processes use wavelets to decompose a stochastic process with respect to an orthogonal increment process in the time-scale plane. Due to the overcompleteness of the non-decimated basis, LSW processes are not uniquely determined by the sequence $\{w_{jk;T}\}$. However, we can build a theory which ensures the existence of a unique wavelet spectrum. Similarly to the situation in Chapter 2, this property is a consequence of the local stationarity setting which introduces a *rescaled time* $z = t/T \in (0,1)$ on which $W_j(z)$ is defined. The rescaled time permits increasing amounts of data about the local structure of $W_j(z)$ to be collected as the observed time T tends to infinity. Even though LSW processes are not uniquely determined by the sequence $\{w_{jk;T}\}$, the model allows to identify (asymptotically) the model coefficients determined by uniquely

defined $W_j(z)$. Then, the *evolutionary wavelet spectrum* (EWS) of an LSW process $\{X_{t,T}\}_{t=0,\ldots,T-1}$, with respect to ψ, is given by

$$S_j(z) = |W_j(z)|^2\,,\ z \in (0,1) \tag{3.26}$$

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 by Definition 3.1, $\sum_{j=-\infty}^{-1} S_j(z) < \infty$ uniformly in $z \in (0,1)$.

The evolutionary wavelet spectrum $S_j(z)$ is related to the time-depending autocorrelation function of the LSW process. Observe that the autocovariance function of an LSW process can be written as

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

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 autocovariance converges asymptotically to a *local autocovariance* defined by

$$c_X\,(z,\tau) = \sum_{j=-\infty}^{-1} S_j(z)\Psi_j\,(\tau) \tag{3.27}$$

where $\Psi_j(\tau)$ is the autocorrelation wavelet function defined above.

Proposition 3.1. *Under the assumptions of Definition 3.1, if* $T \to \infty$

$$\sum_{\tau=-\infty}^{\infty} \int_0^1 dz\ \left|c_{X,T}\,(z,\tau) - c_X\,(z,\tau)\right| = O\left(T^{-1}\right)$$

for all LSW process.

Proof. In the following, we adopt the convention that $|w_{jm;T}|^2$ is zero if $m \geqslant T$ or $m < 0$, and $S_j(u) = 0$ if u is not in the interval $[0,1]$. On one hand, due to Definition 3.1, and equation (3.23), we have

$$\begin{aligned} c_{X,T}(z,\tau) &= \mathrm{Cov}\left(X_{[zT],T}, X_{[zT]+\tau,T}\right) \\ &= \sum_{j=-\infty}^{-1} \sum_{k=-\infty}^{\infty} |w_{j,k+[zT];T}|^2 \psi_{jk}(0)\psi_{jk}(\tau) \\ &= \sum_{j=-\infty}^{-1} \sum_{k=-\infty}^{\infty} S_j\left(\frac{k+[zT]}{T}\right) \psi_{jk}(0)\psi_{jk}(\tau) + \mathrm{Rest}_T(z,\tau) \end{aligned}$$

where the remainder is such that

$$|\operatorname{Rest}_T(z,\tau)| \leqslant T^{-1} \sum_{j=-\infty}^{-1} \sum_{k=-\infty}^{\infty} C_j |\psi_{jk}(0)\psi_{jk}(\tau)|$$

by Assumption (3.23). On the other hand, we have

$$c_X(z,\tau) = \sum_{j=-\infty}^{-1} \sum_{k=-\infty}^{\infty} S_j(z)\, \psi_{jk}(0)\psi_{jk}(\tau).$$

Then,

$$\begin{aligned}
&\sum_{\tau=-\infty}^{\infty} \int_0^1 dz |c_{X,T}(z,\tau) - c_X(z,\tau)| \\
&\leqslant \sum_{\tau=-\infty}^{\infty} \int_0^1 dz \sum_{j=-\infty}^{-1} \sum_{k=-\infty}^{\infty} \left| S_j\left(\frac{k+[zT]}{T}\right) - S_j(z) \right| |\psi_{jk}(0)\psi_{jk}(\tau)| \\
&\qquad\qquad + \sum_{\tau=-\infty}^{\infty} \int_0^1 dz\, |\operatorname{Rest}_T(z,\tau)|
\end{aligned}$$

With appropriate changes of variables, this bound may be written

$$\begin{aligned}
&\sum_{\tau=-\infty}^{\infty} \sum_{j=-\infty}^{-1} \sum_{k=-\infty}^{\infty} \sum_{t=0}^{T-1} \int_0^{1/T} dz \left| S_j\left(\frac{k+[zT]+t}{T}\right) - S_j\left(z+\frac{t}{T}\right) \right| \\
&\qquad\qquad \times |\psi_{jk}(0)\psi_{jk}(\tau)| \\
&\qquad\qquad + \sum_{\tau=-\infty}^{\infty} \int_0^1 dz\, |\operatorname{Rest}_T(z,\tau)|
\end{aligned}$$

which is bounded by

$$\begin{aligned}
&T^{-1} \sum_{\tau=-\infty}^{\infty} \sum_{j=-\infty}^{-1} \sum_{k=-\infty}^{\infty} |k|\, \mathrm{TV}(S_j)\, |\psi_{jk}(0)\psi_{jk}(\tau)| + \\
&\qquad\qquad \sum_{\tau=-\infty}^{\infty} \int_0^1 dz\, |\operatorname{Rest}_T(z,\tau)|
\end{aligned}$$

where we have used the following property of the Total Variation norm:

$$\sum_{t=0}^{T-1} \left| S_p\left(\frac{t}{T}+\frac{\alpha}{T}\right) - S_p\left(\frac{t}{T}+\frac{\beta}{T}\right) \right| \leqslant |\alpha-\beta| \operatorname{TV}(S_p) \tag{3.28}$$

for all $\alpha, \beta \in \mathbb{N}$. As the support of $\psi_{jk}(0)$ is of length N_j, we get $|k| \leqslant N_j$. Together with condition (3.24) of Definition 3.1, this leads to

$$\begin{aligned}\sum_{\tau=-\infty}^{\infty} \int_0^1 dz |c_{X,T}(z,\tau) - c_X(z,\tau)| \\ \leqslant T^{-1} \sum_{j=-\infty}^{-1} (C_j + N_j L_j) \sum_{\tau=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} |\psi_{jk}(0)\psi_{jk}(\tau)|.\end{aligned}$$

The compact support of ψ_{jk} limits the sums over k and τ as follows:

$$\sum_{\tau,k=-\infty}^{\infty} |\psi_{jk}(0)\psi_{jk}(\tau)| = \sum_{\tau=-N_j+1}^{N_j-1} \sum_{-\infty}^{\infty} |\psi_{jk}(0)\psi_{jk}(\tau)| \leqslant 2N_j - 1 \tag{3.29}$$

by the Cauchy-Schwarz inequality for the sum over k. We get the result by Assumption (3.25). □

In Section 3.3.2, we have studied some mathematical properties of the ACW system appearing in (3.27). Like wavelets themselves, this system enjoys good localisation properties. Consequently, we observe that equation (3.27) is a multiscale decomposition of the autocovariance structure of the process over time: The larger the wavelet spectrum $S_j(z)$ is at a particular scale j and point z in the rescaled time, the more dominant is the contribution of scale j in the variance at time z. Thus, the evolutionary wavelet spectrum describes the distribution of the (co)variance at a particular scale and time location.

Moreover, the symmetry of the ACW implies the symmetry of the local autocovariance function, i.e. $c(z,\tau) = c(z,-\tau)$, as expected. Also, Lemma 3.1 shows that the local autocovariance function is uniquely defined. Furthermore, a consequence of the resolution of the identity (Lemma 3.3) is that the EWS of a White Noise process is proportional to 2^j for all scales $j < -1$.

It is worth mentioning that a stationary process with an absolutely summable autocovariance function is an LSW process [83, Proposition

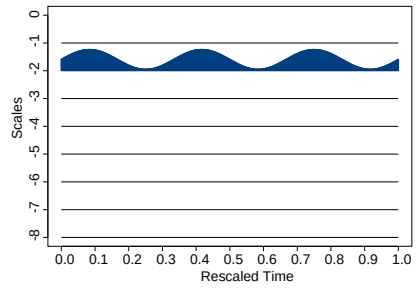
3]. Stationarity is characterized by a wavelet spectrum which is constant over time: $S_j(z) = S_j$ for all $z \in (0,1)$. As a consequence, the local autocovariance function of a time-modulated process (see equation (1.1) of Chapter 1) has the multiplicative structure

$$c(z,\tau) = \sigma^2(z)\rho_Y(\tau) \ ,$$

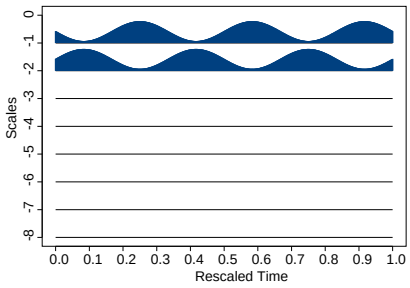
where ρ_Y is the autocorrelation function of the process Y_t in (1.1).

However, our motivation to study LSW processes lies in the modelling of more complex time-varying spectra and time-varying covariance functions. The regularity of the EWS in time is determined by the smoothness of $W_j(z)$ with repect to z. Figure 3.1 shows two examples of realisations of an LSW process with a smooth wavelet spectrum. Figure 3.1(a) represents the wavelet spectrum of a time-modulated process, while Figure 3.1(d) is a process which cannot be modelled by a time-modulated process. This last example is such that $c(z,0) = \sum_j S_j(z) = 2.2$ so that the variance of the process is constant over time. Note however that its covariance is not constant over time, which may be observed on the realisation of the process: The regime of the process clearly goes from high to low frequence regions and vice versa.

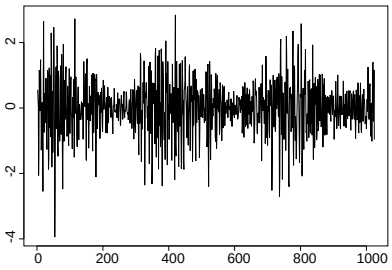
The wavelet spectra of Figure 3.1 are smooth in time. They correspond to the spectra defined by Nason et al. [83], who assumed $S_j(z)$ to be Lipschitz continuous in z. In our definition of LSW processes, it is worth mentioning that we only require the total variation norm of W_j^2 to be bounded. This weaker assumption is not only considered in order to work with minimal assumptions, but also to allow a discontinuous evolution of the wavelet spectrum in time. Consequently, our definition of nonstationary processes includes many more interesting processes, as piecewise stationary signals for instance. Figure 3.2 shows a simulated example of a more general nonstationary process, now included in our class of processes.



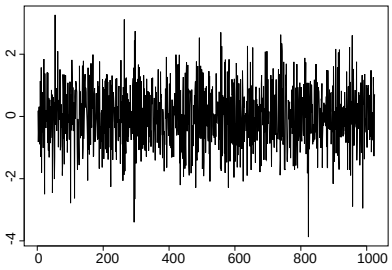
(a) Theoretical wavelet spectrum equal to zero everywhere except scale -2 where $S_{-2}(z) = 0.1 + \cos^2(3\pi z + 0.25\pi)$.



(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 3.1: These simulated examples show two examples of locally stationary processes with a smooth continuous theoretical evolutionary wavelet spectrum. The left-hand column shows an example of a smooth time-varying variance function of a time-modulated process. 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 use Gaussian innovations ξ_{jk} and Haar wavelets.

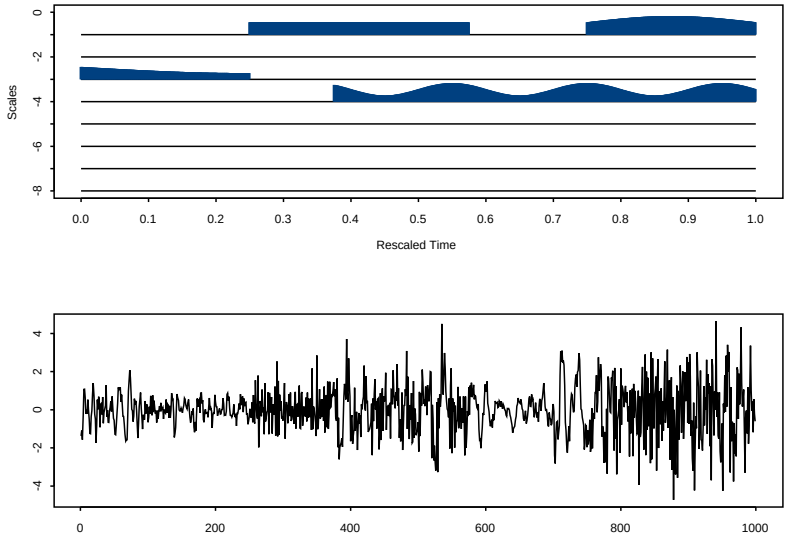


Figure 3.2: The first figure is an example of theoretical spectrum $S_j(z)$. This spectrum is used to simulate the locally stationary wavelet process plotted at the bottom. This simulation uses Gaussian innovations ξ_{jk} and non-decimated Haar wavelets.

3.5 The corrected wavelet periodogram

To end this chapter, we present a preliminary estimator of the EWS. This estimator is constructed by taking the squared empirical coefficients from the non-decimated transform:

$$I_{j;T}\left(\frac{k}{T}\right) = \left(\sum_{t=0}^{T-1} X_{t,T}\psi_{jk}(t)\right)^2$$

for all $j = -1, \ldots, -\log_2 T$ and $k = 0, \ldots, T-1$. $I_{j;T}(z)$ is called the *wavelet periodogram*, as it is analogous to the formula for the classical periodogram in traditional Fourier spectral analysis of stationary processes, see (2.7).

Some asymptotic properties of this estimator have been studied by Nason et al. [83], who showed that the wavelet periodogram is *not* an asymptotic unbiased estimator of the wavelet spectrum. Indeed, Proposition 4 of Nason et al. [83] states that, for all fixed scales $j < 0$,

$$\mathrm{E} I_{j;T}(z) = \sum_{\ell=-\log_2 T}^{-1} A_{j\ell} S_\ell(z) + O(T^{-1}), \tag{3.30}$$

uniformly in $z \in (0,1)$, where the matrix A is defined in (3.17).

Equation (3.30) motivates the definition of another preliminary estimator of the EWS, given by an appropriate correction of the periodogram in order to get an asymptotically unbiased estimator. First, define $J := [\log_2 T]$ and define the $J \times J$ matrix $A_T := (A_{j\ell})_{-1 \geqslant j, \ell \geqslant -J}$. Then, define the *corrected wavelet periodogram* (CWP)

$$L_{j;T}\left(\frac{k}{T}\right) = \sum_{\ell=-\log_2 T}^{-1} (A_T)^{-1}_{j\ell} \left(\sum_{t=0}^{T-1} X_{t,T}\psi_{\ell k}(t)\right)^2 \tag{3.31}$$

as a preliminary estimator of the EWS.

The two next chapters are dealing with testing and estimation problems on the EWS and make use of the CWP as a preliminary estimator.

3.6 Final remarks

We end this chapter by some remarks on the LSW model.

First, one could ask if it would not be easier to define LSW processes using a decimated wavelet system because, for this system, the matrix A

reduces to the identity. Unfortunately, the answer is negative: The use of non-decimated wavelets, as described in von Sachs et al. [98], would not allow to write the local autocovariance function as a wavelet-type transform of an evolutionary spectrum, as in (3.27). Moreover, classical stationary processes are not included in the model based on decimated wavelets.

Another point is the link between the LSW class and the locally stationary Fourier (LSF) processes of Chapter 2. For each LSF process with an evolutionary spectral density (ESD) f, we denote by c its local autocovariance function, which is given by

$$c(z,\tau) = \int_{-\pi}^{\pi} d\lambda \; f(z,\lambda) \exp(i\lambda\tau).$$

It is easy to show that if an LSF process has a local autocovariance function such that $\sum_\tau |c(z,\tau)| < \infty$ uniformly in $z \in (0,1)$, then it is in the class of LSW processes. In that case, the EWS of the process is

$$S_j(z) = \sum_\ell A^{-1}_{j\ell} \int_{-\pi}^{\pi} d\lambda \; f(z,\lambda) |\hat{\psi}_\ell(\lambda)|^2 \; ,$$

where $\hat{\psi}_\ell$ denotes the Fourier transform of $\psi_{\ell 0}$.

CHAPTER 4

Locally adaptive estimation in the wavelet model

4.1 Introduction

The aim of this chapter is to propose a new estimation procedure for the evolutionary wavelet spectrum (EWS). The estimator is based on the corrected wavelet periodogram (CWP) introduced in the previous chapter, and can be seen as a local smoothing procedure of this periodogram.

In this chapter, we also study a test of significance for the coefficients of the CWP. The basic idea here is to test if the CWP is significant over a given segment of time, at a given scale. This test is important for practical purposes because a scale of the EWS can be active (i.e. nonzero) at a given time and not active at another time, and this evolution corresponds to physical changes in the process.

This test is defined and studied in Section 4.2. The test statistic is based on a functional of the wavelet periodogram. It is actually a quadratic form of the increments, which are assumed to be Gaussian, and the test rule is provided through a nonasymptotic result on the deviation of the quadratic form of Gaussian processes. However, the variance of the test statistic crucially depends on the unknown spectrum, and we present a pre-estimator of this nuisance parameter. Finally, we establish a nonasymptotic bound for the deviation of the test statistics. This bound is constructed with our pre-estimator of the variance. A theoretical study of the test power concludes Section 4.2. In particular,

we discuss the consistency and the local alternatives of the proposed test procedure.

In the following Section 4.3, we show how the results of Section 4.2 may be useful for the local estimation of the EWS. We then derive an estimation procedure following the locally adaptive method of Lepski [59]. The behaviour of this estimator is studied for the two cases where the evolutionary wavelet spectrum is either regular or irregular near the point of estimation. For sake of clarity, all the technical material is deferred to Section 4.4.

The last Section 4.5 concludes with some possible directions for future research. The practical evaluation and the computational aspects of the proposed procedures will be studied in the next chapter.

4.2 Testing the local significance of the corrected wavelet periodogram

4.2.1 Local significance

As already observed in Nason et al. [83], the possibility of having an EWS with many zero segments is a major advantage of LSW processes, in comparison with other locally stationary models. The exploratory analysis of such wavelet spectra is easier, for instance if we want to detect significant variations in the multiscale structure of the process (co)variance.

In this chapter, we address the problem how to test the significance of the corrected wavelet periodogram (CWP, see Section 3.5) over a given interval at a given scale. More formally, we will test the null hypothesis

$$\mathrm{H}_0 : S_j(z) = 0 \text{ for a fixed scale } j < 0 \text{ and for all } z \in \mathcal{R}, \qquad (4.1)$$

where $\mathcal{R} \subseteq (0,1)$ is an interval with non zero measure. It is then possible to test if, for instance, a whole scale is "active" or not, or if it is non zero before or after a fixed time point.

The next subsection defines a preliminary estimator of the wavelet spectrum. Then, derivation of a test statistic is considered in subsections 4.2.2 and 4.2.3. This test is discussed in Subsection 4.2.4, where we also study its power under some alternatives.

A practical description of the algorithm and a practical evaluation of its perforances on simulated examples, is given in the next chapter.

The current chapter will derive the theoretical properties of our test procedure.

4.2.2 Derivation of the test statistic and its properties

Suppose we want to test (4.1), i.e. to check if the wavelet spectrum is zero at a fixed scale j and on a given segment of time $\mathcal{R} = (s_1, s_2) \subseteq (0,1)$ for $s_1 < s_2$. Under the null (4.1), the averaged wavelet spectrum

$$Q_{j,\mathcal{R}} = |\mathcal{R}|^{-1} \int_{\mathcal{R}} dz \; S_j(z) \tag{4.2}$$

is zero. If we observe $\underline{X}_T = (X_{0,T}, \ldots, X_{T-1,T})'$, a natural estimate of $Q_{j,\mathcal{R}}$ is

$$Q_{j,\mathcal{R};T} = |\mathcal{R}T|^{-1} \sum_{k \in \mathcal{R}T} L_{j;T}\left(\frac{k}{T}\right) \tag{4.3}$$

where $L_{j;T}(k/T)$ is the corrected wavelet periodogram (3.31) and $k \in \mathcal{R}T$ means $k/T \in \mathcal{R}$. $Q_{j,\mathcal{R};T}$ is the test statistic we use to test H_0. In this section, we will study the statistical properties of $Q_{j,\mathcal{R};T}$ under a set of assumptions.

Assumption 4.1. The autocovariance function $c_{X,T}$ and the local autocovariance function c_X of the LSW process are such that

$$\|c_{X,T}\|_{1,\infty} := \sum_{\tau=-\infty}^{\infty} \sup_{t=0,\ldots,T-1} \left| c_{X,T}\left(\frac{t}{T}, \tau\right) \right| \tag{4.4}$$

is uniformly bounded in T, and

$$\|c_X\|_{1,\infty} := \sum_{\tau=-\infty}^{\infty} \sup_{z \in (0,1)} |c_X(z,\tau)| < \infty. \tag{4.5}$$

$\Diamond$

This assumption is needed to control the spectral norm of the covariance matrix of the process (Lemma 4.3 in Section 4.4). For a stationary process, it reduces to absolute summability of the autocovariance of the process (short memory property).

Assumption 4.2. There exists an $\varepsilon > 0$ such that, for all $z \in (0,1)$, $\sum_{j=-\infty}^{-1} S_j(z) \geqslant \varepsilon$. $\Diamond$

According to equation (3.27), the sum over scales of $S_j(z)$ is the local variance of the process at time $[zT]$, and this assumption says that the local variance of the process is nowhere zero.

Assumption 4.3. The increment process $\{\xi_{jk}\}$ involved in Definition 3.1 is Gaussian. ◊

This assumption allows substantial simplifications in the proofs. It is also assumed to establish some results in Nason et al. [83]. However, Fryźlewicz [38] mentions that non-Gaussian increment processes would be more appropriate to capture some stylised facts of economic processes, such as the leptokurtic behaviour of the data. To this end, the extension of our results to non-Gaussian processes would be a feasible task using the methodology presented in Neumann and von Sachs [85] or Spokoiny [106] for instance.

Assumption 4.4. The evolutionary wavelet spectrum $S_j(z)$ defined in (3.26) is such that

$$\sum_{\ell=-\infty}^{-[\log_2(T)]-1} \sup_{z\in(0,1)} S_\ell(z) = O\left(T^{-1}\right). \qquad \lozenge$$

This assumption is necessary in order to control the difference between the EWS and the CWP at lower scales. Recall that our definition of the LSW processes involves the scales $j = -1$ up to $-\infty$, while the CWP is defined for scales $j = -1$ to $j = -[\log T]$ only.

The following proposition describes the asymptotic properties of $Q_{j,\mathcal{R};T}$.

Proposition 4.1. *Suppose Assumption 4.1 to 4.4 hold true. For all LSW process (Definition 3.1), and for all $\mathcal{R} \subseteq (0,1)$,*

$$\begin{aligned} \mathrm{E}Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}} &= \frac{K_0\, 2^{j/2}}{\sqrt{T}} \sum_{m=-\log_2 T}^{-1} N_m \,\mathrm{TV}\left(S_m\right) \\ &\qquad + O\left(2^{j/2}\,|\mathcal{R}T|^{-1}\right) \qquad (4.6) \\ &= O\left(\frac{2^{j/2}}{\sqrt{T}}\right), \end{aligned}$$

for all $j = -1, \ldots, -J_T$ *with* $J_T = O(\log_2 T)$*, and where* K_0 *is a constant independent of* j, T *and* $|\mathcal{R}|$*. Moreover, if Assumptions 4.1 to 4.3 hold, then there exists* $T_0 > 1$ *such that, for all* $T \geqslant T_0$*,*

$$K_1\, 2^{2j}\, |\mathcal{R}T|^{-1}\,(1 + o_T(1)) \leqslant \operatorname{Var} Q_{j,\mathcal{R};T} \leqslant K_2\, 2^j |\mathcal{R}|^{-2} T^{-1}$$

for all $j = -1, \ldots, -J_T$ *with* $J_T = O(\log_2 T)$*, and where* K_1 *and* K_2 *are two constants independent of* j, T *and* $|\mathcal{R}|$*.*

The proof of this proposition is in Section 4.4.1. Note that the squared bias and the variance of the estimator have the same rate of convergence. This phenomenon is due to the nonstationary behaviour of the process. Indeed, for a stationary process, the total variation norm of S_m is zero at all scales, and then the rate of the bias is T^{-1}. This is not the case for a general nonstationary process: When the wavelet spectrum is not constant over time, an additional term resulting from nonstationarity reduces considerably this rate of convergence. Moreover, even when we are dealing with a *local* estimator of the wavelet spectrum at a fixed scale $j < 0$ and a fixed time interval $\mathcal{R}$, the nonstationarity term in the bias involves the variation of the *global* wavelet spectrum. This may be observed in equation (4.6), which involves a sum over all scales $m = -1, \ldots, -\log_2 T$ and the total variation norm of all S_m over the whole rescaled time interval $(0, 1)$.

This slow rate of convergence of the bias poses a problem to establish the asymptotic normality of $Q_{j,\mathcal{R};T}$. In the next proposition, we circumvent this problem and derive a non asymptotic exponential bound for the deviation of $Q_{j,\mathcal{R};T}$.

Proposition 4.2. *Assume that (4.4) and Assumption 4.1 to 4.4 hold. If* $\sigma^2_{j,\mathcal{R},T} = \operatorname{Var} Q_{j,\mathcal{R};T}$*, then, for all* $\eta > 0$ *and for all scales* $j = -1, \ldots, -J_T$*, where* $J_T = O(\log_2 T)$*,*

$$\begin{aligned}&\Pr\left(|Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}| > \sigma_{j,\mathcal{R},T}\eta\right)\\ &\quad\leqslant c_0 \exp\left\{-\frac{1}{8}\cdot\frac{\eta^2}{1 + \frac{\eta L_j}{|\mathcal{R}T|\sigma_{j,\mathcal{R},T}} + \frac{2^{j/2}\eta\nu}{|\mathcal{R}|\sqrt{T}\sigma_{j,\mathcal{R};T}}\left(\|c_X\|_{1,\infty} + c_1\rho\right)}\right\}\end{aligned}$$

with the positive constants $c_0 = 1 + e$ *and* $c_1 = (2 + \sqrt{2})/2$*, where* ρ *is given in Definition 3.1 and where* ν *is a universal positive constant depending only on the wavelet* ψ*.*

The proof of this proposition is to be found in Section 4.4.2. This proposition gives a non asymptotic bound for the deviation of the test statistics $Q_{j,\mathcal{R};T}$. It can be used in order to construct a test rule, i.e. to choose η such that the exponential function in the proposition is the nominal level of the test (see Section 4.2.4 below, where the test rule is given explicitely). From an asymptotic viewpoint, i.e. as $T \to \infty$, we note that this exponential bound does not tend to zero, meaning that the standardised statistic $Q_{j,\mathcal{R},T}$ is asymptotically non degenerate. This phenomenon is well-known in the context of pointwise estimation [16, 59]. In order to have a consistent result when $T \to \infty$, it is then necessary to impose that $\eta = \eta_T$ grows with T. The appropriate rate for η_T is derived in the next corollary. The proof is given in Section 4.4.2 and is essentially based on the bounds derived in Proposition 4.1.

Corollary 4.1. *Under the assumptions of Propositions 4.1 and 4.2, if k_T tends to infinity and is such that*

$$J_T \exp(-k_T) = o_T(1), \tag{4.7}$$

then there exists a $T_0 > 1$ such that, for all $T \geqslant T_0$,

$$\Pr\left(\sup_{-J_T \leqslant j < 0} |Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}| \geqslant k_T \sqrt{K_2 |\mathcal{R}|^{-2} T^{-1}} \right) = o_T(1)$$

where K_2 is as in the assertion of Proposition 4.1.

Remark 4.1. An example of admissible rates is $J_T \sim \log_2 T$ and $k_T \sim \log_2 T$. Here, the sequence k_T will play a crucial role in Section 4.3. $\Diamond$

4.2.3 Estimation of the variance

If we want to use Proposition 4.2 to test H_0, an estimator of the variance $\sigma^2_{j,\mathcal{R},T} = \operatorname{Var} Q_{j,\mathcal{R};T}$ is needed. This variance depends on the unknown autocovariance function of the LSW process in the following way (see Lemma 4.1 with equation (4.21)):

$$\sigma^2_{j,\mathcal{R},T} = 2 \, \|U'_{j,\mathcal{R};T} \Sigma_T\|_2^2,$$

where Σ_T is the $T \times T$ (non-Toeplitz) covariance matrix of the LSW process $(X_{0,T}, \ldots, X_{T-1,T})'$, and $U_{j,\mathcal{R};T}$ is the $T \times T$ matrix with entry

(s,t) equal to

$$U_{st}^{(j)} = |\mathcal{R}T|^{-1} \sum_{\ell=-[\log_2 T]}^{-1} A_{j\ell}^{-1} \sum_{k \in \mathcal{R}T} \psi_{\ell k}(s)\psi_{\ell k}(t). \tag{4.8}$$

where the matrix A is actually the finite matrix A_T (see Section 3.5). We also denote by $\sigma_{s,s+u}$ the entry $(s, s+u)$ of the matrix Σ_T. Some useful properties of Σ are derived in Section 6.3 below.

We will estimate $\sigma^2_{j,\mathcal{R},T}$ by:

$$\tilde{\sigma}^2_{j,\mathcal{R},T} = 2 \, \|U'_{j,\mathcal{R};T}\tilde{\Sigma}_T\|_2^2$$

where $\tilde{\Sigma}_T$ is an estimate of the covariance matrix Σ_T. A first idea is to define the elements $\tilde{\sigma}_{s,s+u}$ of $\tilde{\Sigma}_T$ by plugging $Q_{j,\mathcal{R};T}$ into the local autocovariance function (3.27), i.e.

$$\tilde{\sigma}_{s,s+u} = \sum_{j=-[\log_2 T]}^{-1} Q_{j,\mathcal{R}(s);T}\Psi_j(u),$$

where $\mathcal{R}(s)$ denotes an interval which contains the time point s/T. However, the convergence in probability of $\tilde{\sigma}_{s,s+u}$ to $\sigma_{s,s+u}$ is not faster than the rate of $\sigma_{s,s+u}$ itself, and we need to modify the estimator in two ways.

(i) Assumption 4.1 indicates that the covariance $|\sigma_{s,s+u}|$ is small for large $|u|$. Then, following the method of Giurcanu and Spokoiny [40], we set $\tilde{\sigma}_{s,s+u}$ to zero when $|u| \geqslant M_T$, for an appropriate sequence M_T tending to infinity with T;

(ii) It is necessary to control the distance in rescaled time between the spectrum $S_j(z)$, for $z \in \mathcal{R}(s)$, and $S_j(s/T)$. To do so, we allow the window $\mathcal{R}(s)$ to depend on T, which is denoted by $\mathcal{R}_T(s)$, in such a way that its length $|\mathcal{R}_T|$ shrinks to zero when T tends to infinity. This is analogous to the estimation of a regression function by kernel smoothing, where the window usually depends on the length of the data set.

With these two ingredients, we propose to estimate $\sigma_{s,s+u}$ by

$$\tilde{\sigma}_{s,s+u} = \sum_{j=-[\log_2 T]}^{-1} Q_{j,\mathcal{R}_T(s);T}\Psi_j(u)1_{|u|\leqslant M_T}, \tag{4.9}$$

and the following assumption makes precise the appropriate rates for the sequences $|\mathcal{R}_T|$ and M_T.

Assumption 4.5. The sequence J_T (i.e., the sequence that defines the lower scale for the test of significance) is such that $J_T = o(\log_2 T)$. Moreover, $|\mathcal{R}_T|$ tends to zero and the sequence k_T (which appears in Corollary 4.1) is such that $J_T \exp(-k_T\sqrt{|\mathcal{R}_T|}) = o_T(1)$. Finally, the sequence M_T (involved in the preliminary estimator for the variance, see (4.9)) tends to infinity such that

$$2^{J_T}|\mathcal{R}_T T|^{-1/2} M_T k_T \log_2^3 T = o_T(1). \qquad \Diamond$$

Admissible rates for this last assumption are $J_T \sim \log_2 \log_2^2 T$, $k_T \sim \log^2 T$, $|\mathcal{R}_T| \sim \log_2^{-3} T$ and $M_T \sim \log_2^\alpha T$ with $\alpha > 0$. It is worth mentioning that, with this assumption, $|\mathcal{R}_T|$ shrinks to zero in the *rescaled* time, whereas, in the *observed* time, the interval length $|T\mathcal{R}_T|$ tends to infinity. This means that our estimate of $S_j(s/T)$ is built using an increasing amount of data in the observed time, but, at the same time, with an average around $S_j(s/T)$ in the rescaled time on a shrinking segment around s/T.

The next proposition shows that on the random set where the estimator $Q_{j,\mathcal{R}_T(s);T}$ is near $Q_{j,\mathcal{R}_T(s)}$, the estimator (4.9) has a good quality. Our proof of this proposition may be found in Section 4.4.3 and needs the following technical assumption, which is a slightly stronger condition than the point 2(a) of Definition 3.1, in the sense that we need to control the decay of $S_j(z)$ with respect to j and uniformly in z.

Assumption 4.6. The local autocovariance function $c(z,\tau)$ is such that

$$\sum_{u=-\infty}^{\infty} \sup_z |c_X(z,u)| 1_{|u|>M_T} = o_T\left(2^{-J_T}\right). \qquad \Diamond$$

This last assumption on the decay of the local autocovariance function uniformly in z, is very sensible in a context of short-memory processes, i.e. when $c(z,u)$ does not depend on z. With the rates specified above, a typical condition is to assume $|c_X(z,u)| \leqslant c \cdot 2^{-|u|}$ uniformly in $z \in (0,1)$. This assumption is sensible even for stationary short-memory processes (it holds for instance for a stationary AR(1) process).

Proposition 4.3. *Suppose Assumptions 4.1 to 4.6 hold. Then, there exists a positive number T_0 and a random set $\mathcal{A}$ independent of j and such that $\Pr(\mathcal{A}) \geqslant 1 - o_T(1)$ and*

$$|Q_{j,\mathcal{R}_T(s);T} - Q_{j,\mathcal{R}_T(s)}| \leqslant \frac{k_T}{|\mathcal{R}_T|}\sqrt{\frac{K_2}{T}}$$

for all $T > T_0$. Moreover, on $\mathcal{A}$,

$$2^{J_T - j}\, T\, |\tilde{\sigma}^2_{j,\mathcal{R},T} - \sigma^2_{j,\mathcal{R},T}| = o_P(1) \tag{4.10}$$

holds for all $j = -1, \ldots, -J_T$, where $o_P(1)$ does not depend on $\mathcal{R}$.

Finally, Proposition 4.3 together with Proposition 4.2 leads to the following result, which will be used to construct the test in practice.

Theorem 4.1. *Suppose Assumptions 4.1 to 4.6 hold. Then, there exists a $\varphi_T = o_T(2^j T^{-1})$ and a positive number T_0 such that, for all $T > T_0$,*

$$\Pr\left(|Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}| > \tilde{\sigma}_{j,\mathcal{R},T}\eta'\right) \leqslant c_0 \exp\left\{-\frac{1}{8}\cdot\frac{\eta^2}{1 + \frac{\eta}{|\mathcal{R}T|\sigma_{j,\mathcal{R},T}}L_j + \frac{2^{j/2}\eta\nu(\|c_X\|_{1,\infty} + c_1\rho)}{|\mathcal{R}|\sqrt{T}\sigma_{j,\mathcal{R};T}}}\right\} + o_T(1)$$

for all $j = -1, \ldots, -J_T$, where $\eta' = \eta(1 - \varphi_T/\sigma^2_{j,\mathcal{R};T})^{1/2}$, and the positive constants c_0, c_1 are defined in the assertion of Proposition 4.2.

Remark 4.2. Theorem 4.1 gives an upper bound for the deviation of the normalized loss $|Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}|/\tilde{\sigma}_{j,\mathcal{R},T}$. This bound depends on the unknown quantities $\|c_X\|_{1,\infty}$ and ρ, cf. (3.25). These two quantities may be understood as nuisance parameters of the problem, depending on the global spectrum. The estimation of these quantities is based on a preliminary smoothing of $L_{j;T}(z)$ with respect to z, which we denote by $L^*_{j;T}(z)$. Here, we think about using a kernel smoothing procedure, or a wavelet transform shrinkage as studied in Nason et al. [83]. Then, a preliminary estimate of $\|c_X\|_{1,\infty}$ is obtained by plugging $L^*_{j;T}(z)$ into $\|c_X\|_{1,\infty}$, cf. (3.27) and (4.5). Next, the preliminary estimation of ρ necessitates the estimation of $\mathrm{TV}(S_j)$, cf. (3.24). We estimate $\mathrm{TV}(S_j)$ by $\sum_i |L^*_{j;T}(z_i^{\max}) - L^*_{j;T}(z_i^{\min})|$, where the sum is over the local minima and maxima of $L^*_{j;T}(z)$, with $z_i^{\max} < z_{i+1}^{\min} < z_{i+1}^{\max}$ for all i. $\Diamond$

4.2.4 Discussion of the test procedure

We now propose our test procedure. Under H_0, see (4.1), the approximation of the deviation of the test statistic is given by

$$\Pr\left(|Q_{j,\mathcal{R};T}| > \eta' \tilde{\sigma}_{j,\mathcal{R},T} \middle| \mathrm{H}_0\right) \leqslant h(\eta') \tag{4.11}$$

for T sufficiently large, and where h is the exponential function following from Theorem 4.1. Let α be the nominal level of the test. We reject H_0 if

$$|Q_{j,\mathcal{R};T}| > \eta_\star \tilde{\sigma}_{j,\mathcal{R},T}, \tag{4.12}$$

where $\eta_\star$ is such that $h(\eta_\star) = \alpha$.

We now discuss the power of this test and, for this, we need to be more specific about the alternative hypothesis H_1. We will work with the sensible alternative hypothesis that there exists a strictly positive real number θ and a measurable set with a non zero measure $\mathcal{U} \subseteq \mathcal{R}$ such that $S_j(z) \geqslant \theta$ for all z in $\mathcal{U}$. Figure 4.1 illustrates this situation.

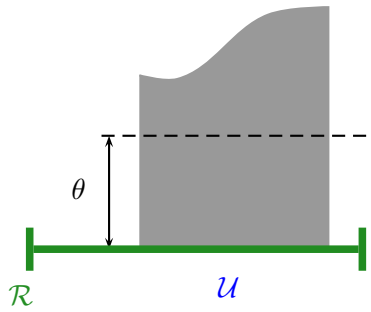


Figure 4.1: Alternative hypothesis.

Formally, if $|\mathcal{U}|$ denotes the Lebesgue measure of $\mathcal{U}$:

$$\mathrm{H}_1 : \exists\, \theta > 0 \text{ and } \mathcal{U} \subseteq \mathcal{R} \text{ with } |\mathcal{U}| > 0 \text{ and } S_j(z) \geqslant \theta \quad \forall z \in \mathcal{U}. \tag{4.13}$$

The next proposition evaluates the type II error of the test. The proof is to be found in Section 4.4.5.

Proposition 4.4. *Suppose Assumption 4.1 to 4.6 hold true. Let the null hypothesis (4.1) against the alternative hypothesis (4.13) be given and consider the test rule (4.12) with*

$$\eta_\star < 2^{(1-j)/2} \|c_X\|_{1,\infty}^{-1} |\mathcal{R}T|^{1/2} Q_{j,\mathcal{R}}.$$

Then, there exists $T_0 > 1$ such that, for all $T \geqslant T_0$, the type II error of the test is bounded as follows:

$$\Pr\left(H_0 \text{ is not rejected} \,\middle|\, H_1\right) \leqslant C' \cdot \exp\left\{-c' \frac{T}{\log_2^2 T} \frac{\theta^2 |\mathcal{U}|^2}{|\mathcal{R}|^2}\right\} + C'' \cdot \exp\left\{-c'' \frac{\sqrt{T}}{\log_2^2 T} \frac{\theta |\mathcal{U}|}{|\mathcal{R}|}\right\} + o_T(1),$$

where the positive constants c', C', c'', C'' and the $o_T(1)$ term do not depend on $\mathcal{R}$, $\mathcal{U}$ and θ.

The last result shows the consistency of the test procedure. Moreover, it allows to discuss the local alternative of the test. We first note that the alternative hypothesis (4.13) depends on the two parameters θ and $\mathcal{U}$. Consequently, to study the local alternative of the test, we need to investigate both cases $\theta = \theta_T \to 0$ and $\mathcal{U} = \mathcal{U}_T$ such that $|\mathcal{U}_T| \to 0$. However, the upper bound of the type II error in Proposition 4.4 depends on the product $\theta_T |\mathcal{U}_T|$, and then the local alternative of the test is studied when this product tends to 0 when $T \to \infty$. By straightforward considerations, we see that if

$$\frac{\log_2^2 T}{\theta_T |\mathcal{U}_T| \sqrt{T}}$$

tends to zero as $T \to \infty$, then the type II error of the test aymptotically vanishes.

4.3 Pointwise adaptive estimation

Theorem 4.1 may be also useful for other statistical applications. In this section, we derive one important application given by the pointwise estimation of the wavelet spectrum.

Indeed, the estimator $Q_{j,\mathcal{R};T}$ may be seen as a smoothing over time of the inconsistent corrected wavelet periodogram. It can then be used for the pointwise estimation of the wavelet spectrum. In this problem, we want to estimate $S_j(z_0)$ at a fixed point z_0. This estimation can be done by computing the histogram $Q_{j,\mathcal{R};T}$ constructed on a segment $\mathcal{R}$ containing the fixed time point z_0. Consequently, the question how to choose the best segment $\mathcal{R}$ arises, and the goal of this section is to provide a data-driven procedure to select $\mathcal{R}$ automatically.

The proposed method goes back to the pointwise adaptive estimation theory of Lepski [59, 60, 105]. Suppose that the wavelet spectrum at $S_j(z_0)$ is well approximated by the averaged spectrum $Q_{j,\mathcal{U}}$ for a given interval $\mathcal{U}$ containing the reference point z_0. The idea of the procedure is to consider a second interval $\mathcal{R}$ containing $\mathcal{U}$ and to test if $Q_{j,\mathcal{R}}$ differs significantly from $Q_{j,\mathcal{U}}$. As we describe below, this test procedure is based on Proposition 4.2 or Theorem 4.1. If there exists a subset $\mathcal{U}$ of $\mathcal{R}$ such that $|Q_{j,\mathcal{R}} - Q_{j,\mathcal{U}}|$ is significantly different from zero, then we reject the hypothesis of homogeneity of the wavelet spectrum $S_j(z)$ on $z \in \mathcal{R}$. Finally, the adaptive estimator corresponds to the largest interval $\mathcal{R}$ such that the hypothesis of homogeneity is not rejected.

This section contains a formal description of this algorithm and derives some properties of the estimator.

4.3.1 Testing homogeneity

Let $\mathcal{R}$ be an interval containing z_0, $\mathcal{U}$ a subset of $\mathcal{R}$ and define

$$\Delta_j(\mathcal{R},\mathcal{U}) = |Q_{j,\mathcal{R}} - Q_{j,\mathcal{U}}|. \tag{4.14}$$

Under assumptions 4.1 to 4.3, Proposition 4.2 implies

$$\begin{aligned}\Pr\left[|Q_{j,\mathcal{R},T} - Q_{j,\mathcal{U},T}| > \Delta_j(\mathcal{R},\mathcal{U}) + \eta\left(\sigma_{j,\mathcal{R},T} + \sigma_{j,\mathcal{U},T}\right)k_T\right] \\ \leqslant h(\mathcal{U},\eta) + h(\mathcal{R},\eta)\end{aligned} \tag{4.15}$$

with

$$h(\mathcal{R},\eta) = c_0 \exp\left\{-\frac{1}{8}\cdot\frac{\eta^2 k_T^2}{1+\frac{\eta k_T L_j}{|\mathcal{R}T|\sigma_{j,\mathcal{R},T}}+\frac{2^{j/2}\eta k_T \nu}{\sqrt{|\mathcal{R}T|}\sigma_{j,\mathcal{R};T}}\frac{\sqrt{|\mathcal{R}|}\|c_X\|_{1,\infty}+c_1\rho}{\sqrt{|\mathcal{R}|}}}\right\}$$

and where the sequence k_T is such that (4.7) holds. Under the assumption that the wavelet spectrum S_j is homogeneous on the segment $\mathcal{R}$, the difference $\Delta_j(\mathcal{R},\mathcal{U})$ is negligeable. Then, as a test rule, we reject the homogeneity hypothesis on $\mathcal{R}$ if there exists a subset $\mathcal{U} \subset \mathcal{R}$ such that $|Q_{j,\mathcal{R};T} - Q_{j,\mathcal{U};T}| > \eta(\sigma_{j,\mathcal{R},T} + \sigma_{j,\mathcal{U},T})k_T$ for a given η.

In the case where the variances $\sigma_{j,\mathcal{R},T}$ and $\sigma_{j,\mathcal{U},T}$ are unknown, they may be estimated as in Section 4.2.3 above. In that case, the homogeneity test is based on Theorem 4.1 and the modification of the following results is straightforward.

In practice, we choose a set Λ of interval-candidates $\mathcal{R}$. Then, for each candidate $\mathcal{R}$, we apply the homogeneity test with respect to a given set $\wp(\mathcal{R})$ of subintervals $\mathcal{U}$ of $\mathcal{R}$.

Assumption 4.7. In the estimation procedure described below, we assume the following properties on the test sets Λ and $\wp(\mathcal{R})$:

1. For all $\mathcal{R}$, the shortest interval of $\wp(\mathcal{R})$ is of length at least $\delta > 0$,
2. The cardinality of $\wp(\mathcal{R})$ is such that $\sharp(\wp(\mathcal{R})) \leqslant |\mathcal{R}T|^{\kappa\alpha\sqrt{\delta}}$ for some $0 < \alpha < 1$ and $\kappa \leqslant \sqrt{K_1}/[\nu(\|c\|_{1,\infty} + c_1\rho)]$,
3. When we test the homogeneity of the wavelet spectrum on $\mathcal{R}$, we assume that there exists a subinterval $\mathcal{U} \in \wp(\mathcal{R})$ such that $\mathcal{U} \subset \mathcal{R}$ and $\mathcal{U}$ contains z_0. $\Diamond$

A precise example of test sets Λ and $\wp(\mathcal{R})$ will be described in Chapter 5 (Section 5.4), where the computational aspects of the procedure is discussed with details and evaluated on empirical simulations.

4.3.2 The estimation procedure

The estimation procedure simply starts with the smallest interval in Λ, assuming that the wavelet spectrum is homogeneous on this short interval. Then, it selects iteratively longer intervals in Λ until the homonegeneity assumption is rejected. Finally, the adaptive segment $\tilde{\mathcal{R}}$ is the longest segment $\mathcal{R}$ of Λ for which the homogeneity test is not rejected:

$$\tilde{\mathcal{R}} = \arg\max_{\mathcal{R}\in\Lambda} \Big\{ |\mathcal{R}| \text{ such that } |Q_{j,\mathcal{R};T} - Q_{j,\mathcal{U};T}| \leqslant \eta(\sigma_{j,\mathcal{R},T} + \sigma_{j,\mathcal{U},T})k_T \text{ for all } \mathcal{U} \subset \wp(\mathcal{R}) \Big\}. \tag{4.16}$$

The adaptive estimator of $S_j(z_0)$ is then defined by

$$\tilde{S}_j(z_0) = Q_{j,\tilde{\mathcal{R}},T}. \tag{4.17}$$

4.3.3 Properties of the estimator in homogeneous regions

The next result quantifies the ℓ_p risk ($p \geqslant 2$) when the wavelet spectrum $S_j(z)$ is homogeneous on $z \in \mathcal{R}$. To define this concept of homogeneity, we introduce the bias

$$b(\mathcal{R}) := \sup_{z\in\mathcal{R}} |S_j(z) - Q_{j,\mathcal{R}}|,$$

which measures how well the wavelet spectrum S_j is approximated by $Q_{j,\mathcal{R}}$ on $z \in \mathcal{R}$. We say that the spectrum is *homogeneous* (or *regular*) on $\mathcal{R}$, if the inequality

$$b(\mathcal{R}) \leqslant C_j \, \sigma_{j,\mathcal{R},T} \, k_T \tag{4.18}$$

holds with

$$C_j = 2^{-j/2}\sqrt{\alpha + p}. \tag{4.19}$$

In the inequality (4.18), $\sigma_{j,\mathcal{R},T}$ is the square root of the variance of the estimator $Q_{j,\mathcal{R};T}$ of $S_j(z)$, $z \in \mathcal{R}$. Inequality (4.18) can be viewed as a balance relation between the bias and the variance of this estimate [105]. The k_T term then appears as the correction term necessary in the pointwise estimation in order to bound the normalized loss [59, 60]. In the following results, we set k_T proportional to $\log_2^2 T$.

Proposition 4.5. *Let $\mathcal{R}$ be an interval of $(0,1)$ and consider the test rule (4.16). If the wavelet spectrum S_j is regular on $\mathcal{R}$ in the sense of conditions (4.18)—(4.19), then, with $\lambda = \eta = 2^{-j/2}5(2\alpha + p)$ and $k_T \sim \log_2^2 T$,*

$$\Pr\left(\mathcal{R} \text{ is rejected}\right) = O\left(T^{-cp\sqrt{\delta}}\right)$$

for some positive constant $c = c(\nu, \|c\|_{1,\infty}, \rho)$ depending on $\nu, \|c\|_{1,\infty}$ and ρ only.

Using this proposition, we can evaluate an upper bound for the ℓ_p risk associated to our estimator.

Theorem 4.2. *Assume that the wavelet spectrum at scale j, $S_j(z)$, is homogeneous on the segment $\mathcal{R}$ in the sense of (4.18)–(4.19) with*

$$k_T \sim \log_2^2 T.$$

If $\tilde{S}_j(z)$ is the pointwise estimator of the wavelet spectrum obtained by the estimation procedure (4.16)–(4.17) with

$$\eta = 2^{-j/2}5(2\alpha + p),$$

then there exists T_0 such that the pointwise ℓ_p-loss is bounded as follows

$$\mathrm{E}|\tilde{S}_j(z) - S_j(z)|^p \leqslant c(\delta T)^{-p/2}\left[2^{1+j/2}\delta^{-1} + 11(2\alpha + p)\log_2^2 T\right]^p$$

for all $T \geqslant T_0$ with a positive constant c depending on $p, \nu, \|c_X\|_{1,\infty}$.

The proof is to be found in Section 4.4.7.

4.3.4 Properties of the estimator in inhomogeneous regions

We now describe the behaviour of our estimator near a breakpoint located at a time point $z_\star$. We first need to be more specific about the definition of a breakpoint in the evolutionary spectrum.

For a fixed scale $j \in \{-1, \dots, -J_T\}$, assume the evolutionary wavelet spectrum to be homogeneous on $\mathcal{R}_0 = [z_0, z_\star)$ and on $\mathcal{R}_1 = (z_\star, z_1]$. Let us denote $\mathcal{R} = \mathcal{R}_0 \cup \mathcal{R}_1 = [z_0, z_1]$ and

$$\theta_T := \mathrm{E}\left[Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}_0;T}\right]$$

Figure 4.2 illustrates this situation.

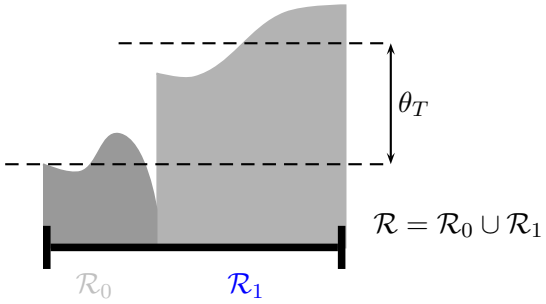


Figure 4.2: Inhomogeneous region.

To prove the next proposition, we assume that the estimation procedure is such that $\mathcal{R}_0$ and $\mathcal{R}_1$ are in $\wp(\mathcal{R})$.

Proposition 4.6. *If the evolutionary wavelet spectrum at scale j contains a breakpoint at $z_\star$ as described above and if $k_T \sim \log_2^2 T$, then*

$$\Pr\left(\mathcal{R} \text{ is not rejected}\right) = O\left(\exp\left[-\frac{T\theta_T^2(|\mathcal{R}_0| \vee |\mathcal{R}_1|)}{\log_2^2 T}\right] + \exp\left[-\frac{\sqrt{T}\theta_T}{\log_2^2 T}\right]\right).$$

where c is a positive constant and $x \vee y = \max(x, y)$.

The proof of this proposition is given in Section 4.4.8. In this result, θ_T may be seen as the level of a jump in the wavelet spectrum. Then, Proposition 4.6 informs about the minimal amplitude of the jump which

may be detected by the estimation procedure. If θ_T is such that

$$\frac{\log_2^2 T}{\theta_T \sqrt{T}} \to 0,$$

then the estimation procedure is consistent in the sense that $\Pr(\mathcal{R}$ is not rejected) is asymptotically zero.

4.4 Proofs

In the sequel, we use the convention $w_{jk;T} = 0$ for $k < 0$ and $k \geqslant T$, which leads to helpful simplifications in the following proofs.

4.4.1 Proof of Proposition 4.1

Our proof of Proposition 4.1 needs the following Lemma quoted from Neumann and von Sachs [85].

Lemma 4.1. *Let $\underline{Z}_n = (Z_1, \ldots, Z_n)'$ be a vector of iid Gaussian random variables with zero mean and* $\operatorname{Var} Z_1 = 1$. *If M_n is an $n \times n$ matrix, then*

$$\begin{aligned} &\mathrm{E}\left(\underline{Z}_n' M_n \underline{Z}_n\right) = \operatorname{tr} M_n, \\ &\operatorname{Var}\left(\underline{Z}_n' M_n \underline{Z}_n\right) = 2 \operatorname{tr} M_n^\star M_n = 2\|M_n\|_2^2, \end{aligned}$$

and, for all $r \geqslant 2$, if Cum_r *denotes the rth cumulant, we have*

$$|\operatorname{Cum}_r\left(\underline{Z}_n' M_n \underline{Z}_n\right)| \leqslant 2^{r-1}(r-1)!\, \|M_n\|_2^2\ \{\lambda_{\max}(M_n)\}^{r-2}\ ,$$

where $\lambda_{\max}$ denotes the maximal eigenvalue.

Define $\underline{X}_T = (X_{0,T}, \ldots, X_{T-1,T})'$. By definition, $Q_{j,\mathcal{R};T}$ is the quadratic form

$$Q_{j,\mathcal{R};T} = \underline{X}_T' U_{j,\mathcal{R};T} \underline{X}_T \tag{4.20}$$

where $U_{j,\mathcal{R};T}$ is the $T \times T$ matrix with entry (s,t) equal to

$$U_{st} = |\mathcal{R}T|^{-1} \sum_{\ell=-\log_2 T}^{-1} A_{j\ell}^{-1} \sum_{k \in \mathcal{R}T} \psi_{\ell k}(s) \psi_{\ell k}(t),$$

where the matrix A is actually the finite-dimensional matrix A_T. For notational convenience, we omit the dependence of U_{st} in j and $\mathcal{R}$. Assuming that the orthonormal increment processes $\{\xi_{jk}\}$ in Definition 3.1

are Gaussian, $\underline{X}_T$ is a multivariate Gaussian random variable with covariance matrix $\Sigma_T = \mathrm{Cov}(\underline{X}_T \underline{X}_T')$. In that case, $Q_{j,\mathcal{R};T}$ is a quadratic form of Gaussian variables and we can apply Lemma 4.1 with

$$M_{j,\mathcal{R};T} = \Sigma_T'^{1/2} U_{j,\mathcal{R};T} \Sigma_T^{1/2} \tag{4.21}$$

in order to prove Proposition 4.1. The following lemmas derive some bounds for the Euclidean and the spectral norm of $U_{j,\mathcal{R};T}$ and Σ_T.

Lemma 4.2. *With fixed $\mathcal{R} \subseteq (0,1)$, there exists a T_0 such that, uniformly in $T \geqslant T_0$,*

$$K_1\, 2^{2j} |\mathcal{R}T|^{-1} (1 + o_T(1)) \leqslant \|U_{j,\mathcal{R};T}\|_2^2 \leqslant K_2\, 2^j |\mathcal{R}|^{-2} T^{-1}$$

for all $j = -1, \ldots, J_T = o_T(\log_2 T)$, where K_1 and K_2 are two constants independent of j, T and $|\mathcal{R}|$.

Proof. The proof is straightforward when $\mathcal{R} = (0,1)$. However, one technical difficulty is to deal with a general interval $\mathcal{R} = (r_1, r_2) \subset (0,1)$. A simple remark which simplifies the proof is to observe that the quadratic form (4.20) involves $X_{[r_1 T],T}, \ldots, X_{T-1,T}$ only. Consequently, the matrix $U_{j,\mathcal{R};T}$ is a $(T - [r_1 T] + 1) \times (T - [r_1 T] + 1)$ matrix, and we can write, from direct computations,

$$\|U_{j,\mathcal{R};T}\|_2^2 = |\mathcal{R}T|^{-2} \sum_{s,t=[r_1 T]}^{T-1} \left(\sum_{\ell=-\log_2 T}^{-1} A_{j\ell}^{-1} \sum_{k \in \mathcal{R}T} \psi_{\ell k}(s) \psi_{\ell k}(t) \right)^2$$

The compact support of $\psi_{\ell k}(s)$ implies that $0 \leqslant k - s$, and, as $k \leqslant [r_2 T]$, we can limit the sum over s, t as follows:

$$\|U_{j,\mathcal{R};T}\|_2^2 = |\mathcal{R}T|^{-2} \sum_{s,t \in \mathcal{R}T} \left(\sum_{\ell=-\log_2 T}^{-1} A_{j\ell}^{-1} \sum_{k \in \mathcal{R}T} \psi_{\ell k}(s) \psi_{\ell k}(t) \right)^2 . \tag{4.22}$$

If we split the sum over ℓ at point ℓ_T such that $|N_\ell| \leqslant |\mathcal{R}T|$ for all $\ell = -1, -2, \ldots, \ell_T$, then $|\ell_T| = O(\log_2 |\mathcal{R}T|)$ and $\|U_{j,\mathcal{R};T}\|_2^2$ is equal to

$$|\mathcal{R}T|^{-2} \sum_{s,t \in \mathcal{R}T} \left(\sum_{\ell=-\ell_T}^{-1} A_{j\ell}^{-1} \Psi_\ell(s-t) + \sum_{m=-\log_2 T}^{-\ell_T - 1} A_{jm}^{-1} \sum_{k \in \mathcal{R}T} \psi_{mk}(s) \psi_{mk}(t) \right)^2 .$$

Expanding the square, we get three terms, herewith denoted by I_T, II_T and III_T. By definition of Ψ_ℓ and $A_{j\ell}$, the first squared term is

$$\begin{aligned}\mathrm{I}_T &= |\mathcal{R}T|^{-2} \sum_{s,t\in\mathcal{R}T} \left(\sum_{\ell=-\ell_T}^{-1} A_{j\ell}^{-1} \Psi_\ell(s-t) \right)^2 \\ &= |\mathcal{R}T|^{-1} \sum_{\ell,m=-\ell_T}^{-1} A_{j\ell}^{-1} A_{jm}^{-1} A_{m\ell}.\end{aligned}$$

If we write the sum over $-\ell_T \leqslant \ell \leqslant -1$ as the sum over $-\log_2 T \leqslant \ell \leqslant -1$ minus the sum over $-\log_2 T \leqslant \ell \leqslant -\ell_T - 1$, then we get

$$\begin{aligned}\mathrm{I}_T &= |\mathcal{R}T|^{-1} \left(\sum_{m=-\ell_T}^{-1} A_{j\ell}^{-1} \delta_{jm} - \sum_{m=-\ell_T}^{-1} \sum_{\ell=-\log_2 T}^{-\ell_T-1} A_{j\ell}^{-1} A_{jm}^{-1} A_{m\ell} \right) \\ &= |\mathcal{R}T|^{-1} \Bigg(A_{jj}^{-1} 1_{\{j \geqslant -\ell_T\}} - \sum_{\ell=-\log_2 T}^{-\ell_T-1} A_{j\ell}^{-1} \delta_{j\ell} \\ &\qquad + \sum_{m,\ell=-\log_2 T}^{-\ell_T-1} A_{j\ell}^{-1} A_{jm}^{-1} A_{\ell m} \Bigg),\end{aligned}$$

and, using that the last sums over m, ℓ contain $\log_2(T) - \ell_T = O(\log_2 |\mathcal{R}|)$ elements,

$$\begin{aligned}\mathrm{I}_T &\leqslant |\mathcal{R}T|^{-1} \left[A_{jj}^{-1} 1_{\{j \geqslant -\ell_T\}} + A_{jj}^{-1} 1_{\{j < -\ell_T\}} + 2^{j+1} \log(|\mathcal{R}|) \nu^2 \right] \\ &\leqslant 2^{j+2} \nu |\mathcal{R}T|^{-1}.\end{aligned}$$

In order to compute a bound for the double product II_T, we exploit the compact support of $\psi_{\ell k}(s)$ implying that $k \leqslant s + N_m \leqslant [r_2 T] + N_m$, and then

$$\begin{aligned}\mathrm{II}_T = 2|\mathcal{R}T|^{-2} \sum_{s,t\in\mathcal{R}T} \sum_{\ell=-\ell_T}^{-1} \sum_{m=-\log_2 T}^{-\ell_T-1} A_{j\ell}^{-1} A_{jm}^{-1} \Psi_\ell(s-t) \times \\ \times \sum_{k=[r_2 T]+1}^{[r_2 T]+N_m} \psi_{mk}(s) \psi_{mk}(t).\end{aligned}$$

With $u := s - t$,

$$\begin{aligned}\mathrm{II}_T = 2|\mathcal{R}T|^{-2} \sum_{\ell=-\ell_T}^{-1} \sum_{m=-\log_2 T}^{-\ell_T-1} A^{-1}_{j\ell} A^{-1}_{jm} \sum_{u=0}^{|\mathcal{R}T|} \Psi_\ell(u) \times \\ \times \sum_{s=u+[r_1T]}^{[r_2T]} \sum_{k=[r_2T]+1}^{[r_2T]+N_m} \psi_{mk}(s)\psi_{mk}(s-u) \\ + 2|\mathcal{R}T|^{-2} \sum_{\ell=-\ell_T}^{-1} \sum_{m=-\log_2 T}^{-\ell_T-1} A^{-1}_{j\ell} A^{-1}_{jm} \sum_{u=-|\mathcal{R}T|}^{-1} \Psi_\ell(u) \times \\ \times \sum_{s=[r_1T]}^{u+[r_2T]} \sum_{k=[r_2T]+1}^{[r_2T]+N_m} \psi_{mk}(s)\psi_{mk}(s-u)\end{aligned}$$

and, applying the Cauchy-Schwarz inequality for the sum over s, we finally get

$$\begin{aligned}\mathrm{II}_T &\leqslant 4|\mathcal{R}T|^{-2} \left(\sum_{m=-\ell_T}^{-1} N_m |A^{-1}_{jm}| \right) \left(\sum_{\ell=-\log_2 T}^{-\ell_T-1} N_\ell |A^{-1}_{j\ell}| \right) \\ &\leqslant 2(6+2\sqrt{2})|\mathcal{R}|^{-2} N_{-1}^2 \nu^2 2^j T^{-1}\end{aligned}$$

using (3.20). Similar calculations lead to

$$\mathrm{III}_T \leqslant (2+\sqrt{2})\nu N_{-1}^2 |\mathcal{R}|^{-2} 2^j T^{-1}.$$

Putting these bounds together gives the upper bound of $\|U_{j,\mathcal{R};T}\|_2^2$.

On the other hand, from (4.22), we can write

$$\|U_{j,\mathcal{R};T}\|_2^2 \geqslant |\mathcal{R}T|^{-2} \sum_{s\in\mathcal{R}T} \left(\sum_{\ell=-\log_2 T}^{-1} A^{-1}_{j\ell} \sum_{k\in\mathcal{R}T} \psi^2_{\ell k}(s) \right)^2 .$$

If we split the sum over ℓ at point ℓ_T such that $|N_\ell| \leqslant |\mathcal{R}T|$ for all $\ell = -1, -2, \ldots, \ell_T$, then $|\ell_T| = O(\log_2 |\mathcal{R}T|)$ and, as $|k-s| \leqslant |\mathcal{R}T|$ and by definition of ℓ_T, $\sum_{k\in\mathcal{R}T} \psi^2_{\ell k}(s) = 1$ in the first term of the parenthesis, and we obtain

$$\|U_{j,\mathcal{R};T}\|_2^2 \geqslant |\mathcal{R}T|^{-2} \sum_{s\in\mathcal{R}T} \left(\sum_{\ell=\ell_T}^{-1} A^{-1}_{j\ell} + \mathrm{Rest}_T \right)^2 .$$

with $|\operatorname{Rest}_T| \leqslant \sum_{\ell=-\log_2 T}^{\ell_T-1} |A_{j\ell}^{-1}| = O(\nu \cdot 2^{j/2}|\mathcal{R}T|^{-1/2}|)$, where the rate follows using the same techniques to prove (3.19), except that here the sum over ℓ goes from $-\log_2 T$ to $\ell_T - 1$ with $\ell_T = O(\log_2 |\mathcal{R}T|)$. On the other hand, (3.18) implies that $\sum_{\ell=\ell_T}^{-1} A_{j\ell}^{-1} = 2^j + O(|\mathcal{R}T|^{-1})$, and we get the result. □

Lemma 4.3. *Under Assumption 4.1, equation (4.4),*

$$\|\Sigma_T\|_{\text{spec}} = \|\Sigma_T^{1/2}\|_{\text{spec}}^2 \leqslant \|c_X\|_{1,\infty} < \infty.$$

On the other hand, under Assumption 4.2, $\|\Sigma_T^{-1}\|_{\text{spec}}$ is uniformly bounded in T.

Proof. For $x \in \mathbb{C}^T$ with $\|x\|_2 = 1$ and if $\sigma_{s,t}$ denotes the element (s,t) of the matrix Σ_T,

$$\begin{aligned}\|\Sigma^{1/2}x\|_2^2 = \sum_{s,t=0}^{T-1} x_s\overline{x}_t\sigma_{s,t} &\leqslant \sum_{s=0}^{T-1}\sum_{u=-(T-1)}^{T-1} |x_s\overline{x}_{s+u}\sigma_{s,s+u}| \\ &\leqslant \sum_{u=-(T-1)}^{T-1} \sup_{s=0,\ldots,T-1} |\sigma_{s,s+u}|\end{aligned}$$

which gives the first result using (4.4). The second result is proved in Lemma 6.3 below. □

We can now prove Proposition 4.1.

Expectation

$$\begin{aligned}\mathrm{E}Q_{j,\mathcal{R};T} &= |\mathcal{R}T|^{-1} \sum_{k\in\mathcal{R}T} \sum_{\ell=-\log_2 T}^{-1} A_{j\ell}^{-1} \sum_{s,t=0}^{T-1} \psi_{\ell k}(s)\psi_{\ell k}(t) \times \\ &\qquad \times \sum_{m=-\infty}^{-1}\sum_{n=-\infty}^{\infty} w_{mn;T}^2 \psi_{mn}(s)\psi_{mn}(t) \\ &= |\mathcal{R}T|^{-1} \sum_{k\in\mathcal{R}T} \sum_{\ell=-\log_2 T}^{-1} A_{j\ell}^{-1} \times \\ &\qquad \times \sum_{m=-\infty}^{-1}\sum_{n=-\infty}^{\infty} w_{mn;T}^2 \left(\sum_{s=0}^{T-1} \psi_{\ell k}(s)\psi_{mn}(s)\right)^2\end{aligned}$$

defining $u := n - k$,

$$\mathrm{E}Q_{j,\mathcal{R};T} = |\mathcal{R}T|^{-1} \sum_{k\in\mathcal{R}T} \sum_{m=-\infty}^{-1} \sum_{u=-\infty}^{\infty} w^2_{m,u+k,T} \times \\ \times \sum_{\ell=-\log_2 T}^{-1} A^{-1}_{j\ell} \left(\sum_{s=-\infty}^{\infty} \psi_{\ell k}(s)\psi_{m,u+k}(s) \right)^2 .$$

By Definition 3.1, we can write $w^2_{m,u+k,T} = S_m(k/T) + R_T(m,u,k)$ with

$$|R_T(m,u,k)| \leqslant \left| S_m\left(\frac{u+k}{T}\right) - S_m\left(\frac{k}{T}\right) \right| + \frac{C_m}{T}$$

which leads to

$$\mathrm{E}Q_{j,\mathcal{R};T} = |\mathcal{R}T|^{-1} \sum_{k\in\mathcal{R}T} \sum_{m=-\infty}^{-1} S_m\left(\frac{k}{T}\right) \sum_{\ell=-\log_2 T}^{-1} A^{-1}_{j\ell} \times \\ \times \sum_{u=-\infty}^{\infty} \left(\sum_{s=-\infty}^{\infty} \psi_{\ell k}(s)\psi_{m,u+k}(s) \right)^2 + \mathrm{Rest}_T$$

By construction of the matrix A_T, we observe that

$$A_{\ell m} = \sum_{u=-\infty}^{\infty} \left(\sum_{s=-\infty}^{\infty} \psi_{\ell k}(s)\psi_{m,u+k}(s) \right)^2 \tag{4.23}$$

which implies with Assumption 4.4

$$\begin{aligned} \mathrm{E}Q_{j,\mathcal{R};T} &= |\mathcal{R}T|^{-1} \sum_{k\in\mathcal{R}T} S_j\left(\frac{k}{T}\right) + \mathrm{Rest}_T \\ &= |\mathcal{R}|^{-1} \int_{\mathcal{R}} dz\; S_j(z) + O\left(|\mathcal{R}T|^{-1} L_j\right) + \mathrm{Rest}_T \end{aligned} \tag{4.24}$$

where the last equality is a standard result on the Total Variation norm [14, Lemma P5.1].

We now bound $|\,\mathrm{Rest}_T\,|$. As s goes from $-\infty$ to ∞, we have

$$|\,\mathrm{Rest}_T\,| \leqslant |\mathcal{R}T|^{-1} \sum_{m=-\infty}^{-1} \sum_{\ell=-\log_2 T}^{-1} |A^{-1}_{j\ell}| \sum_{u=-\infty}^{\infty} \sum_{k\in\mathcal{R}T} \\ \left[\left| S_m\left(\frac{u+k}{T}\right) - S_m\left(\frac{k}{T}\right) \right| + \frac{C_m}{T} \right] \left(\sum_{s=-\infty}^{\infty} \psi_{\ell 0}(s)\psi_{mu}(s) \right)^2 .$$

Using (3.28) for the sum over k, $|\operatorname{Rest}_T|$ is bounded by

$$\sum_{m=-\infty}^{-1}\sum_{u=-\infty}^{\infty}\left[\frac{|u|\operatorname{TV}(S_m)}{|\mathcal{R}T|}+\frac{C_m}{T}\right]\times \sum_{\ell=-\log_2 T}^{-1}|A^{-1}_{j\ell}|\left(\sum_{s=-\infty}^{\infty}\psi_{\ell 0}(s)\psi_{mu}(s)\right)^2$$

In this last expression, the compact support of $\psi_{\ell 0}$ and ψ_{mu} implies that $|u|\leqslant N_\ell\vee N_m$. Together with (3.21) and (4.23), we get

$$\begin{aligned}
&|\operatorname{Rest}_T|\\
&\leqslant|\mathcal{R}T|^{-1}\sum_{m=-\infty}^{-1}\sum_{\ell=-\log_2 T}^{-1}\{\operatorname{TV}(S_m)(N_\ell\vee N_m)+C_m\}\,|A^{-1}_{j\ell}|A_{\ell m}\\
&\leqslant|\mathcal{R}T|^{-1}\sum_{m=-\infty}^{-1}\sum_{\ell=-\log_2 T}^{-1}|A^{-1}_{j\ell}|\Big\{\operatorname{TV}(S_m)N_\ell(2N_m-1)\\
&\qquad\qquad+\operatorname{TV}(S_m)N_m(2N_\ell-1)+C_m(2N_m-1)\Big\}\\
&=(2+\sqrt{2})\nu 2^{j/2}|\mathcal{R}T|^{-1}\sqrt{T}\sum_{m=-\infty}^{-1}(2N_m-1)\operatorname{TV}(S_m)\\
&\qquad\qquad+O\left(2^{j/2}|\mathcal{R}T|^{-1}\right)\qquad(4.25)
\end{aligned}$$

using (3.20) and (3.25).

Variance

Using Lemma 4.1 with (A.5), Lemma 4.2 and Lemma 4.3, we get the upper bound as follows

$$\begin{aligned}
\operatorname{Var}Q_{j,\mathcal{R};T}&=2\|M_{j,\mathcal{R};T}\|_2^2\leqslant 2\|\Sigma_T^{1/2}\|_{\text{spec}}^4\|U_{j,\mathcal{R};T}\|_2^2\\
&\leqslant\|c_X\|_{1,\infty}^2\,2^j\,|\mathcal{R}T|^{-1}\qquad(4.26)
\end{aligned}$$

To obtain the lower bound, we make use of (A.5) two times on $U_{j,\mathcal{R};T}=\Sigma_T^{-1/2}M_{j,\mathcal{R};T}\Sigma_T^{-1/2}$:

$$\operatorname{Var}Q_{j,\mathcal{R};T}=2\|M_{j,\mathcal{R};T}\|_2^2\geqslant 2\|\Sigma_T^{-1/2}\|_{\text{spec}}^{-4}\|U_{j,\mathcal{R};T}\|_2^2.\qquad(4.27)$$

Since Σ_T is a symmetric matrix, $\|\Sigma_T^{-1/2}\|_{\text{spec}}^2 = \|\Sigma_T^{-1}\|_{\text{spec}}$ and Lemma 4.3 shows that $\|\Sigma_T^{-1}\|$ is uniformly bounded under Assumption 4.2. A lower bound for $\|U_{j,\mathcal{R};T}\|_2^2$ is stated in Lemma 4.2 and the result follows. □

4.4.2 Proof of Proposition 4.2 and its consequences

Our proof of Proposition 4.2 needs the use of an exponential bound for quadratic forms of Gaussian random variables. For sake of presentation, we recall now this result and refer to Dahlhaus and Polonik [27, Proposition 6.1].

Proposition 4.7. *Let $\underline{Z}_n = (Z_1, \ldots, Z_n)'$ be a vector of iid Gaussian random variables with zero mean and $\operatorname{Var} Z_1 = 1$. If M_n is an $n \times n$ matrix such that $\|M_n\|_{\text{spec}} \leqslant \tau_\infty$ and $\sigma_n^2 = 2\|M_n\|_2^2$, then for all $\lambda > 0$*

$$\Pr\left((\underline{Z}_n' M_n \underline{Z}_n - \operatorname{tr} M_n) > \sigma_n \lambda\right) \leqslant \exp\left(-\frac{1}{2} \cdot \frac{\lambda^2}{1 + 2\lambda \frac{\tau_\infty}{\sigma_n}}\right).$$

As in the proof of Proposition 4.1, equation (4.21), we write $Q_{j,\mathcal{R};T}$ as a quadratic form of Gaussian variables in order to apply Proposition 4.7 with

$$M_{j,\mathcal{R};T} = \Sigma_T'^{1/2} U_{j,\mathcal{R};T} \Sigma_T^{1/2} \tag{4.28}$$

to prove the assertion.

Proof of Proposition 4.2. Lemma 4.2 and 4.3 imply with (A.2) and (A.4):

$$\|M_{j,\mathcal{R};T}\|_{\text{spec}} \leqslant 2^{j/2} \nu \|c_X\|_{1,\infty} |\mathcal{R}|^{-1} T^{-1/2} \tag{4.29}$$

which, using Proposition 4.7, implies

$$\begin{aligned}
&\Pr\left((Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}) > \eta \sigma_{j,\mathcal{R},T}\right) \\
&\quad \leqslant \Pr\left((Q_{j,\mathcal{R};T} - \mathrm{E} Q_{j,\mathcal{R};T}) > \eta \sigma_{j,\mathcal{R},T}/2\right) \\
&\qquad\qquad + \Pr\left((\mathrm{E} Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}) > \eta \sigma_{j,\mathcal{R},T}/2\right) \\
&\quad \leqslant \exp\left(-\frac{1}{8} \cdot \frac{\eta^2}{1 + \eta \frac{2^{j/2} \nu \|c_X\|_{1,\infty}}{|\mathcal{R}| T^{1/2} \sigma_{j,\mathcal{R},T}}}\right) \\
&\qquad\qquad + \Pr\left((\mathrm{E} Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}) \geqslant \eta \sigma_{j,\mathcal{R},T}/2\right).
\end{aligned}$$

To bound the second probability, we observe that

$$|\mathrm{E}Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}| \leqslant |\mathcal{R}T|^{-1} \left\{ L_j + 2(2+\sqrt{2})\rho\nu 2^{j/2}\sqrt{T} \right\}$$

is obtained using (4.24) and (4.25). This implies

$$\begin{aligned} &\Pr\left((Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}) \geqslant \eta\sigma_{j,\mathcal{R},T}\right) \\ &\qquad \leqslant \exp\left(-\frac{1}{8} \cdot \frac{\eta^2 \sigma_{j,\mathcal{R},T}}{\sigma_{j,\mathcal{R},T} + \eta \frac{2^{j/2}\nu \|c_X\|_{1,\infty}}{|\mathcal{R}|T^{1/2}}} \right) \\ &\qquad + \exp\left(1 - \frac{1}{2\eta} \frac{\eta^2 \sigma_{j,\mathcal{R},T}}{\frac{L_j + 2(2+\sqrt{2})\rho\nu 2^{j/2}\sqrt{T})}{|\mathcal{R}T|}} \right) \end{aligned}$$

and the result follows. □

Proof of Corollary 4.1. In the following proof, K denotes a generic constant and k_T is an increasing function of T. By Proposition 4.1, $\sigma^2_{j,\mathcal{R},T} := \operatorname{Var} Q_{j,\mathcal{R};T} \leqslant 2^j K_2 |\mathcal{R}|^{-2} T^{-1}$ uniformly in j, which implies

$$\begin{aligned} &\Pr\left(\sup_{-J_T \leqslant j < 0} |Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}| \geqslant k_T \sqrt{K_2 |\mathcal{R}|^{-2} T^{-1}} \right) \\ &\quad \leqslant \sum_{j=-J_T}^{-1} \Pr\left(|Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}| \geqslant k_T \sqrt{K_2 |\mathcal{R}|^{-2} T^{-1}} \right) \\ &\quad \leqslant \sum_{j=-J_T}^{-1} \Pr\left(|Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}| \geqslant 2^{-j/2} k_T \sigma_{j,\mathcal{R},T} \right). \end{aligned}$$

Using Proposition 4.2, this probability is bounded by

$$c_0 \, J_T \max_{-J_T \leqslant j < 0} \exp\left(-\frac{1}{8} \cdot \frac{2^{-j} k_T^2}{1 + \frac{2^{-j/2} k_T L_j}{|\mathcal{R}T| \sigma_{j,\mathcal{R};T}} + \frac{k_T \nu (\|c_X\|_{1,\infty} + c_1 \rho)}{|\mathcal{R}| \sqrt{T} \sigma_{j,\mathcal{R};T}}} \right)$$

Proposition 4.1 shows that, $\sigma_{j,\mathcal{R},T} \geqslant 2^j \sqrt{K_1 |\mathcal{R}T|^{-1}}$ for T sufficiently large. This leads to the bound

$$c_0 \, J_T \max_{-J_T \leqslant j < 0} \exp\left(-\frac{1}{8} \cdot \frac{k_T^2}{2^j + \frac{2^{-j/2} L_j k_T}{\sqrt{|\mathcal{R}T| K_1}} + \frac{k_T \nu (\|c_X\|_{1,\infty} + c_1 \rho)}{\sqrt{K_1 |\mathcal{R}|}}} \right)$$

By equation (3.25), there exists a positive constant ρ' such that $L_j \leqslant 2^{j/2}\rho'$. Then, asymptotically, the rate of convergence of the dominant terms in this expression are given by $J_T \cdot \exp(-k_T)$ on k_T. □

4.4.3 Proof of Proposition 4.3

In the following proof K is a generic constant.

Lemma 4.4. *If*

$$U_{ts}^{(j)} = |\mathcal{R}T|^{-1} \sum_{\ell=-\log_2 T}^{-1} A_{j\ell}^{-1} \sum_{k\in\mathcal{R}T} \psi_{\ell k}(s)\psi_{\ell k}(t) \ ,$$

then

$$\sum_{t=-\infty}^{\infty} \sum_{s,u=-\infty}^{\infty} U_{ts}^{(j)} U_{tu}^{(j)} 1_{|s-u|\leqslant q_T} \leqslant |\mathcal{R}|^{-2} q_T T^{-1} 2N_0 \nu^2 2^j \log_2^2 T$$
$$= O\left(2^j \frac{q_T \log_2^2 T}{T}\right)$$

Proof. Direct calculations yields

$$\sum_{t=-\infty}^{\infty} \sum_{s,u=-\infty}^{\infty} U_{ts}^{(j)} U_{tu}^{(j)} 1_{|s-u|\leqslant q_T}$$
$$\leqslant |\mathcal{R}T|^{-2} \sum_{\ell,m=-\log_2 T}^{-1} |A_{j\ell}^{-1}||A_{jm}^{-1}| \sum_{s,u=-\infty}^{\infty} 1_{|s-u|\leqslant q_T} \times$$
$$\times \sum_{t=-\infty}^{\infty} \left(\sum_{k\in\mathcal{R}T} |\psi_{\ell k}(s)\psi_{\ell k}(t)|\right) \left(\sum_{n\in\mathcal{R}T} |\psi_{mn}(u)\psi_{mn}(t)|\right) .$$

Using the Cauchy-Schwarz inequality for the sum over t, we get a product between two terms similar to

$$\sqrt{\sum_t \left(\sum_{k\in\mathcal{R}T} \psi_{\ell k}(s)\psi_{\ell k}(t)\right)^2} = \sqrt{\sum_{k,r\in\mathcal{R}T} \Psi_\ell(k-r)\psi_{\ell k}(s)\psi_{\ell r}(s)}$$

if $k = r+u$, then the range of u is included in $\{-|\mathcal{R}T|, \dots, 0, \dots, |\mathcal{R}T|\}$:

$$\begin{aligned}
&\leqslant \left\{ \sum_{u=-|\mathcal{R}T|}^{|\mathcal{R}T|} \sum_{r \in \mathcal{R}T} |\Psi_\ell(u)| \cdot |\psi_{\ell,r+u}(s)\psi_{\ell r}(s)| \right\}^{1/2} \\
&\leqslant \left\{ \sum_{u=-|\mathcal{R}T|}^{|\mathcal{R}T|} |\Psi_\ell(u)| \right\}^{1/2}
\end{aligned}$$

using Cauchy-Schwarz inequality for the sum over r and $\sum_r \psi_{\ell r}^2(s) = 1$. Finally, using the basic properties of $\Psi_j(\tau)$ described in Section 3.3.2, we get

$$\left\{ \sum_t \left(\sum_{k \in \mathcal{R}T} \psi_{\ell k}(s)\psi_{\ell k}(t) \right)^2 \right\}^{1/2} \leqslant \sqrt{2N_\ell - 1}. \tag{4.30}$$

Then

$$\begin{aligned}
&\sum_{t=-\infty}^{\infty} \sum_{s,u=-\infty}^{\infty} U_{ts}^{(j)} U_{tu}^{(j)} 1_{|s-u| \leqslant q_T} \\
&\qquad \leqslant T q_T |\mathcal{R}T|^{-2} \sum_{\ell,m=-\log_2 T}^{-1} |A_{j\ell}^{-1}||A_{jm}^{-1}| \sqrt{2N_\ell - 1}\sqrt{2N_m - 1}
\end{aligned}$$

and we obtain the result by (3.20). □

In the proof of Proposition 4.3, we need a modification of Corollary 4.1, in which $\mathcal{R}$ is replaced by $\mathcal{R}_T$. The proof of the following result is along the lines of the proof of Corollary 4.1.

Lemma 4.5. *Under Assumptions 4.1 to 4.6, there exists $T_0 \geqslant 1$ such that, for all $T \geqslant T_0$,*

$$\begin{aligned}
&\Pr\left(\sup_{-J_T \leqslant j < 0} |Q_{j,\mathcal{R}_T(s);T} - Q_{j,\mathcal{R}_T(s)}| \geqslant k_T \sqrt{K_2 |\mathcal{R}_T|^{-2} T^{-1}} \right) \\
&\qquad \leqslant c_0 J_T \exp\left\{ -\frac{1}{8} \cdot \frac{k_T^2}{1 + \frac{\rho N_0 k_T}{\sqrt{K_1 |\mathcal{R}_T T|}} + \frac{\nu k_T (\|c_X\|_{1,\infty} + c_1 \rho)}{\sqrt{K_1 |\mathcal{R}_T|}}} \right\} \\
&\qquad = o_T(1) \,.
\end{aligned}$$

Proof of Proposition 4.3. Define

$$\bar{\sigma}_{s,s+u} := \sum_{\ell=-\log_2 T}^{-1} Q_{\ell,\mathcal{R}_T(s)} \Psi_\ell(u) 1_{|u| \leqslant M_T}$$

the entries of a matrix $\bar{\Sigma}$, and set $\bar{\sigma}^2_{j,\mathcal{R},T} := 2\|U'_{j,\mathcal{R};T}\bar{\Sigma}_T\|_2^2$. Our proof is based on the decomposition

$$\tilde{\sigma}^2_{j,\mathcal{R},T} - \sigma^2_{j,\mathcal{R},T} = \left(\tilde{\sigma}^2_{j,\mathcal{R},T} - \bar{\sigma}^2_{j,\mathcal{R},T}\right) + \left(\bar{\sigma}^2_{j,\mathcal{R},T} - \sigma^2_{j,\mathcal{R},T}\right)$$

where the first term is stochastic while the second term is deterministic.

We will first show that the term $|\bar{\sigma}^2_{j,\mathcal{R},T} - \sigma^2_{j,\mathcal{R},T}|$ is $o(2^{j-J_T}T^{-1})$. Using (A.5), we can write

$$\begin{aligned}
\frac{1}{2}\left(\bar{\sigma}^2_{j,\mathcal{R},T} - \sigma^2_{j,\mathcal{R},T}\right) &= \|U'_{j,\mathcal{R};T}\bar{\Sigma}_T\|_2^2 - \|U'_{j,\mathcal{R};T}\Sigma_T\|_2^2 \\
&\leqslant \|U'_{j,\mathcal{R};T}(\bar{\Sigma}_T - \Sigma_T)\|_2^2 + 2 \cdot \|U'_{j,\mathcal{R};T}\Sigma_T\|_2 \cdot \|U'_{j,\mathcal{R};T}(\bar{\Sigma}_T - \Sigma_T)\|_2 \\
&\leqslant \|U_{j,\mathcal{R};T}\|_2^2 \cdot \|\bar{\Sigma}_T - \Sigma_T\|^2_{\text{spec}} \\
&\qquad + 2 \cdot \|U_{j,\mathcal{R};T}\|_2^2 \cdot \|\Sigma_T\|_{\text{spec}} \cdot \|\bar{\Sigma}_T - \Sigma_T\|_{\text{spec}}
\end{aligned}$$

where we know by Lemmas 4.2 and 4.3 that $\|U_{j,\mathcal{R};T}\|_2^2 = O(2^j T^{-1})$ and $\|\Sigma_T\|_{\text{spec}} \leqslant \|c_X\|_{1,\infty}$. Moreover, we can write:

$$\begin{aligned}
\|\bar{\Sigma}_T - \Sigma_T\|_{\text{spec}} &\leqslant \sum_{u=-\infty}^{\infty} \sup_s (\sigma_{s,s+u} - \bar{\sigma}_{s,s+u}) \\
&= \sum_{u=-\infty}^{\infty} \sup_s \sum_{\ell=-\infty}^{-1} \sum_{n=-\infty}^{\infty} \left(w^2_{\ell n;T} - Q_{\ell,\mathcal{R}_T(s)}\right) \cdot \psi_{\ell n}(s)\psi_{\ell n}(s+u) \\
&\quad + \mathrm{R}_1 + \mathrm{R}_2 \qquad (4.31)
\end{aligned}$$

with

$$\mathrm{R}_1 = \sum_{u=-\infty}^{\infty} \sup_s \sum_{\ell=-\infty}^{-1} Q_{\ell,\mathcal{R}_T(s)} \Psi_\ell(u)\ 1_{|u|>M_T}$$

and

$$\mathrm{R}_2 = \sum_{u=-\infty}^{\infty} \sup_s \sum_{\ell=-\infty}^{-[\log_2(T)]-1} Q_{\ell,\mathcal{R}_T(s)} \Psi_\ell(u)\ 1_{|u|<M_T}.$$

As

$$\sum_{u=-\infty}^{\infty} \sup_s \sum_{\ell=-\infty}^{-1} Q_{\ell,\mathcal{R}_T(s)} \Psi_\ell(u) \\ = \sum_{u=-\infty}^{\infty} \sup_s |\mathcal{R}_T|^{-1} \int_{\mathcal{R}_T(s)} dz \; c_X(z,u) \;,$$

the rate of R_1 is $o_T(2^{-J_T})$ by Assumption 4.6. Next, using $|\Psi_\ell(u)| \leqslant 1$ uniformly in $\ell < 0$, we get

$$|\,\mathrm{R}_2\,| \leqslant \sum_{u=-\infty}^{\infty} \sup_s |\mathcal{R}_T|^{-1} \int_{\mathcal{R}_T(s)} dz \sum_{\ell=-\infty}^{-[\log_2(T)]-1} S_\ell(z)\, 1_{|u|<M_T} \\ \leqslant 2M_T \sum_{\ell=-\infty}^{-[\log_2(T)]-1} \sup_z S_\ell(z) = O(M_T/T)$$

using Assumption 4.4. Assumption 4.5 on the rate of of clipping sequence M_T implies $M_T/T = o_T\left(2^{-J_T}\right)$, and then

$$|\,\mathrm{R}_2\,| = o_T\left(2^{-J_T}\right) \;.$$

The main term of (4.31) is bounded by

$$\sum_{u=-\infty}^{\infty} \sup_s \sum_{\ell=-\log_2 T}^{-1} \sum_{n=-\infty}^{\infty} \\ |\mathcal{R}_T|^{-1} \int_{\mathcal{R}_T(s)} dz \; |w^2_{\ell n;T} - S_\ell(z)| \cdot |\psi_{\ell n}(s)\psi_{\ell n}(s+u)|. \quad (4.32)$$

By Definition 3.1, we can write

$$|w^2_{\ell n;T} - S_\ell(z)| \leqslant \frac{C_\ell}{T} + \left|S_\ell\left(\frac{n}{T}\right) - S_\ell\left(\frac{n-s}{T} + z\right)\right| \\ + \left|S_\ell(z) - S_\ell\left(\frac{n-s}{T} + z\right)\right|$$

which, when replaced in (4.32), leads to three terms. By (3.29) and (3.25), the first term is $O(T^{-1})$. For the second term, with a change of

variable z to $z + s/T$, we get:

$$\sum_{u=-\infty}^{\infty} \sup_s \sum_{\ell=-\infty}^{-1} \sum_{n=-\infty}^{\infty} |\mathcal{R}_T|^{-1} \int_{\mathcal{R}_T(0)} dz \left| S_\ell\left(\frac{n}{T}\right) - S_\ell\left(\frac{n}{T} + z\right)\right| \times \\ \times |\psi_{\ell n}(s)\psi_{\ell n}(s+u)|,$$

where $\mathcal{R}_T(0)$ denotes the interval $\mathcal{R}_T(s)$ shifted by $-s$. If we use that $|\psi_{\ell n}(s)|$ is uniformly bounded and that $\sum_{u=-\infty}^{\infty} |\psi_{\ell n}(s+u)| = O(N_\ell)$, the second term is bounded (up to a multiplicative constant) by

$$\begin{aligned} &|\mathcal{R}_T|^{-1} \sum_{\ell=-\infty}^{-1} N_\ell \int_{\mathcal{R}_T(0)} dz \sum_{n=-\infty}^{\infty} \left| S_\ell\left(\frac{n}{T}\right) - S_\ell\left(\frac{n}{T} + z\right)\right| \\ &\quad \leqslant |\mathcal{R}_T|^{-1} \sum_{\ell=-\infty}^{-1} N_\ell \int_{\mathcal{R}_T(0)} dz\, |z|\, \mathrm{TV}(S_\ell) \leqslant |\mathcal{R}_T| \sum_{\ell=-\infty}^{-1} N_\ell L_\ell \\ &\quad = O(|\mathcal{R}_T|) \end{aligned}$$

by assumptions (3.24) and (3.25). The third term is

$$\sum_{u=-\infty}^{\infty} \sup_s \sum_{\ell=-\infty}^{-1} \sum_{n=-\infty}^{\infty} |\mathcal{R}_T|^{-1} \int_{\mathcal{R}_T(s)} dz \left| S_\ell(z) - S_\ell\left(\frac{n-s}{T} + z\right)\right| \times \\ \times |\psi_{\ell n}(s)\psi_{\ell n}(s+u)|.$$

If s_0 denotes the infimum of $\mathcal{R}_T(s)$, we decompose the integral as follows:

$$\sum_{u=-\infty}^{\infty} \sup_s \sum_{\ell=-\infty}^{-1} \sum_{n=-\infty}^{\infty} |\mathcal{R}_T|^{-1} \times \\ \sum_{k=0}^{|\mathcal{R}_T T|-1} \int_{s_0+\frac{k}{T}}^{s_0+\frac{k+1}{T}} dz \left| S_\ell(z) - S_\ell\left(\frac{n-s}{T} + z\right)\right| |\psi_{\ell n}(s)\psi_{\ell n}(s+u)|$$

which can be rewritten with the change of variables $y := z - s_0 - k/T$,

$$\sum_{u=-\infty}^{\infty} \sup_s \sum_{\ell=-\infty}^{-1} \sum_{n=-\infty}^{\infty} |\mathcal{R}_T|^{-1} \times$$
$$\times \sum_{k=0}^{|\mathcal{R}_T T|-1} \int_0^{1/T} dy \left| S_\ell \left(y + s_0 + \frac{k}{T} \right) - S_\ell \left(y + s_0 + \frac{n-s+k}{T} \right) \right| \times$$
$$\times |\psi_{\ell n}(s) \psi_{\ell n}(s+u)|.$$

Assumption (3.24) for the sum over k with (3.28) leads to the bound

$$\sum_{u=-\infty}^{\infty} \sup_s \sum_{\ell=-\infty}^{-1} L_\ell \sum_{n=-\infty}^{\infty} |\mathcal{R}_T T|^{-1} |n-s| |\psi_{\ell n}(s) \psi_{\ell n}(s+u)|.$$

The compact support of $\psi_{\ell n}(s)$ implies $|n-s| < N_\ell$. Therefore, (3.29), (3.24) and (3.25) leads to $O(|\mathcal{R}_T T|^{-1})$. Finally, we summarize all the rates that found we found above:

$$\begin{aligned} 2^{-j} T \left(\bar{\sigma}^2_{j,\mathcal{R},T} - \sigma^2_{j,\mathcal{R},T} \right) & \\ &= O\left(T^{-1} + |\mathcal{R}_T| + |\mathcal{R}_T T|^{-1} \right) + |\,\mathrm{R}_1\,| + |\,\mathrm{R}_2\,| \\ &= O\left(T^{-1} + |\mathcal{R}_T| + |\mathcal{R}_T T|^{-1} \right) + o_T(2^{-J_T}) + o_T(2^{-J_T}) \\ &= o_T(2^{-J_T}) \end{aligned}$$

by Assumption 4.5.

Let us now turn to the stochastic term $|\tilde{\sigma}^2_{j,\mathcal{R},T} - \bar{\sigma}^2_{j,\mathcal{R},T}|$. Lemma 4.5 implies the existence of a random set $\mathcal{A}$ which does not depend on j and such that $\Pr(\mathcal{A}) \geqslant 1 - o_T(1)$ and

$$|Q_{j,\mathcal{R}_T(s);T} - Q_{j,\mathcal{R}_T(s)}| \leqslant k_T \sqrt{K_2 |\mathcal{R}_T|^{-2} T^{-1}} \tag{4.33}$$

almost surely on $\mathcal{A}$, for all $T > T_0$ and $j = -1, \ldots, -J_T$. We can write

$$|\bar{\sigma}^2_{j,\mathcal{R},T} - \tilde{\sigma}^2_{j,\mathcal{R},T}| \leqslant 2 \sum_{h,t=0}^{T-1} \Big| \sum_{s,u=0}^{T-1} U^{(j)}_{ts} U^{(j)}_{tu} \times$$
$$\times \sum_{\ell,m=-\log_2 T}^{-1} \left(Q_{\ell,\mathcal{R}_T(s);T} Q_{m,\mathcal{R}_T(s);T} - Q_{\ell,\mathcal{R}_T(s)} Q_{m,\mathcal{R}_T(s)} \right) \times$$
$$\times \Psi_\ell(s-h) \Psi_m(u-h) \Big| \cdot 1_{|s-h| \leqslant M_T} 1_{|u-h| \leqslant M_T} \tag{4.34}$$

almost surely on $\mathcal{A}$. Using the decomposition

$$\begin{aligned} Q_{\ell,\mathcal{R}_T(s);T}Q_{m,\mathcal{R}_T(s);T} - Q_{\ell,\mathcal{R}_T(s)}Q_{m,\mathcal{R}_T(s)} \\ = \left(Q_{m,\mathcal{R}_T(s);T} - Q_{m,\mathcal{R}_T(s)}\right) Q_{\ell,\mathcal{R}_T(s)} \\ + \left(Q_{\ell,\mathcal{R}_T(s);T} - Q_{\ell,\mathcal{R}_T(s)}\right) Q_{m,\mathcal{R}_T(s);T}, \end{aligned}$$

the first term of the right hand side of (4.34) is split into two terms. On $\mathcal{A}$, the first of these two terms is bounded as follows (the other term is bounded similarly):

$$\begin{aligned} 2\sum_{h,t=0}^{T-1}\sum_{s,u=0}^{T-1}\Big|U_{ts}^{(j)}U_{tu}^{(j)}\sum_{m=-\log_2 T}^{-1}\left(Q_{m,\mathcal{R}_T(s);T} - Q_{m,\mathcal{R}_T(s)}\right)\Psi_m(u-h) \\ \times\sum_{\ell=-M_T}^{-1}Q_{\ell,\mathcal{R}}\Psi_\ell(s-h)\Big|1_{|s-u|\leqslant 2M_T} \\ \leqslant 2k_T\log_2(T)\sqrt{K_2|\mathcal{R}_T T|^{-1}}\sum_{h,t=0}^{T-1}\sum_{s,u=0}^{T-1}|U_{ts}^{(j)}U_{tu}^{(j)}|\times \\ \times\Big|\sum_{\ell=-M_T}^{-1}Q_{\ell,\mathcal{R}_T(s)}\Psi_\ell(s-h)\Big|1_{|s-u|\leqslant 2M_T} \\ \leqslant 2k_T\log_2(T)\sqrt{K_2|\mathcal{R}_T T|^{-1}}\sum_{t=0}^{T-1}\sum_{s,u=0}^{T-1}|U_{ts}^{(j)}U_{tu}^{(j)}|1_{|s-u|\leqslant 2M_T}\times \\ \times\sum_{h=-\infty}^{\infty}\sup_z\Big|\sum_{\ell=-\log_2 T}^{-1}S_\ell(z)\Psi_\ell(h)\Big| \\ = O\left(2^j M_T k_T|\mathcal{R}_T T|^{-1/2}\,T^{-1}\log_2^3 T\right) \quad \text{a.s. on } \mathcal{A} \end{aligned}$$

using Assumption 4.1 (Equation (4.5)) and Lemma 4.4. The result follows from Assumption 4.5. □

4.4.4 Proof of Theorem 4.1

By Lemma 4.5 and Proposition 4.3 and for T large enough, there exists of a random set $\mathcal{A}$ such that $1-\Pr(\mathcal{A}) = o_T(1)$ and (4.10) holds on $\mathcal{A}$.

Then, if $\mathcal{A}^c$ denotes the complementary random set of $\mathcal{A}$, we can write:

$$\begin{aligned}\Pr\left(|Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}| > \tilde{\sigma}_{j,\mathcal{R},T}\eta\right) \\ = \Pr\left(|Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}| > \tilde{\sigma}_{j,\mathcal{R},T}\eta | \mathcal{A}\right) \Pr\left(\mathcal{A}\right) \\ + \Pr\left(|Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}| > \tilde{\sigma}_{j,\mathcal{R},T}\eta | \mathcal{A}^c\right) \left(1 - \Pr\left(\mathcal{A}\right)\right).\end{aligned}$$

The second term of this sum is bounded using Lemma 4.5. To bound the first term, we observe that Proposition 4.3 implies $\tilde{\sigma}^2_{j,\mathcal{R},T} \geqslant \sigma^2_{j,\mathcal{R},T} - \varphi_T$ on $\mathcal{A}$ with $\varphi_T = o_T(2^{j-J_T}T^{-1})$. Together with Proposition 4.1, this implies

$$\frac{\tilde{\sigma}^2_{j,\mathcal{R},T}}{\sigma^2_{j,\mathcal{R},T}} \geqslant 1 - \frac{\varphi_T}{\sigma^2_{j,\mathcal{R},T}} = 1 - o_T(1) \to 1 \tag{4.35}$$

for all $j = -1, \ldots, -J_T$, as T tends to infinity. Then, we can write:

$$\begin{aligned}\Pr\left(|Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}| > \tilde{\sigma}_{j,\mathcal{R},T}\eta\right) \\ \leqslant \Pr\left(|Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}| > \sigma_{j,\mathcal{R},T}\eta\sqrt{1 - \frac{\varphi_T}{\sigma^2_{j,\mathcal{R},T}}} \,\middle|\, \mathcal{A}\right) \\ + o_T(1).\end{aligned}$$

and Proposition 4.2 leads to the result. □

4.4.5 Proof of Proposition 4.4

We first prove the following lemma, stating an exponential inequality for quadratic forms of Gaussian random variables. This result is a generalisation of a similar result obtained by Laurent and Massart [57] for chi-squared distributions, and is proved in the spirit of Spokoiny [106].

Lemma 4.6. *Let $\underline{Z}_T = (Z_1, \ldots, Z_T)'$ be a vector of iid Gaussian random variables with zero mean and $\operatorname{Var} Z_1 = 1$. If M_T is a $T \times T$ symmetric and positive definite matrix, then*

$$\Pr\left(\underline{Z}'_T M_T \underline{Z}_T \leqslant \eta\right) \leqslant \exp\left(-\frac{(\eta - \operatorname{tr} M_T)^2}{4\|M_T\|_2^2}\right).$$

provided that $\eta \leqslant \operatorname{tr} M_T$.

Proof. By assumption on the matrix M_T, the decomposition $M_T = O_T' \Lambda_T O_T$ holds with a diagonal $T \times T$ matrix Λ_T and an orthonormal matrix O_T. If we denote $\underline{Y}_T = U_T' \underline{Z}_T$, then $\underline{Y}_T$ is a vector of iid Gaussian random variables with zero mean and $\operatorname{Var} Y_1 = 1$. We can write $\underline{Z}_T' M_T \underline{Z}_T = \underline{Y}_T' \Lambda_T \underline{Y}_T = \sum_{i=1}^T \lambda_i Y_i^2$ with $\lambda_i > 0$. Moreover, $\operatorname{tr} M_T = \operatorname{tr} \Lambda_T$, $\operatorname{tr} \Lambda_T^2 = \operatorname{tr} M_T^2 = \|M_T\|_2^2$ and $\|M_T\|_{\text{spec}} = \max\{\lambda_1, \dots, \lambda_T\}$. A Chernoff bound on $\underline{Y}_T$ leads to

$$\begin{aligned}
\Pr\left(\underline{Z}_T' M_T \underline{Z}_T \leqslant \eta\right) &= \Pr\left(\underline{Y}_T' \Lambda_T \underline{Y}_T \leqslant \eta\right) \\
&\leqslant \exp\left[\inf_{t<0} \left(-t\eta + \log \mathrm{E} \exp(t \underline{Y}_T' \Lambda_T \underline{Y}_T)\right)\right] \\
&= \exp\left[\inf_{t<0} \left(-t\eta + \sum_{i=1}^T \log \mathrm{E} \exp(\lambda_i t Y_i^2)\right)\right]
\end{aligned}$$

and, using that

$$\log \mathrm{E} \exp(\alpha_i Y_i^2) = -\frac{1}{2} \log(1 - 2\alpha_i) \leqslant \alpha_i + \alpha_i^2$$

holds for $\alpha_i \leqslant 0$, we get

$$\Pr\left(\underline{Z}_T' M_T \underline{Z}_T \leqslant \eta\right) \leqslant \exp\left[\inf_{t<0} \left(-t\eta + t \operatorname{tr} \Lambda_T + t^2 \operatorname{tr} \Lambda_T^2\right)\right].$$

The result follows by taking $t = (\eta - \operatorname{tr} \Lambda_T)/(2 \operatorname{tr} \Lambda_T^2)$. □

Lemma 4.6 is not directly applicable on the quadratic form $Q_{j,\mathcal{R};T} = \underline{Z}_T' M_{j,\mathcal{R};T} \underline{Z}_T$ because the matrix $M_{j,\mathcal{R};T}$ is not definite positive in general. In the next lemma, we show how this matrix can be approximated by the matrix $M_{j,\mathcal{R};T}^\star$, defined as

$$M_{j,\mathcal{R};T}^\star = \Sigma_T^{1/2\,\prime} U_{j,\mathcal{R};T}^\star \Sigma_T^{1/2},$$

where the entry (s,t) of the matrix $U_{j,\mathcal{R};T}^\star$ is

$$u_{st}^\star = \gamma_0 |\mathcal{R}T|^{-1} \sum_{\ell=-\log_2 T}^{-1} 2^{\ell/2} \Psi_\ell(s-t),$$

with $\gamma_0 \geqslant \sup_{j<0} \sup_{\ell<0} 2^{-\ell/2} |A_{j\ell}^{-1}| > 0$. The matrix $M^{\star}_{j,\mathcal{R};T}$ is clearly symmetric. It is also positive definite because $U^{\star}_{j,\mathcal{R};T}$ is positive definite: For all sequences $\underline{x} = (x_1, \ldots, x_T)'$ of ℓ^2, the quadratic form

$$x' U^{\star}_{j,\mathcal{R};T} x = \gamma_0 |\mathcal{R}T|^{-1} \sum_{\ell=-\log_2 T}^{-1} 2^{\ell/2} \sum_{s=0}^{T-1} \left(\sum_{k \in \mathcal{R}T} x_s \psi_{\ell k}(s) \right)^2$$

is strictly positive.

Lemma 4.7. *Assume Assumptions 4.1 to 4.3 and Assumption 4.4 hold true. Define γ_1 such that*

$$0 < \gamma_1 < \gamma_0 \inf_{m<0} \sum_{\ell=-\log_2 T}^{-1} 2^{\ell/2} A_{m\ell}.$$

The following properties hold true for T sufficiently large:

$$\gamma_1 |\mathcal{R}|^{-1} \varepsilon \leqslant \operatorname{tr}(M^{\star}_{j,\mathcal{R};T} - M_{j,\mathcal{R};T}) \leqslant 2 \|c_{X,T}\|_{1,\infty} \gamma_0 |\mathcal{R}|^{-1} \tag{4.36}$$

where ε is defined in Assumption 4.2,

$$\begin{aligned} \|M^{\star}_{j,\mathcal{R};T} - M_{j,\mathcal{R};T}\|^2_{\text{spec}} &\leqslant \|M^{\star}_{j,\mathcal{R};T} - M_{j,\mathcal{R};T}\|_2^2 \\ &\leqslant 4 N_{-1} \gamma_0^2 |\mathcal{R}|^{-2} \|c_X\|^2_{1,\infty} T^{-1} \log_2^2(T) + O(T^{-1}), \end{aligned} \tag{4.37}$$

and, if $\underline{Z}_T = (Z_1, \ldots, Z_T)'$ is a vector of iid Gaussian random variables with zero mean and $\operatorname{Var} Z_1 = 1$, then

$$\begin{aligned} &\Pr\left(\underline{Z}'_T (M^{\star}_{j,\mathcal{R};T} - M_{j,\mathcal{R};T}) \underline{Z}_T > \lambda_T \right) \\ &\qquad = O\left(\exp\left[- \frac{\sqrt{T} \operatorname{tr} M_{j,\mathcal{R};T}}{\log_2^2 T} \right] \right) \end{aligned} \tag{4.38}$$

where $\lambda_T = \operatorname{tr} M^{\star}_{j,\mathcal{R};T} - \operatorname{tr} M_{j,\mathcal{R};T} + \operatorname{tr} M_{j,\mathcal{R};T} \log_2^{-1} T$.

Proof. 1. We prove (4.36). Write $\operatorname{tr}(M^{\star}_{j,\mathcal{R};T} - M_{j,\mathcal{R};T}) = \operatorname{tr}(M^{\star}_{j,\mathcal{R};T}) - \operatorname{tr}(M_{j,\mathcal{R};T})$, where, from Lemma 4.1 and Proposition 4.1, the sec-

ond term is $\mathrm{E}(\underline{Z}_T' M_{j,\mathcal{R};T} \underline{Z}_T) = Q_{j,\mathcal{R}} + O(2^{j/2} T^{-1/2})$. Moreover,

$$\begin{aligned}
\mathrm{tr}(M^\star_{j,\mathcal{R};T}) &= \mathrm{tr}\left(\Sigma_T' U^\star_{j,\mathcal{R};T}\right) \\
&= \gamma_0 |\mathcal{R}T|^{-1} \sum_{s,u=\infty}^{\infty} c_{X,T}\left(\frac{s}{T}, u\right) \sum_{\ell=-\log_2 T}^{-1} 2^{\ell/2} \Psi_\ell(u) \qquad (4.39) \\
&= \gamma_0 |\mathcal{R}T|^{-1} \sum_{s,u=\infty}^{\infty} c_X\left(\frac{s}{T}, u\right) \sum_{\ell=-\log_2 T}^{-1} 2^{\ell/2} \Psi_\ell(u) \\
&\quad + \mathrm{Rest}_T \,. \qquad (4.40)
\end{aligned}$$

We now derive a bound for Rest_T. First we denote $\Delta_T(s/T, u) := c_{X,T}(s/T, u) - c_X(s/T, u)$. We first show that $\mathrm{TV}(\Delta_T(\cdot, u))$ is uniformly bounded in u. For all $I \in \{1, \ldots, T\}$ and for every sequence $0 < a_1 < a_2 < \ldots < a_I < 1$, we can write

$$\begin{aligned}
&\Delta_T(a_i, u) - \Delta_T(a_{i-1}, u) \\
&\quad = \sum_{j=-\infty}^{-1} \sum_{k=-\infty}^{\infty} \left\{ S_j\left(\frac{k}{T}\right) - S_j(a_i) \right\} \times \\
&\qquad\qquad \times \psi_{jk}([a_i T]) \psi_{jk}([a_i T] + u) \\
&\quad\quad - \sum_{j=-\infty}^{-1} \sum_{k=-\infty}^{\infty} \left\{ S_j\left(\frac{k}{T}\right) - S_j(a_{i-1}) \right\} \times \\
&\qquad\qquad \times \psi_{jk}([a_{i-1} T]) \psi_{jk}([a_{i-1} T] + u) \\
&\quad\quad + O(T^{-1}) \,,
\end{aligned}$$

where the $O(T^{-1})$ term comes from the approximation (3.23). Now, substitute k by $k + [a_i T]$ in the first sum, and by $k + [a_{i-1} T]$ in the second one. This leads to

$$\begin{aligned}
&\Delta_T(a_i, u) - \Delta_T(a_{i-1}, u) \\
&\quad = \sum_{j=-\infty}^{-1} \sum_{k=-\infty}^{\infty} \left\{ S_j\left(\frac{k}{T} + a_i\right) - S_j\left(\frac{k}{T} + a_{i-1}\right) + \right. \\
&\qquad\qquad \left. + S_j(a_{i-1}) - S_j(a_i) \right\} \psi_{jk}(0) \psi_{jk}(u) + O(T^{-1}).
\end{aligned}$$

Consequently, using the Cauchy-Schwarz inequality and (3.25),

$$\sum_{i=1}^{I} \left[\Delta_T\left(a_i, u\right) - \Delta_T\left(a_{i-1}, u\right)\right] \leqslant 2 \sum_{j=-\log_2 T}^{-1} L_j \sum_{k=-\infty}^{\infty} |\psi_{jk}(0)\psi_{jk}(u)| \leqslant 2\rho + K,$$

where K is a constant (because $I \leqslant T$), leading to $\mathrm{TV}(\Delta_T(\cdot, u)) \leqslant 2\rho + K$ uniformly in u. We can now bound Rest_T in (4.40) as follows:

$$\begin{aligned}\mathrm{Rest}_T &= \gamma_0 |\mathcal{R}T|^{-1} \sum_{s,u=-\infty}^{\infty} \Delta_T\left(\frac{s}{T}, u\right) \sum_{\ell=-\log_2 T}^{-1} 2^{\ell/2} \Psi_\ell(u) \\ &= \gamma_0 |\mathcal{R}|^{-1} \sum_{s,u=-\infty}^{\infty} \int_{s/T}^{(s+1)/T} dz \left\{ \Delta_T\left(z, u\right) + \Delta_T\left(\frac{s}{T}, u\right) + \right. \\ &\qquad \left. - \Delta_T\left(z, u\right) \right\} \sum_{\ell=-\log_2 T}^{-1} 2^{\ell/2} \Psi_\ell(u) \ ,\end{aligned}$$

as $|\Psi_\ell(u)|$ is uniformly bounded by 1,

$$\begin{aligned}&\leqslant \gamma_0 |\mathcal{R}|^{-1} \int_0^1 dz \sum_{u=-\infty}^{\infty} |\Delta_T\left(z, u\right)| \\ &\quad + \gamma_0 |\mathcal{R}|^{-1} \sum_{s,u=-\infty}^{\infty} \int_0^{1/T} dz \\ &\qquad \left| \Delta_T\left(\frac{s}{T}, u\right) - \Delta_T\left(z + \frac{s}{T}, u\right) \right| \ .\end{aligned}$$

From Proposition 3.1, the first term is $O(|\mathcal{R}T|^{-1})$. Using (3.28) and that $\mathrm{TV}(\Delta_T(\cdot, u))$ is uniformly bounded in u, the second term is also $O(|\mathcal{R}T|^{-1})$.

Now, using (3.27) and the definition of the matrix A, the first term

of (4.40) is bounded from below as follows:

$$\begin{aligned}
&\gamma_0|\mathcal{R}T|^{-1}\sum_{s,u=-\infty}^{\infty} c_X\left(\frac{s}{T},u\right)\sum_{\ell=-\log_2 T}^{-1} 2^{\ell/2}\Psi_\ell(u)\\
&\quad=\gamma_0|\mathcal{R}T|^{-1}\sum_{s,u=-\infty}^{\infty}\sum_{m=-\infty}^{-1} S_m\left(\frac{s}{T}\right)\Psi_m(u)\times\\
&\qquad\qquad\times\sum_{\ell=-\log_2 T}^{-1} 2^{\ell/2}\Psi_\ell(u)\\
&\quad=\gamma_0|\mathcal{R}T|^{-1}\sum_{s=0}^{T-1}\sum_{m=-\infty}^{-1} S_m\left(\frac{s}{T}\right)\inf_{m<0}\left[\sum_{\ell=-\log_2 T}^{-1} 2^{\ell/2}A_{m\ell}\right],
\end{aligned}$$

and we get the lower bound with Assumption 4.2. The upper bound is derived from (4.39), using Assumption 4.1, and $|\Psi_\ell(u)| \leqslant 1$ uniformly in $\ell < 0$ and $u \in \mathbb{Z}$.

2. We prove (4.37). The first inequality is (A.2). From (A.5), we write $\|M^\star_{j,\mathcal{R};T} - M_{j,\mathcal{R};T}\|_2^2 \leqslant \|\Sigma^{1/2}\|^4_{\text{spec}}\|U^\star_{j,\mathcal{R};T} - U_{j,\mathcal{R};T}\|_2^2$. Then, with Lemma 4.2,

$$\begin{aligned}
\|U^\star_{j,\mathcal{R};T} - U_{j,\mathcal{R};T}\|_2^2 &\leqslant 2\|U^\star_{j,\mathcal{R};T}\|_2^2 + 2\|U_{j,\mathcal{R};T}\|_2^2\\
&\leqslant \gamma_0^2|\mathcal{R}|^{-2}T^{-1}\sum_{m,\ell=-\log_2 T}^{-1} 2^{(\ell+m)/2}A_{\ell m} + K_2 2^j|\mathcal{R}|^{-2}T^{-1}
\end{aligned} \tag{4.41}$$

with (3.21) and $\sqrt{N_\ell N_m} \leqslant 2^{-(\ell+m)/2}4N_{-1}$,

$$\leqslant 4N_{-1}\gamma_0^2|\mathcal{R}|^{-2}T^{-1}\log_2^2(T) + O(T^{-1}). \tag{4.42}$$

The result follows from Lemma 4.3.

3. We prove (4.38). For T large enough, λ_T is strictly positive. Using Proposition 4.7 and, from Lemma 4.1, using

$$p_T^2 = \text{Var}(\underline{Z}'_T(M^\star_{j,\mathcal{R};T} - M_{j,\mathcal{R};T})\underline{Z}_T) = 2\|M^\star_{j,\mathcal{R};T} - M_{j,\mathcal{R};T}\|_2^2$$

and denoting

$$q_T = \|M^\star_{j,\mathcal{R};T} - M_{j,\mathcal{R};T}\|_{\text{spec}} ,$$

we can write

$$\begin{aligned}\Pr\left(\underline{Z}'_T(M^\star_{j,\mathcal{R};T} - M_{j,\mathcal{R};T})\underline{Z}_T > \lambda_T\right) \\ \leqslant \exp\left(-\frac{1}{2}\cdot\frac{(\operatorname{tr} M_{j,\mathcal{R};T})^2}{p_T^2\log_2^2 T + 2q_T\operatorname{tr}(M_{j,\mathcal{R};T})\log_2 T}\right) .\end{aligned}$$

(4.37) gives the rates for p_T and q_T, leading to the result. □

Proof of Proposition 4.4. We can write

$$\Pr\left(|Q_{j,\mathcal{R};T}| \leqslant \eta_\star\tilde\sigma_{j,\mathcal{R},T}\right) \leqslant \Pr\left(Q_{j,\mathcal{R};T} \leqslant \eta_\star\tilde\sigma_{j,\mathcal{R},T}\right).$$

In the proof of Proposition 4.3, we define a random set $\mathcal{A}$ such that $\Pr(\mathcal{A}) \geqslant 1-o_T(1)$ and (4.10) holds. Proposition 4.3 implies that $\tilde\sigma^2_{j,\mathcal{R},T} \leqslant \sigma^2_{j,\mathcal{R},T} + \gamma_T$ on $\mathcal{A}$ with $\gamma_T = o(2^{j-J_T}T^{-1})$. Together with Proposition 4.1, this implies

$$\frac{\tilde\sigma^2_{j,\mathcal{R},T}}{\sigma^2_{j,\mathcal{R},T}} \leqslant 1 + \frac{\gamma_T}{\sigma^2_{j,\mathcal{R},T}} \to 1,$$

and then the probability is bounded by

$$\Pr\left(Q_{j,\mathcal{R};T} \leqslant \eta_T\sigma_{j,\mathcal{R},T}\right) + o_T(1)$$

with $\eta_T = \eta_\star\sqrt{1+\gamma_T/\sigma^2_{j,\mathcal{R},T}}$.

With the notations of Lemma 4.7, we can write

$$\begin{aligned}&\Pr\left(Q_{j,\mathcal{R};T} \leqslant \eta_T\sigma_{j,\mathcal{R},T}\right) \\ &\quad= \Pr\left\{\underline{Z}'_T M^\star_{j,\mathcal{R};T}\underline{Z}_T \leqslant \eta_T\sigma_{j,\mathcal{R},T} + \underline{Z}'_T(M^\star_{j,\mathcal{R};T} - M_{j,\mathcal{R};T})\underline{Z}_T\right\}\end{aligned} \tag{4.43}$$

where $\underline{Z}_T = (Z_1,\dots,Z_T)'$ is a vector of iid Gaussian random variables with zero mean and $\operatorname{Var} Z_1 = 1$. We now define the random set $\mathcal{P}_T = \{\underline{Z}'_T(M^\star_{j,\mathcal{R};T} - M_{j,\mathcal{R};T})\underline{Z}_T \leqslant \lambda_T\}$ with $\lambda_T = \operatorname{tr} M^\star_{j,\mathcal{R};T} - \operatorname{tr} M_{j,\mathcal{R};T}(1 -$

$\log_2^{-1} T$). Lemma 4.7, equation (4.38), gives an upper bound for $\Pr(\mathcal{P}^c)$. Conditioning on $\mathcal{P}_T$, we can write

$$\Pr\left(Q_{j,\mathcal{R};T} \leqslant \eta_T \sigma_{j,\mathcal{R},T}\right) \leqslant \Pr\left(\underline{Z}_T' M^\star_{j,\mathcal{R};T} \underline{Z}_T \leqslant \eta_T \sigma_{j,\mathcal{R},T} + \lambda_T\right) + O\left(\exp\left\{-\frac{\sqrt{T}\,\mathrm{tr}\, M_{j,\mathcal{R};T}}{\log_2^2 T}\right\}\right).$$

Then, we are in position to apply Lemma 4.6 and we get

$$\Pr\left(\underline{Z}_T' M^\star_{j,\mathcal{R};T} \underline{Z}_T \leqslant \eta_T \sigma_{j,\mathcal{R},T} + \lambda_T\right) \leqslant \exp\left\{-\frac{\left(\eta_T \sigma_{j,\mathcal{R},T} - \mathrm{tr}\, M_{j,\mathcal{R};T}(1 - \log_2^{-1} T)\right)^2}{4\|M^\star_{j,\mathcal{R};T}\|_2^2}\right\}$$

provided that $\eta_T \sigma_{j,\mathcal{R};T} + \lambda_T \leqslant \mathrm{tr}\, M_T^\star$ (which holds true for T large enough by definition of λ_T). Proposition 4.1 allows to write $\mathrm{E} Q_{j,\mathcal{R};T} = \mathrm{tr}\, M_{j,\mathcal{R};T} = Q_{j,\mathcal{R}} + r_T$ with $r_T = O(T^{-1/2})$, and then

$$\Pr\left(|Q_{j,\mathcal{R};T}| \leqslant \eta_\star \tilde{\sigma}_{j,\mathcal{R},T}\right) \leqslant O\left(\exp\left\{-\|M^\star_{j,\mathcal{R};T}\|_2^{-2} Q^2_{j,\mathcal{R}}\right\}\right) + O\left(\exp\left\{-\sqrt{T} Q_{j,\mathcal{R}} \log_2^{-2} T\right\}\right) + o_T(1).$$

We conclude using that, under H_1, we can write

$$Q_{j,\mathcal{R}} = |\mathcal{U}||\mathcal{R}|^{-1} Q_{j,\mathcal{U}} + (|\mathcal{R}| - |\mathcal{U}|)|\mathcal{R}|^{-1} Q_{j,\mathcal{R}\setminus\mathcal{U}} > |\mathcal{U}||\mathcal{R}|^{-1}\theta,$$

from (4.13). Moreover, the following bound can be derived similarly to the derivation of (4.26), but using now (4.42):

$$\begin{aligned}\|M^\star_{j,\mathcal{R};T}\|_2^2 &\leqslant \|c_X\|_{1,\infty}^2 \|U^\star_{j,\mathcal{R};T}\|_2^2 \\ &\leqslant 4\|c_X\|_{1,\infty}^2 N_{-1} \gamma_0^2 |\mathcal{R}|^{-2} T^{-1} \log_2^2(T) + O(T^{-1}).\end{aligned}$$

Then

$$\Pr\left(|Q_{j,\mathcal{R};T}| \leqslant \eta_\star \tilde{\sigma}_{j,\mathcal{R},T}\right) \leqslant O\left(\exp\left\{-c' \frac{T}{\log_2^2 T} \frac{\theta^2 |\mathcal{U}|^2}{|\mathcal{R}|^2}\right\}\right) + O\left(\exp\left\{-c'' \frac{\sqrt{T}}{\log_2^2 T} \frac{\theta |\mathcal{U}|}{|\mathcal{R}|}\right\}\right) + o_T(1)$$

□

4.4.6 Proof of Proposition 4.5

Let $\mathcal{U}$ be a segment of $\wp(\mathcal{R})$. Consider the a.s. inequality

$$|Q_{j,\mathcal{R};T} - Q_{j,\mathcal{U};T}| \leqslant |Q_{j,\mathcal{U};T} - Q_{j,\mathcal{U}}| + |Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}| + \Delta_j(\mathcal{R},\mathcal{U})$$

where $\Delta_j(\mathcal{R},\mathcal{U})$ is defined in (4.14). In the regular case, $\Delta_j(\mathcal{R},\mathcal{U}) \leqslant b(\mathcal{U}) + b(\mathcal{R}) \leqslant C_j(\sigma_{j,\mathcal{U},T} + \sigma_{j,\mathcal{R},T})k_T$. Consequently, in the regular case,

$$\begin{aligned}
&\Pr(\mathcal{R} \text{ is rejected}) \\
&\quad\leqslant \sum_{\mathcal{U}\in\wp(\mathcal{R})} \Pr\left(|Q_{j,\mathcal{U};T} - Q_{j,\mathcal{R};T}| > (\eta\sigma_{j,\mathcal{U},T} + \eta\sigma_{j,\mathcal{R};T})\, k_T\right) \\
&\quad\leqslant \sum_{\mathcal{U}\in\wp(\mathcal{R})} \Pr\left(|Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}| > -C_j\sigma_{j,\mathcal{U};T}k_T + \eta\sigma_{j,\mathcal{U};T}k_T\right) \\
&\qquad+ \sum_{\mathcal{U}\in\wp(\mathcal{R})} \Pr\left(|Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}}| > -C_j\sigma_{j,\mathcal{R},T}k_T + \eta\sigma_{j,\mathcal{R},T}k_T\right)
\end{aligned}$$

Proposition 4.2 implies

$$\begin{aligned}
&\Pr(\mathcal{R} \text{ is rejected}) \\
&\quad\leqslant (\sharp\wp(\mathcal{R}))\, c_0 \exp\left\{-\frac{1}{8}\cdot\frac{\eta_T^2}{1 + \frac{\eta_T L_j}{|\mathcal{R}T|\sigma_{j,\mathcal{R},T}} + \frac{2^{j/2}\eta_T\nu(\|c_X\|_{1,\infty}+c_1\rho)}{\sigma_{j,\mathcal{R};T}|\mathcal{R}_T|\sqrt{T}}}\right\} \\
&\quad+ c_0 \sum_{\mathcal{U}\in\wp(\mathcal{R})} \exp\left\{-\frac{1}{8}\cdot\frac{\eta_T^2}{1 + \frac{\eta_T L_j}{|\mathcal{U}T|\sigma_{j,\mathcal{U},T}} + \frac{2^{j/2}\eta_T\nu(\|c_X\|_{1,\infty}+c_1\rho)}{\sigma_{j,\mathcal{U};T}|\mathcal{U}|\sqrt{T}}}\right\}
\end{aligned}$$

with

$$\begin{aligned}
\eta_T &:= \eta k_T\sqrt{1-\varphi_T} - C_j k_T \\
&= k_T 2^{-j/2}\left[5(2\alpha+p) - \sqrt{\alpha+p}\right].
\end{aligned}$$

Proposition 4.1 leads to $\sigma_{j,\mathcal{R},T}^{-1} \leqslant 2^{-j}\sqrt{K_1^{-1}|\mathcal{R}T|}$ and similarly for $\sigma_{j,\mathcal{U},T}^{-1}$. As $\delta \leqslant |\mathcal{R}| \leqslant |\mathcal{U}| \leqslant 1$, we consider the dominant terms in the sum, and

we can write, for T large enough, and with $2^{-j}L_j \leqslant \rho N_{-1}$,

$$\Pr(\mathcal{R} \text{ is rejected}) \leqslant 2c_0\left(\sharp_\wp(\mathcal{R})\right)\exp\left\{-\frac{1}{8}\cdot\frac{\eta_T^2}{1+\frac{\eta_T\rho N_{-1}}{\sqrt{K_1|\mathcal{R}T|}}+\frac{\eta_T\nu(\|c_X\|_{1,\infty}+c_1\rho)}{\sqrt{2^jK_1\delta}}}\right\}.$$

Replacing η_T and using $2\alpha + p \geqslant \sqrt{\alpha + p}$ lead to the result. □

4.4.7 Proof of Theorem 4.2

We first prove the following technical lemma.

Lemma 4.8. *Let $\underline{Z}_T = (Z_1, \ldots, Z_T)'$ be a vector of iid Gaussian random variables with zero mean and* $\operatorname{Var} Z_1 = 1$. *If $M_{j,\mathcal{R},T}$ is the matrix (4.28), v is a positive constant and $p \geqslant 2$, then, there exists T_0 such that*

$$\mathrm{E}\left(\underline{Z}_T' M_{j,\mathcal{R},T}\underline{Z}_T - \operatorname{tr} M_{j,\mathcal{R},T} + vk_T T^{-1/2}\right)^p \leqslant C\left(\kappa, \nu, \|c_X\|_{1,\infty}, p\right) T^{-p/2}\left(2^{1+j/2}|\mathcal{R}|^{-1} + vk_T\right)^p$$

for all $T \geqslant T_0$.

Proof. First, we write

$$\mathrm{E}\left(\underline{Z}_T' M_{j,\mathcal{R},T}\underline{Z}_T - \operatorname{tr} M_{j,\mathcal{R},T} + vk_T T^{-1/2}\right)^p = \sum_{r=0}^{p}\binom{p}{r}\mathrm{E}\left(\underline{Z}_T' M_{j,\mathcal{R},T}\underline{Z}_T - \operatorname{tr} M_{j,\mathcal{R},T}\right)^r \times v^{p-r}k_T^{p-r}T^{-(p-r)/2}. \quad (4.44)$$

Due to the relationship between the centered moments of a random variable and its cumulants, we can write

$$\mathrm{E}\left(\underline{Z}_T' M_{j,\mathcal{R},T}\underline{Z}_T - \operatorname{tr} M_{j,\mathcal{R},T}\right)^r = \sum_{m=0}^{r}\sum C(p_1, \ldots, p_m, m, \pi_1, \ldots, \pi_m, r)\kappa_{p_1}^{\pi_1}\ldots\kappa_{p_m}^{\pi_m},$$

where the second sum is over $p_1, \ldots, p_m, \pi_1, \ldots, \pi_m$ in $\{1, \ldots, r\}$ such that $\sum_{i=1}^m p_i\pi_i = r$, κ_{p_i} is the p_ith cumulant of $\underline{Z}_T M_{j,\mathcal{R},T}\underline{Z}_T$ and C

denotes a generic constant in this proof. From Lemma 4.1, we can write, using (4.29) and Proposition 4.1:

$$\begin{aligned}\kappa_{p_i} &\leqslant 2^{p_i-2}(p_i-1)!\,\mathrm{Var}(\underline{Z}_T' M_{j,\mathcal{R},T}\underline{Z}_T)\|M_{j,\mathcal{R},T}\|_{\mathrm{spec}}^{p_i-2}\\ &\leqslant 2^{p_i-2}(p_i-1)!K_2\nu^{p_i-2}\|c_X\|_{1,\infty}^{p_i-2}2^{jp_i/2}|\mathcal{R}|^{-p_i}T^{-p_i/2}.\end{aligned}$$

Consequently,

$$\begin{aligned}&\mathrm{E}\left(\underline{Z}_T' M_{j,\mathcal{R},T}\underline{Z}_T - \mathrm{tr}\, M_{j,\mathcal{R},T}\right)^r\\ &\qquad\qquad\leqslant C(\kappa,\|c_X\|_{1,\infty},r,\nu)2^{r(1+j/2)}|\mathcal{R}|^{-r}T^{-r/2}\ ,\end{aligned}$$

and using this inequality in (4.44) leads to the result. □

Proof of Theorem 4.2. Let $\tilde{\mathcal{R}}$ be the interval selected by the estimation procedure. We consider two cases: $|\tilde{\mathcal{R}}| < |\mathcal{R}|$ or $|\tilde{\mathcal{R}}| \geqslant |\mathcal{R}|$ and split the expectation into two parts:

$$\begin{aligned}\mathrm{E}|\tilde{S}_j(z_0) - S_j(z_0)|^p &= \mathrm{E}|\tilde{S}_j(z_0) - S_j(z_0)|^p\, 1_{|\tilde{\mathcal{R}}|<|\mathcal{R}|}\\ &\quad + \mathrm{E}|\tilde{S}_j(z_0) - S_j(z_0)|^p\, 1_{|\tilde{\mathcal{R}}|\geqslant|\mathcal{R}|}. \quad (4.45)\end{aligned}$$

First term ($|\tilde{\mathcal{R}}| < |\mathcal{R}|$)

In the first case, we make use of the inequality $|a-b|^p \leqslant 2^{p-1}|a|^p + 2^{p-1}|b|^p$ and write

$$\begin{aligned}\mathrm{E}|\tilde{S}_j(z_0) - S_j(z_0)|^p\, 1_{|\tilde{\mathcal{R}}|<|\mathcal{R}|} &\leqslant 2^{p-1}\mathrm{E}|S_j(z_0) - Q_{j,\tilde{\mathcal{R}}}|^p 1_{|\tilde{\mathcal{R}}|<|\mathcal{R}|}\\ &\quad + 2^{p-1}\mathrm{E}|Q_{j,\tilde{\mathcal{R}};T} - Q_{j,\tilde{\mathcal{R}}}|^p 1_{|\tilde{\mathcal{R}}|<|\mathcal{R}|}.\end{aligned}$$

As $|\tilde{\mathcal{R}}| < |\mathcal{R}|$, the evolutionary wavelet spectrum is homogeneous over $\mathcal{R}$ and $\tilde{\mathcal{R}}$ and then property (4.18) holds for $\tilde{\mathcal{R}}$. Then, using Proposition 4.1 on the variance, and the first point of Assumption 4.7, the first term of the right hand side is bounded as follows:

$$\begin{aligned}2^{p-1}\mathrm{E}|S_j(z_0) - Q_{j,\tilde{\mathcal{R}}}|^p 1_{|\tilde{\mathcal{R}}|<|\mathcal{R}|} &\leqslant 2^{p-1}\mathrm{E}(C_j\sigma_{j,\mathcal{R},T}k_T)^p\\ &\leqslant 2^{p-1}C_j^p k_T^p\{K_2 2^j/(T\delta)\}^{p/2}\\ &= 2^{p-1}\{(\alpha+p)K_2k_T^2\}^{p/2}(T\delta)^{-p/2} \quad (4.46)\end{aligned}$$

by definition of C_j (see Equation (4.19)). Now, if we denote $G_T = \underline{Z}_T' M_{j,\mathcal{R},T} \underline{Z}_T - \operatorname{tr} M_{j,\mathcal{R},T}$, then the second term may be written

$$2^{p-1}\mathrm{E}|G_T + \mathrm{bias}_T|^p 1_{|\tilde{\mathcal{R}}|<|\mathcal{R}|} \leqslant 2^{2p-2}(\mathrm{E}\{|G_T|^p 1_{|\tilde{\mathcal{R}}|<|\mathcal{R}|}\} + |\mathrm{bias}_T|^p)$$

where, using Lemma 4.1 and Proposition 4.1 for T large enough,

$$\begin{aligned} |\mathrm{bias}_T|^p &= |\operatorname{tr} M_{j,\mathcal{R},T} - Q_{j,\mathcal{R}}|^p \\ &= |\mathrm{E}Q_{j,\mathcal{R},T} - Q_{j,\mathcal{R}}|^p \\ &\leqslant C(p) 2^{jp/2} T^{-p/2} \end{aligned} \tag{4.47}$$

with a constant $C(p)$ depending on p. Finally, we now show that $\mathrm{E}|G_T|^p$ is uniformly bounded in T. Using $\delta < |\tilde{\mathcal{R}}| < |\mathcal{R}|$, we first note that Propositions 4.1 and 4.7 imply

$$\Pr\left(|G_T| > \lambda \sqrt{\frac{2^j K_2}{\delta^2 T}}\right) \leqslant \exp\left(-\frac{1}{2} \cdot \frac{\lambda^2}{1 + 2\lambda \frac{\tau_\infty \sqrt{|\mathcal{R}T|}}{2^j \sqrt{K_1}}}\right). \tag{4.48}$$

Denote $\mu_T = 2^j K_2/(\delta^2 T)$ and clip the integral

$$\mathrm{E}|G_T|^p = \int_0^\infty dx \ \Pr\left(|G_T|^p \geqslant x\right)$$

at point $\mu_T^{p/2}$. This leads to

$$\mathrm{E}|G_T|^p \lesssim \mu_T^{p/2} + \int_{\mu_T^{p/2}}^\infty dx \ \Pr\left(|G_T|^p > x\right)$$

and, with the change of variable $x = y^p \mu_T^{p/2}$, we get

$$\begin{aligned} \mathrm{E}|G_T|^p &\lesssim \mu_T^{p/2} + p\, \mu_T^{p/2} \int_1^\infty dy\ y^{p-1} \ \Pr\left(|G_T|^p > y\mu_T^{1/2}\right) \\ &\leqslant \mu_T^{p/2} + p\, \mu_T^{p/2} \int_1^\infty dy\ y^{p-1} \ \exp\left(-\frac{1}{2} \frac{y^2}{1 + 2y\tau_\infty \frac{\sqrt{|\mathcal{R}T|}}{2^j \sqrt{K_1}}}\right). \end{aligned}$$

For computing the integral, we note that $1 \leqslant y$, and then we have to compute $\int_1^\infty dy y^{p-1} \exp(-\alpha_T y)$. Straightforward computations show that this integral is bounded, up to a constant, by

$$\left(2 + \frac{4\tau_\infty \sqrt{\mathcal{R}T}}{2^j \sqrt{K_1}}\right)^p$$

which is multiplied by $\mu_T^{p/2}$ and then gives a constant bound, independent of T and j.

In conclusion, in the first case, we get the bound

$$\begin{aligned}
&\mathrm{E}|\tilde{S}_j(z_0) - S_j(z_0)|^p \, 1_{|\tilde{\mathcal{R}}|<|\mathcal{R}|} \\
&\leqslant 2^{p-1}\{(\alpha+p)K_2 k_T^2\}^{p/2}(T\delta)^{-p/2} \\
&\qquad\qquad + 2^{2p-2}C\left(\mathrm{E}1_{|\tilde{\mathcal{R}}|<|\mathcal{R}|} + 2^{jp/2}T^{-p/2}\right)
\end{aligned}$$

from (4.46) and (4.47). As

$$\mathrm{E}1_{|\tilde{\mathcal{R}}|<|\mathcal{R}|} = \Pr(|\tilde{\mathcal{R}}| < |\mathcal{R}|) \leqslant \Pr(\mathcal{R} \text{ is rejected}) = O\left(T^{-cp\sqrt{\delta}}\right)$$

by Proposition 4.5, we can write

$$\mathrm{E}|\tilde{S}_j(z_0) - S_j(z_0)|^p \, 1_{|\tilde{\mathcal{R}}|<|\mathcal{R}|} = O(T^{-cp\sqrt{\delta}} + k_T^p(T\delta)^{-p/2}).$$

Second term ($|\tilde{\mathcal{R}}| \geqslant |\mathcal{R}|$)

We consider now the second case. Select a subinterval $\mathcal{U}$ in $\mathcal{R} \cap \wp(\tilde{\mathcal{R}})$ containing z_0. Then, consider the decomposition

$$\begin{aligned}
\mathrm{E}|\tilde{S}_j(z_0) - S_j(z_0)|^p \, 1_{|\tilde{\mathcal{R}}|\geqslant|\mathcal{R}|} \leqslant \mathrm{E}\Bigg\{ &|Q_{j,\mathcal{U}} - S_j(z_0)| + \\
&+ |Q_{j,\mathcal{U};T} - Q_{j,\mathcal{U}}| + |Q_{j,\tilde{\mathcal{R}};T} - Q_{j,\mathcal{U};T}| \Bigg\}^p .
\end{aligned}$$

As the wavelet spectrum is regular on $\mathcal{U} \subset \mathcal{R}$, the term $|Q_{j,\mathcal{U}} - S_j(z_0)|$ is bounded by $C_j \sigma_{j,\mathcal{U},T} k_T$. On the other hand, using Proposition 4.1, $|Q_{j,\mathcal{U};T} - Q_{j,\mathcal{U}}| = |Q_{j,\mathcal{U};T} - \operatorname{tr} M_{j,\mathcal{U};T}| + R_T$ with $R_T = O(2^{j/2}T^{-1/2})$. Moreover, as $\tilde{\mathcal{R}}$ is selected by the estimation procedure, it holds $|Q_{j,\tilde{\mathcal{R}};T} - Q_{j,\mathcal{U};T}| \leqslant (\eta\tilde{\sigma}_{j,\tilde{\mathcal{R}},T} + \eta\tilde{\sigma}_{j,\mathcal{U},T})k_T$. Finally,

$$\begin{aligned}
\mathrm{E}|\tilde{S}_j(z_0) - S_j(z_0)|^p \, 1_{|\tilde{\mathcal{R}}|\geqslant|\mathcal{R}|} \leqslant \mathrm{E}\Bigg\{ &|Q_{j,\mathcal{U};T} - \operatorname{tr} M_{j,\mathcal{U};T}| \\
&+ R_T + C_j\sigma_{j,\mathcal{U},T}k_T + \left(\eta\sigma_{j,\tilde{\mathcal{R}},T} + \eta\sigma_{j,\mathcal{U},T}\right)k_T \Bigg\}^p .
\end{aligned}$$

With $2\alpha + p \geqslant \sqrt{\alpha + p}$, we can write

$$C_j \sigma_{j,\mathcal{U},T} k_T + \left(\eta \sigma_{j,\tilde{\mathcal{R}},T} + \eta \sigma_{j,\mathcal{U},T} \right) k_T \leqslant 11(2\alpha + p) K_2^{1/2} |\mathcal{U}T|^{-1/2} k_T \; .$$

Lemma 4.8 proves the existence of a constant c_5 depending on κ, ν, p, K_2 and on $\|c_X\|_{1,\infty}$, such that, for $T \geqslant T_0$,

$$\begin{aligned} \mathrm{E}\Big\{ & |Q_{j,\mathcal{U};T} - \operatorname{tr} M_{j,\mathcal{U};T}| \\ & + R_T + C_j \sigma_{j,\mathcal{U},T} \alpha_T + \left(\eta \sigma_{j,\tilde{\mathcal{R}},T} + \lambda \sigma_{j,\mathcal{U},T} \right) k_T \Big\}^p \\ & \leqslant c_5 |\mathcal{U}T|^{-p/2} \left\{ 2^{1+j/2} \delta^{-1} + 11(2\alpha + p) k_T \right\}^p \end{aligned}$$

since $|\tilde{\mathcal{R}}| \geqslant |\mathcal{U}| \geqslant \delta$, and the result follows. □

4.4.8 Proof of Proposition 4.6

Suppose $|\mathcal{R}_0| \vee |\mathcal{R}_1| = |\mathcal{R}_0|$ w.l.o.g. and write

$$\begin{aligned} & \Pr\left(\mathcal{R} \text{ is not rejected}\right) \\ & \qquad \leqslant \Pr\left\{ Q_{j,\mathcal{R},T} - Q_{j,\mathcal{R}_0,T} \leqslant \eta(\sigma_{j,\mathcal{R},T} + \sigma_{j,\mathcal{R}_0,T}) k_T \right\} . \end{aligned}$$

As in the proof of Proposition 4.4, we approximate $M_T = \Sigma_T'(U_{j,\mathcal{R};T} - U_{j,\mathcal{R}_0;T})\Sigma_T$ by $M_T^\star = \Sigma_T'(U_{j,\mathcal{R};T}^\star - U_{j,\mathcal{R}_0;T}^\star)\Sigma_T$, where $U_{j,\mathcal{R};T}^\star$ is defined in Lemma 4.7. Define the random set $\mathcal{P}_T = \{\underline{Z}_T'(M_T^\star - M_T)\underline{Z}_T \leqslant \lambda_T\}$ with $\lambda_T = \operatorname{tr} M_T^\star - (1 - \log_2^{-1} T) \operatorname{tr} M_T$, where $\underline{Z}_T = (Z_1, \ldots, Z_T)'$ is a vector of iid Gaussian random variables. As in the proof of Lemma 4.7, equation (4.38), we have

$$\Pr(\mathcal{P}_T^c) = O\left(\exp\left\{ -\frac{\sqrt{T} \operatorname{tr} M_T}{\log_2^2 T} \right\} \right) \; .$$

Conditioning on $\mathcal{P}_T$, we can write

$$\begin{aligned} & \Pr\left(\mathcal{R} \text{ is not rejected}\right) \\ & \qquad \leqslant \Pr\left\{ \underline{Z}_T' M_T^\star \underline{Z}_T \leqslant \eta(\sigma_{j,\mathcal{R},T} + \sigma_{j,\mathcal{R}_0,T}) k_T + \lambda_T \right\} \\ & \qquad\qquad + O\left(\exp\left\{ -\frac{\sqrt{T} \operatorname{tr} M_T}{\log_2^2 T} \right\} \right) \; . \end{aligned}$$

We are now in position to apply Lemma 4.6 with the symmetric, positive definite matrix $M^{\star}_{j,\mathcal{R};T}$. As $\eta(\sigma_{j,\mathcal{R},T} + \sigma_{j,\mathcal{R}_0,T})k_T + \lambda_T \leqslant \operatorname{tr} M^{\star}_T$ for T large enough, we can write, with Lemma 4.6,

$$\begin{aligned}\Pr(\mathcal{R} \text{ is not rejected}) &\leqslant \exp\left\{-\frac{1}{2}\frac{(\eta(\sigma_{j,\mathcal{R},T} + \sigma_{j,\mathcal{R}_0,T})k_T + \lambda_T - \operatorname{tr} M^{\star}_T)^2}{\operatorname{Var}\left(\underline{Z}'_T M^{\star}_T \underline{Z}\right)}\right\} \\ &\quad + O\left(\exp\left\{-\frac{\sqrt{T}\operatorname{tr} M_T}{\log_2^2 T}\right\}\right).\end{aligned}$$

Lemma 4.7 leads to $\operatorname{Var}(\underline{Z}'_T M^{\star}_T \underline{Z}) = O(|\mathcal{R}_0| T^{-1} \log_2^2 T)$, and then, replacing λ_T, the rate of the probability becomes

$$O\left(\exp\left\{-\frac{T(\operatorname{tr} M_T)^2 |\mathcal{R}_0|}{\log_2^2 T}\right\} + \exp\left\{-\frac{\sqrt{T}\operatorname{tr} M_T}{\log_2^2 T}\right\}\right).$$

The result follows using $\operatorname{tr} M_T = \theta_T$. □

4.5 Possible extension

We end this chapter with an open question for future research.

It is well-known that wavelets are suited to decompose certain inhomogeneous signals into a sparse wavelet coefficient vector. Consequently, the simultaneous localisation of wavelets in time and scale leads to possibly very sparse representations of evolutionary wavelet spectra. In this open section, we would like to develop this idea and to give an insight how the question of testing sparsity can be addressed.

In the multiscale representation (3.27) the coefficients $S_j(z)$ are depending on the continuous rescaled time $z \in (0,1)$. As $\Psi_j(0) = 1$ for all scales j, (3.27) decomposes the instantaneous variance as

$$c_X(z,0) = \sum_{j=-\infty}^{-1} S_j(z). \tag{4.49}$$

If we assume this variance to be non zero, it then follows that, at each time z, there exists a scale j where $S_j(z)$ is non zero. If only *few* scales are non zero for each z, we say that the wavelet spectrum is *sparse*.

There are many approaches in the literature where the notion of sparsity is quantified. In the context of wavelet decompositions of signals, we refer to Abramovich et al. [1] and the references therein for some possible definitions. One possibility to define a sparse wavelet spectrum is to test the null hypothesis that at most M scales are active, i.e.

$$\mathrm{H}_0 : \#\{j : Q_{j,\mathcal{R}\neq 0}\} \leqslant M$$

for a given time interval $\mathcal{R}$ and a given integer M. In this formulation, the number M quantifies the degree of sparsity.

For testing this null hypothesis, what kind of test statistic can be derived? A natural procedure is to consider all the histograms

$$v_j = Q_{j,\mathcal{R};T} \qquad j = -1, \ldots, -J_T \ ,$$

and to consider the corresponding rank vector $v_{(-1)} \geqslant v_{(-2)} \geqslant \ldots \geqslant v_{(-J_T)}$. According to our definition of sparsity, it is expected that the $J_T - M$ first elements of this rank vector differ significantly from zero. Then the idea is to test the significance of the next component, namely $v_{(-J_T+M-1)}$.

However, other definitions of sparsity could be possible. Sparsity can also mean that there is a relatively small proportion of relatively large histograms (as in Abramovich et al. [1]). Rewrite the sequence of theoretical histogram as follows:

$$\theta_j = Q_{j,\mathcal{R}} \qquad j = -1, -2, \ldots$$

We then have to control the decrease of the rank vector $\theta_{(j)}$, and sparsity would mean

$$\mathrm{H}_0 : \theta_{(j)} \leqslant M|j|^{-1/p} \qquad j = -1, -2, \ldots$$

Such constraint corresponds to an ℓ^p constraint, and the interesting range is for p small (substantial sparsity).

In that case, one way to test the sparsity is to study the decrease of the empirical vector $v_{(j)}$, and this can be done for instance through the False Detection Rate principle [6]. However, this methodology leads to new challenging theoretical problems, in particular for the control of the correlation between different scales of the CWP.

CHAPTER 5

Computational aspects and applications

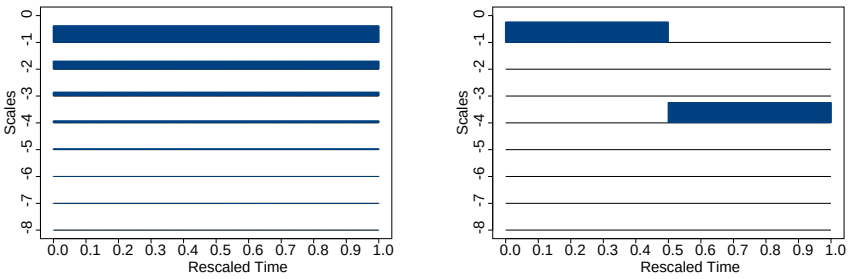
Above, we have considered two main problems, namely a local test of significance (4.1) for an evolutionary wavelet spectrum (EWS), and the pointwise estimation of the EWS. From a theoretical viewpoint, there exists a link between the solution of these two problems, since they are based on a non-asymptotic result on the deviation of a linear functional of the wavelet periodogram (Theorem 4.1).

The first aim of this chapter is to study the performance of these procedures on some simulated examples. Moreover, we show how they can be applied on some practical examples. We also provide an application of the method given by a new test of stationarity. This test is also illustrated on real data.

5.1 Preliminary remarks

In the following, the performance of the tests is evaluated on two specific models. The first one is a white noise with mean zero and variance 4, herewith denoted by $W_t \sim \mathrm{WN}(0, \sigma^2 = 4)$. This is of course a stationary process and, from the resolution of identity (3.12), its spectrum is given by $S_j^W(z) = S_j^W = 2^{j+2}$ for all $z \in (0,1)$ and $j < 0$. A plot of this theoretical spectrum is given in Figure 5.1(a). As the energy is spread evenly on all scales at each time point, this model corresponds to a non sparse spectrum.

The second model used in our simulation is nonstationary. It is given



(a) Wavelet spectrum of the stationary process $W_t \sim \mathrm{WN}(0, 4)$.

(b) Wavelet spectrum of the concatenated Haar process H_t.

Figure 5.1: The two theoretical wavelet spectra used in this chapter in order to evaluate the performance of the procedures.

by the concatenation of two simple stationary processes and corresponds to the wavelet process generated with the spectrum

$$S_j^H(z) = \begin{cases} 1 & \text{if } z \leqslant 0.5 \text{ and } j = -1 \ , \\ 1 & \text{if } z > 0.5 \text{ and } j = -4 \ , \\ 0 & \text{elsewhere} \end{cases}$$

and nondecimated Haar wavelets. A plot of this spectrum is given in Figure 5.1(b). The process is stationary on the two segments of time $(0, 0.5)$ and $(0.5, 1)$. Each of these segments corresponds to a special MA(q) process. More precisely, the whole process, which is denoted by H_t, is given by the concatenation of the MA(1) process

$$\frac{1}{\sqrt{2}}\varepsilon_t - \frac{1}{\sqrt{2}}\varepsilon_{t-1} \qquad \{\varepsilon_t\} \sim \mathrm{WN}(0, 1)$$

on the first segment, with the MA(15) process

$$\frac{1}{4}(\varepsilon_t + \cdots + \varepsilon_{t-7}) - \frac{1}{4}(\varepsilon_{t-8} + \cdots + \varepsilon_{t-15}) \qquad \{\varepsilon_t\} \sim \mathrm{WN}(0, 1).$$

on the second segment. The EWS of the process H_t is especially sparse, because, at each time point, only one scale is nonzero.

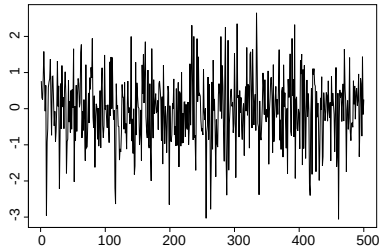
In the second model, the breakpoint between the two concatenated processes occurs at time $z = 0.5$. Here, it is worth mentioning that the

breakpoint may be placed at *any* time point in the rescaled time, and not only on dyadic time points. This is absolutely not a limitation of the model. Our choice of $z = 0.5$ here is simply for the sake of convenience. Later in this chapter, we will define a process with a breakpoint at a non dyadic time point (see Section 5.4).

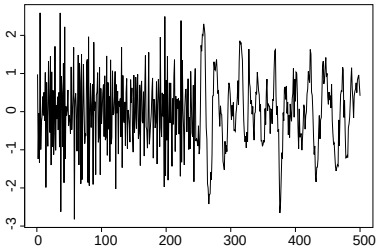
In the following, our evaluations are based on simulations of sample size $T = 500$. Indeed, the model does not limit the sample size to be a power of two. Our experience let us think that the sample size $T = 500$ is quite small in a context of nonstationarity modelling. This should be taken in mind in the interpretation of the following results. We also mention that the sample size of the real data in our study case will be significantly larger.

Figure 5.2(a) and (b) shows a realisation of the two processes W_t and H_t with Gaussian innovations ξ_{jk} and nondecimated Haar wavelets. We also compute the mean of 100 wavelet periodograms I_{jk} and corrected wavelet periodograms L_{jk}, computed with Haar wavelets in the two cases, from 100 independent realisations of the processes W_t and H_t. Figure 5.2(c) and (d) empirically show that, without the correction by A^{-1} in (3.31), the wavelet periodogram is biased for the estimation of the wavelet spectrum.

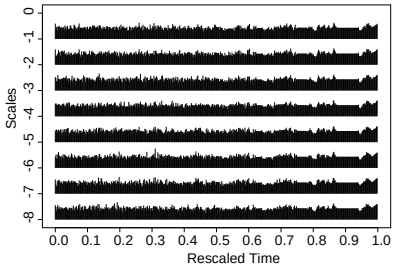
What is not showed in Figure 5.2 is the high variability of the corrected wavelet periodogram. Figure 5.3 plots a single corrected wavelet periodogram computed from one realisation of H_t with $T = 500$. From this figure, we can also see that the correction by A^{-1} in the computation of the CWP in (3.31) may lead to negative values at some time points. This happens because, in pratical situations, the correction is provided with a finite-dimensional matrix A_T of dimension $T \times T$. In our simulations below, this point does not pose a practical problem, since we are smoothing the CWP using histograms with a large window, and the negative values disappear.



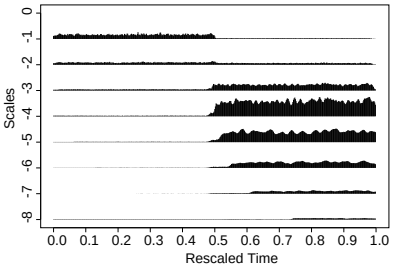
(a) One realisation of W_t.



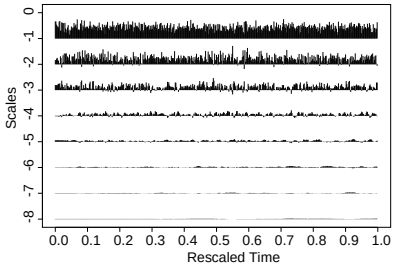
(b) One realisation of H_t.



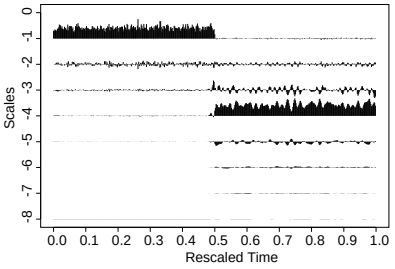
(c) Mean of wavelet periodograms (computed with Haar wavelets) from 100 independent simulations of W_t.



(d) Mean of wavelet periodograms (computed with Haar wavelets) from 100 independent simulations of H_t.



(e) Mean of CWP's from 100 independent simulations of W_t.



(f) Mean of CWP's from 100 independent simulations of H_t.

Figure 5.2: These plots concern 100 paths of W_t and H_t with sample size $T = 500$. H_t corresponds to the spectrum of Figure 5.1(b) using Gaussian innovations ξ_{jk} and nondecimated Haar wavelets.

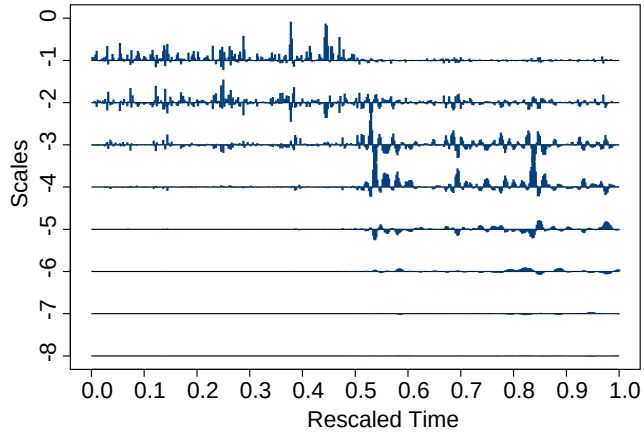


Figure 5.3: Corrected wavelet periodogram of one single realisation of H_t ($T = 500$).

However, if for some reasons it is desirable to guarantee the CWP to be nonnegative, one may use a refined algorithm for the correction, ensuring the positiveness of the CWP. One example is provided in [38], where the CWP is computed from the Linear Complementarity Problem [78], i.e. from the system

$$\begin{cases} I_{jk;T} \leqslant \sum_{m<0} A_{mj} L_{mk;T} \\ L_{mk;T} \geqslant 0 \\ \sum_{j<0} \left(\sum_{m<0} A_{mj} L_{mk;T} - I_{jk;T} \right) L_{jk;T} = 0. \end{cases}$$

We do not use this option later in our work.

5.2 Test of local significance

We now apply the test of local signifiance of the CWP developed in Section 4.2 to the two models described in the preceeding section. In this section, we assume that all nuisance parameters of the test procedure are known. The question of the preliminary estimation of these parameters is deferred to the next section.

In theory, the test is based on the exponential inequality given by Proposition 4.2. However, this bound is not sharp enough for our appli-

cations. The reason is revealed by some careful reading of the proof of this proposition. The proof is based on the decomposition

$$Q_{j,\mathcal{R};T} - Q_{j,\mathcal{R}} = \{Q_{j,\mathcal{R};T} - \mathrm{E}(Q_{j,\mathcal{R};T})\} + \{\mathrm{E}(Q_{j,\mathcal{R};T}) - Q_{j,\mathcal{R}}\}\,.$$

The bound for the deviation of the first term (the stochastic term) comes from a general result giving an exponential bound for quadratic forms of Gaussian random variables (Proposition 4.7). As we shall see in this section, this bound leads to sensible results in our simulations. In contrast, the bound for the second term (the deterministic term) comes from an upper bound for the bias, derived as in the proof of Proposition 4.1. This bound is not sharp.

One solution adopted in this chapter is to base the test on the null hypothesis

$$\mathrm{H}_0' : \mathrm{E}(Q_{j,\mathcal{R};T}) = 0 \text{ for a fixed scale } j < 0 \text{ and for all } z \in \mathcal{R}.$$

Then, the test may be based on the inequality

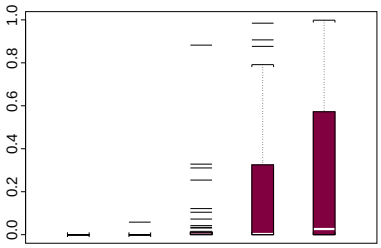
$$\Pr\left(|Q_{j,\mathcal{R};T} - \mathrm{E}Q_{j,\mathcal{R};T}| > \sigma_{j,\mathcal{R},T}\eta\right) \leqslant \exp\left(-\frac{1}{8}\cdot\frac{\eta^2}{1+\eta\frac{2^{j/2}\nu\|c_X\|_{1,\infty}}{|\mathcal{R}|T^{1/2}\sigma_{j,\mathcal{R},T}}}\right), \quad (5.1)$$

which is derived along the lines of the proof of Propostion 4.2.

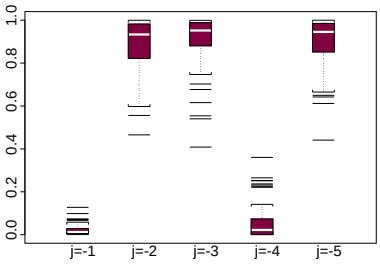
The test has been applied to simulated series from the two models described above. From each simulated series, we have computed the test statistics $\mathcal{T}_T := |Q_{j,\mathcal{R};T}|/\sigma_{j,\mathcal{R};T}$ for scales $j = -1$ up to -6 and for different choices of intervals $\mathcal{R}$. In the case of the spectrum of Figure 5.1(a) (standard white noise), we consider only the interval $\mathcal{R} = (0, 1)$. In the case of the second model, in Figure 5.1(b), we have considered the most relevant intevals $\mathcal{R} = (0, 1/2)$, $\mathcal{R} = (1/2, 1)$ and $\mathcal{R} = (0, 1)$. Having computed $\mathcal{T}_T$, we then evaluate the upper bound (5.1) for $\eta = \mathcal{T}_T$, i.e. we compute an approximated p-value of the test.

The results are presented in Figure 5.4. In this figure, we present a box plot of the 50 p-values obtained from 50 independent replications of $T = 500$ data. These show that very small p-values are obtained where the spectrum is not zero, and large p-values correspond to regions of sparsity. If we focus on the results for the white noise process, we can

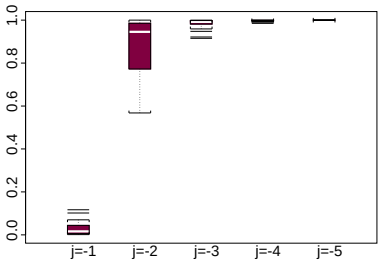
see that the test rejects very well the finer scales. However, the test does very often not reject the null hypothesis of sparsity for coarsest scales. This is due to the fact that the true EWS of the white noise at scale j is proportional to 2^j ($j < 0$). Then, the signal-to-noise ratio at lower scales is very low. This is in contrast with the process H_t, which is 1 at scale -1 and -4, for the rescaled time $z < 0.5$, and $z > 0.5$ respectively.



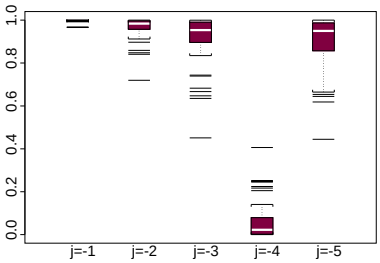
(a) p-values of the test of significance for $\{W_t\}$ computed on $\mathcal{R} = (0, 1)$ for all scales.



(b) p-values of the test of significance for $\{H_t\}$ computed on $\mathcal{R} = (0, 1)$ for all scales.



(c) p-values of the test of significance for $\{W_t\}$ computed on $\mathcal{R} = (0, 1/2)$ for all scales.



(d) p-values of the test of significance for $\{W_t\}$ computed on $\mathcal{R} = (1/2, 1)$ for all scales.

Figure 5.4: p-values of the test of significance applied on the processes $\{W_t\}$ and $\{H_T\}$, whose the theoretical spectrum is respectively in Figure 5.1(a) and (b). The results are based on 50 independent similations of times series with 500 data.

We also note from these results that the test makes a correct distinction between zero and nonzero regions of H_t on the whole scale (Figure

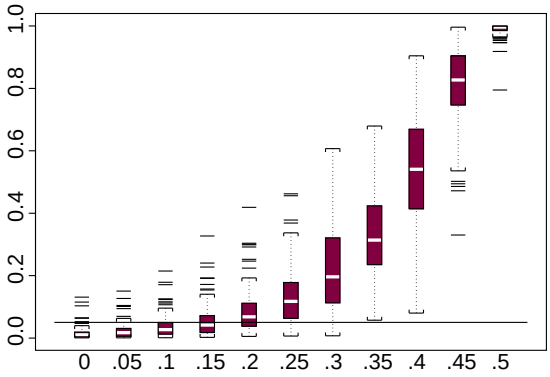
5.4(b)). Remember that, at scale $j = -1$ and -4, only half of the true EWS is active. This remarks leads to the following interesting question of how to study the behaviour of the test over an interval containing some inactive regions. For this, we consider again the concatenated process $\{H_t\}$. Let us consider the test at scale -1 for instance. The idea is to study the behaviour of the test statistics $\mathcal{T}_T$ when testing the signifiance over the interval $(z, 1)$, where we let the time point z vary between 0 and 1/2 in the rescaled time. The behaviour of the p-values is plotted in Figure 5.5(a) as a function of z on a grid of $(0, 0.5)$. If we fix the nominal level of the test at $\alpha = 0.05$, this figure shows that the test rejects mostly the null hypothesis on $(z, 1)$ when $z \leqslant 0.15$. In other words, 175 data corresponding to a non-zero segment of the spectrum were necessary in order to have a correct control of the Type II error of the test.

Similarly, in Figure 5.5(b) the test is also computed at scale -4 from the same process, for testing significance on $(0, z)$, where z takes its values on a grid of $(0.5, 1)$. This figure shows a nice empirical result, because the conclusion is the same as for Figure 5.5(a): If we fix the nominal level of the test at $\alpha = 0.05$, 175 data corresponding to a nonzero EWS at scale -4 are necessary in order to have a correct control of the Type II error of the test. This observation will be useful later, for the practical applications of the pointwise adaptive estimator of the EWS, because this simulation helps to understand how many data points are needed if in practice one wishes to detect a breakpoint in the spectrum with high probability.

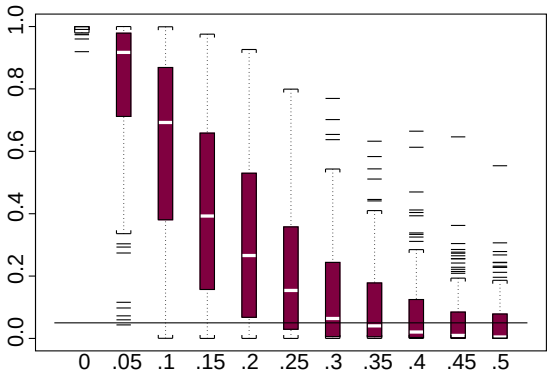
Remark 5.1. In our software, we have also used the following approximation for the deterministic matrix $U_{j,\mathcal{R};T}$ defined in (4.8): The entry (s, t) of $U_{j,\mathcal{R};T}$ is approximated by

$$|\mathcal{R}T|^{-1} \sum_{\ell=-[\log_2 T]}^{-1} A_{j\ell}^{-1} \Psi_\ell(s-t).$$

Since the approximated matrix is now Toeplitz, its encoding is significantly quicker than the encoding of the true matrix $U_{j,\mathcal{R};T}$. $\Diamond$



(a) p-values of the test of significance for $\{H_t\}$ at scale $j = -1$ computed on $\mathcal{R} = (z, 1)$, where z varies on a grid between 0 and 0.5.



(b) p-values of the test of significance for $\{H_t\}$ at scale $j = -4$ computed on $\mathcal{R} = (0, z)$, where z varies on a grid between 0.5 and 1.

Figure 5.5: p-values of the test of significance for $\{H_t\}$ with respect to an interval $\mathcal{R}$ with a varying length (based on 50 independent simulations of length 500). The horizontal line indicates the value 0.05.

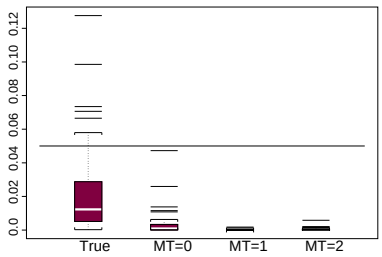
5.3 Estimation of the variance

In this section, the preliminary estimator of the variance described in Section 4.2.3 is evaluated on simulations of the process H_t. Recall that this estimator depends on two quantities: the interval $\mathcal{R}_T$ and the number M_T (see Assumption 4.5). We have tested the robustness of the test of significance with respect to these two parameters, and we present here a typical result.

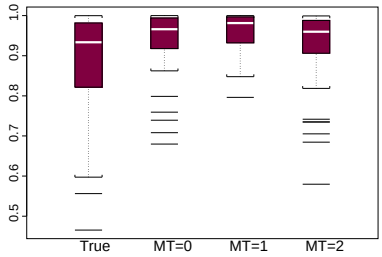
In the following experiment, we simulate 50 times a realisation of the process H_t, with $T = 500$. Suppose we would like to test the significance of the spectrum at scale $j = -1$, over the interval $\mathcal{R} = (0, 1)$. For each sample path, we compare the approximated p-values of the test obtained from four different test statistics $\mathcal{T}_T$. The first statistic uses the theoretical variance, as in the previous section. The three others use an interval $\mathcal{R}_T$ with length $|\mathcal{R}_T| = [\log_2 T]$, and with $M_T = 0, 1, 2$ respectively. Note that $M_T = 0$ corresponds to the case where the covariance matrix Σ_T is estimated by a diagonal matrix (see equation (4.9)). In our simulations, the interval $\mathcal{R}_T(s)$ is centered around s.

Thus, this experiment considers an interval $\mathcal{R}_T$ with a constant length $[\log_2(500)]$, and studies the variability of the test procedures with respect to the clipping parameter M_T. Figure 5.6 presents the results of the experiment for two scales ($j = -1$ and -2) and for the most relevant testing intervals $\mathcal{R} = (0, 1)$, $\mathcal{R} = (0, 1/2)$ or $\mathcal{R} = (1/2, 1)$. An observation of these results show that there is no significant difference between the four situations. This can be considered as a general conclusion, also for scales lower than -2 which are not reported here.

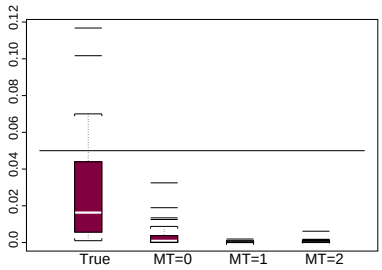
More surprisingly, we note that the best result is not obtained with the test statistic computed with the theoretical variance. In most of the plots, this statistic leads to a more spread out box plot than the one based on the estimated variance. A possible explanation for this phenomenon is that the theoretical variance $\sigma^2_{j,\mathcal{R}}$ is an asymptotic quantity depending on the local covariance function $c_X(z, \tau)$ (see (3.27)). From a practical viewpoint, this local covariance may be far from the true finite-sample covariance given by $\mathrm{Cov}(X_{[zT],T}, X_{[zT]+\tau,T})$. This argument especially holds in our case, where the simulations contain only $T = 500$ data.



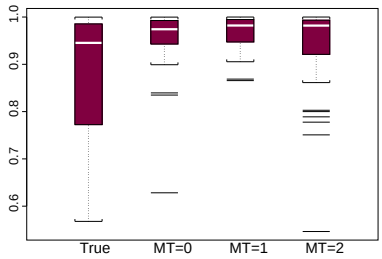
(a) Scale -1, $\mathcal{R} = (0, 1)$.



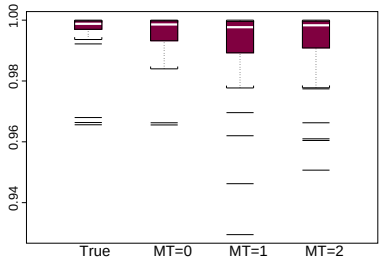
(b) Scale -2, $\mathcal{R} = (0, 1)$.



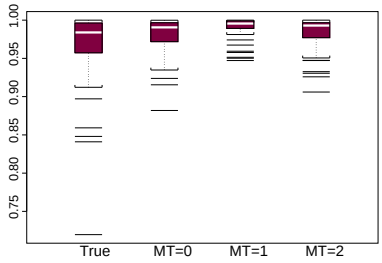
(c) Scale -1, $\mathcal{R} = (0, 1/2)$.



(d) Scale -2, $\mathcal{R} = (0, 1/2)$.



(e) Scale -1, $\mathcal{R} = (1/2, 1)$.



(f) Scale -2, $\mathcal{R} = (1/2, 1)$.

Figure 5.6: p-values for the test of significance applied on the process $\{H_t\}$ (50 independent replications of $T = 500$ data. Each figure presents four box-plots. The first correponds to the test statistic computed with the theoretical variance. The three others are with an estimated variance matrix, with $M_T = 0, 1, 2$ respectively with $|\mathcal{R}_T| = [\log_2 500]$, see equation (4.9).

In conclusion, even if the choice of R_T and M_T is important for the preliminary estimation of the covariance matrix Σ_T, their impact on the quality of the test of sparsity is limited. In practice, we recommend to set $R_T = \log_2 T$. In our applications, we use the following rule for the choice of M_T. We start with a large value for M_T (around 10) and examine the decreasing (or increasing) of the off-diagonals of the covariance matrix. Very often, in our applications, these off-diagonals start with an abrupt decreasing, sometimes followed by an increasing trend. When this behaviour is observed in several off-diagonals of the matrix, we recommand to clip the matrix just before the abrupt decreasing. This procedure has been followed for choosing M_T in the case studies below.

We end this section by mentioning that the pre-estimation of the variance is not computationaly expensive. The key point here is that we only need to estimate the covariance matrix Σ_T by $\tilde{\Sigma}_T$ of the whole process and to store it. Then, as explained in Section 4.2.3, the estimated variance of a histogram $Q_{j,\mathcal{R};T}$ is obtained by computing $2\|U'_{j,\mathcal{R};T}\tilde{\Sigma}_T\|_2^2$, where the matrix $U_{j,\mathcal{R};T}$ is purely deterministic.

Remark 5.2. One possible way to avoid the estimation of the variance would be to use the bounds of the variance derived in Proposition 4.1 in the test of sparsity. However, some preliminary simulation studies showed that these bounds are not sharp in practice. ◊

5.4 Adaptive estimation of the wavelet spectrum

The pointwise adaptive estimator (4.16–4.17) of $S_j(z_0)$ is based on the detection of the interval $\mathcal{R}$ around z_0 where the evolutionary wavelet spectrum (EWS) is homogeneous. In this section, we want to be more specific about all the quantities which are needed in the procedure.

5.4.1 Choice of the sets Λ, $\wp(\mathcal{R})$

We first need to choose the sets Λ and $\wp(\mathcal{R})$. Recall that Λ defines a family of interval-candidates $\mathcal{R}$ and that, for each candidate $\mathcal{R}$, the test of homogeneity compares the behaviour of the corrected wavelet periodogram (CWP) on $\mathcal{R}$ and several subintervals $\mathcal{U}$. The set $\wp(\mathcal{R})$ contains these subintervals $\mathcal{U}$.

Several propositions have been proposed in the literature for choosing these two sets [74, 105]. In our computations, we used the following sets. For each scale $j < 0$, the CWP (3.31) is evaluated on a grid k/T,

$k = 0, \ldots, T-1$ in time. Let us fix an integer K that will be discussed later on and define the set $\mathcal{K} = \{iT/K, T : i = 0, \ldots, (K-1)\}$. The set $\mathcal{K}$ depends on K and defines a grid of $\{0, \ldots, T-1\}$ including the time points 0 and $T-1$. We choose the set Λ as

$$\Lambda = \{[r_0/T, r_1/T] : r_0, r_1 \in \mathcal{K} \text{ and } r_0 < [z_0 T] < r_1\}\,.$$

Finally, the set Λ depends only on the parameter K that will be discussed below. Next, for every interval $\mathcal{R} = [r_m/T, r_n/T]$ in Λ, we define the set $\wp(\mathcal{R})$ of subintervals $\mathcal{U}$ by taking all smaller subintervals $[r_k/T, r_n/T]$ with the right end point r_n/T and similarly all smaller intervals $[r_m/T, r_\ell/T]$ with the left end-point r_m/T:

$$\wp(\mathcal{R}) = \{\mathcal{U} = [r_k/T, r_n/T] \text{ or } \mathcal{U} = [r_m/T, r_\ell/T] : m < \ell, k < n\}$$

if $\mathcal{R} = [r_m/T, r_n/T]$.

5.4.2 The procedure

The test of homogeneity of the EWS over an interval-candidate $\mathcal{R}$ is based on the test statistic

$$\mathcal{T}_T(\mathcal{R}, \mathcal{U}) := \frac{|Q_{j,\mathcal{R};T} - Q_{j,\mathcal{U};T}|}{\log_2^2(T)\sqrt{\mathrm{Var}\,|Q_{j,\mathcal{R};T} - Q_{j,\mathcal{U};T}|}}$$

for each subinterval $\mathcal{U}$ in $\wp(\mathcal{R})$. An upper bound for the deviation of this statistic is given in (4.15). In other words, under the assumption of homogeneity, that is if the difference $\Delta(\mathcal{R}, \mathcal{U})$ is small, then

$$\Pr(\mathcal{T}_T(\mathcal{R}, \mathcal{U}) > \eta) \leqslant g(\mathcal{R}, \mathcal{U}, \eta)\,, \tag{5.2}$$

where g follows from equation (4.15). Then the adaptive iterative procedure for selecting the interval of homogeneity may be summarized as follows:

Initialization. Select the smallest interval $\mathcal{R}$ in Λ.

Iteration. Select the next interval $\mathcal{R}$ and calculate the corresponding estimate $Q_{j,\mathcal{R},T}$ and the estimated variance $\tilde{\sigma}^2_{j,\mathcal{R},T}$.

Testing homogeneity. Reject $\mathcal{R}$ if there exists one $\mathcal{U} \in \wp(\mathcal{R})$ such that

$$g\,(\mathcal{R}, \mathcal{U}, \mathcal{T}_T(\mathcal{R}, \mathcal{U})) > g_0\,.$$

Loop. If $\mathcal{R}$ is not rejected, then iterate using a larger interval. Otherwise, select the latest non rejected interval.

This procedure requires the preselection of the sets Λ and $\wp(\mathcal{R})$, but also the choice of a constant g_0 which will be discussed later on.

Remark 5.3. With the approximation explained in Remark 5.1, the denominator of the test statistics is simple to compute. Indeed, following the reasoning at the beginning of Section 4.2.3, the variance of $|Q_{j,\mathcal{R};T} - Q_{j,\mathcal{U};T}|$ is simply $2\|U'_{j,\mathcal{U},\mathcal{R};T}\Sigma_T\|_2^2$, where the entry (s,t) of $U_{j,\mathcal{U},\mathcal{R};T}$ is given by

$$\big(|\mathcal{R}T|^{-1} - |\mathcal{U}T|^{-1}\big) \sum_{\ell=-[\log_2 T]}^{-1} A_{j\ell}^{-1}\Psi_\ell(s-t).$$

In our computations, the matrix Σ_T and the sum over ℓ are first computed and stored. Then, the variance is computed by a normalisation with the factor $(|\mathcal{R}T|^{-1} - |\mathcal{U}T|^{-1})$. $\Diamond$

To end this subsection, let us discuss two technical points of the algorithm. The first point is the choice of the first interval-candidate $\mathcal{R}$ needed to initialize the procedure. One first recommendation is to not take this interval too small. Here, the study of Figure 5.5 above is very helpful. From this study, we conclude that this interval should not have less than 175 data, such that we have a correct control of the Type II error of the test.

The second point is: having accepted an interval $\mathcal{R}$ in the procedure, how can we define the next interval-candidate in the iteration step? The point here is that increasing of $\mathcal{R}$ may be done to the right of the interval, or to the left of the interval. In our code, we consider the two possibilities at each iteration, that is we consider an increasing to the right, denote it by $\mathcal{R}_r$, and to the left, denote it by $\mathcal{R}_\ell$. We test the homogeneity of the EWS over $\mathcal{R}_r$ and $\mathcal{R}_\ell$ seperately. If the test rejects the homogeneity over one of the two intervals, then we reject this interval and do no longer increase the interval-candidate in this direction. If the test does not reject any of the two possibilities, then we compare the minimal probability obtained in the two situations, that is we compare

$$m_r = \min_{\mathcal{U}\in\wp(\mathcal{R}_r)} g(\mathcal{R}_r,\mathcal{U},\mathcal{T}_T(\mathcal{R}_r,\mathcal{U}))$$

and

$$m_\ell = \min_{\mathcal{U}\in\wp(\mathcal{R}_\ell)} g(\mathcal{R}_\ell,\mathcal{U},\mathcal{T}_T(\mathcal{R}_\ell,\mathcal{U}))$$

and we select the new interval corresponding to the maximum of m_r and m_ℓ.

5.4.3 Discussion of the constants K and g_0

Now the procedure basically depends on the choice of two parameters K and g_0. The number K defines the set of subintervals in the test of homogeneity, and g_0 is a kind of level for this test. As a matter of fact, g_0 plays the role of a smoothing parameter.

It is worth mentioning that the parameters K and g_0 are *global*, in the sense that they do not depend on the time point z_0 where we are testing the homogeneity. They no longer depend on $\mathcal{R}$ or $\mathcal{U}$, and they are fixed by the model only.

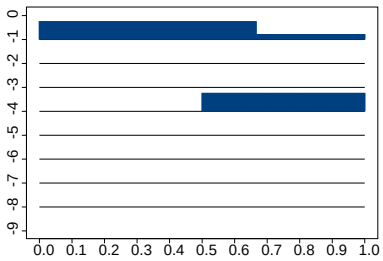
The choice of K and g_0 follows from some preliminary study of the CWP. In a different context, Mercurio and Spokoiny [74] propose to select these nuisance parameters by minimisation of the mean square prediction error. This could also be possible in our context, since the prediction theory of LSW processes is developed in the next chapter.

5.4.4 Simulated example

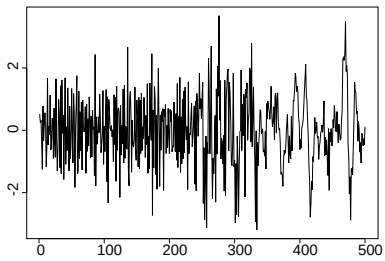
To illustrate the pointwise adaptive estimation procedure on a simulated example, we consider a new process, called `ghaar`, whose theoretical spectrum is given in Figure 5.7(a). This process has two active scales on $j = -1$ and -4. The scale $j = -4$ is active only from the middle of the time series, while the scale $j = -1$ is active at each time point, but with a breakpoint occuring at time point $2T/3$.

In this experiment, we focus on the estimation of scale $j = -1$. The procedure described above is for a pointwise estimator. If we wish to estimate a whole scale, the algorithm has to be performed on several points of the scale. In the following, we have applied the estimator on 19 equidistant time points. If one wants to have an estimator of the whole scale, these 19 estimators may be linked, for instance by segments or using a more sophisticated interpolation algorithm.

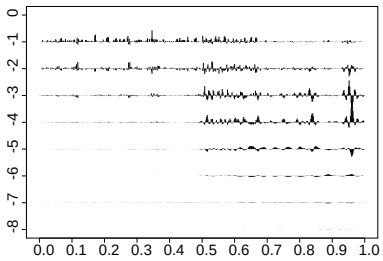
Moreover, when the estimation procedure is performed on several points of the scale, it may happen that some homogeneity tests are computed several times for the same intervals $\mathcal{R}$ and $\mathcal{U}$. For instance,



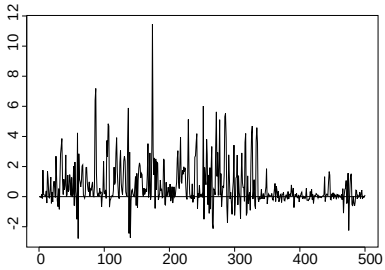
(a) EWS of the `ghaar` process.



(b) One realisation of length $T = 500$.



(c) CWP of the `ghaar` process.



(d) Scale $j = -1$ of the corrected wavelet periodogram

Figure 5.7: The `ghaar` process uses the spectrum in (a) with nondecimated Haar wavelets and Gaussian increments in the Definition of the locally stationary wavelet process.

this situation arises if we estimate the EWS at two points included in one true interval of homogeneity. In theory, we expect that the estimated interval of homogeneity $\tilde{\mathcal{R}}$ will be the same for the two points. As a consequence, in this situation, many identical tests of homogeneity will be computed in the estimation at the first and second point. To overcome this loss of efficiency in the procedure, our code stores all tests of homogeneity for each $\mathcal{R}$ and $\mathcal{U}$, such that each test is computed one single time in the global estimation. In a sense, the algorithm for the global estimation gets more and more informations when the estimation goes from one point to another point of estimation. In terms of computation time, the gain with this self-learning procedure is highly significant.

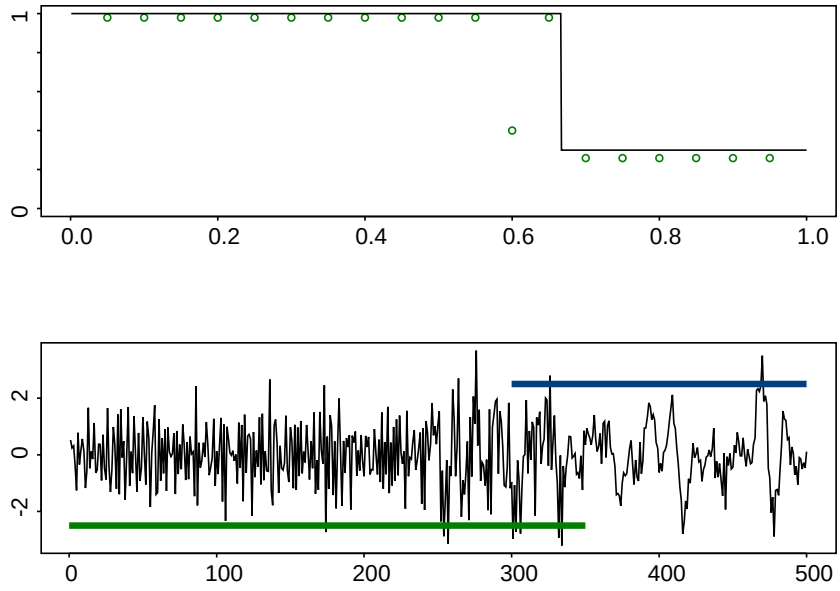


Figure 5.8: Above is the result of the pointwise adaptive estimator at scale $j = -1$ for the `ghaar` process (the solid line is the true spectrum, and the dots are the pointwise adaptive estimators). The estimation is obtained from the realisation of the process plotted at the bottom. The two horizontal lines on the time series correspond to some intervals of homogeneity selected by the adaptive procedure (see text).

The results are presented in Figure 5.8, and correspond to the following choices of parameters: $K = 20$, $g_0 = 0.4$. In this simulation, these two parameters are chosen by hand, but the result is robust to this choice (that is, a small change of the parameters K and g_0 does not change the estimation in a significant way). It is worth mentioning that these results are given with the pre-estimator of the variance. For this pre-estimation, we take $|\mathcal{R}_T| = [\log_2(500)]$ and $M_T = 1$. The above figure plots the true EWS at scale $j = -1$ (solid line) together with our estimators at 19 equidistant time points (dots). The constancy of the EWS on $(0, 2/3)$ and $(2/3, 1)$ is well detected by the estimator, except at one time point near the breakpoint. If we consider the CWP at scale

$j = -1$, see Figure 5.7(d), one may observe that the CWP has a short period with lower amplitudes before the time break. We believe that the single bad pointwise estimator comes from this phenomenon. Once again, it is important to note that the CWP is a highly variable quantity, and our estimation is provided with only 500 data.

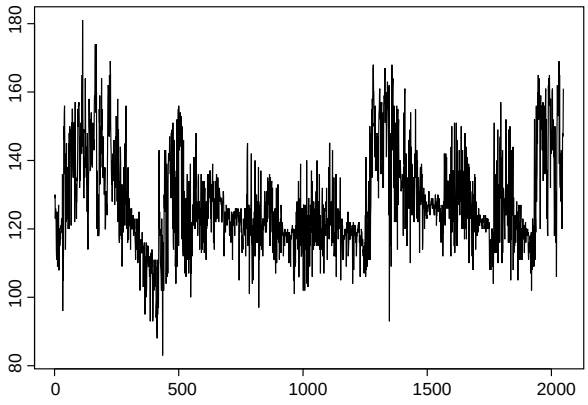
In the estimation plotted in Figure 5.8, all the pointwise estimators near 1 select the same interval of homogeneity $\tilde{\mathcal{R}}$, as expected. This interval is showed together with the original time series at the bottom of Figure 5.8. Similarly, the estimators near 0.25 select the same interval of homogeneity, which is plotted on the second plot.

5.5 Case study: Baby heart rate

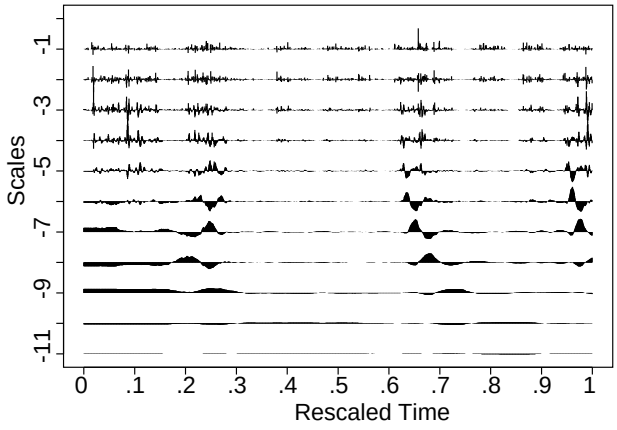
The test of significance and the pointwise adaptive estimator are devoted to different specific classes of statistical problems. On one hand, the test of significance is of use when we would like to measure a change of regime in an observed process. Many examples arise where the effect of an input is measured on a time series, for instance the effect of a drug on the heart rate measured by an electrocardiogram recording. The resulting time series is expected to be globally nonstationary, and we think our test procedure may be used to detect if a scale that is not active before the input becomes active after the input. Moreover, as already said above, the test of significance may be applied on a whole scale, in order to test the significance of one given scale in an observed process. On the other hand, the pointwise adaptive estimator of the wavelet spectrum may be applied time point by time point, leading to an estimator of the whole wavelet spectrum of the process.

In this case study, we illustrate the possibility to combine the two procedures. The key idea is to test the significance of some whole scales over the whole time, before performing the estimation procedure on the scales which are significantly different from zero.

Our study concerns a heart rate (electrocardiogram (ECG)) recording on a 66-day-old infant. Figure 5.9(a) plots the series, sampled at 1/16 Hz and recorded from 21:17:59 to 06:27:18 ($T = 2048$ observations). This series is considered in Nason and von Sachs [82] and Nason et al. [83] as a motivating example for the exploratory analysis using the LSW model. First of all, it is unlikely that this time series will be a stationary time series. The heart rate varies considerably over time and changes significantly between periods of sleep and waking. These



(a) ECG recording. Series is sampled at 1/16 Hz and is recorded from 21:17:59 to 06:27:18 ($T = 2048$ observations).



(b) CWP of the data, computed with Haar wavelets.

Figure 5.9: ECG recording of a 66-day-old infant, and its CWP. (Data courtesy Institute of Child Health, Royal Hospital for Sick Children, Bristol, and Guy P. Nason [79])

changes are of a big interest for the paediatricians. These are related to other variables of interest which are not easily observable nor easy to quantify. For instance, the paediatricians are interested to measure the sleep states (quiet, active, awake,...) using some objective measures and the question is how the heart rate may be used as a tool for measuring the sleep state. We shall come back later to this question of the link between the ECG and the sleep states.

Another argument for applying the LSW model to these data is because a preliminary analysis in the Fourier domain did not give very informative results in terms of exploratory analysis. The Fourier analysis we are mentioning here is a localised analysis, for instance using the windowed Fourier transform (WFT) [70]. Another drawback of the analysis in the Fourier domain is the choice of the length of window in the WFT. Moreover, the result of this analysis is a two-dimensional surface which is not always easy to interpret.

In contrast, the approach using the LSW model leads to a multiscale representation of the nonstationary process. Figure 5.9(b) shows the CWP of the heart rate, computed with nondecimated Haar wavelets. The CWP is not smoothed and highly variable, and our goal now is to extract some useful information from it.

As a first step, we need a pre-estimator of Σ_T and, for this, need to choose an appropriate parameter M_T and segment $\mathcal{R}_T$. From the conclusions of Section 5.3 above, we choose $\mathcal{R}_T(z)$ centered in z and of length $[\log_2 T]$. The selection of M_T is provided as described in Section 5.3. We first compute $\tilde{\Sigma}_T$ with $M_T = 10$ and then analyse the behaviour of its off-diagonals. Figure 5.10 superimposes the value of 10 different off-diagonals, that is we superimpose $\tilde{\Sigma}_{s,s+u}$ for $u = 0, \ldots, 10$ and for 10 different s. This shows a similar behaviour between the off-diagonals, which decrease quickly to $M_T = 2$ then vary slowly. We then choose $M_T = 2$ in the pre-estimation of $\tilde{\Sigma}_T$.

To start the analysis, we want to detect if some scales of the CWP are not significant. For this, we apply our test of significance over the interval $\mathcal{R} = (0, 1)$ at each scale. The results of the test are given in Table 5.1. From this table, we conclude that *the only active scales of the data are given by* $j = -1, -7, -9$ *and* -10, and the other scales are not significantly different from zero. To our knowledge, such conclusion is new for these data, and also very helpful since it indicates that the analysis should focus on 4 active scales only.

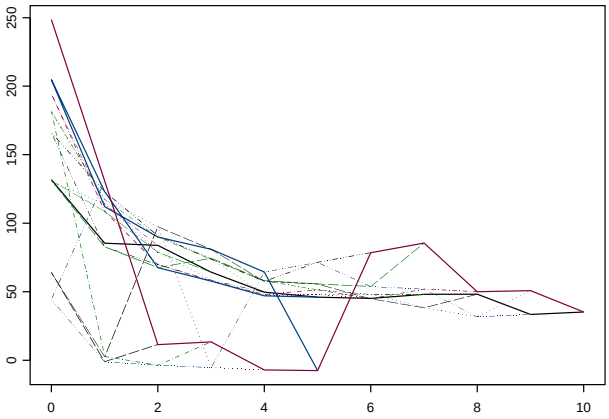
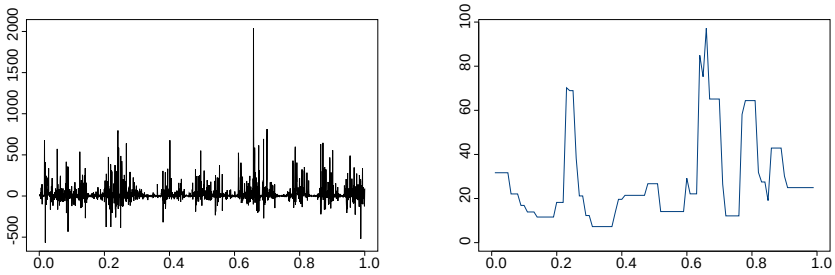


Figure 5.10: Ten off-diagonals of the estimated matrix $\tilde{\Sigma}_T$.

We now focus on the significant scale $j = -1$ and apply our estimation procedure. The results are given in Figure 5.11. In our estimation, we estimate the EWS at $K = 100$ points. In fact, this choice does not follow the "rule of the 175 data" discovered by Figure 5.2. The reason is that a careful observation of the CWP at scale -1 (see Figure 5.11(a)) shows that some changes of regime in the data occur very often, and some regions of the spectrum have clearly an interval of homogeneity containing less than 175 data. In the estimation procedure, we also set $g_0 = 0.95$, which is quite large. In terms of homogeneity tests, this means that we perhaps reject the homogeneity assumption very often. In our opinion, this is in fact very sensible, because this error (Type I error) is not as serious than the complementary type II error. Indeed, in the pointwise estimation, it is not so bad if the homogeneity interval is too small, since the corresponding histogram will not differ too much from the histogram based on a possible larger truly homogeneous interval. In contrast, if we choose a too large interval, for instant if we choose an interval which contains some discontinuity, then the estimator will be significantly different. Once again, it could be possible to select g_0 automatically by minimisation of a criterion like the mean square prediction error (see Chapter 6).

Scale	$Q_{j,\mathcal{R};T}$	$\tilde{\sigma}^2_{j,\mathcal{R};T}$	approximated p-value
-1	31.66	12.01	$1.68 \cdot 10^{-4}$
-2	9.71	22.50	0.60
-3	9.31	101.98	0.90
-4	19.69	104.83	0.63
-5	17.25	66.95	0.57
-6	20.14	35.90	0.24
-7	44.66	18.08	$1.44 \cdot 10^{-6}$
-8	3.67	8.79	0.83
-9	33.53	4.07	$5.77 \cdot 10^{-15}$
-10	17.56	1.69	$5.66 \cdot 10^{-10}$

Table 5.1: Results of the test of significance over the interval $\mathcal{R} = (0, 1)$ performed at each scale j between -1 and -10 for the heart rate data.



(a) Scale $j = -1$ of the CWP.

(b) The estimator ($j = -1$).

Figure 5.11: Pointwise adaptive estimator performed at scale $j = -1$ for the baby ECG. The estimator is computed at 100 different points, and we line up two consecutive points.

Simultaneously to the ECG recording, some experts on the analysis of brain-waves and eye movements recorded the sleep states of the infant. These states are recorded independently of the ECG. They are plotted in Figure 5.12 together with our estimator of the EWS at scale $j = -1$. The observers classifies the sleep states as quiet sleep (A), between quite

and active sleep (B), active sleep (C) and awake (D).

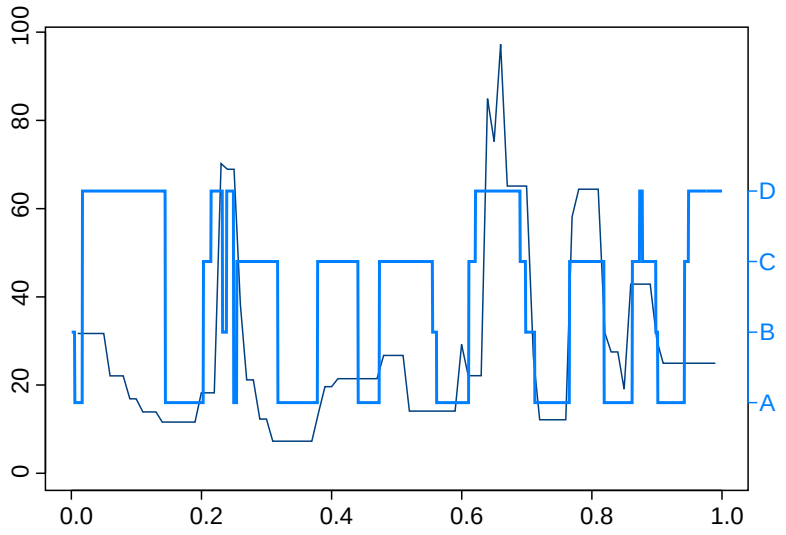
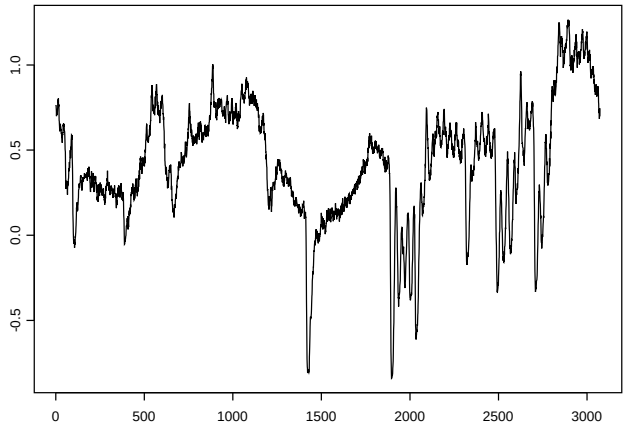
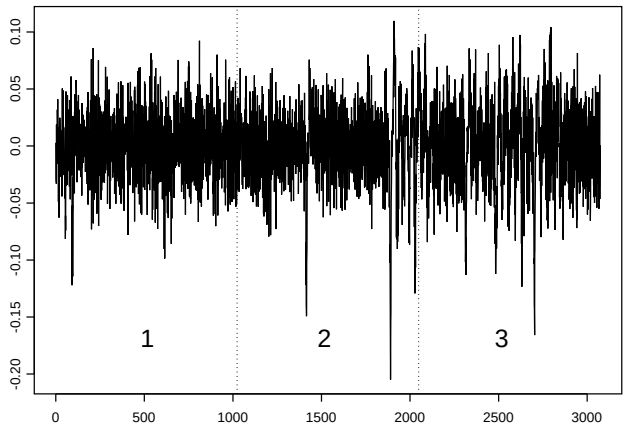


Figure 5.12: Pointwise adaptive estimator of the EWS at scale $j = -1$ together with the sleep states (A, quiet sleep; B, state between A and C; C, active sleep; D, awake).

It is clear that there is some relationship between our estimator and the sleep states. In particular, periods of activity occur whilst the estimate of $S_{-1}(z)$ is large, and periods of quiet sleep when it is small. It is worth mentioning that the ECG is easy to measure, while the sleep states is more tricky and less objective. With our estimator, we are also able to detect some changes in the sleep states, sometimes with a small delay. Then, we believe that our estimator may help to provide some objective measurement of the activity during sleep. Finally, we mention that a finer analysis of the fitting between the sleep states and our adaptive estimator is also one possibile way to derive an automatic choice of the parameter g_0.



(a) Tremor data ($T = 3072$ data)



(b) First-order difference

Figure 5.13: Tremor data (Data courtesy Cognitive Neuroscience Laboratory of the University of Quebec, Anne Beuter and Roderick Edwards).

5.6 Test of stationarity

We end this chapter with an interesting and simple application of the previous theory given by a new test of covariance stationarity for time series.

5.6.1 The basic idea

A covariance stationary process is characterised by an EWS which is constant over time, i.e. $S_j(z) = S_j$ for each scale j (see Chapter 3). The key idea for the test of stationarity is to test if the EWS is constant over time at each scale or not. With this idea, testing the stationarity of a time series is equivalent to test the homogeneity of the EWS at each scale over $\mathcal{R} = (0, 1)$. This procedure is illustrated in the following case study.

5.6.2 Case study: Tremor data

The data shown in Figure 5.13(a) are the first 3072 observations of a set of tremor data. The object of the study is to compare different regions of tremor activity coming from a subject with Parkinson's disease. These data have been considered by von Sachs and Neumann [99] who apply their test of stationarity over three consecutive segments of length 1024 of the first-order differenced series, shown in Figure 5.13(b). As in [99], we have added a Gaussian white noise of standard deviation 0.01 to the originial data, in order to break the discrete nature of the data.

We apply the test of homogeneity over $\mathcal{R} = (0, 1)$ scale by scale for the three segments (1), (2) and (3). The parameters for the pre-estimation of the variance are $|\mathcal{R}_T| = \log_2(1024)$ and $M_T = 2$. For each scale, we test the homogeneity between the EWS on $(0, 1)$ and on 20 subintervals. Table 5.2 reports the results. The number reported in the table is the minimum probability value obtained between the 20 tests. This value is computed using g in (5.2). (*) indicates a value less than or equal to 0.05 and (**) indicates a value less than 0.01.

The conclusion of this study is a lack of stationarity for segments (2) and (3) of the tremor data. The test of von Sachs and Neumann [99] concludes also to a lack of stationarity for series (2). However, they do not detect any change of regime in the series (3), and our conclusion seems to be a new observation. A careful inspection of the time series shows that some changes of regime indeed occur in segment (2), and also in segment (3) (around the time point 2300). This is also in accordance

Scale	(1)	(2)	(3)
-1	0.08	0.39	0.05 (*)
-2	0.35	0.44	0.46
-3	0.74	0.42	0.68
-4	0.34	0.001 (**)	0.28
-5	0.07	$3 \cdot 10^{-21}$ (**)	0.002 (**)
-6	0.54	$8 \cdot 10^{-5}$ (**)	0.24
-7	0.72	0.02 (*)	0.60
-8	0.91	0.84	0.81
-9	0.92	0.94	0.88
-10	0.98	0.98	0.97

Table 5.2: Results of the test of stationarity for the three segments taken from the tremor data. (1), (2) and (3) refer to the first, second and third segment in Figure 5.13(b). At each scale, for each segment, we perform 20 tests of homogeneity between $\mathcal{R} = (0, 1)$ and 20 subintervals $\mathcal{U}$. The number reported in the table is the minimum probability value obtained among the 20 tests. This value is computed using g in (5.2). (*) indicates a value less than or equal to 0.05 and (**) indicates a value less than 0.01.

with the findings of the neurologists who attributed two different regimes of tremor activity for this part of data. Our conclusion is that (at least) one change of regime occurs in segment (2) and also in segment (3).

The difference with the conclusion of von Sachs and Neumann [99] may be explained by Table 5.2. Indeed, the lack of stationarity for segment (3) is due to an inhomogeneity at scale -5 only (and perhaps also at scale -1). This is certainly a very subtle behaviour to be detected, and our multiscale approach succeeded to find this lack of homogeneity.

Our analysis offers a more precise interpretation of the nonstationarity of the tremor data. Moreover, we would like to recall that, unlike the test of von Sachs and Neumann [99], our approach is not limited to time series with a length equal to a power of 2. Furthermore, it is of course possible with our method to detect some intervals of homogeneity in the tremor data, such that we can say exactly where the changes of regime occur in the time series.

To illustrate this last point, let us consider the scale $j = -5$ of the CWP of tremor data. From Table 5.2, we have seen that a change of regime in the tremor data occurs at this scale in both segments 2 and 3. Figure 5.14 plots the corrected wavelet periodogram at scale $j = -5$ for the 2048 last data (i.e. for the two last segments of the data).

Figure 5.15 presents the result of the pointwise adaptive estimator from the CWP showed in Figure 5.14. In our estimation, we estimate the EWS at $K = 50$ points from time 1025 to 3042 and $g_0 = 0.95$ as in the baby heart rate case study. This estimation procedure confirms the nonstationarity of the data, because the estimated spectrum is far from being constant over time. In addition, this result shows that clear breakpoints at scale $j = -5$ occur around the time points 1880, 2130, 2500 and 2830. This information is easily obtained from our methodology, and gives valuable details that are of a great interest for the neurologists.

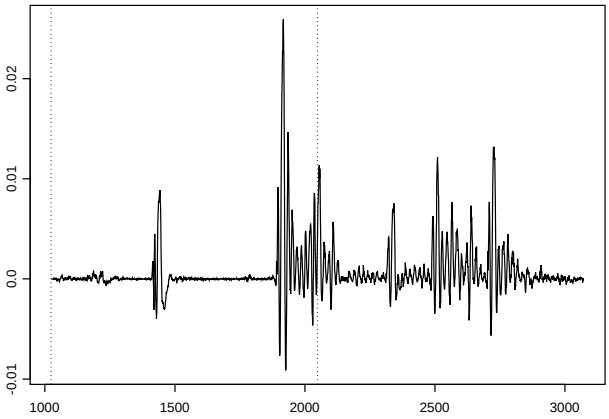


Figure 5.14: Fifth scale of the corrected wavelet periodogram of 2048 last data of the differenciated tremor data. The dotted lines indicate the change between the segments.

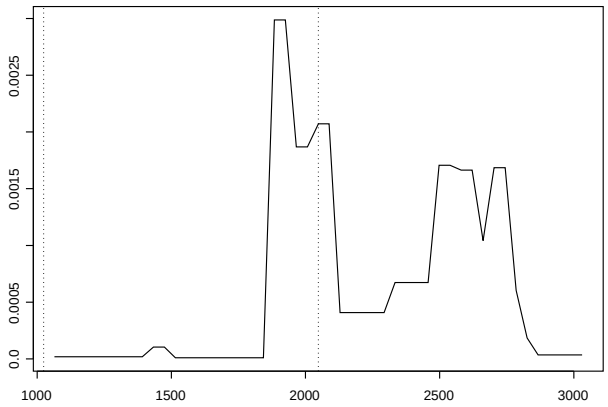


Figure 5.15: Adaptive estimation of the fifth scale of tremor data (segments 2 and 3).

CHAPTER 6

Forecasting locally stationary wavelet processes

6.1 Introduction

In this chapter, we address the problem how to use the locally stationary wavelet (LSW) model for forecasting. In Sections 6.2 and 6.3, we develop a theory of prediction for these processes. Since the LSW processes are linear with respect to their random increments, our predictor is linear. It is motivated by the approach in the stationary case, i.e. it follows from the minimisation of the mean-square prediction error (MSPE). This leads to a generalisation of the Yule-Walker equations [15], which can be solved numerically by matrix inversion or standard iterative algorithms such as the innovations algorithm [15], provided that the non-stationary covariance structure is known.

Of course, the local autocovariance function $c(z, \tau)$ needs to be estimated in order to compute the forecasts in practice. In the previous chapters, we have presented a new pointwise adaptive estimator of the evolutionary wavelet spectrum (EWS), and this estimator may be used to derive an estimator of $c(z, \tau)$, due to the equation

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

giving the relation between the local autocovariance function and the EWS $S_j(z)$. The regularity assumption on $S_j(z)$ with respect to z is fixed

by Definition 3.1, that imposes that the EWS is bounded in the total variation norm. In other words, the EWS can for instance have a finite number of jumps between 0 and 1. As a matter of fact, this assumption is very mild and probably too general for our forecasting task. In order to forecast the process, we intuitively need more continuity on $S_j(z)$ with respect to z.

Consequently, even if the prediction equations derived in Section 6.3 are valid under the general assumptions of Definition 3.1, it seems more appropriate to us to assume some continuity regularity in time for the EWS as long as forecasting is considered. That is the reason why, from Section 6.4 on, we impose a Lipschitz-continuity assumption on the EWS. This assumption is useful in order to extrapolate the EWS into the future for computing the forecasted value, similarly to the situation of Chapter 1. We then derive a new algorithm for the estimation of the local autocovariance and for the forecasting of locally stationary wavelet processes. This algorithm takes advantage of the continuity assumption that we have introduced in this section. The whole forecasting procedure is then applied in Section 6.5 on a case study in meteorology.

6.2 The prediction equations

In this section, we define and analyse the general linear predictor for non-stationary data which are modelled to follow the LSW process representation.

6.2.1 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 *h-step-ahead predictor* of $X_{t-1+h,T}$ by

$$\hat{X}_{t-1+h,T} = \sum_{s=0}^{t-1} b^{(h)}_{t-1-s;T} \, X_{s,T} \ , \tag{6.1}$$

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

$$\mathrm{MSPE}(\hat{X}_{t-1+h,T}, X_{t-1+h,T}) = \mathrm{E}\left(\hat{X}_{t-1+h,T} - X_{t-1+h,T}\right)^2 \ .$$

The predictor (6.1) is a linear combination of doubly-indexed observations where the weights need to follow the same doubly-indexed

framework. This means that as $T \to \infty$, we augment our knowledge about the local structure of the process, which allows us to fit coefficients $b^{(h)}_{t-1-s;T}$ more and more accurately. 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 (6.1) is a relation which is written in rescaled time.

The following assumption holds in the sequel of the chapter.

Assumption 6.1. If h is the prediction horizon and t is the number of observed data, then we set $T = t + h$ and we assume $h = o(T)$. $\Diamond$

With this assumption, the last observation of the LSW process is denoted by $X_{t-1,T} = X_{T-h-1,T}$, while $\hat{X}_{T-1,T}$ is the last possible forecast (h steps ahead). Consequently, as in Chapter 1 where a local variance $\sigma^2(z)$ is estimated, the evolutionary wavelet spectrum $S_j(z)$ can only be estimated on the interval

$$\left[0, 1 - \frac{h+1}{T}\right]. \tag{6.2}$$

The rescaled-time segment

$$\left(1 - \frac{h+1}{T}, 1\right) \tag{6.3}$$

accommodates the predicted values of $S_j(z)$. With Assumption 6.1, the estimation domain (6.2) asymptotically tends to $[0, 1)$ while the prediction domain (6.3) shrinks to an empty set in the rescaled time. Thus, Assumption 6.1 ensures that asymptotically, we acquire knowledge of the wavelet spectrum over the full interval $[0, 1)$.

6.2.2 Prediction in the wavelet domain

There is an interesting link between the above definition of the linear predictor (6.1) and another, "intuitive" definition of a predictor in the LSW model. For ease of presentation, let us suppose the forecasting horizon is $h = 1$, so that $T = t + 1$. 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 moving 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 $j = -1, \ldots, -J := -[\log_2 T]$ and $k \in \mathbb{Z}$. Then, the one-step-ahead predictor is constructed as

$$\hat{X}_{t,T} = \sum_{j=-J}^{-1} \sum_{k \in \mathbb{Z}} d_{jk;T} \; a^{(1)}_{jk;T} \; \psi_{jk}(t) \; , \tag{6.4}$$

where the coefficients $a^{(1)}_{jk}$ have to be estimated and are such that they minimise the MSPE. This predictor (6.4) may be viewed as a projection of $X_{t,T}$ on the space of random variables spanned by $\{d_{j,k;T} | j = -1, \ldots, -J \text{ and } k = 0, \ldots, T-1\}$.

It turns out that due to the redundancy of the non-orthogonal wavelet system $\{\psi_{jk}(t)\}$, the predictor (6.4) does not have a unique representation: there exists more than one solution $\{a^{(1)}_{jk}\}$ 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 (6.4) is related to the linear predictor (6.1) by

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

Because of the redundancy of the non-decimated wavelet system, for a fixed sequence $b^{(1)}_{t-s;T}$, there exists more than one sequence $a^{(1)}_{jk;T}$ such that this relation holds. For this reason, we prefer to work directly with the general linear predictor (6.1), bearing in mind that it can also be expressed as a (non-unique) projection onto the wavelet domain.

6.2.3 One-step ahead prediction equations

In this subsection, we consider a forecasting horizon $h = 1$ (so that $T = t + 1$) and want to minimise the mean square prediction error $\mathrm{MSPE}(\hat{X}_{t;T}, X_{t;T})$ with respect to $b^{(1)}_{t-s;T}$. 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 $(b^{(1)}_{t-1;T}, \ldots, b^{(1)}_{0;T}, -1)$ 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 (3.22) 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

$$\sum_{j=-J}^{-1} Q_{j,\mathcal{R}_T(\frac{n+m}{2T})} \Psi_j(n-m) \ ,$$

where $\mathcal{R}_T$ is an interval such that $|\mathcal{R}_T| = o_T(1)$ and $Q_{j,\mathcal{R}_T}$ is defined in (4.2) and can be estimated by estimating the (uniquely defined) wavelet spectrum S_j. We first consider the following assumptions on the evolutionary wavelet spectrum.

Assumption 6.2. The evolutionary wavelet spectrum is such that

$$\|c_X\|_{1,\infty} := \sum_{\tau=-\infty}^{\infty} \sup_z |c(z,\tau)| < \infty, \tag{6.5}$$

$$C_1 := \operatorname{ess\,inf}_{z,\omega} \sum_{j<0} S_j(z) |\hat{\psi}_j(\omega)|^2 > 0, \tag{6.6}$$

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

Note that if (6.5) holds, then

$$C_2 := \operatorname{ess\,sup}_{z,\omega} \sum_{j<0} S_j(z) |\hat{\psi}_j(\omega)|^2 < \infty. \tag{6.7}$$

Assumption (6.5) ensures that for each z, the local covariance $c(z,\tau)$ is absolutely summable, so the process is short-memory. This assumption is also present in Chapter 4 (Assumption 4.1). Assumption (6.6) and formula (6.7) become more transparent when we recall that for a stationary process X_t with spectral density $f(\omega)$ and wavelet spectrum S_j, we have $f(\omega) = \sum_j S_j |\hat{\psi}_j(\omega)|^2$ (the Fourier transform of equation (3.27) for stationary processes). In this sense, (6.6) and (6.7) are "time-varying" counterparts of the classical assumptions of the (stationary) spectral density being bounded away from zero, as well as bounded from above.

Proposition 6.1. *Under Assumptions (6.5) and (6.6), the mean square one-step-ahead prediction error may be written as*

$$\mathrm{MSPE}(\hat{X}_{t;T}, X_{t;T}) = \boldsymbol{b}_t' \boldsymbol{B}_{t;T} \boldsymbol{b}_t \ (1 + o_T(1)) \ . \tag{6.8}$$

Moreover, if $\{b^{(1)}_{s;T}\}$ *are the coefficients which minimise* $\boldsymbol{b}_t' \boldsymbol{B}_{t;T} \boldsymbol{b}_t$*, then* $\{b^{(1)}_{s;T}\}$ *solve the following linear system*

$$\sum_{m=0}^{t-1} b^{(1)}_{t-1-m;T} \left\{ \sum_{j=-J}^{-1} Q_{j,\mathcal{R}_T(\frac{n+m}{2T})} \Psi_j(m-n) \right\} = \sum_{j=-J}^{-1} Q_{j,\mathcal{R}_T(\frac{t+n}{2T})} \Psi_j(t-n) \tag{6.9}$$

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

The proof of the first result can be found in the next section (see Lemma 6.5) and uses standard approximations of covariance matrices of locally stationary processes. The second result is simply the minimisation of the quadratic form (6.8) and the system of equations (6.9) 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 the relation of formula (3.27) between the wavelet spectrum and the local autocovariance function, the prediction equations can also be written as

$$\sum_{m=0}^{t-1} b^{(1)}_{t-1-m;T} \, c\left(\frac{n+m}{2T}, m-n\right) = c\left(\frac{n+t}{2T}, t-n\right). \tag{6.10}$$

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

Remark 6.1 (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 (6.10) become

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

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

Remark 6.2 (Time-modulated processes). For the processes considered in Chapter 1 (equation (1.1)), the local autocovariance function has a multiplicative structure: $c(z,\tau) = \sigma^2(z)\rho_Y(\tau)$. Therefore, for these processes, prediction equations (6.10) become

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

We will now study the inversion of the system (6.9) 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 [56, Theorem 5.3]) the measure of this stability is given by the so-called *condition number*, which is defined by $\mathrm{cond}\,(\boldsymbol{P}_t) = \|\boldsymbol{P}_t\|\,\|\boldsymbol{P}_t^{-1}\|$. It can be proved along the lines of Lemma 6.3 below that, under Assumptions (6.5) and (6.6), $\mathrm{cond}\,(\boldsymbol{P}_t) \leqslant C_1\,C_2$.

6.2.4 The prediction error

The next result generalises the classical Kolmogorov formula for the theoretical one-step-ahead prediction error (Theorem 5.8.1 in Brockwell and Davis [15]). It is a direct modification of Theorem 3.2(i) of Dahlhaus [22], who states a similar result for locally stationary Fourier processes.

Proposition 6.2. *Suppose that Assumptions (6.5) and (6.6) hold true. Given t observations $X_{0,T}, \ldots, X_{t-1,T}$ of the LSW process $\{X_{t,T}\}$ (with $T = t+1$), the one-step ahead mean square prediction error σ^2_{OSPE} in forecasting $\hat{X}_{t,T}$ is given by*

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

Proof. The proof uses Lemmas 6.1 to 6.3 below and is along the lines of the proof of Theorem 3.2(i) of Dahlhaus [22]. The idea is to reduce the problem to a 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 an absolutely summable autocovariance and with Fourier spectrum $f(\cdot)$, then its wavelet spectrum is given by

$$S_j = \sum_{\ell} A^{-1}_{j\ell} \int d\lambda \; f(\lambda) |\hat{\psi}_\ell(\lambda)|^2 \tag{6.11}$$

for any fixed non-decimated system of compactly supported wavelets $\{\psi_{jk}\}$. We refer to [22] for details. □

Note that due to Assumption (6.6), the sum $\sum_j Q_{j,\mathcal{R}_T(t/T)} |\hat{\psi}_j(\omega)|^2$ is strictly positive, except possibly on a set of measure zero.

6.2.5 h-step-ahead prediction

The one-step-ahead prediction equations have a natural generalisation to the h-step-ahead prediction problem with $h > 1$. The mean square prediction error can be written as

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

where $\boldsymbol{\Sigma}_{t+h-1;T}$ is the covariance matrix of $X_{0,T}, \ldots, X_{t+h-1,T}$ and the vector $\boldsymbol{b}_{t+h-1}$ is given by $(b^{(h)}_{t-1}, \ldots, b^{(h)}_0, b^{(h)}_{-1}, \ldots, b^{(h)}_{-h})$, and is such that $b^{(h)}_{-1}, \ldots, b^{(h)}_{-h+1} = 0$ and $b^{(h)}_{-h} = -1$. 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

$$\sum_{j=-J}^{-1} Q_{j,\mathcal{R}_T\left(\frac{n+m}{2T}\right)} \; \Psi_j(n-m) \; .$$

Proposition 6.3. *Under Assumptions (6.5) and (6.6), the mean square prediction error may be written as*

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

The proof of this Proposition is to be found in the next section.

6.3 Theoretical properties of the predictor

Let $\boldsymbol{X}_{t;T} = (X_{0;T}, \ldots, X_{t-1;T})'$ be a realisation of an LSW process. In this section, we study the theoretical properties of the covariance matrix $\boldsymbol{\Sigma}_{t;T} = \mathrm{E}(\boldsymbol{X}_{t;T}\boldsymbol{X}'_{t;T})$. As we need upper bounds for the spectral norms $\|\boldsymbol{\Sigma}_{t;T}\|$ and $\|\boldsymbol{\Sigma}_{t;T}^{-1}\|$, we base the following results and their proofs on methods developed in Section 4 of Dahlhaus [22] for approximating covariance matrices of locally stationary Fourier processes. However, in our setting these methods need important modifications. A first reason is since we are working in the wavelet domain. A second reason is that we are dealing with non smooth evolutionary spectra, which differs from Dahlhaus [22], who considers time-varying spectrum which are Lipschitz-continuous in time only. The first idea of the proof 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:

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

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

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

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:

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

and, analogously, define $\mathcal{P}_3, \mathcal{P}_4, \ldots$ up to $\mathcal{P}_M$ where $M = (L/\delta) - 1$. Consider the union of all these partitions $\mathcal{P} = \{\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_M\}$, 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 M_p, $p = 1, \ldots, P$.

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

$$D^{(p)}_{nm} = \sum_{j<0} Q_{j,\mathcal{R}_T(\nu_p)} \Psi_j(n-m) \mathbb{I}_{n,m \in M_p}$$

where $\mathbb{I}_{n,m\in M_p}$ means that we only include those n, m that are in M_p, and the interval $\mathcal{R}_T$ is such that $|\mathcal{R}_T| = o_T(1)$. Observe that each ν_p is contained exactly in L/δ segments. 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 6.1. *Assume that (6.5) holds. If $L \to \infty$, $\delta/L \to 0$ and $L^2/T \to 0$ as $T \to \infty$, then*

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

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

$$\boldsymbol{x}' \left(\boldsymbol{\Sigma}_{t;T} - \boldsymbol{D}\right) \boldsymbol{x} = \boldsymbol{x}' \left\{ \frac{\delta}{L} \sum_{p=1}^{P} (\boldsymbol{\Sigma}^{(p)}_{t;T} - \boldsymbol{D}^{(p)}) \right\} \boldsymbol{x} + \mathrm{Rest}_T \qquad (6.12)$$

where

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

Let us first bound this remainder. Replace $\left(\Sigma_{t;T}\right)_{nm}$ by

$$\sum_j Q_{j,\mathcal{R}_T(n-m)} \Psi_j(n-m)$$

and denote $b(k) := \sup_z |\sum_j Q_{j,\mathcal{R}_T(z)}\Psi_j(k)| = \sup_z |c(z,k)|$. We have

$$
\begin{aligned}
|\operatorname{Rest}_T| &\leqslant 2\boldsymbol{x}'\boldsymbol{x} \sum_{d=1}^{\frac{T}{\delta}-1} \min\left(d\frac{\delta}{L}, 1\right) \sum_{k=(d-1)\delta+1}^{d\delta} b(k) + \operatorname{Rest}'_T \\
&\leqslant 2\boldsymbol{x}'\boldsymbol{x} \left(\frac{\delta+\sqrt{L}}{L} \sum_{k=1}^{\infty} b(k) + \sum_{k>\sqrt{L}} b(k) \right) + \operatorname{Rest}'_T
\end{aligned}
$$

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 (6.5). Let us now turn to the remainder Rest'_T. We have

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

which may be bounded by

$$
|\mathcal{R}_T|^{-1} \int_{\mathcal{R}_T(n-m)} dz \sum_{n,m} \sum_{j,k} |x_n x_m| \cdot \left| w^2_{jk;T} - S_j(z) \right| \cdot |\psi_{jk}(n)\psi_{jk}(m)|.
$$

Using the decomposition

$$
\begin{aligned}
\left| w^2_{jk;T} - S_j(z) \right| \leqslant \frac{C_j}{T} &+ \left| S_j\left(z + \frac{k-n}{T} \right) - S_j(z) \right| \\
&+ \left| S_j\left(z + \frac{k-n}{T} \right) - S_j\left(\frac{k}{T} \right) \right|,
\end{aligned}
$$

we get three terms that we will bound from above separately. Following the proofs developed in the above Chapter 4, it is an easy task to show that the first term, that is the term containing C_j/T, is $O(\boldsymbol{x}'\boldsymbol{x}T^{-1})$. The second term leads to

$$
\begin{aligned}
&\sum_{j,k} \sum_{n,m} |x_n x_m| \cdot |\psi_{jk}(n)\psi_{jk}(m)| \times \\
&\qquad \times |\mathcal{R}_T| \sum_{\ell=0}^{|\mathcal{R}_T T|-1} \int_{z_0+\ell/T}^{z_0+(\ell+1)/T} dz \left| S_j\left(z + \frac{k-n}{T} \right) - S_j(z) \right|
\end{aligned}
$$

where z_0 denotes the left point of the interval $\mathcal{R}_T(\nu_p)$. With the change of variables $y = z - \ell/T$, this last bound is equal to

$$\sum_{j,k} \sum_{n,m} |x_n x_m| \cdot |\psi_{jk}(n)\psi_{jk}(m)| \times \\ \times |\mathcal{R}_T| \sum_{\ell=0}^{|\mathcal{R}_T T|-1} \int_{z_0}^{z_0+1/T} dy \left| S_j\left(y + \frac{\ell + k - n}{T}\right) - S_j\left(y + \frac{\ell}{T}\right)\right|.$$

We bound this last quantity using the total variation constraint on the EWS. This allows to replace the sum over ℓ of the integral by $|k - n| \operatorname{TV}(S_j) \leqslant N_j L_j$ since the sum over n goes from n to $n + N_j$. Then we get the upper bound

$$|\mathcal{R}_T T|^{-1} \sum_{jk} N_j L_j \left(\sum_n x_n \psi_{jk}(n)\right)^2 = O(\boldsymbol{x}'\boldsymbol{x}|\mathcal{R}_T T|^{-1}).$$

The last term is bounded using the change of variables $y = z - (n-m)/T$, which leads to the upper bound

$$|\mathcal{R}_T|^{-1} \int_{\mathcal{R}_T(0)} dy \sum_{m,n} |x_n x_m| \sum_{jk} \left| S_j\left(y + \frac{k-n}{T}\right) - S_j\left(\frac{k}{T}\right)\right| \times \\ \times |\psi_{jk}(n)\psi_{jk}(m)|$$

An application of the Cauchy-Schwarz inequality for the sums over n and m leads to the product of two terms like $\sum_n x_n \psi_{jk}(n)$ which is bounded by $\sqrt{\boldsymbol{x}'\boldsymbol{x}}$. Then, using the total variation constraint on the EWS derived from the sum over k, we get the upper bound

$$|\mathcal{R}_T|^{-1} \boldsymbol{x}'\boldsymbol{x} \operatorname{TV}(S_j) \int_{\mathcal{R}_T(0)} dy\, |y| = O(\boldsymbol{x}'\boldsymbol{x}|\mathcal{R}_T|)\,.$$

Then, $\operatorname{Rest}'_T = O(\boldsymbol{x}'\boldsymbol{x}|\mathcal{R}_T|)$.

Finally, we can consider the main term in (6.12). We have

$$\begin{aligned}
&\boldsymbol{x}' \left(\frac{\delta}{L} \sum_{p=1}^{P} \boldsymbol{\Sigma}_{t;T}^{(p)} - \boldsymbol{D}^{(p)} \right) \boldsymbol{x} \\
&\quad \leqslant \frac{\delta}{L} \sum_{p=1}^{P} \sum_{jk} \left| w_{jk;T}^2 - Q_{j,\mathcal{R}_T(\nu_p)} \right| \left(\sum_u \psi_{jk}(u) x_u \mathbb{I}_{u \in M_p} \right)^2 \\
&\quad \leqslant \frac{\delta}{L} \sum_{p=1}^{P} \sum_{jk} \frac{1}{|\mathcal{R}_T|} \int_{\mathcal{R}_T(\nu_p)} \left| w_{jk;T}^2 - S_j(z) \right| \times \\
&\qquad\qquad \times \left(\sum_u \psi_{jk}(u) x_u \mathbb{I}_{u \in M_p} \right)^2
\end{aligned}$$

and use the decomposition

$$\begin{aligned}
\left| w_{jk;T}^2 - S_j(z) \right| \leqslant \frac{C_j}{T} &+ \left| S_j\left(\frac{k}{T}\right) - S_j\left(z + \frac{k - \nu_p}{T}\right) \right| \\
&+ \left| S_j\left(z\right) - S_j\left(z + \frac{k - \nu_p}{T}\right) \right|
\end{aligned}$$

for each $z \in \mathcal{R}_T(\nu_p)$, and we get three terms $\mathrm{I}_T + \mathrm{II}_T + \mathrm{III}_T$ that we bound separately. Again, we easily show that the term I_T, containing C_j/T, is $O(\boldsymbol{x}'\boldsymbol{x}/T)$. To bound the second term, II_T, we first note that, by the Cauchy-Schwarz inequality and using (3.11), $(\sum_{u \in M_p} x_u \psi_{jk}(u))^2$ is bounded by $\sum_{u \in M_p} x_u^2$, and this upper bound does no longer depend on k. Consequently, the sum over k together with the total variation constraint on the EWS leads to the bound

$$\mathrm{II}_T \leqslant \frac{\delta}{L} \sum_{p=1}^{P} \int_{\mathcal{R}_T(\nu_p)} dz \sum_j \mathrm{TV}(S_j)\, |\nu_p - z| \sum_{n \in M_p} x_n^2.$$

By construction, each x_n is contained in exactly L/δ segments of the coverage. Then, with (3.25), we get $\mathrm{II}_T \leqslant \rho \boldsymbol{x}'\boldsymbol{x} \int_{\mathcal{R}_T(0)} dz\ |z|$ and we

obtain $\mathrm{II}_T = \boldsymbol{x}'\boldsymbol{x} o_T(1)$. The term III_T is:

$$\frac{\delta}{L}\sum_{p=1}^{P}\int_{\mathcal{R}_T(\nu_p)} dz \sum_{jk}\left|S_j\left(z\right) - S_j\left(z+\frac{k}{T}-\nu_p\right)\right| \left(\sum_{n\in M_p} x_n\psi_{jk}(n)\right)^2 .$$

Let us denote by z_0 the left point of the interval $\mathcal{R}_T(\nu_p)$ (note that z_0 depends on p). Deriving like above, we get

$$\mathrm{III}_T \leqslant \frac{\delta}{L}\sum_{p=1}^{P}\sum_{\ell=0}^{|\mathcal{R}_T|-1}\int_{z_0}^{z_0+1/T} dz \sum_{jk}\left(\sum_{n\in M_p} x_n\psi_{jk}(n)\right)^2 \times \times\left|S_j\left(z+\frac{\ell}{T}\right) - S_j\left(z+\frac{k+\ell-\nu_p}{T}\right)\right|.$$

The sum over ℓ leads to $L_j|k-\nu_p|$ which is bounded by $L_j(N_j+L)$ due to the compact support of the wavelets. We finally get

$$\begin{aligned}\mathrm{III}_T &\leqslant \frac{\delta}{TL}\sum_{p}\sum_{jk} L_j(N_j+L)\left(\sum_{n\in M_p} x_n\psi_{jk}(n)\right)^2 \\ &= O(T^{-1})\boldsymbol{x}'\boldsymbol{x}\sum_j (C_j + L_j(N_j+L))(N_j+L)\end{aligned}$$

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. □

Lemma 6.2. *Assume that (6.5) holds and there exists a t^* such that $x_u = 0$ for all $u \notin \{t^*,\dots,t^*+L\}$. Then for each $t_0 \in \{t^*,\dots,t^*+L\}$,*

$$\boldsymbol{x}'\Sigma_{t,T}\boldsymbol{x} = \sum_j Q_{j,\mathcal{R}_T(t_0)}\sum_k\left(\sum_{u=t^*}^{t^*+L} x_u\psi_{j,k}(u)\right)^2 + \boldsymbol{x}'\boldsymbol{x}O\left(\frac{L^2}{T}\right). \quad (6.13)$$

Proof. Identical to the part of the proof of Lemma 6.1. □

In what follows, the matrix norm $\|\boldsymbol{M}\|_{\mathrm{spec}}$ denotes the spectral norm of the matrix $\boldsymbol{M}$ (cf. Appendix A).

Lemma 6.3. *Assume that (6.5) holds. The spectral norm* $\|\mathbf{\Sigma}_{t;T}\|_{\text{spec}}$ *is bounded in* t*. Also, if (6.6) holds, then the spectral norm* $\|\mathbf{\Sigma}_{t;T}^{-1}\|_{\text{spec}}$ *is bounded in* t*.*

Proof. Lemma 6.1 implies

$$\|\mathbf{\Sigma}_{t;T}\|_{\text{spec}} = \sup_{\|\boldsymbol{x}\|_2^2=1} \frac{\delta}{L} \sum_{p=1}^{P} \sum_{j<0} Q_{j,\mathcal{R}_T(\nu_p)} \sum_k \left(\sum_{n \in M_p} x_n \psi_{j,k-n} \right)^2 + o_T(1)$$

Using Parseval formula, $\|\Sigma\|_{\text{spec}}$ is equal to

$$\sup_{\|\boldsymbol{x}\|_2^2=1} \frac{\delta}{2\pi L} \sum_{p=1}^{P} \int_{-\pi}^{\pi} d\omega \sum_{j<0} Q_{j,\mathcal{R}_T(\nu_p)} \left| \hat{\psi}_j(\omega) \right|^2 \times \\ \times \left| \sum_n x_n \exp(-i\omega n) \mathbb{I}_{n \in M_p} \right|^2 + o_T(1)$$

which may be bounded from above by

$$\begin{aligned} &\operatorname*{ess\,sup}_{z,\omega} \sum_j S_j(z) \left| \hat{\psi}_j(\omega) \right|^2 \sup_{\|\boldsymbol{x}\|_2^2=1} \|\boldsymbol{x}\|_2^2 + o_T(1) \\ &\quad = \operatorname*{ess\,sup}_{z,\omega} \sum_j S_j(z) \left| \hat{\psi}_j(\omega) \right|^2 + o_T(1) \end{aligned}$$

which is bounded by (6.5) (as (6.5) implies (6.7)). Using (A.1) with $\boldsymbol{M} = \mathbf{\Sigma}_{t;T}$, the boundedness of $\|\mathbf{\Sigma}_{t;T}^{-1}\|$ is shown in exactly the same way. □

We will now study the approximation of $\mathbf{\Sigma}_{t;T}$ by $\boldsymbol{B}_{t;T}$.

Lemma 6.4. *Under the assumptions of Proposition 6.1 and 6.3,*

$$\begin{aligned} &\text{MSPE}(\hat{X}_{t+h-1;T}, X_{t+h-1;T}) \\ &\qquad = \boldsymbol{b}'_{t+h-1} \boldsymbol{B}_{t+h-1;T} \boldsymbol{b}_{t+h-1} + \boldsymbol{b}'_{t+h-1} \boldsymbol{b}_{t+h-1} \, o_T(1) \end{aligned}$$

and, in particular,

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

Proof. Write

$$\boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1} = \sum_{jk}\sum_{n,m=0}^{t+h-1} b_n b_m \psi_{jk}(n)\psi_{jk}(m)|w_{jk;T}|^2$$
$$= \sum_{jk}\sum_{n,m=0}^{t+h-1} b_n b_m \Psi_j(n-m) Q_{j,\mathcal{R}_T(\frac{n+m}{2T})} + \mathrm{Rest}_1 \qquad (6.14)$$

where Rest_1 is such that

$$|\,\mathrm{Rest}_1\,| \leqslant \sum_{jk}\sum_{n,m=0}^{t+h-1} b_n b_m \psi_{jk}(n)\psi_{jk}(m)\left(|w_{jk;T}|^2 - Q_{j,\mathcal{R}_T(\frac{n+m}{2T})}\right)$$
$$\leqslant |\mathcal{R}_T|^{-1}\int_{\mathcal{R}_T(\frac{n+m}{2T})} dz \sum_{jk}\sum_{n,m=0}^{t+h-1} b_n b_m \psi_{jk}(n)\psi_{jk}(m)\times$$
$$\times\left(|w_{jk;T}|^2 - S_j(z)\right).$$

Now, we use again the decomposition

$$|w_{jk;T}|^2 - S_j(z) \leqslant \frac{C_j}{T} + \left|S_j(k) - S_j\left(z - \frac{n+m}{2T}\right)\right|$$
$$+ \left|S_j(z) - S_j\left(z - \frac{n+m}{2T}\right)\right|$$

and bound each term as in the above proofs, using Assumption (6.5).□

Lemma 6.5. *Under the assumptions of Proposition 6.3, we have*

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

Proof of Lemma 6.5. By Lemma 6.4, we have $\boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1} = \boldsymbol{b}'_{t+h-1}\boldsymbol{B}_{t+h-1;T}\boldsymbol{b}_{t+h-1} + \boldsymbol{b}'_{t+h-1}\boldsymbol{b}_{t+h-1}\; o_T(1)$ By Lemma 6.3, 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

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

which is equivalent to

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

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. □

6.4 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 (6.9). One possibility is to use the previous pointwise adaptive estimation to solve this problem. In the following, we present another way to solve this problem. For that, we will be more specific about the regularity of the EWS in time. Then, we introduce some new estimators of the local autocovariance function. Finally, we propose a complete algorithm for forecasting non-stationary time series. This algorithm allows to select adaptively all the parameters needed for the estimation and for the forecasting.

6.4.1 LSW model under the Lipschitz constraint

From now on, we assume the EWS of the spectrum to be Lipschitz-continuous in time. More precisely, we replace the constraint (3.24) in Definition 3.1 by the folllowing assumption.

Assumption 6.3. For each scale $j < 0$, the evolutionary wavelet spectrum $S_j(z)$ is Lipschitz-continuous on $(0, 1)$ with Lipschitz constants L_j. ♢

With this assumption, we assume that the EWS is smoothly varying in time. Under this assumption, the prediction equations can be written in a simpler form. The proof of result, stated in the next proposition, is a straightforward adaptation of the previous proof. A detailed proof may be found in [39].

Proposition 6.4. *Under Assumption 6.3 and if (6.5) and (6.6) hold, then the coefficients* $\{b^{(1)}_{s;T}\}$ *which minimise* $\boldsymbol{b}_t' \boldsymbol{B}_{t;T} \boldsymbol{b}_t$ *in Proposition 6.1, are such that*

$$\sum_{m=0}^{t-1} b^{(1)}_{t-1-m;T} \left\{ \sum_{j=-J}^{-1} S_j \left(\frac{n+m}{2T} \right) \Psi_j(m-n) \right\} = \sum_{j=-J}^{-1} S_j \left(\frac{t+n}{2T} \right) \Psi_j(t-n) \quad (6.15)$$

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

In the following, we develop an algorithm which exploits the new assumption and allows to select all the parameters of the forecasting in a data-driven way. We first need to introduce an estimator of the local autocovariance function.

6.4.2 Estimation of the time-varying second-order structure

Our estimator of the local autocovariance function $c(z, \tau)$, with $0 < z < t/T$, is constructed by estimating the unknown wavelet spectrum $S_j(z)$ in the multiscale representation (3.27). Let us first define the function $J(t) = -\min\{j : N_j \leqslant t\}$.

We define the multiscale estimator of the local variance function (3.27) as

$$\tilde{c}\left(\frac{k}{T}, 0\right) = \sum_{j=-J}^{-1} 2^j \; I_j\left(\frac{k}{T}\right). \quad (6.16)$$

where I is the periodogram defined in Section 3.5. The next proposition concerns the asymptotic behaviour of the first two moments of this estimator.

Proposition 6.5. *The estimator (6.16) satisfies*

$$\mathrm{E}\,\tilde{c}\left(\frac{k}{T},0\right) = c\left(\frac{k}{T},0\right) + O\left(T^{-1}\log(T)\right).$$

If, in addition, the increment process $\{\xi_{jk}\}$ in Definition 3.1 is Gaussian and (6.5) holds, then

$$\begin{aligned}&\mathrm{Var}\;\tilde{c}\left(\frac{k}{T},0\right)\\ &\quad= 2\sum_{i,j=-J}^{-1} 2^{i+j}\left(\sum_{\tau} c(k/T,\tau)\sum_{n}\psi_{in}(\tau)\psi_{jn}(0)\right)^2 + O(T^{-1}).\end{aligned}$$

Proof. We will first show

$$\begin{aligned}&\mathrm{Cov}\left(\sum_s X_{s,T}\psi_{ik}(s), \sum_s X_{s,T}\psi_{jk}(s)\right)\\ &\quad= \sum_{\tau} c\left(\frac{k}{T},\tau\right)\sum_n \psi_{in}(\tau)\psi_{jn}(0) + O(2^{-(i+j)/2}T^{-1}). \quad (6.17)\end{aligned}$$

We have

$$\begin{aligned}&\mathrm{Cov}\left(\sum_s X_{s,T}\psi_{i,k}(s), \sum_s X_{s,T}\psi_{j,k}(s)\right)\\ &\qquad= \sum_{\ell,u}\left\{S_\ell\left(\frac{k}{T}\right) + O\left(\frac{C_\ell + L_\ell(u-k)}{T}\right)\right\}\\ &\qquad\qquad \sum_{s,t}\psi_{\ell s}(u)\psi_{jk}(s)\psi_{\ell t}(u)\psi_{ik}(t).\end{aligned}$$

Using $N_j = O(M2^{-j})$ in the first step, and the Cauchy inequality in the

second one, we bound the reminder as follows:

$$\begin{aligned}
&\left|\sum_{\ell,u} O\left(\frac{C_\ell + L_\ell(u-k)}{T}\right)\sum_{s,t}\psi_{\ell s}(u)\psi_{jk}(s)\psi_{\ell t}(u)\psi_{ik}(t)\right| \\
&\quad\leqslant \sum_\ell \frac{C_\ell + ML_\ell(2^{-\ell}+\min(2^{-i},2^{-j}))}{T}\times \\
&\qquad\times\sum_u\left|\sum_{s,t}\psi_{\ell s}(u)\psi_{jk}(s)\psi_{\ell t}(u)\psi_{ik}(t)\right| \\
&\quad\leqslant \sum_\ell \frac{C_\ell + ML_\ell(2^{-\ell}+2^{-i/2}2^{-j/2})}{T}(A_{\ell j})^{1/2}(A_{\ell i})^{1/2} \\
&\quad= \frac{2^{-(i+j)/2}}{T}\sum_\ell (C_\ell + ML_\ell 2^{-\ell})2^{(i+j)/2}(A_{\ell j})^{1/2}(A_{\ell i})^{1/2} \\
&\qquad+\frac{2^{-(i+j)/2}}{T}\sum_\ell ML_\ell(A_{\ell j})^{1/2}(A_{\ell i})^{1/2} \\
&\quad= \frac{2^{-(i+j)/2}}{T}\{\mathrm{I}+\mathrm{II}\}.
\end{aligned}$$

By formula (3.18),

$$\begin{aligned}
\mathrm{I} &\leqslant \sum_\ell (C_\ell + ML_\ell 2^{-\ell})(2^i A_{\ell i} + 2^j A_{\ell j}) \\
&\leqslant \sum_\ell (C_\ell + ML_\ell 2^{-\ell})2\sum_i 2^i A_{\ell i} \\
&\leqslant D_1.
\end{aligned}$$

As $\sum_i L_i 2^{-i} < \infty$, we must have $L_i \leq C2^i$ so $\sum_i L_i A_{ij} \leqslant C$ again by (3.18). This and the Cauchy inequality give

$$\mathrm{II} \leqslant 2M\left(\sum_\ell L_\ell A_{\ell i}\right)^{1/2}\left(\sum_\ell L_\ell A_{\ell j}\right)^{1/2} \leqslant D_2.$$

The bound for the reminder is therefore $O(2^{-(i+j)/2}T^{-1})$. For the main

term, straightforward computation gives

$$\sum_{\ell,u} S_\ell\left(\frac{k}{T}\right) \sum_{s,t} \psi_{\ell s}(u)\psi_{jk}(s)\psi_{\ell t}(u)\psi_{ik}(t) = \sum_{\tau} c(k/T,\tau) \sum_{n} \psi_{in}(\tau)\psi_{jn}(0),$$

which yields formula (6.17). Using Lemma 3.3 and (6.17) with $i = j$, we obtain

$$\begin{aligned} \mathrm{E}(\tilde{c}(k/T,0)) &= \sum_{j=-J}^{-1} 2^j \left\{ \sum_{\tau} c(k/T,\tau)\Psi_j(\tau) + O(2^{-j}/T) \right\} \\ &= \sum_{\tau} c(k/T,\tau)\delta_0(\tau) + O(\log(T)/T) \\ &= c(k/T,0) + O(\log(T)/T), \end{aligned}$$

which proves the expectation. For the variance, observe that, using Gaussianity, we have

$$\begin{aligned} &\mathrm{Cov}\left(I_i\left(\frac{k}{T}\right), I_j\left(\frac{k}{T}\right)\right) \\ &\quad = 2\left(\sum_{\tau} c(k/T,\tau) \sum_{n} \psi_{in}(\tau)\psi_{jn}(0) + O(2^{-(i+j)/2}T^{-1})\right)^2 \\ &\quad = 2\left(\sum_{\tau} c(k/T,\tau) \sum_{n} \psi_{in}(\tau)\psi_{jn}(0)\right)^2 + O(2^{-(i+j)/2}T^{-1}), \end{aligned} \tag{6.18}$$

provided that (6.5) holds. Using (6.18), we finally obtain

$$\begin{aligned} &\mathrm{Var}(\tilde{c}(k/T,0)) = \\ &\quad 2\sum_{i,j=-J}^{-1} 2^{i+j}\left(\sum_{\tau} c(k/T,\tau) \sum_{n} \psi_{in}(\tau)\psi_{jn}(0)\right)^2 + O(T^{-1}). \quad \square \end{aligned} \tag{6.19}$$

Remark 6.3 (Time-modulated processes). For Gaussian time-modulated processes with a Lipschitz-continuous time-modulated variance, the variance of estimator (6.16) reduces to

$$\begin{aligned}&\text{Var } \tilde{c}\left(\frac{k}{T},0\right)\\&= 2\sigma^4(k/T)\sum_{i,j=-J}^{-1} 2^{i+j}\left(\sum_{\tau}\rho(\tau)\sum_{n}\psi_{in}(\tau)\psi_{jn}(0)\right)^2 + O(T^{-1}),\end{aligned} \tag{6.20}$$

where $\rho_Y(\tau)$ is the autocorrelation function of Y_t (see equation (1.1)). If $X_{t,T} = \sigma(t/T)Z_t$, where Z_t are i.i.d. $N(0,1)$, then the leading term in (6.20) reduces to $(2/3)\sigma^4(k/T)$ for all compactly supported wavelets ψ. ◊

Remark 6.4. Proposition 6.5 can be generalised for the estimation of $c(z,\tau)$ for $\tau \neq 0$. Define the estimator

$$\tilde{c}\left(\frac{k}{T},\tau\right) = \sum_{j=-J}^{-1}\left(\sum_{\ell=-J}^{-1} A^{-1}_{j\ell}\Psi_\ell(\tau)\right) I_j\left(\frac{k}{T}\right)\ , \tag{6.21}$$

for $k = 0,\ldots,t-1$ and $\tau \neq 0$. Using Lemma 3.4, it is possible to generalise the proof of Proposition 6.5 for Haar wavelets and to show that

$$\text{E } \tilde{c}\left(\frac{k}{T},\tau\right) = c\left(\frac{k}{T},\tau\right) + O\left(T^{-1/2}\right)$$

for $\tau \neq 0$ and, if Assumption (6.5) holds and if the increment process $\{\xi_{jk}\}$ in Definition 3.1 is Gaussian, then

$$\begin{aligned}&\text{Var } \tilde{c}\left(\frac{k}{T},\tau\right)\\&\qquad= 2\sum_{i,j=-J}^{-1} h_i(\tau)h_j(\tau)\left\{\sum_{\tau} c\left(\frac{k}{T},\tau\right)\sum_{n}\psi_{in}(\tau)\psi_{jn}(0)\right\}^2\\&\qquad\qquad + O\left(T^{-1}\log^2(T)\right)\end{aligned}$$

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

These results show the inconsistency of the estimator of the local (co)variance, which needs to be smoothed w.r.t. the rescaled time z (i.e. $\tilde{c}(\cdot, \tau)$ needs to be smoothed for all τ). We use standard kernel smoothing where the problem of the choice of the bandwidth parameter g arises. The goal of Subsection 6.4.4 is to provide a fully automatic procedure for choosing g.

To compute the linear predictor in practice, we invert the generalised Yule-Walker equations (6.15) in which the theoretical local autocovariance function is replaced by the smoothed version of $\tilde{c}(k/T, \tau)$. However, in equations (6.16) and (6.21), our estimator is only defined for $k = 0, \ldots, t-1$ while the prediction equations (6.15) require the local autocovariance up to $k = t$ (for $h = 1$). This problem is inherent to our non-stationary framework. We denote the predictor of $c(t/T, \tau)$ by $\hat{c}(t/T, \tau)$ and, motivated by the slow evolution of the local autocovariance function, 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 gT, where g is the smoothing parameter (bandwidth) for estimation. They can be obtained by the data-driven procedure described in Subsection 6.4.4.

6.4.3 Future observations in rescaled time

For clarity of presentation, we restrict ourselves (in this and the following subsection) to the case $h = 1$.

In Chapter 1, we explained the mechanics of rescaled time for non-stationary processes (Remark 1.2). An important ingredient of this concept is that the data come in the form of a triangular array whose rows correspond to *different* stochastic processes, only linked through the asymptotic wavelet spectrum sampled on a finer and finer grid. This mechanism is inherently different to what we observe in practice, where, typically, observations arrive one by one and neither the values of the "old" observations, nor their corresponding second-order structure, change when a new observation arrives.

One way to reconcile the practical setup with our theory is to assume that for an observed process $X_0, \ldots, X_{t-1}$, there exists a doubly-indexed LSW process $\mathbf{Y}$ such that $X_k = Y_{k,T}$ for $k = 0, \ldots, t-1$. When a new observation X_t arrives, the underlying LSW process changes, i.e. there exists another LSW process $\mathbf{Z}$ such that $X_k = Z_{k,T+1}$ for $k =$

$0, \ldots, t$. An essential point underlying our adaptive algorithm of the next subsection is that the spectra of $\mathbf{Y}$ and $\mathbf{Z}$ are close to each other, due to the above construction and the regularity assumptions imposed by Definition 1 (in particular, the Lipschitz continuity of $S_j(z)$).

The objective of our algorithm is to choose appropriate values of certain nuisance parameters (see the next subsection) in order to forecast X_t from $X_0, \ldots, X_{t-1}$. Assume that these parameters have been selected well, i.e. that the forecasting has been successful. The closeness of the two spectra implies that we can also expect to successfully forecast X_{t+1} from $X_0, \ldots, X_t$ using the same, or possibly "neighbouring", values of the nuisance parameters.

Bearing in mind the above discussion, we introduce our algorithm with a slight abuse of notation: we drop the second subscript when referring to the observed time series.

6.4.4 Data-driven choice of parameters

In theory, the best one-step-ahead linear predictor of $X_{t,T}$ is given by (6.1), where $\boldsymbol{b}_t = (b^{(1)}_{t-1-s;T})_{s=0,\ldots,t-1}$ solves the prediction equations (6.9). In practice, each of the t components of the vector $\boldsymbol{b}_t$ is estimated using our estimator of the local autocovariance function based on the 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 an index p such that the "clipped" predictor

$$\hat{X}^{(p)}_{t,T} = \sum_{s=t-p}^{t-1} b^{(1)}_{t-1-s;T} X_{s,T} \tag{6.22}$$

gives a good compromise between the theoretical prediction error and the estimation error. The construction (6.22) is reminiscent of the classical idea of AR(p) approximation for stationary processes.

We propose an automatic procedure for selecting the two nusiance parameters: the order p in (6.22) and the bandwidth g, necessary to smooth the inconsistent estimator $\tilde{c}(z, \tau)$ using a kernel method. The idea of this procedure is to start with some 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* [58].

Suppose that we observe the series up to X_{t-1} and 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 some 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 pair (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). As the true value of X_{t-s} 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, the updated pair is used as the initial parameters, and itself updated by predicting X_{t-s+1} from $X_0, \ldots, X_{t-s}$. 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 perform the actual prediction.

Many different criteria can be used to compare the quality of the pairs of parameters at each step. Denote by $\hat{X}_{t-i}(p, g)$ the predictor of X_{t-i} computed using pair (p, g), and by $I_{t-i}(p, g)$ the corresponding 95% *prediction interval* based on the assumption of Gaussianity:

$$\begin{aligned} &I_{t-i}(p, g) \\ &\quad = \left[-1.96\hat{\sigma}_{t-i}(p, g) + \hat{X}_{t-i}(p, g)\ , 1.96\hat{\sigma}_{t-i}(p, g) + \hat{X}_{t-i}(p, g)\right], \end{aligned} \tag{6.23}$$

where $\hat{\sigma}^2_{t-i}(p, g)$ is the estimate of $\mathrm{MSPE}(\hat{X}_{t-i}(p, g), X_{t-i})$ computed using formula (6.8) with the remainder neglected. The criterion which we use in the simulations reported in the next section is to compute

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

for each of the 9 pairs at each step of the procedure and select the updated pair as the one which 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.

In order to be able to use our procedure completely on-line, we do not have to repeat the whole algorithm. Indeed, when observation X_t becomes available, we only have to update the pair (p_1, g_1) by predicting X_t, and we directly obtain the "optimal" pair for predicting X_{t+1}.

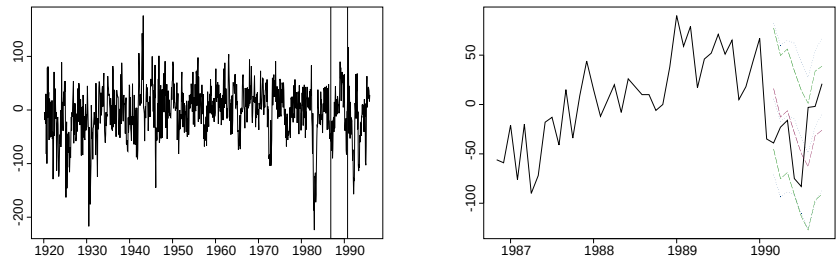
There are, obviously, many possible variants of our algorithm. Possible modifications include, for example, using a different criterion, restricting the allowed parameter space for (p, g), penalising certain regions of the parameter space, or allowing more than one parameter update at each time point.

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.

6.5 Case study: Wind speed anomaly index

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 [91] 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 with Lipschitz-continuous EWS to model the speed. In this section, we 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 6.1(a) shows the plot of the series.

Throughout this section, we use Haar wavelets to estimate the local (co)variance. Having provisionally made a safe assumption of the possible non-stationarity of the data, we first attempt to find a suitable



(a) The wind anomaly index (in cm/s). The two vertical lines indicate the segment shown in Figure 6.1(b).

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

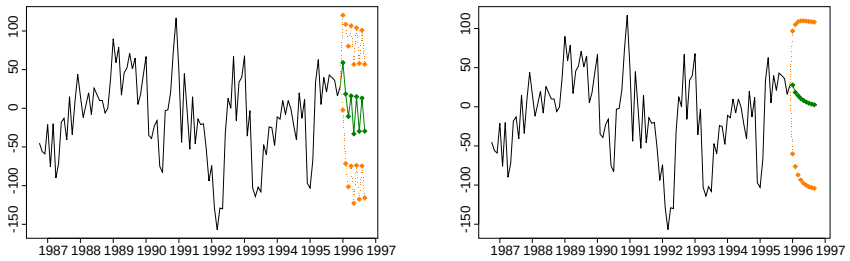
Figure 6.1: The wind anomaly data (910 observations from March 1920 to December 1995).

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 (formula (6.23)).

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 observations 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 6.1(b), which also compares our results to those obtained by means of AR 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 also similar in 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.



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

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

Figure 6.2: The last observations of the wind anomaly series and its 1- up to 9-step-ahead forecasts (in cm/s). The first predicted value in Figure (b) corresponds to March 1990.

We end this section by applying our general prediction method to compute multi-step-ahead forecasts. Figure 6.2 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 6.2(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 6.2(b) shows the 9-step-ahead prediction based on AR modelling (here, AR(2)). The prediction in Figure 6.2(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 forecasts look 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.

6.6 Conclusion

In this last chapter, 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. These results are 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. Moreover, 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 [82].

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 in order 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 as 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 a meteorological data set. Our non-parametric forecasting algorithm shows interesting advantages over the classical parametric alternative (AR forecasting). 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.

Conclusions

Overview of the contributions

The contributions of this thesis on locally stationary processes are of three types.

Modelling. The contribution is threefold. First, we present a simple and meaningful model for economic time series data in Chapter 1. This is a new model in econometry, and the empirical evaluation of its fitting is new in this context. Second, Chapter 3 extends the class of locally stationary wavelet processes in a significant way. We are now able to model intermittent phenomena, and not only smoothly varying evolutions in the spectrum. Finally, the test of significance of Chapters 4 and 5 allows to remove some scales which are not significant in the evolutionary wavelet spectrum (EWS). In that sense, we get an EWS with some inactive scales, which offers apparent advantages compared to the "full EWS". In particular, the EWS is now easier to analyse as we can focus on those scales and locations that provide a significant contribution.

Estimating. Chapter 2 presents a new estimating (or fitting) procedure for semiparametric models. Moreover, Chapter 4 presents a new pointwise estimator of the EWS. In both cases, our results are based on non asymptotic risk bounds. These bounds are more delicate to derive in our context than in the iid case. Moreover, the estimation procedures depend on the spectral norm of the covariance matrix Σ_T and we provide a new consistent pre-estimator of this quantity.

Forecasting. It may seem contradictory to forecast data which are not assumed to be stationary. We address this problem and give a pre-

cise explanation of the forecasting mechanics for locally stationary processes. Our forecasting procedure concerns the LSW models and we provide a new data-driven algorithm for forecasting these processes.

The estimation and forecasting procedures in this thesis are *adaptive* in the sense that all the parameters needed in the algorithm are chosen in a data-driven way. Applications are provided on economic, biomedical and meteorological time series.

Possible directions for future research

Some challenging problems remain for future research. Some of these have already been mentioned above in the thesis.

A first problem, which is not addressed in our work, is the problem of *model selection*. This problem arises in two situations. First, in the semiparametric fitting of Chapter 2, we assumed that the number of components of θ were known. In the context of time-varying ARMA(p,q) fitting, this means that the orders p and q of the model are known. Of course, it would be of a considerable interest to develop procedures for selecting these orders from data. Another model selection problem arises for the LSW model. LSW processes are constructed using a fixed wavelet system, e.g. Haar or another Daubechies' system. It would be interesting to develop a method for choosing the wavelet system automatically. This is a relevant question because an EWS could be sparse with respect to one given wavelet system, and not sparse with respect to another one.

To solve these two model selection problems, one current possibility offered by this thesis is to compare the fitting quality of each model by comparing its prediction performance on the observed data. This has to be investigated in more details. In particular, this is perhaps not the best solution in terms of computing time.

Another nice problem is the use of the results of Chapter 5 in order to construct *tests of second-order stationarity*. We provided an illustration on Tremor data, but we do not believe that this is the definitive answer to this question. This test is a complicated multiple tests procedure, with highly dependent tests. A theoretical investigation of this procedure could be useful in order to derive an appropriate level of the test. Another possibility is to use the false discovery rate in order to improve the quality of the test [1]. This is not straightforward, since it requires

to analyse the correlations of the corrected wavelet periodogram (CWP) between scale. And this inter-scales correlation cannot be written in a simple way from a mathematical viewpoint.

To end with, we also mention a possible extension of the pointwise adaptive estimator of Chapter 4 if we replace the histogram-based pointwise estimation by a *smooth kernel estimate*. The behaviour of the resulting estimator in terms of the smoothness of the kernel would be of interest. As our procedure is based on nonasymptotic approximations and is fully adaptive, a practical evaluation of the kernel-based estimator may also be provided through simulations.

APPENDIX A

Standard results in matrix theory

In this appendix, we recall some standard results in matrix theory. These are used in the proofs of the thesis.

Suppose M is an $n \times n$ matrix and $M^\star$ is the conjugate transpose of M. We denote

$$\|M\|_2 := \sqrt{\mathrm{tr}\,(M^\star M)}$$

the Euclidean norm of M and

$$\|M\|_{\mathrm{spec}} := \max\{\sqrt{\lambda} : \lambda \text{ is eigenvalue of } M^\star M\}$$

the spectral norm of M. If M is symmetric and nonnegative definite, by standard theory we have

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

We will also use the following standard relations which hold for all symmetric matrices B, C:

$$\|B\|_{\mathrm{spec}} \leqslant \|B\|_2 \quad \text{(A.2)}$$

$$\|B\|_{\mathrm{spec}} = \max\{\lambda : \lambda \text{ is eigenvalue of } B\} \quad \text{(A.3)}$$

$$\|BC\|_{\mathrm{spec}} \leqslant \|B\|_{\mathrm{spec}} \|C\|_{\mathrm{spec}} \quad \text{(A.4)}$$

$$\|BC\|_2 \leqslant \|B\|_{\mathrm{spec}} \|C\|_2 \leqslant \|B\|_2 \|C\|_2 \quad \text{(A.5)}$$

Moreover, if we suppose that the elements of the matrix B are continuously differentiable functions of t, then we shall also use

$$\frac{\partial}{\partial t} \log \det B = \operatorname{tr}\left(B^{-1} \frac{\partial}{\partial t} B \right). \tag{A.6}$$

APPENDIX B

Functional spaces

We will first define Hölder spaces C^s and Sobolev spaces $W^{s,p}$, where s is a positive number (not necessarily an integer) which is associated to the regularity of the function space. After that, we will define the Besov spaces $B_q^{s,p}$.

B.1 Hölder spaces

If m is a nonnegative integer, C^m is the space of continuous functions which are bounded up to the order m (i.e. f and $f^{(k)}$ are continuous and bounded for all integer $k \leqslant m$). With the norm

$$\sup_{x\in\mathbb{R}} |f(x)| + \sup_{x\in\mathbb{R}} |f^{(m)}(x)|,$$

C^m is a Banach space.

To define Hölder spaces with noninteger m, we note that, for all f in $C^1(\mathbb{R})$ and for all $h \in \mathbb{R}$, we have $\sup_{x\in\mathbb{R}} |f(x) - f(x-h)| \leqslant C \cdot |h|$ and, for all f in $C^0(\mathbb{R})$, $\sup_{x\in\mathbb{R}} |f(x) - f(x-h)|$ tends to zero as $|h| \to 0$ arbitrarily slowly. These observations lead us to define, if $0 < s < 1$:

$$C^s(\mathbb{R}) = \left\{ f \in C^0(\mathbb{R}) \text{ such that } \sup_{x\in\mathbb{R}} |f(x) - f(x-h)| \leqslant C \cdot |h|^s \right\}$$

and, if $m < s < m+1$:

$$C^s(\mathbb{R}) = \left\{ f \in C^m(\mathbb{R}) \text{ such that } f^{(m)} \in C^{s-m}(\mathbb{R}) \right\}.$$

This property can be rewritten

$$\sup_{x\in\mathbb{R}} |\Delta_h^n f(x)| \leqslant C \cdot |h|^s$$

where $n > s$ and Δ_h^n is such that

$$\begin{cases} \Delta_h^1 f(x) = f(x) - f(x-h) \\ \Delta_h^n f(x) = \Delta_h^1 \left(\Delta_h^{n-1}\right) f(x). \end{cases}$$

B.2 Sobolev spaces

Now, we will recall the definition of the Sobolev spaces in the frequency domain. If s is nonnegative, we define

$$H^s(\mathbb{R}) = \Bigg\{ f \in L^2(\mathbb{R}) \text{ such that } \|f\|_{H^s} = \left(\int_{\mathbb{R}} (1+|\lambda|)^{2m} \left| \hat{f}(\lambda) \right|^2 d\lambda \right)^2 < \infty \Bigg\}.$$

More generally, in L^p, the Sobolev space $W^{s,p}$ is defined through the norm

$$\|f\|_{W^{s,p}} = \|f\|_{L^p} + \left(\int_{\mathbb{R}^2} \frac{\left| f^{(m)}(x) - f^{(m)}(y) \right|^p}{|x-y|^{sp+1}} dxdy \right)^{1/p}$$

where m is a nonnegative integer, $0 < s < 1$ and the derivatives of f are weak derivatives (see Adams [2]). With these definitions, $W^{s,2} = H^s$.

B.3 Besov spaces

Before defining the Besov spaces, we have to define the moduli of continuity, that are such that, for all $t \geqslant 0$:

$$\begin{aligned} &\omega_p^1(f,t) = \sup_{|h| \leqslant t} \left\| \Delta_h^1 f \right\|_{L^p} \\ &\vdots \\ &\omega_p^n(f,t) = \sup_{|h| \leqslant t} \left\| \Delta_h^n f \right\|_{L^p} \end{aligned}$$

where n is a nonnegative integer. Now, we are able to define the Besov spaces $B_q^{s,p}(\mathbb{R})$, for $s > 0$ and $1 \leqslant p, q \leqslant \infty$. These are spaces of functions f in $L^p(\mathbb{R})$ such that

$$\left(2^{sj} \omega_p^n\left(f, 2^{-j}\right)\right)_{j \geqslant 0} \in \ell^q(\mathbb{N}). \tag{B.1}$$

A natural norm for these spaces is

$$\|f\|_{B_q^{s,p}} = \|f\|_{L^p} + \left\|2^{sj}\omega_p^n\left(f, 2^{-j}\right)\right\|_{\ell^q(\mathbb{N})}$$

with $s < n$. (It can be shown that the definition of Besov spaces does not depend on n, more presicely if $t \leqslant m, n$, then $\omega_p^n(f,t) \sim \omega_p^m(f,t)$).

Besov spaces contain a third parameter, q. If $q = \infty$, (B.1) leads to $\|\Delta_h^n f\| \leqslant C \cdot |h|^{-s}$ with $|h| < 1$. If $q < \infty$, this decay is faster. Roughly speaking, the parameter q determines the regularity rate which is given by s. Indeed, we have that $B_{q_1}^{s,p} \subseteq B_{q_2}^{s,p}$ if $q_1 \leqslant q_2$. But this new parameter is of secondary importance with respect to s, because of $B_{q_1}^{s_1,p} \subseteq B_{q_2}^{s_2,p}$ for $s_1 \leqslant s_2$ and for *all* q_1, q_2.

There exists numerous relations between Hölder, Sobolev and Besov spaces. For example, if s is not an integer, we have $B_\infty^{s,\infty} = C^s$ and $B_p^{s,p} = W^{s,p}$. This relation does not hold if s is not an integer, except for $B_2^{s,2} = W^{s,2} = H^s$. Other embeddings are illustrated in Figure B.1.

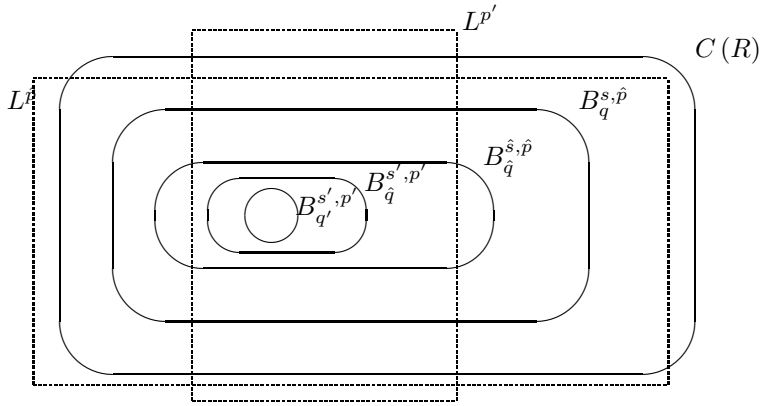


Figure B.1: Some embedding results in Besov spaces ($s' - 1/p' = \hat{s} - 1/\hat{p}$, $q' \leqslant \hat{q}$, $q' \leqslant q$, $s \geqslant \hat{s}$, $s > 1/\hat{p}$, $p' \leqslant \hat{p}$)

.

Index

Bibliography

[1] F. Abramovich, Y. Benjamini, D. Donoho, and I. Johnstone. Adapting to unknown sparsity by controlling the False Discovery Rate. Technical Report 2000-19, Dept. of Statistics, Stanford University, 2000. `http://www.math.tau.ac.il/~felix/ltx/PAPERS/Annals.ps.gz`.

[2] R. Adams. *Sobolev Spaces*. Academic Press, New York, 1975.

[3] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automat. Control*, 19:716–723, 1974.

[4] Y. Baraud, F. Comte, and G. Viennet. Adaptive estimation in autoregression or β-mixing regression via model selection. *Ann. Statist.*, 29:839–875, 2001.

[5] A. R. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related fields*, 113:301–413, 1999.

[6] Y. Benjamini and Y. Hochberg. Controlling the false detection rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B*, 57:289–300, 1995.

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

[8] K. Berkner and R. Wells. Smoothness estimates for soft-threshold denoising via translation-invariant wavelet transforms. *Appl. Comput. Harmon. Anal.*, 12:1–24, 2002.

[9] L. Birgé and P. Massart. Rates of convergence for minimum contrast estimators. *Probab. Theory and Related Fields*, 97:113–150, 1993.

[10] L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgersen, and G. Yang, editors,

Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics, pages 55–87, New York, 1997. Springer-Verlag.

[11] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4:329–375, 1998.

[12] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *J. Econometrics*, 31:307–327, 1986.

[13] T. Bollerslev and E. Ghysel. Periodic autoregressive conditional heteroscedasticity. *J. Bus. Econ. Statist.*, 14:139–151, 1996.

[14] D. Brillinger. *Time Series. Data Analysis and Theory.* Holt, Rinehart and Winston, Inc., 1975.

[15] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods.* Springer Series in Statistics. Springer, New York, second edition, 1991.

[16] L. D. Brown and M. G. Low. A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.*, 24: 2524–2535, 1996.

[17] P. F. Christoffersen. Evaluating interval forecasts. *Internat. Econom. Rev.*, 39:841–862, 1998.

[18] M. P. Clements and D. F. Hendry. *Forecasting economic time series.* Cambridge University Press, 1998.

[19] R. Coifman and D. Donoho. Time-invariant de-noising. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, volume 103 of *Lecture Notes in Statistics*, pages 125–150, New York, 1995. Springer-Verlag.

[20] F. Comte. Adaptive estimation of the spectrum of a stationary Gaussian sequence. *Bernoulli*, 7:267–298, 2001.

[21] R. Dahlhaus. Asymptotic statistical inference for nonstationary processes with evolutionary spectra. In P. Robinson and M. Rosenblatt, editors, *Athens conference on Applied Probability and Time Series Analysis*, volume 2. Springer, New York, 1996.

[22] R. Dahlhaus. On the Kullback-Leibler information divergence of locally stationary processes. *Stochastic Process. Appl.*, 62:139–168, 1996.

[23] R. Dahlhaus. Fitting time series models to nonstationary processes. *Ann. Statist.*, 25:1–37, 1997.

[24] R. Dahlhaus. A likelihood approximation for locally stationary processes. *Ann. Statist.*, 28:1762–1794, 2000.

[25] R. Dahlhaus and L. Giraitis. On the optimal segment length for parameter estimates for locally stationary time series. *J. Time Ser. Anal.*, 19:629–655, 1998.

[26] R. Dahlhaus and M. H. Neumann. Locally adaptive fitting of semiparametric models to nonstationary time series. *Stochastic Process. Appl.*, 91:277–308, 2001.

[27] R. Dahlhaus and W. Polonik. Empirical spectral processes and nonparametric maximum likelihood estimation for time series. In H. Dehling, T. Mikosch, and M. Sørensen, editors, *Empirical Process Techniques for Dependent Data*, Boston, 2002. Birkhäuser.

[28] R. Dahlhaus and W. Polonik. Nonparametric maximum likelihood estimation of parameter curves for locally stationary processes. Manuscript, 2003.

[29] R. Dahlhaus and S. Subba Rao. Statistical inference for time-varying ARCH processes. Preprint, Universität Heidelberg, 2003.

[30] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 41:909–996, 1988.

[31] I. Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE Trans. Inform. Theory*, 36:961–1005, 1990.

[32] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, Philadelphia, 1992.

[33] R. A. De Vore and G. G. Lorentz. *Constructive Approximation*. Springer-Verlag, Berlin, 1993.

[34] F. X. Diebold and R. S. Mariano. Comparing predictive accuracy. *J. Bus. Econ. Statist.*, 13:253–263, 1995.

[35] K. Dzhaparidze. *Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series*. Springer, New York, 1986.

[36] R. Engle. Autoregressive conditional heteroskedasticity with estimates of the variance of the U.K. inflation. *Econometrica*, 50:987–1008, 1982.

[37] K. B. Eom. Time-varying autoregressive modeling of high rande resolution radar signatures for classification of noncooperative targets. *IEEE Trans. Aerosp. Electron. Syst.*, 35:974–988, 1999.

[38] P. Fryźlewicz. Modelling and forecasting financial log-returns as locally stationary wavelet processes. Research report, Department of Mathematics, University of Bristol, 2002. `http://www.stats.bris.ac.uk/pub/ResRept/2002/14.pdf`.

[39] P. Fryźlewicz, S. Van Bellegem, and R. von Sachs. Forecasting non-stationary time series by wavelet process modelling. *Ann. Inst. Statist. Math.*, 2003. To appear.

[40] M. Giurcanu and V. Spokoiny. Confidence estimation of the covariance function of stationary and locally stationary processes. Preprint 726, WIAS, Berlin, 2002. `http://www.math.tu-berlin.de/~giurcanu/document.pdf`.

[41] U. Grenander. *Abstract Inference*. Wiley, New York, 1981.

[42] Y. Grenier. Time-dependent ARMA modeling of nonstationary signals. *IEEE Trans. Acoust. Speech Signal Process*, 31:899–911, 1983.

[43] C. Grillenzoni. Forecasting unstable and nonstationary time series. *Int. J. Forecasting*, 14:469–482, 1998.

[44] C. Grillenzoni. Time-varying parameters prediction. *Ann. Inst. Statist. Math.*, 52:108–122, 2000.

[45] A. Grossmann and J. Morlet. Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM J. Math. Anal.*, 15:723–736, 1984.

[46] M. Hallin. Mixed autoregressive moving average multivariate processes with time-dependent coefficients. *J. Multivariate Anal.*, 8:567–572, 1978.

[47] B. E. Hansen. The new econometrics of structural change: Dating breaks in U.S. labor productivity. *J. Econ. Perspect.*, 15(4):117–128, 2001.

[48] P. R. Hansen and A. Lunde. A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? Working paper 01-04, Department of Economics, Brown University, 2001.

[49] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelets, Approximation and Statistical Applications*, volume 129 of *Lectures Notes in Statistics*. Springer-Verlag, New York, 1998.

[50] D. Hoffman and A. R. Pagan. Post-sample prediction tests for generalized method of moment estimators. *Oxford Bull. Econ. Statist.*, 51: 333–343, 1989.

[51] I. Johnstone. Wavelet shrinkage for correlated data and inverse problems: Adaptivity results. *Statist. Sinica*, 9:51–83, 1999.

[52] I. Johnstone and B. W. Silverman. Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc. Ser.* B, 59:319–351, 1997.

[53] M. Juntunen and J. Kaipio. Stabilization of TVAR models: A regularization approach. *Circuits Systems Signal Process.*, 19:423–435, 2000.

[54] S. Kawasaki and R. Shibata. Weak stationarity of a time series with wavelet representation. *Japan J. Indust. Appl. Math.*, 12:37–45, 1995.

[55] G. Kitagawa and W. Gersch. A smoothness priors time-varying AR coefficient modeling of the nonstationary covariance time series. *IEEE Trans. Automat. Control.*, 30:48–56, 1985.

[56] R. Kress. *Numerical Analysis.* Graduate Texts in Mathematics # 181. Springer, 1991.

[57] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28:1302–1338, 2000.

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

[59] O. Lepski. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probab. Appl.*, 35:454–470, 1990.

[60] O. Lepski and V. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.*, 25:2512–2546, 1997.

[61] K.-S. Lii and M. Rosenblatt. Spectral analysis for harmonizable processes. *Ann. Statist.*, 30:258–297, 2002.

[62] A. W. Lo. Long-term memory in stock market prices. *Econometrica*, 59:1279–1313, 1991.

[63] M. Loretan and P. C. Phillips. Testing the covariance stationarity of heavy-tailed time series. *J. Empirical Finance*, 1:211–248, 1994.

[64] C. A. Los. Nonparametric efficiency testing of Asian stock markets using weekly data. *Adv. in Econometrics*, 14:329–363, 2000.

[65] R. Loynes. On the concept of the spectrum for non-stationary processes (with discussions). *J. Roy. Statist. Soc. Ser.* B, 30:1–30, 1968.

[66] S. Mallat. Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Trans. Amer. Math. Soc.*, 315:69–87, 1989.

[67] S. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11:674–693, 1989.

[68] S. Mallat. *A Wavelet Tour on Signal Processing.* Academic Press, San Diego, 1998.

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

[70] W. Martin and P. Flandrin. Wigner-Ville spectral analysis of nonstationary processes. *IEEE Trans. Acoust. Speech Signal Process.*, 33: 1461–1470, 1985.

[71] G. Matz, F. Hlawatsch, and W. Kozek. Generalized evolutionary spectral analysis and the Weyl spectrum of nonstationary random processes. *IEEE Trans. Signal Processing*, 45:1520–1534, 1997.

[72] R. A. Meese and K. Rogoff. Was it real? The exchange rate-interest differential relation over the modern floating rate period. *J. Finance*, 43: 933–948, 1988.

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

[74] D. Mercurio and V. Spokoiny. Statistical inference for time-inhomogeneous volatility models. *Ann. Statist.*, 2003. To appear.

[75] Y. Meyer. Principe d'incertitude, bases hilbertiennes et algèbres d'opérateurs. *Séminaire Bourbaki*, 662, 1986.

[76] Y. Meyer. *Ondelettes et Opérateurs 1.* Hermann, Paris, 1990.

[77] T. Mikosch and C. Stărică. Change of structure in financial data, long-range dependence and GARCH modelling. Technical report, University of Groningen, 1999. `http://www.math.ku.dk/~mikosch/preprint.html`.

[78] K. G. Murty. *Linear Complementarity, Linear and Nonlinear Programming.* 1997. Internet edition available at `http://ioe.engin.umich.edu/people/fac/books/murty`.

[79] G. P. Nason. *WaveThresh3 Software*. Department of Mathematics, University of Bristol, Bristol, UK, 1998.

[80] G. P. Nason and T. Sapatinas. Wavelet packet transfer function modelling of nonstationary time series. *Statist. Comput.*, 12:45–56, 2002.

[81] G. P. Nason and B. W. Silverman. The stationary wavelet transform and some statistical applications. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, volume 103 of *Lecture Notes in Statistics*, pages 281–299, New York, 1995. Springer-Verlag.

[82] G. P. Nason and R. von Sachs. Wavelets in time series analysis. *Phil. Trans. R. Soc. Lond.* A, 357:2511–2526, 1999.

[83] G. P. Nason, R. von Sachs, and G. Kroisandt. Wavelet processes and adaptive estimation of evolutionary wavelet spectra. *J. Roy. Statist. Soc. Ser.* B, 62:271–292, 2000.

[84] D. B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59:347–370, 1991.

[85] M. Neumann and R. von Sachs. Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *Ann. Statist.*, 25:38–76, 1997.

[86] W. K. Newey and K. D. West. A simple, positive semi-definite, heterosketasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55:703–708, 1987.

[87] H. Ombao, J. Raz, R. von Sachs, and W. Guo. The SLEX Model of a non-stationary random process. *Ann. Inst. Statist. Math.*, 54:171–200, 2002.

[88] A. R. Pagan and G. W. Schwert. Alternative models for conditional stock volatility. *J. Econometrics*, 45:267–290, 1990.

[89] A. R. Pagan and G. W. Schwert. Testing for covariance stationarity in stock market data. *Econ. Letters*, 33:165–170, 1990.

[90] C. H. Page. Instantaneous power spectra. *J. Appl. Phys.*, 23:103–106, 1952.

[91] S. Philander. *El Niño, La Niña and the Southern Oscillation*. Academic Press, San Diego, 1990.

[92] P. C. Phillips. Time series regression with a unit root. *Econometrica*, 55:277–301, 1987.

[93] D. Pollard. *Convergence of stochastic processes*. Springer Verlag, 1984.

[94] M. Priestley. Evolutionary spectra and non-stationary processes. *J. Roy. Statist. Soc. Ser.* B, 27:204–237, 1965.

[95] M. Priestley. *Spectral Analysis of Time Series*. Probability and Mathematical Statistics. Academic Press, London, 1981.

[96] J. B. Ramsey and Z. Zhang. The analysis of foreign exchange data using waveform dictionaries. *J. Empirical Finance*, 4:341–372, 1997.

[97] R. von Sachs and B. MacGibbon. Nonparametric curve estimation by wavelet thresholding with locally stationary errors. *Scand. J. Statist.*, 27:475–499, 2000.

[98] R. von Sachs, G. P. Nason, and G. Kroisandt. Adaptive estimation of the evolutionary wavelet spectrum. Technical Report 516, Department of Statistics, Stanford, 1997. `http://www.stats.bris.ac.uk/pub/reports/Wavelets/StanTechRep516.ps.gz`.

[99] R. von Sachs and M. H. Neumann. A wavelet-based test for stationarity. *J. Time Ser. Anal.*, 21:597–613, 2000.

[100] R. von Sachs and K. Schneider. Wavelet smoothing of evolutionary spectra by non-linear thresholding. *Appl. Comput. Harmon. Anal.*, 3: 268–282, 1996.

[101] N. Saito and G. Beylkin. Multiresolution representations using the autocorrelation functions of compactly supported wavelets. *IEEE Trans. Signal Process.*, 41:3584–3590, 1993.

[102] G. W. Schwert. Indexes of United States stock prices from 1802 to 1987. *J. Bus.*, 63:399–426, 1990.

[103] R. H. Shumway. Time-frequency clustering and discriminant analysis. *Statist. & Prob. Letters*, 63:307–314, 2003.

[104] R. A. Silverman. Locally stationary random process. *IRE Trans. Information Theory*, IT-3:182–187, 1957.

[105] V. Spokoiny. Estimation of a function with discontinuities via local polynomial fit with an adaptive choice. *Ann. Statist.*, 26:1356–1378, 1998.

[106] V. Spokoiny. Data driven testing the fit of linear models. *Math. Methods Statist.*, 10:465–497, 2001.

[107] T. Subba Rao. The fitting of non-stationary time-series models with time-dependent parameters. *J. Roy. Statist. Soc. Ser.* B, 32:312–322, 1970.

[108] N. R. Swanson and H. White. Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. *Int. J. Forecasting*, 13:439–461, 1997.

[109] G. Thonet, T. Duvanel, J.-M. Vesin, E. Pruvot, and M. Fromer. Assessment of stationarity horizon of the heart rate. In *Proceedings of the 18th Annual International Conference on the IEEE Engineering in Medicine and Biology*, Amsterdam, October 1996. Paper # 764.

[110] S. Van Bellegem, P. Fryźlewicz, and R. von Sachs. A wavelet-based model for forecasting non-stationary processes. In J.-P. Gazeau, R. Kerner, J.-P. Antoine, S. Métens, and J.-Y. Thibon, editors, *GROUP 24: Physical and Mathematical Aspects of Symmetries*, Bristol, 2003. IOP Publisher. To appear.

[111] S. Van Bellegem and R. von Sachs. Forecasting economic time series with unconditional time-varying variance. *Int. J. Forecasting*, 2003. To appear.

[112] S. Van Bellegem and R. von Sachs. Locally adaptive estimation of sparse evolutionary wavelet spectra. Discussion Paper 0310, Université catholique de Louvain, Institut de statistique, 2003. `ftp://www.stat.ucl.ac.be/pub/papers/dp/dp03/dp0310.ps`. Submitted for publication.

[113] S. Van Bellegem and R. von Sachs. On adaptive estimation for locally stationary processes and its applications. Forthcoming discussion paper, Université catholique de Louvain, Institut de statistique, 2003. Submitted for publication.

[114] A. van der Vaart and J. Wellner. *Weak convergence and empirical processes*. Springer Verlag, New York, 1996.

[115] J. Verlaak, C. Andrieu, A. Doucet, and S. J. Godsill. Particle methods for bayesian modeling and enhancement of speech signals. *IEEE Trans. Speech, Audio Process.*, 10:173–185, 2002.

[116] B. Vidakovic. *Statistical modeling by wavelets*. Wiley, New York, 1999.

[117] H. L. White. A reality check for data snooping. *Econometrica*, 68: 1097–1126, 2000.

[118] H. L. White and I. Domowitz. Nonlinear regression with dependant observations. *Econometrica*, 52:143–161, 1984.

[119] P. Whittle. Estimation and information in stationary time series. *Ark. Mat.*, 2:423–434, 1953.

[120] Y. Yajima. On estimation of a regression model with long-memory stationary errors. *Ann. Statist.*, 19:158–177, 1988.

[121] T. Zeng and N. R. Swanson. Predictive evaluation of econometric forecasting models in commodity futures markets. *Stud. Nonlinear Dynam. Econometrics*, 2:159–177, 1998.