Bernoulli **5**(5), 1999, 873–906

Nonlinear wavelet estimation of time-varying autoregressive processes

RAINER DAHLHAUS,[1] MICHAEL H. NEUMANN[2] and RAINER VON SACHS[3]

[1]*Institut für Angewandte Mathematik, Universität Heidelberg, Im Neuenheimer Feld 294, D-69120 Heidelberg, Germany. E-mail: dahlhaus@statlab.uni-heidelberg.de*
[2]*SFB 373, Humboldt-Universität zu Berlin, Spandauer Strasse 1, D-10178 Berlin, Germany. E-mail: neumann@wiwi.hu-berlin.de*
[3]*Institut de Statistique, Université Catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium. E-mail: rvs@stat.ucl.ac.be*

We consider nonparametric estimation of the parameter functions $a_i(\cdot)$, $i = 1, \ldots, p$, of a time-varying autoregressive process. Choosing an orthonormal wavelet basis representation of the functions a_i, the empirical wavelet coefficients are derived from the time series data as the solution of a least-squares minimization problem. In order to allow the a_i to be functions of inhomogeneous regularity, we apply nonlinear thresholding to the empirical coefficients and obtain locally smoothed estimates of the a_i. We show that the resulting estimators attain the usual minimax L_2 rates up to a logarithmic factor, simultaneously in a large scale of Besov classes. The finite-sample behaviour of our procedure is demonstrated by application to two typical simulated examples.

Keywords: nonlinear thresholding; non-stationary processes; time series; time-varying autoregression; wavelet estimators

1. Introduction

Stationary models have always been the main focus of interest in the theoretical treatment of time series analysis. For several reasons autoregressive models form a very important class of stationary models: they can be used for modelling a wide variety of situations (for example, data which show a periodic behaviour); there exist several efficient estimates which can be calculated via simple algorithms (Levinson–Durbin algorithm, Burg algorithm); and the asymptotic properties, including the properties of model selection criteria, are well understood.

Frequently, people have also tried to use autoregressive models for modelling data that show a certain type of non-stationary behaviour by fitting such models on small segments. This method is often used, for example, in signal analysis for coding a signal (linear predictive coding) or for modelling data in speech analysis. The underlying assumption then is that the data are coming from an autoregressive process with time-varying coefficients.

Suppose we have some observations $\{X_1, \ldots, X_T\}$ from a zero-mean autoregressive process with time-varying coefficients $a_1(\cdot), \ldots, a_p(\cdot)$. To obtain a tractable framework for our asymptotic analysis we assume that the functions a_i are supported on the interval [0, 1]

1350–7265

and connected to the underlying time series by an appropriate rescaling. This leads to the model

$$X_{t,T} + \sum_{i=1}^{p} a_i(t/T) X_{t-i,T} = \sigma(t/T)\varepsilon_t, \qquad t = 1, \ldots, T, \tag{1.1}$$

where the ε_ts are independent, identically distributed with $\mathrm{E}\varepsilon_t = 0$ and $\mathrm{var}(\varepsilon_t) = 1$. To make this definition complete, assume that $X_0, \ldots, X_{1-p}$ are random variables from a stationary AR(p) process with parameters $a_1(0), \ldots, a_p(0)$. As usual in nonparametric regression, we focus on estimating the functions a_i in full, although, strictly speaking, the intermediate values $a_i(s)$, for $(t-1)/T < s < t/T$, are not identifiable. This time-varying autoregressive model is a special locally stationary process as defined in Dahlhaus (1997). However, for the main results of this paper we only use the representation (1.1) and not the general properties, like analogue of Cramér's representation a non-stationary (see Dahlhaus (1997, p. 3)), for example, of a locally stationary process.

The estimation problem now consists of estimating the parameter functions $a_i(\cdot)$. Very often these functions are estimated at a fixed time point t_0/T by fitting a stationary model in a neighbourhood of t_0, for example, by estimating $a_1(t_0/T), \ldots, a_p(t_0/T)$ with the classical Yule–Walker (or Burg) estimate over the segment $X_{t_0-N,T}, \ldots, X_{t_0+N,T}$, where N/T is small. This method has the disadvantage that it automatically leads to a smooth estimate of $a_i(\cdot)$. Sudden changes in the $a_i(\cdot)$, common as they are for example in signal analysis, cannot be detected by this method. Moreover, the performance of this method depends on the appropriate choice of the segmentation parameter N. Instead, in this paper we develop an automatic alternative, which avoids this a priori choice and adapts to local smoothness characteristics of the $a_i(\cdot)$.

Our approach consists in a nonlinear wavelet method for the estimation of the coefficients $a_i(\cdot)$. This concept, based on orthogonal series expansions, has recently been brought into the nonparametric regression estimation problem by Donoho and Johnstone (1998), and has been proven very useful if the class of functions to be estimated exhibits a varying degree of smoothness. Some generalizations can be found in Brillinger (1994), Johnstone and Silverman (1997), Neumann and Spokoiny (1995) and Neumann and von Sachs (1995). As usual, the unknown functions, $a_i(u)$, are expanded by orthogonal series with respect to a specially chosen orthonormal basis of $L_2[0, 1]$, a *wavelet* basis. Basically, the basis functions are generated by dilations and translations of the so-called scaling function ϕ and wavelet function ψ, which are both localized in spatial position (here temporal) and frequency. These basis functions, unlike most of the 'traditional' ones (Fourier, (non-local) polynomials, and so on), are able to optimally compress both functions with quite homogeneous smoothness over the whole domain (like Hölder or L_2-Sobolev) as well as members of certain inhomogeneous smoothness classes (like L_p-Sobolev or Besov $B^m_{p,q}$ with $p < 2$). Note that the better compressed a signal is (that is, the smaller the number of coefficients representing it), the better the performance of an estimator of the signal which is optimally tuned with respect to bias–variance trade-off. A strong theoretical justification for the merits of using wavelet bases in this context has been given by Donoho (1993): it was shown that wavelets provide unconditional bases for a wide variety of these inhomogeneous smoothness classes with the result that wavelet estimators can be optimal in the above-mentioned sense.

To actually achieve this optimality there is need to nonlinearly modify traditional linear series estimation rules which are known to be optimal only in the case of homogeneous smoothness: there the coefficients of each resolution level j are essentially of the same order of magnitude, and the loss due to a levelwise inclusion/exclusion rule, as opposed to a componentwise rule, is only small. However, under strong inhomogeneity, the coefficients of each fixed level might not only differ considerably in their orders of magnitude but also have significant values on higher levels to be included by a suitably chosen inclusion rule. Surprisingly enough, this is possible by simple and intuitive schemes which are based on comparing the size of the empirical coefficients with their variability. Such nonlinear rules can dramatically outperform linear ones in cases where the vector of coefficients forms a sparse signal (that is, in cases of inhomogeneous function classes represented in an unconditional basis).

In this work, we apply these locally adaptive estimation procedures to the particular problem of estimating autoregression coefficients which are functions of time. A basic problem in this situation is to obtain adequate emprirical wavelet coefficients. For example, if one made a wavelet expansion of the function $\hat{a}_i(\cdot)$ where $\hat{a}_i(t_0/T)$ was the Yule–Walker estimate on a segment (as described above), then the information on irregularities of the functions $a_i(\cdot)$ would already be lost (since the segment is smooth). No thresholding procedure of the empirical wavelet coefficients would recover it.

To overcome this problem, we suggest in this paper using the empirical wavelet coefficients obtained from the solution of a least-squares minimization problem. In a second step, soft or hard thresholding is applied. We show that in this situation our nonlinear wavelet estimator attains the usual near-optimal minimax rate of L_2 convergence, in a large scale of Besov spaces, that is, classes of functions with different degrees of smoothness and different norms in which smoothness is measured. The full procedure requires consistent estimators for the variance of the empirical coefficients. In particular, a consistent estimator of the variance function is needed (cf. Section 3), for example the squared residuals of a local autoregressive model fit.

Finally, with this adaptive estimation of the time-varying autoregression coefficients, we immediately provide a semi-parametric estimate for the resulting time-dependent spectral density of the process given by (1.1). An alternative, fully nonparametric approach for estimating the so-called evolutionary spectrum of a general locally stationary process (as defined in Dahlhaus 1997) has been delivered by Neumann and von Sachs (1997), which is based on nonlinear thresholding in a two-dimensional wavelet basis.

The content of our paper is organized as follows. While in the next section we describe details of our set-up and present this main result, in Section 3 the statistical properties of the empirical coefficients are given. Section 4 shows the finite-sample behaviour of our procedure applied to two typical (simulated) time-varying autogressive processes. Section 5 deals with the proof of the main theorem. The remaining Sections 6–7 and the Appendix present some auxiliary results, both of interest in their own right and in this particular context used to derive the main proof (of Section 5).

2. Assumptions and the main result

Before we develop nonlinear wavelet estimators for the functions a_i, we describe the general set-up. First we introduce an appropriate orthonormal basis of $L_2[0, 1]$. Assume that we have a scaling function ϕ and a so-called wavelet ψ such that, defining $\phi_{lk}(x) = 2^{l/2}\phi(2^l x - k)$ and $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$, $\{\phi_{lk}(\cdot)\}_{k\in\mathbb{Z}} \cup \{\psi_{jk}(\cdot)\}_{j\geq l; k\in\mathbb{Z}}$ forms an orthonormal basis of $L_2(\mathbb{R})$. The construction of such functions ϕ and ψ, which are compactly supported, is described in Daubechies (1988). It is well known that the boundary-corrected Meyer wavelets (Meyer 1990) and those developed by Cohen, Daubechies and Vial (1993) form orthonormal bases of $L_2[0, 1]$. In both approaches Daubechies' wavelets are used to construct an orthonormal basis of $L_2[0, 1]$, essentially by truncation of the above functions to the interval $[0, 1]$ and a subsequent orthonormalization step. Throughout this paper either of these bases can be used, which we write as $\{\phi_{lk}\}_{k\in I_l^0} \cup \{\psi_{jk}\}_{j\geq l; k\in I_j}$. It is known that $\#I_j = 2^j$, and that $\#I_l^0 = 2^l$ for the Cohen–Daubechies–Vial (CDV) bases, whereas for the Meyer bases, $\#I_l^0 = 2^l + N$ for some integer N depending on the regularity of the wavelet basis. For reasons of notational simplicity, in what follows we restrict our attention to the CDV bases.

Accordingly, we can expand a_i in an orthogonal series

$$a_i = \sum_{k\in I_l^0} \alpha_{lk}^{(i)}\phi_{lk} + \sum_{j\geq l}\sum_{k\in I_j} \beta_{jk}^{(i)}\psi_{jk}, \tag{2.1}$$

where $\alpha_{lk}^{(i)} = \int \alpha_i(u)\phi_{lk}(u)\,\mathrm{d}u$, $\beta_{jk}^{(i)} = \int \alpha_i(u)\psi_{jk}(u)\,\mathrm{d}u$ are the usual Fourier coefficients, also called wavelet coefficients.

Assume a degree of smoothness m_i for the function a_i, that is, a_i is a member of a Besov class $B_{p_i,q_i}^{m_i}(C)$ defined below. In accordance with this, we choose compactly supported wavelet functions of regularity $r > m := \max\{m_i\}$, that is:

Assumption 1.

(i) ϕ and ψ are $C^r[0, 1]$ and have compact support.
(ii) $\int \phi(t)\,\mathrm{d}t = 1$, $\int \psi(t)t^k\,\mathrm{d}t = 0$ for $0 \leq k \leq r$.

The first step in each wavelet analysis is the definition of empirical versions of the wavelet coefficients. We define the empirical coefficients simply as a least-squares estimator, that is, as a minimizer of

$$\sum_{t=p+1}^{T}\left(X_{t,T} + \sum_{i=1}^{p}\left[\sum_{k\in I_l^0}\alpha_{lk}^{(i)}\phi_{lk}(t/T) + \sum_{j=l}^{j^*-1}\sum_{k\in I_j}\beta_{jk}^{(i)}\psi_{jk}(t/T)\right]X_{t-i,T}\right)^2, \tag{2.2}$$

where the choice of j^* will be specified below. Since $\{\phi_{lk}\}_k \cup \{\psi_{jk}\}_{l\leq j\leq j^*-1;k}$ forms a basis of the subspace V_{j^*} of $L_2[0, 1]$, this amounts to an approximation of a_i in just this space V_{j^*}.

In the present paper we propose to apply nonlinear smoothing rules to the coefficients $\tilde{\beta}_{jk}^{(i)}$. It is well known (cf. Donoho and Johnstone 1998) that linear estimators can be optimal with respect to the optimal rate of convergence as long as the underlying smoothness of a_i is not too inhomogeneous. This situation changes considerably if the smoothness varies strongly over the domain. Then we have the new effect that even at higher resolution scales a small number of coefficients cannot be neglected, whereas the overwhelming majority of them are much smaller than the noise level. This kind of sparsity of non-negligible coefficients is responsible for the need for a nonlinear estimation rule. Two commonly used rules to treat the coefficients are: hard thresholding,

$$\delta^{(\mathrm{h})}(\tilde{\beta}_{jk}^{(i)}, \lambda) = \tilde{\beta}_{jk}^{(i)} I(|\tilde{\beta}_{jk}^{(i)}| \geqslant \lambda);$$

and soft thresholding,

$$\delta^{(\mathrm{s})}(\tilde{\beta}_{jk}^{(i)}, \lambda) = (|\tilde{\beta}_{jk}^{(i)}| - \lambda)_+ \operatorname{sgn}(\tilde{\beta}_{jk}^{(i)}).$$

To treat these coefficients in a statistically appropriate manner, we have to tune the estimator in accordance with their distribution. It turns out that, at the finest resolution scales, this distribution actually depends on the (unknown) distribution of the $X_{t,T}$s, whereas we can hope to have asymptotic normality if $2^j = o(T)$. We show in Section 3 that we do not lose asymptotic efficiency of the estimator if we truncate the series at some level $j = j(T)$ with $2^{j(T)} \asymp T^{1/2}$. To give a definite rule, we choose the highest resolution level $j^* - 1$ such that $2^{j^*-1} \leqslant T^{1/2} < 2^{j^*}$, that is to say, we restrict our analysis to coefficients $\tilde{\alpha}_{lk}^{(i)}$ $(k \in I_l^0,\ i = 1, \ldots, p)$ and $\tilde{\beta}_{jk}^{(i)}$ $(j \geqslant l,\ 2^j \leqslant T^{1/2},\ k \in I_j,\ i = 1, \ldots, p)$. Unlike in ordinary regression, it is not possible in the autocorrelation problem considered here to include coefficients from resolution scales j up to $2^j = o(T)$. This is due to the fact that the empirical coefficients cannot be reduced to sums of independent (or sufficiently weakly dependent) random variables, which results in some additional bias term.

Finally, we build an estimator of a_i by applying the inverse wavelet transform to the nonlinearly modified coefficients.

Before we state our main result, we introduce some more assumptions. The constant C used here and in the following is assumed to be positive, but need not be the same at each occurrence.

Assumption 2. There exists some $\gamma \geqslant 0$ such that

$$|\mathrm{cum}_n(\varepsilon_t)| \leqslant C^n (n!)^{1+\gamma}, \qquad \text{for all } n,\ t.$$

Assumption 3. There exists a $\rho > 0$ with

$$1 + \sum_{i=1}^{p} a_i(s) z^i \neq 0, \qquad \text{for all } |z| \leqslant 1 + \rho \text{ and all } s \in [0, 1].$$

Furthermore, σ is assumed to be continuous with $C_1 \leqslant \sigma(s) \leqslant C_2$ on $[0, 1]$.

Remark 2.1. Note that, besides the obvious case of the normal distribution, many of the distributions that can be found in textbooks satisfy Assumption 2 for an appropriate choice of γ. In Johnson and Kotz (1970) we can find closed forms of higher-order cumulants of the exponential, gamma and inverse Gaussian distribution, which show that this condition is satisfied for $\gamma = 0$. The need for a positive γ occurs in the case of a heavier-tailed distribution, which could arise as the distribution of a sum of weakly dependent random variables.

Assumption 3 implies uniform continuity of the covariances of $\{X_{t,T}\}$ (Lemma 8.1). We conjecture that the continuity in Assumption 3 can, for example, be relaxed to piecewise continuity.

In the following we derive a rate for the risk of the proposed estimator uniformly over certain smoothness classes. It is well known that nonlinearly thresholded wavelet estimators have the potential to adapt to spatial inhomogeneity. Accordingly, we consider Besov classes as functional classes which admit functions with this feature. Furthermore, Besov spaces represent the most convenient scale of functional spaces in the context of wavelet methods, since the corresponding norm is equivalent to a certain norm in the sequence space of coefficients of a sufficiently regular wavelet basis. For an introduction to the theory of Besov spaces $B^m_{p,q}$ see, for example, Triebel (1990). Here $m \geqslant 1$ denotes the degree of smoothness and p, q ($1 \leqslant p, q \leqslant \infty$) specify the norm in which smoothness is measured. These classes contain traditional Hölder and L_2-Sobolev smoothness classes by setting $p = q = \infty$ and $p = q = 2$, respectively. Moreover, they embed other interesting functional spaces such as Sobolev spaces W^m_p, for which the inclusions $B^m_{p,p} \subseteq W^m_p \subseteq B^m_{p,2}$ (in the case $1 < p \leqslant 2$) and $B^m_{p,2} \subseteq W^m_p \subseteq B^m_{p,p}$ (if $2 \leqslant p < \infty$) hold true; see, for example, Theorem 6.4.4 in Bergh and Löfström (1976).

For convenience, we define our functional class by constraints on the sequences of wavelet coefficients. Fix any positive constants C_{ij}, $i = 1, \ldots, p$; $j = 1, 2$. We will assume that a_i lies in the set of functions

$$\mathscr{F}_i = \left\{ f = \sum_k \alpha_{lk}\phi_{lk} + \sum_{j,k} \beta_{jk}\psi_{jk} \,\middle|\, \|\alpha_{l.}\|_\infty \leqslant C_{i1},\ \|\beta_{..}\|_{m_i,p_i,q_i} \leqslant C_{i2} \right\},$$

where

$$\|\beta_{..}\|_{m,p,q} = \left(\sum_{j \geqslant l} \left[2^{jsp} \sum_{k \in I_j} |\beta_{jk}|^p \right]^{q/p} \right)^{1/q},$$

$s = m + 1/2 - 1/p$. It is well known that the class $\mathscr{F}_i$ lies between functional classes $B^{m_i}_{p_i,q_i}(c)$ and $B^{m_i}_{p_i,q_i}(C)$, for appropriate constants c and C; see Theorem 1 in Donoho and Johnstone (1998) for the Meyer bases, and Theorem 4.2 of Cohen, Dahmen and DeVore (1995) for the CDV bases.

To ensure sufficient regularity, we restrict ourselves to the following:

Assumption 4. $\tilde{s}_i > 1$, where $\tilde{s}_i = m_i + 1/2 - 1/\tilde{p}_i$, with $\tilde{p}_i = \min\{p_i, 2\}$.

In the case of normally distributed coefficients $\tilde{\beta}^{(i)}_{jk} \sim N(\beta^{(i)}_{jk}, \sigma^2)$, a very popular method is to apply thresholds $\lambda = \sigma\sqrt{2 \log n}$, where n is the number of these coefficients. As shown in Donoho *et al.* (1995), the application of these thresholds leads to an estimator which is simultaneously near-optimal in a wide variety of smoothness classes. Because of the heteroscedasticity of the empirical coefficients in our case, we have to modify the above rule slightly. Let $\mathscr{J}_T = \{(j, k) | l \leqslant j,\ 2^j \leqslant T^{1/2},\ k \in I_j\}$ and let σ^2_{ijk} be the variance of the empirical coefficient $\tilde{\beta}^{(i)}_{jk}$. Then any threshold λ_{ijk} satisfying

$$\sigma_{ijk}\sqrt{2\log(\#\mathscr{J}_T)} \leq \lambda_{ijk} = O(T^{-1/2}\sqrt{\log(\mathrm{T})}) \tag{2.3}$$

would be appropriate. Particular such choices are the 'individual thresholds'

$$\lambda_{ijk} = \sigma_{ijk}\sqrt{2\log(\#\mathscr{J}_T)}$$

and the 'universal threshold'

$$\lambda_T^{(i)} = \sigma_T^{(i)}\sqrt{2\log(\#\mathscr{J}_T)}, \qquad \sigma_T^{(i)} = \max_{(j,k)\in\mathscr{J}_T}\{\sigma_{ijk}\}.$$

Let $\hat{\lambda}_{ijk}$ be estimators of λ_{ijk} or $\lambda_T^{(i)}$, respectively, which satisfy at least the following minimal condition:

Assumption 5.

(i) $\sum_{(j,k)\in\mathscr{J}_T} P(\hat{\lambda}_{ijk} < \gamma_T\lambda_{ijk}) = O(T^{\eta})$, where $\eta < 1/(2m_i+1)$ for some $\gamma_T \to 1$.
(ii) $\sum_{(j,k)\in\mathscr{J}_T} P(\hat{\lambda}_{ijk} > CT^{-1/2}\sqrt{\log(T)}) = O(T^{-1})$.

With such thresholds $\hat{\lambda}_{ijk}$ we build the estimator

$$\hat{a}_i(u) = \sum_{k\in I_l^0}\tilde{\alpha}_{lk}^{(i)}\phi_{lk}(u) + \sum_{(j,k)\in\mathscr{J}_T}\delta^{(\cdot)}(\tilde{\beta}_{jk}^{(i)}, \hat{\lambda}_{ijk})\psi_{jk}(u), \tag{2.4}$$

where $\delta^{(\cdot)}$ stands for $\delta^{(\mathrm{h})}$ or $\delta^{(\mathrm{s})}$, as appropriate.

Finally, we wish to impose an additional condition on the matrix D defined by (7.4) in Section 7.1. Basically, this matrix is the analogue of the $p\times(T-p)$ matrix $((X_{t-m}))_{t=p+1,\ldots,T;m=1,\ldots,p}$, as arising in the classical Yule–Walker equations, which describe the corresponding least-squares problem for a stationary $AR(p)$ process $\{X_t\}$. Here, we assume additionally the following:

Assumption 6. $\mathrm{E}\|(D'D)^{-1}\|^{2+\delta} = O(T^{-2-\delta})$, for some $\delta > 0$.

Theorem 2.2. *(i) If Assumptions 1–5 hold, then*

$$\sup_{a_i\in\mathscr{F}_i}\{\mathrm{E}(\|\hat{a}_i - a_i\|^2_{L_2[0,1]} \wedge C)] = O((\log(T)/T)^{2m_i/(2m_i+1)}).$$

(ii) If, in addition, Assumption 6 is fulfilled, then

$$\sup_{a_i\in\mathscr{F}_i}\{\mathrm{E}\|\hat{a}_i - a_i\|^2_{L_2[0,1]}\} = O((\log(T)/T)^{2m_i/(2m_i+1)}).$$

Remark 2.3. Even without Assumption 6 we can show that $D'D$ is close to its expectation $\mathrm{E}D'D$, >and hence $\lambda_{\min}(D'D)$ is bounded away from zero, except for an event with a very small probability. To take this event into account, the somewhat unusual truncated loss function is introduced in part (i) of Theorem 2.2.

Remark 2.4. In our estimator (2.4) we restricted ourselves to a fixed primary resolution level l, that is, l does not change with growing sample size T. In principle, we could allow l to increase with T at a sufficiently slow rate. This has already been considered, for example by Hall and Patil (1995), in a different context. We expect the same rate for the risk of our estimator (2.4) as long as $2^{l(T)} \leqslant T^{1/(2m+1)}$, which can be shown similarly to methods in Hall and Patil (1995).

It is known that the rate $T^{-2m/(2m+1)}$ is minimax for estimating a function with degree of smoothness m in a variety of settings (regression, density estimation, spectral density estimation). Although we do not have a rigorous proof for its optimality in the present context, we conjecture that we cannot do better in estimating the a_is.

Analogously to Donoho and Johnstone (1998), we can obtain exactly the rate $T^{-2m_i/(2m_i+1)}$ by the use of level-dependent thresholds $\lambda^{(i)}(j, T, \mathscr{F}_i)$. These thresholds, however, would depend on the assumed degree of smoothness m_i, and it seems to be difficult to determine them in a fully data-driven way. In a simple model with Gaussian white noise, Donoho and Johnstone (1995) showed that full adaptivity can be reached by minimization of an empirical version of the risk, using Stein's unbiased estimator of risk. Because of our really strong version of asymptotic normality, we are convinced that we could attain this optimal rate of convergence in the same way.

Let us, however, note that the '*log*-thresholds' are much easier to apply, with only the small loss of a logarithmic factor in the rate. The surprising fact that a single estimator is optimal within some logarithmic factor in a large scale of smoothness classes can be explained by methodology quite different from conventional smoothing techniques: rather than aiming at an asymptotic balance relation between squared bias and variance of the estimator, which usually leads to the optimal rate of convergence, we perform something like an informal significance test on the coefficients. This leads to a slightly oversmoothed, but nevertheless near-optimal estimator.

3. Statistical properties of the empirical coefficients

Before we prove the main theorem in Section 5, we give an exact definition of the empirical coefficients and state some statistical properties of them.

First, note that our estimator, as a truncated orthogonal series estimator with nonlinearly modified empirical coefficients, involves two smoothing methodologies: one part of the smoothing is due to the truncation above some level j^*. Whereas such a truncation amounts to some linear, spatially non-adaptive technique, the more important smoothing is due to the pretest-like thresholding step applied to the coefficients below the level j^*. This step aims to select those coefficients which are in absolute value significantly above the noise level and eliminating the others.

From the definition of the Besov norm we obtain that (cf. Theorem 8 in Donoho *et al.* 1995)

$$\sup_{a_i \in \mathscr{F}_i} \left\{ \sum_{j \geqslant j^*} \sum_k |\beta_{jk}^{(i)}|^2 \right\} = O(2^{-2j^* \tilde{s}_i}), \tag{3.1}$$

where $\tilde{s}_i = m_i + 1/2 - 1/\min\{p_i, 2\}$. Hence, our loss due to the truncation is of order $T^{-2m_i/(2m_i+1)}$, if j^* is chosen such that $2^{-2j^*\tilde{s}_i} = O(T^{-2m_i/(2m_i+1)})$. According to our assumption that $\tilde{s}_i > 1$, it can be shown by simple algebra that j^* with $2^{j^*-1} \leqslant T^{1/2} < 2^{j^*}$ is large enough.

A first observation about the statistical behaviour of the empirical coefficients is stated by the following assertion.

Proposition 3.1. *If Assumptions 1–4 and 6 hold, then*

$$(i)\ \mathrm{E}(\tilde{\alpha}_{lk}^{(i)} - \alpha_{lk}^{(i)})^2 = O(T^{-1}),$$

$$(ii)\ \mathrm{E}(\tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)})^2 = O(T^{-1})$$

hold uniformly in i, k and $j < j^*$.

In view of the nonlinear structure of the estimator, the above assertion will not be strong enough to derive an efficient estimate for the rate of the risk of the estimator. If the empirical coefficients were Gaussian, then the number of $O(2^{j^*})$ coefficients would be dramatically reduced by thresholding with thresholds that are larger by a factor of $\sqrt{2\log(\#\mathscr{J}_T)}$ than the noise level. If we want to tune this thresholding method in accordance to our particular case with non-Gaussian coefficients, we have to investigate their tail behaviour. Hence, we state asymptotic normality of the coefficients with a special emphasis on moderate and large deviations. To prove the following theorem we decompose the empirical coefficients into a certain quadratic form and some remainder terms of smaller order of magnitude. Then we derive upper estimates for the cumulants of these quadratic forms, which provide asymptotic normality in terms of large deviations due to a lemma by Rudzkis *et al.* (1978); see Lemma 6.2 below.

It turns out that we can state asymptotic normality for empirical coefficients $\tilde{\beta}_{jk}^{(i)}$ with (j, k) from the following set of indices. Let, for arbitrarily small δ, $0 < \delta < 1/2$,

$$\tilde{\mathscr{J}}_T = \{(j, k) | 2^j \geqslant T^{\delta}, j < j^*, k \in I_j\}.$$

Proposition 3.2. *If Assumptions 1–4 hold, then*

$$P((\tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)})/\sigma_{ijk} \geqslant x) = (1 - \Phi(x)) + o(\min\{1 - \Phi(x), \Phi(x)\}) + O(T^{-\lambda})$$

uniformly in $(j, k) \in \tilde{\mathscr{J}}_T$, $x \in \mathbb{R}$ *for arbitrary* $\lambda < \infty$.

We now derive the asymptotic variances of the $\tilde{\beta}_{jk}^{(i)}$s. For notational simplicity, again, restricting ourselves without loss of generality to the treatment of CDV bases, we identify $\psi_1, \ldots, \psi_\Delta$ $(\Delta = 2^{j^*})$ with $\phi_{l1}, \ldots, \phi_{l,2^l}, \psi_{l1}, \ldots, \psi_{l,2^l}, \ldots, \psi_{j^*-1,1} \ldots, \psi_{j^*-1,2^{j^*}-1}$ and $\tilde{\theta}_1^{(i)}, \ldots, \tilde{\theta}_\Delta^{(i)}$ with $\tilde{\alpha}_{l1}^{(i)}, \ldots, \tilde{\alpha}_{l,2^l}^{(i)}, \tilde{\beta}_{l1}^{(i)}, \ldots, \tilde{\beta}_{l,2^l}^{(i)}, \ldots, \tilde{\beta}_{j^*-1,1}^{(i)}, \ldots, \tilde{\beta}_{j^*-1,2^{j^*}-1}^{(i)}$, respectively.

Furthermore, let

$$c(s,\ k) := \int_{-\pi}^{\pi} \frac{\sigma^2(s)}{2\pi} \left| 1 + \sum_{j=1}^{p} a_j(s) \exp(\mathrm{i}\lambda j) \right|^{-2} \exp(\mathrm{i}\lambda k)\, d\lambda. \tag{3.2}$$

Here $c(s,\ k)$ is the local covariance of lag k at time $s \in [0,\ 1]$ (cf. Lemma 8.1).

Proposition 3.3. *If Assumptions 1–4 and 6 hold, then*

$$\operatorname{var}(\tilde{\theta}_u^{(i)}) = T^{-1}(A^{-1}BA^{-1})_{p(u-1)+i,\,p(u-1)+i} + o(T^{-1}), \tag{3.3}$$

where

$$A_{p(u-1)+k,\,p(v-1)+l} = \int \psi_u(s)\psi_v(s)c(s,\ k-l)\,\mathrm{d}s,$$

$$B_{p(u-1)+k,\,p(v-1)+l} = \int \psi_u(s)\psi_v(s)\sigma^2(s)c(s,\ k-l)\,\mathrm{d}s.$$

Furthermore, $A^{-1}BA^{-1} \geqslant E^{-1}$, *where*

$$E_{p(u-1)+k,\,p(v-1)+l} = \int \psi_u(s)\psi_v(s)(\sigma^2(s))^{-1}c(s,\ k-l)\,\mathrm{d}s.$$

The eigenvalues of E are uniformly bounded.

Remark 3.4. The above form of A and B suggests different estimates for the variances of $\tilde{\theta}_u^{(i)}$ and therefore also for the thresholds. One possibility is to use (3.3) and plug in a preliminary estimate ($\sigma^2(s)$ may be estimated by a local sum of squared residuals). Another possibility is to use a nonparametric estimate of the local covariances $c(s,\ k)$. However, these suggestions require further investigation.

4. Some numerical examples

Before proving our main theorem, we wish to apply the procedure to two simulated autoregressive processes of order $p = 2$, both of length $T = 1024 = 2^{10}$:

$$X_{t,T} + a_1(t/T)X_{t-1,T} + a_2(t/T)X_{t-2,T} = \varepsilon_t, \qquad t = 1,\ \ldots,\ T,$$

where the ε_t are i.i.d. standard normal, $\mathrm{E}\varepsilon_t = 0$ and $\operatorname{var}(\varepsilon_t) = 1$. In both examples, the autoregressive parameters $a_i = a_i(t/T)$, $i = 1,\ 2$, are functions which change over time, that is to say, our simulated examples are realizations of a non-stationary process which follows the model (1.1).

Example 1. Here $a_1(u) = -1.69$ for $u \leqslant 0.6$, $a_1(u) = -1.38$ for $u > 0.6$, whereas $a_2(u) = 0.81$ for all $0 \leqslant u \leqslant 1$; that is, the first coefficient is a piecewise constant function with a jump at $u = 0.6$ and the second coefficient is constant over time. This gives a time-varying spectral density of the process $\{X_{t,T}\}$ which has a peak at $\pi/9$ for $t \leqslant 0.6T$ and at $4\pi/9$ for $t > 0.6T$ (see Figure 1, bottom right-hand plot).

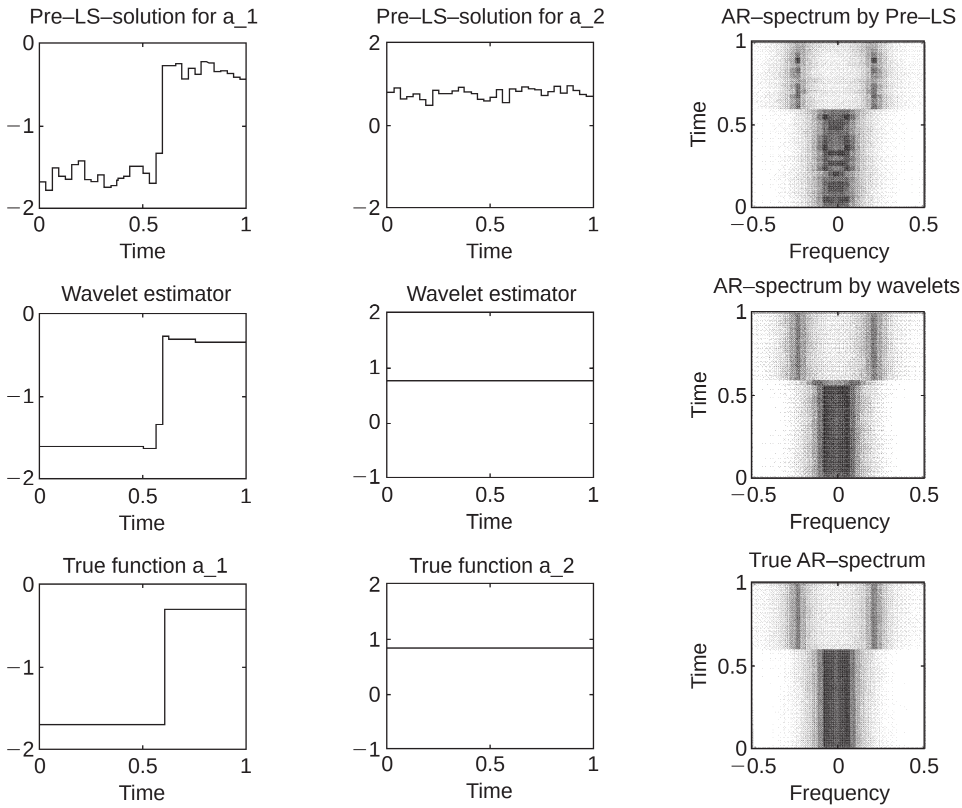

Figure 1. Example 1: preliminary LS solution, wavelet threshold estimator and true function for a_1, a_2 and resulting AR(2) spectrum

We have applied our estimation procedure using Haar wavelets and fixing the scale of our least-squares (LS) procedure to be $j^* = 5$, that is, $\Delta = 32$. Then we feed the resulting solution $\tilde{\alpha}^{(i)}$, for each $i = 1, 2$, a vector of length Δ (cf. also equation (7.2)) into our fast wavelet transform, apply hard thresholding on all resulting wavelet coefficients $\tilde{\beta}^{(i)}_{jk}$ on scales $j = 0, \ldots, 4$, and apply fast inverse wavelet transform up to scale 10, our original sample scale. Hereby, we use a universal data-driven $\sqrt{2 \log \Delta}$ threshold based on an empirical variance estimator of the finest wavelet scale $j^* - 1 = 4$.

In Figure 1 we show, for a_1 (left column) and a_2 (middle column), in the upper row the solution $\tilde{\alpha}^{(i)}$ of the LS procedure (without performing nonlinear wavelet thresholding). In the middle row the nonlinear wavelet estimators are shown, and in the bottom row the corresponding true function, all on an equispaced grid of resolution $T^{-1} = 2^{-10}$ of the interval [0, 1]. In the right-hand column, by grey-scale images in the time-frequency plane, we plot the resulting time-varying semi-parametric spectral density, based on the respective (estimated and true) autoregressive coefficient functions. Note that the darker the scale the higher the value of the $2-d$ object as a function of time and frequency.

Note that although the number of samples used for denoising by nonlinear wavelet thresholding is comparatively small ($\Delta = 32$ only), this second step delivers an additional significant contribution, which differs in its smoothness considerably from the LS solution alone.

We did not try different threshold rules, which possibly could improve a bit on the denoising. We found that the simple automatic universal rule is quite satisfactory, as it is also in accordance with the theoretically possibly range of thresholds as given by (2.3). Of course, in one or the other realization we observed that randomly one of the coefficients contributing only by noise was not set to zero, which, not surprisingly, had some disturbing effect on the visual appearance of the estimator, in particular of the constant autoregressive coefficient. Also, both in this and the next example we did not observe any significant difference between using hard or soft thresholding.

Example 2. This is a slight modification of both Example 1 and the example to be found in Dahlhaus (1997). The second autogressive coefficient is again constant over time; however, the first shows a smooth time variation of different phase and oscillation between the imposed jumps at $u = 0.25$ and $u = 0.75$. This was achieved by choosing $a_1(u) = -1.8\cos(1.5 - \cos(4\pi u + \pi))$ for $u \leqslant 0.25$ and for $u > 0.75$, and $a_1(u) = -1.8\cos(3 - \cos(4\pi u + \pi/2))$ for $0.25 < u \leqslant 0.75$, whereas again $a_2(u) = 0.81$ for all $0 \leqslant u \leqslant 1$.

A simulation of this process with $T = 1024$ is shown in Figure 2. It is the same realization that was used for the estimation procedure. Clearly one can observe the non-stationary behaviour of this process.

Here, we chose as wavelet basis a (periodic) Daubechies with $N = 4$ vanishing moments (filter length $2N = 8$), and we chose $\Delta = 64$ ($j^* = 6$). Note that for this specific example we replaced wavelets on the interval by a traditional periodic basis simply for reasons of computational convenience, as our chosen example is periodic with respect to time. However, we do not expect a big difference in performance between these two bases. In Figure 3 we have again plotted the LS solutions, the estimators based on wavelet hard thresholding with the same universal threshold rule as before, and the true functions, both for a_1, a_2 and for the resulting time-varying autoregressive spectrum.

5. Proof of the main theorem

To simplify the treatment of some particular remainder terms which occasionally arise in the following proofs, as for example in the decomposition (7.5), we introduce the following notation.

Definition 5.1. *We write*

$$Z_T = \tilde{O}(\eta_T),$$

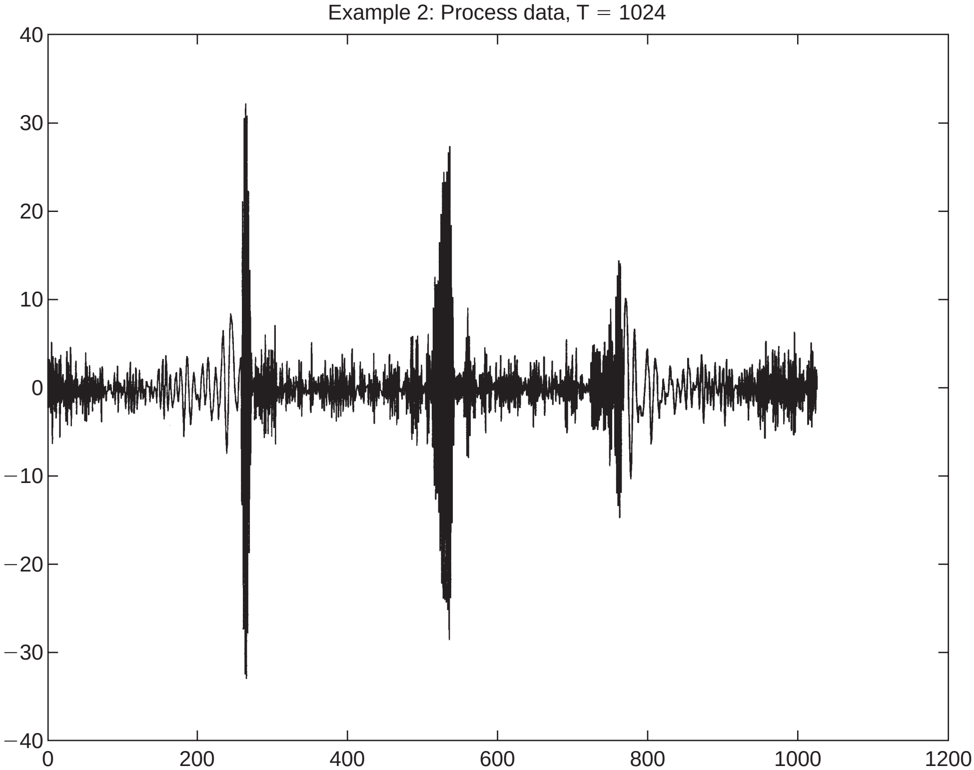

Figure 2. Example 2: realization of a stretch of length $T = 1024$

if for each $\lambda < \infty$ *there exists a* $C = C(\lambda)$ *such that*

$$P(|Z_T| > C\eta_T) \leqslant CT^{-\lambda}.$$

(If we use this notation simultaneously for an increasing number of random variables, we mean the existence of a *universal* constant only depending on λ.)

Proof of Theorem 2.2. We prove only (ii). The proof of (i) without the additional assumption (A6) is very similar, because the stochastic properties of the $\tilde{\beta}_{jk}^{(i)}$s are then nearly the same. The only difference is that we cannot guarantee the finiteness of moments of the $\tilde{\beta}_{jk}^{(i)}$s, and therefore we need the truncation in the loss function.

Using the monotonicity of $\delta^{(\cdot)}(\tilde{\beta}_{jk}^{(i)}, .)$ in the second argument we obtain

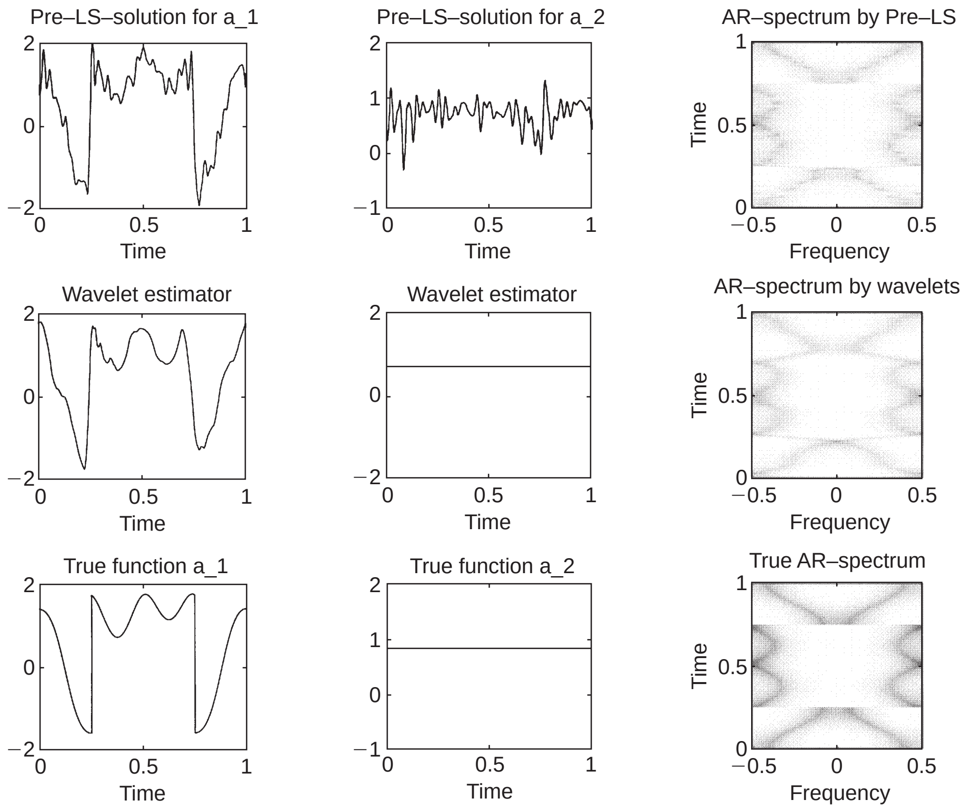

Figure 3. Example 2: preliminary LS solution, wavelet threshold estimator and true function for a_1, a_2 and resulting AR(2) spectrum

$$
(\delta^{(\cdot)}(\tilde{\beta}_{jk}^{(i)}, \hat{\lambda}_{ijk}) - \beta_{jk}^{(i)})^2 \leqslant \begin{cases} (\tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)})^2 + (\delta^{(\cdot)}(\tilde{\beta}_{jk}^{(i)}, \gamma_T \lambda_{ijk}) - \beta_{jk}^{(i)})^2, \\ \qquad\qquad \text{if } \hat{\lambda}_{ijk} > \gamma_T \lambda_{ijk}, \\ (\delta^{(\cdot)}(\tilde{\beta}_{jk}^{(i)}, \gamma_T \lambda_{ijk}) - \beta_{jk}^{(i)})^2 + (\delta^{(\cdot)}(\tilde{\beta}_{jk}^{(i)}, CT^{-1/2}\sqrt{\log(T)}) - \beta_{jk}^{(i)})^2, \\ \qquad\qquad \text{if } \gamma_T \lambda_{ijk} \leqslant \hat{\lambda}_{ijk} \leqslant CT^{-1/4}\sqrt{\log(T)}, \\ (\delta^{(\cdot)}(\tilde{\beta}_{jk}^{(i)}, CT^{-1/2}\sqrt{\log(T)}) - \beta_{jk}^{(i)})^2 + (\beta_{jk}^{(i)})^2, \\ \qquad\qquad \text{if } \hat{\lambda}_{ijk} > CT^{-1/2}\sqrt{\log(T)}, \end{cases}
$$

which implies the decomposition

$$
\begin{aligned}
\mathrm{E}\|\hat{a}_i - a_i\|^2 &\leqslant \sum_k \mathrm{E}(\tilde{\alpha}_{lk}^{(i)} - \alpha_{lk}^{(i)})^2 + \sum_{(j,k)\in\mathscr{J}_T} \mathrm{E}(\delta^{(\cdot)}(\beta_{jk}^{(i)}, \hat{\lambda}_{ijk}) - \beta_{jk}^{(i)})^2 + \sum_{j\geqslant j^*}\sum_{k\in I_j}(\beta_{jk}^{(i)})^2 \\
&\leqslant \sum_k \mathrm{E}(\tilde{\alpha}_{lk}^{(i)} - \alpha_{lk}^{(i)})^2 + \sum_{(j,k)\in\mathscr{J}_T} \mathrm{E}(\delta^{(\cdot)}(\tilde{\beta}_{jk}^{(i)}, \gamma_T\lambda_{ijk}) - \beta_{jk}^{(i)})^2 \\
&\quad + \sum_{(j,k)\in\mathscr{J}_T} \mathrm{E}(\delta^{(\cdot)}(\tilde{\beta}_{jk}^{(i)}, CT^{-1/2}\sqrt{\log T}) - \beta_{jk}^{(i)})^2 \\
&\quad + \sum_{(j,k)\pi\mathscr{J}_T} \mathrm{E}I(\hat{\lambda}_{ijk} < \gamma_T\lambda_{ijk})(\tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)})^2 \\
&\quad + \sum_{(j,k)\in\mathscr{J}_T} (\beta_{jk}^{(i)})^2 P(\hat{\lambda}_{ijk} > CT^{-1/2}\sqrt{\log T}) + \sum_{j\geqslant j^*}\sum_{k\in I_j}(\beta_{jk}^{(i)})^2 \\
&= S_1 + \ldots + S_6. \qquad (5.1)
\end{aligned}
$$

By (i) of Proposition 3.1 we immediately obtain

$$S_1 = O(T^{-1}). \qquad (5.2)$$

Let $(j, k) \in \tilde{\mathscr{J}}_T$. We choose a constant γ_{ijk} such that

$$
\begin{aligned}
\delta^{(\cdot)}(\beta, \gamma_T\lambda_{ijk}) &\geqslant \beta_{jk}^{(i)}, \qquad \text{if } \beta - \beta_{jk}^{(i)} > \gamma_{ijk}, \\
\delta^{(\cdot)}(\beta, \gamma_T\lambda_{ijk}) &\leqslant \beta_{jk}^{(i)}, \qquad \text{if } \beta - \beta_{jk}^{(i)} < \gamma_{ijk}.
\end{aligned}
$$

Without loss of generality, we assume $\delta^{(\cdot)}(\gamma_{ijk} + \beta_{jk}^{(i)}, \gamma_T\lambda_{ijk}) \geqslant \beta_{jk}^{(i)}$.

Let $\eta_T = CT^{-1/2}\sqrt{\log T}$ for some appropriate C. Then we decompose the terms occurring in the sum S_2 as follows:

$$
\begin{aligned}
S_{21}^{jk} &= \mathrm{E}I(\gamma_{ijk} \leqslant \tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)} < \eta_T)(\delta^{(\cdot)}(\tilde{\beta}_{jk}^{(i)}, \gamma_T\lambda_{ijk}) - \beta_{jk}^{(i)})^2, \\
S_{22}^{jk} &= \mathrm{E}I(-\eta_T < \tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)} < \gamma_{ijk})(\delta^{(\cdot)}(\tilde{\beta}_{jk}^{(i)}, \gamma_T\lambda_{ijk}) - \beta_{jk}^{(i)})^2
\end{aligned}
$$

and

$$S_{23}^{jk} = \mathrm{E}I(|\tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)}| \geqslant \eta_T)(\delta^{(\cdot)}(\tilde{\beta}_{jk}^{(i)}, \gamma_T\lambda_{ijk}) - \beta_{jk}^{(i)})^2.$$

Using Proposition 3.2 we obtain, with $\xi_{jk}^{(i)} \sim N(\beta_{jk}^{(i)}, \sigma_{ijk}^2)$, due to integration by parts with respect to x,

$$S_{21}^{jk} = -\int [I(\gamma_{ijk} \leqslant x < \eta_T)(\delta^{(.)}(\beta_{jk}^{(i)} + x, \gamma_T \lambda_{ijk}) - \beta_{jk}^{(i)})^2] \, \mathrm{d}\Big\{ P(\tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)} \geqslant x) \Big\}$$

$$= \int \Big\{ P(\tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)} \geqslant x) \Big\} \, \mathrm{d}[I(\gamma_{ijk} \leqslant x < \eta_T)(\delta^{(.)}(\beta_{jk}^{(i)} + x, \gamma_T \lambda_{ijk}) - \beta_{jk}^{(i)})^2]$$

$$+ P(\tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)} \geqslant \gamma_{ijk})(\delta^{(.)}(\beta_{jk}^{(i)} + \gamma_{ijk}, \gamma_T \lambda_{ijk}) - \beta_{jk}^{(i)})^2$$

$$\leqslant C_T \Big\{ \int \Big\{ P(\xi_{jk}^{(i)} - \beta_{jk}^{(i)} \geqslant x) \Big\} \mathrm{d}[I(\gamma_{ijk} \leqslant x < \eta_T)(\delta^{(.)}(\beta_{jk}^{(i)} + x, \gamma_T \lambda_{ijk}) - \beta_{jk}^{(i)})^2]$$

$$+ P(\xi_{jk}^{(i)} - \beta_{jk}^{(i)} \geqslant \gamma_{ijk})(\delta^{(.)}(\beta_{jk}^{(i)} + \gamma_{ijk}, \gamma_T \lambda_{ijk}) - \beta_{jk}^{(i)})^2 \Big\}$$

$$+ O(T^{-\lambda})$$

$$= C_T \mathrm{E} I(\gamma_{ijk} \leqslant \xi_{jk}^{(i)} - \beta_{jk}^{(i)} < \eta_T)(\delta^{(.)}(\xi_{jk}^{(i)}, \gamma_T \lambda_{ijk}) - \beta_{jk}^{(i)})^2 + O(T^{-\lambda})$$

for some $C_T \to 1$. Analogously, we obtain

$$S_{22}^{jk} \leqslant C_T \mathrm{E} I(-\eta_T \leqslant \xi_{jk}^{(i)} - \beta_{jk}^{(i)} < \gamma_{ijk})(\delta^{(.)}(\xi_{jk}^{(i)}, \gamma_T \lambda_{ijk}) - \beta_{jk}^{(i)})^2 + O(T^{-}\lambda).$$

Finally, we have, for any δ_1 with $0 < \delta_1 < \delta$ and δ as in Assumption 6, that

$$S_{23}^{jk} \leqslant (P(|\tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)}| \geqslant \eta_T))^{1-2/(2+\delta_1)} (\mathrm{E}|\delta^{(.)}(\tilde{\beta}_{jk}^{(i)}, \gamma_T \lambda_{ijk}) - \beta_{jk}^{(i)}|^{2+\delta_1})^{2/(2+\delta_1)} = O(T^{-\lambda}),$$

which implies

$$\mathrm{E}(\delta^{(.)}(\tilde{\beta}_{jk}^{(i)}, \gamma_T \lambda_{ijk}) - \beta_{jk}^{(i)})^2 \leqslant C_T \mathrm{E}(\delta^{(.)}(\xi_{jk}^{(i)}, \gamma_T \lambda_{ijk}) - \beta_{jk}^{(i)})^2 + O(T^{-\lambda}). \tag{5.3}$$

From Lemma 1 in Donoho and Johnstone (1994) we can immediately derive the formula

$$\mathrm{E}(\delta^{(.)}(\xi_{jk}^{(i)}, \lambda) - \beta_{jk}^{(i)})^2 \leqslant C \left(\sigma_{ijk}^2 \varphi \left(\frac{\lambda}{\sigma_{ijk}} \right) \left(\frac{\lambda}{\sigma_{ijk}} + 1 \right) + \min\{(\beta_{jk}^{(i)})^2, \lambda^2\} \right), \tag{5.4}$$

where φ denotes the standard normal density. This implies, by Theorem 7 in Donoho *et al.* (1995), that

$$\sum_{(j,k) \in \tilde{\mathscr{J}}_T} \mathrm{E}(\delta^{(.)}(\xi_{jk}^{(i)}, \gamma_T \lambda_{ijk}) - \beta_{jk}^{(i)})^2$$

$$= O\left(T^{-1} (\# \tilde{\mathscr{J}}_T)^{1-\gamma_T^2} \sqrt{\log(T)} + \sum_{(j,k) \in \tilde{\mathscr{J}}_T} \min\{(\beta_{jk}^{(i)})^2, (\gamma_T \lambda_{ijk})^2\} \right)$$

$$= O\big((\log(\mathrm{T})/T)^{2m_i/(2m_i+1)} \big).$$

Therefore, in conjunction with (5.3), we obtain that

$$\sum_{(j,k)\in\tilde{\mathscr{J}}_T} \mathrm{E}\Big(\delta^{(\cdot)}(\tilde{\beta}_{jk}^{(i)}, \gamma_T\lambda_{ijk}) - \beta_{jk}^{(i)}\Big)^2 = O\big((\log(T)/T)^{2m_i/(2m_i+1)}\big). \tag{5.5}$$

Further we obtain, because of $|\delta^{(\cdot)}(\beta, \lambda) - \beta| \leqslant \lambda$, that

$$\begin{aligned}\sum_{(j,k)\in\mathscr{J}_T\backslash\tilde{\mathscr{J}}_T} \mathrm{E}(\delta^{(\cdot)}(\tilde{\beta}_{jk}^{(i)}, \gamma_T\lambda_{ijk}) - \beta_{jk}^{(i)})^2 &\leqslant \sum_{(j,k)\in\mathscr{J}_T\backslash\tilde{\mathscr{J}}_T} [2\mathrm{E}(\tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)})^2 + 2(\gamma_T\lambda_{ijk})^2]\\ &= \#(\mathscr{J}_T\backslash\tilde{\mathscr{J}}_T)O(T^{-1}\log(T)).\end{aligned}$$

If we choose δ in the definition of $\tilde{\mathscr{J}}_T$ in such a way that $\delta < 1/(2m_i+1)$, we obtain, by $\#(\mathscr{J}_T\backslash\tilde{\mathscr{J}}_T) = O(T^\delta)$, that

$$\sum_{(j,k)\in\mathscr{J}_T\backslash\tilde{\mathscr{J}}_T} \mathrm{E}(\delta^{(\cdot)}(\tilde{\beta}_{jk}^{(i)}, \gamma_T\lambda_{ijk}) - \beta_{jk}^{(i)})^2 = O(T^{-2m_i/(2m_i+1)}). \tag{5.6}$$

By analogous considerations we can show that

$$S_3 = O((\log(T)/T)^{2m_i/(2m_i+1)}). \tag{5.7}$$

From (7.14) and (7.22) we have

$$\tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)} = \tilde{O}(T^{-1/2}\sqrt{\log(T)} + 2^{-j/2}T^{-1/2}\log(T)),$$

which implies by Assumption 5(i) and Lemma A.2 that

$$\begin{aligned}S_4 &= O(T^{-1}(\log(T))^2) \sum_{(j,k)\in\mathscr{J}_T} P(\hat{\lambda}_{ijk} < \gamma_T\lambda_{ijk})\\ &\quad + C\sum_{(j,k)\in\mathscr{J}_T} (P(|\tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)}| > CT^{-1/2}\log(T)))^{2/(2+\delta_1)}(\mathrm{E}|\tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)}|^{2+\delta_1})^{2/(2+\delta_1)}\\ &= O(T^{-2m_i/(2m_i+1)}).\end{aligned} \tag{5.8}$$

The relation

$$S_5 = O(T^{-2m_i/(2m_i+1)}) \tag{5.9}$$

is obvious, due to Assumption 5 (ii). Finally, it can be shown by simple algebra that

$$S_6 = O(2^{-2j^*\tilde{s}_i}) = O(T^{-2m_i/(2m_i+1)}), \tag{5.10}$$

which completes the proof. □

6. Asymptotic normality of quadratic forms

In this section we list the basic technical lemmas which are necessary to prove asymptotic normality or to find stochastic estimates for quadratic forms. First, we quote a lemma that provides upper estimates for the cumulants of quadratic forms that satisfy a certain condition on their cumulant sums. This result is a generalization of Lemma 2 in Rudzkis (1978), which was formulated specifically for quadratic forms that occur in periodogram-based kernel estimators of a spectral density. We obtain a slightly improved estimate, which turns out to be important, for example, for certain quadratic forms with sparse matrices.

We consider the quadratic form

$$\eta_T = \underline{X}_T' A \underline{X}_T,$$

where

$$\underline{X}_T = (X_1, \ldots, X_T)',$$

$$A = ((a_{ij}))_{i,j=1,\ldots,T}, \qquad a_{ij} = a_{ji}.$$

Further, let

$$\xi_T = \underline{Y}_T' A \underline{Y}_T,$$

where $\underline{Y}_T = (Y_1, \ldots, Y_T)'$ is a zero-mean Gaussian vector with the same covariance matrix as $\underline{X}_T$.

Lemma 6.1. *Assume* $\mathrm{E}X_t = 0$ *and, for some* $\gamma \geqslant 0$,

$$\sup_{1 \leqslant t_1 \leqslant T} \left\{ \sum_{t_2,\ldots,t_k=1}^{T} |\mathrm{cum}(X_{t_1}, \ldots, X_{t_k})| \right\} \leqslant C^k (k!)^{1+\gamma}, \qquad \text{for all } T \text{ and } k = 2, 3, \ldots .$$

Then, for $n \geqslant 2$,

$$\mathrm{cum}_n(\eta_T) = \mathrm{cum}_n(\xi_T) + R_n,$$

where

(i) $|\mathrm{cum}_n(\xi_T)| \leqslant \mathrm{var}(\xi_T) 2^{n-2} (n-1)! [\lambda_{\max}(A) \lambda_{\max}(\mathrm{cov}(\underline{X}_T))]^{n-2}$,
(ii) $R_n \leqslant 2^{n-2} C^{2n} ((2n)!)^{1+\gamma} \max_{s,t}\{|a_{st}|\} \tilde{A} \|A\|_\infty^{n-2}$,

$$\tilde{A} = \sum_s \max_t \{|a_{st}|\}, \qquad \|A\|_\infty = \max_s \left\{ \sum_k |a_{st}| \right\}.$$

The proof of this lemma is given in Neumann (1996).

Using Lemma 6.1 we obtain useful estimates for the cumulants, which can be used to derive asymptotic normality. For the reader's convenience we quote two basic lemmas on the asymptotic distribution of η_T. Lemma 6.2, which is due to Rudzkis *et al*. (1978), states asymptotic normality under a certain relation between variance and the higher-order cumulants of η_T. Even if such a favourable relation is not given, we can still obtain estimates for probabilities of large deviations on the basis of the Lemma 6.3, which is due to Bentkus and Rudzkis (1980).

Lemma 6.2. *Assume, for some* $\Delta_T \to 0$, *that*

$$|\mathrm{cum}_n(\eta_T/\sqrt{\mathrm{var}(\pi_T)})| \leqslant \frac{(n!)^{1+\gamma}}{\Delta_T^{n-2}}, \qquad \text{for } n = 3, 4, \ldots .$$

Then

$$\frac{P(\pm(\eta_T - \mathrm{E}\eta_T)/\sqrt{\mathrm{var}(\eta_T)} \geqslant x)}{1 - \Phi(x)} \to 1$$

holds uniformly over $0 \leqslant x \leqslant \nu_T$, *where* $\nu_T = o(\Delta_T^{1/(3+6\gamma)})$.

Lemma 6.3. *Assume, for some* $\overline{\Delta}_T \to 0$, *that*

$$|\mathrm{cum}_n(\eta_T)| \leqslant \left(\frac{n!}{2}\right)^{1+\gamma} \frac{H_T}{\overline{\Delta}_T^{n-2}}, \qquad \textit{for } n = 2, 3, \ldots .$$

Then, for $x \geqslant 0$,

$$\begin{aligned} P(\pm\eta_T \geqslant x) &\leqslant \exp\left(-\frac{x^2}{2[H_T + (x/\overline{\Delta}_T^{1/(1+2\gamma)})^{(1+2\gamma)/(1+\gamma)}]}\right) \\ &\leqslant \begin{cases} \exp(-x^2/4H_T), & \textit{if } 0 \leqslant x \leqslant (H_T^{1+\gamma}\overline{\Delta}_T)^{1/(1+2\gamma)}, \\ \exp(-\frac{1}{4}(x\overline{\Delta}_T)^{1/(1+\gamma)}), & \textit{if } x \geqslant (H_T^{1+\gamma}\overline{\Delta}_T)^{1/(1+2\gamma)}. \end{cases} \end{aligned}$$

7. Derivation of the asymptotic distribution of the empirical coefficients

7.1. Preparatory considerations

Before we turn directly to the proofs Propositions 3.1–3.3, we represent the empirical coefficients in a form that allows the nature of every remainder term to be easily recognized. Note that throughout the rest of the paper, for notational convenience we now omit the double index in the sequence $\{X_{t,T}\}$; that is, in the following let $X_t := X_{t,T}$.

Although it is essential for our procedure to have a *multiresolution* basis, that is, empirical coefficients from different resolution levels, it turns out to be easier to analyse the statistical behaviour of such coefficients coming from a single level. Since the empirical coefficients of the multiresolution basis can be obtained as linear combinations of coefficients of an appropriate monoresolution basis, we are able to derive their asymptotic distribution.

Since both $\{\phi_{l1}, \ldots, \phi_{l,2^l}, \psi_{l1}, \ldots, \psi_{l,2^l}, \ldots, \psi_{j^*-1,1}, \ldots, \psi_{j^*-1,2^{j^*-1}}\}$ and $\{\phi_{j^*1}, \ldots, \phi_{j^*,2^{j^*}}\}$ are orthonormal bases of the same space V_{j^*}, the minimization of (2.2) is equivalent to that of

$$\sum_{t=p+1}^{T}\left(X_t+\sum_{i=1}^{p}\left[\sum_{k\in I^0_{j^*}}\alpha^{(i)}_{j^*k}\phi_{j^*k}(t/T)\right]X_{t-i}\right)^2. \tag{7.1}$$

Assume for a moment that $D'D$ is positive definite, which is indeed true with a probability exceeding $1-O(T^{-\lambda})$. The solution $\tilde{\alpha}=(\tilde{\alpha}^{(1)}_{j^*1},\ldots,\tilde{\alpha}^{(p)}_{j^*1},\ldots,\tilde{\alpha}^{(p)}_{j^*\Delta},\ldots,\tilde{\alpha}^{(1)}_{j^*\Delta})'$, $\Delta=\#I^0_{j^*}=2^{j^*}$, can be written as the least-squares estimator

$$\tilde{\alpha}=(D'D)^{-1}D'Y \tag{7.2}$$

in the linear model

$$Y=D\alpha+\gamma, \tag{7.3}$$

where

$$Y=(X_{p+1},\ldots,X_T)',$$

$$D=-\begin{pmatrix}
\phi_{j^*1}\left(\frac{p+1}{T}\right)X_p & \cdots & \phi_{j^*1}\left(\frac{p+1}{T}\right)X_1 & \cdots & \phi_{j^*\Delta}\left(\frac{p+1}{T}\right)X_p & \cdots & \phi_{j^*\Delta}\left(\frac{p+1}{T}\right)X_1\\
\phi_{j^*1}\left(\frac{p+2}{T}\right)X_{p+1} & \cdots & \phi_{j^*1}\left(\frac{p+2}{T}\right)X_2 & \cdots & \phi_{j^*\Delta}\left(\frac{p+2}{T}\right)X_{p+1} & \cdots & \phi_{j^*\Delta}\left(\frac{p+2}{T}\right)X_2\\
\vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots\\
\phi_{j^*1}\left(\frac{T}{T}\right)X_{T-1} & \cdots & \phi_{j^*1}\left(\frac{T}{T}\right)X_{T-p} & \cdots & \phi_{j^*\Delta}\left(\frac{T}{T}\right)X_{T-1} & \cdots & \phi_{j^*\Delta}\left(\frac{T}{T}\right)X_{T-p}
\end{pmatrix}, \tag{7.4}$$

$$\alpha=(\alpha^{(1)}_{j^*1},\ldots,\alpha^{(p)}_{j^*1},\ldots,\alpha^{(1)}_{j^*\Delta},\ldots,\alpha^{(p)}_{j^*\Delta})'$$

and

$$\gamma=(\gamma_{p+1},\ldots,\gamma_T)'.$$

The residual term in (7.3) can, for $t=p+1,\ldots,T$, be written as

$$\begin{aligned}\gamma_t&=X_t-(D\alpha)_{t-p}\\
&=-\sum_{i=1}^{p}a_i(t/T)X_{t-i}+\varepsilon_t+\sum_{i=1}^{p}\sum_{k\in I^0_{j^*}}\alpha^{(i)}_{j^*k}\phi_{j^*k}(t/T)X_{t-i}=\sum_{i=1}^{p}R_i(t/T)X_{t-i}+\varepsilon_t,\end{aligned}$$

where

$$R_i(u)=-a_i(u)+\sum_{k\in I^0_{j^*}}\alpha^{(i)}_{j^*k}\phi_{j^*k}(u)=-\sum_{j\geqslant j^*}\sum_{k\in I_j}\beta^{(i)}_{jk}\psi_{jk}(u).$$

With the definitions

$$S=\left(\sum_{i=1}^{p}R_i\left(\frac{p+1}{T}\right)X_{p+1-i},\ldots,\sum_{i=1}^{p}R_i\left(\frac{T}{T}\right)X_{T-i}\right)'$$

and

$$e = (\varepsilon_{p+1}, \ldots, \varepsilon_T)',$$

we decompose the right-hand side of (7.2) as

$$\begin{aligned}\tilde{\alpha} &= (D'D)^{-1}D'D\alpha + (\mathrm{E}D'D)^{-1}D'e + [(D'D)^{-1} - (\mathrm{E}D'D)^{-1}]D'e + (D'D)^{-1}D'S \\ &= \alpha + T_1 + T_2 + T_3. \end{aligned} \tag{7.5}$$

Because of the above-mentioned relation between the two orthonormal bases of V_{j^*}, there exists an orthonormal $(\Delta \times \Delta)$ matrix Γ with

$$(\phi_{l1}, \ldots, \phi_{l,2^l}, \psi_{l1}, \ldots, \psi_{l,2^l}, \ldots, \psi_{j^*-1,1}, \ldots, \psi_{j^*-1,2^{j^*-1}})' = \Gamma(\phi_{j^*1}, \ldots, \phi_{j^*\Delta})'.$$

This implies

$$(\alpha_{j^*1}^{(i)}, \ldots, \alpha_{j^*\Delta}^{(i)})\begin{pmatrix}\phi_{j^*1}\\ \vdots \\ \phi_{j^*\Delta}\end{pmatrix} = (\alpha_{j^*1}^{(i)}, \ldots, \alpha_{j^*\Delta}^{(i)})\Gamma'\begin{pmatrix}\phi_{l1}\\ \vdots \\ \phi_{l,2^l}\\ \psi_{l1}\\ \vdots \\ \psi_{j^*-1,2^{j^*-1}}\end{pmatrix}.$$

Hence, having the least-squares estimator $(\tilde{\alpha}_{j^*1}^{(i)}, \ldots, \tilde{\alpha}_{j^*\Delta}^{(i)})$ according to the basis $\{\phi_{j^*1}, \ldots, \phi_{j^*\Delta}\}$, we obtain the least-squares estimator in model (2.2) as

$$(\tilde{\alpha}_{l1}^{(i)}, \ldots, \tilde{\alpha}_{l,2^l}^{(i)}, \tilde{\beta}_{l1}^{(i)}, \ldots, \tilde{\beta}_{l,2^l}^{(i)}, \ldots, \tilde{\beta}_{j^*-1,1}^{(i)}, \ldots, \tilde{\beta}_{j^*-1,2^{j^*-1}}^{(i)})' = \Gamma(\tilde{\alpha}_{j^*1}^{(i)}, \ldots, \tilde{\alpha}_{j^*\Delta}^{(i)})'.$$

In other words, every empirical coefficient $\tilde{\beta}_{jk}^{(i)}$ which is part of the solution to (2.2) can be written as

$$\tilde{\beta}_{jk}^{(i)} = \Gamma'_{ijk}\tilde{\alpha}, \tag{7.6}$$

where $\|\Gamma_{ijk}\|_{l_2} = 1$. (Analogously, $\tilde{\alpha}_{lk}^{(i)} = \Gamma'_{ik}\tilde{\alpha}$.)

7.2. Proofs of the Propositions 3.1, 3.2 and 3.3

Proof of Proposition 3.1. For notational convenience we write down the proof for empirical coefficients $\tilde{\beta}_{jk}^{(i)}$ only. The proof for the $\tilde{\alpha}_{lk}^{(i)}$s is analogous.

According to (7.5), we have

$$\tilde{\beta}_{jk}^{(i)} = \beta_{jk}^{(i)} + \Gamma'_{ijk}T_1 + \Gamma'_{ijk}T_2 + \Gamma'_{ijk}T_3. \tag{7.7}$$

From (i) and (iii) of Lemma A.3 we conclude

$$\begin{aligned}\mathrm{E}(\Gamma'_{ijk}T_1)^2 &= \Gamma'_{ijk}(\mathrm{E}D'D)^{-1}\operatorname{cov}(D'e)(\mathrm{E}D'D)^{-1}\Gamma_{ijk} \\ &\leqslant \|\Gamma_{ijk}\|_2^2\|(\mathrm{E}D'D)^{-1}\|_2^2\|\operatorname{cov}(D'e)\|_2 = O(T^{-1}).\end{aligned} \tag{7.8}$$

The vector Γ_{ijk} has a length of support of $O(2^{j^*-j})$, which implies

$$\sum_l |(\Gamma_{ijk})_l| \leqslant \|\Gamma_{ijk}\|_2 \sqrt{\#\{l|(\Gamma_{ijk})_l \neq 0\}} = O(2^{(j^*-j)/2}). \tag{7.9}$$

We have, by Taylor expansion of the matrix $(D'D)^{-1}$, $T_2 = T_{21} + T_{22}$, where

$$T_{21} = (\mathrm{E}D'D)^{-1}((\mathrm{E}D'D) - D'D)(\mathrm{E}D'D)^{-1}D'e$$

and

$$\|T_{22}\|_2 = \tilde{O}(\|(\mathrm{E}D'D)^{-1}\|_2^3 \|(\mathrm{E}D'D) - D'D\|_2^2 \|D'e\|_2).$$

Using (i) of Lemma A.3, (A.8) and (A.9) we obtain

$$\begin{aligned}\|T_{21}\|_\infty &\leqslant \|(\mathrm{E}D'D)^{-1}\|_\infty^2 \|(\mathrm{E}D'D) - D'D\|_\infty \|D'e\|_\infty \\ &= \tilde{O}(2^{j^*/2}T^{-1}\log(T)).\end{aligned} \tag{7.10}$$

Since we have enough moment assumptions, we obtain the analogous rate, but without the logarithmic factor, for the second moment of $\Gamma'_{ijk}T_{21}$, that is,

$$\mathrm{E}(\Gamma'_{ijk}T_{21})^2 = O(2^{j^*-j}2^{j^*}T^{-2}). \tag{7.11}$$

Further, we have

$$\Gamma'_{ijk}T_{22} = \tilde{O}(2^{3j^*/2}T^{-3/2}\log(T)). \tag{7.12}$$

Using (i) of Lemma A.3 and (i) of Lemma A.4, we obtain

$$\|(D'D)^{-1}\|_2 \leqslant \|(\mathrm{E}D'D)^{-1}\|_2 + \|(D'D)^{-1} - (\mathrm{E}D'D)^{-1}\|_2 = O(T^{-1}) + \tilde{O}(2^{j^*/2}T^{-3/2}\sqrt{\log(T)}),$$

which yields, in conjunction with Lemma A.5, that

$$\begin{aligned}\Gamma'_{ijk}T_3 &= O(\|(D'D)^{-1}\|_2 \|D'S\|_2) \\ &= \tilde{O}((2^{-j^*\min\{\tilde{s}_i\}} + T^{-1/2}2^{-j^*\min\{m_i-1/2-1/(2p_i)\}})\sqrt{\log(T)}) \\ &= \tilde{O}(T^{-1/2-\tau}),\end{aligned} \tag{7.13}$$

for some $\tau > 0$. Now we infer from (7.7), (7.8) and (7.11)–(7.13), which are in part $\tilde{O}$-results rather than estimates for the expectations, that

$$\mathrm{E}I(\Omega_0)((\tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)})^2) = O(T^{-1}),$$

where Ω_0 is an appropriate event with $P(\Omega_0) \geqslant 1 - O(T^{-\lambda})$ for $\lambda < \infty$ chosen arbitrarily large. This implies in conjunction with Lemma A.2, with $0 < \delta_1 < \delta$, that

$$\mathrm{E}I(\Omega_0^c)((\tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)})^2) \leqslant (\mathrm{E}|\tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)}|^{2+\delta_1})^{2/(2+\delta_1)}(P(\Omega_0^c))^{1-2/(2+\delta_1)} = O(T^{-1}),$$

which finishes the proof. □

Proof of Proposition 3.2. It will turn out that the asymptotic distribution of $\tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)}$ is essentially determined by the behaviour of $\Gamma'_{ijk}T_1$. By (7.9), (7.10), (7.12) and (7.13) from the proof of Proposition 3.1 we infer that

$$\Gamma'_{ijk}(T_2 + T_3) = \tilde{O}(2^{-j/2}T^{-1/2}\log(T) + T^{-1/2-\kappa}), \tag{7.14}$$

for some $\kappa > 0$.

First, note that the process $\{X_{t,T}\}$ admits an MA(∞) representation

$$X_{t,T} = \sum_{s=0}^{\infty} \gamma_{t,T}(s)\varepsilon_{t-s}, \tag{7.15}$$

with

$$\sum_{s=0}^{\infty} \sup_{t,T}\{|\gamma_{t,T}(s)|\} \leqslant C, \qquad \text{for all } T;$$

see Künsch (1995).

Now we turn to the derivation of the asymptotic distribution of $\Gamma'_{ijk}T_1$. It is clear that, because of the MA(∞) representation of the process, $\Gamma'_{ijk}T_1$ can be rewritten as $\sum_{u,v}A_{u,v}\varepsilon_u\varepsilon_v$ for some symmetric matrix $A = A(i, j, k)$. In the following, without writing down the explicit form of this matrix, we derive upper estimates for $\|A\|_\infty$ and $\tilde{A} = \sum_u \max_v\{|A_{u,v}|\}$.

We have

$$\begin{aligned}\Gamma'_{ijk}T_1 &= -\sum_{t=p+1}^{T} \varepsilon_t \sum_{l=1}^{p} X_{t-l} \sum_{u=1}^{\Delta} \phi_{j^*u}(t/T) \sum_v ((\mathrm{E}D'D)^{-1})_{p(u-1)+l,v}(\Gamma_{ijk})_v \\ &= \sum_{l,s}\left[\sum_t \varepsilon_t\varepsilon_{t-l-s}w_t(l, s)\right],\end{aligned} \tag{7.16}$$

where

$$w_t(l, s) = \gamma_{t-l}(s)\sum_{u=1}^{\Delta} \phi_{j^*u}(t/T)\sum_v ((\mathrm{E}D'D)^{-1})_{p(u-1)+l,v}(\Gamma_{ijk})_v.$$

If we write the expression in brackets on the right-hand side of (7.16) as $\sum_{ij}\tilde{W}_{ij}\varepsilon_i\varepsilon_j$, we obtain, by $\sup_v\{|(\Gamma_{ijk})_v|\} = O(2^{-(j^*-j)/2})$, that

$$\|\tilde{W}\|_\infty = O(T^{-1}\sup_t\{|\gamma_{t-l}(s)|\}2^{j/2}). \tag{7.17}$$

We can also rewrite $w_t(l, s)$ as

$$w_t(l, s) = -\gamma_{t-l}(s)\sum_v (\Gamma_{ijk})_v \sum_u ((\mathrm{E}D'D)^{-1})_{v,p(u-1)+l}\phi_{j^*u}(t/T),$$

which implies, by $\sum_v|(\Gamma_{ijk})_v| = O(2^{(j^*-j)/2})$ and by $\sum_t\phi_{j^*u}(t/T) = O(2^{-j^*/2}T)$, that

$$\sum_i \sup_j \{|\tilde{W}_{ij}|\} = \sum_t |w_t(l, s)| = O(2^{-j/2}). \tag{7.18}$$

Because of Assumption 3, the summation over s does not affect the rates in (7.17) and (7.18), and neither does the (finite) sum over l. Hence, with the notation of Lemma 6.1, we obtain

$$\|A\|_\infty = O(T^{-1}2^{j/2}), \tag{7.19}$$

$$\tilde{A} = O(2^{-j/2}). \tag{7.20}$$

Let $(j, k) \in \tilde{\mathscr{J}}_T$. Using Lemma 6.1, we obtain

$$|\mathrm{cum}_n(\Gamma'_{ijk}T_1)| \leqslant C^n T^{-1}(n!)^{2+2\gamma}(T^{-1}2^{j/2})^{n-2}, \tag{7.21}$$

which implies, by Lemma 6.2,

$$P(\pm(\Gamma'_{ijk}T_1)/\sigma_{ijk} \geqslant x) = (1 - \Phi(x))(1 + o(1)) \tag{7.22}$$

uniformly in $0 \leqslant x \leqslant \kappa_T$, $\kappa_T \asymp T^\nu$ for some $\nu > 0$. This relation can obviously be extended to $x \in (-\infty, \kappa_T]$.

Recall that

$$\tilde{\beta}^{(i)}_{jk} - \beta^{(i)}_{jk} = \Gamma'_{ijk}T_1 + \tilde{O}(T^{-1/2-\kappa}) \tag{7.23}$$

holds for some $\kappa > 0$. Therefore we have, for arbitrarily large $\lambda < \infty$, that

$$\begin{aligned} P(\pm(\Gamma'_{ijk}T_1)/\sigma_{ijk} - CT^{-\kappa} \geqslant x) - CT^{-\lambda} &\leqslant P(\pm(\tilde{\beta}^{(i)}_{jk} - \beta^{(i)}_{jk})/\sigma_{ijk} \geqslant x) \\ &\leqslant P(\pm(\Gamma'_{ijk}T_1)/\sigma_{ijk} + CT^{-\kappa} \geqslant x) + CT^{-\lambda}, \end{aligned}$$

which implies

$$\begin{aligned} P(\pm(\tilde{\beta}^{(i)}_{jk} - \beta^{(i)}_{jk})/\sigma_{ijk} \geqslant x) &= [1 - \Phi(x)](1 + o(1)) + O(|\Phi(x) - \Phi(x + CT^{-\kappa})|) \\ &\quad + O(|\Phi(x) - \Phi(x - CT^{-\kappa})|) + O(T^{-\lambda}). \end{aligned} \tag{7.24}$$

Fix any $c > 1$. For $x \leqslant c$ we obviously have

$$|\Phi(x) - \Phi(x + CT^{-\kappa})| \leqslant CT^{-\kappa}\phi(0) = o(1 - \Phi(x)). \tag{7.25}$$

For $c < x \leqslant (2\lambda\log(T))^{1/2}$ we obtain by a formula for Mill's ratio (see Johnson and Kotz 1970, Vol. 2, p. 278) that

$$\begin{aligned} |\Phi(x) - \Phi(x + CT^{-\kappa})| &\leqslant CT^{-\kappa}\phi(x) \\ &\leqslant CT^{-\kappa}x\left(1 - \frac{1}{x^2}\right)^{-1}(1 - \Phi(x)) \\ &\leqslant CT^{-\kappa}x\left(1 - \frac{1}{c^2}\right)^{-1}(1 - \Phi(x)) = o(1 - \Phi(x)). \end{aligned} \tag{7.26}$$

The third term on the right-hand side of (7.24) can be treated analogously.

For $x > C(2\lambda\log(T))^{1/2}$ we obviously have

$$P(\pm(\tilde{\beta}_{jk}^{(i)} - \beta_{jk}^{(i)})/\sigma_{ijk} \geqslant x) = O(T^{-\lambda}) = (1 - \Phi(x))(1 + o(1)) + O(T^{-\lambda}), \tag{7.27}$$

which completes the proof. □

Proof of Proposition 3.3. Because of $\mathrm{E}T_1 = 0$ we have

$$\operatorname{cov}(\mathrm{T}_1) = \mathrm{E}T_1 T_1' = (\mathrm{E}D'D)^{-1} \operatorname{cov}(D'e)(\mathrm{E}D'D)^{-1},$$

which implies by (ii) and (iii) of Lemma A.3 that

$$\|\operatorname{cov}(T_1) - F^{-1}GF^{-1}\|_\infty = o(T^{-1}),$$

where

$$F = \left(\left\{T\int \phi_{j^*u}(s)\phi_{j^*v}(s)c(s,\ k-l)\mathrm{d}s\right\}_{p(u-1)+k,\, p(v-1)+l}\right)$$

and

$$G = \left(\left\{T\int \phi_{j^*u}(s)\phi_{j^*v}(s)\sigma^2(s)c(s,\ k-l)\mathrm{d}s\right\}_{p(u-1)+k,\, p(v-1)+l}\right).$$

This yields

$$\|\operatorname{cov}(\Gamma T_1) - \Gamma F^{-1}\Gamma'\Gamma G\Gamma'\Gamma F^{-1}\Gamma'\|_\infty = \|\operatorname{cov}(\Gamma T_1) - A^{-1}BA^{-1}\|_\infty = o(T^{-1}).$$

Further, due to (6.13), we have

$$\mathrm{E}(\Gamma'_{ijk}(T_2 + T_3))^2 = o(T^{-1}),$$

which proves the first assertion (3.3).

The matrix $\begin{pmatrix} B & A \\ A & E \end{pmatrix}$ is non-negative definite, which leads, with Theorem 12.2.21(5) of Graybill (1983), to $A^{-1}BA^{-1} \geqslant E^{-1}$. Furthermore, we have, with $x \in \mathbb{C}^{\Delta p}$,

$$\begin{aligned} x^* Ex &= \int_0^1\int_{-\pi}^{\pi} |A(s,\lambda)|^2(\sigma^2(s))^{-1}\left|\sum_{u,k} x_{p(u-1)+k}\psi_u(s)\exp(\mathrm{i}\lambda k)\right|^2 \mathrm{d}\lambda\, \mathrm{d}s \\ &\leqslant C\int_0^1\int_{-\pi}^{\pi}\left|\sum_{u,k} x_{p(u-1)+k}\psi_u(s)\exp(\mathrm{i}\lambda k)\right|^2 \mathrm{d}\lambda\, \mathrm{d}s \\ &= 2\pi C\|x\|^2, \end{aligned}$$

which implies that the eigenvalues of E are uniformly bounded.

Appendix

In order to preserve a clear presentation of our results, we include some of the technical calculations into this separate section. We suppose throughout this section that Assumptions 1–5 are satisfied.

Let $\Sigma_{t,T} = \operatorname{cov}((X_{t-1,T}, \ldots, X_{t-p,T})')$.

Lemma A.1. *By Assumption 3, with some constants* C_1, $C_2 > 0$,

(i) $\lambda_{\max}(\Sigma_{t,T}) \leqslant C_2$ *and* $\lambda_{\min}(\Sigma_{t,T}) \geqslant C_1 + o(1)$, *where the* $o(1)$ *is uniform in* t;
(ii) *there exists some function* g, *with* $g(s) \to 0$ *as* $s \to 0$, *such that*

$$\|\Sigma_{t_1,T} - \Sigma_{t_2,T}\| \leqslant g\left(\frac{t_1 - t_2}{T}\right), \qquad \text{for all } t_1,\ t_2,\ T;$$

(iii) $c(s, k-l)$ *is uniformly continuous in* s *and*

$$\lim_{T\to\infty, t/T\to s} \operatorname{cov}(X_{t-l,T}, X_{t-k,T}) = c(s, k-l).$$

Proof. Completely analogously to the proof of Theorem 2.3 in Dahlhaus (1996), we can show that $X_{t,T}$ has the representation

$$X_{t,T} = \int_{-\pi}^{\pi} \exp(\mathrm{i}\lambda t) A^0_{t,T}(\lambda)\, \mathrm{d}\xi(\lambda),$$

with

$$\sup_{t,\lambda} |A^0_{t,T}(\lambda) - A(t/T, \lambda)| = o(1),$$

where $\xi(\lambda)$ is a process with mean zero and orthonormal increments,

$$A^0_{t,T}(\lambda) = \frac{1}{\sqrt{2\pi}} \sum_{l=0}^{\infty} \gamma_{t,T}(l) \exp(-\mathrm{i}\lambda l),$$

with $\gamma_{t,T}(l)$ given by the MA(∞) representation (7.15), and

$$A(s, \lambda) = \frac{\sigma(s)}{\sqrt{2\pi}} \left(1 + \sum_{j=1}^{p} a_j(s) \exp(-\mathrm{i}\lambda j)\right)^{-1}.$$

Then

$$\operatorname{cov}(X_{t-l,T}, X_{t-k,T}) = \int_{-\pi}^{\pi} \exp(\mathrm{i}\lambda(k-l)) A^0_{t-l,T}(\lambda) A^0_{t-k,T}(-\lambda)\, \mathrm{d}\lambda.$$

Since $A(s, \lambda)$ is uniformly continuous in s, this is equal to

$$\int_{-\pi}^{\pi} \exp(\mathrm{i}\lambda(k-l)) |A(s, \lambda)|^2\, \mathrm{d}\lambda + o(1) = c(s, k-l) + o(1), \qquad \text{for } t/T \to s,$$

which implies (iii). We obtain (ii) analogously. Furthermore, we have, for $x = (x_1, \ldots, x_p) \in \mathbb{C}^p$,

$$x^*\Sigma_{t,T}x = \int_{-\pi}^{\pi}\left|\sum_{j=1}^{p} x_j \exp(-\mathrm{i}\lambda j)A^0_{t-j,T}(\lambda)\right|^2 \mathrm{d}\lambda$$

$$= \int_{-\pi}^{\pi}|A(t/T, \lambda)|^2\left|\sum_{j=1}^{p} x_j \exp(-\mathrm{i}\lambda j)\right|^2 \mathrm{d}\lambda + \|x\|^2 o(1).$$

Under Assumption 3 there exist constants with $C_1 \leqslant |A(s, \lambda)| \leqslant C_2$ uniformly in s and λ, which implies (i). □

Lemma A.2. *Suppose additionally that Assumption 6 is satisfied, and let* $0 < \delta_1 < \delta$. *Then*

(i) $\mathrm{E}|\tilde{\alpha}^{(i)}_{lk} - \alpha^{(i)}_{lk}|^{2+\delta_1} = O(1)$,

(ii) $\mathrm{E}|\tilde{\beta}^{(i)}_{jk} - \beta^{(i)}_{jk}|^{2+\delta_1} = O(1)$

hold uniformly in i, k *and* $j < j^*$.

Proof. (i) In this part we derive estimates for the moments of $\|D'e\|$ and $\|D'S\|$, which will be used later in this proof.

Using the MA(∞) representation of $\{X_t\}$, we can write $(D'e)_{p(u-1)+k}$ as a quadratic form $\underline{\varepsilon}'A\underline{\varepsilon}$ for some $A = A(p, k)$, where $\underline{\varepsilon} = (\varepsilon_T, \ldots, \varepsilon_1, \varepsilon_0, \varepsilon_{-1}, \ldots)'$ is an infinite-dimensional vector according to the MA(∞) representation of $\{X_t\}$. Since, however, the proof of Lemma 6.1 does not depend on the dimension of the matrix A, we can apply this lemma also to this infinite-dimensional case.

We obtain, using the notation of Lemma 6.1, that

$$\tilde{A} = O(2^{-j^*/2}T),$$

$$\max\{|a_{st}|\} \leqslant \|A\|_\infty = O(2^{j^*/2}),$$

which implies

$$|\mathrm{cum}_n((D'e)_{p(u-1)+k})| \leqslant C^n(n!)^{2+2\gamma}T(2^{j^*/2})^{n-2}, \qquad \text{for } n \geqslant 2.$$

Since $\mathrm{E}(D'e)_{p(u-1)+k} = 0$, we obtain, for even s, that

$$\mathrm{E}|(D'e)_{p(u-1)+k}|^s = O\left(\sum_{r=1}^{n}\prod_{\substack{i_1,\ldots,i_r:\\ i_1+\ldots+i_r=n,\\ i_j\geqslant 1}}|\mathrm{cum}_{i_j}((D'e)_{p(u-1)+k})|\right) \leqslant C(s)T^{s/2}.$$

We obtain, with $\Delta = O(2^{j^*})$,

$$
\begin{aligned}
\mathrm{E}\|D'e\|^s &= \mathrm{E}\left(\sum_{u,k}(D'e)^2_{p(u-1)+k}\right)^{s/2} \\
&\leqslant (\Delta p)^{s/2-1}\sum_{u,k}\mathrm{E}(D'e)^s_{p(u-1)+k} \\
&= O((\Delta p)^{s/2}\max_{u,k}\{\mathrm{E}(D'e)^s_{p(u-1)+k}\}) \\
&= O(2^{j^* s/2}T^{s/2}). \qquad\qquad (\mathrm{A.1})
\end{aligned}
$$

We now treat the quantity $\|D'S\|$ in an analogous way. $(D'S)_{p(u-1)+k}$ is a quadratic form in $\underline{X} = (X_1, \ldots, X_T)'$ with a matrix A, which satisfies, according to (A.11) below,

$$
\begin{aligned}
\tilde{A} &= O\left(\sum_t |\phi_{j^*u}(t/T)|\sum_i |R_i(t/T)|\right) \\
&= O\left(\sum_i \sqrt{\sum_t \phi_{j^*u}(t/T)^2}\sqrt{\sum_t R_i(t/T)^2}\right) \\
&= O(T(2^{-j^*\min\{\tilde{s}_i\}} + T^{-1/2}2^{-j^*\min\{m_i-1/2-1/(2p_i)\}})) = O(T^{1/2})
\end{aligned}
$$

and, by (A.10),

$$
\|A\|_\infty = O\left(2^{j^*/2}\sum_i \|R_i\|_\infty\right) = O(2^{j^*/2}).
$$

Therefore, we obtain by Lemma 6.1, that

$$
|\mathrm{cum}_n((D'S)_{p(u-1)+k})| \leqslant C^n(n!)^{2+2\gamma}T(2^{j^*/2})^{n-2}, \qquad \text{for } n \geqslant 2,
$$

which implies, in conjunction with $\mathrm{E}(D'S)_{p(u-1)+k} = O(\tilde{A}) = O(T^{1/2})$, that

$$
\mathrm{E}\|D'S\|^s = O(2^{j^* s/2}T^{s/2}). \qquad\qquad (\mathrm{A.2})
$$

(ii) According to (7.5), we have $\tilde{\alpha} - \alpha = (D'D)^{-1}(D'e + D'S)$, which yields that

$$
\begin{aligned}
\mathrm{E}|\tilde{\beta}^{(i)}_{jk} - \beta^{(i)}_{jk}|^{2+\delta_1} &= \mathrm{E}|\Gamma'_{ijk}(\tilde{\alpha}-\alpha)|^{2+\delta_1} \\
&\leqslant \mathrm{E}(\|(D'D)^{-1}\|_2(\|D'e\|_2 + \|D'S\|_2))^{2+\delta_1} \\
&\leqslant (\mathrm{E}\|(D'D)^{-1}\|^{2+\delta})^{\frac{2+\delta_1}{2+\delta}}\left(\mathrm{E}(\|D'e\| + \|D'S\|)^{\frac{(2+\delta_1)(2+\delta)}{\delta-\delta_1}}\right)^{1-\frac{2+\delta_1}{2+\delta}} \\
&= O(T^{-(2+\delta_1)})O((2^{j^*/2}T^{1/2})^{2+\delta_1}) \\
&= O((2^{j^*/2}T^{-1/2})^{2+\delta_1}) = O(1). \qquad \square
\end{aligned}
$$

Lemma A.3. *Let $j^* = j^*(T) \to \infty$ and $j^* = o(T)$. Then*

(*i*) $\|(\mathrm{E}D'D)^{-1}\|_\infty = O(T^{-1})$,

$$(ii)\quad \left\| (\mathrm{E}D'D)^{-1} - \left(\left\{ T \int \phi_{j^*u}(s)\phi_{j^*v}(s) c(s,\, k-l)\,\mathrm{d}s \right\}_{p(u-1)+k,\, p(v-1)+l} \right)^{-1} \right\|_\infty = o(T^{-1}),$$

$$(iii)\quad \left\| \mathrm{cov}(D'e) - \left(\left\{ T \int \phi_{j^*u}(s)\phi_{j^*v}(s)\sigma^2(s) c(s,\, k-l)\,\mathrm{d}s \right\}_{p(u-1)+k,\, p(v-1)+l} \right) \right\|_\infty = o(T)$$

hold uniformly in u, v, k, l.

Proof. (i) Let $M = T\,\mathrm{diag}[M_1, \ldots, M_\Delta]$, where $M_u = \Sigma_t$ for any t with $t/T \in \mathrm{supp}(\phi_{j^*u})$. Because of $M^{-1} = T^{-1}\,\mathrm{diag}[M_1^{-1}, \ldots, M_\Delta^{-1}]$ we obtain by (i) and (ii) of Lemma A.1, that

$$\|M^{-1}\|_\infty = O(T^{-1}). \tag{A.3}$$

Further, we have, by $j^* = j^*(T) \to \infty$ and $j^* = o(T)$, that

$$\begin{aligned}(\mathrm{E}D'D - M)_{p(u-1)+k,\, p(v-1)+l} &= \sum_{t=p+1}^{T} \phi_{j^*u}\left(\frac{t}{T}\right)\phi_{j^*v}\left(\frac{t}{T}\right)[(\Sigma_t)_{kl} - (M_u)_{kl}] \\ &\quad + \left[\sum_{t=p+1}^{T} \phi_{j^*u}\left(\frac{t}{T}\right)\phi_{j^*v}\left(\frac{t}{T}\right) - T\delta_{uv} \right](M_u)_{kl} \\ &= o(T)\end{aligned} \tag{A.4}$$

holds uniformly in u, v, k, l. Since ϕ_{j^*u} and ϕ_{j^*v} have disjoint support for $|u-v| \geqslant C$, we obtain $(\mathrm{E}D'D)_{kl} = 0$ for $|k-l| \geqslant Cp$. Therefore we obtain, by (A.4),

$$\|\mathrm{E}D'D - M\|_\infty = o(T). \tag{A.5}$$

Because of (A.4) and (A.5) there exists a T_0 such that

$$\|M^{-1/2}(\mathrm{E}D'D - M)M^{-1/2}\| \leqslant C < 1, \qquad \text{for all } T \geqslant T_0.$$

Therefore, by the spectral decomposition of $(I + M^{-1/2}(\mathrm{E}D'D - M)M^{-1/2})$, the following inversion formula holds:

$$\begin{aligned}(\mathrm{E}D'D)^{-1} &= [M^{1/2}(I + M^{-1/2}(\mathrm{E}D'D - M)M^{-1/2})M^{1/2}]^{-1} \\ &= M^{-1/2}\left[I + \sum_{s=1}^{\infty}(-1)^s (M^{-1/2}(\mathrm{E}D'D - M)M^{-1/2})^s \right] M^{-1/2},\end{aligned} \tag{A.6}$$

which implies (i).

(ii) It can be shown in the same way as (A.4) that

$$\left\| (\mathrm{E}D'D) - \left(\left\{ T \int \phi_{j^*u}(s)\phi_{j^*v}(s)c(s,\ k-l)\,\mathrm{d}s \right\}_{p(u-1)+k,\,p(v-1)+l} \right) \right\|_\infty = o(T), \qquad \text{(A.7)}$$

which implies, analogously to (A.6),

$$\left\| (\mathrm{E}D'D)^{-1} - \left(\left\{ T \int \phi_{j^*u}(s)\phi_{j^*v}(s)c(s,\ k-l)\,\mathrm{d}s \right\}_{p(u-1)+k,\,p(v-1)+l} \right)^{-1} \right\|_\infty$$

$$= \left\| (\mathrm{E}D'D)^{-1} \sum_{s=1}^{\infty} (-1)^s [(\mathrm{E}D'D - (\{\ldots\}))(\mathrm{E}D'D)^{-1}]^s \right\|_\infty = o(T^{-1}).$$

(iii) Obviously we have

$$\mathrm{E}D'e = 0,$$

which implies

$$\begin{aligned} \operatorname{cov}((D'e)_{p(u-1)+k},\ (D'e)_{p(v-1)+l}) &= \sum_{s,t=p+1}^{T} \phi_{j^*u}\left(\frac{s}{T}\right)\phi_{j^*v}\left(\frac{t}{T}\right)\mathrm{E}\varepsilon_s\varepsilon_t X_{s-k}X_{t-l} \\ &= \sum_{s=p+1}^{T} \phi_{j^*u}\left(\frac{s}{T}\right)\phi_{j^*v}\left(\frac{s}{T}\right)\mathrm{E}\varepsilon_s^2 \mathrm{E}X_{s-k}X_{s-l} \\ &= T\int \phi_{j^*u}(s)\phi_{j^*v}(s)\sigma^2(s)c(s,\ k-l)\,\mathrm{d}s + o(T). \end{aligned}$$

The corresponding result in the $\|.\|_\infty$-norm follows from the same reasoning, leading to (A.5). □

Lemma A.4. *We have:*

(i) $\|(D'D)^{-1} - (\mathrm{E}D'D)^{-1}\|_\infty = \tilde{O}(2^{j^*/2}T^{-3/2}\sqrt{\log(T)})$;
(ii) $\|D'e\|_2^2 = \tilde{O}(2^{j^*}T\log(T))$.

Proof. (i) First, observe that by Assumption 2 and the MA(∞) representation of $\{X_t\}$,

$$\begin{aligned}
&\sum_{t_2,\ldots,t_k=1}^{T} |\mathrm{cum}(X_{t_1},\ldots,X_{t_k})| \\
&\quad = \sum_{t_2,\ldots,t_k=1}^{T} \left| \mathrm{cum}\left(\sum_{s_1=-\infty}^{t_1} \gamma_{t_1}(t_1-s_1)\varepsilon_{s_1},\ldots,\sum_{s_k=-\infty}^{t_k} \gamma_{t_k}(t_k-s_k)\varepsilon_{s_k} \right) \right| \\
&\quad \leqslant \sum_{s=-\infty}^{t_1} \sum_{t_2,\ldots,t_k=s\vee 1}^{T} |\gamma_{t_1}(t_1-s)|\cdots|\gamma_{t_k}(t_k-s)|\,|\mathrm{cum}_k(\varepsilon_s)| \\
&\quad \leqslant \sup_s\{|\mathrm{cum}_k(\varepsilon_s)|\} \sum_{s=0}^{\infty} |\gamma_{t_1}(s)| \left(\sum_{t=s\vee 1}^{T} |\gamma_t(t-s)| \right)^{k-1} \\
&\quad \leqslant C^{2k}(k!)^{1+\gamma}.
\end{aligned}$$

We see that

$$(D'D)_{p(u-1)+k,\,p(v-1)+l} = \sum_{t=p+1}^{T} \phi_{j^*u}(t/T)\phi_{j^*v}(t/T)X_{t-k}X_{t-l}$$

is a quadratic form with a matrix A satisfying, in the notation of Lemma 6.1,

$$\|A\|_\infty = O(2^{j^*}), \qquad \tilde{A} = O(T).$$

This implies, by Lemma 6.1, that

$$\begin{aligned}
|\mathrm{cum}_n((D'D)_{p(u-1)+k,\,p(v-1)+l})| &\leqslant C^n(n!)^{2+2\gamma}(2^{j^*})^{n-1}T \\
&\leqslant \left(\frac{n!}{2}\right)^{1+(1+2\gamma)} \frac{H_T}{\overline{\Delta}_T^{n-2}},
\end{aligned}$$

where $H_T \asymp 2^{j^*}T$, $\overline{\Delta}_T \asymp 2^{-j^*}$. Hence, by Lemma 6.3, we obtain that

$$P(|(D'D)_{p(u-1)+k,\,p(v-1)+l} - (\mathrm{E}D'D)_{p(u-1)+k,\,p(v-1)+l}| \geqslant x) \leqslant \exp\left(-C\frac{x^2}{2^{j^*/2}T}\right),$$

for $0 \leqslant x \leqslant (H_T^{1+\gamma}\overline{\Delta}_T)^{1/(1+2\gamma)}$. Since $(H_T^{1+\gamma}\overline{\Delta}_T)^{1/(1+2\gamma)} \asymp 2^{j^*\gamma/(1+2\gamma)}T^{(1+\gamma)/(1+2\gamma)} \gg 2^{j^*/2}T^{1/2}$, we obtain

$$(D'D)_{p(u-1)+k,\,p(v-1)+l} - (\mathrm{E}D'D)_{p(u-1)+k,\,p(v-1)+l} = \tilde{O}(2^{j^*/2}T^{1/2}\sqrt{\log(T)}).$$

Since ϕ_{j^*u} and ϕ_{j^*v} have disjoint support for $|u-v| \geqslant C$, we immediately obtain

$$\|D'D - \mathrm{E}D'D\|_\infty = \tilde{O}(2^{j^*/2}T^{1/2}\sqrt{\log(T)}), \tag{A.8}$$

which yields, in conjunction with (i) of Lemma A.3,

$$\|(D'D)^{-1} - (\mathrm{E}D'D)^{-1}\|_\infty \leqslant \|(\mathrm{E}D'D)^{-1}\|_\infty \sum_{s=1}^{\infty}(\|D'D - \mathrm{E}D'D\|_\infty \|(\mathrm{E}D'D)^{-1}\|_\infty)^s$$
$$= O(T^{-1})\tilde{O}(2^{j^*/2}T^{1/2}\sqrt{\log(T)}T^{-1})$$
$$= \tilde{O}(2^{j^*/2}T^{-3/2}\sqrt{\log(T)}).$$

(ii) From similar arguments we obtain

$$(D'e)_{p(u-1)+k} = \tilde{O}(T^{1/2}\sqrt{\log(T)}), \tag{A.9}$$

which implies (ii). □

Lemma A.5. *We have*

$$\|D'S\|_2^2 = \tilde{O}(T^2(2^{-2j^*\min\{\tilde{s}_i\}} + T^{-1}2^{-j^*\min\{sm_i-1-1/p_i\}})\log(T)).$$

Proof. Because of our assumption $m_i + 1/2 - 1/\tilde{p}_i > 1$, we get

$$\|R_i\|_\infty = O\left(\sum_{j\geqslant j^*} 2^{j/2}\max_k\{|\beta_{jk}^{(i)}|\}\right)$$
$$= O\left(\sum_{j\geqslant j^*} 2^{j/2}2^{-js_i}\right) = O(2^{-j^*(m_i-1/p_i)}) \tag{A.10}$$

and

$$TV(R_i) = O\left(\sum_{j\geqslant j^*} 2^{j/2}\sum_k |\beta_{jk}^{(i)}|\right)$$
$$= O\left(\sum_{j\geqslant j} 2^{j/2}\left(\sum_k |\beta_{jk}^{(i)}|^{p_i}\right)^{1/p_i} 2^{j(1-1/p_i)}\right)$$
$$= O\left(\sum_{j\geqslant j^*} 2^{j/2}2^{-js_i}2^{j(1-1/p_i)}\right) = O(2^{-j^*(m_i-1)}),$$

where TV(f) denotes the total variation of a function f. This implies

$$T^{-1}\sum_{t=1}^{T}(R_i(t/T))^2 - \|R_i\|_{L_2[0,1]}^2 \leqslant \sum_{t=1}^{T}\int_{(t-1)/T}^{t/T}|R_i(t/T) + R_i(u)\|R_i(t/T) - R_i(u)|\,\mathrm{d}u$$
$$= \sum_t O(T^{-1}\|R_i\|_\infty TV(R_i)|_{[\frac{t-1}{T},\frac{t}{T}]})$$
$$= O(T^{-1}2^{-j^*(2m_i-1-1/p_i)}).$$

Since we know from Theorem 8 in Donoho *et al.* (1995) that

$$\|R_i\|^2_{L_2[0,1]} = \sum_{j \geqslant j^*} \sum_k |\beta^{(i)}_{jk}|^2 = O(2^{-2j^* \tilde{s}_i}),$$

we have that

$$T^{-1} \sum_{t=1}^{T} (R_i(t/T))^2 = O(2^{-2j^* \tilde{s}_i} + T^{-1} 2^{-j^*(2m_i - 1 - 1/p_i)}). \tag{A.11}$$

Now,

$$\begin{aligned}(D'S)_{p(u-1)+k} &= \sum_{t=p+1}^{T} \phi_{j^* u}(t/T) X_{t-k} \sum_{i=1}^{p} X_{t-i} R_i(t/T) \\ &= \tilde{O}(2^{j^*/2} \sqrt{\log(T)}) \sum_{t/T \in \operatorname{supp}(\phi_{j^* u})} \sum_{i=1}^{p} |R_i(t/T)|,\end{aligned}$$

which implies

$$\begin{aligned}\|D'S\|_2^2 &= \tilde{O}(2^{j^*} \log(T)) \sum_{i=1}^{p} \sum_{u=1}^{\Delta} \left(\sum_{t/T \in \operatorname{supp}(\phi_{j^* u})} |R_i(t/T)| \right)^2 \\ &= \tilde{O}(2^{j^*} \log(T)) \sum_{i=1}^{p} \sum_{u=1}^{\Delta} \left(\sum_{t/T \in \operatorname{supp}(\phi_{j^* u})} R_i(t/T)^2 \right) T 2^{-j^*} \\ &= \tilde{O}(T^2 (2^{-2j^* \min\{\tilde{s}_i\}} + T^{-1} 2^{-j^* \min\{2m_i - 1 - 1/p_i\}}) \log(T)). \qquad \square\end{aligned}$$

Acknowledgement

We thank three anonymous referees and the Associate Editor for suggestions which helped to improve the presentation of our work. Also we are indebted to Markus Monnerjahn who contributed the software for the numerical simulation results.

References

Bentkus, R. and Rudzkis, R. (1980) On exponential estimates of the distribution of random variables. *Liet. Mat. Rink.*, **20**, 15–30.

Bergh, J. and Löfström, J. (1976) *Interpolation Spaces: An Introduction.* Berlin: Springer-Verlag.

Brillinger, D. (1994) Some asymptotics of wavelet fits in the stationary error case. Technical Report 415, Department of Statistics, University of California at Berkeley.

Cohen, A., Dahmen, W. and DeVore, R. (1995) Multiscale decompositions on bounded domains.

IGPM Preprint 114, TH Aachen.
Cohen, A., Daubechies, I. and Vial, P. (1993) Wavelets on the interval and fast wavelet transform. *Appl. Comput. Harmon. Anal.*, **1**, 54–81.
Dahlhaus, R. (1996) On the Kullback–Leibler information divergence of locally-stationary processes. *Stochastic Process. Appl.*, **62**, 139–168.
Dahlhaus, R. (1997) Fitting time series models to nonstationary processes. *Ann. Statist.*, **25**, 1–37.
Daubechies, I. (1988) Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, **41**, 909–996.
Donoho, D. (1993) Unconditional bases are optimal bases for data compression and statistical estimation. *Appl. Comput. Harmon. Anal.*, **1**, 100–115.
Donoho, D. and Johnstone, I. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.
Donoho, D. and Johnstone, I. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, **90**, 1200–1224.
Donoho, D. and Johnstone, I. (1998) Minimax estimation via wavelet shrinkage. *Ann. Statist.*, **26**, 879–921.
Donoho, D., Johnstone, I., Kerkyacharian, G. and Picard, D. (1995) Wavelet shrinkage: asymptopia? (with discussion). *J. Roy. Statist. Soc. Ser. B*, **57**, 301–337.
Graybill, F. A. (1983) *Matrices with Applications in Statistics*, 2nd edn. Wadsworth, Belmont, CA.
Hall, P. and Patil, P. (1995) Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Ann. Statist.*, **23**, 905–928.
Johnson, N. and Kotz, S. (1970) *Distributions in Statistics. Continuous Univariate Distributions*. New York: Wiley.
Johnstone, I. and Silverman, B. (1997) Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc., Ser. B*, **59**, 319–351.
Künsch, H. (1995) A note on causal solutions for locally stationary AR-processes. Manuscript, ETH Zürich.
Meyer, Y. (1990) *Ondelettes et Opérateurs I*. Paris: Herman.
Neumann, M. (1996) Spectral density estimation via nonlinear wavelet methods for stationary non-Gaussian time series. *J. Time Ser. Anal.*, **17**, 601–633.
Neumann, M. and Spokoiny, V. (1995) On the efficiency of wavelet estimators under arbitrary error distributions. *Math. Methods Statist.*, **4**, 137–166.
Neumann, M. and von Sachs, R. (1995) Wavelet thresholding: beyond the Gaussian i.i.d. situation. *In* A. Antoniadis and G. Oppenheim (eds), *Wavelets and Statistics*, Lecture Notes in Statistics 103, pp. 301–329. New York: Springer-Verlag.
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.
Rudzkis, R. (1978) Large deviations for estimates of spectrum of stationary series. *Lithuanian Math. J.*, **18**, 214–226.
Rudzkis, R., Saulis, L. and Statulevicius, V. (1978) A general lemma on probabilities of large deviations. *Lithuanian Math. J.*, **18**, 226–238.
Triebel, H. (1990) *Theory of Function Spaces II*. Basel: Birkhäuser.

Received October 1995 and revised January 1998.