INSTITUT DE STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION PAPER

0706

SHRINKAGE ESTIMATION IN THE FREQUENCY DOMAIN OF MULTIVARIATE TIME SERIES

BÖHM, H. and R. von SACHS

Shrinkage estimation in the frequency domain of multivariate time series

Hilmar Böhm and Rainer von Sachs

Institut de statistique, Université catholique de Louvain, Voie du Roman Pays 20, 1348 Louvain-la-Neuve, Belgium
boehm@stat.ucl.ac.be

30 March 2007

Abstract

In this paper on developing shrinkage for spectral analysis of multivariate time series of high dimensionality, we propose a new non-parametric estimator of the spectral matrix with two appealing properties: compared to the traditional smoothed periodogram our shrinkage estimator has a smaller L_2 risk and is numerically more stable due to a smaller condition number. We use the concept of "Kolmogorov" asymptotics where simultaneously the sample size and the dimensionality tend to infinity, to show that the smoothed periodogram is not consistent and to derive the asymptotic properties of our regularized estimator. This estimator is shown to have asymptotically minimal risk among all linear combinations of the identity and the averaged periodogram matrix. Compared to existing work on shrinkage in the time domain, our results show that in the frequency domain it is necessary to take the size of the smoothing span as "effective sample size" into account. Furthermore, we perform extensive Monte Carlo studies showing the overwhelming gain in terms of lower L_2 risk of our shrinkage estimator, even in situations of oversmoothing the periodogram by using a large smoothing span.

Key words:
multivariate time series, shrinkage, spectral analysis, regularization, condition number

We acknowledge finanical support from the IAP research network grant P 5/24 of the Belgian government (Belgian Science Policy).

1. Introduction

Spectral analysis is a method that is common to all scientists and most practitioners that work on time series. The spectrum of a stationary stochastic process is the Fourier transform of its autocovariance function. There are many ways to estimate the spectrum. The standard nonparametric approach is to smooth the periodogram, which is the square of the discrete Fourier transform of the data, around a frequency ω to obtain a local estimator of the spectrum. It is most elegant to use a kernel function for smoothing, but already a local average of the periodogram guarantees consistency and asymptotic unbiasedness. This is treated extensively in standard books on time series analysis, like [Bri75], [Pri81], [BD87] or [SS00]. In a quite straightforward way, most existing smoothing methods can be generalized to multivariate time series. This paper is concerned with improving upon the averaged periodogram as an estimator for the multivariate spectrum using regularization, i.e. shrinkage, techniques.

Estimation in the case of a $p-$variate time series suffers from a drawback that does not have an analogue in the univariate case: the result may have a bad condition number. The averaged periodogram at frequency ω is a sum over the $p \times p$ periodogram matrices at the m Fourier frequencies nearest to ω; each of the m periodogram matrices is singular, see (4). The condition number of this estimator, defined as the ratio $l_{\max}/l_{\min}$ of its largest to its smallest eigenvalue, depends not only on the condition number of the true spectrum; it is also influenced by the smoothing span m. The condition number is higher for the averaged periodogram than for the spectrum, this effect becoming negligible only if $m \gg p$.

In practice, it will only seldom be the case that we have enough data to neglect this effect. In many applications, it is severe if the estimator of the spectrum is badly conditioned. For instance, in [KST98], the authors use the Kullback-Leibler discrimination information [KL52] as a measure of disparity between several estimated multivariate spectra. Computing the Kullback-Leibler discrimination information does, however, involve inverting the estimate of one of the spectra, resulting in possibly high inaccuracy due a to bad condition number of the estimated spectrum.

In many fields of application, including economic panel data [BN02], [FHLR00], but also genetic engineering or neuropsychology, the dimension of the data may match or even exceed the sample size; in the latter case, the averaged periodogram is even singular.

In [GOvS05] finally, the authors search for the optimal partitioning of a multivariate time series into segments of approximate stationarity using a singular value decomposition of the estimated spectrum. It is a well-known (but in practice often neglected) phenomenon that, in the process of estimation, the dispersion of the sample eigenvalues is systematically larger than the dispersion of the population eigenvalues: the larger eigenvalues are biased upwards, the smaller downwards ([Jol02]). Thus, estimation can be improved by shrinking the eigenvalues towards one another.

There is indeed a large literature, e.g. [BD98], showing that in the situation of a high-dimensional target, the quality of an estimator can be improved by shrinkage not only numerically but even on the level of some theoretical criterion, such as the mean squared error. However, to the best of our knowledge, virtually all the literature is concentrated on the time domain of i.i.d. data, for which we like to cite approaches based on a decision theoretic background [Ste75], or quite differently, on "double" or *Kolmogorov* asymptotics [LW04] where simultaneously the sample size T and the dimensionality p tend to infinity.

In this paper, we address the problem of shrinkage in the frequency domain of multivariate time series. We will show that simply choosing the smoothing span of a conventional smoother, a periodogram matrix averaged over frequency, is no reasonable solution to the problem: on one hand, using the methods we will develop in this paper, even a strongly oversmoothed estimator can still be improved upon in terms of its L_2 risk. On the other hand we will show by the use of "double asymptotics", which is *the* proper theoretical framework to address the problem, that the conventional smoothed periodogram is not merely suboptimal, but not even mean square consistent.

For our proposed shrinkage estimator we follow a linear approach that combines the averaged periodogram $\hat{f}^0(\omega)$ at frequency $\omega \in (0, 2\pi]$ with the identity matrix:

$$\hat{f}_T(\omega) := r_T(\omega)\,\mathrm{Id} + s_T(\omega)\hat{f}^0_T(\omega)$$

To take on the afore-mentioned idea of reducing the dispersion of the eigenvalues of $\hat{f}_T^0(\omega)$, the factor r_T is chosen such that the sample eigenvalues are shrunk towards each other linearly. The amount of shrinkage is determined by a data driven approach that has a double asymptotic background. It is inspired by the work of [LW04], where such a framework is developed to estimate a covariance matrix based on a sample of iid data. While some of those techniques can be extended to work for non-iid data, here we face an essentially different problem: we have to develop a pointwise curve estimator $\hat{f}_T(\omega)$, which can be seen as kind of a localization of the concept of shrinkage. Compared to existing work on shrinkage in the time domain, we show that in the frequency domain it is necessary to take the size of the smoothing span m as "effective sample size" into account.

In classical asymptotic theory of frequency domain time series analysis, the smoothing span m is a function m_T of the length of the time series T that is assumed to converge to infinity, but less fast than T. In our approach, we let the dimension p grow with T, too, and the challenge is to balance the three parameters T, m_T and p_T. Such a framework is necessary because the need to shrink would vanish asymptotically if the number of dimensions were constant.

It may seem unnatural to some readers to let the dimension p_T grow with the sample size, but this is not only an indispensable tool for theory, but may as well describe what happens in practice. If you think, e.g., of a panel of economic data, it is likely that not only more and more observations are made, but also that new variables are added over time [FHLR00]. In neuropsychology, when analyzing EEG data [GOvS05], not only the observation period can be extended, but also the number of channels that are analysed may be increased to better capture localized features of the signal once sufficient observations are available. Finally, in the build-up of a monitoring system for a nuclear test ban treaty, more data may be available as more and more institutions and governments open their seismological databases [DSH02].

The remainder of the paper is organized as follows: In the next section we present our theoretical results outlined above, and construct a data driven spectral shrinkage estimator (DDSSE). This main section is followed by a presentation of extensive simulation studies where we evaluate the performance of the DDSSE and compare it with both the unshrunk averaged periodogram and a benchmark shrinkage estimator that is optimal in a certain sense, but only available if the true spectrum is known. We will see that, even for very small sample size, the improvement by our new, data driven estimator is overwhelming; using the background information needed for the benchmark estimator improves it only slightly more. The fourth section discusses both the theoretical and simulation results, links our work to existing approaches for iid data, and discusses the remaining problems and challenges for future research. Furthermore, there are two appendices. The first gives the proofs for the results of section 2, the second gives asymptotic properties of discrete Fourier transforms of random data under Kolmogorov asymptotics as well as some probabilistic lemmata, both of which are needed for the proofs.

2. Theoretical results

2.1. *Introduction to spectral analysis of multivariate stationary time series*

We assume that we observe a realisation $(X_t)_{t=1}^T$ of a p-dimensional real-valued, centered Gaussian time series (X_t). We aim at estimating the $p \times p$ spectral density matrix

$$g(\omega) = \frac{1}{2\pi T} \sum_{u \in \mathbb{Z}} \operatorname{Cov}(X_t, X_{t+u}) \exp(-i\omega u), \quad \omega \in (0, 2\pi) \tag{1}$$

The most common nonparametric estimators of (1) are based on the *periodogram*. If we denote by

$$d(\omega) = \frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} X_t \exp(-i\omega t), \quad \omega \in (0, 2\pi) \tag{2}$$

the vector-valued *discrete Fourier transformation* of the realization $(X_t)_{t=1}^T$, then the $p \times p$ periodogram matrix is defined as

$$I(\omega) := d(\omega) d^*(\omega) \tag{3}$$

where * means conjugate complex transpose. Furthermore, we will denote conjugate complex (for a scalar value) by overline. The periodogram is not a consistent estimator of the spectrum (1), and it is only asymptotically unbiased. Moreover, for $p > 1$, the periodogram is a singular matrix: if $d(\omega) = (d_1(\omega), \ldots, d_p(\omega))'$, then (3) can be expressed as

$$I(\omega) = \left(\overline{d_1(\omega)} \begin{pmatrix} d_1(\omega) \\ \vdots \\ d_p(\omega) \end{pmatrix} \quad \ldots \quad \overline{d_p(\omega)} \begin{pmatrix} d_1(\omega) \\ \vdots \\ d_p(\omega) \end{pmatrix} \right) \tag{4}$$

and thus has almost surely rank 1. If the periodogram is smoothed over frequency, the estimators derived this way are consistent under a classical asymptotical framework. In our paper, we will restrict ourselves to the simplest form of smoothing, the *averaged periodogram* with smoothing span m_T, where the conditions $m_T/T \to 0$ and $m_T \to \infty$ as $T \to \infty$ guarantee consistency and asymptotic unbiasedness:

$$\hat{f}_T^0(\omega) := \frac{1}{m_T} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} I_T(\omega + \omega_k) \tag{5}$$

where ω_k denotes the Fourier frequency $2\pi k/T$. Furthermore, we eliminate the influence of the bias of the averaged periodogram with respect to the spectrum by setting

$$f_T(\omega) := \mathrm{E}\, \hat{f}_T^0(\omega) \tag{6}$$

and constructing estimators for the expected averaged periodogram instead for the spectrum $g(\omega)$. The expected averaged periodogram (6) is a deterministic function of the spectrum. Under assumptions we will specify in section 2.4, the difference is negligible, even under general, i.e. Kolmogorov, asymptotics, see (17).

2.2. *Basic concepts and definitions*

The aim of our paper is to find an estimator of the multivariate spectrum that has less deviation from the true spectrum and better condition number than the averaged periodogram. We measure the deviation of our estimators from the true spectrum in terms of the Frobenius or Hilbert-Schmidt risk.
We will first introduce some notation and give some definitions. The loss of an estimator $\hat{f}(\omega)$ of the spectrum $g(\omega)$ at frequency ω,

$$\mathcal{L}(\hat{f}(\omega), g(\omega)) := \|\hat{f}(\omega) - g(\omega)\|^2$$

and its risk

$$\mathcal{R}(\hat{f}(\omega), g(\omega)) := \mathrm{E}\, \|\hat{f}(\omega) - g(\omega)\|^2$$

are measured in terms of a normalized Hilbert-Schmidt (HS) norm

$$\|A\|^2 := \frac{1}{p} \operatorname{tr}(AA^*) = \frac{1}{p} \sum_{i,j=1}^{p} |a_{ij}|^2. \tag{7}$$

The normalization by the dimension p enables us to set up a double asymptotic framework where the dimension p and the smoothing span m are both functions of the length T of the time series. See section 2.3 for a more detailed motivation and treatment of this. Associated to the normalized HS norm is a scalar product

$$\langle A, B \rangle := \frac{1}{p} \operatorname{tr}(AB^*)$$

which will be used as well throughout this chapter.
The enhanced estimator is chosen from the class of linear combinations of the averaged periodogram at frequency ω and the identity matrix:

$$\hat{f}_T(\omega) := r_T(\omega) \operatorname{Id} + s_T(\omega) \hat{f}_T^0(\omega) \tag{8}$$

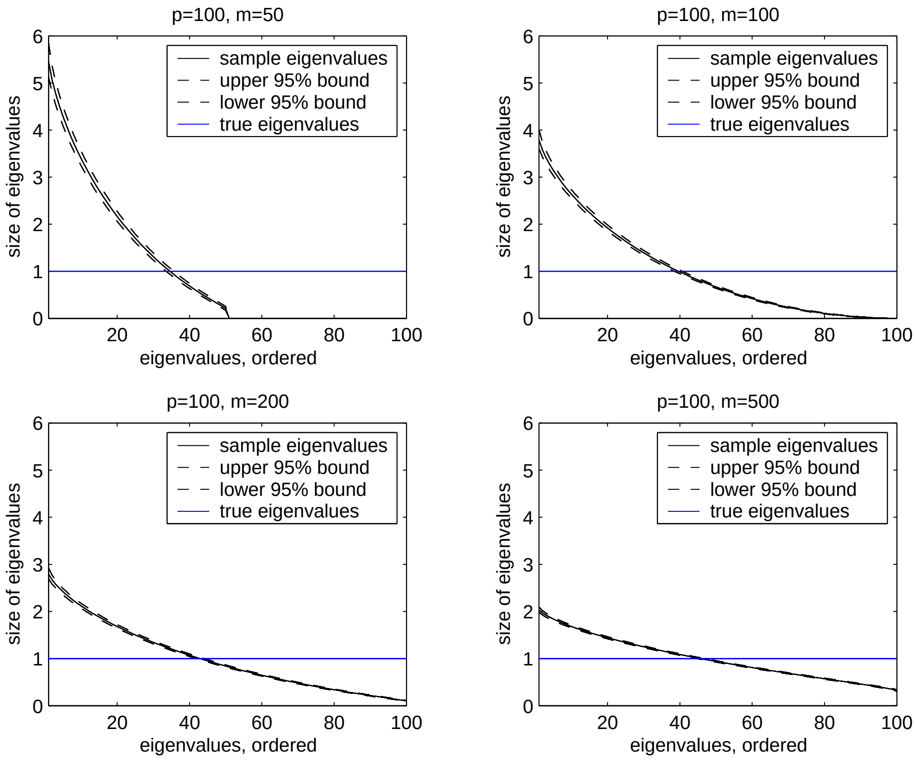

Fig. 1. Eigenvalues of the spectrum of a p-variate White Noise process for different smoothing spans m and different p. Average over $M = 1000$ simulations. The solid line shows the median, the dashed lines mark the 95% confidence bounds

The reason to choose this class is best understood when we paraphrase the problem in terms of the singular value decomposition of the averaged periodogram. Figure 1 shows the median of the sample eigenvalues at frequency $\pi/2$ of the averaged periodogram computed for a $p = 100$ dimensional multivariate white noise process of length $T = 100,000$ with innovations $\sim \mathcal{N}(0, \sqrt{2\pi}\,\mathrm{Id}_{100})$. The four plots refer to smoothing spans of $m = 50, 100, 200$ and 500. Due to the the Gaussian iid structure of the time series, the smoothed periodogram is an unbiased estimator of the spectrum here [BD87]. This does, on the other hand, not imply that the estimated eigenvalues of the spectral matrix are unbiased: The true spectrum here is, independent of frequency, the identity

$$g(\omega) \equiv \mathrm{Id}\,.$$

Thus all true eigenvalues are equal to 1. However, we see in the different subplots of figure 1 that the sample eigenvalues are strongly biased - the larger ones upwards, the smaller ones downwards - and that this bias grows with the ratio p/m. This bias is inherent to the method of the Singular Value Decomposition, which rotates a given matrix in such a way that the diagonal elements of the rotated matrix have maximum dispersion. It depends on two factors only: the ratio p/m and the true eigenvalues.
We will from now on speak of the 'estimation bias' when we mean the bias that is introduced by estimating the spectrum by the smoothed periodogram. It originates from the biasedness of the periodogram and from smoothing. We will speak of the 'sampling bias' when we mean the biasedness of the sample eigenvalues with respect to the true eigenvalues of the spectrum. We have seen that, even when there is no estimation bias, there may still be a large sampling bias, because the latter depends on the ratio p/m. What our method does is to correct the sampling bias at the price of increasing the estimation bias. In Figure 1 we see that the shift in the eigenvalues due to estimation is not linear, but may be reasonably well approximated by a linear function. Choosing the appropriate weights in (8), we linearly shrink the eigenvalues back towards one another. The reasons to prefer a linear shrinkage to a nonlinear are: First, even in the much easier case of iid data, no general results on the distribution function of the sample eigenvalues are available [Jol02], so

it would be technically difficult to prove optimality properties for a nonlinear shrinkage procedure. Then, we see in figure 1 that a linear function is a fairly good approximation of the distortion in the eigenvalues. We choose the identity matrix as a shrinkage target because it has the best possible condition number and there is no other 'natural' candidate.
It is evident from (8) that the proposed shrinkage estimator will never be worse conditioned than the averaged periodogram. The price of this is that we increase estimation bias. However, we will see that the obtained estimator is the linear combination balancing the bias-variance decomposition perfectly, thus at the same time minimizing L_2 risk in the class of linear shrinkage estimators (8).

2.3. *Kolmogorov asymptotic framework*

A proper theoretical framework is essential when looking for the optimal weights in (8). Under classical asymptotics, the sampling bias vanishes, which corresponds to consistency of the averaged periodogram. This is of no use for choosing the weights $r_T(\omega), s_T(\omega)$. Instead, we set up a double asymptotic framework where both the smoothing span and the dimension are allowed to grow with the length of the time series T. With this our estimand, the spectral matrix, becomes dependent on T, too, i.e. $g(\omega) = g_T(\omega)$. We impose the following assumption:
Assumption 1 *There exists a constant K_1 such that*

$$\frac{p_T}{m_T} \leq K_1 \qquad \forall\, T \in \mathbb{N}$$

Assumption 1 allows for the classical asymptotic framework, $p_T/m_T \to 0$, in which the averaged periodogram is consistent, as a special case, but in general, the averaged periodogram will not be consistent. This will permit us to show that our constructed shrinkage estimator has asymptotically lower risk than $\hat{f}_T^0(\omega)$.

We must, furthermore, guarantee that when increasing the dimension p_T, the overall energy in the sample does not grow too fast. We will do this by an appropriate moment condition in the frequency domain the form of which we will motivate now. First of all, in order to ensure comparability over spectra of different dimension, we have introduced a normalization in the norm (7). Second, a convenient formulation for our bound on moments will be based on the use of the basis defined by the eigenvectors of the true spectrum. Let

$$\Gamma_T(\omega)\Lambda_T(\omega)\Gamma_T^*(\omega)$$

be the eigendecomposition of the true spectrum $g_T(\omega)$ at frequency ω, the eigenvalues $\lambda_{(.)}$ in $\Lambda_T(\omega)$ ordered from the biggest to the smallest, the eigenvectors in $\Gamma_T(\omega)$ normalized. We rotate the data from the discrete Fourier transform to the eigensystem spanned by $\Gamma_T(\omega)$, defining

$$y_T(\omega) := \Gamma_T^*(\omega) d_T(\omega) \tag{9}$$

to be the rotated Fourier transform. This rotation is useful because the essential features of the cross-dimensional intercorrelation structure of the DFT and the periodogram are, asymptotically, captured in the eigenbasis $\Gamma_T(\omega)$. Making use of this, we can control both the total variance and the amount of dependence with the help of a single tool. As multiplying by $\Gamma_T(\omega)$ is an orthonormal transformation, the sum of the diagonal of both spectrum and estimate is preserved when doing so, i.e.

$$\sum_{i=1}^{p_T} g_T^{(ii)}(\omega) = \sum_{i=1}^{p_T} \lambda_T^{(ii)}(\omega)$$

and

$$\sum_{i=1}^{p_T} I_T^{(ii)}(\omega) = \sum_{i=1}^{p_T} |y_T^{(i)}(\omega)|^2$$

The challenge of the technique to use in our proofs on "double" (or Kolmogorov) asymptotics is the following. Obviously with the dimensionality $p = p_T$ to be allowed to tend to infinity with $T \to \infty$ we need some conditions on the underlying time series to be able to place ourselves into a meaningful framework. We chose to work in a framework where the Frobenius (or Hilbert-Schmidt) norm of the $p_T \times p_T$ identity matrix remains bounded. Hence we normalize the Frobenius norm by the dimensionality. As a consequence we want both our estimators and our target, the spectral matrix, to remain bounded in this normalized norm for $T \to \infty$. Quite naturally this entails the need of conditions on the correlation structure of the stationary time series (bounded sums of higher-order covariances and cross-covariances) which we prefer to give, as aforementioned, by a convenient sufficient condition in the frequency domain

Assumption 2 *There exists a constant K_2 such that for all ω and T,*

$$\sum_{i=1}^{p_T} \frac{1}{p_T} \mathrm{E}\, |y_T^{(i)}(\omega)|^8 \leq K_2$$

This assumption leads in particular to the boundedness of $||f_T(\omega)||^2$ uniformly over ω. It is convenient for two reasons - it allows for direct control of the off-diagonal contribution in the occurring spectral matrices, and it avoids to put an explicit bound on the norm $||f_T(\omega)^2||^2$ which typically occurs as nuisance in the variance of our spectral estimator: we recall that the variance of a periodogram-based estimator is proportional to the square of the target (the spectrum) itself, as it is a highly heteroskedastic nonparametric curve estimation problem. Although its control is fully understood in a classical multivariate framework, to the best of our knowledge this work is the first to contribute a rigorous development under double asymptotics. Imposing restrictions on the average eighth moment of the ys is more than imposing restrictions on the average eighth moment of the DFT. The ys take into account not only the overall variance on the diagonal of the periodogram matrix, but also the intercorrelation structure between the dimensions. Thus, imposing assumption 2, we control the whole stochastic structure of the periodogram over frequency and dimension.

2.4. *The oracle*

We now have the prerequisites to construct a shrinkage estimator with better risk than the averaged periodogram. We need two further assumption:

Assumption 3 *The real and imaginary parts of all components of the true spectrum $g(\omega)$ are twice continuously differentiable.*

Assumption 4 *The product of smoothing span and dimension grows slower than the sample size T:*

$$\frac{m_T p_T}{T} \to 0$$

In a classical asymptotic framework, asymptotic unbiasedness of the averaged periodogram, *i.e.* $f_T(\omega) \to g(\omega)$, is guaranteed by $m_T/T \to 0$ [Bri75]; in our double asymptotic framework, we need assumption 4 in addition, as the number of remainder terms in $\|f_T(\omega) - g_T(\omega)\|^2$ grows dynamically with T at a rate p_T, see lemma 4. Demanding that the second derivatives exist and are continuous allows us to keep the proofs simple. For the technical details, we point to [Bri75], [BD87] and [SS00].

We will first derive a benchmark estimator that depends on some functions of the true spectrum. This benchmark is shown to have asymptotically minimal risk. We refer to it as the *oracle*, as it cannot be derived from the data alone.

First we define

$$\mu_T(\omega) := \langle f_T(\omega), \mathrm{Id} \rangle. \tag{10}$$

This is a scale parameter, as $\langle f_T(\omega), \mathrm{Id} \rangle = \frac{1}{p_T} \sum_{i=1}^{p_T} f_T^{(ii)}(\omega)$. The optimal shrinkage parameters can now be derived by a very simple geometric argument. $f_T(\omega)$, $\hat{f}_T^0(\omega)$ and identity matrix are all entities in the Hilbert space of Hermitian p-dimensional random matrices with finite HS norm. The optimal shrinkage at frequency ω is the projection of $f_T(\omega)$ to the line spanned by the properly scaled identity matrix $\mu_T(\omega)\,\mathrm{Id}$ and the averaged periodogram $\hat{f}_T^0(\omega)$, which is illustrated in figure 2. To derive an algebraic expression for this, we first calculate the side lengths of the right-angled triangle spanned by $\mu_T(\omega)\,\mathrm{Id}$, $\hat{f}_T^0(\omega)$ and $f_T(\omega)$ as

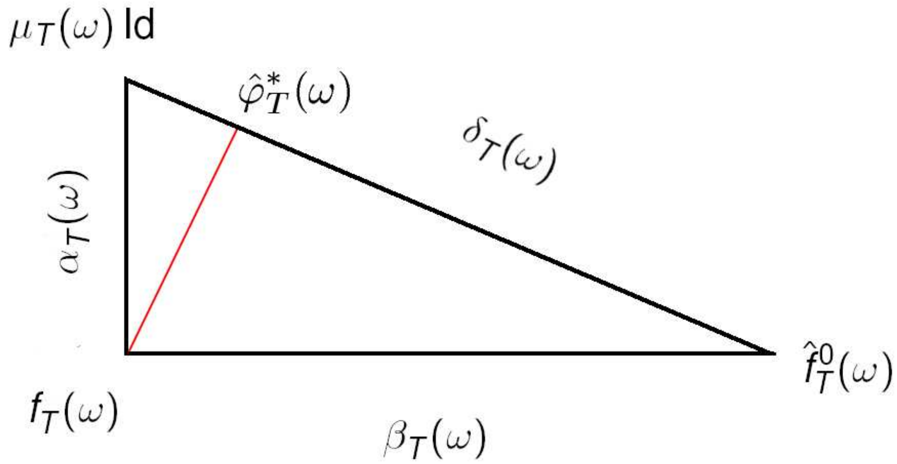

Fig. 2. Geometrical derivation of the optimal shrinkage parameters. A Pythagorean relationship $\alpha^2 + \beta^2 = \delta^2$ holds true. The optimal shrinkage parameters are derived by projecting $\mathrm{E}\,\hat{f}^0_T(\omega)$ to the line spanned by the scaled identity matrix and the averaged periodogram.

$$\alpha^2_T(\omega) := \left\| f_T(\omega) - \mu_T(\omega)\,\mathrm{Id} \right\|^2 \tag{11}$$

$$\beta^2_T(\omega) := \mathrm{E}\left\| f_T(\omega) - \hat{f}^0_T(\omega) \right\|^2 \tag{12}$$

$$\delta^2_T(\omega) := \mathrm{E}\left\| \hat{f}^0_T(\omega) - \mu_T(\omega)\,\mathrm{Id} \right\|^2 \tag{13}$$

Using the pythagorean relationship

$$\alpha^2_T(\omega) + \beta^2_T(\omega) = \delta^2_T(\omega), \tag{14}$$

we derive the optimal shrinkage parameters $r_T(\omega)$ and $s_T(\omega)$ as

$$r_T(\omega) := \frac{\beta^2_T(\omega)}{\delta^2_T(\omega)} \mu_T(\omega)$$

$$s_T(\omega) := \frac{\alpha^2_T(\omega)}{\delta^2_T(\omega)}$$

Thus, we have derived a first shrinkage estimator

$$\hat{\varphi}^*_T(\omega) := \frac{\beta^2_T(\omega)}{\delta^2_T(\omega)} \mu_T(\omega)\,\mathrm{Id} + \frac{\alpha^2_T(\omega)}{\delta^2_T(\omega)} \hat{f}^0_T(\omega) \tag{15}$$

to which we refer as the oracle, as it requires four functions of the expected averaged periodogram to be known. The asymptotic behavior of these four functions under the assumptions 1 and 2 is given by the following

Lemma 1 *As $T \to \infty$, all four functions $\mu_T(\omega)$, $\alpha^2_T(\omega)$, $\beta^2_T(\omega)$ and $\hat{\delta}^2_T(\omega)$ remain bounded.*

The proofs of this lemma and of all following results in this chapter can be found in appendix A. The first important result that we show is the asymptotic behavior of the averaged periodogram under Kolmogorov asymptotics:

Theorem 1

$$\lim_{T\to\infty} \mathrm{E}\left\|\hat{f}_T^0(\omega) - f_T(\omega)\right\|^2 - \frac{p_T}{m_T}\mu_T^2(\omega) = 0 \tag{16}$$

This result is essential. We see immediately that, under a Kolmogorov asymptotic framework, the averaged periodogram is no longer necessarily consistent. This is the most essential feature of the Kolmogorov framework: by increasing the dimension with the sample size, the asymptotic risk of different estimators can be compared, whereas under the classical framework, the possibly bad finite sample size properties of these estimators are hidden by the fact that they are consistent.
The conditions under which $\hat{f}_T^0(\omega)$ remains consistent in the more general framework are either if $\frac{p_T}{m_T} \to 0$, which is a special case including the classical framework, or when $\mu_T(\omega) \to 0$. The latter means that, asymptotically, the total variance of the periodogram becomes negligible with respect to the dimension p_T, as the variance of the periodogram is determined by the trace of the spectrum.

The risk of (15) at estimating the expected averaged periodogram $f_T(\omega)$ is, due to construction, minimal amongst all estimators of the type (8). Asymptotically, the oracle constitutes the minimal risk estimator of the spectrum $g_T(\omega)$, too, as

$$\|f_T(\omega) - g_T(\omega)\|^2 \to 0 \tag{17}$$

at a rate of $\mathrm{O}\left(\frac{m_T p_T}{T}\right)$ due to assumption 4 and lemma 4. Now, the minimal risk estimator in the class of linear shrinkage estimators (8) could be the averaged periodogram itself, the oracle not providing a real improvement. The following theorem shows that this is not the case:

Theorem 2 *The risk of the oracle with respect to the expected averaged periodogram is given by*

$$\mathrm{E}\,\|\hat{\varphi}_T^*(\omega) - f_T(\omega)\|^2 = \frac{\alpha_T^2(\omega)\beta_T^2(\omega)}{\delta_T^2(\omega)}$$

As the risk of $\hat{f}_T^0(\omega)$ is $\beta_T^2(\omega)$, this means that

$$\mathcal{R}(\hat{\varphi}_T^*(\omega), f_T(\omega)) = \frac{\alpha_T^2(\omega)}{\delta_T^2(\omega)}\mathcal{R}(\hat{f}_T^0(\omega), f_T(\omega)).$$

However, as the pythagorean relationship (14) holds true,

$$\frac{\alpha_T^2(\omega)}{\delta_T^2(\omega)} < 1 \quad \text{a.s.}$$

and so the oracle has almost surely lower risk than $\hat{f}_T^0(\omega)$.

2.5. *Data driven shrinkage estimation*

Our next step is to derive an estimator that no longer requires knowledge of functional parameters of the true spectrum. This is done by estimating the parameters in (10), (11), (12), (13) and plugging in the estimators in (15). The trace $\mu_T(\omega)$ of $f_T(\omega)$, is estimated by the trace of the averaged periodogram:

$$\hat{\mu}_T(\omega) := \langle \hat{f}_T^0(\omega), \mathrm{Id}\rangle. \tag{18}$$

Likewise, the estimator of $\delta_T^2(\omega)$ is a sample version of $\delta_T^2(\omega)$, derived by omitting the expected value:

$$\hat{\delta}_T^2(\omega) := \left\|\hat{f}_T^0(\omega) - \hat{\mu}_T(\omega)\,\mathrm{Id}\right\|^2 \tag{19}$$

We cannot, however, derive estimators of $\alpha_T^2(\omega)$ and $\beta_T^2(\omega)$ in an as straighforward way. $\beta_T^2(\omega)$ is the local variance of the averaged periodogram at frequency ω. It can be estimated by some kind of sample variance, where the data are only asymptotically uncorrelated and only have approximately identical first and second moments. We neglect the deviation from iid and construct this estimator of $\beta_T^2(\omega)$ as follows:

$$\bar{\beta}_T^2(\omega) := \frac{1}{m_T^2} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \left\| I_T(\omega+\omega_k) - \hat{f}_T^0(\omega) \right\|^2$$

We must ensure that our estimator of $\beta_T^2(\omega)$ is not greater than $\hat{\delta}_T^2(\omega)$, thus we define

$$\hat{\beta}_T^2(\omega) := \min(\bar{\beta}_T^2(\omega), \hat{\delta}_T^2(\omega)) \tag{20}$$

Finally, we use the Pythagorean relationship $\alpha_T^2(\omega) + \beta_T^2(\omega) = \delta_T^2(\omega)$ and estimate $\alpha_T^2(\omega)$ by

$$\hat{\alpha}_T^2(\omega) := \hat{\delta}_T^2(\omega) - \hat{\beta}_T^2(\omega) \tag{21}$$

These estimators are consistent under the double asymptotic framework, which is ensured by the following lemma:

Lemma 2 *If Assumptions 1 and 2 hold true, then we have, for any ω and $T \to \infty$*

$$\begin{aligned}
\mathrm{E}(\hat{\mu}_T(\omega) - \mu_T(\omega))^4 &\to 0\\
\mathrm{E}(\hat{\alpha}_T^2(\omega) - \alpha_T^2(\omega))^2 &\to 0\\
\mathrm{E}(\hat{\beta}_T^2(\omega) - \beta_T^2(\omega))^2 &\to 0\\
\mathrm{E}(\hat{\delta}_T^2(\omega) - \delta_T^2(\omega))^2 &\to 0
\end{aligned}$$

Lemma 2 permits us to construct a *data driven spectral shrinkage estimator* (DDSSE), which requires no background knowledge of the true spectrum. It is derived by simply plugging in the estimators (18),(19),(20) and (21) for their estimands in the definition of the oracle (15):

$$\hat{f}_T^*(\omega) := \frac{\hat{\beta}_T^2(\omega)}{\hat{\delta}_T^2(\omega)} \hat{\mu}_T(\omega)\,\mathrm{Id} + \frac{\hat{\alpha}_T^2(\omega)}{\hat{\delta}_T^2(\omega)} \hat{f}_T^0(\omega) \tag{22}$$

The central result of this paper is that, asymptotically, the difference between the DDSSE and the oracle vanishes:

Theorem 3 *f_T^* is a mean square consistent estimator of $\hat{\varphi}_T^*$, i.e.*

$$\mathrm{E}\left\| \hat{f}_T^*(\omega) - \hat{\varphi}_T^*(\omega) \right\|^2 \to 0.$$

As a result, the risk of the DDSSE is, in the limit, the same as the risk of the oracle:

$$\mathrm{E}\left\| \hat{f}_T^*(\omega) - f_T(\omega) \right\|^2 - \mathrm{E}\left\| \hat{\varphi}_T^*(\omega) - f_T(\omega) \right\|^2 \to 0$$

Thus, we have derived an estimator that asymptotically is optimal in the class of linear shrinkage estimators (8), and which can be calculated without any knowledge of the true spectrum. As asymptotically also

$$\| f_T(\omega) - g_T(\omega) \|^2 \to 0$$

the DDSSE is an estimator for the multivariate spectrum with asymptotically minimal risk.

In the following empirical section we will see that the DDSSE performs extremely well even for very small datasets.

3. Monte Carlo results

How does the DDSSE perform in practice? If we have a finite time series, is it justified to rely on the DDSSE rather than on a conventional estimator of the spectrum? A comprehensive Monte Carlo study we have run shows that we should indeed use the DDSSE, even for very short multidimensional time series.

3.1. *Setup*

The simulations aim at comparing the DDSSE with the averaged periodogram, upon which we want to improve, on one hand, and on the other hand with the oracle, which we have used in theory as a benchmark. For each of these estimators, we compute risk, bias and variance.
In the process of developing the DDSSE, we have performed simulations that have shown which aspects of the underlying time series do or do not matter for the MC study. We have chosen a number of 5-dimensional time series of different lengths. To examine the influence of the condition number of the true spectrum, we use $T = 128$. To examine the influence of the smoothing span, we have also simulated longer time series. The product $m_T p_T$ can be seen as a measure of distance from 'infinity' under double asymptotics. Analogously to the classical framework, increasing the smoothing span enhances precision. Moreover increasing p_T for fixed ratio p_T/m_T means that the confidence intervals for the eigenvalues become smaller, improving the precision of the estimators $\hat{\mu}_T(\omega), \hat{\alpha}_T^2(\omega), \hat{\beta}_T^2(\omega)$ and $\hat{\delta}_T^2(\omega)$. Figure 3 illustrates this effect.

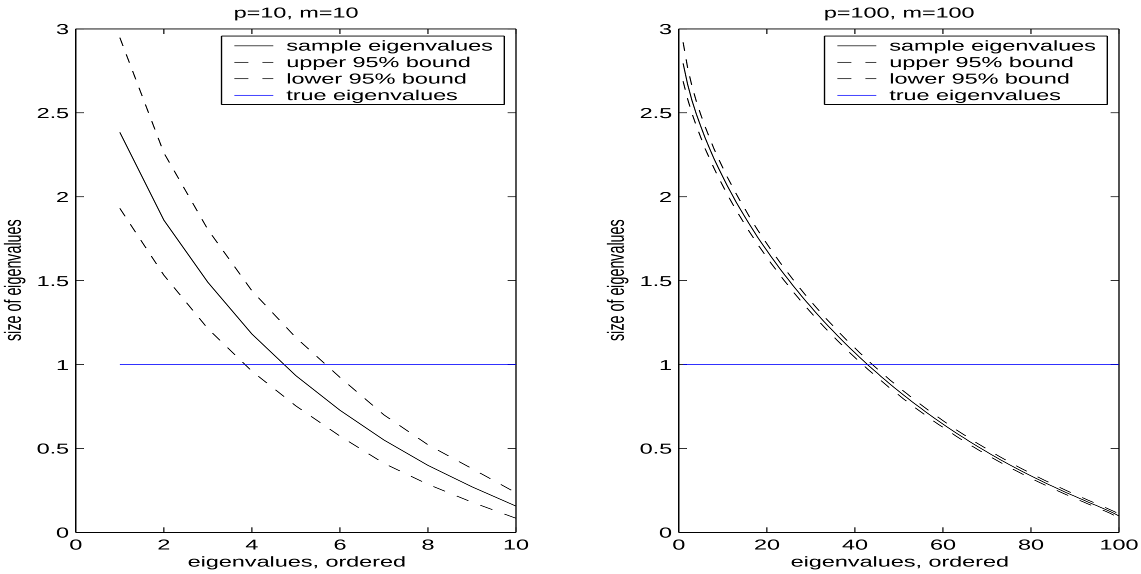

Fig. 3. Distribution of the eigenvalues for different dimensionality: on the left $p_T = 10$, on the right $p_T = 100$.

The underlying process is a vector valued MA(2) with normal innovations:

$$X_t = \theta_0 e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}, \quad e_t \sim \mathcal{N}(0, \mathrm{Id}_5) \tag{23}$$

The coefficients θ are chosen differently to enable different condition numbers c. We give as an example the coefficients for condition number $c = 100$.

$$\theta_0 = (1.4072 \;\; 1.2207 \;\; 1 \;\; .7141 \;\; .1407)\,\mathrm{Id} \tag{24}$$

$$\theta_1 = (.3391 \;\; .2941 \;\; .2409 \;\; .1721 \;\; .0339)\,\mathrm{Id} \tag{25}$$

$$\theta_2 = -(.7750 \;\; .6723 \;\; .5508 \;\; .3933 \;\; .0775)\,\mathrm{Id} \tag{26}$$

The spectrum of the process (23) is a diagonal matrix function with condition number, at each frequency, equal to c. In case of the weights chosen in (24) to (26) $c = 100$. Moreover, for any frequency ω, the eigenvalues of the true spectrum are equidistantly apart in our setup:

$$\lambda_1(\omega) - \lambda_2(\omega) = \lambda_2(\omega) - \lambda_3(\omega) = \ldots = \lambda_4(\omega) - \lambda_5(\omega)$$

A picture of the spectrum is given in figure 4. It may seem unintuitive that the cross-spectra are set to zero; yet this may be done without loss of generality. The results of the algorithms only depend on the true eigenvalues of the spectrum; it makes no difference in which basis the data are represented as the knowledge of a diagonal spectrum is not used in the algorithm. E.g., the parameters $r_T(\omega), s_T(\omega)$ for the DDSSE depend on scalar products only, which are invariant to orthonormal rotations.

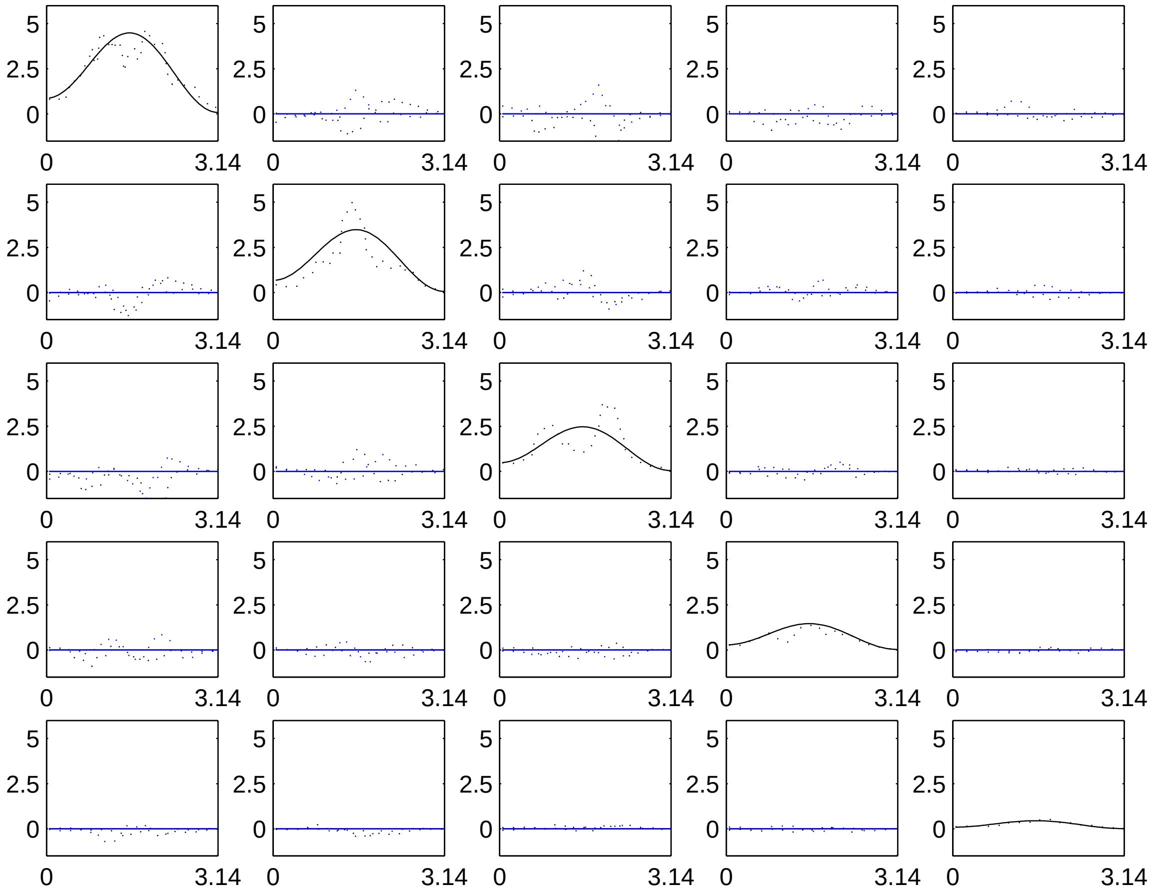

Fig. 4. The spectrum (solid lines) and the averaged periodogram (dotted lines) with smoothing span $m_T = 9$ of the 5-dimensional process. The condition number is $c = 10$ here, the underlying time series has length $T = 128$. The real components are in black, the imaginary components in blue.

3.2. *Influence of the condition number*

We will first study the influence of the condition number of the true spectrum on the risk, bias and variance of the estimators $\hat{f}_T^0(\omega)$(averaged periodogram), $\hat{f}_T^*(\omega)$(DDSSE), $\hat{\varphi}_T^*(\omega)$(oracle). We choose a smoothing span of $m = 7$, which is very small compared to the dimension $p = 5$, and vary the condition number c from 1 to 10^9. We expect that our DDSSE performs best for small condition number, as the sampling bias is maximal in this case [Jol02], so we are interested in its asymptotic behavior for $c \to \infty$. We present the results graphically in figure 5; for $c = 10$, we also give a graphical representation in figure 6, which shows that variance, squared bias and risk are proportional to the energy in the spectrum at the respective frequency. The results of the simulations show that the DDSSE does indeed perform remarkably better than the averaged periodogram. We first discuss the results given in figure 5. For condition number greater or equal to 10, the risk of the DDSSE is approximately half as big as the risk of the averaged periodogram. The improvement is slightly better for $c = 10$, then converges quickly to its limit, which seems already obtained at $c = 100$. Moreover, the oracle, which is our benchmark here, performs better than the DDSSE, as expected, and has asymptotic risk approximately equal to 37% of the averaged periodogram. We expected the oracle, which uses background knowledge of the true spectrum, to perform better than the DDSSE.

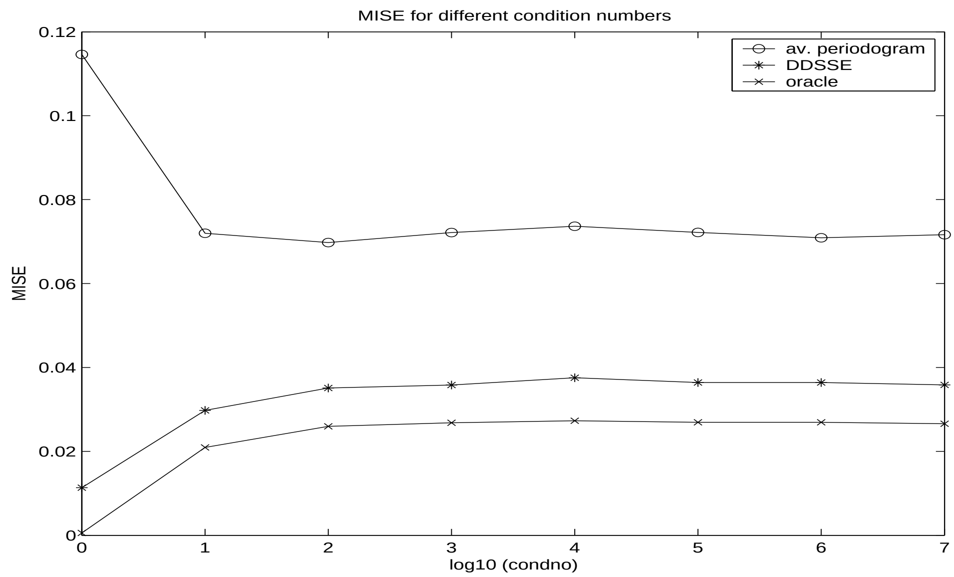

Fig. 5. MISE in function of condition number for the averaged periodogram, DDSSE and oracle.

Yet, the improvement in terms of the risk that the oracle offers over the DDSSE is clearly smaller than the improvement in terms of the risk that the DDSSE offers over the averaged periodogram.
The case $c = 1$ is distinct. We see that here the improvement by both the DDSSE and the oracle is huge, the risk of the latter being only 0.5% of the risk of the averaged periodogram. This is however an artifact: for $c = 1$, the spectrum is just a multiple of the identity matrix. Thus, shrinking in this case can be seen as a special, parametric case of the otherwise nonparametric shrinking procedure, resulting in an abnormally huge improvement. Next, we look at figure 6, which shows on one hand the bias-variance decomposition of

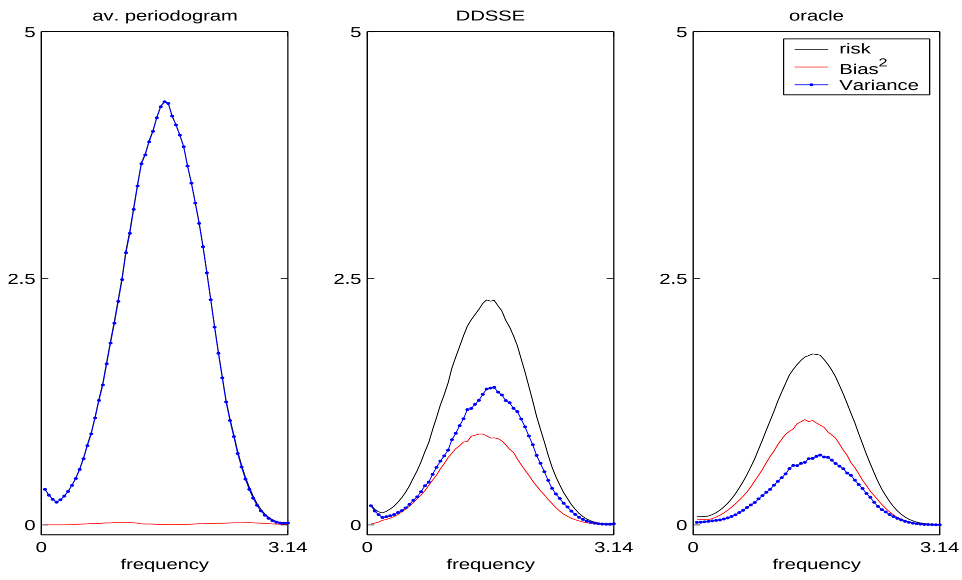

Fig. 6. risk, squared bias and variance in function of frequency

the estimators, on the other hand their dependency on the frequency, for $c = 10$. The latter shows the same shape as the spectrum - squared bias, variance and MSE are proportional to the energy of the spectrum at the respective frequency. This makes it easier to interpret our results, and justifies our use of the MISE as a measure of risk above. Looking at the bias-variance decomposition of the estimators, we first remark that the averaged periodogram is almost all variance and not bias. This is because the bias is only due to smoothing, and the smoothing span is very small here. Also, the true spectrum is not too peaky, which would increase the bias. The risk of the DDSSE, on the other hand, is about equally squared bias and variance. It is the idea behind the DDSSE to introduce a bias in order to reduce the risk, and this is confirmed by the Monte Carlo results. Proceeding to the oracle, we see that the bias here is about the same size as for the DDSSE, whereas the variance is much smaller. This is because the shrinkage parameters ρ_1, ρ_2 are deterministic for the oracle, eliminating one major source of variance compared to the DDSSE. Overall, the oracle still improves on the DDSSE.

3.3. *Influence of the smoothing span*

Next we examine the influence of the smoothing span on the performance of the three estimators. We fix the condition number at $c = 100$ and vary the smoothing span from $m = 7$ to $m = 23$, which corresponds to roughly a smoothing span of 10% to 33% of the time series. We first see in figure 7 that the optimal

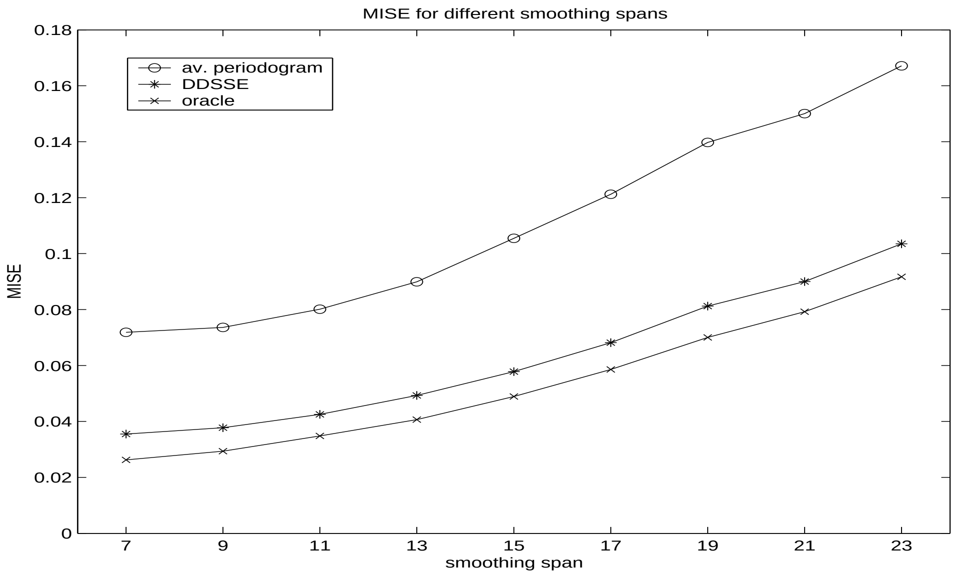

Fig. 7. MISE in function of smoothing span, for a time series of length $T = 128$

smoothing span in terms of the MISE is $m = 7$ here for the averaged periodogram. Increasing the smoothing span results in a worse overall quality of estimation. Moreover, the relative improvement of the DDSSE and oracle over the averaged periodogram here are best for the smallest chosen smoothing span $m = 7$. Thus, the optimal smoothing span is the same for all estimators here. Yet, even when oversmoothing a lot, we still can improve significantly on the results by replacing the averaged periodogram by the DDSSE. The DDSSE thus shows a certain degree of robustness, which is important to remark, as the estimators $\hat{\mu}, \hat{\alpha}, \hat{\beta}, \hat{\delta}$ are biased with respect to the parameters μ, α, β and δ, and the size of this bias depends on the smoothing span. As these estimators determine the amount of shrinkage, it would not be contradictory to theory that this might result in the DDSSE performing worse than the averaged periodogram for finite sample size. Yet the simulations show that the opposite is true. We have performed additional MC runs on the same time series as in (23), with $c = 100$, but for lengths of $T = 256$ and $T = 512$, and with varying smoothing span to empirically choose its optimum. The results are given in figure 8. First, we remark that the DDSSE never has higher risk than $\hat{f}^0_T(\omega)$. This again confirms our observation that in practice, the DDSSE may just be used to replace a conventional estimator without concerns about increasing the risk.
Moreover, we see that, surprisingly, in each MC study the optimal smoothing span for $\hat{f}^0_T(\omega)$, $\hat{\varphi}^*_T(\omega)$ and

$\hat{f}^*_T(\omega)$ almost coincide. We assume that there is some link between the optimal smoothing spans that we have not yet discovered. However, if it turns out to be true that the two optimal smoothing spans are identical, this would be a very good feature, as it would enable us to deploy our shrinkage estimator in a simple two step procedure: use existing theory for bandwidth choice as derived for a conventional estimator [OYR86], [Lee97] and then replace the estimator by the DDSSE. This will be subject to further research.

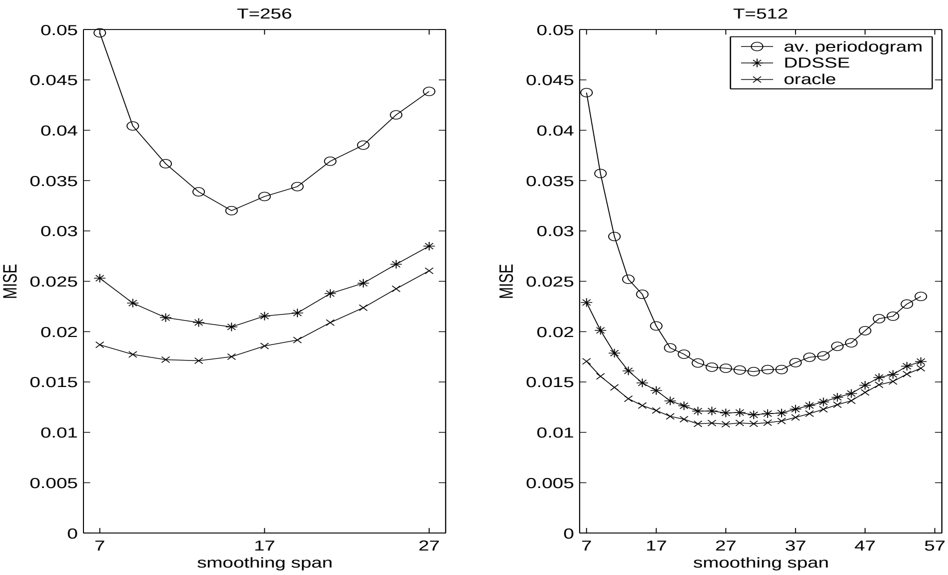

Fig. 8. MISE in function of smoothing span, for time series of different length.

4. Discussion

To the best of the authors' knowledge, there exists no approach to regularize an estimate of the spectrum using shrinkage techniques so far. The idea of shrinkage, however, is not new. The earliest ideas go back to a lecture by Stein [Ste75]. Various authors have based shrinking techniques for iid data on these concepts, among them [Haf79], [Haf80] and [DS85]. The theoretical background of these estimators is not classical or double asymptotics; they rather follow a decision theoretic approach: in a class $\mathcal{D}$ of estimators for $\theta \in \Theta$, an estimator has the *minimax* property iff

$$\sup_{\theta\in\Theta} \mathrm{R}(\theta, \delta^*) = \inf_{\delta\in\mathcal{D}} \sup_{\theta\in\Theta} \mathrm{R}(\theta, \delta). \tag{27}$$

The minimax property is not unique, so the progress in this area of research focuses on finding new minimax estimators that *dominate*, i.e. have uniformly smaller or equal risk than, an estimator that has been shown to have the minimax property. While a minimax estimator from a class that includes $f_T(\omega)$ would be guaranteed to have risk uniformly smaller than or equal to that of the averaged periodogram, the minimax property is often a too conservative restriction to guarantee that an estimator offers substantially lower risk in practice. A different approach is to examine shrinkage with the aid of double asymptotics, as introduced in [LW04] for covariance matrices of iid data. That paper also includes a comprehensive Monte Carlo study comparing the newly introduced shrinkage estimator based on double asymptotics to various minimax estimators.
The authors of this paper also started to investigate this comparison by an explorative Monte Carlo study. It is rather straightforward to adapt the minimax estimators to the time series case in an ad hoc manner. In the explorative simulations, it turned out that the minimax estimators performed by far not as good as the shrinkage estimator presented here, constructed with the aid of Kolmogorov asymptotics. This is why we decided to follow the latter approach; furthermore, it might be technically difficult to generalize the minimax estimators to the frequency domain, as the iid assumption is essential in the derivation of these.

A double asymptotic framework has become an almost common tool in recent research on time series and panel data. It [FHLR00], it is used to distinguish between idiosyncratic and global common components in the analysis of economic panel data. The authors of [MHvS06] use it to identify time variant factors driving a nonstationary time series, where time is rescaled according to the Dahlhaus model of locally stationary time series [Dah96a], [Dah96b], [Dah00]. In our work, we use it to derive an enhanced estimator of the spectrum that asymptotically has minimal risk in a class of linear estimators that is chosen to approximately compensate for the bias of the eigenvalues of the averaged periodogram. We have shown in section 2 that the DDSSE has asymptotically the same risk as the oracle; the latter is the optimal estimator of the expected periodogram with respect to Hilbert-Schmidt norm. We have also shown in theorem 2 that the risk of the oracle is truly smaller than that of $\hat{f}_T^0(\omega)$. Asymptotically, all these properties are attained by the DDSSE as well as by the oracle, giving rise to a simple data driven approach to enhanced spectral estimation, reducing the risk and improving the condition number at the same time. Moreover, is is computationally cheap, as the floating point operations needed to calculate the DDSSE are of the same order as the floating point operations needed to calculate $\hat{f}_T^0(\omega)$. This is another reason why this approach is superior to minimax theory, which always involves an expensive singular value decomposition.

What we are doing can be seen as finding a new bias-variance balance for an estimator of the spectrum. The bias in $\hat{f}_T^0(\omega)$ is due to smoothing and due to the biasedness of the periodogram. We add another source of bias, the shrinkage target $\mu_T(\omega)\,\mathrm{Id}$, reducing the variance. The oracle constitutes the optimal balance between bias and variance, and the DDSSE constitutes an approximate optimum.

What is more important yet is the fact that the DDSSE performs well for finite sample size, too. This is not guaranteed by theory, which for finite sample size only shows that the oracle has minimal risk. To gain the DDSSE, however, the parameters $\mu_T(\omega), \alpha_T^2(\omega), \beta_T^2(\omega)$ and $\delta_T^2(\omega)$ need to be estimated to gain (15). We have no theoretical results about how precisely these parameters are estimated. For sure is that their estimators $\hat{\mu}_T(\omega)$ etc. are biased, as they require smoothing over neighbouring periodogram frequency. It would not be contradictory to theory if, for finite sample size, the estimator $\hat{f}_T^*(\omega)$ were not only worse than the oracle, but even worse than the averaged periodogram $f_T(\omega)$. Fortunately, none of the simulations confirm these concerns. There is one single simulation in which the risk of the DDSSE becomes larger than that of $\hat{f}_T^0(\omega)$, namely when, for a sample size of $T = 256$, the smoothing span $m_T > 71$, which is clearly oversmoothing. In all other scenarios, the DDSSE has smaller risk, making it an excellent alternative to conventional estimators that is at the same time more precise and robust to use in all areas where the inverse of the spectrum needs to be estimated.

Another question we have adressed empirically is that of the choice of the smoothing span. We had expected the optimal smoothing span to be larger for the averaged periodogram than for the DDSSE; however, in all simulations the smoothing spans that have minimal risk coincide for the averaged periodogram and the DDSSE. We therefore assume that classical methods for choosing the smoothing span might be transferred to the DDSSE; however, this will require future research.

The oracle has minimal risk among all linear combinations

$$r_T(\omega)\,\mathrm{Id} + s_T(\omega)\hat{f}_T^0(\omega) \tag{28}$$

with nonrandom coefficients $r_T(\omega), s_T(\omega)$. It would be more interesting to allow for these coefficients to be random, too. In fact, our theory can be extended to allow for nonrandom coefficients. In this case, we have another benchmark replacing the oracle, which is optimal in the larger class of all linear combinations (28), to which we refer as the *optimal spectral shrinkage estimator* (OSSE). We have been investigating this, and we have shown that in this case all three, oracle, OSSE and DDSSE have, asymptotically, the same risk. Moreover, simulations we have run seem to point that, for finite sample size, the risk of OSSE is roughly of the same size as that of the oracle. These results will be reported elsewhere.

Appendix A. Proofs

This section gives the proofs of the results in the theory section. The proofs are given in the order in which the results are given in the text. Basic results on the asymptotic rates of function of the discrete Fourier transform under Kolmogorov asymptotics, which are needed throughout the proofs, are given separately in section B. We will make frequent use of the abbreviations $\tilde{\omega}$, which means the Fourier frequency nearest ω, and $\tilde{\omega}_k := \tilde{\omega} + \omega_k$. Furthermore, we mean by '$\approx$' that the difference of the terms on the left and right hand side of '$\approx$' is asymptotically negligible.

A.1. *Proof of Theorem 2*

This can be shown by a simple geometrical argument: The risk $\mathcal{R}(\hat{\varphi}^*_T(\omega), \mathrm{E}\,\hat{f}^0_T(\omega))$ is the distance between the two respective points in the Hilbert space of Hermitian p-dimensional random matrices. In figure A.1 we see that the two triangles $\mathrm{E}\,\hat{f}^0_T(\omega) \longleftrightarrow \mu_T(\omega)\,\mathrm{Id} \longleftrightarrow \hat{f}^0_T(\omega)$ and $\hat{\varphi}^*_T(\omega) \longleftrightarrow \mathrm{E}\,\hat{f}^0_T(\omega) \longleftrightarrow \mu_T(\omega)\,\mathrm{Id}$ are similar. The length of the perpendicular dropped on the hypotenuse of the large triangle, which constitutes the longer cathetus of the smaller triangle, is thus equal to $\alpha^2_T(\omega) \times \frac{\beta^2_T(\omega)}{\delta^2_T(\omega)}$.

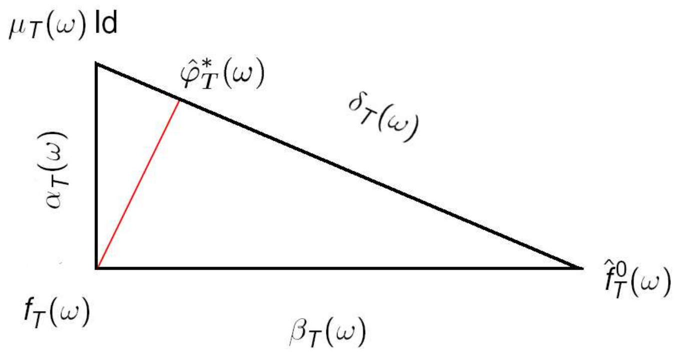

Fig. A.1. Geometrical derivation of the risk of the oracle. The risk $\mathcal{R}(\hat{\varphi}^*_T(\omega), \mathrm{E}\,\hat{f}^0_T(\omega))$ is the distance between the two respective points.

□

A.2. *Proof of lemma 1*

We will first show that the norm of $\mathrm{E}\,\hat{f}^0_T(\omega)$ is uniformly bounded in ω. This is then used to show that $\mu_T(\omega)$ and $\alpha^2_T(\omega)$ are bounded, too. The boundedness of $\beta^2_T(\omega)$ is implied by theorem 1. Due to the relationship $\alpha^2_T(\omega) + \beta^2_T(\omega) = \delta^2_T(\omega)$, this is sufficient to show the boundedness of $\mu_T(\omega), \alpha^2_T(\omega), \beta^2_T(\omega)$ and $\delta^2_T(\omega)$.

$$\left\|\mathrm{E}\,\hat{f}^0_T(\omega)\right\|^2 = \left\|\frac{1}{m_T}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \mathrm{E}\,I_T(\tilde{\omega}_k)\right\|^2 \le \frac{2m_T-1}{m_T^2}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \|\mathrm{E}\,I_T(\tilde{\omega}_k)\|^2$$
$$\le 2\sup_\omega \|\mathrm{E}\,I_T(\omega)\|^2 = 2\sup_\omega \|\mathrm{E}\,y_T(\omega)y^*_T(\omega)\|^2 = 2\sup_\omega \frac{1}{p_T}\sum_{i,j=1}^{p_T} |\,\mathrm{E}\,y_i(\omega)y^*_j(\omega)|^2$$
$$= \underbrace{2\sup_\omega \frac{1}{p_T}\sum_{i=1}^{p_T}\left(\mathrm{E}\,|y_i(\omega)|^2\right)^2}_{\mathrm{I}} + \underbrace{2\sup_\omega \frac{1}{p_T}\sum_{\substack{i,j=1\\ i\neq j}}^{p_T}\left|\mathrm{E}\,y_i(\omega)y^*_j(\omega)\right|^2}_{\mathrm{II}}$$

Part I is bounded because

$$\mathrm{I} \leq 2 \sup_\omega \frac{1}{p_T} \sum_{i=1}^{p_T} \mathrm{E}\,|y_i(\omega)|^4 \leq 2 \sup_\omega \sqrt{\frac{1}{p_T} \sum_{i=1}^{p_T} \mathrm{E}\,|y_i(\omega)|^8} \overset{\text{Ass.2}}{\leq} 2\sqrt{K_2}$$

Part II vanishes asymptotically. According to lemma 7 in the Appendix B, no matter if $i = j$ or $i \neq j$, we have $\mathrm{E}\, y_i(\omega) y_j^*(\omega) - \lambda_{ij}(\omega) = \mathrm{O}\left(\frac{p_T}{T}\right)$, uniformly in ω. Here $\lambda_{ij}(\omega) = 0$ because $i \neq j$, so $\mathrm{E}\, y_i(\omega) y_j^*(\omega) = \mathrm{O}\left(\frac{p_T}{T}\right)$. We obtain

$$\mathrm{II} = 2 \frac{1}{p_T} \sum_{\substack{i,j=1 \\ i \neq j}}^{p_T} \left| \mathrm{O}\left(\frac{p_T}{T}\right) \right|^2 = \mathrm{O}\left(\frac{p_T^3}{T^2}\right)$$

which according to assumption 2 converges to zero. Thus $\left\| \mathrm{E}\, \hat{f}_T^0(\omega) \right\|^2$ is bounded. Using this, we can easily show that $\mu_T(\omega)$ and $\alpha_T^2(\omega)$ are bounded:

$$\mu_T(\omega) = \langle \mathrm{E}\, \hat{f}_T^0(\omega), \mathrm{Id} \rangle \leq \sqrt{\left\| \mathrm{E}\, \hat{f}_T^0(\omega) \right\|^2 \|\mathrm{Id}\|^2} \leq \left\| \mathrm{E}\, \hat{f}_T^0(\omega) \right\|$$

$$\alpha_T^2(\omega) = \left\| \mathrm{E}\, \hat{f}_T^0(\omega) - \mu_T(\omega)\, \mathrm{Id} \right\|^2 \leq \left\| \mathrm{E}\, \hat{f}_T^0(\omega) \right\|^2 + 2\mu_T(\omega) \left\| \mathrm{E}\, \hat{f}_T^0(\omega) \right\| + \mu_T^2(\omega)$$

It remains to show that $\beta_T^2(\omega)$ and $\delta_T^2(\omega)$ are bounded. As the Pythagorean relationship $\alpha_T^2(\omega) + \beta_T^2(\omega) = \delta_T^2(\omega)$ holds true, it suffices to show that $\beta_T^2(\omega)$ is bounded. Since $\beta_T^2(\omega) = \mathrm{E} \left\| \hat{f}_T^0(\omega) - \mathrm{E}\, \hat{f}_T^0(\omega) \right\|^2$, this follows from theorem 1. This completes the proof.

□

A.3. *Proof of theorem 1*

We make use of Theorem 7.3.2. in [Bri75], according to which

$$\lim_{T \to \infty} \sum_{i,j=1}^{p} \mathrm{Var}\, \hat{f}_T^{0(ij)}(\omega) = \frac{1}{m_T} \left(\mathrm{tr}(g_T(\omega))\right)^2. \tag{A.1}$$

Our assumption 3 implies the assumptions made on the covariance structure of the time series made there. We obtain, using (A.1):

$$\begin{aligned}
\lim_{T \to \infty} \left(\beta_T^2(\omega) - \frac{p_T}{m_T} \mu_T^2(\omega) \right) &= \lim_{T \to \infty} \left(\mathrm{E} \left\| \hat{f}_T^0(\omega) - \mathrm{E}\, \hat{f}_T^0(\omega) \right\|^2 - \frac{p_T}{m_T} \mu_T^2(\omega) \right) \\
&= \lim_{T \to \infty} \left(\frac{1}{p_T} \sum_{i,j=1}^{p_T} \mathrm{Var}\, \hat{f}_T^{0(ij)}(\omega) - \frac{1}{p_T m_T} \left(\mathrm{tr}(\mathrm{E}\, \hat{f}_T^0(\omega) \right)^2 \right) \\
&= \lim_{T \to \infty} \left(\frac{1}{p_T} \frac{1}{m_T} \left(\mathrm{tr}\, g_T(\omega)\right)^2 - \frac{1}{p_T m_T} \left(\mathrm{tr}(\mathrm{E}\, \hat{f}_T^0(\omega) \right)^2 \right) \\
&= \lim_{T \to \infty} \frac{1}{p_T} \frac{1}{m_T} \left((\mathrm{tr}\, g_T(\omega) + \mathrm{tr}\, \mathrm{E}\, \hat{f}_T^0(\omega))(\mathrm{tr}\, g_T(\omega) - \mathrm{tr}\, \mathrm{E}\, \hat{f}_T^0(\omega)) \right) \\
&= \lim_{T \to \infty} \frac{1}{p_T} \frac{1}{m_T} \left((\mathrm{tr}\, g_T(\omega) + \mathrm{tr}\, \mathrm{E}\, \hat{f}_T^0(\omega))\, \mathrm{O}\left(\frac{m_T p_T}{T} \right) \right) = 0
\end{aligned}$$

□

A.4. *Proof of lemma 2*

We first show that $\mathrm{E}(\hat{\mu}_T(\omega) - \mu_T(\omega))^4 \to 0$: Using the ys defined in (9), we can expand the quartic mean as

$$
\begin{aligned}
\mathrm{E}(\hat{\mu}_T(\omega)-\mu_T(\omega))^4 &= \mathrm{E}\left(\frac{1}{m_T}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\frac{1}{p_T}\sum_{i=1}^{p_T}(|y_i(\tilde{\omega}_k)|^2-\mathrm{E}\,|y_i(\tilde{\omega}_k)|^2)\right)^4 \\
&= \frac{1}{m_T^4}\sum_{k_1}\sum_{k_2}\sum_{k_3}\sum_{k_4}\mathrm{E}\left[\left(\frac{1}{p_T}\sum_{i=1}^{p_T}(|y_i(\tilde{\omega}_{k_1})|^2-\mathrm{E}\,|y_i(\tilde{\omega}_{k_1})|^2)\right)\right. \qquad \text{(A.2)}\\
&\times\left.\left(\frac{1}{p_T}\sum_{i=1}^{p_T}(|y_i(\tilde{\omega}_{k_2})|^2-\mathrm{E}\,|y_i(\tilde{\omega}_{k_2})|^2)\right)\times\ldots\times\left(\frac{1}{p_T}\sum_{i=1}^{p_T}(|y_i(\tilde{\omega}_{k_4})|^2-\mathrm{E}\,|y_i(\tilde{\omega}_{k_4})|^2)\right)\right]
\end{aligned}
$$

Here, we must distinguish two cases: The first is that all the $k_{(\cdot)}$ are distinct. In this case, we use Lemmata 10 and 4 to obtain that (A.2) is $\mathrm{O}\left(\frac{p_T^4}{T^2}\right)$, which is sufficient due to assumptions 1 and 4. The second case is that two of the $k_{(\cdot)}$ are equal. There are six symmetric conditions, and making extensive use of the Cauchy-Schwarz inequality and Assumption 2, and abbreviating

$$
D_{k_{(\cdot)}} := \frac{1}{p_T}\sum_{i=1}^{p_T}(|y_i(\tilde{\omega}_{k_{(\cdot)}})|^2-\mathrm{E}\,|y_i(\tilde{\omega}_{k_{(\cdot)}})|^2),
$$

we obtain:

$$
\begin{aligned}
\mathrm{E}(\hat{\mu}_T(\omega)-\mu_T(\omega))^4 &\le \frac{6}{m_T^4}\sum_{k_1}\sum_{k_3}\sum_{k_4}\left|\mathrm{E}\,D_{k_1}^2 D_{k_3} D_{k_4}\right| \le \frac{6}{m_T^4}\sum_{k_1}\sum_{k_3}\sum_{k_4}\sqrt{\mathrm{E}\,D_{k_1}^4}\,\sqrt[4]{\mathrm{E}\,D_{k_3}^4}\,\sqrt[4]{\mathrm{E}\,D_{k_4}^4} \\
&\le \frac{6}{m_T}\,\mathrm{E}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}(|y_i(\tilde{\omega}_m)|^2-\mathrm{E}\,|y_i(\tilde{\omega}_m)|^2)\right)^4,
\end{aligned}
$$

where $\tilde{\omega}_m$ denotes the $\tilde{\omega}_{(\cdot)}$ for which the fourth moment above is maximal. Using a binomial expansion, we proceed:

$$
\begin{aligned}
&\frac{6}{m_T}\,\mathrm{E}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}(|y_i(\tilde{\omega}_m)|^2-\mathrm{E}\,|y_i(\tilde{\omega}_m)|^2)\right)^4 \\
&= \frac{6}{m_T}\sum_{q=0}^{4}(-1)^q\binom{4}{q}\,\mathrm{E}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_m)|^2\right)^q\mathrm{E}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_m)|^2\right)^{4-q} \\
&\le \frac{6}{m_T}\sum_{q=0}^{4}\binom{4}{q}\,\mathrm{E}\left(\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_m)|^2\right)^4\right)^{q/4}\mathrm{E}\left(\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_m)|^2\right)^4\right)^{(4-q)/4} \\
&\le \frac{96}{m_T}\,\mathrm{E}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_m)|^2\right)^4 \le \frac{96K_2}{m_T},
\end{aligned}
$$

which completes the quartic mean convergence of $\hat{\mu}_T(\omega)$. We proceed with the quadratic convergence of $\hat{\delta}_T^2(\omega)$. First, we decompose the difference:

$$
\begin{aligned}
\hat{\delta}_T^2(\omega)-\delta_T^2(\omega) &= \left(\left\|\hat{f}_T^0(\omega)-\hat{\mu}_T(\omega)\,\mathrm{Id}\right\|^2-\left\|\hat{f}_T^0(\omega)-\mu_T(\omega)\,\mathrm{Id}\right\|^2\right) \qquad \text{(A.3)}\\
&+\left(\left\|\hat{f}_T^0(\omega)-\mu_T(\omega)\,\mathrm{Id}\right\|^2-\mathrm{E}\left\|\hat{f}_T^0(\omega)-\mu_T(\omega)\,\mathrm{Id}\right\|^2\right)
\end{aligned}
$$

It suffices to show that both summands on the right side of (A.3) converge to zero in quadratic mean. The first summand does because of the quartic mean convergence of $\hat{\mu}_T(\omega)$:

$$\left\|\hat{f}_T^0(\omega)-\hat{\mu}_T(\omega)\,\mathrm{Id}\right\|^2-\left\|\hat{f}_T^0(\omega)-\mu_T(\omega)\,\mathrm{Id}\right\|^2=\left\|\mu_T(\omega)\,\mathrm{Id}-\hat{\mu}_T(\omega)\,\mathrm{Id}\right\|^2=(\mu_T(\omega)-\hat{\mu}_T(\omega))^2$$

The second summand on the right side of (A.3) consists of a deterministic part and a stochastic part. It suffices to treat the stochastic part:

$$\left\|\hat{f}_T^0(\omega)-\mu_T(\omega)\,\mathrm{Id}\right\|^2=\mu_T^2(\omega)-2\mu_T(\omega)\hat{\mu}_T(\omega)+\left\|\hat{f}_T^0(\omega)\right\|^2 \tag{A.4}$$

Again, we treat the summands one by one; only $\left\|\hat{f}_T^0(\omega)\right\|^2$ needs further effort:

$$\begin{aligned}\left\|\hat{f}_T^0(\omega)\right\|^2&=\frac{1}{p_T}\sum_{i,j=1}^{p_T}\left|\frac{1}{m_T}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}y_i(\tilde{\omega}_k)y_j^*(\tilde{\omega}_k)\right|^2\\&=\underbrace{\frac{p_T}{m_T^2}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_k)|^2\right)^2}_{I}+\underbrace{\frac{p_T}{m_T^2}\sum_{\substack{k,l=-(m_T-1)/2\\k\neq l}}^{(m_T-1)/2}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_k)y_i(\tilde{\omega}_l)|\right)^2}_{II}\end{aligned}$$

We show that the variance of both I and II vanishes asymptotically, where we make essentially use of assumptions 1 and 2 and the asymptotic behavior of cumulants of $y(\cdot)$'s at different Fourier frequencies. We start with I:

$$\begin{aligned}\mathrm{Var}(\mathrm{I})&=\frac{p_T^2}{m_T^4}\,\mathrm{Var}\left(\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_k)|^2\right)^2\right)\\&=\underbrace{\frac{p_T^2}{m_T^4}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\mathrm{Var}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_k)|^2\right)^2}_{\mathrm{Ia}}\\&\quad+\underbrace{\frac{p_T^2}{m_T^4}\sum_{\substack{k,l=-(m_T-1)/2\\k\neq l}}^{(m_T-1)/2}\mathrm{Cov}\left(\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_k)|^2\right)^2,\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_l)|^2\right)^2\right)}_{\mathrm{Ib}}\end{aligned}$$

First, we treat Ia. The order of magnitude of Ia does not change if we multiply by m_T, leave the sum over k out and replace $\tilde{\omega}_k$ by ω, which leads to:

$$\begin{aligned}\mathrm{Ia}&\approx\frac{p_T^2}{m_T^3}\,\mathrm{Var}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\omega)|^2\right)^2\leq\frac{p_T^2}{m_T^3}\,\mathrm{E}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\omega)|^2\right)^4\leq\frac{1}{m_T}\left(\frac{p_T}{m_T}\right)^2\left(\frac{1}{p_T}\sum_{i=1}^{p_T}\mathrm{E}\,|y_i(\omega)|^8\right)\\&\leq\frac{K_1^2K_2}{m_T}\to 0\end{aligned}$$

where we have used assumptions 1 and 2. To treat Ib, we first expand:

$$\begin{aligned}&\mathrm{Cov}\left(\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_k)|^2\right)^2,\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_l)|^2\right)^2\right)=\frac{1}{p_T^4}\,\mathrm{Cov}\left(\left(\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_k)|^2\right)^2,\left(\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_l)|^2\right)^2\right)\\&=\frac{1}{p_T^4}\sum_{i_1,i_2,i_3,i_4=1}^{p_T}\mathrm{Cov}\left(|y_{i_1}(\tilde{\omega}_k)|^2|y_{i_2}(\tilde{\omega}_k)|^2,|y_{i_3}(\tilde{\omega}_l)|^2|y_{i_4}(\tilde{\omega}_l)|^2\right)\end{aligned} \tag{A.5}$$

Now we use Lemma 9. We must distinguish three cases: firstly, if the covariance in (A.5) is decomposed into a product of covariances of y's which are distinct we have a product of four times $\mathrm{O}\left(\frac{p_T}{T}\right)$, thus $\mathrm{O}\left(\frac{p_T^4}{T^4}\right)$. Secondly, one pair of the y's may match, and thirdly and worst, there may be two matches. Still, in this worst case (A.5) is decomposed into terms of the kind

$$\frac{1}{p_T^4}\sum_{i_1,i_2,i_3,i_4=1}^{p_T}\operatorname{Var}(y_{i_1}(\tilde{\omega}_k))\operatorname{Var}(y_{i_3}(\tilde{\omega}_l))\left(\operatorname{Cov}(y_{i_2}(\tilde{\omega}_k),y_{i_4}(\tilde{\omega}_l))\right)^2. \tag{A.6}$$

Now, using Assumption 2, we have that (A.6) is $\mathrm{O}\left(\frac{p_T^2}{T^2}\right)$, which is also the convergence rate of the whole term Ib.
It remains to be shown that the variance of II converges to zero, too. We will skip this part of the proof as it is similar to the proof in [LW04] on pp 395-399, the difference being that in our case there are many cross-terms that must be shown to converge to zero separately, which is done with the help of lemmata 6, 9 and 10 as demonstrated for I.
We proceed with the mean square convergence of $\hat{\beta}_T^2(\omega)$. First, we look at the unconstrained estimator $\bar{\beta}_T^2(\omega)$, which will, like in the proofs before, successively be decomposed into terms that are easier to study:

$$\begin{aligned}
&\bar{\beta}_T^2(\omega)-\beta_T^2(\omega)\\
&=\left[\underbrace{\frac{1}{m_T^2}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\left\|I_T(\tilde{\omega}_k)-\mathrm{E}\,\hat{f}_T^0(\omega)\right\|^2}_{Ia}-\underbrace{\mathrm{E}\left\|\hat{f}_T^0(\omega)-\mathrm{E}\,\hat{f}_T^0(\omega)\right\|^2}_{Ib}\right]\\
&\quad+\underbrace{\left[\frac{1}{m_T^2}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\left\|I_T(\tilde{\omega}_k)-\hat{f}_T^0(\omega)\right\|^2-\frac{1}{m_T^2}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\left\|I_T(\tilde{\omega}_k)-\mathrm{E}\,\hat{f}_T^0(\omega)\right\|^2\right]}_{II}
\end{aligned}$$

Now we have to show that both terms I and II converge to zero in quadratic mean. We start with I. First we show that the asymptotic expectation of I is zero. The expectation of Ia is:

$$\begin{aligned}
\mathrm{E}\,\frac{1}{m_T^2}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\left\|I_T(\tilde{\omega}_k)-\mathrm{E}\,\hat{f}_T^0(\omega)\right\|^2&=\frac{1}{p_T}\sum_{i,j=1}^{p_T}\frac{1}{m_T^2}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\left|I_{ij}(\tilde{\omega}_k)-\sum_{l=-(m_T-1)/2}^{(m_T-1)/2}\mathrm{E}\,I_{ij}(\tilde{\omega}_l)\right|^2\\
&=\frac{1}{p_T}\sum_{i,j=1}^{p_T}\frac{1}{m_T^2}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\left|I_{ij}(\tilde{\omega}_k)-\mathrm{E}\,I_{ij}(\tilde{\omega}_k)+\mathrm{O}\left(\frac{1}{m_T}\right)\right|^2\\
&\approx\frac{1}{p_T}\sum_{i,j=1}^{p_T}\frac{1}{m_T^2}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\left|I_{ij}(\tilde{\omega}_k)-\mathrm{E}\,I_{ij}(\tilde{\omega}_k)\right|^2\\
&=\frac{1}{p_T}\sum_{i,j=1}^{p_T}\frac{1}{m_T^2}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\operatorname{Var}I_{ij}(\tilde{\omega}_k)
\end{aligned}$$

Ib can be expanded as

$$
\begin{aligned}
\mathrm{E}\left\|\hat{f}_T^0(\omega)-\mathrm{E}\,\hat{f}_T^0(\omega)\right\|^2 &= \frac{1}{p_T}\sum_{i,j=1}^{p_T}\mathrm{Var}\left|\frac{1}{m_T}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2} I_{ij}(\tilde{\omega}_k)\right|^2 \\
&= \frac{1}{p_T}\sum_{i,j=1}^{p_T}\frac{1}{m_T^2}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\mathrm{Var}\, I_{ij}(\tilde{\omega}_k) + \mathrm{O}\left(\frac{p_T}{T}\right)
\end{aligned}
$$

Thus, the expectation of I is asymptotically zero. We now show that the variance of I vanishes asymptotically, too. As $\mathrm{E}\left\|\hat{f}_T^0(\omega)-\mathrm{E}\,\hat{f}_T^0(\omega)\right\|^2$ is deterministic, we restrict ourselves to investigating the variance of Ia. As before, contributions in the covariance are shown to be of smaller order, the derivation of which will be skipped here:

$$
\begin{aligned}
&\mathrm{Var}\left(\frac{1}{m_T^2}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\left\|I_T(\tilde{\omega}_k)-\mathrm{E}\,\hat{f}_T^0(\omega)\right\|^2\right) = \frac{1}{m_T^4}\mathrm{Var}\left(\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\left\|I_T(\tilde{\omega}_k)-\mathrm{E}\,\hat{f}_T^0(\omega)\right\|^2\right) \\
&\approx \frac{1}{m_T^4}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\mathrm{Var}\left\|I_T(\tilde{\omega}_k)-\mathrm{E}\,\hat{f}_T^0(\omega)\right\|^2 \approx \frac{1}{m_T}\mathrm{Var}\left(\frac{1}{m_T}\left\|I_T(\omega)-\mathrm{E}\,\hat{f}_T^0(\omega)\right\|^2\right) \\
&= \frac{1}{m_T}\mathrm{Var}\left(\frac{1}{m_T}\left\|y_T(\omega)y_T^*(\omega)-\mathrm{E}\,y_T(\omega)y_T^*(\omega)+\mathrm{O}\left(\frac{1}{m_T^2}\right)\right\|^2\right) \\
&\approx \frac{1}{m_T}\mathrm{Var}\left(\frac{1}{m_T}\left\|y_T(\omega)y_T^*(\omega)-\mathrm{E}\,y_T(\omega)y_T^*(\omega)\right\|^2\right) \\
&= \frac{1}{m_T^3}\frac{1}{p_T^2}\sum_{i_1,i_2,i_3,i_4=1}^{p_T}\mathrm{Cov}\left(y_{i1}y_{i2}^*, y_{i3}y_{i4}^*\right) = \frac{p_T^2}{m_T^3}\mathrm{O}\left(\frac{p_T^2}{T^2}\right)
\end{aligned}
$$

where we have used lemma 6 and 8. We can now move on to the term II. Some elementary transformations result in $\mathrm{II} = \frac{1}{m_T}\left\|\hat{f}_T^0(\omega)-\mathrm{E}\,\hat{f}_T^0(\omega)\right\|^2$. To show mean square convergence of II, it is sufficient to show that its second moment vanishes:

$$
\begin{aligned}
\mathrm{E}\left(\frac{1}{m_T}\left\|\hat{f}_T^0(\omega)-\mathrm{E}\,\hat{f}_T^0(\omega)\right\|^2\right)^2 &\le \frac{1}{m_T^2}\mathrm{E}\left(\left\|\hat{f}_T^0(\omega)-\mu_T(\omega)\,\mathrm{Id}\right\|^2+\left\|\mu_T(\omega)\,\mathrm{Id}-\mathrm{E}\,\hat{f}_T^0(\omega)\right\|^2\right)^2 \\
&\le \frac{2}{m_T^2}\left(\mathrm{E}\,\|\hat{f}_T^0(\omega)-\mu_T(\omega)\,\mathrm{Id}\,\|^4+\|\mu_T(\omega)\,\mathrm{Id}-\mathrm{E}\,\hat{f}_T^0(\omega)\|^4\right)
\end{aligned}
$$

Now, $\mathrm{E}\,\|\hat{f}_T^0(\omega)-\mu_T(\omega)\,\mathrm{Id}\,\|^4$ is bounded, which we have shown above, see (A.4), and $\|\,\mathrm{E}\,\hat{f}_T^0(\omega)-\mu_T(\omega)\,\mathrm{Id}\,\|^4 = \delta^4(\omega)$ is bounded, too, so II converges to zero in quadratic mean. We have thus shown that the unconstrained estimator $\bar{\beta}_T^2(\omega)$ converges to $\beta_T^2(\omega)$ in quadratic mean. Elementary calculations show that this is the case, too, for the constrained estimator $\hat{\beta}_T^2(\omega)$.
We finally remark that mean square convergence of $\hat{\alpha}_T^2(\omega)$ follows trivially from that of $\hat{\beta}_T^2(\omega)$ and $\hat{\delta}_T^2(\omega)$, which completes the proof.

□

The following lemma will be needed for the proof of theorem 3:
Lemma 3 *If u_T^2 is a sequence of nonnegative random variables whose expectations converge to zero, τ_1, τ_2 are two nonnegative constants, and $\frac{u_T^2}{\hat{\delta}_T^2(\omega)^{\tau_1}\delta_T^2(\omega)^{\tau_2}} \le 2(\hat{\delta}_T^2(\omega)+\delta_T^2(\omega))$ a.s., then*

$$
\mathrm{E}\,\frac{u_T^2}{\hat{\delta}_T^2(\omega)^{\tau_1}\delta_T^2(\omega)^{\tau_2}} \to 0.
$$

The proof is similar to pp 402-404 in [LW04].

A.5. *Proof of theorem 3*

We will first show that $\mathrm{E}\left\|\hat{f}_T^*(\omega)-\hat{\varphi}_T^*(\omega)\right\|^2 \to 0$. We decompose

$$\begin{aligned}
\left\|\hat{f}_T^*(\omega)-\hat{\varphi}_T^*(\omega)\right\|^2 &= \left\|\frac{\beta_T^2(\omega)}{\delta_T^2(\omega)}(\hat{\mu}_T(\omega)-\mu_T(\omega))\,\mathrm{Id}+\left(\frac{\hat{\alpha}_T^2(\omega)}{\hat{\delta}_T^2(\omega)}-\frac{\alpha_T^2(\omega)}{\delta_T^2(\omega)}\right)(\hat{f}_T^0(\omega)-\hat{\mu}_T(\omega)\,\mathrm{Id})\right\|^2 \\
&= \frac{\beta_T^4(\omega)}{\delta_T^4(\omega)}(\hat{\mu}_T(\omega)-\mu_T(\omega))^2+\left(\frac{\hat{\alpha}_T^2(\omega)}{\hat{\delta}_T^2(\omega)}-\frac{\alpha_T^2(\omega)}{\delta_T^2(\omega)}\right)^2\left\|\hat{f}_T^0(\omega)-\hat{\mu}_T(\omega)\,\mathrm{Id}\right\|^2 \\
&\quad +2\frac{\beta_T^2(\omega)}{\delta_T^2(\omega)}(\hat{\mu}_T(\omega)-\mu_T(\omega))\left(\frac{\hat{\alpha}_T^2(\omega)}{\hat{\delta}_T^2(\omega)}-\frac{\alpha_T^2(\omega)}{\delta_T^2(\omega)}\right)\langle\hat{f}_T^0(\omega)-\hat{\mu}_T(\omega)\,\mathrm{Id},\mathrm{Id}\rangle \\
&= \frac{\beta_T^4(\omega)}{\delta_T^4(\omega)}(\hat{\mu}_T(\omega)-\mu_T(\omega))^2+\left(\frac{\hat{\alpha}_T^2(\omega)}{\hat{\delta}_T^2(\omega)}-\frac{\alpha_T^2(\omega)}{\delta_T^2(\omega)}\right)^2\hat{\delta}_T^2(\omega) \\
&\le (\hat{\mu}_T(\omega)-\mu_T(\omega))^2+\frac{(\hat{\alpha}_T^2(\omega)\delta_T^2(\omega)-\alpha_T^2(\omega)\hat{\delta}_T^2(\omega))^2}{\hat{\delta}_T^2(\omega)\delta_T^4(\omega)}
\end{aligned}$$

Now, $(\hat{\mu}_T(\omega)-\mu_T(\omega))^2$ converges to zero in quadratic mean by lemma 2. The expectation of the second term converges to zero by lemma 3 as

$$\frac{(\hat{\alpha}_T^2(\omega)\delta_T^2(\omega)-\alpha_T^2(\omega)\hat{\delta}_T^2(\omega))^2}{\hat{\delta}_T^2(\omega)\delta_T^4(\omega)} \le \hat{\delta}_T^2(\omega) \le 2(\hat{\delta}_T^2(\omega)+\delta_T^2(\omega)) \qquad \text{a.s.}$$

Thus, the first statement of theorem 3 is shown. The second statement follows immediately from the first as

$$\begin{aligned}
\mathrm{E}\left\|\hat{f}_T^*(\omega)-\mathrm{E}\,\hat{f}_T^0(\omega)\right\|^2-\mathrm{E}\left\|\hat{\varphi}_T^*(\omega)-\mathrm{E}\,\hat{f}_T^0(\omega)\right\|^2 &= \mathrm{E}\langle\hat{f}_T^*(\omega)-\hat{\varphi}_T^*(\omega),\hat{f}_T^*(\omega)+\hat{\varphi}_T^*(\omega)-2\,\mathrm{E}\,\hat{f}_T^0(\omega)\rangle \\
&\le \sqrt{\mathrm{E}\left\|\hat{f}_T^*(\omega)-\hat{\varphi}_T^*(\omega)\right\|^2}\sqrt{\mathrm{E}\left\|\hat{f}_T^*(\omega)+\hat{\varphi}_T^*(\omega)-2\,\mathrm{E}\,\hat{f}_T^0(\omega)\right\|^2}
\end{aligned}$$

which completes the proof.

□

Appendix B. Properties of Fourier transforms under Kolmogorov asymptotics

This last section gives properties of functions of discrete Fourier transforms under conventional and Kolmogorov asymptotics. Many results from the first case can be adapted to the latter case, but the convergence rates usually deteriorate. The first lemma is on the change of the convergence rates:

Lemma 4 *Suppose that $R_T=[r_{ij}]_{i,j\in 1,\dots,p_T}$ is a sequence of $p_T\times p_T$ matrices that consists of entries that are uniformly $\mathrm{O}(f(T))$. Let $\tilde{R}_T:=\Gamma_T R_T$, where Γ_T may be any orthonormal $p_T\times p_T$ matrix. Then each of the entries $\tilde{r}_{ij}$ of $\tilde{R}_T$ is $\mathrm{O}(p_T f(T))$.*

Proof.

$$\tilde{r}_{ij}=\sum_{s=1}^{p_T}\bar{\gamma}_{si}\sum_{u=1}^{p_T}r_{su}\gamma_{uj}\le\sum_{s=1}^{p_T}|\gamma_{si}|\sum_{u=1}^{p_T}|r_{su}||\gamma_{uj}| \tag{B.1}$$

Orthonormality of Γ_T implies that $\sum_{i=1}^{p_T}|\gamma_{ij}|^2=\sum_{j=1}^{p_T}|\gamma_{ij}|^2=1$. Using the Cauchy-Schwarz inequality, we obtain $\forall\ i,j$

$$\sum_{i=1}^{p_T} |\gamma_{ij}| \le \sqrt{\sum_{i=1}^{p_T} |\gamma_{ij}|^2} \sqrt{p_T} = \sqrt{p_T}$$
$$\sum_{j=1}^{p_T} |\gamma_{ij}| \le \sqrt{\sum_{j=1}^{p_T} |\gamma_{ij}|^2} \sqrt{p_T} = \sqrt{p_T}$$

Plugging this into (B.1) *completes the proof.*

Lemma 5 *Let Γ be an orthonormal matrix, i.e. $\Gamma\Gamma^* = \Gamma^*\Gamma = \mathrm{Id}$. Then the Hilbert Schmidt scalar product and the Hilbert Schmidt norm of any matrices A, B remain unchanged if multiplied by Γ:*

$$\|A\|^2 = \|\Gamma A\|^2$$

and

$$\langle A, B \rangle = \langle \Gamma A, \Gamma B \rangle$$

Lemma 6 *For centered, possibly complex, jointly normal random variables A, B, C, D, we have*

$$\mathrm{Cov}(AB, CD) = \mathrm{Cov}(A, C)\,\mathrm{Cov}(B, D) + \mathrm{Cov}(A, D)\,\mathrm{Cov}(B, D)$$

and

$$\mathrm{E}\, AB\overline{CD} = (\mathrm{E}\, A\overline{C})(\mathrm{E}\, B\overline{D}) + (\mathrm{E}\, A\overline{D})(\mathrm{E}\, B\overline{C}) + (\mathrm{E}\, AB)\overline{(\mathrm{E}\, CD)}$$

Lemma 7 (Convergence rate of bias of periodogram) *The expected value of the periodogram is componentwise*

$$\mathrm{E}\, I_T^{(ij)}(\omega) - g_T^{(ij)}(\omega) = \mathrm{O}\left(\frac{1}{T}\right). \tag{B.2}$$

For the ys, the convergence rate is

$$\mathrm{E}\, y_i(\omega) y_j^*(\omega) - \lambda_{ij}(\omega) = \mathrm{O}\left(\frac{p_T}{T}\right) \tag{B.3}$$

In the norm, both the bias of the unrotated and the rotated data have the same convergence rate:

$$\|\mathrm{E}\, I_T(\omega) - g_T(\omega)\|^2 = \frac{1}{p_T} \sum_{i,j=1}^{p_T} \left(\mathrm{O}\left(\frac{1}{T}\right)\right)^2 = \mathrm{O}\left(\frac{p_T}{T^2}\right) \tag{B.4}$$

and

$$\|\mathrm{E}\, y_T(\omega) y_T^*(\omega) - \Lambda(\omega)\|^2 = \mathrm{O}\left(\frac{p_T}{T^2}\right) \tag{B.5}$$

In all cases, the remainder term is uniform in ω.
Proof. *This follows from theorem 5.2.8. in [Bri75] and lemmata 4 and 5.*

Lemma 8 (Asymptotic properties of the ys) *If $i \ne j$, we have*

$$\mathrm{E}\, y_i(\omega) y_j^*(\omega) = \mathrm{O}\left(\frac{p_T}{T}\right). \tag{B.6}$$

If $\tilde{\omega}_k \ne \pm\tilde{\omega}_l \mod 2\pi$, we have

$$\mathrm{E}\, y_i(\tilde{\omega}_k) y_j^*(\tilde{\omega}_l) = \mathrm{O}\left(\frac{p_T}{T}\right). \tag{B.7}$$

Proof. (B.6) *follows directly from* (B.3) *and the fact that the matrix $\Lambda(\omega)$ is diagonal.* (B.7) *follows from p239ff in [SS00] and lemma 4.*

Lemma 9 *Let $y_{ab}, a \in \{1,2\},\ b \in \{1,\dots,4\}$, be any of the rotated Fourier transforms at any frequency $\omega \in (0, 2\pi)$. Define*

$$Y_1 := y_{11}y_{12}y_{13}y_{14}$$
$$Y_2 := y_{21}y_{22}y_{23}y_{24}$$

Then

$$\operatorname{Cov}(Y_1, Y_2) = \sum_{\nu} \prod_{i=1}^{4} \operatorname{Cov}(y_{c_i d_i}, y_{c'_i, d'_i})$$

where the summation is over all possible partitions ν of ab into subsets of size two such that in each partition at most two of the c_i are identical to the c'_i.

Proof. *This is a special case of Theorem 2.3.2 of [Bri75], where we use that the ys are centered and (complex) Gaussian, due to which the cumulants of order 1 and those of order ≥ 3 are zero.*

Lemma 10 *Let $\lambda_1, \dots, \lambda_4 \neq 0$ be Fourier frequencies such that, for $i \neq j$,*

$$\lambda_i \pm \lambda_j \neq 0 \mod 2\pi. \tag{B.8}$$

Let $d_T(\omega)$ be the discrete Fourier transform of a univariate Gaussian time series $X_1, \dots, X_T$, and $I_T(\omega) = d_T(\omega)\overline{d_T(\omega)}$ the periodogram. Then

$$\operatorname{cum}(I_T(\lambda_1), I_T(\lambda_2), I_T(\lambda_3), I_T(\lambda_4)) = \mathrm{O}\left(\frac{1}{T^4}\right) \tag{B.9}$$

and

$$\mathrm{E}\left(\dot{I}_T(\lambda_1)\dot{I}_T(\lambda_2)\dot{I}_T(\lambda_3)\dot{I}_T(\lambda_4)\right) = \mathrm{O}\left(\frac{1}{T^2}\right), \tag{B.10}$$

where $\dot{I}_T(\lambda) := I_T(\lambda) - \mathrm{E}\, I_T(\lambda)$.

References

[BD87] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods.* Springer, 1987.

[BD98] R. Beran and L. Dümbgen. Modulation of estimators and confidence sets. *The Annals of Statistics*, 26(5):1826–1856, 1998.

[BN02] Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, January 2002.

[Bri75] David R. Brillinger. *Time Series. Data Analysis and Theory.* Holt, Reinhart and Winston, 1975.

[Dah96a] Rainer Dahlhaus. Asymptotic statistical inference for nonstationary processes with evolutionary spectra. In P.M. Robinson and M. Rosenblatt, editors, *Athens Conference on Applied Probability and Time Series Analysis*, volume 2. Springer Verlag, 1996.

[Dah96b] Rainer Dahlhaus. On the Kullback-Leibler information divergence for locally stationary processes. *Stochastic Processes and their Applications*, 62(2):139–168, 1996.

[Dah00] Rainer Dahlhaus. A likelihood approximation for locally stationary processes. *The Annals of Statistics*, 28(6):1762–1794, December 2000.

[DS85] D.K. Dey and C. Srinivasan. Estimation of a covariance matrix under Stein's loss. *The Annals of Statistics*, 13(4):1581–1591, 1985.

[DSH02] Z.A. Der, R.H. Shumway, and E.T. Herrin. *Monitoring the Comprehensive Nuclear-Test-Ban Treaty: Data Processing and Infrasound.* Birkhauser, 2002.

[FHLR00] M. Forni, M. Hallin, M. Lippi, and L. Reichlin. The generalized dynamic-factor model: Identification and estimation. *The Review of Economics and Statistics*, 82(4):540–554, November 2000.

[GOvS05] W. Guo, Hernando C. Ombao, and Rainer von Sachs. SLEX analysis of multivariate nonstationary time series. *Journal of the American Statistical Association*, 100(470):519–531, June 2005.

[Haf79] L.R. Haff. Estimation of the inverse covariance matrix: Random mixtures of the inverse wishart matrix and the identity. *The Annals of Statistics*, 7(6):1264–1276, 1979.

[Haf80] L.R. Haff. Empirical bayes estimation of the multivariate normal covariance matrix. *The Annals of Statistics*, 8:586–597, 1980.

[Jol02] I.T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[KL52] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1952.

[KST98] Y. Kakizawa, R.H. Shumway, and M. Taniguchi. Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association*, 93(441):328–340, 1998.

[Lee97] T.C.M. Lee. A simple span selector for periodogram smoothing. *Biometrika*, 84:965–969, 1997.

[LW04] O. Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411, 2004.

[MHvS06] G. Motta, C. Hafner, and R. von Sachs. Locally stationary factor models: identification and nonparametric estimation. Discussion paper 0624, Institut de Statistique, Université Catholique de Louvain, 2006.

[OYR86] F. O'Sullivan, B. Yandell, and W. Raynor. Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association*, 81:96–103, 1986.

[Pri81] M. Priestley. *Spectral Analysis and Time Series*. London: Academic Press, 1981.

[SS00] Robert H Shumway and David S. Stoffer. *Time Series Analysis and its Applications*. Springer, 2000.

[Ste75] C. Stein. Rietz lecture. 39th annual meeting IMS, 1975.