Semiparametric estimation by model selection for locally stationary processes[1]

Sébastien Van Bellegem[2]   Rainer Dahlhaus[3]

Submitted: March, 3, 2005
This revision: April 9, 2006

Abstract

Over the last decades more and more attention has been paid to the problem how to fit a parametric model of time series with time-varying parameters. A typical example is given by autoregressive models with time-varying parameters (tvAR processes). We propose a procedure to fit such time-varying models to general nonstationary processes. The estimator is a maximum Whittle likelihood estimator on sieves. The results do not assume that the observed process belongs to a specific class of time-varying parametric models. We discuss in more details the fitting of tvAR(p) processes for which we treat the problem of the selection of the order p, and propose an iterative algorithm for the computation of the estimator. Comparison with model selection by AIC is provided through simulations.

Keywords: time-varying autoregressive process, locally stationary process, sieve estimator, Whittle likelihood, model selection, empirical spectral process.

AMS 1991 subject classification: Primary 60G12; secondary 62M10, 62G05, 41A45

[1] This work was supported by the IAP research network nr P5/24 of the Belgian Government (Belgian Science Policy) and by the "Fonds Spciaux de Recherche" (FSR 05) of the Universit catholique de Louvain, Louvain-la-Neuve, Belgium. Part of this research was done when S. Van Bellegem visited the Institut für Angewandte Mathematik, Universität Heidelberg. Financial supports from the Deutscher Akademischer Austausch Dienst (DAAD), and from the Deutsche Forschungsgemeinschaft (DA 187/12-1+2) are gratefully acknowledged.

[2] *Corresponding author.* Collaborateur scientifique du F.N.R.S. Université catholique de Louvain, Institut de statistique. Voie du Roman Pays, 20, B-1348 Louvain-la-Neuve, Belgium, `vanbellegem@stat.ucl.ac.be`

[3] Universität Heidelberg, Institut für Angewandte Mathematik, Im Neuenheimer Feld, 294, D-69120 Heidelberg, Germany, `dahlhaus@statlab.uni-heidelberg.de`

1 Introduction

In recent years several estimation methods have been derived for locally stationary time series models, that is for models whose parameters change slowly in time and which can locally be approximated by stationary processes. Out of the large literature we mention the work of Priestley (1965) on oscillatory processes, Dahlhaus (1997) on locally stationary processes, Nason, von Sachs and Kroisandt (2000) on a wavelet-based model of evolutionary spectra, and more recent works such as Ombao, von Sachs and Guo (2005) on multivariate time series or Davis, Lee and Rodriguez-Yam (2006) on piecewise stationary processes.

In this paper we address the problem of model selection for sieve estimates for such models in a rigorous way. As a contrast function we use an approximation of the Kullback-Leibler divergence. Below we assume that the true process is locally stationary in the sense of Dahlhaus (1997) (see Definition 2.1 below). The models we will study are parametrized by a D-dimensional function $\theta(u)$. One example is given by the time-varying autoregressive (tvAR(p)) model

$$X_{t,T} + \sum_{j=1}^{p} a_j \left(\frac{t}{T} \right) X_{t-j,T} = \varepsilon_{t,T} \ , \quad t = 0, \ldots, T-1 \ , \quad T > 0 \ , \tag{1.1}$$

where $\varepsilon_{t,T}$ are independent normal random variables $\mathcal{N}(0, \sigma^2(t/T))$. In this example, $D = p + 1$ and $\theta(u) = (\sigma^2(u), a_1(u), \ldots, a_p(u))$ for $u \in [0, 1]$. As usual for locally stationary processes the parameters are rescaled to the unit interval $[0, 1]$ in order to obtain a meaningful asymptotic theory.

A usual assumption is that the coefficients $a_j(u)$ may be approximated satisfactorily by a linear combination of a small number of known functions. For instance, Subba Rao (1970) assumes that the first three terms of the Taylor expansion give a good approximation for the parameters, i.e. $a_j(u) = a_{j,0} + a_{j,1}u + a_{j,2}u^2/2$. Similar ideas with various approximations in a finite-dimensional linear space of approximation may be found in the literature, see for instance the review in Grenier (1983). In summary one approximates the time-varying parameters in a suitable orthonormal basis $\{\varphi_j\}$ and assumes that the expansion

$$a_j(u) = \sum_{i=1}^{m} a_{ji} \varphi_i(u) \tag{1.2}$$

holds true for each $j = 1, \ldots, D$.

The problem of choosing the number m of elements in the sum (1.2) occurs and we propose in this paper a data-driven method for selecting this parameter. More specifically, the goal of this paper 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 different m. These estimators are constructed as minimum contrast estimators where the contrast function is an approximation of the Gaussian likelihood of the model. The estimator $\hat{\theta}_{\hat{m}}$ is obtained from a model selection procedure.

The proposed procedure is inspired by the work of Barron *et al.* (1999); Birgé and Massart (1998), 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 with an L^2 contrast function to standard time series problems may be found in the literature (Baraud *et al.*, 2001; Comte, 2001). Our situation is different and more complex in the sense that we are dealing with dependent, covariance nonstationary data. The example of tvAR(p) models shows that the estimation procedure is complicated by the fact that the curve θ is not observed "directly". This is in contrast to classical nonparametric regression, where the curve $\theta(\cdot)$ is observed plus some noise. In our context, the characteristics of the process (such as the spectral density) may depend on the parameter curves in a highly nonlinear way. An additional difficulty is that our contrast function is the Whittle likelihood, which is more natural than, e.g., quadratic loss in the context of spectral density estimation.

The paper is organized as follows. In the next section, we recall the formal definition of locally stationary processes and their evolutionary spectral density. Then, in Section 3, we address the problem of semiparametric estimation. This problem is presented in a general setting including the tvAR(p) model as a particular example. This section summarizes the main results of the paper. In Section 4, we focus on the particular problem of fitting tvAR(p) models, including the question of the selection of p, and propose an algorithm for the estimation of the curve $\theta(u)$. This section also includes simulation results and compares the proposed model selection method with a method based on the AIC. The proof of the main results are to be found in Section 5. They are based on two maximal inequalities for the deviation of the empirical process of locally stationary processes that are proved in a technical appendix.

2 The model of local stationarity

2.1 Locally stationary processes

We assume that the observed data $X_1, \dots, X_T$ follow a general locally stationary processes as introduced in Dahlhaus (1997).

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

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

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), and where

2. *there exists a positive constant K and a 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,*

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

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

This definition of covariance nonstationary processes is a straightforward extension of the spectral (Cramér) representation for stationary time series. The difference comes from the transfer function $A(z,\lambda)$ that is depending on both time and frequency and is defined on $[0,1]\times[-\pi,\pi)$. The smoothness of A in u defines the departure from stationarity and ensures the locally stationary behavior of the process. The smoothness assumptions on A are formulated via the total variation norm. Recall that the *total variation norm* of a univariate function f defined on an interval $[a,b]$ is given by

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

If there is no risk of ambiguity on 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 (1997).

Assumption 2.1. The function A in Definition 2.1 is such that

(a) $\sup_u \mathrm{TV}_{[-\pi,\pi]}\left(A(u,\cdot)\right)\leqslant C_1<\infty$

(b) $\sup_\lambda \mathrm{TV}_{[0,1]}\left(A(\cdot,\lambda)\right)\leqslant C_2<\infty$

(c) $\sup_{u,\lambda}|A(u,\lambda)|\leqslant\kappa_s<\infty$

(d) $\inf_{u,\lambda}|A(u,\lambda)|\geqslant\kappa_1$ for some $\kappa_1>0$

(e) $\sup_u\sum_{s\in\mathbb{Z}}|\tilde{A}(u,s)|<\infty$, where $\tilde{A}(u,s):=(2\pi)^{-1}\int_{-\pi}^{\pi}d\lambda\ A(u,\lambda)\exp(i\lambda s)$ for $s\in\mathbb{Z}$ and $u\in[0,1]$.

This assumption presents mild conditions under which there is a unique spectral representation in the class of locally stationary processes, see Section 2.2 below. Part of these conditions might be fulfilled simply by restricting A to be a member of a specific smoothness class (Sobolev, Hölder, etc.), see Section 3 of Neumann and von Sachs (1997) for details.

In the above definition two different functions $A_{t,T}^{\circ}(\lambda)$ 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 $A_{t,T}^{\circ}(\lambda)=A(t/T,\lambda)$ in the above definition, then the class does not include the class of tvAR(p) processes (see Theorem 2.3 in Dahlhaus (1996b)).

2.2 Evolutionary spectral density

If $\{X_{t,T}\}$ is a locally stationary process then the *Wigner-Ville* spectrum is given by

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

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 $t > T$. Neumann and von Sachs (1997) have shown under Assumption 2.1 that

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

where

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

(see also Dahlhaus (1996b)). The function $f(u,\lambda)$ is called the *evolutionary spectral density* (ESD) of the process. The above result is important because it shows the uniqueness of the evolutionary spectral density $f(u,\lambda)$.

Remark 2.1. The uniqueness property of the ESD is a major difference between the theory of locally stationary processes and other approaches to model nonstationary time series, such as the theory of oscillatory processes (Priestley, 1965). Contrary to Priestley's definition, locally stationary time series are doubly indexed and their time-varying spectral density is rescaled on the time interval $[0,1]$. This is the key point that allows to make use of asymptotic considerations. A deeper comparison between the two approaches can be found in Dahlhaus (1996a).

3 Semiparametric estimation

The model we like to fit is characterized by a D-dimensional parameter function $\theta(u)$, $u \in [0,1]$, which defines the evolutionary spectral density $f_{\theta(u)}(\lambda)$. Dahlhaus and Neumann (2001) suggested to use a minimum distance method for the estimation of $\theta(\cdot)$, based on a contrast function between some nonparametric estimate of the evolutionary spectral density and the (model) evolutionary spectral density. We follow this method and have to define a suitable nonparametric estimate and the contrast function.

3.1 Contrast functions

Suppose we observe data $\{X_{1,T}, \ldots, X_{T,T}\}$ from a locally stationary process with evolutionary spectral density $f(u,\lambda)$. Motivated by the above convergence result for the Wigner-Ville spectrum, Neumann and von Sachs (1997) define the *preperiodogram* as

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

where the sum over k if for $k \in \mathbb{Z}$ such that $1 \leqslant [uT - (k-1)/2], [uT + (k+1)/2] \leqslant T$. 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 tends to infinity. 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. The decision about the degree of smoothing in each of these directions is left to the major smoothing step itself.

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). However, in the present paper, our goal is to fit a semiparametric model $f_{\theta(u)}(\lambda)$ to the data. It is worth mentioning that f is not assumed to obey the structure of the semiparametric 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

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

which is up to a constant the asymptotic Kullback-Leibler information divergence of a locally stationary process (Dahlhaus, 1996b). Then, we define the *empirical contrast function* by

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

where $J_T(t/T,\lambda)$ is the preperiodogram. $\mathcal{L}_T(f_\theta, J_T)$ is an approximation to the negative log-likelihood of locally stationary process (Dahlhaus, 2000).

3.2 The sieve estimator

Our aim is to develop a nonparametric estimator of the multivariate curve $\theta(\cdot) = (\theta^{(1)}(\cdot), \cdots, \theta^{(D)}(\cdot))$. Theoretically, an estimator can be constructed by minimizing the empirical contrast function $\mathcal{L}_T(f_\theta, J_T)$ over a class Θ of parameter curves. 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).

The approach we follow is a suitable adaptation of the *method of sieves* (Birgé and Massart, 1998). Each component $\theta^{(i)}(\cdot)$ of the target vector curve is approximated in a finite-dimensional, linear space of approximation S_{m_i}. As our aim is to estimate a D-dimensional curve, we set $\mathcal{N}_{D,T} = \{m = (m_1, \ldots, m_D), m_j \in \mathcal{M}_T\}$ and, for each multi-index $m = (m_1, \ldots, m_D)$, we define $\mathcal{F}_m = S_{m_1} \otimes \ldots \otimes S_{m_D}$.

The estimation procedure has two steps:

1. On each space $\mathcal{F}_m$, we minimize the empirical contrast function and compute the minimum contrast estimator
$$\hat{\theta}_m = \arg\min_{\theta\in\mathcal{F}_m} \mathcal{L}_T\left(f_\theta, J_T\right) \tag{3.1}$$
for each $m \in \mathcal{N}_{D,T}$.

2. From the set $\{\hat{\theta}_m : m \in \mathcal{N}_{D,T}\}$ of estimators, we choose $\hat{m}$ among the family $\mathcal{N}_{D,T}$ such that
$$\hat{m} = \arg\min_{m\in\mathcal{N}_{D,T}} \left\{\mathcal{L}_T\left(f_{\hat{\theta}_m}, J_T\right) + \text{pen}(m)\right\}$$
where pen(m) is a penalty function to be specified later.

Finally, the sieve estimator is
$$\hat{\theta} = \hat{\theta}_{\hat{m}}. \tag{3.2}$$
The explicit form of the penalty function is derived in Theorem 3.2 below.

Note that, in the above procedure, we assumed that the order D is fixed and known. A discussion about the selection of this parameter for the fitting of time-varying autoregressive models is presented in Section 4 below.

3.3 The collection of models

Before stating the main results and assumptions, we have to introduce some notations. If $g(u,\lambda)$ is a function over $[0,1]\times(-\pi,\pi)$, then we define the Fourier transform w.r.t. u as
$$\tilde{g}(u,j) := \int_{-\pi}^{\pi} d\lambda\ g(u,\lambda)\exp(i\lambda j)\ ,$$
and define
$$\rho_2(g) := \left(\int_0^1 du \int_{-\pi}^{\pi} d\lambda\ |g(u,\lambda)|^2\right)^{1/2}\ , \qquad \rho_\infty(g) := \sum_{j=-\infty}^{\infty} \sup_u |\tilde{g}(u,j)|\ ,$$
$$\tilde{v}(g) := \sup_j \text{TV}\left(\tilde{g}(\cdot,j)\right).$$
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:
$$\|\theta\|_2^2 := \sum_{i=1}^{D}\int_0^1 du\ \left(\theta^{(i)}(u)\right)^2\ , \qquad \|\theta\|_\infty := \sup_{i=1,\dots,D}\ \sup_{u\in[0,1]} |\theta^{(i)}(u)|\ .$$

The choice of a family of models $\{\mathcal{F}_m, m \in \mathcal{N}_{D,T}\}$ (i.e. the choice of a sieve) is basically guided by approximation theory. Typical examples are trigonometric polynomials, wavelet expansions or piecewise polynomials, because their approximation properties are well studied in the literature.

In this paper, each space $\mathcal{S}_{m_i}$ is a linear finite-dimensional subspace of $L^2([0,1]) \cap L^\infty([0,1])$ spanned by some orthonormal basis $\{\varphi_j; j \in \Lambda_{m_i}\}$ with $|\Lambda_{m_i}| = d_{m_i}$. 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 (1998). However, in our context, their definition is slightly different due to our specific framework.

Consider the expansion of $\theta^{(i)}$ in the basis S_{m_i}:

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

and set

$$\overline{r}_m = \frac{1}{\sqrt{d_m}} \sup_{\beta \neq 0} \sup_{1 \leqslant i \leqslant D} \frac{\left\| \sum_{j \in \Lambda_{m_i}} \beta_{ij} \varphi_j \right\|_\infty}{\sup_j |\beta_{ij}|}\ , \quad \Phi_m = \frac{1}{\sqrt{d_m}} \sup_{1 \leqslant i \leqslant D} \left\| \sum_{j \in \Lambda_{m_i}} \varphi_j^2 \right\|_\infty^{1/2}\ , \tag{3.3}$$

where $d_m = \sum_{i=1}^D d_{m_i}$ is the dimension of $\mathcal{F}_m$.

The two indices $\overline{r}_m$ and Φ_m describe the relationships between the L^2 and the L^∞ structure of the sieve $\mathcal{F}_m$. An extension of Lemma 1 of Birgé and Massart (1998) leads to the inequalities $\Phi_m \leqslant \overline{r}_m \leqslant \Phi_m \sqrt{d_m}$.

We consider the following assumptions on the collection of models:

Assumption 3.1. **(a)** For all $m_i \in \mathcal{M}_T$, S_{m_i} is a linear subspace of $L^2([0,1]) \cap L^\infty([0,1])$ with finite dimension d_{m_i}. It is generated by the orthonormal system of functions $\{\varphi_j; j \in \Lambda_{m_i}\}$.

(b) For all $m = (m_1, \ldots, m_D)$ in $\mathcal{N}_{D,T}$, $\mathcal{F}_m$ denotes the product space $S_{m_1} \otimes \ldots \otimes S_{m_D}$ of dimension $d_m = \sum_{i=1}^D d_{m_i}$. Each $\mathcal{F}_m$ is such that $\overline{r}_m \leqslant C_{\overline{r}} \sqrt{T/d_m}$ for all $m \in \mathcal{N}_{D,T}$.

(c) The collection of models $\mathcal{F} = \{\mathcal{F}_m : m \in \mathcal{N}_{D,T}\}$ is nested and such that $\mathcal{F}_m^\star = \{1/f; f \in \mathcal{F}_m\}$ are convex. Moreover, $\max_{m \in \mathcal{N}_{D,T}} d_m \leqslant T$, $\sup_{\mathcal{F}} \|\theta\|_2 \leqslant k_2 < \infty$, $\sup_{\mathcal{F}} \|\theta\|_\infty \leqslant k_\infty < \infty$ and $\sup_{\mathcal{F}} \tilde{v}(1/f_\theta) \leqslant \tilde{v} < \infty$.

To fix the ideas, we now present two examples of models which fulfill Assumption 3.1 . Other examples of models may be found in the standard literature (see for instance Barron *et al.* (1999) or Comte (2001)).

Example 3.1 (Trigonometric functions)**.** Consider spaces S_{m_i} generated from the functions $\varphi_j(u) = \sqrt{2}\cos(2\pi j u)$ for $j = 0, \ldots, m_i - 1$. The dimension of S_{m_i} is $d_{m_i} = m_i$. This collection is such that $\Phi_m^2 \leqslant 1$, hence $\overline{r}_m \leqslant \sqrt{d_m}$ and Assumption 3.1 holds with $C_{\overline{r}} = 1$ provided that $d_m \leqslant \sqrt{T}$.

Example 3.2 (Piecewise polynomials)**.** Consider 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 S_{m_i} is defined as the space of piecewise polynomials with degree bounded by $s-1$ on the partition I_m. The dimension of S_{m_i} is $r2^{m_i}$ and it follows from Barron *et al.* (1999) that $\overline{r}_m \leqslant \sqrt{(r+1)(2r+1)} = C_{\overline{r}}$ provided that $d_m \leqslant T$.

3.4 Main results

We first consider the following assumptions on distances.

Assumption 3.2. For all θ and ν in $\cup_{m \in \mathcal{N}_{D,T}} \mathcal{F}_m$, there exists finite strictly positive constants K_2, K_2', K_∞ (which may depend on D) such that

(a) $K_2'^{-1} \|\theta - \nu\|_2 \leqslant \rho_2\left(1/f_\theta - 1/f_\nu\right) \leqslant K_2 \|\theta - \nu\|_2$;

(b) $\rho_\infty\left(1/f_\theta - 1/f_\nu\right) \leqslant K_\infty \|\theta - \nu\|_\infty$.

With these two conditions, we assume a convenient connection between the L^2 and L^∞ norms of spectral densities and their corresponding time-varying curves. These conditions are mild compared, e.g., to Assumption (A.4) of Dahlhaus and Neumann (2001). To illustrate, we can consider for a tvAR (1) process with variance $\sigma^2 = 1$. In that situation one can derive $K_2 = 48\pi^3$, $K_2' = (16\pi^3)^{-1}$ and $K_\infty = 4\pi^2(1 + k_\infty)$.

The first result is on the Kullback-Leibler divergence between f and $f_{\hat{\theta}_m}$ 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 $\mathrm{Cov}(X_{s,T}, X_{t,T})$. Its spectral norm is

$$\|\Sigma\|_{\text{spec}} := \max\left\{\sqrt{\lambda} : \lambda \text{ eigenvalue of } \Sigma^\star\Sigma\right\}$$

where $\Sigma^\star$ is the transpose of Σ.

Theorem 3.1. *Suppose that we observe data $X_{1,T}, \ldots, X_{T,T}$ from a Gaussian locally stationary process. Fix a sieve $\mathcal{F}_m$ according to Assumptions 2.1 and define*

$$\theta_m = \arg\min_{\theta \in \mathcal{F}_m} \mathcal{L}\left(f_\theta, f\right)$$

Under Assumptions 3.1 and 3.2, the minimum contrast estimator $\hat{\theta}_m$ over a fixed sieve $\mathcal{F}_m$ is such that

$$\mathrm{E}\mathcal{L}\left(f_{\hat{\theta}_m}, f\right) \leqslant \mathcal{L}\left(f_{\theta_m}, f\right) + c_1(1 + \|\Sigma\|_{\text{spec}}^2)\frac{d_m}{T}$$

where c_1 is a positive finite constant depending on $\kappa_1, K, k_\infty, \tilde{v}, K_2, K_2', K_\infty$.

The proof is to be found in Section 5. The second result is about the estimator $\hat{\theta}_{\hat{m}}$ computed from the model selection procedure described above. We first need the following assumption on the number of sieves.

Assumption 3.3. There exists some weights L_m and a finite constant Υ such that

$$\sum_{m \in \mathcal{N}_{D,T}} \exp(-L_m d_m) \leqslant \Upsilon < \infty \, .$$

When the collection of models has at most one model per dimension, the weights L_m can be constant. A nonconstant L_m is essentially needed to prevent the situation where the dimension of the models does not grow fastly enough. See Barron *et al.* (1999) for details.

Theorem 3.2. *Suppose that we observe data $X_{1,T}, \dots, X_{T,T}$ from a Gaussian locally stationary process and suppose that Assumptions 2.1 and 3.1 to 3.3 hold true. For all $m \in \mathcal{N}_{D,T}$, define $\theta_m = \arg\min_{\theta \in \mathcal{F}_m} \mathcal{L}\left(f_\theta, f\right)$. If the penalty function* pen$(\cdot)$ *is such that*

$$\text{pen}(m) \geqslant c_3 \frac{d_m}{T} + c_4 \frac{d_m(1+L_m)}{T} \|\Sigma\|^2_{\text{spec}} \tag{3.4}$$

then the estimator $\hat{\theta}_{\hat{m}}$ defined in (3.2) is such that

$$\mathrm{E}\mathcal{L}\left(f_{\hat{\theta}_{\hat{m}}}, f\right) \leqslant \inf_{m \in \mathcal{N}_{D,T}} \left\{\mathcal{L}\left(f_{\theta_m}, f\right) + \text{pen}(m)\right\} + c_5 \frac{\Upsilon \|\Sigma\|^2_{\text{spec}} + 1}{T}$$

where c_3, c_4 are positive finite coefficients depending on $\kappa_1, K, k_\infty, \tilde{v}, K_2, K_2', K_\infty, C_{\mathcal{T}}$ and c_5 is a positive, finite constance depending on $\kappa_1, K, k_\infty, \tilde{v}, K_2, K_2', K_\infty$.

This theorem shows that the selection of the sieve $\mathcal{F}_{\hat{m}}$ among all sieves $\{\mathcal{F}_m; m \in \mathcal{N}_{D,T}\}$ leads to an estimator of the spectral density $\hat{f}_{\hat{m}}$ that performs as well as the best estimator $f_{\hat{\theta}_m}$ among $m \in \mathcal{N}_{D,T}$. The price to pay for this adaptation appears through the sequence L_m and the constant of the T^{-1} term, which is different in the two theorems. Note that the specific form for c_3 and c_4 is complicated but can be derived from the proof section, see in particular (5.3). It is important to note that they do not depend on the second-order quantities (ESD, time-varying autocovariance function) of the time series.

We now discuss the important situation where the true ESD f takes the semiparametric form f_{θ° for a given $\theta^\circ \in \Theta$, where Θ is a class of time-varying curves with D component (D is known). This corresponds to the correctly specified situation and one way to measure the quality of the estimation procedure is to consider the norm $\|\cdot\|_2$ defined above instead of the Kullback-Leibler divergence. With the L^2 norm, similar results than Theorems 3.1–3.2 can be derived. However, as there is no equivalence between the Kullback-Leibler divergence and the L^2 norm, it is worth saying that this result is not a corollary of the two theorems. Therefore an explicit proof is needed, but this proof can be adapted from the proof of the two above theorems.

Proposition 3.1. *Suppose that we observe data $X_{1,T},\dots,X_{T,T}$ from a Gaussian locally stationary process with evolutionary spectral density f_{θ°, where θ° is a time-varying D-dimensional curve. Suppose that Assumptions 2.1 and 3.1 to 3.3 hold true and set*

$$\theta_m = \arg\min_{\theta\in\mathcal{F}_m} \mathcal{L}(f_\theta, f_{\theta^\circ})$$

for all $m \in \mathcal{N}_{D,T}$. Then,

(a) *the minimum contrast estimator $\hat\theta_m$ over a fixed sieve $\mathcal{F}_m$ (see (3.1)) is such that*

$$\mathrm{E}\|\hat\theta_m - \theta^\circ\|_2^2 \leqslant \|\theta_m - \theta^\circ\|_2^2 + c_6(1 + \|\Sigma\|_{\text{spec}}^2)\frac{d_m}{T}$$

for all $m \in \mathcal{N}_{D,T}$;

(b) *if the penalty function* $\text{pen}(\cdot)$ *is such that (3.4) holds true, then the estimator $\hat\theta_{\hat m}$ defined in (3.2) is such that*

$$\mathrm{E}\|\hat\theta_{\hat m} - \theta^\circ\|_2^2 \leqslant \inf_{m\in\mathcal{N}_{D,T}} \left\{\|\theta_m - \theta^\circ\|_2^2 + \text{pen}(m)\right\} + c_7\frac{\Upsilon\|\Sigma\|_{\text{spec}}^2 + 1}{T}$$

where c_6, c_7 are positive, finite coefficients depending on $\kappa_1, K, k_\infty, \tilde v, K_2, K_2', K_\infty$.

Again, a comparison between the two results (a) and (b) shows that the automatic selection of the index m does not increase the estimation error significantly. Moreover, from this proposition, it is easy to derive an adaptation result with respect to the unknown smoothness of the curve θ°. Let $\beta > 0$, we recall that a function $g \in L_2([0,1])$ belongs to the Besov space $B_\infty^{\beta,2}$ if it satisfies

$$\|g\|_{\beta,2} = \sup_{u>0} u^{-\beta}\omega_d(g,u)_2 < \infty \qquad d = [\beta] + 1$$

where $\omega_d(g,u)_2$ is the modulus of continuity defined by $\omega_d(g,u)_2 = \sup_{|h|\leqslant u}\|\Delta_h^2 g\|_2$ where $\Delta_h g(x) = g(x-h) - g(x)$ and $\Delta_h^2 g = \Delta_h\Delta_h g$. Let us suppose that each component $\theta^{\circ(i)}$ of the target curve belongs to a Besov space $B_\infty^{\beta_i,2}$. If we consider the trigonometric orthogonal systems or the piecewise polynomial model, it is known from approximation theory (De Vore and Lorentz, 1993) that if $r \geqslant \beta$, then $\|\theta^{\circ(i)} - \theta_m^{(i)}\|_2 \leqslant C(\beta)\|\theta^{\circ(i)}\|_{\beta,2} d_{m_i}^{-\beta_i}$, where r is the regularity of the polynomial model. For these models, $L_m = 1$ and the proposition leads to the following corollary.

Corollary 3.1. *Suppose that we observe data $X_{1,T},\dots,X_{T,T}$ from a Gaussian locally stationary process with evolutionary spectral density f_{θ°, where θ° is a time-varying D-dimensional curve. Suppose in addition that each component $\theta^{\circ(i)}$ of the target curve belongs to a Besov space $B_\infty^{\beta_i,2}$. Under Assumptions 2.1 and 3.1 to 3.3, the estimator (3.2) is such that*

$$\mathrm{E}\|\hat\theta_{\hat m} - \theta^\circ\|_2^2 \leqslant c_8 T^{-\frac{2\beta}{2\beta+1}}$$

where $\beta = \min\{\beta_1,\dots,\beta_D\}$ and c_8 depends on $\kappa_1, K, k_\infty, \tilde v, K_2, K_2', K_\infty$ and $\|\Sigma\|_{\text{spec}}$.

If the model is correctly specified, this result gives the rate of convergence of the estimator to the true target curve. If only one curve has to be estimated ($D = 1$), this result gives the usual rate of convergence in Besov smoothness classes. If more that one curve has to be estimated, the global risk is bounded at a rate corresponding to the least smooth class $\beta = \min\{\beta_1, \ldots, \beta_D\}$. Moulines, Priouret and Roueff (2006, Theorem 4) have proved that this is the optimal rate of convergence for time-varying AR models in certain Lipschitz-spaces. We conjecture that this is also the optimal rate in the above Besov-spaces if all β_i are the same.

4 Fitting time-varying autoregressive models

In this section, we focus on the particular situation of fitting a tvAR(p) model to non-stationary data. The model then takes the form (1.1) and the target curve is denoted by $\theta(\cdot) = (\theta^{(0)}(\cdot), \theta^{(1)}(\cdot), \ldots, \theta^{(p)}(\cdot))$ with $\theta^{(0)}(\cdot) = \sigma^2(\cdot)$ and $\theta^{(i)}(\cdot) = a_i(\cdot)$, $i = 1, \ldots, p$. The model selection procedure presented in Section 3 can be adapted to the situation where the order p is unknown but bounded from above by a given nonnegative integer P. In such a case, we need to define the sieve as follow: Each component $\theta^{(i)}(\cdot)$ is approximated in a linear finite dimensional space S_m, $m \in \mathcal{M}_T$ where, with a slight change of notation, we set $\mathcal{N}_{p,T} = \{(p, m_0, \ldots, m_p), m_j \in \mathcal{M}_T\}$ and, for each $m = (p, m_0, \ldots, m_p) \in \mathcal{N}_{p,T}$ we define $\mathcal{F}_m = S_{m_0} \otimes \ldots \otimes S_{m_p}$. The set of approximation spaces is then defined by the set of indexes $\mathcal{N}_T = \cup_{j=1}^{P} \mathcal{N}_{j,T}$ (see also Baraud *et al.* (2001)).

The evolutionary spectral density of a tvAR(p) is

$$f_{\theta(u)}(\lambda) = \frac{\sigma^2(u)}{2\pi} \cdot \frac{1}{|\sum_{j=0}^{p} a_j(u) \exp(i\lambda j)|^2}$$

see Dahlhaus (1996b). With this particular form of spectrum and the Kolmogorov's formula, we obtain after some straightforward calculations

$$\mathcal{L}_T(f_\theta, J_T) = \frac{1}{2T} \sum_{t=1}^{T} \left[\log \sigma^2 \left(\frac{t}{T}\right) + \frac{1}{\sigma^2\left(\frac{t}{T}\right)} \times \right.$$
$$\left. \times \left\{ \left(\Gamma_{t,T} a\left(\frac{t}{T}\right) + C_{t,T} \right)' \Gamma_{t,T}^{-1} \left(\Gamma_{t,T} a\left(\frac{t}{T}\right) + C_{t,T} \right) + c_T\left(\frac{t}{T}, 0\right) - C'_{t,T} \Gamma_{t,T}^{-1} C_{t,T} \right\} \right] \tag{4.1}$$

with

$$a\left(\frac{t}{T}\right) = \left(a_1\left(\frac{t}{T}\right), \ldots, a_p\left(\frac{t}{T}\right) \right)',$$

$$c_T\left(\frac{t}{T}, j\right) = \int_{-\pi}^{\pi} d\lambda\ J_T\left(\frac{t}{T}, \lambda\right) \exp(i\lambda j) = X_{[t+\frac{j+1}{2}]} X_{[t-\frac{j-1}{2}]},$$

$$C_{t,T} = \left(c_T\left(\frac{t}{T}, 1\right), \ldots, c_T\left(\frac{t}{T}, p\right)\right)' ,$$

$$\Gamma_{t,T} = \left\{c_T\left(\frac{t}{T}, j-k\right)\right\}_{j,k=1,\ldots,p} .$$

In the following we consider the practical implementation of the model selection procedure.

4.1 Model selection with a stationary innovation process

If we assume that $\sigma^2(u)$ is constant over time, an explicit formula can be written for the estimator $\hat{\theta}_m(u)$ with a fixed $m \in \mathcal{N}_T$. This derivation is an extension of the expansion of Dahlhaus (1997, equations (4.3)-(4.4)), where the localized periodogram of Dahlhaus (1997) is replaced by the preperiodogram J_T. We therefore skip the details of the derivation that leads to the estimator.

For the sake of simplicity, we start by considering the case where the dimension of S_{m_j} does not depend on j, i.e. we fit the model $a_j(u) = \sum_{k=0}^{d_m-1} \theta_{jk}\varphi_k(u)$, where $d_m = \dim(S_m)$. Let $\theta = (\theta_{1,0}, \ldots, \theta_{1,d_m-1}, \ldots, \theta_{p,d_m-1})'$. Let $\Phi(\cdot)$ be the matrix $\{\varphi_j(\cdot)\varphi_k(\cdot)\}_{j,k=0,\ldots,d_m-1}$ and set $\varphi(\cdot) = (\varphi_0(\cdot), \ldots, \varphi_{d_m-1}(\cdot))'$. If $A \otimes B$ denotes the left direct product of the matrices A and B, the parameters that minimize $\mathcal{L}_T(f_\theta, J_T)$ are given by

$$\hat{\theta}_m = -\left(\frac{1}{T}\sum_{t=1}^{T} \Phi\left(\frac{t}{T}\right) \otimes \Gamma_{t,T}\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} \varphi\left(\frac{t}{T}\right) \otimes C_{t,T}\right) \tag{4.2}$$

and

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T} c_T\left(\frac{t}{T}, 0\right) + \frac{1}{T}\,\hat{\theta}_m' \sum_{t=1}^{T} \varphi\left(\frac{t}{T}\right) \otimes C_{t,T} . \tag{4.3}$$

The resulting system is similar to the Yule-Walker equations. If the dimension of S_{m_j} depends on j, i.e. if different spaces S_j are used to fit different curves $a_j(u)$, the estimator is obtained similarly after deleting the corresponding columns and rows in

$$\frac{1}{T}\sum_{t=1}^{T} \Phi\left(\frac{t}{T}\right) \otimes \Sigma_{t,T}$$

and

$$\frac{1}{T}\sum_{t=1}^{T} \varphi\left(\frac{t}{T}\right) \otimes C_{t,T} .$$

We now apply the estimation procedure based on model selection developed in Section 3. The contrast function is given by (4.1) and, from Theorem 3.2, the penalty is set to

$$\mathrm{pen}(m) = c_3\frac{d_m}{T} + c_4\frac{d_m(1+L_m)}{T}\|\Sigma\|_{\mathrm{spec}} .$$

The implementation of our procedure requires the pre-estimation of $\|\Sigma\|_{\text{spec}}$ and the computation of c_3 and c_4. In our simulations, we compute $\|\Sigma\|_{\text{spec}}$ following the method of Van Bellegem and von Sachs (2003, 2004). First, we notice that the entry (t,j) of the matrix Σ is given by the covariance operator $c(t/T,j) = \int d\lambda f(t/T,\lambda)\exp(i\lambda j)$. For a given lag j, a pre-estimator of $c(u,j)$ is then given by the smoothing of $c_T(u,j)$ with respect to $u \in [0,1]$. If $\hat{c}_T(u,j)$ denotes this smoothed curve, we then estimate Σ by

$$\hat{\Sigma}_{s,t} = \hat{c}_T\left(\frac{s+t}{2T}, |s-t|\right) I(|s-t| \leqslant M)$$

where M is a prescribed nonnegative integer. The indicator function $I(|s-t| \leqslant M)$ sets to zero all $\hat{\Sigma}_{s,t}$ with $|s-t| > M$. The indicator function appears because it is expected that the covariance function $c(u,j)$ tends to zero for large lags j (Assumption 2.1(e)). Therefore it reduces the variance of the pre-estimation. $\|\Sigma\|_{\text{spec}}$ is then estimated by computing the largest eigenvalue of the symmetric matrix $\hat{\Sigma}$. For more details about the choice of the tuning parameters (choice of M, bandwidth selection in the estimation of $\hat{c}_T(u,j)$) and the properties of this pre-estimator, we refer to the more exhaustive study of Van Bellegem and von Sachs (2003, 2004).

Constants c_3 and c_4 are explicitely given in the proof of the results but they are difficult to compute explicitely since they depend on some constants of the model (such as K_2 for instance). Therefore, we first need an initial calibration step to fix c_3 and c_4. However, Theorem 3.2 ensures that the results are optimal if we consider upper bounds for c_3 and c_4. This means that the results are very robust to a large choice of c_3, c_4. In the simulations of this paper, we used $c_3 = c_4 = 1$. One option is to select these constants from a grid of prescribed values in a data-driven way, based for instance on the out-of-sample properties of the estimator.

We now compare our model selection procedure with a selection based on the Akaike information criterion

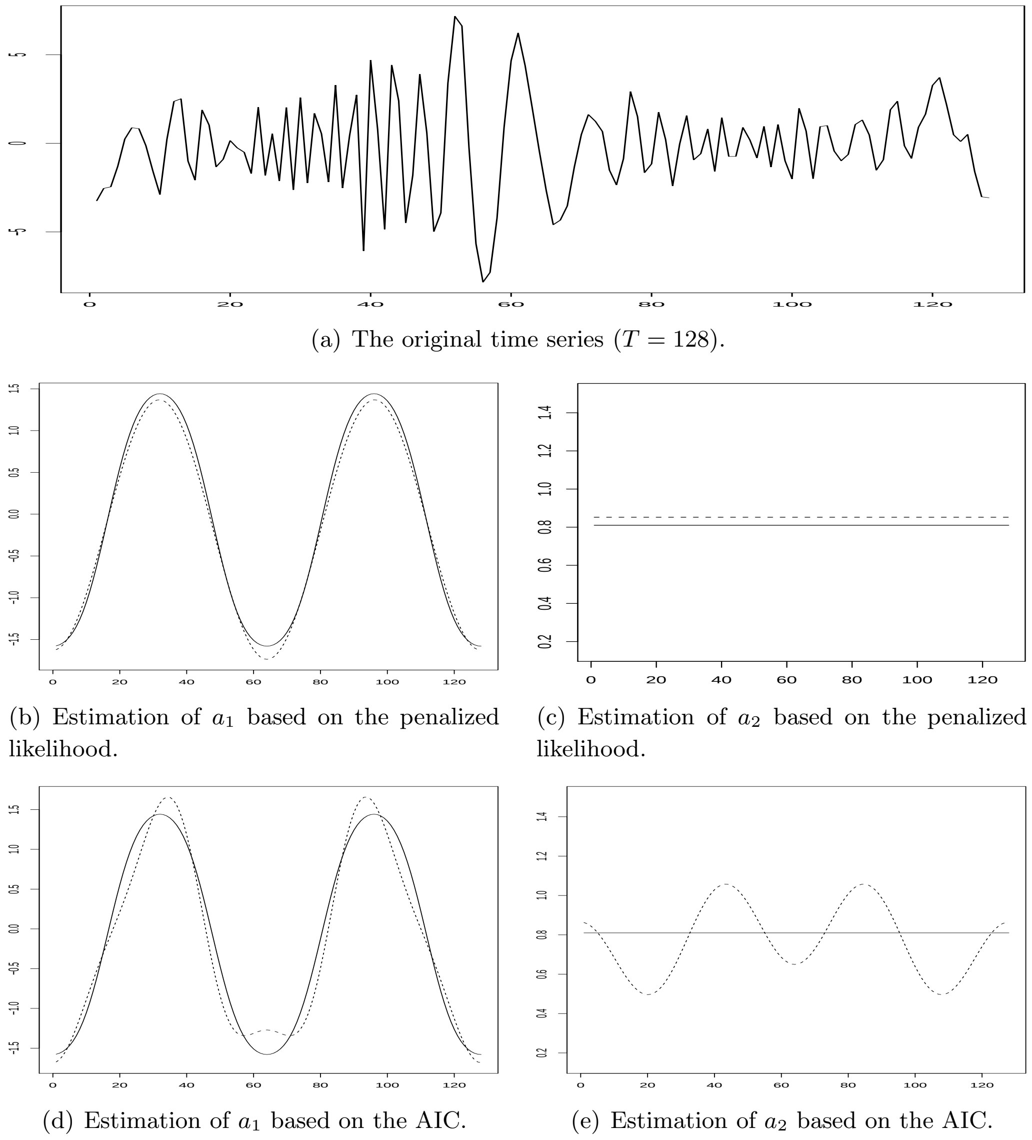
$$\text{AIC}(m) := \log\hat{\sigma}^2(m) + \frac{2}{T}\left(1 + p + \sum_{i=1}^{p} d_{m_i}\right), \qquad m = (p, m_1, \ldots, m_p).$$

This form of the AIC has been proposed in Dahlhaus (1997, Section 6) and illustrated through simulations on one example of a tvAR(2) process. We consider the same example in our first simulation.

Example 4.1. Consider the tvAR(2) process (model (1.1) with $p = 2$) with parameters $\sigma(u) \equiv 1$, $a_1(u) = -1.8\cos(1.5 - \cos 4\pi u)$ and $a_2(u) \equiv 0.81$ and with a stationary Gaussian innovation process ε_t with unit variance. One realisation of this process is given in Figure 1(a) with $T = 128$ data.

		$T = 64$		$T = 128$	
		Pen. Lik.	AIC	Pen. Lik.	AIC
Model selection:					
	d_{m_1}	2.80	4.67	2.99	4.47
		(0.424)	(1.658)	(0.010)	(1.625)
	d_{m_2}	0.84	3.32	1.02	3.44
		(0.136)	(5.008)	(0.101)	(4.289)
Mean quadratic error:					
	$a_1(u)$	0.18	0.31	0.07	0.21
		(0.150)	(0.320)	(0.296)	(0.512)
	$a_2(u)$	1.87	2.34	2.037	2.218
		(0.127)	(2.063)	(0.113)	(1.172)
Mean absolute deviation:					
	$a_1(u)$	0.26	0.37	0.13	0.24
		(0.073)	(0.042)	(0.046)	(0.861)
	$a_2(u)$	1.14	1.21	1.17	1.18
		(0.005)	(0.048)	(0.003)	(0.023)
Mean square prediction error:		1.53	2.01	1.47	1.54
		(3.55)	(5.58)	(3.23)	(4.54)
Mean absolute prediction error:		0.92	1.03	0.89	0.91
		(0.13)	(0.19)	(0.14)	(0.15)

Table 1: Simulations are based on 100 generations of a tvAR(2) process of sample size $T = 64$ and $T = 128$ (Example 4.1) . The "Model selection" row presents the mean of the orders d_{m_i} $(i = 1, 2)$ for our method (Pen. Lik) and the AIC method. The "Mean quadratic error" row shows the mean of the square error $T^{-1} \sum_t (\hat{a}_i(t/T) - a_i(t/T))^2$ $(i = 1, 2)$ while the "Mean absolute deviation" presents the mean of the error $T^{-1} \sum_t |\hat{a}_i(t/T) - a_i(t/T)|$ $(i = 1, 2)$. The "Mean square prediction error" evaluates the mean over all samples of $(T - p)^{-1} \hat{\sigma}^{-2} \sum_{t=p+1}^{T} (X_{t,T} - \sum_{j=1}^{p} \hat{a}_j(t/T) X_{t-j,T})^2$ and the "Mean absolute prediction error" computes the mean over all samples of $(T-p)^{-1} |\hat{\sigma}|^{-1} \sum_{t=p+1}^{T} |X_{t,T} - \sum_{j=1}^{p} \hat{a}_j(t/T) X_{t-j,T}|$. In all rows, numbers in parenthesis are the variance computed from the 100 samples. Remark: In row d_{m_2}, the value $0.84 < 1$ appears since the procedure sometimes select the order $p = 1$. In such a case, the dimension d_{m_2} is set to zero.



(a) The original time series ($T = 128$).

(b) Estimation of a_1 based on the penalized likelihood.

(c) Estimation of a_2 based on the penalized likelihood.

(d) Estimation of a_1 based on the AIC.

(e) Estimation of a_2 based on the AIC.

Figure 1: In each figure, the solid line plots the true curves a_i $(i = 1, 2)$ and the dotted line is the estimator. This example is based on a simulation of length $T = 128$. The orders selected by the penalized likelihood method are $(d_{m_1}, d_{m_2}) = (3, 1)$ while the AIC method selected $(6, 4)$.

In our estimation procedure we use trigonometric sieves as described in Example 3.1 above. It is worth mentioning that the curve $a_1(u)$ cannot be written as a finite linear combination of trigonometric functions and we are therefore in a misspecified case. Table 1 presents the results of a Monte-Carlo simulation based on 100 generations of the tvAR process of sample size $T = 64$ and 128, and with $P = 2$. This table compares the order selection based on our method with the selection based on the AIC. It also computes the error of estimation, based on the mean squared error, the mean absolute deviation error and the mean square prediction error.

Table 1 shows that the models selected by our penalized likelihood method are near $(d_{m_1}, d_{m2}) = (3, 1)$ while the models selected by AIC are around $(4.5, 3.5)$. From the specific tvAR(2) we considered, it is clear that the true order of the second curve is $d_{m_2} = 1$. This table then reveals how the AIC overfits, while the penalized likelihood does not.

If we consider the estimation of the curve $a_1(u)$, there is no true order. To have an idea which order provides the best fitting, we consider our estimator applied on the simulation plotted in Figure 1(a). In this particular simulation, the penalized likelihood method selected the models $(3, 1)$ while the AIC selected $(6, 4)$. This figure confirms that the model $(3, 1)$ gives a better fit, while the AIC overfits. The mean quadratic error and mean absolute deviation computed in Table 1 confirm the better performance of the estimator based on the penalized likelihood criterion.

4.2 Model selection with a nonstationary innovation process

If the function $\sigma^2(u)$ is not constant over time, explicit formula for the estimators cannot be easily derived. For this general situation, we propose an iterative procedure of estimation that we will now describe.

Step I (Initialisation). Compute initial estimators $\hat{\theta}_T$ and $\hat{\sigma}_T^2$ from the above formula (4.2–4.3).

Step II (Update of $\hat{\sigma}_T^2$). Given $\hat{\theta}_T^{(1)}$, compute the vector $s_T^2 = (s_T^2(t/T))_{t=1,\dots,T}$ that minimizes the likelihood (4.1) evaluated with $a_j(u) = \hat{a}_j^{(1)}(u) := \sum_k \hat{\theta}_{jk}^{(1)} \varphi_j(u)$:

$$s_T^2\left(\frac{t}{T}\right) = \left(\Gamma_{t,T}\hat{a}^{(1)}\left(\frac{t}{T}\right) + C_{t,T}\right)' \Gamma_{t,T}^{-1}\left(\Gamma_{t,T}\hat{a}^{(1)}\left(\frac{t}{T}\right) + C_{t,T}\right) + c_T\left(\frac{t}{T}, 0\right) - C_{t,T}'\Gamma_{t,T}^{-1}C_{t,T}.$$

If d_0 denotes the dimension of the sieve $\mathcal{F}_{d_0}$ on which σ^2 is estimated, then update $\hat{\sigma}_T^2$ to the curve that smoothes s_T^2 over the space $\mathcal{F}_{d_0}$, i.e.

$$\hat{\sigma}_T^2(t/T) = \sum_{j=0}^{d_0-1} \hat{\alpha}_j \varphi_j(t/T)$$

where the vector $\hat{\alpha} = (\hat{\alpha}_0, \dots, \hat{\alpha}_{d_0-1})$ is such that $\hat{\alpha} = (\Delta'\Delta)^{-1}\Delta'\hat{s_T}^2$ with $\Delta_{it} = \varphi_i(t/T)$.

Step III (Update of $\hat{\theta}_T$). Given $\hat{\sigma}_T^2$, update $\hat{\theta}_T$ to the value that minimizes the likelihood (4.1) computed with $\sigma^2 = \hat{\sigma}_T^2$. This leads to

$$\hat{\theta}_T^{(2)} = -\left(\frac{1}{T}\sum_{t=1}^{T}\frac{\Phi\left(\frac{t}{T}\right)\otimes\Gamma_{t,T}}{\hat{\sigma}_T^2\left(\frac{t}{T}\right)}\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\frac{\varphi\left(\frac{t}{T}\right)\otimes C_{t,T}}{\hat{\sigma}_T^2\left(\frac{t}{T}\right)}\right)$$

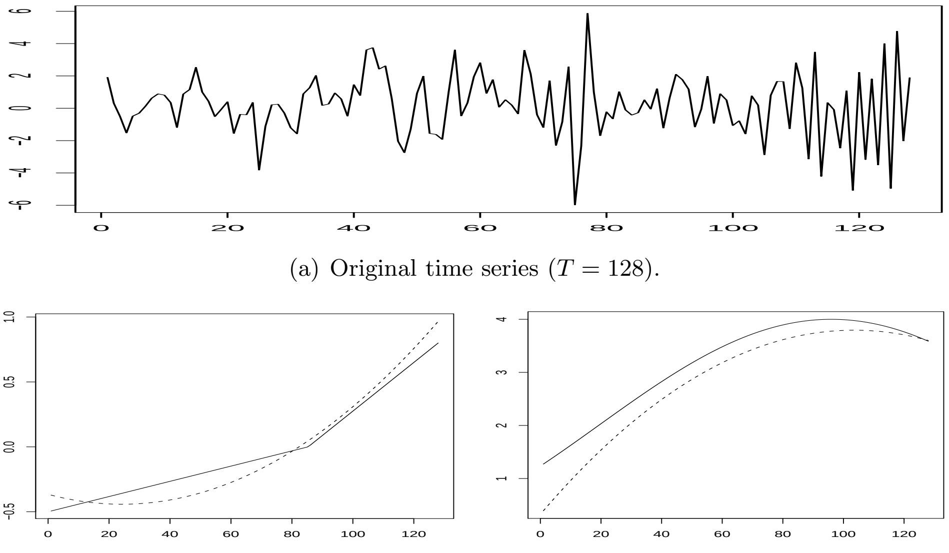
Step IV (Loop). Iterate steps II and III until convergence.

We illustrate this procedure on a second simulation, based on the following tvAR model with time-varying innovations.

Example 4.2. Consider the tvAR(1) model (1.1) with $p = 1$, $a_1(u) = (3u/4 - 1/2)I(u \leqslant 2/3) + (12u/5 - 8/5)I(u > 2/3)$ and the evolutionary variance given by $\sigma^2(u) = -2\cos(6\pi(u + .45)/5) + 2$. Figure 2 shows the result of one simulation based on $T = 128$ data.

The analysis considers a sieve generated by the Legendre polynomials, which define an orthonormal basis of polynomial functions. In this example, we then work once again in a misspecified case since $a_1(u)$ and $\sigma^2(u)$ cannot be written as a finite linear combination of Legendre polynomials. For the simulation of Figure 2(a), the procedure selected the models $d_0 = 3$ for the estimation of σ^2 and $d_1 = 3$ for the estimation of a_1. As we can see from the plot, the quality of the fit is remarkable, given that the estimators are computed from $T = 128$ data only.

In Table 2 we report the results of a Monte-Carlo simulation that aims to study what model is selected by the procedure, and what is the influence of the sample sizes T for this model selection. We consider three different sample size, $T = 64, 128$ and 256 and simulate 100 times the TVAR(1) process with nonstationary innovations. The table indicates the frequency of selection of a given model (d_0, d_1). The corresponding error associated with this Monte-Carlo simulation is showed in Figure 3.



(a) Original time series ($T = 128$).

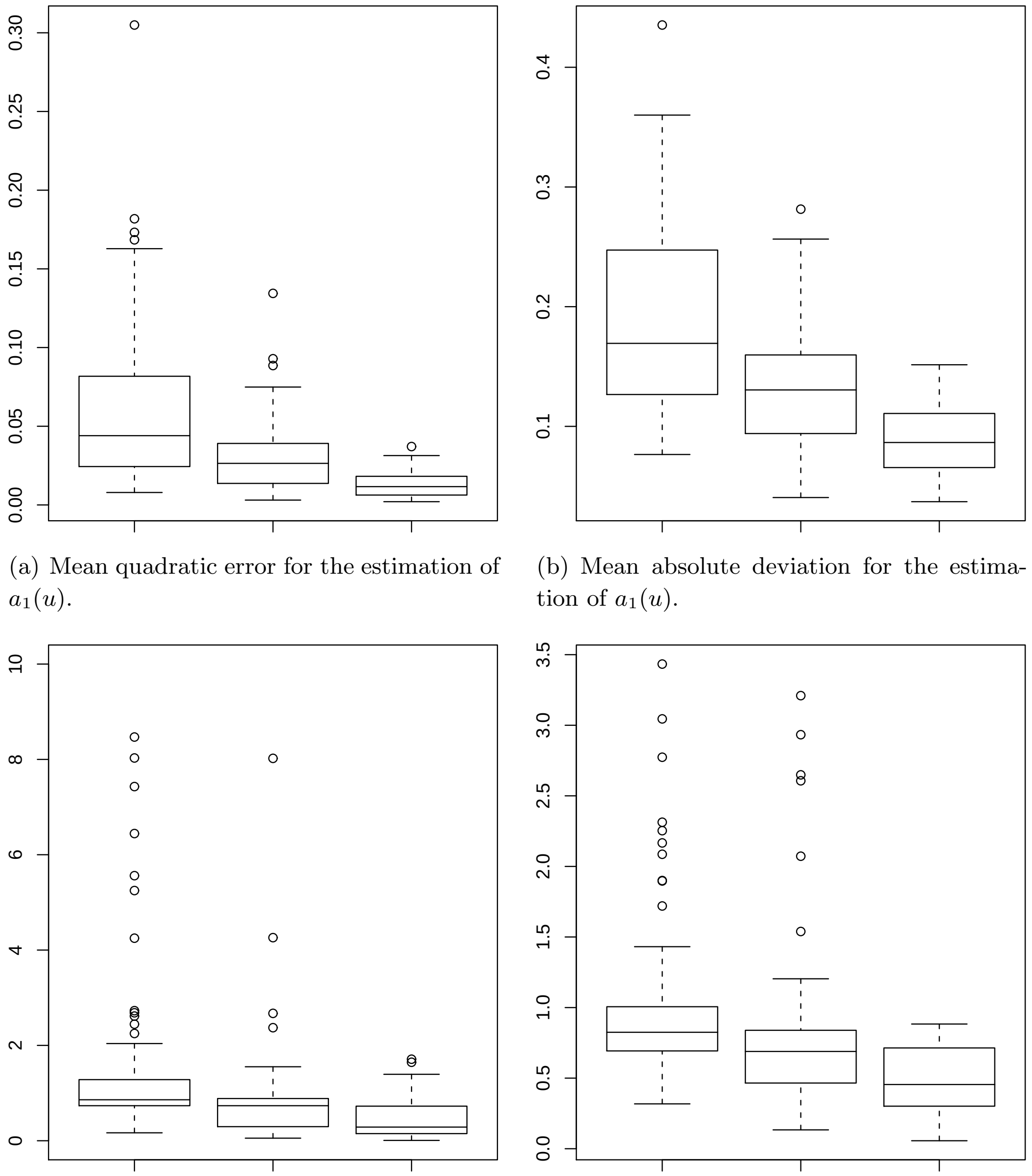
(b) Time-varying coefficient $a_1(u)$ and its estimator.

(c) Time-varying variance of the innovations ($\sigma^2(u)$) and its estimator.

Figure 2: These plots show the result of one simulation of a tvar(1) process with time-varying innovations. The solid lines show the curves $a_1(u)$ and $\sigma^2(u)$. Estimators based on the sample (a) are superimposed in dotted lines. The models considered in this estimation procedure are constructed using Legendre polynomials.

		Order of $\hat{a}_1(u)$			
		1	*2*	*3*	*4*
Order of $\hat{\sigma}^2(u)$	*1*	1 0 0	20 7 1	13 13 9	21 17 14
	2	1 0 0	7 8 7	8 14 14	3 10 6
	3	1 1 0	7 7 5	8 9 11	3 7 15
	4	0 0 0	1 2 3	0 3 10	1 2 5

Table 2: The table shows the frequency of selection of a given model from data. Simulations are based on 100 generations of a tvAR(1) process with nonstationary innovations. The coefficients of the process are plotted in Figure 2 (b) and (c). The three numbers in each cell correspond to three sample sizes: $T = 64$ (upper number), $T = 128$ or $T = 256$ (bottom number).



(a) Mean quadratic error for the estimation of $a_1(u)$.

(b) Mean absolute deviation for the estimation of $a_1(u)$.

(c) Mean quadratic error for the estimation of $\sigma^2(u)$.

(d) Mean absolute deviation for the estimation of $\sigma^2(u)$.

Figure 3: Error of estimation from 100 generations of the tvAR(1) process with nonstationary innovations. The three boxplot in each subfigure correspond to three different sample sizes ($T = 64, 128$ and 256).

5 Proofs

5.1 Proof of Theorem 3.1

As usual in the context of minimum contrast estimation on sieves, the key point is to establish maximal exponential bounds for the fluctuation of the empirical process. In the context of locally stationary processes, the *empirical spectral process* is defined as

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

where

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

and

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

The connection between this empirical process and the contrast functions has been derived by Dahlhaus and Polonik (2006): as $f_{\hat{\theta}_m}$ minimizes $\mathcal{L}_T(f_\theta, J_T)$ and f_{θ_m} minimizes $\mathcal{L}(f_\theta, f)$ over $\theta \in \mathcal{F}_m$, we can write

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

where

$$R(\theta) := \frac{1}{4\pi}\int_{-\pi}^{\pi} d\lambda\ \left\{\frac{1}{T}\sum_{t=1}^{T}\log f_{\theta(t/T)}\left(\lambda\right) - \int_0^1 du\ \log f_{\theta(u)}(\lambda)\right\}.$$

Assumption 2.1 implies the existence of a positive, finite constant κ such that $\sup_{\theta\in\mathcal{F}_m}|R(\theta)| \leqslant \kappa/(2T)$.Then we can write

$$\mathcal{L}\left(f_{\hat{\theta}_m}, f\right) \leqslant \mathcal{L}\left(f_{\theta_m}, f\right) + \frac{1}{4\pi\sqrt{T}} E_T\left(\frac{1}{f_{\theta_m}} - \frac{1}{f_{\hat{\theta}_m}}\right) + \frac{\kappa}{T}.$$

We now decompose the empirical spectral process as $E_T = \tilde{E}_T + \overline{E}_T$, where $\tilde{E}_T = \sqrt{T}\,(F_T - \mathrm{E}F_T)$ is a stochastic term while $\overline{E}_T = \sqrt{T}\,(\mathrm{E}F_T - F)$ is a deterministic term. Theorem 4.1 of Dahlhaus and Polonik (2006) implies $|\overline{E}_T(\phi)| \leqslant \overline{K}(\rho_\infty(\phi) + \tilde{v}(\phi))/(2\sqrt{T})$ for some finite, positive constant $\overline{K}$. Then we get with Assumption 3.1

$$\mathcal{L}\left(f_{\hat{\theta}_m}, f\right) \leqslant \mathcal{L}\left(f_{\theta_m}, f\right) + \frac{1}{4\pi\sqrt{T}} \tilde{E}_T\left(\frac{1}{f_{\theta_m}} - \frac{1}{f_{\hat{\theta}_m}}\right) + \frac{\kappa + \overline{K}(k_\infty + \tilde{v})}{2T}.$$

Let $1 - p(\omega)$ be the probability of the event $\mathcal{A}$ defined by

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

where k is a positive constant that will be specified later on. On $\mathcal{A}_\omega$, we can write with Assumption 3.2(a):

$$\mathcal{L}\left(f_{\hat{\theta}_m}, f\right) - \mathcal{L}\left(f_{\theta_m}, f\right) \leqslant \frac{k}{4\pi} \left\{ \omega^2 + K_2' \rho_2 \left(\frac{1}{f_{\theta_m}} - \frac{1}{f_{\hat{\theta}_m}} \right)^2 \right\} + \frac{\kappa + \overline{K}(k_\infty + \tilde{v})}{2T}. \quad (5.1)$$

We now make use of the following lemma, quoted from Dahlhaus and Polonik (2006).

Lemma 5.1. *If the class $\mathcal{F}_m$ is such that θ_m exists and is unique, if $\rho_\infty(1/f_\theta)$ and $\rho_2(1/f_\theta)$ are uniformly bounded under $\theta \in \mathcal{F}_m$, if the set $\mathcal{F}_m^\star = \{1/f; f \in \mathcal{F}_m\}$ is convex then there exists a constant $\alpha > 0$ such that*

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

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

If we choose $k = 2\pi(\alpha K_2')^{-1}$ and rearrange the inequality (5.1) to get

$$\mathcal{L}\left(f_{\hat{\theta}_m}, f\right) - \mathcal{L}\left(f_{\theta_m}, f\right) \leqslant \frac{\omega^2}{K_2'\alpha} + \frac{\kappa_1 + K(k_\infty + \tilde{v})}{T}$$

a.s. on $\mathcal{A}_\omega$. If we denote $V = \mathcal{L}(f_{\hat{\theta}_m}, f) - \mathcal{L}\left(f_{\theta_m}, f\right) - T^{-1}(\kappa_1 + K(k_\infty + \tilde{v}))$, then $V \leqslant \omega^2 (K_2'\alpha)^{-1}$ a.s. on $\mathcal{A}_\omega$ with $\Pr(\mathcal{A}_\omega) = 1 - p(\omega)$. In consequence, $\Pr(V > \omega^2) \leqslant p(\omega\sqrt{K_2'\alpha})$ and we get

$$\mathrm{E}(V) = \int_0^\infty dx \Pr(V > x) \leqslant \int_0^\infty dx p\left(\sqrt{K_2'\alpha x}\right) = \frac{2}{K_2'\alpha} \int_0^\infty dy\, yp(y).$$

We now make use of a maximal inequality for the empirical process stated in the appendix (Lemma A.1), which implies, with $k = 2\pi(\alpha K_2')^{-1}$,

$$\mathrm{E}(V) \leqslant \frac{k}{\pi} \left\{ \int_0^{\omega_{d_m}(k)} dy\, yp(y) + \int_{\omega_{d_m}(k)}^\infty dy\, yp(y) \right\} \leqslant \frac{k}{\pi} \omega_{d_m}^2(k) + \frac{8e^2}{(e-1)^2} \frac{K_2^2 \|\Sigma\|_{\text{spec}}^2}{Tk^2 K_\infty^2 \overline{r}_m^2}$$

where the function $\omega_d(\cdot)$ is defined in equation (A.1). □

5.2 Proof of Theorem 3.2

Set $m^\star = \arg\min_{m \in \mathcal{N}_{D,T}} \mathcal{L}(f_{\theta_m}, f)$ and fix (ν, m') such that $\nu \in \mathcal{F}_m$ and $\mathcal{L}_T(f_\nu, J_T) + \mathrm{pen}(m') \leqslant \mathcal{L}_T(f_{\theta_m}, J_T) + \mathrm{pen}(m)$ for all $m \in \mathcal{N}_{D,T}$. For all $m \in \mathcal{N}_{D,T}$ we can write

$$\begin{aligned} 0 &\leqslant \mathcal{L}\left(f_\nu, f\right) - \mathcal{L}\left(f_{\theta_{m^\star}}, f\right) \\ &\leqslant \mathcal{L}\left(f_\nu, f\right) - \mathcal{L}\left(f_{\theta_{m^\star}}, f\right) + \mathcal{L}_T\left(f_{\theta_m}, J_T\right) - \mathcal{L}_T\left(f_\nu, J_T\right) + \mathrm{pen}(m) - \mathrm{pen}(m') \\ &\leqslant \frac{1}{4\pi\sqrt{T}} \tilde{E}_T \left(\frac{1}{\theta_m} - \frac{1}{f_\nu} \right) + R_T + U_m + \mathrm{pen}(m) - \mathrm{pen}(m') \qquad (5.2) \end{aligned}$$

where $U_m = \mathcal{L}(f_{\theta_m}, f) - \mathcal{L}(f_{\theta_{m^\star}}, f) \geqslant 0$ and $R_T = (4\pi T)^{-1}(\kappa_1 + K(k_\infty + \tilde{v}))$, as in the proof of Theorem 3.1.

Now, we fix $m \in \mathcal{N}_{D,T}$. For all $m' \in \mathcal{N}_{D,T}$, define

$$\underline{\omega}^2_{m'}(y) = \omega^2_{d_m}\left(\frac{2\pi}{\alpha K_2'}\right) \vee \omega^2_{d_{m'}}\left(\frac{2\pi}{\alpha K_2'}\right) \vee \left\{\zeta\left(\frac{2\pi}{\alpha K_2'}\right)\frac{L_m d_m \vee L_{m'} d_{m'}}{T}\right\} + \frac{y}{T}$$

for $y \geqslant 1$, where $\omega_{d_m}(\cdot)$ is defined in (A.1), α is defined in Lemma 5.1 and $\zeta(\cdot)$ is defined in Lemma A.2 of the Appendix. Let $p(y)$ be the probability of the set

$$\mathcal{A}_y = \left\{\sup_{m' \in \mathcal{N}_{D,T}} \sup_{\nu \in \mathcal{F}_{m'}} \frac{\left|\tilde{E}_T\left(\frac{1}{f_{\theta_m}} - \frac{1}{f_\nu}\right)\right|}{\|\theta_{m^\star} - \theta_m\|_2^2 \vee \|\theta_{m^\star} - \nu\|_2^2 \vee \underline{\omega}^2_{m'}(y)} > \frac{2\pi\sqrt{T}}{\alpha K_2'}\right\}.$$

Using Assumption 3.2(a) and Lemma 5.1, we can write, a.s. on $\mathcal{A}_y^c$,

$$\frac{1}{4\pi\sqrt{T}}\left|\tilde{E}_T\left(\frac{1}{\theta_m} - \frac{1}{f_\nu}\right)\right| \leqslant \frac{1}{2}\mathcal{L}(f_\nu, f) + \frac{1}{2}\mathcal{L}(f_{\theta_m}, f) - \mathcal{L}(f_{\theta_{m^\star}}, f) + \frac{1}{2\alpha K_2'}\underline{\omega}^2_{m'}(y)$$

and, if we rearrange the inequality (5.2), this implies that the minimum penalized likelihood estimator $\hat{\theta}_{\hat{m}}$ satisfies

$$\mathcal{L}\left(f_{\hat{\theta}_{\hat{m}}}, f\right) \leqslant \mathcal{L}(f_{\theta_m}, f) + \frac{1}{2\alpha K_2'}\underline{\omega}^2_{\hat{m}}(y) + 2R_T + 2U_m + 2\,\mathrm{pen}(m) - 2\,\mathrm{pen}(\hat{m})$$

a.s. on $\mathcal{A}_y^c$. We choose c_3 and c_4 in the theorem such that the penalty function fullfils

$$2\,\mathrm{pen}(m) \geqslant (2\alpha K_2')^{-1}\left\{\omega^2_{d_m}\left(\frac{16\pi^2}{\tau}\right) \vee \frac{\zeta L_m d_m}{T}\right\} + 2R_T \tag{5.3}$$

which implies $(2\alpha K_2')^{-1}\underline{\omega}^2_{\hat{m}}(y) + 2R_T \leqslant 2\,\mathrm{pen}(m) + 2\,\mathrm{pen}(\hat{m}) + y/(\alpha K_2' T)$. Then we get

$$\mathcal{L}\left(f_{\hat{\theta}_{\hat{m}}}, f\right) \leqslant \mathcal{L}(f_{\theta_m}, f) + 4\,\mathrm{pen}(m) + \frac{y}{2\alpha K_2' T} + 2U_m$$

a.s. on $\mathcal{A}_y^c$. The random variable

$$V = \left\{\mathcal{L}\left(f_{\hat{\theta}_{\hat{m}}}, f\right) - \mathcal{L}(f_{\theta_m}, f) - 2U_m - 4\,\mathrm{pen}(m)\right\} \vee 0.$$

is such that $V \leqslant y/(2\alpha K_2' T)$ a.s. on $\mathcal{A}_y^c$ with $\Pr(\mathcal{A}_y) = p(y)$. Thus, if $y > 1$, $\Pr(V > y/(2\alpha K_2' T)) \leqslant p(y)$ and we can write, for any $m \in \mathcal{N}_{D,T}$,

$$\mathrm{E}(V) = (2\alpha K_2' T)^{-1}\left(1 + \int_1^\infty dy p(y)\right).$$

Lemma A.2 of the appendix and Assumption 3.3 allow to bound $p(y)$ as follows:

$$p(y) \leqslant 3.6 \sum_{m' \in \mathcal{N}_{D,T}} \exp\left(-\frac{(\zeta L_{m'} d_{m'} + y)}{\zeta}\right) \leqslant 3.6\Upsilon \exp\left(-\frac{y}{\zeta}\right)$$

where $\zeta := \zeta(2\pi/\alpha K_2')$. This implies $\mathrm{E}(V) \leqslant (2\alpha K_2' T)^{-1}(1 + 3.6\Upsilon)$ and we conclude. □

APPENDIX

A Auxiliary results

The following lemma is a maximal inequality on the empirical spectral process over a certain class of functions $\mathcal{F}_m$. $\mathcal{F}_m$ is a finite-dimensional linear space of the form $S_{m_1} \otimes \ldots \otimes S_{m_d}$ such that Assumption 3.1 is fulfilled. We also denote by d_m the dimension of $\mathcal{F}_m$.

Lemma A.1 (Maximal Inequality I)**.** *Under Assumptions 2.1, 3.1 and 3.2, for all* $\gamma \in \mathcal{F}_m$*,*

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

provided that $\omega^2 \geqslant \omega_{d_m}^2(\tau)$ *with*

$$\omega_{d_m}^2(\tau) = \frac{d_m}{T} \left\{ 1 \vee \frac{c_7}{\tau^2} \|\Sigma\|_{\text{spec}}^2 \right\} \tag{A.1}$$

where c_7 *is a positive, finite coefficient depending on* $\tilde{v}, K_2, K_\infty$ *and* $\overline{r}_m$*.*

This key result helps for controlling the fluctuation of the empirical spectral process. It is a generalisation of Theorem 5 of Birgé and Massart (1998), who proved a similar result for the empirical process of an i.i.d. sequence.

The next lemma states a maximal exponential inequality when the empirical spectral process involves vectors in two different sieves $\mathcal{F}_m$ and $\mathcal{F}_{m'}$.

Lemma A.2 (Maximal inequality II)**.** *Define* $m^\star = \arg\min_{m \in \mathcal{N}_{D,T}} D(f_{\theta_m}, f)$*. Under Assumptions 2.1, 3.1 and 3.2, for all indices* $m, m' \in \mathcal{N}_{p,T}$ *and for all* $\theta \in \mathcal{F}_m$*, the inequality*

$$\Pr\left\{ \sup_{\nu \in \mathcal{F}_{m'}} \frac{\left| \tilde{E}_T \left(\frac{1}{f_\theta} - \frac{1}{f_\nu} \right) \right|}{\|\theta_{m^\star} - \theta\|_2^2 \vee \|\theta_{m^\star} - \nu\|_2^2 \vee \omega^2} > \tau \sqrt{T} \right\} \leqslant 3.6 \exp\left(- \frac{T \omega^2}{\zeta(\tau)} \right)$$

holds true provided that $\omega > \omega_{d_m}(\tau) \vee \omega_{d_{m'}}(\tau)$ *(where the function* $\omega_{d_m}(\cdot)$ *is defined in (A.1)), where*

$$\zeta(\tau) := \frac{k_\infty K_\infty}{\tau} \|\Sigma\|_{\text{spec}} + \frac{4}{\tau^2} \left(\frac{4 K_2^2}{K_\infty^2 \overline{r}_m^2} + 2\pi K_2 + 4\pi K_\infty k_\infty \tilde{v} \right) \|\Sigma\|_{\text{spec}} \, .$$

The usual way for proving maximal inequalities is to start with a Bernstein inequality and use the chaining technique, provided that the complexity (entropy) of $\mathcal{F}_m$ is well controlled. We follow this scheme in our proof, and start by quoting two useful results. The first one is a Bernstein inequality derived in Dahlhaus and Polonik (2006) and the second one allows to control the complexity of the approximation space.

Lemma A.3 (Dahlhaus and Polonik, 2006). *Suppose that $\{X_{t,T}\}$ is a Gaussian 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 $\tilde{v}(\phi) < \infty$. 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 $\tilde{E}_T = \sqrt{T}(F_T - \mathrm{E}F_T)$ (see Section 5). Then the inequality

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

holds true for all $\epsilon > 0$.

The lemma is actually not exactly formulated as in Dahlhaus and Polonik (2006), but is a straightforward application of their Theorem 4.1. Note that

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

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 A.3:

$$\Pr\left(|\tilde{E}_T(\phi)| \geqslant \eta \right) \leqslant \exp\left(-\frac{1}{4} \cdot \frac{\eta^2}{2\pi\|\Sigma^{1/2}\|_{\text{spec}}^4 \rho_{2,T}^2(\phi) + \|\Sigma^{1/2}\|_{\text{spec}}^2 \frac{\rho_\infty(\phi)\eta}{\sqrt{T}}} \right) \tag{A.3}$$

for all $\eta > 0$.

The next lemma is a straightforward extension of Lemma 9 in Barron *et al.* (1999).

Lemma A.4. *Suppose that $\mathcal{F}_m$ is a finite-dimensional linear space of the form $S_{m_1} \otimes \ldots \otimes S_{m_D}$ such that Assumption 3.1 holds and denote by d_m the dimension of $\mathcal{F}_m$. Then, for any positive δ one can find a countable set $\mathcal{E}(\delta) \subset \mathcal{F}_m$ and a mapping $\mu : \mathcal{F}_m \to \mathcal{E}(\delta)$ such that*

(a) *For each ball $\mathcal{B}$ in $\mathbb{R}^d$ with radius $\omega \geqslant 5\delta$, $|\mathcal{E}(\delta) \cap \mathcal{B}| \leqslant (5\omega/\delta)^{d_m}$,*

(b) $\|\theta - \mu(\theta)\|_2 \leqslant \delta$ *for all $\theta \in \mathcal{F}_m$,*

(c) $\sup_{t \in \mathcal{E}(\delta)} \|t - \mu^{-1}(t)\|_\infty \leqslant \overline{r}_m \delta$ *for all $t \in \mathcal{E}(\delta)$, where $\overline{r}_m$ is defined in (3.3).*

where the norms are defined in Section 3.3.

We can now prove the two maximal inequalities. The following proofs use chaining argument and contains similar techniques to the proofs of Barron *et al.* (1999); Birgé and Massart (1998).

B Proof of Lemma A.1

Fix γ in $\mathcal{F}_m$. The proof proceeds in two steps. We shall first prove a maximal inequality on a ball $\mathcal{B}(\gamma,\omega)$ centered in γ with radius $\omega>0$, included in $\mathcal{F}_m$, i.e. an exponential bound for

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

In a second step, we extend the exponential inequality to the whole space $\mathcal{F}_m$.

B.1 Inequality on a ball $\mathcal{B}(\gamma,\omega)$

Chaining. We first define all ingredients of the chaining argument, in which we use a sequence $\delta_k = 2^{-k}\delta_0$, $k=0,1,\ldots$, where δ_0 will be fixed below. From the above Lemma A.4, there exists a sequence of subsets $\mathcal{E}(\delta_k)\subset\mathcal{F}_m$ such that $5\delta_k\leqslant\omega$ and

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

Given some point $\theta\in\mathcal{B}(\gamma,\omega)$, the sequence (θ_k) is such that $\theta_k\to\theta$ in the L^2 and the L^∞ norms. Therefore by Assumption 3.2 we have the decomposition

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

If we choose a sequence $(\xi_k)_{k\geqslant 0}$ such that

$$\sum_{k=0}^{\infty}\xi_k \leqslant \frac{K_2}{K_\infty\overline{r}_m}\|\Sigma\|_{\mathrm{spec}}\xi\omega^2, \tag{B.1}$$

we can write, by linearity of $\phi\to\tilde{E}_T(\phi)$,

$$\begin{aligned}\mathcal{P} &\leqslant \sum_{\theta_0\in\mathcal{E}(\delta_0)} \Pr\left\{\left|\tilde{E}_T\left(\frac{1}{f_{\theta_0}}-\frac{1}{f_\gamma}\right)\right| > \xi_0\sqrt{T}\right\} \\ &\qquad + \sum_{k=1}^{\infty} \sum_{\substack{\theta_k\in\mathcal{E}(\delta_k)\\ \theta_{k-1}\in\mathcal{E}(\delta_{k-1})}} \Pr\left\{\left|\tilde{E}_T\left(\frac{1}{f_{\theta_k}}-\frac{1}{f_{\theta_{k-1}}}\right)\right| > \xi_k\sqrt{T}\right\} \\ &=: P_0 + \sum_{k=1}^{\infty} P_k. \end{aligned}\tag{B.2}$$

In the following, we define a particular sequence (ξ_k) such that (B.1) holds and that leads to the required exponential bound for $\mathcal{P}$.

Define $H_k := \ln |\mathcal{E}(\delta_k)|$ and consider a positive sequence $(\eta_k)_{k=0,1,2,\ldots}$ that will be fixed below. Using the Bernstein inequality (Lemma A.3), we get $P_0 \leqslant \exp(H_0 - T\eta_0)$ provided that η_0 is such that

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

and $P_k \leqslant \exp(H_k + H_{k-1} - T\eta_k)$ provided that η_k, $k \geqslant 1$ are such that

$$\xi_k = 2\|\Sigma^{1/2}\|_{\text{spec}}^2 \left\{ 2\eta_0 \, \rho_\infty \left(\frac{1}{f_{\theta_k}} - \frac{1}{f_{\theta_{k-1}}} \right) + \sqrt{2\pi\eta_k} \, \rho_{2,T} \left(\frac{1}{f_{\theta_k}} - \frac{1}{f_{\theta_{k-1}}} \right) \right\}$$

for $k \geqslant 1$.

We now fix the sequence $(\eta_k)_{k=0,1,2,\ldots}$ as follows. Set L such that the inequality $L \geqslant \xi^2 \vee L'$ holds where L' is implicitely given by the equation $L' = 2\ln\{5\alpha(L')\}$ with

$$\alpha(L') := (1 \vee c_* \vee \beta_m) + \frac{K_\infty \overline{r}_m}{K_2} \sqrt{c_* \left(1 + \frac{d_m L'}{T} \overline{r}_m^2 \right)}$$

where $c_* := (6 \cdot 50^2 \pi \tilde{v} K_\infty) \vee (120^2 K_\infty^2) \vee (2 \cdot 54^2 \pi K_2^2)$ and

$$\beta_m := C_{\overline{r}} \frac{K_\infty^4 \overline{r}_m^4}{K_2^4} \left\{ 1 + \frac{K_2^2 c_*}{K_\infty^2 \overline{r}_m^2} + \overline{r}_m^2 \frac{d_m L}{T} \right\}.$$

Define $\eta_0 = (H_0 + d_m L)/T$ and $\eta_k = \{H_k + H_{k-1} + (k+1) d_m L\}/T$ for $k \geqslant 1$.

With the above definitions, one can check by straightforward (but long) algebra that if we choose $\delta_0 = \omega/\alpha(L)$, then the inequality

$$\sum_{k \geqslant 0} \xi_k \leqslant \|\Sigma\|_{\text{spec}} \frac{\omega K_2}{K_\infty \overline{r}_m} \sqrt{\frac{d_m L}{T}} \tag{B.3}$$

holds true provided that $\omega \geqslant \overline{r}_m/(C_{\overline{r}} T)$.

Maximal inequality if $\omega = \xi^{-1}\sqrt{d_m L/T}$. Assume that the radius of the ball is such that $\xi\omega = \sqrt{d_m L/T}$. It implies $\omega \geqslant \sqrt{d_m/T}$, thus it fulfills the constraint $\omega \geqslant \overline{r}_m/(C_{\overline{r}} T)$ by Assumption 3.1 (b) and using that $d_m \geqslant 1$. Therefore, by (B.3), the condition (B.1) is satisfied. Under this constraint, we can derive a maximal exponential inequality. From (B.2) and the above calibration of the chaining,

$$\begin{aligned} \mathcal{P} &\leqslant \exp(-d_m L) \left\{ 1 + \sum_{k=1}^{\infty} \exp(-k d_m L) \right\} \leqslant \exp(-d_m L) \left\{ 1 - \exp(-d_m L) \right\}^{-1} \\ &\leqslant e(e-1)^{-1} \exp(-d_m L) = e(e-1)^{-1} \exp(-\omega^2 \xi^2 T) \end{aligned} \tag{B.4}$$

where we used $d_m L/2 \geqslant 1$ since $d_m \geqslant 1$, $\alpha \geqslant 1$ and then $L \geqslant 2$.

Exponential bound for $\mathcal{P}$. The previous paragraph shows a maximal inequality on the ball $\mathcal{B}(\gamma,\omega)$ under the constraint $\omega\xi = \sqrt{d_m L/T}$, i.e. $\omega^2 \geqslant d_m T^{-1}\{1 \vee \xi^{-2}L'\}$. Using the inequalities $\ln(|x|+|y|) \leqslant (\ln|2x|)\vee(\ln|2y|)$ and $\ln|x| \leqslant |x|/e$, we derive

$$L' \leqslant \frac{2e}{e-1}\ln\left\{10(1+c_*+\beta_m)+\frac{10K_\infty \overline{r}_m c_*^{1/2}}{K_2}\right\} \vee \frac{2e}{e-1}\ln\left\{\frac{10K_\infty\overline{r}_m^2}{K_2}\sqrt{c_*\frac{d_m}{T}}\right\} .$$

Assumption 3.1 implies $\sqrt{d_m/T} \leqslant C_{\overline{r}}/\overline{r}_m$ then we conclude that the exponential inequality holds for all $\omega^2 \geqslant d_m T^{-1}\{1 \vee \xi^{-2}A\}$ where $A = (2e/e-1)\ln\{10(1+c_*+\beta_m) + 10K_\infty K_2^{-1}\overline{r}_m c_*^{1/2}(1+C_{\overline{r}})\}$.

B.2 Inequality over $\mathcal{F}_m$

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

$$\begin{aligned}
&\Pr\left\{\sup_{\theta\in\mathcal{F}_m}\frac{\left|\tilde{E}_T\left(\frac{1}{f_\theta}-\frac{1}{f_\gamma}\right)\right|}{\omega^2\vee\|\theta-\gamma\|_2^2} > \tau\sqrt{T}\right\}\\
&\qquad\leqslant \sum_{j=0}^{\infty}\Pr\left\{\sup_{\theta\in\mathcal{F}_m;\omega_j^2\leqslant\|\theta-\gamma\|_2^2<\omega_{j+1}^2}\frac{\left|\tilde{E}_T\left(\frac{1}{f_\theta}-\frac{1}{f_\gamma}\right)\right|}{\omega_{j+1}^2/4} > \tau\sqrt{T}\right\}\\
&\qquad\leqslant \sum_{j=0}^{\infty}\Pr\left\{\sup_{\theta\in\mathcal{B}(\gamma,\omega_{j+1})}\left|\tilde{E}_T\left(\frac{1}{f_\theta}-\frac{1}{f_\gamma}\right)\right| > \omega_{j+1}^2\tau\sqrt{T}/4\right\} . \qquad \text{(B.5)}
\end{aligned}$$

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

$$\omega^2 \geqslant \frac{d_m}{T}\left\{1\vee\frac{K_2^2\|\Sigma\|_{\text{spec}}^2}{K_\infty^2\overline{r}_m^2\tau^2}A\right\} \qquad \text{(B.6)}$$

holds true, then we bound (B.5) from above by:

$$\frac{e}{e-1}\sum_{j=0}^{\infty}\exp\left(-\frac{T\tau^2K_\infty^2\overline{r}_m^2\omega_{j+1}^2}{4K_2^2\|\Sigma\|_{\text{spec}}^2}\right) \leqslant \frac{e^2}{(e-1)^2}\exp\left(-\frac{T\tau^2K_\infty^2\overline{r}_m^2\omega^2}{4K_2^2\|\Sigma\|_{\text{spec}}^2}\right).$$

since (B.6) with $d_m \geqslant 1$ and $A \geqslant 1$ implies that $K_\infty^2\overline{r}_m^2\tau^2\omega^2T \geqslant K_2^2\|\Sigma\|_{\text{spec}}^2$. The lemma follows with $c_7 := AK_2^2/(K_\infty^2\overline{r}_m^2)$. □

C Proof of Lemma A.2

Write $s := \|\Sigma^{1/2}\|_{\text{spec}}^2 = \|\Sigma\|_{\text{spec}}$. From Lemma A.1, it holds

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

for all $\gamma \in \mathcal{F}_{m'}$ provided that $\omega^2 \geqslant \omega^2_{d_{m'}}(\tau)$. Moreover, the Bernstein inequality (A.3) allows to write

$$\Pr\left\{ \frac{\left|\tilde{E}_T\left(\frac{1}{f_\theta} - \frac{1}{f_\gamma}\right)\right|}{\|\theta-\gamma\|_2^2 \vee \omega^2} > \tau\sqrt{T} \right\} \leqslant \exp\left(-\frac{1}{4} \cdot \frac{T\tau^2(\|\theta-\gamma\|_2^2 \vee \omega^2)}{2\pi s^2 A^\circ_{m,m'} + Bs\tau} \right)$$

for all $\gamma \in \mathcal{F}_{m'}$ where, using Assumption 3.2 and (A.2),

$$A^\circ_{m,m'} := \frac{\rho^2_{2,T}\left(\frac{1}{f_\theta} - \frac{1}{f_\gamma}\right)}{\|\theta-\gamma\|_2^2 \vee \omega^2} \leqslant \frac{K_2\|\theta-\gamma\|_2^2 + T^{-1}K_\infty\|\theta-\gamma\|_\infty \tilde{v}}{\|\theta-\gamma\|_2^2 \vee \omega^2}$$

and with $B := \rho_\infty(1/f_\gamma - 1/f_\theta) \leqslant k_\infty K_\infty$ by Assumptions 3.1 and 3.2. Moreover, as $\omega^2 \geqslant d_{m'}/T$ and $d_m \geqslant 1$, we get $A^\circ_{m,m'} \leqslant K_2 + 2K_\infty k_\infty \tilde{v}$.

Then, we can write

$$\Pr\left\{ \frac{\left|\tilde{E}_T\left(\frac{1}{f_\theta} - \frac{1}{f_\gamma}\right)\right|}{\|\theta-\gamma\|_2^2 \vee \omega^2} > \tau\sqrt{T} \right\} \leqslant \exp\left(-\frac{1}{4} \cdot \frac{T\tau^2\omega^2}{2\pi A_{m,m'} s^2 + Bs\tau} \right).$$

We finally get, for all $\gamma \in \mathcal{F}_{m'}$,

$$\Pr\left\{ \sup_{\nu \in \mathcal{F}_{m'}} \frac{\left|\tilde{E}_T\left(\frac{1}{f_\theta} - \frac{1}{f_\nu}\right)\right|}{\|\gamma-\theta\|_2^2 \vee \|\gamma-\nu\|_2^2 \vee \omega^2} > \tau\sqrt{T} \right\}$$
$$\leqslant \left(1 + \frac{e^2}{(e-1)^2}\right) \exp\left(-\frac{1}{4} \cdot \frac{T\tau^2\omega^2}{\left(\frac{4K_2^2}{K_\infty^2 \bar{r}_m^2} \vee 2\pi A_{m,m'}\right) s^2 + Bs\tau} \right).$$

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

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

and this implies

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

and this argument holds for an arbitrary $\varepsilon > 0$. □

References

Baraud, Y., Comte, F. and Viennet, G. (2001). Adaptive estimation in autoregression or β-mixing regression via model selection. *Ann. Statist.*, *29*, 839–875.

Barron, A. R., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields*, *113*, 301–413.

Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, *4*, 329–375.

Comte, F. (2001). Adaptive estimation of the spectrum of a stationary Gaussian sequence. *Bernoulli*, *7*, 267–298.

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

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

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

Dahlhaus, R. (2000). A likelihood approximation for locally stationary processes. *Ann. Statist.*, *28*, 1762–1794.

Dahlhaus, R. and Neumann, M. H. (2001). Locally adaptive fitting of semiparametric models to nonstationary time series. *Stoch. Process. Applic.*, *91*, 277–308.

Dahlhaus, R. and Polonik, W. (2006). Nonparametric quasi maximum likelihood estimation for Gaussian locally stationary processes. *Ann. Statist.* (forthcoming)

Davis, R., Lee, T. and Rodriguez-Yam, G. (2006). Structural break estimation for nonstationary time series models. *J. Am. Statist. Ass.*, *101*, 223–239.

De Vore, R. A. and Lorentz, G. G. (1993). *Constructive approximation.* Berlin: Springer.

Grenier, Y. (1983). Time-dependent ARMA modeling of nonstationary signals. *IEEE Trans. Acoust. Speech Signal Process*, *31*, 899–911.

Moulines, É., Priouret, P. and Roueff, F. (2006). On recursive estimation for time-varying autoregressive processes. *Ann. Statist.* (forthcoming)

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

Neumann, M. and von Sachs, R. (1997). Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *Ann. Statist.*, *25*, 38–76.

Ombao, H., von Sachs, R. and Guo, W. (2005). SLEX analysis of multivariate nonstationary time series. *J. Am. Statist. Ass.*, *470*, 519–531.

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

Subba Rao, T. (1970). The fitting of non-stationary time-series models with time-dependent parameters. *J. R. Statist. Soc.* B, *32*, 312–322.

Van Bellegem, S. and von Sachs, R. (2003). *Locally adaptive estimation of evolutionary wavelet spectra.* (Tech. Rep.)

Van Bellegem, S. and von Sachs, R. (2004). On adaptive estimation for locally stationary wavelet processes and its applications. *Int. J. Wavelets Multiresolut. Inf. Process.*, *2*, 545–565.