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

SHRINKAGE METHODS FOR MULTIVARIATE SPECTRAL ANALYSIS

Membres du jury:

Prof. Christian Hafner *(Président du jury)*
Prof. Jens-Peter Kreiß
Prof. Hernando Ombao
Prof. Rainer von Sachs *(Promoteur)*
Prof. Léopold Simar
Prof. Sébastien Van Bellegem

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

Louvain-la-Neuve
Défense du 29 janvier 2008

With love, I dedicate this work to Grit, who let me go far away but always stood by my side. Thank you!

Acknowledgements

I would like to thank all the people and institutions who have helped me write this thesis:
My Ph.D. advisor, Rainer von Sachs, has supported me with enormous energy and zest. He has spared incredibly much time for countless scientific discussions that were always inspiring and mostly fruitful. All the time, he has been spreading enthusiasm and optimism.
The Institute of Statistics has been the most inspiring, challenging and promoting working environment I could possibly have chosen for my thesis. The graduate school in statistics and actuarial sciences has allowed me to become familiar with a broad variety of up-to-date statistical techniques that turned out to be a great asset, both for this work and my further career.
Hernando Ombao, who has travelled a very long way to attend my défense, has helped me understand the details of non-stationary time series analysis. Christian Hafner, Léopold Simar and Sébastien Van Bellegem have given many a useful comment for my research. Jens-Peter Kreiß, who has travelled a long way for my prédéfense, has asked helpful, in-depth questions.
Last but not least, my friends from Louvain-la-Neuve have helped me with encouragement and diversion through the long and sometimes frustrating process of writing a Ph.D. thesis.

*

This research was supported by the "Fonds Spéciaux de Recherche" from the Université catholique de Louvain and by the IAP research network No. P5/24 of the Belgian State (Federal Office for Scientific, Technical and Cultural Affairs).

Contents

List of Symbols

$\mathcal{N}^C$	Complex normal distribution	50
m_T	Smoothing span	8
p	Dimension of data (classical asymptotic framework)	7
p_T	Dimension of data (general asymptotics)	31
$\mathcal{R}$	Risk function	27

List of Figures

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, about 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, if the length of the averaging window grows with the sample size. This is treated extensively in standard books on time series analysis, such as the books by Brillinger (1975), Priestley (1981), Brockwell & Davis (1987), Shumway & Stoffer (2000) or Kreiß & Neuhaus (2006). In a quite straightforward way, most existing smoothing methods can be generalized to multivariate time series.

This work is concerned with improving upon the averaged periodogram as an estimator for the multivariate spectrum. The methods we will apply are commonly referred to as regularization or shrinkage techniques. This is because one of the main effects of the improvements they offer concerns the numerical stability of the estimator that will be derived; however, we will also see that the shrinkage estimators are, at the same time, more precise in the mean-squared error sense.

Estimation of the spectrum 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 condition number c of a quadratic matrix is the ratio between its largest and its smallest eigenvalue. It is a measure of numerical stability; a high condition number means that, when inverting the matrix, the results may be imprecise.

The classical estimator at frequency ω is obtained by averaging the $p \times p$ periodogram matrices at the m Fourier frequencies nearest to ω; each of the m periodogram matrices is singular, as the periodogram at frequency ω is the product of the discrete Fourier transform vector at frequency ω with its transpose. The condition number of the averaged periodogram 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 ignore this effect. In many applications, it is severe if the estimator of the spectrum is badly conditioned. For instance, Kakizawa, Shumway & Taniguchi (1998) use the Kullback-Leibler discrimination information (Kullback & Leibler 1952) as a measure of disparity between several estimated multivariate spectra. Computing the Kullback-Leibler discrimination information, however, does involve inverting the estimate of one of the spectra, resulting in possibly high inaccuracy due to a bad condition number of the estimated spectrum. Moreover, poor estimators can lead to unacceptably large rates of misclassification.
In many fields of application, including economic panel data (Bai & Ng 2002, Forni, Hallin, Lippi & Reichlin 2000), but also genetic engineering (Deonier, Tavaré & Waterman 2005) or neuropsychology (Guo, Ombao & von Sachs 2005), the dimension of the data may match or even exceed the sample size; in the latter case, the averaged periodogram is even singular.
Guo et al. (2005) 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 (Jolliffe 2002). Thus, estimation can be improved by shrinking the eigenvalues towards one another.

There is indeed a large literature (e.g. Beran & Dümbgen 1998) 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 er-

ror. 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 (Stein 1975, Haff 1977, 1979, 1980, Dey & Srinivasan 1985, 1986, Kubokawa & Konno 1990, Loh 1991, Evans & Stark 1996, Kubokawa & Srivastava 1999, Sheena 2002), or quite differently, on asymptotic theory. There is less literature following the latter concept, and the few existing publications are more recent. They are partly based on classical asymptotics (Ledoit & Wolf 2003), partly on "double" or *Kolmogorov* asymptotics (Ledoit & Wolf 2004) where simultaneously the sample size T and the dimension p tend to infinity.

In this work, 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 not an optimal solution to the problem. On one hand, using the methods we will develop in the following chapters, even a strongly oversmoothed estimator can still be improved upon in terms of its L_2 risk.
On the other hand we will develop a "double asymptotic" framework in chapter 2, under which the conventional smoothed periodogram is not merely suboptimal, but not even mean square consistent. This will enable us to construct competing estimators that, asymptotically and for finite sample size, have smaller L_2 risk than the averaged periodogram and are, at the same time, numerically more stable. For these shrinkage estimators we follow a linear approach that combines the averaged periodogram $\hat{f}_T^0(\omega)$ at frequency $\omega \in (0, 2\pi]$ with the identity matrix:

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

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 Ledoit & Wolf (2004), 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. The results of chapter 2 can also be found in Böhm & von Sachs (2007).
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 slower than T. In chapter 2, 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 useful 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 (e.g. Forni et al. 2000). In neuropsychology, when analyzing EEG data, not only the observation period can be extended, but also the number of channels that are analyzed may be increased to better capture localized features of the signal once sufficient observations are available (compare Guo et al. 2005). 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 (see Der, Shumway & Herrin 2002).
The double or Kolmogorov asymptotic framework in chapter 2 contains the classical asymptotic framework as a special case. When the dimension p is fixed, the necessity to shrink vanishes asymptotically, resulting in the shrinkage estimator converging to the averaged periodogram. However, for finite sample size, the averaged periodogram is still shrunk. The next chapter 3 demonstrates that it is also possible to use a classical asymptotic framework directly. The results are, technically, easier to obtain, at the price of being less comprehensive. In chapter 3, we also introduce a different shrinkage target. We will shrink to a matrix that is obtained by fitting a one-factor model to the data. This model is, however, assumed not to be true. Rather, it is just a tool to describe the local spectral structure in a very parsimonious way, like the identity matrix in chapter 2. Chapter 3 will allow us to see shrinkage methods from different perspectives. One angle of view is to regard shrinkage as

a way to circumvent the dilemma of model choice in factor analysis. A frequent problem in using factor models is the problem of choice of the *true* number of factors underlying the data. As will be discussed in chapter 3, shrinkage techniques may offer a new solution to the problem, by building an asymptotically optimal linear combination of an underfitted factor model and a nonparametric estimator. Moreover, the concept of shrinkage may also be interpreted as a theoretical approach to merge two different kinds of estimators - a parametric and a nonparametric one - to a new estimator that inherits the good properties of both its 'ancestors'.

Before we move to the construction of the shrinkage estimators, we will give more motivating and explicative theoretical background information in chapter 1. Then, we will construct the two shrinkage estimators in chapters 2 and 3. The following chapter 4 will present a real data example as well as the results of comprehensive simulation studies we have performed to compare the shrinkage estimators with each other and with various benchmark estimators, among them the averaged periodogram, shrinkage estimators obtained under a different theoretical background, and benchmark estimators that have certain optimality properties, but are only available when the true spectrum is known. We will also give simulation results that are based on tapered data. Finally, in chapter 5, we will discuss the theoretical and practical results. The main part of the work is followed by appendices that give the proofs of most of the theorems in chapters 2 and 3, and also a series of technical lemmata needed for these proofs.

CHAPTER 1

Preliminaries

1.1 Multivariate spectral analysis

We assume that we observe a realization $(X_t)_{t=1}^T$ of a p-dimensional real-valued, centered stationary Gaussian time series (X_t). We aim at estimating the $p \times p$ spectral density matrix function at frequency $\omega \in (0, 2\pi]$

$$f(\omega) = \frac{1}{2\pi T} \sum_{u \in \mathbb{Z}} \mathrm{Cov}(X_t, X_{t+u}) \exp(-\iota \omega u), \quad \omega \in (0, 2\pi) \tag{1.1}$$

where $\iota = \sqrt{-1}$. The most common nonparametric estimators of (1.1) are based on the *periodogram*. If we denote by

$$d_T(\omega) = \frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} X_t \exp(-\iota \omega t), \quad \omega \in (0, 2\pi) \tag{1.2}$$

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

$$I_T(\omega) := d_T(\omega) d_T^*(\omega) \tag{1.3}$$

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 (1.1), and it is only asymptotically unbiased. Moreover, for $p > 1$, the periodogram

is a singular matrix: if $d_T(\omega) = (d_1(\omega), \ldots, d_p(\omega))'$, then (1.3) can be expressed as

$$I_T(\omega) = \left(\overline{d_1(\omega)} \begin{pmatrix} d_1(\omega) \\ \vdots \\ d_p(\omega) \end{pmatrix} \quad \ldots \quad \overline{d_p(\omega)} \begin{pmatrix} d_1(\omega) \\ \vdots \\ d_p(\omega) \end{pmatrix} \right) \tag{1.4}$$

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 this work, 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:

$$\hat{f}_T^0(\omega) := \frac{1}{m_T} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} I_T(\omega + \omega_k) \tag{1.5}$$

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

$$f_T^0(\omega) := \mathrm{E}\, \hat{f}_T^0(\omega) \tag{1.6}$$

instead of constructing estimators for the spectrum $f(\omega)$. The expected averaged periodogram (1.6) is a deterministic function of the spectrum. Under assumptions we will specify in section 2.3, the difference is negligible, even under general, i.e. Kolmogorov, asymptotics, see (2.13).

1.2 Limitations of traditional multivariate spectral analysis

In the literature on multivariate spectral analysis, estimates of the spectrum are used in applications such as classification and discrimination, or the partitioning of a non-stationary time series into segments of approximate stationarity. This subsection is aimed at providing insight into these fields of research and into the specific problems that originate from multi-dimensionality therein.

1.2.1 Discrimination and classification

Classification, discrimination and cluster analysis in multivariate spectral analysis is of high practical relevance in seismology. Here, the differences between realizations of vector time series allow to classify events as e.g. an earthquake, a nuclear detonation or a mining explosion (Kakizawa et al. 1998). Monitoring nuclear proliferation depends critically on the ability to discriminate reliably between small nuclear explosions and the other two categories (Der et al. 2002). When large events, such as the detonation of a hydrogen bomb or a major earthquake, are observed, discrimination can be based on the first moments of the observations, even at teleseismic distances (10,000-15,000 km). However, this is not possible for small nuclear explosions, even at regional (100 - 2,000 km) distances. In the latter case, discrimination analysis must be based on estimates of the multivariate spectra. Figures 1.1 and 1.2 show examples of the seismic signal of an earthquake and an explosion at regional distances. There is a large variety of criteria that are used to develop

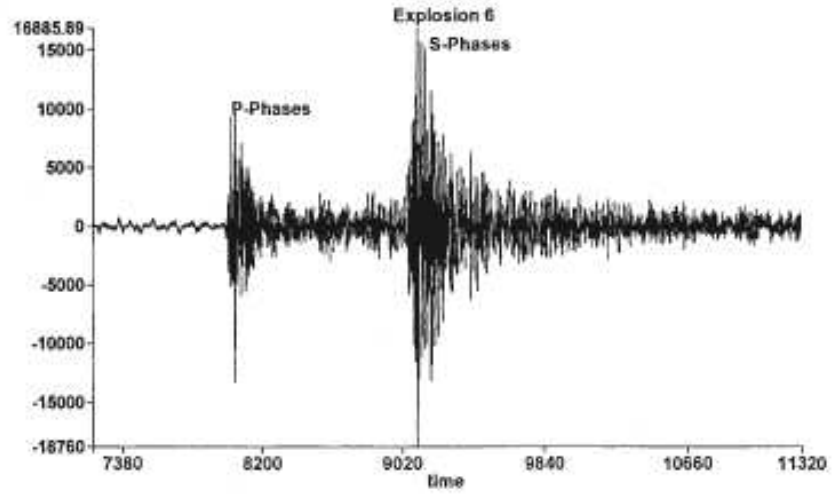

Figure 1.1: Seismic signal of a typical earthquake, recorded from regional distance. From Kakizawa et al. (1998).

discrimination rules on. One property they all have in common is that they use operations that depend on the spectrum being well conditioned. Out of this large number of possible criteria, we will give three for the

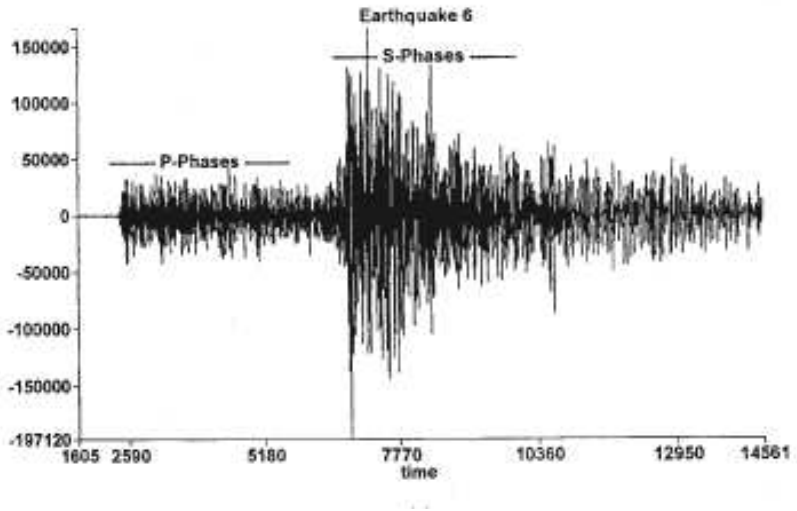

Figure 1.2: Seismic signal of a typical explosion, recorded from regional distance. From Kakizawa et al. (1998).

purpose of demonstration.
One frequently used criterion is the discriminant rule based on Whittle's asymptotic likelihood (compare Taniguchi & Kakizawa 2000). In this approach, the log-likelihood of a p-dimensional realization of a process of length T with $p \times p$ spectral density matrix $f(\omega)$ is approximated by

$$L(X, f) = -\frac{1}{2} \sum_{t=1}^{T} \left(\ln \det(f(\omega_t)) + \operatorname{tr} \left(I_T(\omega_t) f^{-1}(\omega_t) \right) \right), \quad \omega_t = \frac{2\pi t}{T}. \tag{1.7}$$

where 'det' refers to the determinant. A realization X is then classified into a class

$$\Pi_j : \{ f(\omega) = f_j(\omega) \} \tag{1.8}$$

if

$$L(X, f_j) > L(X, f_k) \quad \forall k \neq j.$$

Other approaches are based on the *Kullback-Leibler discrimination information* (Kullback & Leibler 1952). Shumway & Stoffer (2000) derive a discrimination criterion for spectra based on this measure of disparity as

$$I(f_1, f_2) = \frac{1}{T}\sum_{t=1}^{T}\left(\operatorname{tr}\left(f_1(\omega_t)f_2^{-1}(\omega_t)\right) - \ln\frac{\det(f_1(\omega_t))}{\det(f_2(\omega_t))} - p\right) \quad (1.9)$$

for $\omega_t = \frac{2\pi t}{T}$. Kakizawa et al. (1998) propose a large sample spectral approximation to the Chernoff information measure (Chernoff 1952, Renyi 1961), indexed by a parameter $0 < \alpha < 1$, as

$$B_\alpha(f_1, f_2) = \frac{1}{2T}\sum_{t=1}^{T}\left(\ln\frac{\det(\alpha f_1(\omega_t) + (1-\alpha)f_2(\omega_t))}{\det f_2(\omega_t)} - \alpha\ln\frac{\det f_1(\omega_t)}{\det f_2(\omega_t)}\right) \quad (1.10)$$

These three criteria have in common that they require calculating the determinant of a spectral matrix or inverting it. As far as the true spectrum is known, this is not problematic. However, in practice classification is done by replacing the unknown spectra in (1.8) by estimates of spectra from events that are known to be representatives of the respective classes. In this case, computing one of the criteria involves inverting an estimated spectral matrix ((1.7), (1.9)), and/or computing its determinant ((1.7), (1.9) and (1.10)). Both operations depend crucially on the estimate being well-conditioned. Parametric estimators of the spectrum are usually well-conditioned, but when the data are high dimensional, they suffer either from a curse of dimensionality, as too many parameters have to be estimated, or are under-parameterized. On the other hand, the condition number of a non-parametrically estimated spectral matrix is in mean always higher than the condition number of the true spectrum. This is a general result from random matrix theory: if the dimension of the data is not negligible with respect to the sample size, then the sample eigenvalues are always biased estimators of the true eigenvalues - the small ones are biased downwards, the large ones upwards. In consequence, the condition number is amplified. This phenomenon was first examined mathematically by Marčenko & Pastur (1967). A good summary is given in Frahm & Jaekel (2007).
Even for long time series, the true condition number may be overestimated by a large factor. This leads to amplification of estimation error

and, in consequence, misclassification. It would therefore be desirable to develop an estimator that is numerically more stable, even if this were to a certain extent at the price of precision. In chapter 2 we will construct an estimator of the spectrum that is both numerically stable and more precise. In section 4.4, we will compare the performance of this estimator in classification to the performance of the averaged periodogram. Before, we will give another example of how amplification of the condition number affects application in multivariate spectral analysis.

1.2.2 Analysis of non-stationary multivariate time series

For many empirical multivariate time series, the challenge is to address both the problem of non-stationarity and the problem of high dimensionality simultaneously. E.g., in neuropsychology, applications such as the diagnosis of epilepsy are based on electroencephalograms (EEGs). EEGs are electric signals of the cumulative neural activity of the brain that are recorded on multiple locations of the surface of the head simultaneously, and that exhibit a clearly non-stationary behavior. Figure 1.3 shows a realization of an EEG time series. Due to non-stationarity, these signals cannot be analyzed using classical Fourier methods. Guo et al. (2005) propose to use a collection of bases consisting of waveforms that are time-localized and orthogonal generalizations of the Fourier complex exponentials, the SLEX (smooth localized complex exponentials) library (Guo, Ombao, Raz & von Sachs 2002), for highly collinear, multivariate data. This library is used to develop a family of models that can explicitly characterize the time-changing spectral and coherence features of the p-variate time series. The members of this family of models are all possible dyadic partitions of the time axis up to a certain level. E.g., if the maximal level is $J = 1$, this means that the algorithm decides whether a spectral analysis is performed for the time series as a whole or for the first and second half of the data separately. In the latter case, the result is two different spectral densities. If the maximal level is $J = 2$, then the algorithm may divide the first or second half of the time series into segments of one quarter the length of the whole time series. Figure 1.4 shows a dyadic SLEX library with level $J = 2$. In this case, the algorithm has chosen to split the time series into segments (marked yellow) of length $T/2, T/4$ and $T/4$.

Choosing the optimal segmentation is a core part of the problem in this context. The approach the authors follow is based on a complexity-

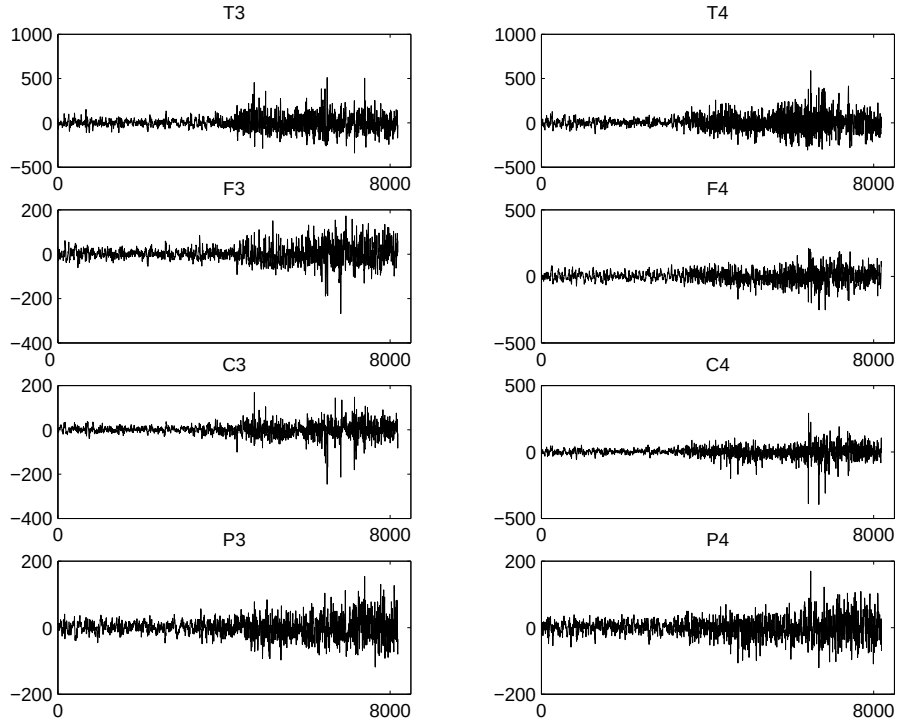

Figure 1.3: Eight channels of an EEG recorded during an epileptic seizure. T=8,192, sampling rate is 100 Hz. The abbreviations 'T', 'F', and 'P' refer to the temporal, frontal and parietal lobe of the brain, 'C' to a central position. The odd numbers mark recordings from its left, the even numbers recordings from its right side. From Guo et al. (2005).

penalized log energy cost function. For all possible dyadic segmentations, the cost is derived as the sum over the cost functions of the time segments the segmentation consists of. Then, the segmentation that minimizes the global cost function is chosen. The cost of the block number b at level j is derived as

$$\text{Cost}(j,b) = \sum_{k=-M_j/2+1}^{M_j/2} \log\det \tilde{I}_{j,b}(\tilde{\omega}_k) + \beta_{j,b}(p)\sqrt{M_j} \tag{1.11}$$

where $M_j = T/2^j$ is the length of a segment of level j, $\tilde{I}_{j,b}(\omega)$ is the smoothed SLEX periodogram matrix at frequency ω, and $\beta_{j,b}(p)$

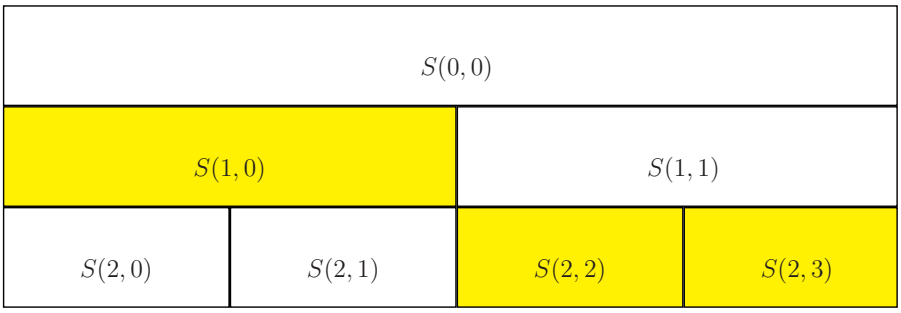

Figure 1.4: A SLEX library with level $J = 2$. The yellow blocks represents one basis from the SLEX library. Adapted from Guo et al. (2005).

is a data driven complexity parameter. Like the averaged periodogram $\hat{f}_T^0(\omega)$, the smoothed SLEX periodogram matrix $\tilde{I}_{j,b}(\omega)$ is an estimator of the spectrum, but it uses SLEX basis functions instead of Fourier basis functions, and is a weighted average instead of a simple average.
The cost function (1.11) suffers from the same drawback due to ill conditioning that has been addressed in the preceding subsection. Like before, this is because computing the cost function involves calculating the determinant of the smoothed SLEX periodogram matrix $\tilde{I}_{j,b}(\tilde{\omega}_k)$. Unlike in many other publications on multivariate spectral analysis, however, the authors have seen this problem and developed two heuristical methods to cope with it. We will sketch both methods here, because these heuristical approaches precede the rigorous methods we will present in chapters 2 and 3.
Both approaches are based on a principal component decomposition of the smoothed SLEX periodogram matrix $\tilde{I}_{j,b}(\omega)$. The determinant of a matrix is invariant to orthogonal rotation, and rotating to the basis spanned by the eigenvectors of a matrix is an orthonormal rotation. Let us denote by $\nu_{j,b}^d(\omega)$ the dth largest eigenvalue of $\tilde{I}_{j,b}(\omega)$. Then the representation of $\tilde{I}_{j,b}(\omega)$ in the basis of eigenvectors corresponding to $\nu_{j,b}^1(\omega), \ldots, \nu_{j,b}^p(\omega)$ is

$$\overline{I}_{j,b}(\omega) = \begin{pmatrix} \nu_{j,b}^1(\omega) & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \nu_{j,b}^p(\omega) \end{pmatrix}$$

Thus, the determinant of both $\tilde{I}_{j,b}(\omega)$ and $\overline{I}_{j,b}(\omega)$ is the product $\nu^1_{j,b}(\omega) \times \ldots \times \nu^p_{j,b}(\omega)$, and equation (1.11) can be written equivalently as

$$\text{Cost}(j,b) = \sum_{k=-M_j/2+1}^{M_j/2} \sum_{d=1}^{p} \log \nu^d_{j,b}(\tilde{\omega}_k) + p\beta_{j,b}\sqrt{M_j} \tag{1.12}$$

Now, the first heuristical approach is to take only the first $q \leq p$ of the eigenvalues $\nu^1_{j,b}(\tilde{\omega}_k), \ldots, \nu^p_{j,b}(\tilde{\omega}_k)$ of $\tilde{I}_{j,b}(\tilde{\omega}_k)$ into consideration and to modify the cost function as

$$\text{Cost}(j,b) = \sum_{k=-M_j/2+1}^{M_j/2} \sum_{d=1}^{q} \log \nu^d_{j,b}(\tilde{\omega}_k) + q\beta_{j,b}\sqrt{M_j} \tag{1.13}$$

The problem here is the choice of q. Principal components analysis is a tool to describe data, and using principal components of reduced dimension is basically a data compression technique. The number of dimensions to take into consideration in the cost function is thus an arbitrary choice.

The second heuristical approach takes into account the numerical instability of the small eigenvalues of the spectrum under highly collinear data by weighing the eigenvalues. Using the weights

$$w^d_{j,b}(\tilde{\omega}_k) = \frac{\nu^d_{j,b}(\tilde{\omega}_k)}{\sum_{c=1}^{p} \nu^c_{j,b}(\tilde{\omega}_k)}, \tag{1.14}$$

the cost is computed as

$$\text{Cost}(j,b) = \sum_{k=-M_j/2+1}^{M_j/2} \sum_{d=1}^{p} w^d_{j,b}(\tilde{\omega}_k) \log \nu^d_{j,b}(\tilde{\omega}_k) + \beta_{j,b}(p)\sqrt{M_j} \tag{1.15}$$

The effect of this procedure is that the influence of the smaller eigenvalues is diminished. However, there is no formal justification for using the weights (1.14). In chapter 2, we will develop a rigorous theoretical framework for this problem.

1.3 Addressing the problems of ill conditioning and collinearity in multivariate analysis

The problems of ill conditioning and collinearity we have described in the context of multivariate spectral analysis in the previous section have,

as far as we know, hardly been addressed in this context. However, in multivariate analysis for iid data, a number of different approaches exist that will be shortly described in this section: Ridge regression, sparse principal component analysis and shrinkage techniques.

1.3.1 Ridge regression

In the context of multiple regression, the 'conventional' approach in estimating the parameters is based on least squares. Consider the standard model

$$Y = X\beta + \epsilon \tag{1.16}$$

where X is a $n \times p$ matrix of rank p of predictors, β is an unknown $p \times 1$ vector, and the idiosyncratic components ϵ fulfill the conditions $\mathrm{E}\,\epsilon = 0$ and $\mathrm{E}\,\epsilon\epsilon' = \sigma^2\,\mathrm{Id}_n$ with unknown σ^2. When the random vector ϵ is Gaussian, it is well known that the best linear unbiased estimator (BLUE) of β is

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{1.17}$$

However, estimation based on (1.17) suffers from a severe drawback: the whole class of unbiased estimators implicitly depends on the predictors being nearly orthogonal. To demonstrate this, we will investigate the quadratic distance L^2 between $\hat{\beta}$ and β:

$$L^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta) \tag{1.18}$$

Its expected value is

$$\mathrm{E}\,L^2 = \sigma^2\,\mathrm{tr}(X'X)^{-1} \tag{1.19}$$

and, when the idiosyncratic components are Gaussian, its variance is

$$\mathrm{Var}\,L^2 = 2\sigma^4\,\mathrm{tr}(X'X)^{-2} \tag{1.20}$$

If the eigenvalues of $X'X$ are denoted by

$$l_{\max} = l_1 \geq \ldots \geq l_p = l_{\min} > 0$$

we can, as the trace of a matrix is invariant under orthogonal transformations, express (1.19) and (1.20) as

$$\mathrm{E}\,L^2 = \sigma^2 \sum_{i=1}^{p} l_i^{-1} \tag{1.21}$$

and

$$\operatorname{Var} L^2 = 2\sigma^4 \sum_{i=1}^{p} l_i^{-2}. \tag{1.22}$$

Equations (1.21) and (1.22) show that in case of non-orthogonality of the predictors X, the error in estimating β is amplified, in case of high collinearity dramatically, as in the latter case eigenvalues of $X'X$ near zero do exist. Moreover, the classical regression setting assumes that the predictors X are non-random. If we drop that assumption and assume w.l.o.g. that $\mathrm{E}\, X = 0$, the matrix $X'X$ is a multiple of the least squares covariance matrix estimator $S = X'X/n$. In this case, another problem may arise, even if the true covariance matrix Σ is well conditioned, from plugging in S for $\Sigma = \mathrm{E}\, X'X$ in equation (1.17). As we will see in chapter 2, the eigenvalues of S are biased estimates of the eigenvalues of Σ, which may result in S being ill conditioned even if Σ is well conditioned. Hoerl & Kennard (1970*a*,*b*) introduced the concept of *ridge regression*, which is a special case of *Tikhonov regularization* (Tikhonov 1963), to compensate for non-orthogonality in the predictors. The idea is to regularize the matrix $X'X$ by adding $c \geq 0$ times the identity matrix, which leads to the ridge estimator

$$\hat{\beta}^r = (X'X + c\,\mathrm{Id})^{-1} X'Y. \tag{1.23}$$

The eigenvalues l_i^W of the matrix $W = (X'X + c\,\mathrm{Id})^{-1}$ are related to the eigenvalues l_i of $X'X$ by

$$l_i^W = \frac{1}{l_i + c} \tag{1.24}$$

which has the effect of shifting the upper bounds $\sigma^2/l_{\min}$ of (1.21) and $2\sigma^4/l_{\min}^2$ of (1.22), respectively, downwards. Of course, for $c > 0$, the estimator $\hat{\beta}^r$ is no longer unbiased as the unbiasedness is traded for a lower risk. Hoerl and Kennard, however, derive no optimal solution for this bias-variance tradeoff. The concept of ridge regression can be further refined by replacing $c \times \mathrm{Id}$ in (1.23) by a diagonal matrix with different elements:

$$\hat{\beta}^R = (X'X + \operatorname{diag}(c_1, \ldots, c_p))^{-1} X'Y. \tag{1.25}$$

This is referred to as *generalized ridge regression*.
Solving (1.23) is equivalent to solving the minimization problem

$$\hat{\beta}^r = \arg\min_{\tilde{\beta}} \left\| Y - X\tilde{\beta} \right\|^2 + c \left\| \tilde{\beta} \right\|^2 \tag{1.26}$$

for given $c > 0$. In the following subsection, we will use (1.26) to link the theory of ridge regression to principal component analysis and refinements thereof.

1.3.2 Sparse PCA

Principle component analysis, although widely used in data processing and dimensionality reduction, has the flaw of producing results that are difficult to interpret. This is because each component is a linear combination of *all* underlying variables. To overcome this, practitioners often perform a secondary rotation after selecting a limited number of principle components (Jolliffe 2002) to obtain a 'simple structure'. Recent research (James & Sugar 2003, Johnstone & Lu 2004, Zou, Hastie & Tibshirani 2004) has lead to the development of a new method called *sparse principal component analysis* (SPCA) that produces principal components with sparse loadings. 'Sparse' in this context means that there are no small loadings, as one frequently finds in abundance after performing classical PCA. Rather, the method is designed to produce loadings that are either high or exactly zero. This renders the method a useful tool, as the results are easy to interpret. It is widely applied in areas such as genetic engineering, or in portfolio optimization, where it is used to find near-optimal investment strategies that require buying a smaller number of assets, minimizing transaction costs.
Ridge regression is directly linked to classical PCA via the minimization problem (1.26). Consider a centered $n \times p$ matrix X of n iid observations of p variables. Denote by

$$X = UDV' \tag{1.27}$$

the singular value decomposition of X with the entries of the diagonal matrix D ordered from largest to smallest. The singular value decomposition decomposes the $n \times p$ matrix X into a unitary $n \times n$ matrix Y, a $n \times p$ matrix D with zero off-diagonal entries and a unitary $p \times p$ matrix V. $Y = UD$ are the principal components (PC) of unit length, the ith

principal component is denoted by $Y_i = U_i D_{ii}$, and the columns V_i of V are the corresponding loadings of the PCs. Then V_i is given by

$$V_i = \frac{\hat{\beta}^r}{\left\|\hat{\beta}^r\right\|} \tag{1.28}$$

if ridge regression is performed like in the previous subsection, with the Y_is in place of Y:

$$\hat{\beta}^r = \arg\min_{\tilde{\beta}} \left\|Y_i - X\tilde{\beta}\right\|^2 + c\left\|\tilde{\beta}\right\|^2 \tag{1.29}$$

for $c > 0$. When $n > p$ and X has full rank, the ridge penalty parameter $c\|\beta\|^2$ is not needed; in any other case, any choice of $c > 0$ will yield the ith principal component, due to the renormalization (1.28). A proof of this can be found in Zou et al. (2004).
An extension of ridge regression is the *elastic net* (Hastie, Tibshirani & Friedman 2001), which adds an L_1 penalty, the so-called *lasso*, to the ridge penalty. Consider the L_1 matrix (vector) norm

$$\|A\|_1 = \sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}| \tag{1.30}$$

of an $m \times n$ ($m \times 1$) matrix (vector) A. Adding the lasso penalty to (1.29), we obtain

$$\hat{\beta}^e = \arg\min_{\tilde{\beta}} \left\|Y_i - X\tilde{\beta}\right\|^2 + c\left\|\tilde{\beta}\right\|^2 + d\left\|\tilde{\beta}\right\|_1 \tag{1.31}$$

The L_1 penalty has the effect of a nonlinear threshold; increasing the lasso penalty parameter d results in a continuous subset selection. This can be used to perform a linear regression with sparse weights, or, likewise, to obtain sparse principal components. In the latter case, it is important to remark that equations (1.29) and (1.31) are not self-contained in the sense that they require the principal components Y_i to be known. However, it is possible to derive self-contained regression type criteria that allow to obtain directly sparse principal components and loadings simultaneously. We will not present these criteria here, as they do not contribute to a deeper understanding of sparse PCA. Instead, we refer

the reader to Zou et al. (2004).
While sparse PCA is extremely useful in data presentation and interpretation, and has many useful applications in areas such as image processing or genetic engineering, it contributes little to the problems of classification or choice of best basis for non-stationary time series.

1.3.3 Shrinkage estimators of the covariance matrix

The multivariate spectrum is a sophisticated tool for describing the interdependence structure of non-iid data. Estimating the spectrum is the analogue in the frequency domain to estimating the covariance function $\gamma(h)$ in the time domain. As we have mentioned before, to the best of our knowledge, there is no literature on shrinkage estimation of the spectrum. However, the concept of shrinkage has been widely used in the context of estimation of the covariance matrix, and the insights gained in this context are useful for understanding the more general approach we will develop in the following chapter.

1.3.3.1 Stein's example

The concept of shrinkage dates back to Stein (1956), who discussed a surprising effect that is known as *Stein's paradox* or *Stein's example*. Suppose that we have a p-variate ($p > 2$) Gaussian random variable X that is distributed

$$X \sim \mathcal{N}(\mu, \mathrm{Id}_p),$$

and want to estimate the unknown mean vector μ, for which we have a sample of size $N = 1$ only. The usual parametric estimator would be to take the sample mean,

$$\hat{\mu} = X. \tag{1.32}$$

Stein's paradox comes into play when we follow a decision theoretic approach to evaluate the performance of the estimator (1.32). Using the mean squared error as a risk function,

$$\mathcal{R}(\hat{\mu}, \mu) = \mathrm{E}\,\|\hat{\mu} - \mu\|^2, \tag{1.33}$$

where $\|\cdot\|^2$ means the Hilbert-Schmidt norm

$$\left\|A_{(m\times n)}\right\|^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2,$$

it is possible to show that $\hat{\mu}$ has suboptimal risk. In decision theoretic terminology, an estimator $\hat{\mu}_1$ is *dominated* by another estimator $\hat{\mu}_2$ if, for all possible μ,

$$\mathcal{R}(\hat{\mu}_1, \mu) \geq \mathcal{R}(\hat{\mu}_2, \mu).$$

If an estimator is dominated by no other estimator, it is said to be *admissible*. Thus, Stein's paradox states that the estimator (1.32) is *inadmissible* under mean squared error risk (for $p > 2$). Stein has also shown that, for $p \leq 2$, the estimator (1.32) is admissible.
James & Stein (1961) have developed an estimator, the *James-Stein estimator*, of μ that, for $p > 2$, dominates the estimator (1.32):

$$\hat{\mu}_{JS} = \left(1 - \frac{(p-2)}{\|X\|^2}\right) X \tag{1.34}$$

This is, on first sight, a little bit surprising. As we deal with Gaussian data, the estimator (1.32) consists of one dimensional least squares or maximum likelihood estimators, respectively, of independent random variables $X_1, \ldots, X_p$, that are admissible. However, if more than two of these admissible estimators are combined, the risk can be improved upon by using global information (in the form of the squared norm of X in the denominator in (1.34)). This effect may be misunderstood as "combining the estimates of independent observations may improve the quality of the individual estimate". However, what is improved upon is the global risk (1.33), not the individual, one dimensional risk functions. A rigorous treatment of this for first moments of a wide class of distribution functions and different loss functions was developed by Brown (1966). Bock (1975) extended these results to statistically dependent observations (over dimension) that are allowed to have different variances. A comprehensive overview can be found in the book by Lehmann & Casella (1998).

1.3.3.2 Covariance matrix estimators based on decision theory

Stein's paradox has stimulated the development of new estimators not only of first (Efron & Morris 1973), but also of second moments (Stein 1975, Efron & Morris 1976), that follow a decision theoretic approach. These estimators have the appealing property that they are not only more precise in terms of risk, but are also numerically more stable than the respective maximum likelihood estimators. The latter property is, of

course, indispensable in applications where the inverse of an estimated covariance matrix is needed, e.g. in portfolio selection (Markowitz 1952). The reason why these estimators are numerically more stable is that they produce estimates of the covariance matrix with eigenvalues that are less dispersed than the eigenvalues of the least squares estimator S. Consider the James-Stein estimator of the first moment (1.34). It is gained by multiplying the usual least squares or maximum likelihood estimator of the first moment by a factor that is < 1. For this reason, the techniques that are based on this approach are called 'shrinkage' techniques. In case of the James-Stein estimator, the sample mean is shrunk towards zero. In covariance matrix estimation, an analogue is done: the covariance matrix estimator S is manipulated in a way such that its eigenvalues are shifted towards each other. We will discuss the behaviour of sample vs. true eigenvalues in more detail in chapter 2. For the moment, it suffices to state that the sample eigenvalues are, in mean, more dispersed than the true ones, which causes the condition number, defined as the ratio $l_{\max}/l_{\min}$ of largest to smallest eigenvalue, of an estimator to be greater than the condition number of the true covariance matrix.
All publications on shrinkage of covariance matrices in decision theory that we cite in our work are based on a loss function of Kullback-Leibler type:

$$\mathcal{L}(\hat{\Sigma}, \Sigma) = \operatorname{tr}(\hat{\Sigma}\Sigma^{-1}) - \log\det(\hat{\Sigma}\Sigma^{-1}) - p \tag{1.35}$$

and the respective risk function

$$\mathcal{R}(\hat{\Sigma}, \Sigma) = \mathrm{E}\,\mathcal{L}(\hat{\Sigma}, \Sigma). \tag{1.36}$$

The decision theoretic approach follows a *minimax* philosophy: it compares estimators $\hat{\theta}$ from a set F of eligible estimators. Then, an estimator $\hat{\theta} \in \mathit{F}$ of a parameter $\theta \in \Theta$ is minimax if it satisfies

$$\sup_{\theta\in\Theta} \mathcal{R}(\hat{\theta}, \theta) = \inf_{\tilde{\theta}\in\mathit{F}} \sup_{\theta\in\Theta} \mathcal{R}(\tilde{\theta}, \theta) \tag{1.37}$$

In words, an estimator has the minimax property if its risk is minimal for the worst possible choice of the estimand $\theta \in \Theta$. It is important to remark that the minimax property is not unique: several minimax estimators may exist within the same class F, and a minimax estimator may be dominated by another minimax estimator.

The first minimax estimator of the covariance matrix was derived by James & Stein (1961). It is referred to as the *constant risk minimax estimator*. Shrinkage for covariance matrices under decision theory has then been developed further in a series of publications by Haff (1977, 1979, 1980), who found other minimax estimators that dominate the constant risk minimax estimator. However, all estimators from these early papers are static in the sense that the amount of shrinkage is determined only by the sample size N and the dimension p of the data. The first minimax estimator that dominates the constant risk estimator and uses information from the realizations X was developed by Dey & Srinivasan (1985). Apart from N and p, it uses the eigenvalues of the sample covariance matrix S to determine the amount of shrinkage. After this, the decision theoretic approach of covariance estimation has been developed further by Dey & Srinivasan (1986), Kubokawa & Konno (1990), Loh (1991), Evans & Stark (1996), Kubokawa & Srivastava (1999) and Sheena (2002). However, the minimax approach can be criticized for being too conservative. The minimax property, which demands that an estimator have minimal risk for any choice of the true parameter $\theta \in \Theta$, can in practice be too restrictive and render the improvement attained by constructing a new, dominating estimator almost irrelevant. We will demonstrate this in chapter 4, where we will compare the performance of the constant risk minimax estimator and two estimators from the paper by Dey & Srinivasan (1985) to the performance of estimators that are theoretically based on asymptotic statistics. These estimators will be introduced in the following subsection.

1.3.3.3 Covariance matrix estimators based on asymptotic theory

A more recent approach on shrinkage estimation of covariance matrices has been introduced by Ledoit & Wolf (2003, 2004). Like the estimators in the previous subsection, these estimators rely on iid assumptions on the data. However, there are two basic differences to the decision theoretic approach: first, these results are not based on a Kullback-Leibler type risk (1.35), but on a Frobenius or Hilbert-Schmidt norm, which is technically advantageous, as the algebraic structures imposed by the norm, such as symmetry and non-negative definiteness, can be exploited. Second, and more important, the decision theoretic approach has been replaced by an asymptotic background.
Shrinkage can be realized as constructing a new estimator S^* of the co-

variance matrix that is a linear combination of the maximum likelihood estimator S and a shrinkage target R. The shrinkage target R is an estimator of the covariance matrix Σ that has small variance but large bias, and is well-conditioned. If shrinkage is implemented in this way, the problem is reduced to finding optimal shrinkage weights w_1, w_2 such that

$$S^* = w_1 R + w_2 S. \tag{1.38}$$

Now, the main idea of the concept is, first, to derive the optimal shrinkage weights. Then, consistent estimators w_1^*, w_2^* of the shrinkage weights are constructed. The maximum likelihood estimator S is consistent. Thus, if a classical asymptotic framework is used like in Ledoit & Wolf (2003), $w_1 \to 0$ and $w_2 \to 1$. The alternative is to use a Kolmogorov asymptotic framework, like in Ledoit & Wolf (2004), under which the dimension p is allowed to grow with the sample size N. It is most elegant to make the weak assumption that

$$\frac{p_N}{N} < K$$

for some constant $K < \infty$, as the classical asymptotic framework is then included as a special case.

In the absence of information on the structure of the true covariance matrix Σ, it is reasonable to choose a very simple shrinkage target. In Ledoit & Wolf (2004), the shrinkage target is just a multiple $\mu\,\mathrm{Id}$ of the identity matrix. Thus, the shrinkage target is characterized by extreme simplicity. This means that it is very biased but can be estimated from the data with small variance. The estimator S, on the other hand, is unbiased but has high variance, which offers the interpretation of shrinkage as a method that finds an optimal trade-off between bias, introduced by the shrinkage target, and variance, introduced by S. We will come back to this idea in chapter 2, when we will develop a sophisticated shrinkage estimator for the spectrum.

If the data come from a context where there is some at least vague background knowledge on its structure, it makes sense to adopt the shrinkage target to catch the fundamental properties of the cross-dimensional interdependence structure. E.g., in Ledoit & Wolf (2003), the authors deal with economic data and aim at estimating the covariance matrix of stock returns. This is a fundamental problem of portfolio selection

(Markowitz 1952), where it is crucial to obtain a precise and at the same time numerically stable estimate. These two goals can often be contradictory. A numerically stable estimate is commonly attained by using a highly structured estimator. E.g., Fama (1971) proposes to use a scaled identity matrix as an estimator of the covariance matrix of stock returns. Elton & Gruber (1973) propose a model where the correlation between any two stock returns is the same. These estimators have low variance, but are on the other hand strongly biased.

A more elaborate approach is to fit factor models. The range here spans from simple single index models (Sharpe 1963) to multifactor models with infinite dynamics and nonorthogonal idiosyncratic components (Forni et al. 2000). However, in any factor model context the choice of the number of factors is a crucial question. Here, shrinkage estimation comes into play as an alternative that circumvents the problem of dimension choice completely. A factor model that is, from background knowledge, known to be too parsimonious, e.g. a one-factor model (Ledoit & Wolf 2003), is used as a shrinkage target. In this case, shrinkage can be understood not only as an improvement on S realized by shrinking towards a highly structured target, but also vice versa: the one-factor model, which is too restrictive, is *stretched* towards the estimator S, adding a refinement to the estimator given by the factor model, which is too biased.

Like the estimators based on decision theory, the estimators introduced in Ledoit & Wolf (2003, 2004) cannot cope with data that is dependent or not identically distributed. In the next two chapters, we will present the core of our work, two localized shrinkage estimators for multivariate time series that get rid of all these restrictions by transferring the concept of shrinkage to the domain of spectral analysis of time series.

CHAPTER 2

Model free shrinkage regularization

2.1 Basic concepts and definitions

The aim of this chapter, which has been published as Böhm & von Sachs (2007), 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 smoothing span of the averaged periodogram will be denoted by m_T like in the preceding chapter. The dimension of the time series we examine will be referred to as p_T from now on. The reason for this is that in section 2.2 we will introduce general asymptotics that allow for the dimension to grow with the length T of the time series. We will make the following assumptions on the time series data (X_T):

Assumption 2.1. The data (X_T), $t \in 1, \dots, T$, originate from a p_T-variate centered, stationary time series.

Assumption 2.2. The data (X_T), $t \in 1, \dots, T$, are Gaussian.

To measure the deviation of our estimators from the true spectrum, we make the following definitions: The loss of an estimator $\hat{f}(\omega)$ of the spectrum $f(\omega)$ at frequency ω,

$$\mathcal{L}(\hat{f}(\omega), f(\omega)) := \|\hat{f}(\omega) - f(\omega)\|^2 \tag{2.1}$$

and its risk

$$\mathcal{R}(\hat{f}(\omega), f(\omega)) := \mathrm{E}\,\|\hat{f}(\omega) - f(\omega)\|^2 \tag{2.2}$$

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

$$\|A\|^2 := \frac{1}{p_T}\operatorname{tr}(AA^*) = \frac{1}{p}\sum_{i,j=1}^{p}|a_{ij}|^2. \tag{2.3}$$

The reason to normalize by $1/p_T$ is the general asymptotic framework that will be introduced in the following section. If p_T is fixed, it is obvious that using the norm (2.3) is equivalent to using the Hilbert-Schmidt norm without normalization. Under general asymptotics, we need a modified HS norm that allows us to compare quadratic matrices of different dimension. It is necessary for the norm of the $p_T \times p_T$ identity matrix to be bounded away from both infinity and zero as $p_T \to \infty$. Any normalization function which accomplishes that aim could be used in (2.3) equivalently to $1/p_T$. The latter has been chosen for the sake of simplicity. For the reader to develop some intuition, let us compare the $p_T \times p_T$ null matrix to the $p_T \times p_T$ matrix that has one as its top left entry and zero entries everywhere else. The canonical HS-norm of the difference is always one. However, if $p_T \to \infty$, the norm (2.3) of the difference converges to zero. This reflects the fact that the difference between the two matrices becomes less important with respect to the increased dimensionality: they disagree on one dimension, but they agree on $p_T - 1$ others. If p_T is large, this disagreement becomes negligible, which is reflected in the normalization by $1/p_T$.
Associated to the normalized HS norm is a scalar product

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

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:

$$\hat{f}_T(\omega) := r_T(\omega)\,\mathrm{Id} + s_T(\omega)\hat{f}_T^0(\omega) \tag{2.4}$$

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.

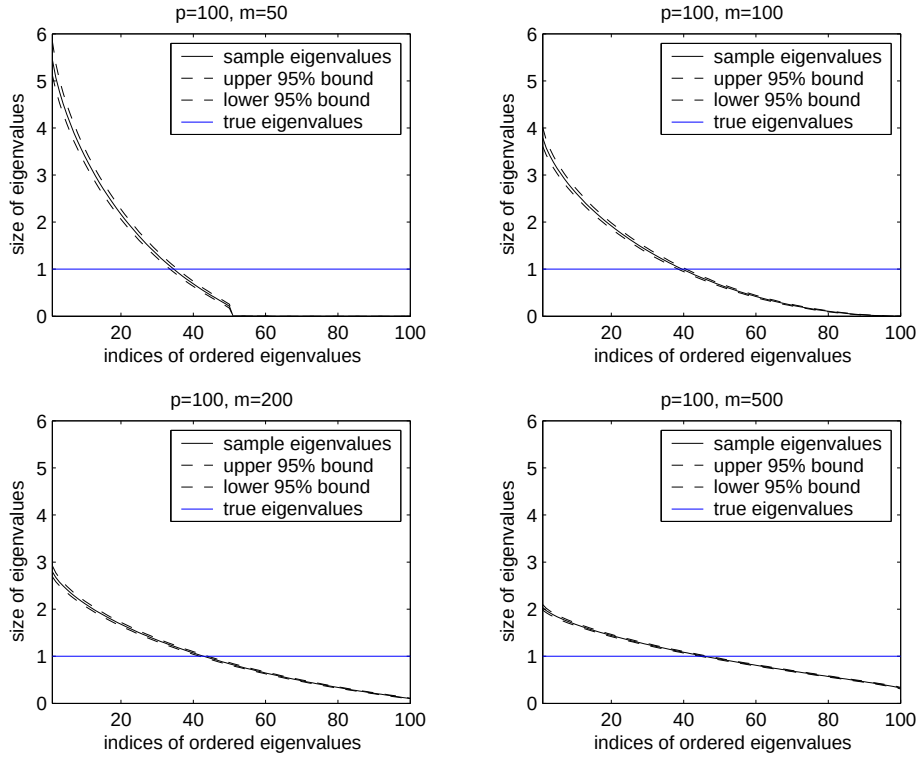

Figure 2.1: Eigenvalues of the spectrum of a $p = 100$-variate white noise process for different smoothing spans m. Average over $M = 10,000$ simulations. The black solid line shows the median, the dashed lines mark its empirical pointwise 95% confidence bounds.

Figure 2.1 shows the distribution of the sample eigenvalues at frequency $\pi/2$ of the averaged periodogram, based on a Monte Carlo simulation (10,000 runs). The underlying model is 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 Gaussian iid structure of the time series, the smoothed periodogram is an unbiased estimator of the spectrum here (see Brockwell & Davis 1987). 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

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

Thus all true eigenvalues are equal to 1. However, we see in the different subplots of figure 2.1 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. For true models which are as easy as this Gaussian IID model, there are theoretical results on the distribution of the sample eigenvalues (Marčenko & Pastur 1967). The distribution of the sample eigenvalues for a multivariate normal distribution with true covariance matrix = Id is given in Yin (1986). A good overview can be found in Frahm & Jaekel (2007). However, there is no result that is general enough to replace the approach of linear shrinkage in this paper.
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:

$$\text{estimation bias} = \mathrm{E}\,\hat{f}_T^0(\omega) - f_T(\omega)$$

It originates from the bias of the periodogram and from smoothing. We will speak of the 'sampling bias' when we mean the bias of the sample eigenvalues with respect to the true eigenvalues of the spectrum:

$$\text{sampling bias} = \mathrm{E}\,\hat{\lambda}_i(\omega) - \lambda_i(\omega)$$

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 2.1 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 (2.4), 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 (Jolliffe 2002), so it would

be technically difficult to prove optimality properties for a nonlinear shrinkage procedure. Then, we see in figure 2.1 that a linear function is a fairly good approximation of the distortion in the eigenvalues. We choose the identity matrix as a shrinkage target for two reasons. First because it has the best possible condition number, $c = 1$. Second, in the absence of any model structure for our time series, there is no other 'natural' candidate.
It is evident from (2.4) 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 (2.4).

2.2 Kolmogorov asymptotic framework

A proper theoretical framework is essential when looking for the optimal weights in (2.4). 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. $f(\omega) = f_T(\omega)$. We impose the following assumption:

Assumption 2.3. There exists a constant K_1 such that

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

Assumption 2.3 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, see theorem 2.3. 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 (2.3). 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 $f_T(\omega)$ at frequency ω, the eigenvalues $\lambda^{(\cdot)}(\omega)$ in $\Lambda_T(\omega)$ ordered from the biggest to the smallest, the eigenvectors in $\Gamma_T(\omega)$ normalized. We rotate the vector of the discrete Fourier transform to the eigensystem spanned by $\Gamma_T(\omega)$, defining

$$y_T(\omega) := \Gamma_T^*(\omega)d_T(\omega) \tag{2.5}$$

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.

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

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 2.4. There exists a constant K_2 such that for all ω and T,

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

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 2.4, we control the whole stochastic structure of the periodogram over frequency and dimension.

2.3 The oracle

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

Assumption 2.5. The real and imaginary parts of all components of the true spectrum $f_T(\omega)$ are twice continuously differentiable.

Assumption 2.6. The product of the smoothing span and the dimension grows slower than the sample size T:

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

In a classical asymptotic framework, asymptotic unbiasedness of the averaged periodogram, *i.e.*

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

is guaranteed by $m_T/T \to 0$ (Brillinger 1975); in our double asymptotic framework, we need assumption 2.6 in addition, as the number of remainder terms in $\left\| f_T^0(\omega) - f_T(\omega) \right\|^2$ grows dynamically with T at a rate p_T, see lemma B.7 and equation (2.13). Demanding that the second derivatives exist and are continuous allows us to keep the proofs simple. For the technical details, we point to Brillinger (1975), Brockwell & Davis (1987) and Shumway & Stoffer (2000).

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

$$\mu_T(\omega) := \langle f_T^0(\omega), \mathrm{Id} \rangle. \tag{2.6}$$

This is a scale parameter, as $\langle f_T^0(\omega), \mathrm{Id} \rangle = \frac{1}{p_T} \sum_{i=1}^{p_T} f_{ii}^0(\omega)$. The optimal shrinkage parameters can now be derived by a very simple geometric argument. $f_T^0(\omega)$, $\hat{f}_T^0(\omega)$ 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 $f_T^0(\omega)$ 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 2.2. 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 $f_T^0(\omega)$ as

$$\alpha_T^2(\omega) := \left\| f_T^0(\omega) - \mu_T(\omega)\,\mathrm{Id} \right\|^2 \tag{2.7}$$

$$\beta_T^2(\omega) := \mathrm{E} \left\| f_T^0(\omega) - \hat{f}_T^0(\omega) \right\|^2 \tag{2.8}$$

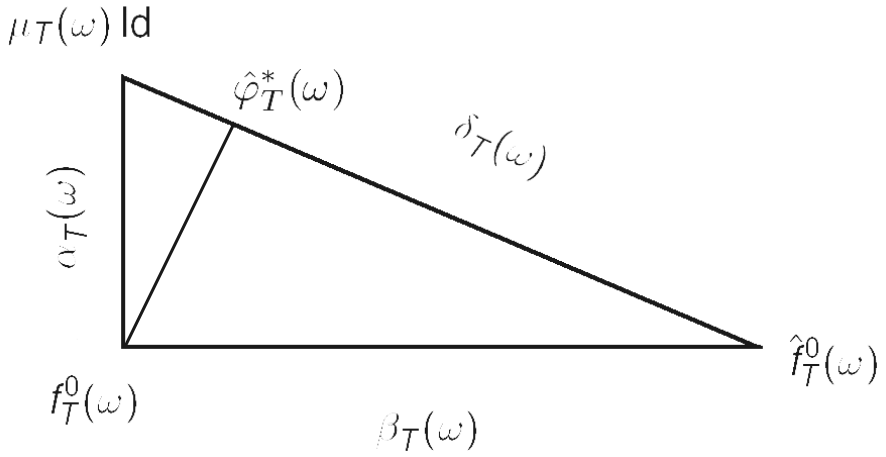

Figure 2.2: Geometrical derivation of the optimal shrinkage parameters. A Pythagorean relationship $\alpha^2+\beta^2 = \delta^2$ holds true. The optimal shrinkage parameters are derived by projecting $f_T^0(\omega) = \mathrm{E}\,\hat{f}_T^0(\omega)$ to the line spanned by the scaled identity matrix and the averaged periodogram.

$$\delta_T^2(\omega) \quad := \quad \mathrm{E}\left\|\hat{f}_T^0(\omega) - \mu_T(\omega)\,\mathrm{Id}\right\|^2 \tag{2.9}$$

These parameters follow a pythagorean relationship, as shown by the following lemma:

Lemma 2.1. The following relationship holds always true:

$$\alpha_T^2(\omega) + \beta_T^2(\omega) = \delta_T^2(\omega), \tag{2.10}$$

Proof.

$$\begin{aligned}
\delta_T^2(\omega) &= \mathrm{E}\left\|\hat{f}_T^0(\omega) - \mu_T(\omega)\,\mathrm{Id}\right\|^2 \\
&= \mathrm{E}\left\|\hat{f}_T^0(\omega) - f_T^0(\omega) + f_T^0(\omega) - \mu_T(\omega)\,\mathrm{Id}\right\|^2 \\
&= \alpha_T^2(\omega) + \beta_T^2(\omega) + \mathrm{E}\left\langle \hat{f}_T^0(\omega) - f_T^0(\omega), f_T^0(\omega) - \mu_T(\omega)\,\mathrm{Id}\right\rangle
\end{aligned}$$

$$
\begin{aligned}
& + \mathrm{E}\left\langle f_T^0(\omega) - \mu_T(\omega)\,\mathrm{Id}, \hat{f}_T^0(\omega) - f_T^0(\omega)\right\rangle \\
= \;& \alpha_T^2(\omega) + \beta_T^2(\omega)
\end{aligned}
$$

where we have used that the expected value of the scalar products is zero because $\mathrm{E}\,\hat{f}_T^0(\omega) = f_T^0(\omega)$. □

Using lemma (2.1), we derive the optimal shrinkage parameters $r_T(\omega)$ and $s_T(\omega)$ as

$$
\begin{aligned}
r_T(\omega) &:= \frac{\beta_T^2(\omega)}{\delta_T^2(\omega)}\mu_T(\omega) \\
s_T(\omega) &:= \frac{\alpha_T^2(\omega)}{\delta_T^2(\omega)}
\end{aligned}
$$

Thus, we have derived a first shrinkage estimator

$$
\hat{\varphi}_T^*(\omega) := \frac{\beta_T^2(\omega)}{\delta_T^2(\omega)}\mu_T(\omega)\,\mathrm{Id} + \frac{\alpha_T^2(\omega)}{\delta_T^2(\omega)}\hat{f}_T^0(\omega) \tag{2.11}
$$

to which we refer as the oracle, as it requires three functions of the expected averaged periodogram to be known. The asymptotic behavior of these four functions under the assumptions 2.3 and 2.4 is given by the following lemma

Lemma 2.2. Under assumption 2.1 and assumptions 2.3 to 2.6, as $T \to \infty$, all four functions $\mu_T(\omega)$, $\alpha_T^2(\omega)$, $\beta_T^2(\omega)$ and $\delta_T^2(\omega)$ remain bounded.

Proof. The proof of this lemma can be found in appendix A.1, in subsection A.1.1. □

The first important result that we show is the asymptotic behavior of the averaged periodogram under Kolmogorov asymptotics:

Theorem 2.3. Under assumption 2.1 and assumptions 2.3 to 2.6

$$
\lim_{T\to\infty} \mathrm{E}\left\|\hat{f}_T^0(\omega) - f_T^0(\omega)\right\|^2 - \frac{p_T}{m_T}\mu_T^2(\omega) = 0 \tag{2.12}
$$

Proof. The proof is given in A.1.2. □

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 (2.11) at estimating the expected averaged periodogram $f_T^0(\omega)$ is, due to construction, minimal amongst all estimators of the type (2.4). Asymptotically, the oracle constitutes the minimal risk estimator of the spectrum $f_T(\omega)$, too, as

$$\left\| f_T^0(\omega) - f_T(\omega) \right\|^2 = \mathrm{O}\left(p_T \frac{m_T^2}{T^2} \right) \tag{2.13}$$

which converges to zero due to assumption 2.6. Now, the minimal risk estimator in the class of linear shrinkage estimators (2.4) could be the averaged periodogram itself, the oracle not providing a real improvement. The following theorem shows that this is not the case:

Theorem 2.4. Under assumption 2.1 and assumptions 2.3 to 2.6, the risk of the oracle with respect to the expected averaged periodogram is given by

$$\mathrm{E} \left\| \hat{\varphi}_T^*(\omega) - f_T^0(\omega) \right\|^2 = \frac{\alpha_T^2(\omega)\beta_T^2(\omega)}{\delta_T^2(\omega)}$$

Proof. The proof is found in A.1.3. □

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

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

However, as the pythagorean relationship (2.10) 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 $\hat{f}_T^0(\omega)$.

2.4 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 (2.6), (2.7), (2.8), (2.9) and plugging in the estimators in (2.11). The trace $\mu_T(\omega)$ of $f_T^0(\omega)$ is estimated by the trace of the averaged periodogram:

$$\hat{\mu}_T(\omega) := \left\langle \hat{f}_T^0(\omega), \mathrm{Id} \right\rangle . \tag{2.14}$$

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

$$\hat{\delta}_T^2(\omega) := \left\| \hat{f}_T^0(\omega) - \hat{\mu}_T(\omega)\,\mathrm{Id} \right\|^2 \tag{2.15}$$

We cannot, however, derive estimators of $\alpha_T^2(\omega)$ and $\beta_T^2(\omega)$ in an equally straightforward 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 have only 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

$$\hat{\beta}_T^2(\omega) := \min(\bar{\beta}_T^2(\omega), \hat{\delta}_T^2(\omega)) \tag{2.16}$$

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

$$\hat{\alpha}_T^2(\omega) := \hat{\delta}_T^2(\omega) - \hat{\beta}_T^2(\omega) \tag{2.17}$$

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

Lemma 2.5. Under assumptions 2.1 to 2.6 we have, for any ω and $T \to \infty$

$$\begin{aligned} \mathrm{E}(\hat{\mu}_T(\omega) - \mu_T(\omega))^4 &\to 0 && (2.18)\\ \mathrm{E}(\hat{\alpha}^2_T(\omega) - \alpha^2_T(\omega))^2 &\to 0 && (2.19)\\ \mathrm{E}(\hat{\beta}^2_T(\omega) - \beta^2_T(\omega))^2 &\to 0 && (2.20)\\ \mathrm{E}(\hat{\delta}^2_T(\omega) - \delta^2_T(\omega))^2 &\to 0 && (2.21) \end{aligned}$$

Proof. The proof is found in A.1.4. □

Lemma 2.5 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 (2.14),(2.15),(2.16) and (2.17) for their estimands in the definition of the oracle (2.11):

$$\hat{f}^*_T(\omega) := \frac{\hat{\beta}^2_T(\omega)}{\hat{\delta}^2_T(\omega)}\hat{\mu}_T(\omega)\,\mathrm{Id} + \frac{\hat{\alpha}^2_T(\omega)}{\hat{\delta}^2_T(\omega)}\hat{f}^0_T(\omega) \tag{2.22}$$

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

Theorem 2.6. Under assumptions 2.1 to 2.6, $f^*_T(\omega)$ is a mean square consistent estimator of $\hat{\varphi}^*_T(\omega)$, 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:

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

Proof. The proof is found in A.1.5 □

Thus, we have derived an estimator that asymptotically is optimal in the class of linear shrinkage estimators (2.4) with nonrandom

coefficients, and which can be calculated without any knowledge of the true spectrum. As asymptotically also

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

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

2.5 Random oracle shrinkage intensities

Our next and final step will be to extend our theory to a larger class of estimators: the class of all linear combinations of the averaged periodogram and the identity matrix with *random* shrinkage intensity.
The shrinkage intensity of the oracle is not random: it depends on the true shape of the spectrum only, but is not affected by the actual realization of the data. Thus, if several realizations of time series with the same spectrum are examined, the oracle shrinkage intensity will always be the same. On the other hand, the shrinkage intensity of the DDSSE does depend on the realizations and is thus random.
According to theorem 2.6, the risk of the DDSSE converges to the risk of the oracle. However, the shrinkage intensity of the oracle is nonrandom. If we construct a larger class of oracle estimators that use both background knowledge on the true spectrum and the information in the data, it is not clear whether the oracle will attain the minimum in this larger class of estimators.
The minimizer of the class of shrinkage estimators with random coefficients is the one that minimizes the quadratic *loss* function (and, thus, almost surely, also the risk function). We will refer to this estimator as $\hat{\varphi}_T^{**}(\omega)$ or the *superoracle*. It is the solution of the optimization problem

$$\min_{\dot{r_T}(\omega), \dot{s_T}(\omega)} \left\| \hat{\varphi}_T^{**}(\omega) - f_T^0(\omega) \right\|^2 \tag{2.23}$$

subject to

$$\hat{\varphi}_T^{**}(\omega) = \dot{r_T}(\omega)\,\mathrm{Id} + \dot{s_T}(\omega)\hat{f}_T^0(\omega) \tag{2.24}$$

where the coefficients $\dot{r_T}(\omega)$ and $\dot{s_T}(\omega)$ are allowed to be random. It is now interesting to compare the DDSSE, the oracle and the superoracle

to each other. If, asymptotically, the risk of the superoracle were smaller than the risk of the oracle, we would have to search for a new data driven estimator that has asymptotically negligible distance to the superoracle. Fortunately, this is not true. The following theorem will show that the DDSSE is a mean square consistent estimator of the superoracle, too:

Theorem 2.7. Under assumptions 2.1 to 2.6, the DDSSE $\hat{f}_T^*(\omega)$ is a mean square consistent estimator of the superoracle $\hat{\varphi}_T^{**}(\omega)$, 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 superoracle:

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

Proof. The proof is found in A.1.6. □

Theorem 2.7 has two important implications. First, all three estimators, DDSSE, oracle and $\hat{\varphi}_T^{**}(\omega)$ are, asymptotically, equivalent. This means that, under Kolmogorov asymptotics, in the limit there is no advantage in having background knowledge of the true spectrum or, for the purpose of estimating the shrinkage coefficients, in using the information contained in the actual realization of the time series. Furthermore, from the proof of theorem 2.7, we are able to give an explicit formula for the superoracle:

Theorem 2.8. Under assumptions 2.1 to 2.6, let

$$\dot{\alpha}_T^2(\omega) = \left\langle f_T^0(\omega), \hat{f}_T^0(\omega)\right\rangle - \mu_T(\omega)\hat{\mu}_T(\omega).$$

Then the superoracle $\hat{\varphi}_T^{**}(\omega)$ has the explicit formula

$$\hat{\varphi}_T^{**}(\omega) = \left(\mu_T(\omega) - \frac{\dot{\alpha}_T^2(\omega)}{\hat{\delta}_T^2(\omega)}\hat{\mu}_T(\omega)\right)\mathrm{Id} + \frac{\dot{\alpha}_T^2(\omega)}{\hat{\delta}_T^2(\omega)}\hat{f}_T^0(\omega) \tag{2.25}$$

From (2.25), we see that the superoracle does, indeed, depend on background knowledge of the underlying process, as well as on information in the respective realization of the time series.

So far, we have seen that the DDSSE has certain, important optimality properties under Kolmogorov asymptotics. In chapter 4, we will show that the DDSSE performs very well in practice, too. This is even true for very small datasets.

CHAPTER 3

Shrinkage based on factor models

3.1 Introduction

In chapter 2, we have developed theory that uses the identity matrix as a shrinkage target. This is reasonable if there is little or no knowledge about the underlying multidimensional structure of the data. In this case, we use a shrinkage target that imposes the least possible amount of structure and which, at the same time, has the best of all possible condition numbers.

In some settings, however, it is reasonable to assume that the covariance or spectral structure of the data is induced by underlying, known or hidden, factors. The general idea underlying factor models is that p observed random variables can be expressed, except for an error term, as linear functions of $q < p$ random *factors*. For instance, in econometrics, markets are usually assumed to be driven by underlying global variables such as interest rate, employment rate, gross national product. The models reach from simple one-factor models to very sophisticated approaches that use multiple global and industry specific factors that may be intercorrelated. A first, very simple model was proposed by Sharpe (1963), who assumed that stock returns can be described by a one-factor model with uncorrelated residuals. Until today, many refinements have been made, reaching up to complex models that use double asymptotics. E.g., Forni et al. (2000) developed a generalized dynamic factor model that allows for non-iid factors, lags and intercorrelated idiosyncratic components.

In the literature, two general types of factor model approaches exist:

hidden or *latent* factor models on one hand, *exogenous* factor models on the other hand. Both have in common that they try to describe the behavior of a p-variate set of variables by $q < p$ factors. The focus in hidden factor models is on estimating the factor values, whereas the focus in exogenous factor models is on predicting future behavior of the variables by future values of the factors. We give an example for each.
Psychometrics is a typical area of application for hidden factor models. Consider as an example an IQ test. An individual that takes the test will respond to a large number of items, typically of multiple choice type, with only one correct answer. Items of the same type are summarized in p scales. These p values constitute the raw data per individual. For a comprehensive IQ test, the number of scales typically lies between fifteen and thirty. An individual's ability to solve the test is assumed to depend on its IQ, which is a hidden set of q factor values, like memory, reasoning or verbal fluency. Typically, q lies between five and seven. Questions that arise in this context are the choice of q, and which linear combination of the p variables performs best in estimating an individual's q IQ factor scores.
In econometrics, exogenous factor models are widely used. Here, a number of known, exogenous variables is used to model the behavior of a larger number of other variables. E.g., the variables to be modelled might be p stock prices, and out of an exogenous data set consisting of global variables like gross national product, employment rate, interest rate, inflation, a subset of q factors that predict the stock prices best is selected.
A disadvantage of factor models is that, usually, the number of factors is a parameter that must be specified *a priori*. In practice, it is seldom if ever possible to do so. This introduces a certain degree of arbitrariness into the domain of factor analysis: Although certain theoretic criteria for choosing the number of factors do exist, they deliver different results. Thus, the choice of the number of factors to actually use is left to the intuition of the data analyst. It may even be suspected that the choice of this parameter is often made *a posteriori*, making the whole analysis prone to overfitting.
On the other hand, the structure imposed by a factor model is a remedy to the curse of dimensionality. This makes factor analysis a powerful tool in multivariate statistics.
We have developed a hybrid approach to circumvent the dilemma of

model choice and still retain the advantages of factor analysis. We combine a nonparametric estimator, in our case the averaged spectrum, with a parametric estimator of the spectral matrix. The latter is our new shrinkage target. It is given by fitting a one-factor model to the data. However, we do not assume that this model is true; rather, we believe that the data follow a more complicated structure. This may be a q-factor model ($q > 1$), a model driven by different layers of factors, or the model may be completely unknown. By combining a shrinkage target, which is actually underfitted, with a nonparametric estimator of the spectrum, we circumvent the problem of model choice. In a data driven approach, weights are chosen such that the new, hybrid estimator is the asymptotically optimal linear combination of two conventional estimators. The first component, the averaged periodogram, is asymptotically unbiased but has high variance. The second component is biased due to misspecification but, by imposing structure, has low variance.
Two things are to be remarked here:
First, instead of choosing a one-factor model as our shrinkage target, we might as well opt for something more complicated, e.g. a q factor model with $q > 1$. The only prerequisite for doing this is having background knowledge that the underlying structure is more complicated than the shrinkage target, e.g. a $\tilde{q}$ factor model, $\tilde{q} > q$. The theory we will give in section 3.2 can easily be adapted to such a case.
Second, for this chapter, we have chosen an approach that uses a conventional asymptotic framework. This means that the optimal shrinkage intensity converges to zero as $T \to \infty$. Estimating the optimal shrinkage intensity under this approach is a 'race against time': as more data becomes available, it becomes less important to shrink as the new information obtained does not have additional dimensions. Our choice to develop shrinkage to identity under double asymptotics and shrinkage to factor model under standard asymptotics is to a certain degree arbitrary, as illustrated in figure 3.1 However, we wanted to demonstrate both possible approaches in this thesis. As in chapter 2 we have the simpler shrinkage target, we decided to introduce the more complicated theoretical framework with it. As the classical asymptotic framework is a special case of Kolmogorov asymptotics, the theory in chapter 2 remains valid for fixed p. In this case, the necessity to shrink vanishes asymptotically, and the shrinkage estimator converges to the averaged periodogram. The reader who, for application, might be interested in

Figure 3.1: Possible asymptotic frameworks and shrinkage targets in spectral analysis. The combinations developed for this work are in bold. The combination in the upper left is a special case of the results in chapter 2.

shrinkage to a factor model under double asymptotics may develop it with the help of the techniques used in the proofs in the appendices. However, the technical details may be tedious to obtain, as migrating from a classical to a Kolmogorov asymptotic framework implies most convergence rates to deteriorate.
Again, to the best of the author's knowledge, there is no literature on

shrinkage to a factor model in time series analysis. In the literature on finance, an approach to shrink to a factor model has been developed in the context of portfolio selection (Ledoit & Wolf 2003) under iid assumptions on the data.
The rest of the chapter is organized as follows: in the next section, we will develop the theoretical background for data driven shrinkage to a 'market' one-factor model, where the term 'market' is just a wildcard term that does not necessarily mean that we are in an economic context. We will first give the basic assumptions and definitions in the following subsection. In section 3.3, we will introduce the shrinkage target, which is a one-factor model. The model assumptions are that the p dimensional process is driven by a dynamic, hidden or known, underlying process with spectral density $f_0(\omega)$. The model assumes that there are no lag effects. We will fit this model to the data; however, at the same time we assume that the model be misspecified. The philosophy behind this is that the model is just a parsimonious tool of describing the data. The reason that we do not allow the model to be true is that otherwise we would encounter identifiability problems with respect to the shrinkage intensity. We will derive the optimal solution for the shrinkage intensity in sections 3.4 and 3.5. It is, like in the previous chapter, a function of the true spectral density $f(\omega)$. In section 3.6, we will examine the asymptotic behaviour of the optimal shrinkage intensity, which will help us to develop a data driven estimator in 3.7. Comprehensive Monte Carlo studies will show the usefulness of our estimator in chapter 4.

3.2 Setup and assumptions

Our aim in this chapter is to estimate the spectrum $f(\omega)$ of a p-variate Gaussian time series. We assume that we have realizations

$$(X_{it})_{t\in\{1,\dots,T\}} = X_{i1},\dots,X_{iT}, \qquad i = 1,\dots,p$$

Moreover, we assume that we have realizations from another, one dimensional time series

$$(X_{0t})_{t\in\{1,\dots,T\}} = X_{01},\dots,X_{0T}$$

to which we refer as the *market* or *exogenous* time series. The market time series is thought to be a process that has a certain explanatory value for the other time series $(X_{it}), i = 1,\dots,p$. One possible choice is

to use the average over dimension in the time domain of the $(X_{it})_{i=1,\dots,p}$,

$$X_{0t} = \frac{1}{p}\sum_{i=1}^{p} X_{it}$$

However, we make no special assumptions on the market time series. It would as well be possible to choose an external variable or the first principal component of the data.
We assume that all our time series, including the market time series, are centered

$$\mathrm{E}\, X_{it} = 0 \qquad i = 0, \dots, p$$

and stationary. Furthermore, if

$$\gamma_{ij}(h) = \mathrm{E}\, X_{it} X_{j,t+h}, \qquad i, j = 0, \dots, p, \quad h \in \mathbb{Z}$$

denotes the autocovariance function, we impose the following summability condition:

$$\sum_{h=-\infty}^{\infty} |h\gamma_{ij}(h)| < \infty \qquad \forall i, j = 0, \dots, p$$

The enhanced estimator we want to construct is gained by linearly combining a standard nonparametric estimator, in our case the averaged periodogram, with a shrinkage target. The latter is gained by fitting a one-factor model to the data, where the time series X_{0t} is assumed to be the underlying factor.
In this chapter, we assume the dimension p to be fixed while still $T \to \infty$. Like in the previous chapter, we denote the ith component of the discrete Fourier transform

$$d_T(\omega) = \frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} X_t \exp(-\iota\omega t), \quad \omega \in (0, 2\pi) \tag{3.1}$$

of the data at frequency ω as $d_i(\omega)$.
We furthermore make the following notational convention: whenever we use vector- or matrix valued terms, we will mean the respective p-dimensional vector or the $p \times p$ matrix unless we explicitly state otherwise. Thus, $f(\omega), f_T^0(\omega)$ and $\hat{f}_T^0(\omega)$ refer to the spectrum, expected

averaged periodogram and averaged periodogram, respectively, of the time series $(X_{it})_{i=1,\dots,p}$. We will also refer to the p-dimensional vector of the time series at time t as X_t. However, when we look at components, we will use the index value zero to refer to the market time series. E.g., we refer to the cross-spectrum between the market series and the first component of X_t as $f_{01}(\omega)$.

3.3 One-factor model

The shrinkage target is given by fitting a one-factor model to the data $(X_{it}), i = 1, \dots, p$, which we will define in this section. We will use a different notation for the random variables to emphasize that this model is not assumed to hold true for the data X_{it}. Rather, we use the model as a parsimonious tool to approximate the spectral structure of the process. Let us assume that we have a univariate *exogenous* time series $\dot{X}_{0t}, t = 1, \dots, T$ with spectrum $\dot{f}_0(\omega)$. When we speak of exogenous, we mean that this data $\dot{X}_{0t}$ can be used as a factor time series that has some explicative value for the data in the sense of the following model:

$$\dot{X}_{it} = \beta_i \dot{X}_{0t} + \epsilon_{it} \quad i = 1, \dots, p \tag{3.2}$$

The weights $\beta_i \in \mathbb{R}$ are non-random. The idiosyncratic components ϵ_{it} are assumed to be normally distributed and independent over time and dimension, and independent of $(\dot{X}_{0t})$:

$$\epsilon_{it} \sim \mathcal{N}\left(0, 2\pi(\sigma_i^{\epsilon})^2\right) \tag{3.3}$$

In this simple factor model, all serial correlation in the data $\dot{X}_{it}$ originates from serial correlation in the exogenous time series $\dot{X}_{0t}$. The serial correlation of the exogenous time series is determined by its spectrum $\dot{f}_0(\omega)$. Furthermore, there are no direct lag effects in the sense that in the notation

$$\dot{X}_{it} = \sum_{k=-\infty}^{\infty} \beta_{i,k} \dot{X}_{0,t-k} + \epsilon_{it}$$

all $\beta_{i,k}$ for $k \neq 0$ would be zero. The fact that the idiosyncratic components are uncorrelated over time and dimension is important, as in either other case, it would be impossible to identify the model under classical asymptotics (Forni et al. 2000). Together with the independence between the idiosyncratic components and the exogenous time series, this

has two more advantages: first, it will allow us to use simple linear regression to estimate the β_i and the $(\sigma_i^\epsilon)^2$. More complicated methods like generalized least squares are not necessary here. Second, this model implies the following relationship for the DFTs of the data :

Lemma 3.1.

$$\dot{d}_i(\omega) = \beta_i \dot{d}_0(\omega) + \dot{d}_i^\epsilon(\omega) \tag{3.4}$$

where $\dot{d}_i^\epsilon(\omega)$ is the DFT of the idiosyncratic components. Furthermore,

$$\dot{d}_i^\epsilon(\omega) \sim \mathcal{N}^C\left(0, (\sigma_i^\epsilon)^2\right) \quad \forall\omega \tag{3.5}$$

Proof. We have

$$\begin{aligned}
\dot{d}_k(\omega) &= \sum_{t=1}^{T} \dot{X}_{kt} \exp(-\iota t\omega) \\
&= \sum_{t=1}^{T} \left(\beta_k \dot{X}_{0t} + \epsilon_{kt}\right) \exp(-\iota t\omega) \\
&= \sum_{t=1}^{T} \beta_k \dot{X}_{0t} \exp(-\iota t\omega) + \sum_{t=1}^{T} \epsilon_{kt} \exp(-\iota t\omega) \\
&= \beta_k \dot{d}_0(\omega) + \dot{d}_k^\epsilon(\omega)
\end{aligned}$$

which proves (3.4). According to Brillinger (1975, p.95), we have the following relationship between the distribution of the discrete Fourier transform of Gaussian data and the spectrum of the underlying process:

$$d(\omega) \sim \mathcal{N}^C\left(0, f(\omega)\right) \tag{3.6}$$

Now, (ϵ_{it}) is Gaussian white noise, which means that its spectrum is constant. According to (3.3),

$$f^\epsilon(\omega) = (\sigma_i^\epsilon)^2$$

which proves (3.5). □

We see from (3.5) that the variance in the idiosyncratic components is independent of frequency. Furthermore, the weights $\beta = (\beta_1, \ldots, \beta_p)'$ are independent of the frequency, too, due to (3.2). This means that the spectrum under the above specified one-factor model (3.2) is

$$\tilde{f}^1(\omega) = \beta\beta' \dot{f}_0(\omega) + \Delta \tag{3.7}$$

where

$$\Delta = \begin{pmatrix} (\sigma_1^\epsilon)^2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & (\sigma_p^\epsilon)^2 \end{pmatrix} \tag{3.8}$$

When it comes to estimation of the one-factor model, we will, like in the previous chapter, identify the spectrum with the expected averaged periodogram. Thus, instead of using the model (3.7), we will use the slightly modified model

$$\dot{f}^1(\omega) = \beta\beta' \dot{f}_0^0(\omega) + \Delta, \tag{3.9}$$

where $\dot{f}_0^0(\omega)$ means the expected averaged periodogram of the factor time series $\dot{X}_{0t}$.

3.3.1 Estimation of one-factor model

The model (3.9) is assumed *not* to hold true. However, even under these circumstances, it is possible to fit it to the time series X_{it} by choosing weights β_i such that the L_2 distance between $f(\omega)$ and $\beta\beta' f_0(\omega)$ becomes minimal. We will refer to this minimum L_2 distance spectral density under one-factor model as to $f^1(\omega)$.
The fact that both weights β_i and idiosyncratic variances $(\sigma_i^\epsilon)^2$ are independent of lag and frequency, respectively, enables us to estimate these parameters with standard methods. We use linear regression to obtain the following estimators b_i for the β_is

$$b_i = \frac{\sum_{t=1}^T (X_{0t} X_{it})}{\sum_{t=1}^T (X_{0t})^2} \tag{3.10}$$

This is just the standard estimator of the slope in linear regression, which can be found in any book on applied statistics, like the books by Neter (1978) and Sachs & Hedderich (2006).

Next, we need to estimate the variances $(\sigma_i^\epsilon)^2$ of the idiosyncratic components. The standard way to do this is again to use the time domain estimator of the residual variance, which we normalize by $1/2\pi$:

$$\widehat{(\sigma_i^\epsilon)^2} = \frac{1}{2\pi} \sum_{t=1}^{T} \frac{(X_{0t} - b_i X_{it})^2}{T} \tag{3.11}$$

Furthermore, both estimators have the convenient property of being consistent, and the stochastic rate of convergence is in both cases $1/\sqrt{T}$ (Sachs & Hedderich 2006), which we will find useful in deriving an approximation of the optimal shrinkage intensity $\zeta_T^*(\omega)$:

$$b_i = \beta_i + \mathrm{O}_p\left(\frac{1}{\sqrt{T}}\right) \tag{3.12}$$

and

$$\widehat{(\sigma_i^\epsilon)^2} = (\sigma_i^\epsilon)^2 + \mathrm{O}_p\left(\frac{1}{\sqrt{T}}\right) \tag{3.13}$$

Plugging the estimators from (3.12) and (3.13) and the averaged periodogram into the definition of the one-factor model (3.9), we obtain an estimator of the multivariate spectrum that is based on a one-factor model:

$$\hat{f}_T^1(\omega) = bb' \hat{f}_0^0(\omega) + D \tag{3.14}$$

where

$$D = \operatorname{diag}\left(\widehat{(\sigma_1^\epsilon)^2} \dots \widehat{(\sigma_p^\epsilon)^2}\right)$$

This estimator is our shrinkage target.

3.4 Optimal shrinkage intensity

Our aim is, like in the previous chapter, to improve upon the averaged periodogram by shrinking to a target matrix function that is more regular, at the prize of possibly having larger bias. In this chapter, however, we make the assumption that a one-factor model is not far too crude an approximation. We do, however, not believe that the underlying structure is totally explained; we even make the opposite assumption, namely that the model is misspecified:

Assumption 3.1. There exists a $\delta > 0$ such that, uniformly over all frequencies $\omega \in [0, 2\pi]$ and all $i, j = 1, \ldots, p$, we have

$$\left| f_{ij}^1(\omega) - f_{ij}^0(\omega) \right| \geq \delta \tag{3.15}$$

Assumption 3.1 is made for technical reasons: the estimator of the shrinkage intensity which we are going to derive will have an estimator of the difference $f_{ij}^1(\omega) - f_{ij}^0(\omega)$ in the denominator. Because of this, assumption 3.1 is needed to avoid problems of identifiability.
We search for a linear combination

$$\hat{f}^+(\omega) = \zeta_T(\omega)\hat{f}_T^1(\omega) + (1 - \zeta_T(\omega))\hat{f}_T^0(\omega)$$

where $\zeta_T(\omega)$ is a data driven estimator of an optimal, oracle shrinkage intensity $\zeta_T^*(\omega)$ that is the solution of the minimization problem

$$\mathrm{E} \left\| \hat{f}_T^+(\omega) - f_T^0(\omega) \right\|^2 = \min! \tag{3.16}$$

In this chapter, we will not normalize the Hilbert Schmidt norm

$$\|A\|^2 = \sum_{i,j=1}^{p_T} |a_{ij}|^2$$

by $1/p_T$ like in the preceding one, as the dimension does not change with T. We will proceed in three steps:
First, in 3.5, we will derive the optimal, *oracle* shrinkage intensity $\zeta_T^*(\omega)$ which depends on background knowledge of the underlying process. Like in the procedure for deriving the oracle estimator in chapter 2, we will identify the estimand not with the true spectrum $f(\omega)$, but with the expected averaged periodogram $f_T^0(\omega)$.
Second, in 3.6 we will derive the asymptotic behaviour of the oracle shrinkage intensity. We will see that, due to the different asymptotic setting used here, the necessity to shrink vanishes asymptotically. This is because the averaged periodogram is a consistent estimator whereas the shrinkage target is misspecified due to assumption 3.1. In consequence, the data driven estimator of $f_T^0(\omega)$ will asymptotically have the same behaviour as the averaged periodogram, as the data driven estimator of the shrinkage intensity will converge to zero. Finally, we will construct a data driven estimator in subsection 3.7.

3.5 Oracle shrinkage intensity

We will derive the *oracle* shrinkage intensity by solving the minimization problem given in formula (3.16). This can simply be done by differentiation. Let $z \in [0,1]$ denote a shrinkage intensity. The risk associated with z is

$$R(z) = \mathrm{E}\left\| z\hat{f}_T^1(\omega) + (1-z)\hat{f}_T^0(\omega) - f_T^0(\omega) \right\|^2 \tag{3.17}$$

(3.17) can be decomposed as follows:

$$\begin{aligned}
& (3.17) \\
= & \sum_{i,j=1}^{p} \mathrm{E}\left| z\hat{f}_{ij}^1(\omega) + (1-z)\hat{f}_{ij}^0(\omega) - f_{ij}^0(\omega) \right|^2 \\
= & \sum_{i,j=1}^{p} \mathrm{Var}\left| z\hat{f}_{ij}^1(\omega) + (1-z)\hat{f}_{ij}^0(\omega) - f_{ij}^0(\omega) \right| \\
& + \sum_{i,j=1}^{p} \left| \mathrm{E}\left(z\hat{f}_{ij}^1(\omega) + (1-z)\hat{f}_{ij}^0(\omega) - f_{ij}^0(\omega) \right) \right|^2 \\
= & \sum_{i,j=1}^{p} \mathrm{Var}\left| z\hat{f}_{ij}^1(\omega) + (1-z)\hat{f}_{ij}^0(\omega) - f_{ij}^0(\omega) \right| \\
& + \sum_{i,j=1}^{p} \left| z f_{ij}^1(\omega) - z f_{ij}^0(\omega) \right|^2 \\
= & \sum_{i,j=1}^{p} \Big(z^2 \,\mathrm{Var}\, \hat{f}_{ij}^1(\omega) + (1-z)^2 \,\mathrm{Var}\, \hat{f}_{ij}^0(\omega) \\
& + z(1-z)\,\mathrm{Cov}\left(\hat{f}_{ij}^1(\omega), \hat{f}_{ij}^0(\omega) \right) \\
& + z(1-z)\,\mathrm{Cov}\left(\hat{f}_{ij}^0(\omega), \hat{f}_{ij}^1(\omega) \right) \\
& + z^2 \left| f_{ij}^1(\omega) - f_{ij}^0(\omega) \right|^2 \Big) \\
= & \sum_{i,j=1}^{p} \Big(z^2 \,\mathrm{Var}\, \hat{f}_{ij}^1(\omega) + (1-z)^2 \,\mathrm{Var}\, \hat{f}_{ij}^0(\omega) \\
& + 2z(1-z)\Re\left(\mathrm{Cov}\left(\hat{f}_{ij}^1(\omega), \hat{f}_{ij}^0(\omega) \right) \right) \\
& + z^2 \left| f_{ij}^1(\omega) - f_{ij}^0(\omega) \right|^2 \Big)
\end{aligned}$$

where we have used that

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

and, according to (3.9),

$$\mathrm{E}\,\hat{f}_T^1(\omega) = f_T^1(\omega).$$

The first derivative with respect to z of this is:

$$\begin{aligned} R'(z) &= 2\sum_{i,j=1}^{p}\Big(z\,\mathrm{Var}\,\hat{f}_{ij}^1(\omega) - (1-z)\,\mathrm{Var}\,\hat{f}_{ij}^0(\omega) \\ &\quad +(1-2z)\Re\left(\mathrm{Cov}\left(\hat{f}_{ij}^1(\omega), \hat{f}_{ij}^0(\omega)\right)\right) + z\left|f_{ij}^1(\omega) - f_{ij}^0(\omega)\right|^2\Big) \end{aligned}$$

Moreover, the second derivative is

$$\begin{aligned} R''(z) &= 2\sum_{i,j=1}^{p}\left(\mathrm{Var}(\hat{f}_{ij}^1(\omega) - \hat{f}_{ij}^0(\omega)) + \left|f_{ij}^1(\omega) - f_{ij}^0(\omega)\right|^2\right) \\ &> 0 \end{aligned} \tag{3.18}$$

where we use that $\hat{f}_T^1(\omega)$ and $\hat{f}_T^0(\omega)$ are hermitian, so that the imaginary parts sum to zero.
Thus, we know that any local extremum will be a minimum. Setting the first derivative equal to zero, we obtain the following theorem.

Theorem 3.2. The optimal shrinkage intensity is given by

$$\zeta_T^*(\omega) = \frac{\sum_{i,j=1}^{p}\left(\mathrm{Var}\,\hat{f}_{ij}^0(\omega) - 2\Re\,\mathrm{Cov}\left(\hat{f}_{ij}^1(\omega), \hat{f}_{ij}^0(\omega)\right)\right)}{\sum_{i,j=1}^{p}\left(\mathrm{Var}(\hat{f}_{ij}^1(\omega) - \hat{f}_{ij}^0(\omega)) + \left|f_{ij}^1(\omega) - f_{ij}^0(\omega)\right|^2\right)} \tag{3.19}$$

Proof. The proof is found in A.2.1. □

3.6 Asymptotic behaviour of optimal shrinkage intensity

Now, we will examine the asymptotic behavior of the optimal shrinkage intensity (3.19). This will enable us to derive a data driven estimator in the following subsection. We first define the following parameters:

$$\pi(\omega) = \sum_{i,j=1}^{p}\pi_{ij}(\omega) \tag{3.20}$$

$$\rho(\omega) = \sum_{i,j=1}^{p} \rho_{ij}(\omega) \tag{3.21}$$

$$\gamma(\omega) = \sum_{i,j=1}^{p} \gamma_{ij}(\omega) \tag{3.22}$$

where the subcomponents are defined, respectively, as:

$$\pi_{ij}(\omega) = \text{AsyVar}\left(\sqrt{m_T}\hat{f}_{ij}^0(\omega)\right) \tag{3.23}$$

$$\rho_{ij}(\omega) = \text{AsyCov}\left(\sqrt{m_T}\hat{f}_{ij}^1(\omega), \sqrt{m_T}\hat{f}_{ij}^0(\omega)\right) \tag{3.24}$$

$$\gamma_{ij}(\omega) = \left|f_{ij}^1(\omega) - f_{ij}^0(\omega)\right|^2 \tag{3.25}$$

using the notation

$$\text{AsyVar}(\cdot) := \lim_{T\to\infty} \text{Var}(\cdot).$$

Now, we can express $\zeta_T^*(\omega)$ as a function of (3.20) to (3.22) plus a remainder term which converges to zero sufficiently swift. The exact expression is given by the following theorem:

Theorem 3.3. The optimal shrinkage intensity can be expressed as the following function of the parameters $\pi(\omega), \rho(\omega)$ and $\gamma(\omega)$:

$$\zeta_T^*(\omega) = \frac{1}{m_T}\frac{\pi(\omega) - 2\Re(\rho(\omega))}{\gamma(\omega)} + \text{O}\left(\frac{1}{T}\right) \tag{3.26}$$

Proof. The proof is found in A.2.2 □

This means that the optimal shrinkage intensity converges to zero at a rate of $1/m_T$. At the same time, it can be approximated by the parameters (3.20) to (3.22) with an error that vanishes at the faster rate of $1/T$. This will allow us to derive a data driven estimator of the shrinkage intensity, and thus of $f_T^0(\omega)$, by plugging in estimators for (3.20) to (3.22) in (3.26).

3.7 Data driven estimation

The final step in deriving a data driven estimator of the spectrum that combines the averaged periodogram with a parsimonious, one-factor model based estimator, is to derive estimators for the parameters

$\pi(\omega), \rho(\omega)$ and $\gamma(\omega)$. We will start by estimating $\pi(\omega)$. According to (3.20), $\pi_{ij}(\omega)$ is the asymptotic variance of the i,jth component of the averaged periodogram, scaled by the smoothing span m_T. The following lemma will give a consistent estimator:

Lemma 3.4. $\pi(\omega)$ is estimated consistently by

$$p(\omega) = \sum_{i,j=1}^{p} p_{ij}(\omega) \tag{3.27}$$

where

$$p_{ij}(\omega) = \frac{1}{m_T} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} |I_{ij}(\tilde{\omega}_k) - \hat{f}_{ij}^0(\omega)|^2 \tag{3.28}$$

i.e. $p_{ij}(\omega)$ is the standard estimator of the local variance of the (i,j)th component of the periodogram at frequency ω.

Proof. The proof is given in A.2.3 □

The next step is to estimate $\rho(\omega)$. We will estimate its components and distinguish between the components on the diagonal and the components on the off-diagonal. On the diagonal, $\hat{f}_T^1(\omega) = \hat{f}_T^0(\omega)$, thus we can use the estimator (3.28). On the off-diagonal, we can use the estimator given by the following lemma:

Lemma 3.5. For $i \neq j$, a consistent estimator of $\rho_{ij}(\omega)$ is given by

$$r_{ij}(\omega) = b_i b_j \hat{f}_{0i}^0(\omega) \hat{f}_{j0}^0(\omega) \tag{3.29}$$

Proof. The proof is given in A.2.4. □

The estimator of the last of the three parameters, $\gamma(\omega)$, is derived in a straightforward way:

Lemma 3.6. $\gamma(\omega)$ is estimated consistently by

$$g(\omega) = \sum_{i,j=1}^{p} g_{ij}(\omega) \tag{3.30}$$

where

$$g_{ij}(\omega) = \left| \hat{f}_{ij}^1(\omega) - \hat{f}_{ij}^0(\omega) \right|^2 \tag{3.31}$$

Proof. Both $\hat{f}_T^0(\omega)$ and $\hat{f}_T^1(\omega)$ are consistent estimators of $f_T^0(\omega)$ and $f_T^1(\omega)$, respectively. □

With the help of lemmata 3.4 to 3.6, we can now construct the data driven market shrinkage estimator of the spectrum, which is given by the following theorem:

Theorem 3.7. The estimator

$$\zeta_T(\omega) = \frac{1}{m_T} \frac{p(\omega) - 2\Re\left(r(\omega)\right)}{g(\omega)} \tag{3.32}$$

is a consistent estimator of

$$\frac{1}{m_T} \frac{\pi(\omega) - 2\Re\left(\rho(\omega)\right)}{\gamma(\omega)}.$$

Proof. This is implied by assumption 3.1 in conjunction with lemmata 3.4, 3.5 and 3.6. □

Thus, we have finally arrived at a shrinkage estimator that depends on the data only, not on background knowledge of the underlying process:

$$\hat{f}_T^+(\omega) = \frac{p(\omega) - 2\Re\left(r(\omega)\right)}{m_T g(\omega)} \hat{f}_T^1(\omega) + \left(1 - \frac{p(\omega) - 2\Re\left(r(\omega)\right)}{m_T g(\omega)}\right) \hat{f}_T^0(\omega) \tag{3.33}$$

We will refer to this estimator as to the DDMSE (data driven market shrinkage estimator). The following theorem gives the asymptotic behavior of the DDMSE $\hat{f}_T^+(\omega)$:

Theorem 3.8. $\hat{f}_T^+(\omega)$ is a consistent estimator of the spectrum.

Proof. Asymptotically, the optimal shrinkage intensity $\zeta_T^*(\omega)$ vanishes according to 3.3. According to 3.7, $\zeta_T^*(\omega)$ is estimated consistently by $\zeta_T(\omega)$. This means that $\zeta_T(\omega)$ converges to zero, too, and thus that $\hat{f}_T^+(\omega)$ converges to the averaged periodogram. □

The performance of the DDMSE in practice will be examined by extensive Monte Carlo simulations in chapter 4.

CHAPTER 4

Simulations and applications

In this chapter, we will give the results of comprehensive Monte Carlo studies we have performed to examine the performance of the two shrinkage estimators we have developed in chapter 2 and 3, and also the performance of the decision theory based estimators that have been presented in chapter 1. Furthermore, we will give the results of application of our method to real data and the results of a study on the performance of the DDSSE in a classification setup.

The DDSSE in chapter 2 has been developed under a Kolmogorov asymptotic framework. This guarantees that, for sufficiently large T, it will give a result that is better than the averaged periodogram. However, we have no theoretical results that prove that, for finite sample size, it will grant a good estimate of the multivariate spectrum. Section 4.1 will, however, show that its performance, even for very small sample size, is exquisite.

The DDMSE in chapter 3 has been developed under a standard asymptotic framework. This means that, for large sample size, the shrinkage intensity will converge to zero. On the other hand, if the sample size is small, the shrinkage intensity may be badly estimated from the data. Thus, using the DDMSE is a 'race against time', as it will be identical to the averaged periodogram for large sample size, and there is no theoretical proof for it performing well for small sample size. However, section 4.2 will show that it performs well in practice.

Moreover, it will be interesting to compare the two estimators. The DDSSE is designed for a situation where there is no knowledge about the structure of the data. Its shrinkage target is the identity matrix,

which is the simplest and simultaneously most structured choice possible. The DDMSE is designed for a situation where a one-factor model may be of some value in approximating the data, without being the true model. This means that, for fixed frequency ω, there is a small number of eigenvalues of the true spectrum (and thus of the averaged periodogram) that are significantly higher than the rest. We will see in section 4.1 that the DDSSE performs well for all condition numbers, but best when the condition number is not extremely high. An extremely high condition number often indicates a high degree of collinearity in the data, and in this case, a parsimonious factor model is often a good approximation, and thus a good choice for a shrinkage target. Section 4.2 will thus also compare the two estimators with each other. It will be shown that the DDMSE outperforms the DDSSE in situations where a one-factor model may be not too crude an approximation, while the DDSSE still beats the benchmark, the averaged periodogram.
In section 4.3, we will use the DDSSE on a real data set from neuropsychology. In section 4.4 finally, we will perform an MC study on classification that shows the bad performance of conventional methods in classification of short multivariate time series and how to improve this by shrinkage.

4.1 Monte Carlo studies for the DDSSE

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.

4.1.1 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.

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 asymp-

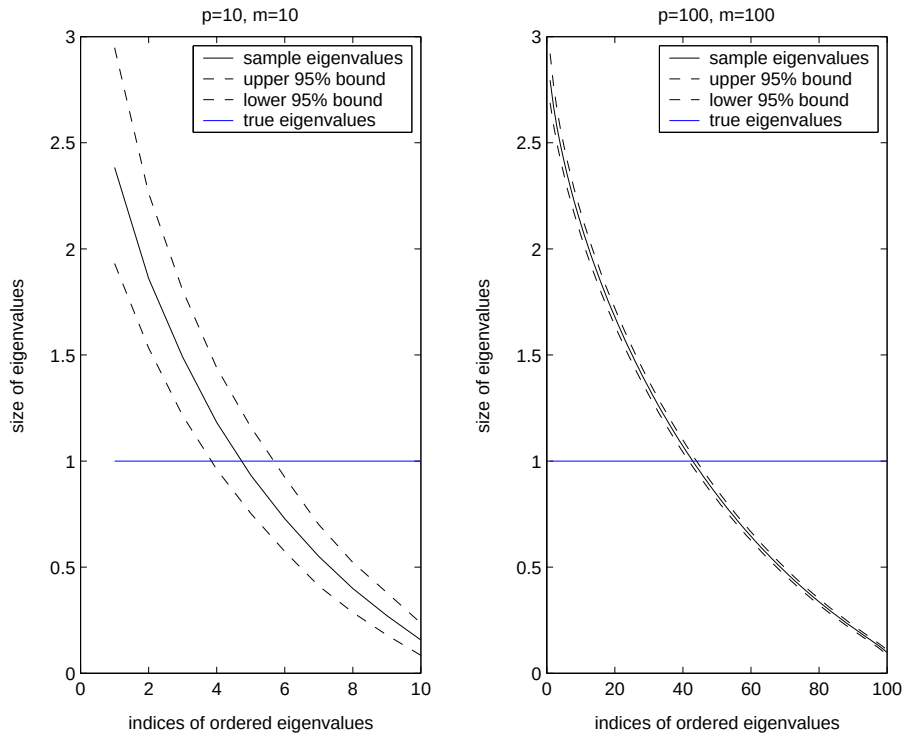

Figure 4.1: Median of sample eigenvalues at frequency $\pi/2$ for white noise processes of different dimensionality: on the left $p_T = 10$, on the right $p_T = 100$. In both subplots, the ratio p_T/m_T is the same. Based on a MC simulation (10,000 runs).

totics. Analogously to the classical framework, increasing the smoothing span enhances precision. Moreover increasing p_T and m_T simultaneously for fixed ratio p_T/m_T means that the confidence intervals for the sample eigenvalues become smaller, improving the precision of the estimators $\hat{\mu}_T(\omega), \hat{\alpha}^2_T(\omega), \hat{\beta}^2_T(\omega)$ and $\hat{\delta}^2_T(\omega)$. This does not mean, however, that the sample eigenvalues are good estimators for the true ones, as the bias depends on the ratio p_T/m_T. Figure 4.1 illustrates this effect.
The underlying process of the following simulations is a vector valued MA(2) time series with normal innovations:

$$X_t = \theta_0 e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}, \quad e_t \sim \mathcal{N}(0, \mathrm{Id}_5) \tag{4.1}$$

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

$$\theta_0 = (1.4072 \;\; 1.2207 \;\; 1 \;\; .7141 \;\; .1407)\,\mathrm{Id} \tag{4.2}$$

$$\theta_1 = (.3391 \;\; .2941 \;\; .2409 \;\; .1721 \;\; .0339)\,\mathrm{Id} \tag{4.3}$$

$$\theta_2 = -(.7750 \;\; .6723 \;\; .5508 \;\; .3933 \;\; .0775)\,\mathrm{Id} \tag{4.4}$$

The spectrum of the process (4.1) is a diagonal matrix function with condition number, at each frequency, equal to c. 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 4.2. 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. Furthermore, we have chosen equally shaped spectra for all dimensions here. This makes it easy in our simulations to obtain extreme true condition numbers - were the one dimensional spectra of different shapes, this would result in the condition number being lower. Moreover, we can obtain a high condition number uniformly over frequency, permitting to calculate the mean integrated square error for constant condition number. As, for a high condition number, the situation is more difficult than for a comparatively low one to improve upon, a good result of the DDSSE in this situation is more meaningful.

4.1.2 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 averaged periodogram $\hat{f}_T^0(\omega)$, DDSSE $\hat{f}_T^*(\omega)$ and oracle $\hat{\varphi}_T^*(\omega)$. We choose a smoothing span of $m = 7$, which is very small compared to the dimension $p = 5$, and the smallest possible without running into numerical problems. We vary the condition number c from 1 to 10^9. We expect that the DDSSE performs best for small condition number, as in this case the bias of the

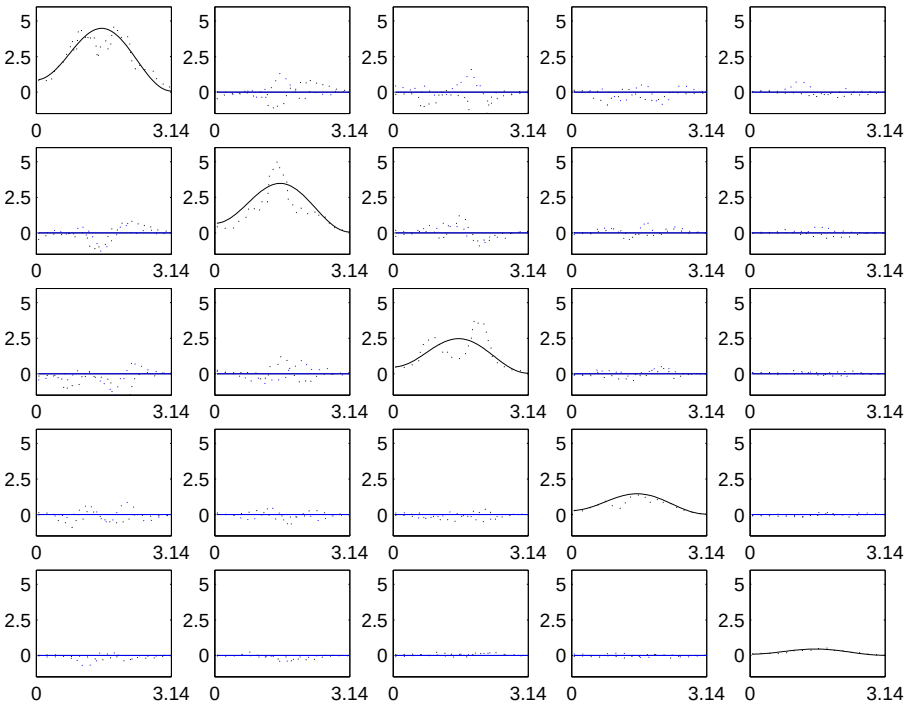

Figure 4.2: The spectrum (solid lines) and the averaged periodogram (dotted lines) with smoothing span $m_T = 9$ of the 5-dimensional process. The condition number is $c = 10$ here, the underlying time series has length $T = 128$. The real components are in black, the imaginary components in blue.

sample eigenvalues with respect to the true eigenvalues is maximal (Jolliffe 2002). Thus, it is particularly interesting to examine its asymptotic behavior for $c \to \infty$. We present the results graphically in figure 4.3; for $c = 10$, we also give a graphical representation in figure 4.4, 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 4.3. For condition number greater than 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

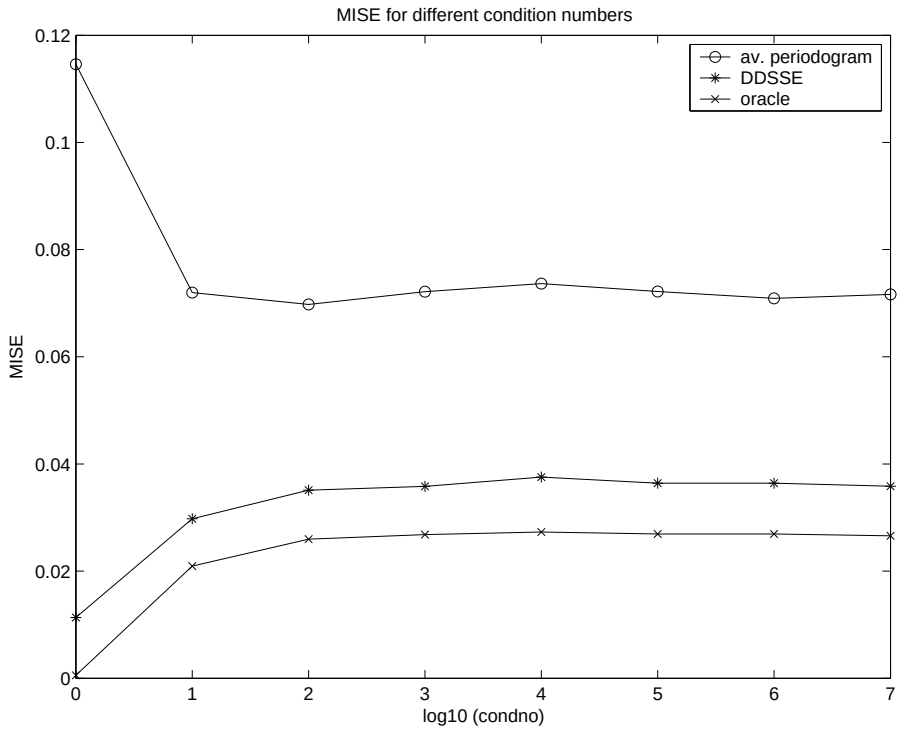

Figure 4.3: MISE as a function of condition number for the averaged periodogram, DDSSE and oracle. Based on M=1,000 Monte Carlo observations. Confidence intervals are not printed as there is no intersection at 99% level.

$c = 10$, 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. 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 (appr. 50%).
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.
We have not plotted Monte Carlo confidence intervals in this and the following figures, as the confidence intervals are very narrow. This is not only due to the Monte Carlo sample size of $M = 1,000$, but also, more, due to the fact that, except for figure 4.4, we average over frequency. Next, we look at figure 4.4, which shows on one hand the bias-variance

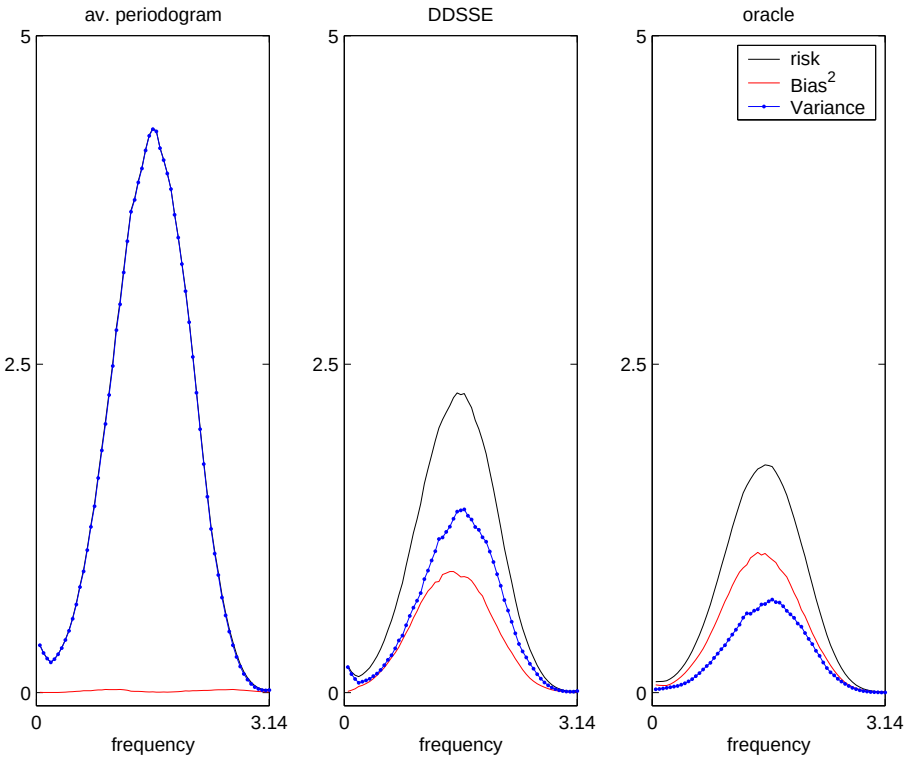

Figure 4.4: Risk, squared bias and variance as a function of frequency for the averaged periodogram, the DDSSE and the oracle. Based on M=1,000 Monte Carlo observations.

decomposition of 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.
Figure 4.4 is also useful in helping the reader to develop intuition on the narrowness of the confidence intervals in this chapter. The risk of the averaged periodogram is proportional to the trace of the true spectrum. In figure 4.4, the shape of the estimated risk for the averaged periodogram mimics the shape of the trace of the true spectrum, as seen in figure 4.2, with high precision. As the estimates of the risk at frequencies that are apart by more than the smoothing span are approximately independent, both the smoothness of the curve and the similarity between the shapes of the estimated risk and the true spectrum indicate that the estimate is highly precise. In all other figures, we have integrated over frequency, enhancing precision even further.

4.1.3 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 4.5 that the optimal 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 peri-

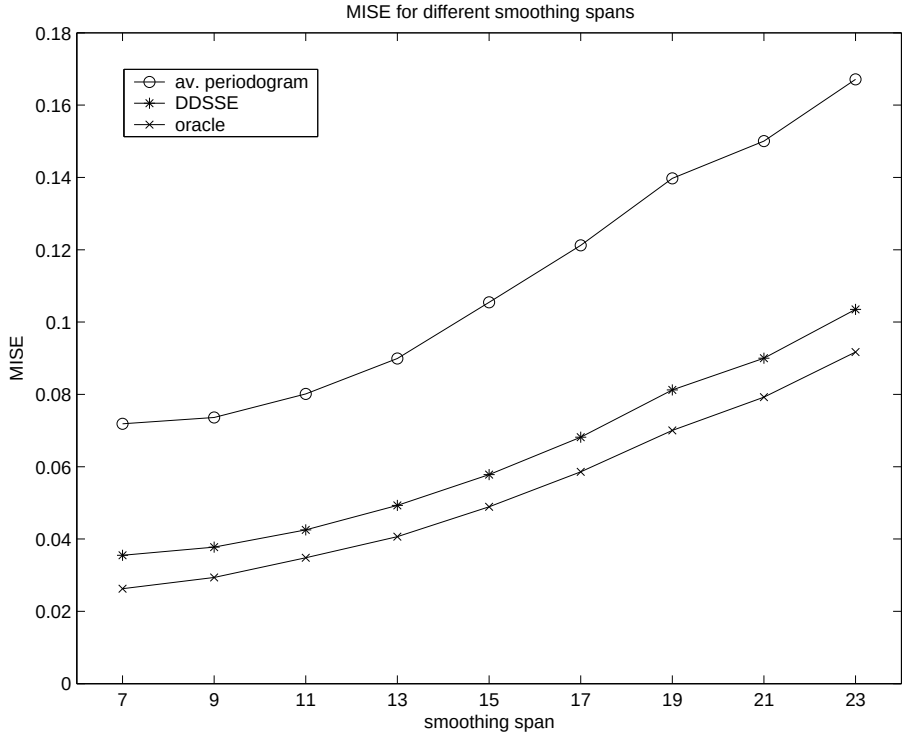

Figure 4.5: MISE of averaged periodogram, DDSSE and oracle as a function of smoothing span, for a time series of length $T = 128$. Based on M=1,000 Monte Carlo runs. Confidence intervals are not printed as there is no intersection at 99% level.

odogram 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

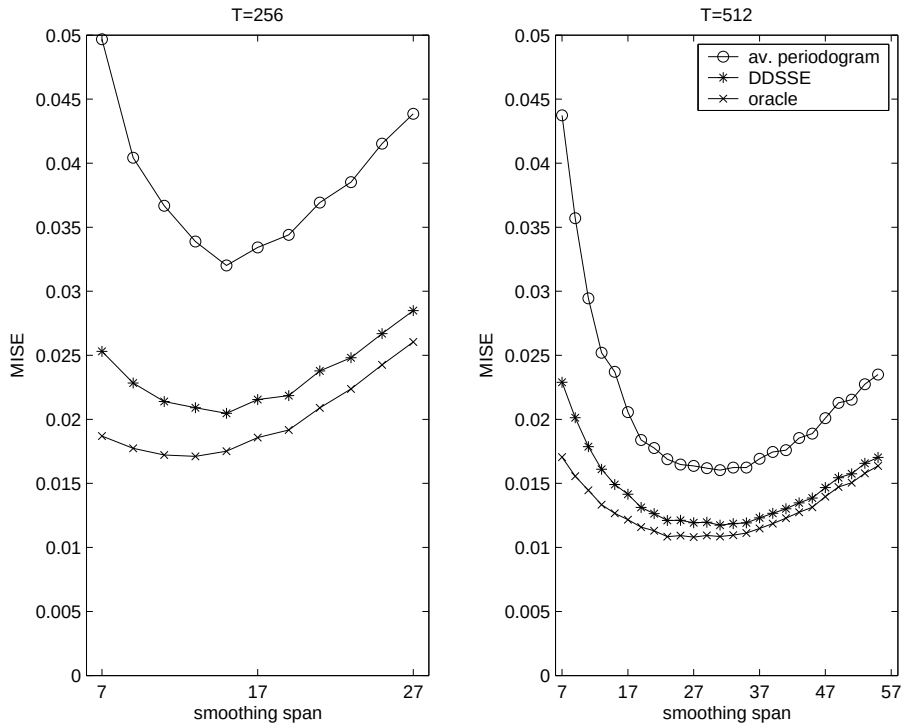

Figure 4.6: MISE of averaged periodogram, DDSSE and oracle as a function of smoothing span, for time series of different length. Based on M=1,000 Monte Carlo runs. Confidence intervals are not printed as there is no intersection at 99% level.

have performed additional MC runs on the same time series as in (4.1), 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 4.6. First, we remark that the DDSSE never has higher risk than $\hat{f}_T^0(\omega)$. 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 $\hat{f}_T^0(\omega)$, $\hat{\varphi}_T^*(\omega)$ 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 (Ombao, Raz, von Sachs & Strawderman 2001) and then replace the estimator by the DDSSE.

4.1.4 Shrinkage for tapered data

The bias of the averaged periodogram can be calculated as

$$\mathrm{E}\, I_T(\omega) = \int_{-\pi}^{\pi} f(\omega - \alpha) K_T(\alpha) d\alpha \tag{4.5}$$

where

$$K_T(\alpha) = \frac{1}{2\pi T} \frac{\sin^2(T\alpha/2)}{\sin^2(\alpha/2)} \tag{4.6}$$

is the Fejér kernel. This kernel has side peaks of magnitude $\mathrm{O}\,(1/T)$. If, in equation (4.5), one of these side peaks is multiplied with a peak in the true spectrum f, the result is a possibly too large expectation at this frequency. Consequently, other, lower peaks can be superimposed and will possibly not be discovered. This phenomenon is referred to as *spectral leakage*.

To reduce this effect, it is common to *taper* the data (Dahlhaus 1983, 1988). Data tapering is a nonparametric method where the data record is multiplied by a function that is near zero at the beginning and the end of the record and near one in the middle. It was first introduced by Tukey (see Brillinger 2002). The expectation of the periodogram of the tapered data is again the convolution of the true spectral density with a kernel, like in (4.5). However, the tapered kernel has a faster rate of convergence to zero at frequencies $\alpha \neq 0 (\mathrm{mod}\ 2\pi)$. This reduces spectral leakage.

We have empirically examined the influence of data tapering on our shrinkage estimator. For this, we have applied the polynomial data taper

$$\begin{aligned} h_\rho(x) = &\quad 16(x/\rho)^2(1 - x/\rho)^2, && x \in [0, \rho/2), \\ = &\quad 1, && x \in [\rho/2, 1/2], \\ = &\quad h_\rho(1 - x), && x \in (1/2, 1] \end{aligned}$$

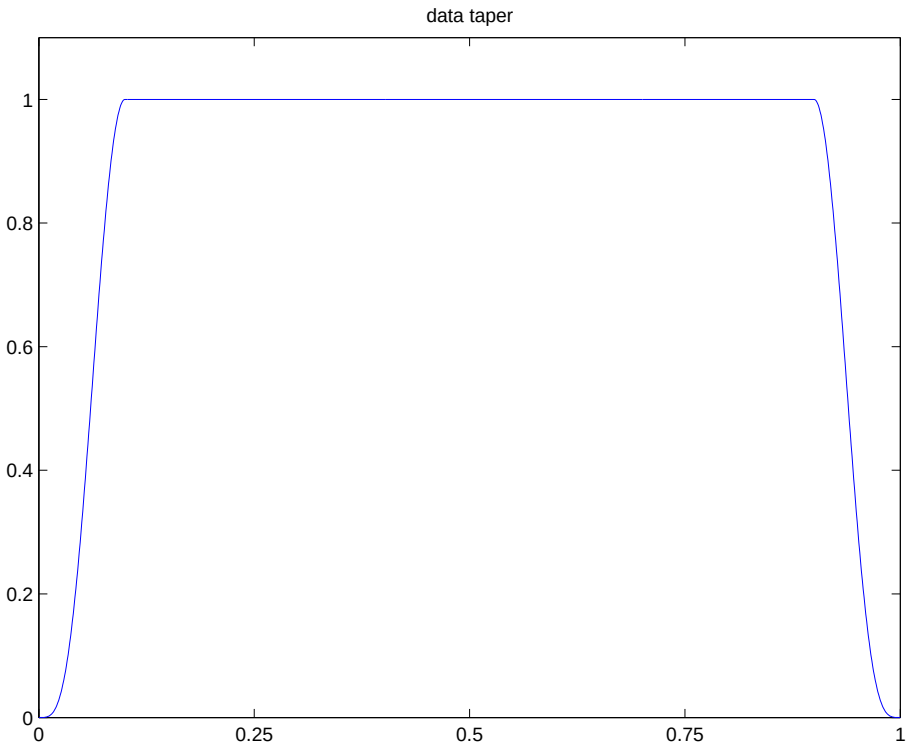

Figure 4.7: Polynomial taper function of order 2, $\rho = 0.2$, that is applied to the data

for a choice of $\rho = 0.2$. The taper function is plotted in figure 4.7. We have repeated the simulations of the preceding subsection, for a time series of length $T = 256$, and this time applied the data taper. The results are given in 4.8. The left side of the figure shows the results for untapered data, the right side the results for tapered data. The taper improves the performance of the averaged periodogram by approximately 15%. The improvement on the DDSSE induced by tapering is much smaller, about 5%, the effect being approximately the same for all smoothing spans. The oracle seems not to be improved upon. These results seem to indicate that shrinking to identity works well for tapered data, too. Although the effect of tapering on the DDSSE is smaller than the effect of tapering on the averaged periodogram, the DDSSE for tapered data is again much more precise than the averaged periodogram

for tapered data.

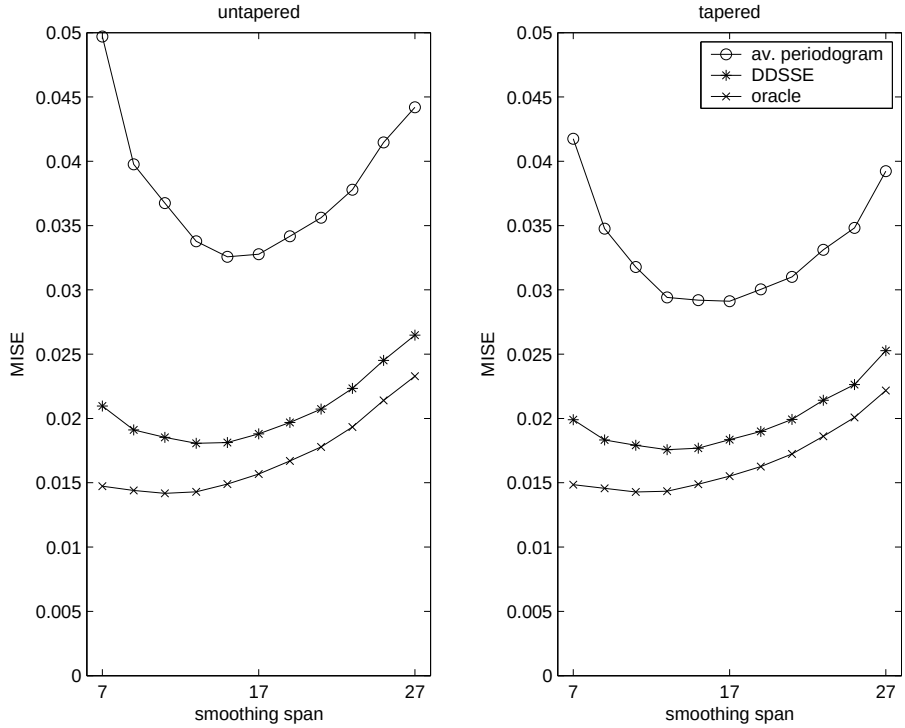

Figure 4.8: MISE of averaged periodogram, DDSSE and oracle for tapered and untapered data. Based on M=1,000 Monte Carlo runs of a time series of length T=256. Confidence intervals are not printed as there is no intersection at 99% level.

4.1.5 Estimators based on decision theory

A decision we had to make during this work was whether we should base our new estimators of the multivariate spectrum on decision theory or rather on asymptotic theory. The decision against the former was partly based on the better performance of the DDSSE in studies we performed before we developed the rigorous framework. For this, we derived the DDSSE in its present form and compared it to estimators of the spectrum that are *ad hoc* generalizations of estimators from the literature based on decision theory. Each of these estimators was adopted by sim-

ply using the smoothing span as the effective sample size. Figure 4.9 gives the results of a study that is based on a 5-dimensional Gaussian white noise series. We have chosen white noise here because, as discussed above, it makes no difference to a 'real' time series except in one aspect: there is no oversmoothing in the case of white noise, and thus the effect of increasing the smoothing span can be studied. The estimators we compare the DDSSE with are the constant risk minimax estimator derived by James & Stein (1961), a minimax estimator from Dey & Srinivasan (1985) that dominates the constant risk estimator but uses no information from the actual realization ("m") and another estimator from the same paper that uses the eigenvalues of the sample covariance matrix to determine the amount of shrinkage ("S"). We see that the constant risk estimator improves only slightly upon the averaged periodogram. The two other estimators offer significant improvement, but the DDSSE shows the best performance. If the true condition number is increased, the relative improvement of all shrinkage estimators is less good, as discussed above, but the order of the performance of the five estimators is preserved.

4.2 Monte Carlo studies for the DDMSE

In this section, we will evaluate the performance of the data driven market shrinkage estimator in practice. For this, we will again perform comprehensive Monte Carlo simulations. The DDMSE will have three benchmark estimators to compete with:

1. the averaged periodogram
2. the one-factor model that is the shrinkage target
3. the DDSSE

In a setting where it is reasonable to use the DDMSE, it should outperform all three benchmarks. Such a setting can be characterized as the frequently encountered situation where one may fit a factor model to the data, but has no background knowledge on how many factors to actually choose. In a screeplot of the eigenvalues, one will typically encounter one or more prominent eigenvalues followed by a longer tail of small eigenvalues. The method developed in chapter 3 will allow us to completely avoid the problem of model choice.

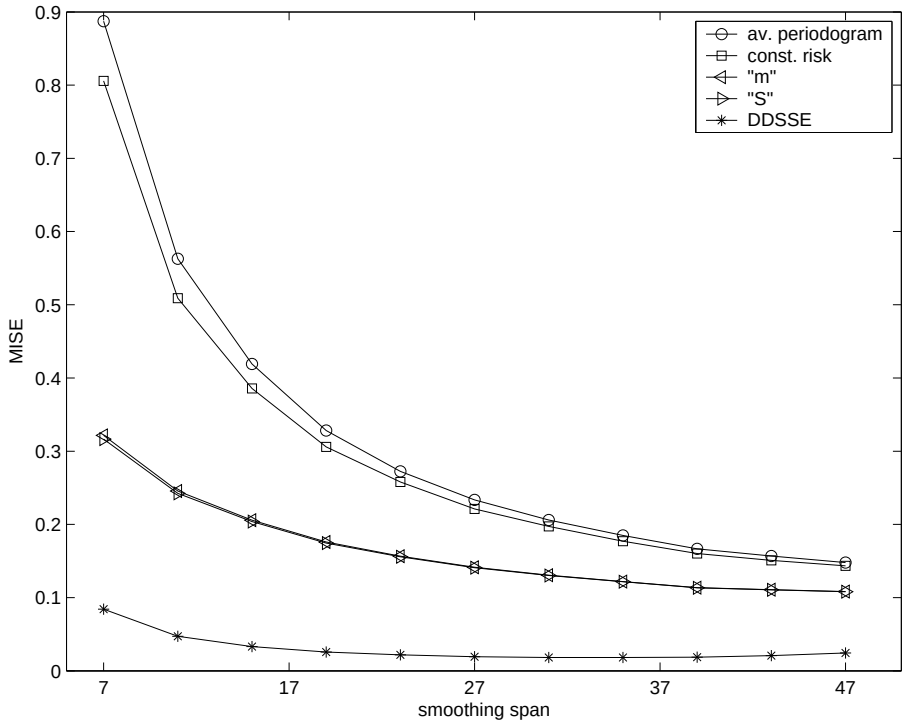

Figure 4.9: Comparison of shrinkage estimators based on asymptotic and decision theory. "m" refers to the estimator based on $\hat{\Sigma}^m$, "S" to the estimator based on $\hat{\Sigma}^S$. Based on $M = 1,000$ Monte Carlo observations of a 5-dimensional Gaussian white noise time series, with condition number $c = 3$ uniformly over frequency.

4.2.1 Setup

For the simulations, we have chosen to use a two-factor model as the true model. The first factor is an MA(2) process. Its spectrum has a peak at $\pi/2$. The second factor driving the process is a Gaussian white noise time series; its variance will be varied in a first simulation study, to examine the performance of the DDMSE on the 'scale' between almost one-factor model to true two-factor model. Figure 4.10 shows the spectrum of the two factors underlying the simulations. These two factors are then projected onto a 5-dimensional time series according to

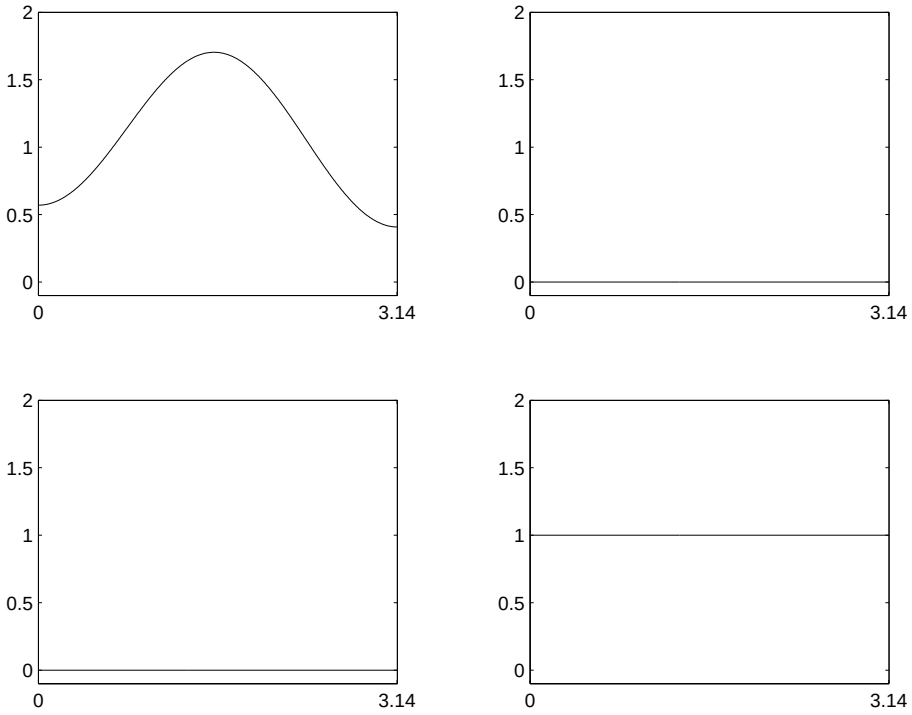

Figure 4.10: True spectrum of the two underlying factors. The imaginary parts are all zero.

the following model:

$$X_t = \Upsilon f_t + \epsilon_t, \qquad \epsilon \sim \mathcal{N}(0, \Omega) \tag{4.7}$$

Here, Υ is a 5×2 weight matrix that was chosen at random initially, then fixed for this section. The initial random distribution for the components of Υ was uniform $\sim U([.3, 1])$, the components being chosen independently.

$$\Upsilon = \begin{pmatrix} .5871 & .4510 \\ .5676 & .9691 \\ .4645 & .7268 \\ .8691 & .5511 \\ .5379 & .4754 \end{pmatrix}$$

The covariance matrix of the idiosyncratic components was obtained likewise: the off-diagonal components were set to zero, the diagonal components were simulated as iid uniform $\sim U([.2, .4])$ and then fixed as

$$\Omega = \text{diag}(.3213 \quad .3726 \quad .2646 \quad .4169 \quad .3257)$$

The market factor time series was obtained as the mean over dimension of the simulated data. All simulations presented in this chapter were repeated for new realizations of $\{\Upsilon, \text{Cov}(\Omega)\}$ without any major changes in the results, which is why we will omit these repetitive studies. A length of $T = 1,024$ was chosen for the time series in this section.

4.2.2 Influence of the true model

The only formal prerequisite for the true model in order for the DDMSE to work is that its true spectrum is not that of a one-factor model like specified in 3.3. In this subsection, we will examine the influence of the 'distance' from a one-factor model. This is accomplished by using the two-factor model (4.7) to generate the data and systematically varying the standard deviation of the second, flat-spectrum factor. For small standard deviation, the data are very close to a one-factor model; as the standard deviation of the second factor increases, so does its influence. The results are given in figure 4.11. The effects we observe in the simulations study confirm our assumptions on the respective behavior of averaged periodogram, one-factor model, DDSSE and DDMSE. First of all, we remark that the DDMSE performs best for all choices of white noise variance in the simulations. The averaged periodogram, upon which we want to improve, exhibits the worst performance. Not only is it outperformed by the DDSSE, which we would have expected based on the results of the preceding section, but also by the one-factor model. This shows that, in this context, the one-factor model is a useful model in itself, even although it is actually misspecified. It even outperforms the DDSSE for most choices of white noise variance. Overall, the MISE increases with the variance of the second factor, and the different estimators follow the MISE in a parallel shape.

4.2.3 Influence of the smoothing span

In the next Monte Carlo study, we have varied the smoothing span and examined its influence on the MISE. The results are given in figure

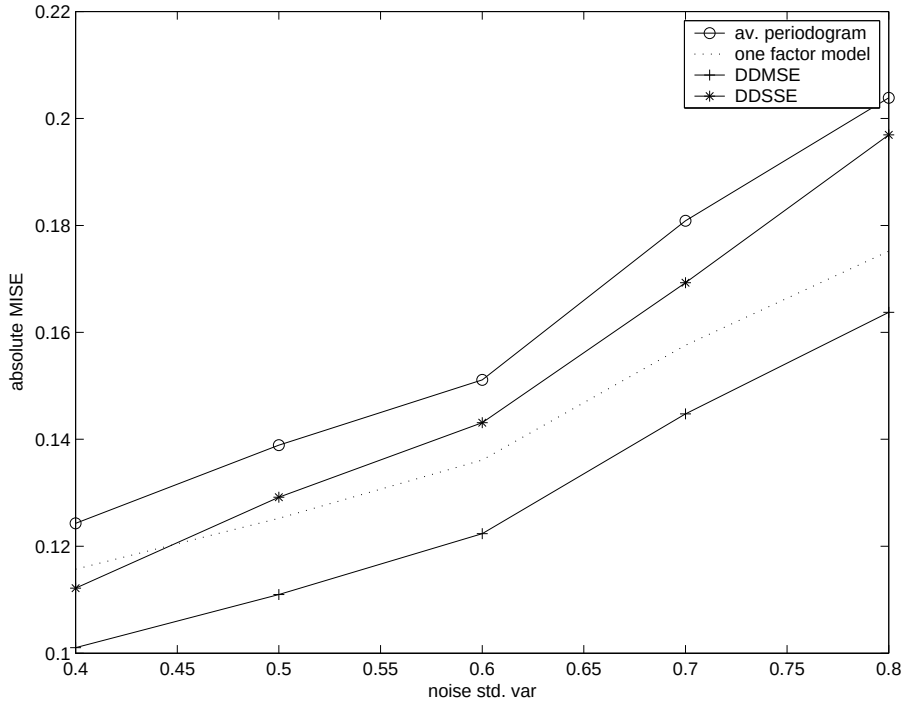

Figure 4.11: MISE of DDMSE, averaged periodogram, 1-factor-model and DDSSE for data from a 2 factor model. T=1,024, smoothing span =19, different standard deviations of second factor. Based on M=1,500 Monte Carlo runs. Confidence intervals are not printed as there is no intersection at 99% level for the solid curves.

4.12. Not surprisingly, we observe that the overall MISE decreases as the sample size is increased for all three estimators. For small smoothing span, the averaged periodogram exhibits the worst performance. The DDSSE performs better than the averaged periodogram for small smoothing span, but is outperformed by the one-factor model and by the DDMSE. For the very small smoothing span $m_t = 7$, the DDMSE and the one-factor model have approximately the same MISE. Then, we have again the ranking averaged periodogram-DDSSE-one-factor model-DDMSE, as in the preceding subsection. Finally, for a comparatively large smoothing span of 31 or more, the DDMSE, DDSSE and averaged periodogram seem to have approximately the same MISE. This

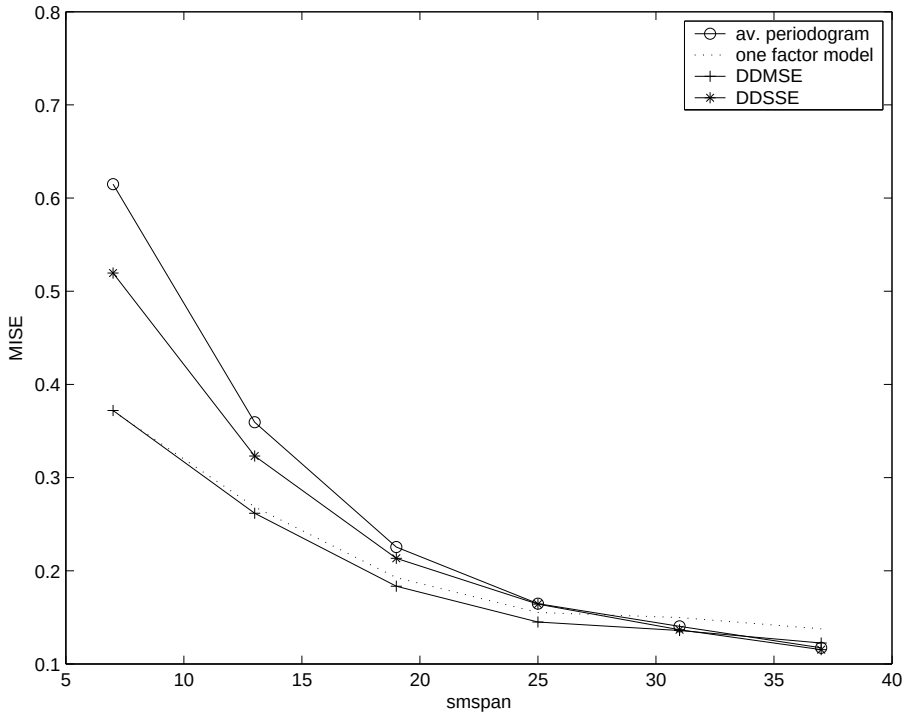

Figure 4.12: MISE of DDMSE, averaged periodogram, 1-factor-model and DDSSE for data from a 2 factor model. T=1,024, different smoothing spans. Based on M=1,500 Monte Carlo runs.

is again not surprising, as for fixed dimension, both data driven estimators converge to the averaged periodogram. Moreover, for large smoothing span, the one-factor model performs worse than the averaged periodogram. This is, however, not due to a loss of performance of the one-factor model, which improves monotonously with m_T, but rather due to the faster improvement of the averaged periodogram in terms of MISE. The fact that the worse improvement of the one-factor model for large smoothing span with respect to the averaged periodogram does not result in the DDMSE exhibiting worse performance than the averaged periodogram can be interpreted as resulting from the fact that, for large m_T, the shrinkage intensity becomes negligibly small.

4.3 A real data example

In this section we will apply the DDSSE to a real multivariate dataset. An electroencephalogram (EEG) was taken from a patient prior to and during an epileptic seizure. An EEG is the neurophysiologic measurement of the cumulative electrical activity of the brain. It is recorded by placing electrodes on the scalp. The signal consists of the sum of the post-synaptic potentials of a large number of neurons situated near the surface of the brain. EEG is widely used in both clinical neurology and research.

The large-scale patterns of synchronized neuronal activity constituting the EEG are changing continuously. This is a result well-known since the beginnings of psychophysiology. However, in practice this has been widely ignored until recently. Researchers have treated the signal as though it were stationary and have applied various smoothing and averaging techniques to eliminate non-stationarity. Under such crude treatment, many of the highly informative features of the raw EEG signal are usually lost. Progress made in both computer sciences and statistics (Guo et al. 2005) has now made it possible to analyze the non-stationary features of the EEG, permitting scientists to relate the non-stationarity of the EEG to underlying mechanisms of brain functioning. As EEG signals are multivariate, we believe that the shrinkage techniques developed in this thesis are an important contribution to EEG analysis, as they permit a more precise estimate of the multivariate spectrum of EEG data based on very small sample sizes, and thus a better resolution.

4.3.1 The dataset

The dataset consists of 10 channels of EEG data recorded from a patient at a sampling rate of 100 Hz. The data were recorded directly before and during an epileptic seizure. Figure 4.13 shows the data. The data are highly non-stationary. Thus, it does not make sense to try to retrieve a global estimate of the spectrum. We will take a small piece of 256 data points from the beginning of the measurements and perform spectral analysis using both the averaged 10 dimensional periodogram and the DDSSE.

4.3.2 Spectral analysis

Figure 4.14 shows the 256 data points we have used for our analysis. We will examine how the eigenvalues of the two estimators differ. We have

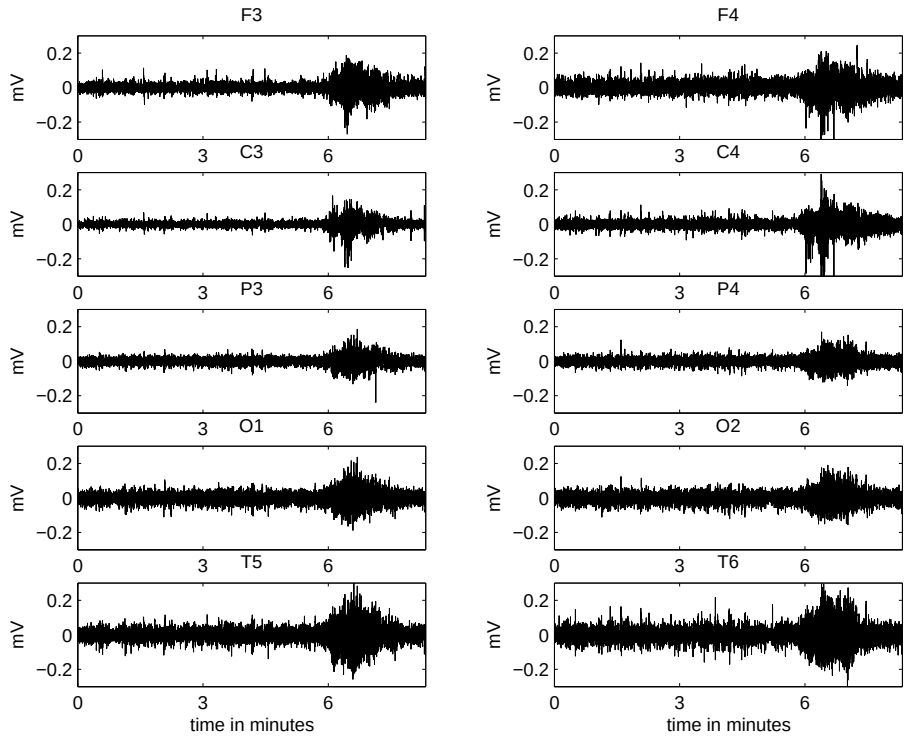

Figure 4.13: 10 channels of an EEG before and during an epileptic seizure. The abbreviations 'F', 'P', 'O' and 'T' refer to the frontal, parietal, occipital and temporal lobe of the brain, 'C' to a central position. The odd numbers mark recordings from its left, the even numbers recordings from its right side.

chosen a smoothing span of $m = 11$ here, which will enable us to capture even high frequency features of the EEG signal. In figure 4.15, we see the influence of shrinkage on the eigenvalues. For this, we have plotted the three largest eigenvalues without and with shrinkage as well as the three smallest eigenvalues. The largest and second largest eigenvalue are both shifted down. The third largest eigenvalue is left almost unaltered and seems to mark the boundary between eigenvalues that are shifted downwards and eigenvalues that are shifted upwards. The three smallest eigenvalues are all shifted upwards and are, after shrinkage, almost indistinguishable.

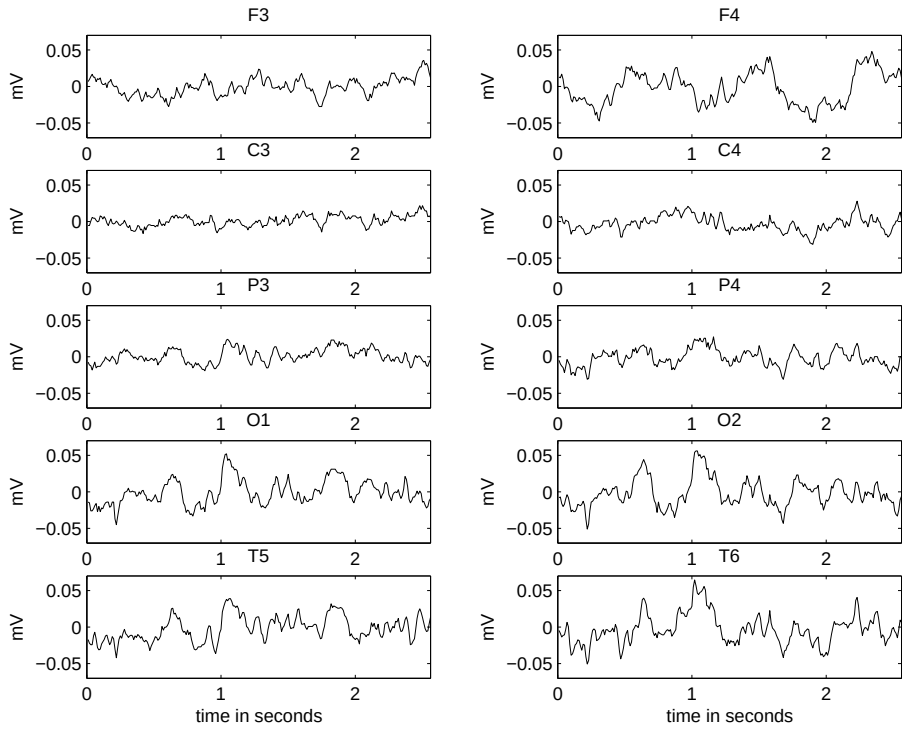

Figure 4.14: Zoom of EEG data from figure 4.13. The abbreviations 'F', 'P', 'O' and 'T' refer to the frontal, parietal, occipital and temporal lobe of the brain, 'C' to a central position. The odd numbers mark recordings from its left, the even numbers recordings from its right side.

It is obvious from figure 4.15 that the condition number is improved upon. The improvement is large: the maximal condition number over frequency is $7,439$ for the averaged periodogram (minimum: 15.31). This maximum is attained for the frequency $\omega = 0.024$. For the data driven spectral shrinkage estimator, the maximal condition number is 33.18 (minimum: 2.035), and this maximum is attained for the frequency $\omega = 1.006$. This shows that, at least as far as numerical stability is concerned, we do obtain a (more than 200 times) better estimate of the spectrum in this real data example. As we do not know the true spectrum, we cannot prove that the DDSSE is, in this case, more precise. However, we believe that further research, especially on using the

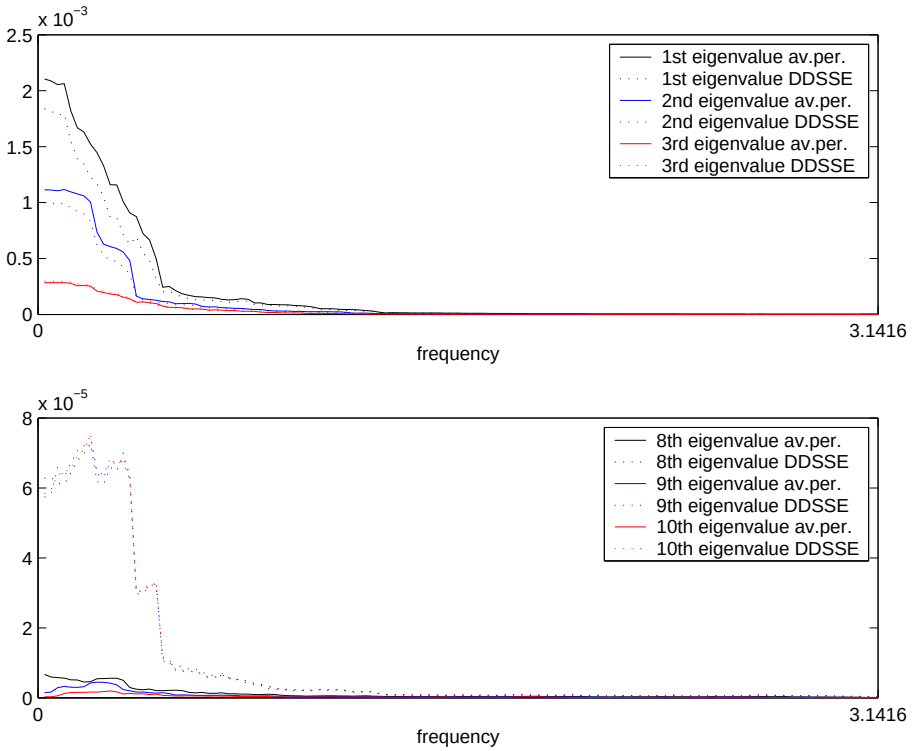

Figure 4.15: Comparison of unshrunken eigenvalues (corresponding to averaged periodogram) and shrunken eigenvalues (corresponding to DDSSE).

shrinkage methods developed in this thesis in the context of finding the optimal segmentation of a non-stationary time series, like in subsection 1.2.2, will show that using the DDSSE is indeed superior to using the averaged or smoothed periodogram without regularization in treating real multivariate time series data.

4.4 An application to classification

In this last section we will present an MC study that shows how classification of multivariate time series fails when conventional methods of spectral analysis are applied, and how this problem can be solved by shrinkage.

4.4.1 Setup

Our aim is to classify short multivariate stationary time series into one of two categories $C1$,$C2$. The inference is based on the Kullback-Leibler discrimination information (1.9). First, we specify the true spectral density for a time series from category $C1$ and for a time series from category $C2$. We choose the spectrum of $C1$ to be the spectrum of the 5-dimensional MA(2) process given in figure 4.2. The spectrum has a peak near $\omega = \pi/2$, and the condition number is uniformly over frequency $c = 10$. The spectrum of $C2$ is chosen to differ from the spectrum of $C1$ only with respect to the condition number, which for $C2$ is $c = 1$ uniformly over frequency. At each frequency, the trace of the spectra of $C1$, $C2$ are equal.
In practice, it will seldom if ever occur that the true spectra are known. Rather, the data analyst will have one or multiple instances of time series of which he knows that they belong to class $C1$ or class $C2$, and will base his classification on estimates of the spectra based on these realizations. This is a crucial point. Consider the Kullback-Leibler discrimination information between the estimated spectrum of an instance X of a time series and the spectrum of class $C\cdot$, which is:

$$\mathrm{KL}(\hat{f}_X, f_{C\cdot}) = \frac{1}{2T}\sum_{t=1}^{T}\left(\mathrm{tr}\left[\hat{f}_X(\omega_t)f_{C\cdot}^{-1}(\omega_t)\right] - \ln\frac{\det(\hat{f}_X(\omega_t))}{\det(f_{C\cdot}(\omega_t))} - p\right) \tag{4.8}$$

If the true spectrum $f_{C\cdot}(\omega)$ is not known, but has to be replaced by an estimate $\hat{f}_{C\cdot}(\omega)$, the estimate is based on

$$\mathrm{KL}(\hat{f}_X, \hat{f}_{C\cdot}) = \frac{1}{2T}\sum_{t=1}^{T}\left(\mathrm{tr}\left[\hat{f}_X(\omega_t)\hat{f}_{C\cdot}^{-1}(\omega_t)\right] - \ln\frac{\det(\hat{f}_X(\omega_t))}{\det(\hat{f}_{C\cdot}(\omega_t))} - p\right) \tag{4.9}$$

instead of equation (4.8). The consequence is that we suffer severe damage if the estimator $\hat{f}_{C\cdot}(\omega)$ is badly conditioned. As this is just what we want to examine in this Monte Carlo study, we will therefore choose the following design:
First, we will simulate two reference time series $Y_{C1} \in C1$ and $Y_{C2} \in C2$. These time series are assumed to be known to belong to class $C1$ or $C2$,

respectively.
Next, averaged periodograms $\hat{f}^0_{C.}(\omega)$ with smoothing span m are calculated for the two classes as well as DDSSEs $\hat{f}^*_{C.}(\omega)$ with the same smoothing span.
In the third step, 500 time series $X \in C1$ and 500 time series $X \in C2$ are simulated, and for each a smoothed periodogram $\hat{f}^0_X(\omega)$ with smoothing span m is calculated.
In the last step, the time series are classified into C_1 if $\mathrm{KL}(\hat{f}^0_X, \hat{f}^0_{C1}) < \mathrm{KL}(\hat{f}^0_X, \hat{f}^0_{C2})$, otherwise they are classified into $C2$. A second classification is done based on the DDSSE: the time series are classified into C_1 if $\mathrm{KL}(\hat{f}^0_X, \hat{f}^*_{C1}) < \mathrm{KL}(\hat{f}^0_X, \hat{f}^*_{C2})$, otherwise they are classified into $C2$. For both classification methods, the false classification rate (FCR), defined as

$$\mathrm{FCR} = \frac{\text{no. of false classifications}}{\text{no. of time series to classify}} \tag{4.10}$$

is computed.

The results are highly dependent on the particular realizations Y_{C1} and Y_{C2}. Therefore, the simulation study will be repeated multiple times: for each of the smoothing spans $m = 7, m = 15, m = 23$, the simulation study is repeated $M = 100$ times. The results are given in figures 4.16 to 4.18. In figure 4.16, where the smoothing span is only $m = 7$, the average false classification rate is as high as 0.502 when classification is based on the averaged periodogram. This means that the classification algorithm, which is exactly the algorithm described in the book by Shumway & Stoffer (2000), fails completely, as we would obtain the same false classification rate by basing classification on the tossing of a coin. By replacing the averaged periodogram with the DDSSE it is possible to obtain a false classification rate that is only = 0.056, i.e. approximately 19 out of 20 time series are classified correctly.
The false classification rate is highly dependent on the smoothing span chosen. This is evident from figures 4.17 and 4.18. For smoothing spans of $m = 15$ and $m = 23$ the false classification rate is, on average, for the averaged periodogram 0.031 and 0.028, respectively, and for the DDSSE 0.002 and 0.001, respectively. The particular shape of the spectrum chosen in this example allows for a large smoothing span to be chosen. However, in practice the spectrum of at least one of the classes $C1, C2$ may be peaky, in case of which a small smoothing span *has* to be chosen.

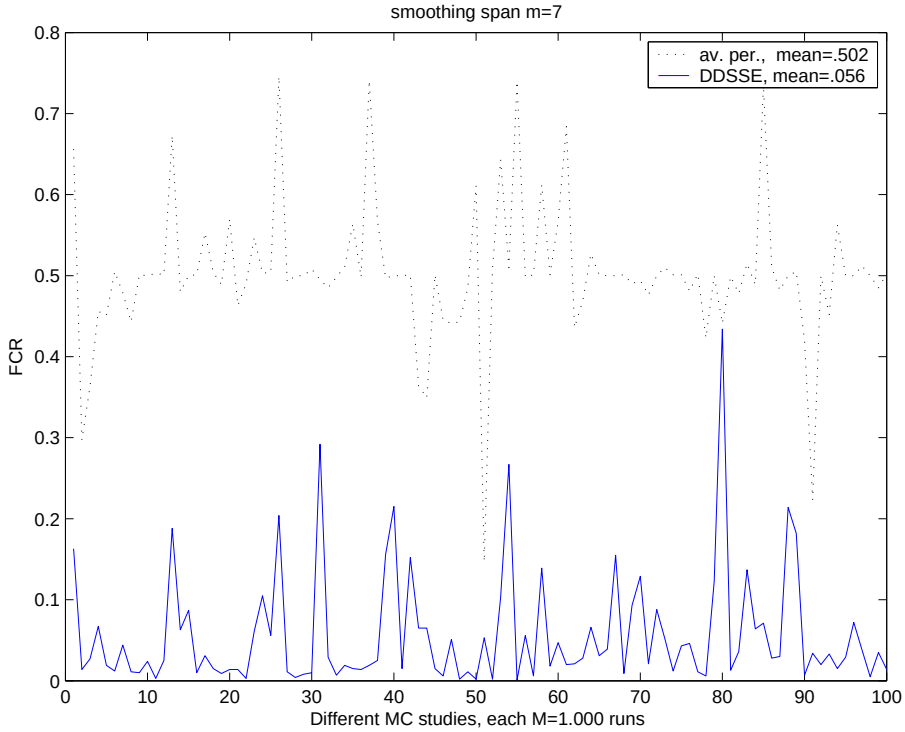

Figure 4.16: Multiple replications for false classification rate for smoothing span m=7. Each of the points in the plot corresponds to the estimated false classification rate, based on 2×500 time series each.

In this case, it seems to be indispensable to shrink. Furthermore, the results seem to indicate that only in few cases the false classification rate is amplified by shrinkage. However, there are such cases, like in figure 4.18, MC-runs 59 and 61, where the false classification rate is higher for the DDSSE than for the averaged periodogram. This may be due to an outlier in the realizations Y_{C1}, Y_{C2}, in case of which the estimator based on shrinkage may, in principle, be further away from the true spectrum than the estimator based on the averaged periodogram.

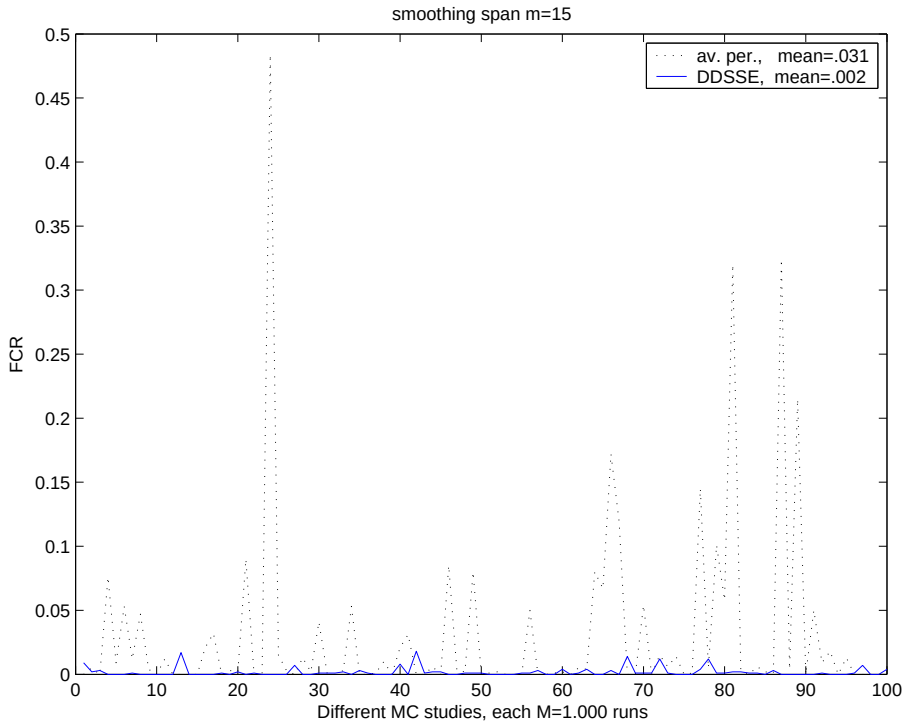

Figure 4.17: Multiple replications for false classification rate for smoothing span m=15. Each of the points in the plot corresponds to the estimated false classification rate, based on 2×500 time series each.

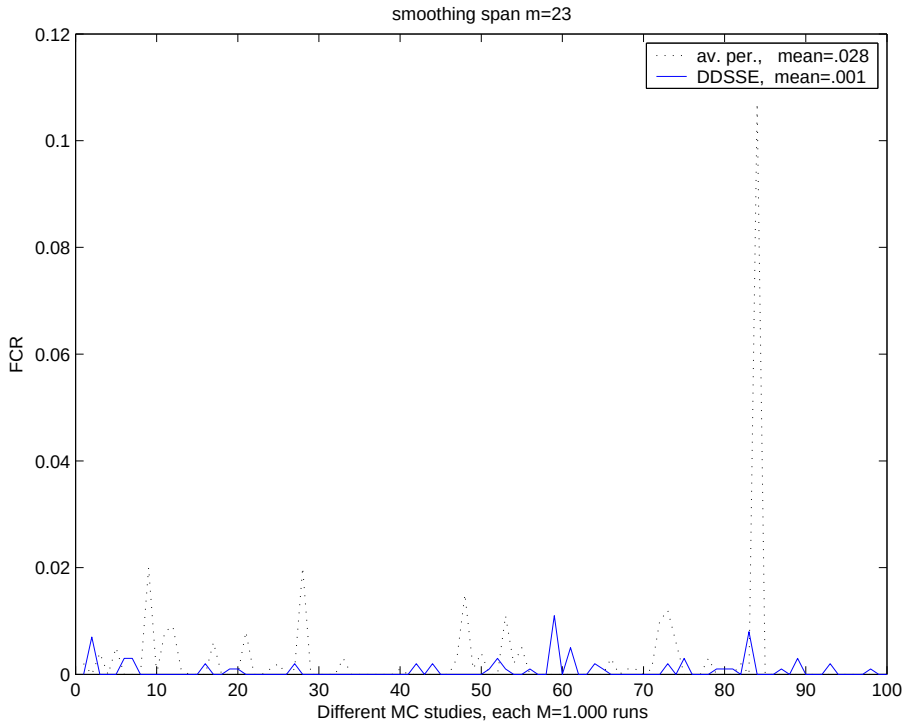

Figure 4.18: Multiple replications for false classification rate for smoothing span m=23. Each of the points in the plot corresponds to the estimated false classification rate, based on 2×500 time series each.

CHAPTER 5

Conclusions

5.1 Contributions

To the best of our knowledge, this thesis presents the first systematic approach to the problems related to ill conditioning in multivariate spectral analysis. In chapter 1, we have aroused sensitivity for this important aspect of multivariate spectral analysis. We have shown how different areas such as classification, clustering, discrimination, and high dimensional non-stationary time series analysis, are affected by this problem. We have furthermore discussed three different approaches, shrinkage, ridge regression and sparse PCA, that have been developed to cope with this problem in case of iid data, and worked out the close relationship between the three.

The main contribution of our work is a new, localized concept of shrinkage that allows for the development of estimators that simultaneously overcome the problem of numerical instability due to high dimensionality or collinearity and have, asymptotically and for finite sample size, lower quadratic risk. Two such estimators have been developed.

In chapter 2, the estimator is a model free convex combination of a nonparametric estimator, the averaged periodogram, and a parametric estimator, $\mu_T(\omega)$ times the identity matrix. We use the interesting framework of general asymptotics, under which we show the averaged periodogram not even to be consistent. Also, we have highlighted two important interpretations of shrinkage. First, it can be understood as a method of trading the asymptotic unbiasedness of the averaged periodogram for a decrease in variance, where the shrinkage intensity de-

rived attains asymptotically the optimal, balancing value. Second, it is a method that minimizes an estimated risk. However, the risk is not estimated directly, but via the triangle $\mu_T(\omega)\,\mathrm{Id} - \hat{f}_T^0(\omega) - f_T^0(\omega)$, the hypotenuse and one of the catheti of which are estimated; the estimate is completed by imposing rectangularity on the triangle.
We have demonstrated the possible flexibility in the choice of the underlying asymptotic framework by showing that it is both possible to develop shrinkage under a rigorous general asymptotic framework or under classical asymptotics, the latter case being more easy to develop. In chapter 3, we also extend the scope of our theory by incorporating background knowledge on the underlying cross dimensional structure of the data in our shrinkage target. This, in conjunction with the relative simplicity in deriving the asymptotically optimal shrinkage intensity under classical asymptotics, opens up the way to designing 'custom made' shrinkage estimators that offer a new answer to problems of model choice. In a given setting where a class of parametric or semi-parametric estimators is eligible, and the order has to be chosen, instead of relying on criteria such as AIC or BIC, the minimum order model can be used as a target towards which to shrink. Instead of calling the method shrinkage, we might as well describe it as *stretching*: a too parsimonious model is fitted and the estimate is then refined by adding the periodogram as a stretching target that has low bias and high variance.
Whereas chapters 2 and 3 provide the theoretical background, the comprehensive Monte Carlo simulations in chapter 4 show the large gain in terms of L_2 risk of the estimators DDSSE and DDMSE even for small sample size. The simulations also demonstrate that shrinkage can be applied to tapered data; as tapering improves the rate of the bias without changing the rate of consistency, it is easy to transfer this to theory. For similar reasons, it is possible to replace the averaging of the periodogram by kernel smoothing.

5.2 Possible technical enhancements

It has been the aim of this work to provide a first approach to shrinkage in multivariate spectral analysis that is both theoretically rigorous and applicable in practice. To accomplish this, we have introduced some restrictions that have allowed to keep the technical expenditure at a reasonable level. Had we not done so, the technical part in appendices A and B might have become longer and more complicated. However, we

would like to communicate our ideas in overcoming these restrictions:
So far, our estimators are based on the averaged periodogram. It is simple to extend the theory to the more general case of estimators based on a kernel smoothed periodogram and can be accomplished by adding kernel weights, at the price of rendering the proofs more difficult to read.
The assumption that the data must be Gaussian is harder to overcome, as it has been used frequently in the proofs. In the context of chapter 2, it has been used throughout all proofs in form of the frequently used lemma B.13. A first step in eliminating the necessity of assumption 2.2 would thus be to check for an alternative each time lemma B.13 has been used. We believe that this is possible, but would come at the price of discussing more remainder terms. Apart from this, assumption 2.2 has been used only in the proof of lemma 2.5, where Gaussianity guarantees that both cumulants and spectra of order higher than two are zero. Here, it would thus be necessary to discuss the higher order spectra and cumulants in detail.
In time series analysis, it is more common to use the Kullback-Leibler discrimination information as a measure of disparity than to use the Hilbert-Schmidt norm. Thus, the question may arise whether it would be possible to base the risk function (2.2) on Kullback-Leibler discrimination information instead of on the Hilbert-Schmidt norm. We doubt that this could be accomplished. The problem is that the norm properties of the Hilbert-Schmidt norm are exploited for each and every proof in the whole work. Kullback-Leibler discrimination information lacks the symmetry property of a norm, which could be mended by symmetrizing it, but it induces no triangular inequality, either. The latter cannot be helped or overcome. Without a triangular inequality, there is no Cauchy-Schwarz inequality which constitutes one of the backbones of the proofs. However, we would like to point to the fact that although the theory on shrinkage is based on the Hilbert-Schmidt norm, is is possible to use these shrinkage estimators in applications that use Kullback-Leibler discrimination information. E.g., in classification, it is Kullback-Leibler discrimination information that is used to classify multivariate time series, and we have shown that our L_2 based shrinkage estimators outperform the averaged periodogram.

5.3 Possible directions of future research

We expect our new proposed method to offer substantial advantages when applied to real data situations, in particular those being discussed in the introduction and in chapter 1. In spectral analysis of multivariate *non-stationary* EEG time series (see section 1.2.2) we believe to be able to contribute to the problem of stabilization of the segmentation criterion, by a less ad-hoc solution than the one proposed by Guo et al. (2005). Guo et al. (2005) address the problem of the estimated eigenvalues of the SLEX spectral matrix being biased by limiting the influence of the small eigenvalues on the cost function. This is achieved by multiplying the estimated eigenvalues with heuristically derived weights. Our method corrects the bias in the estimated eigenvalues directly. This results in a numerically stable, more precise estimate of the true local SLEX spectrum. In consequence, it will permit a better time resolution in the analysis of multivariate, non-stationary data like EEG, as we have shown in the real data example in subsection 4.3.

Our approach also gives improved estimators in related applications such as more general discrimination or classification of time series which are based on Kullback-Leibler or (non-parametric versions of) Whittle estimators. Here, we have seen in subsection 4.4 that shrinkage can offer a substantial advantage. The amount of shrinkage considered optimal for the DDSSE and the DDMSE has been derived with respect to L_2 risk. This is not necessarily the optimal solution for application in classification, and it would be interesting to examine whether it would be possible to further improve by choosing different shrinkage coefficients, perhaps even eliminating the sporadic occurrence of worse classification rates for the DDSSE that we have discussed in subsection 4.4.

A second important field of application to panels of economic time series data is factor modelling. Our achievements in particular of chapter 3 will allow for circumvention of the problem of searching for an appropriate factor dimension - a problem still not satisfactorily solved in the literature, in particular for dynamic factor models. This latter application calls for a possible theoretical direction of future research: the generalization of our approach of chapter 3 to a dynamic factor model setting that allows for lag effects and time-changing weights.

APPENDIX A

Proofs

A.1 Proofs for chapter 2

This appendix gives the proofs of the results in chapter 2. 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, as well as other technical tools needed are given separately in appendix 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.1.1 Proof of lemma 2.2

Proof. We will first show that the norm of $f_T^0(\omega)$ is uniformly bounded in ω. This is then used to show that $\mu_T(\omega)$ and $\alpha_T^2(\omega)$ are bounded, too. The boundedness of $\beta_T^2(\omega)$ is implied by theorem 2.3. Due to the relationship $\alpha_T^2(\omega) + \beta_T^2(\omega) = \delta_T^2(\omega)$, this is sufficient to show the boundedness of $\mu_T(\omega), \alpha_T^2(\omega), \beta_T^2(\omega)$ and $\delta_T^2(\omega)$.

$$\begin{aligned} \left\| f_T^0(\omega) \right\|^2 &= \left\| \frac{1}{m_T} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \mathrm{E}\, I_T(\tilde{\omega}_k) \right\|^2 \\ &\overset{\text{(lemma B.12)}}{\leq} \frac{m_T}{m_T^2} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \left\| \mathrm{E}\, I_T(\tilde{\omega}_k) \right\|^2 \end{aligned}$$

$$
\begin{aligned}
&\leq && \sup_{\omega} \left\| \mathrm{E}\, I_T(\omega) \right\|^2 \\
&\overset{\text{(lemma B.8)}}{=} && \sup_{\omega} \left\| \mathrm{E}\, y_T(\omega) y_T^*(\omega) \right\|^2 \\
&= && \sup_{\omega} \frac{1}{p_T} \sum_{i,j=1}^{p_T} \left| \mathrm{E}\, y_i(\omega) y_j^*(\omega) \right|^2 \\
&= && \underbrace{\sup_{\omega} \frac{1}{p_T} \sum_{i=1}^{p_T} \left(\mathrm{E} \left| y_i(\omega) \right|^2 \right)^2}_{[1]} \\
& && + \underbrace{\sup_{\omega} \frac{1}{p_T} \sum_{\substack{i,j=1 \\ i \neq j}}^{p_T} \left| \mathrm{E}\, y_i(\omega) y_j^*(\omega) \right|^2}_{[2]}
\end{aligned}
$$

[1] is bounded because

$$
\begin{aligned}
[1] \quad &\leq && \sup_{\omega} \frac{1}{p_T} \sum_{i=1}^{p_T} \mathrm{E} \left| y_i(\omega) \right|^4 \\
&\leq && \sup_{\omega} \sqrt{\frac{1}{p_T} \sum_{i=1}^{p_T} \mathrm{E} \left| y_i(\omega) \right|^8} \\
&\overset{\text{Ass.2}}{\leq} && \sqrt{K_2}
\end{aligned}
$$

[2] vanishes asymptotically. According to lemma B.9, 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

$$
\begin{aligned}
[2] \quad &= \quad \frac{1}{p_T} \sum_{\substack{i,j=1 \\ i \neq j}}^{p_T} \left| \mathrm{O}\left(\frac{p_T}{T}\right) \right|^2 \\
&= \quad \mathrm{O}\left(\frac{p_T^3}{T^2}\right)
\end{aligned}
$$

which according to assumption 2.4 converges to zero. We have thus shown that

$$
\sup_{\omega} \left\| f_T^0(\omega) \right\|^2 < \infty. \tag{A.1}
$$

Using (A.1), we can easily show that $\mu_T(\omega)$ is bounded:

$$\begin{aligned}\mu_T(\omega) &= \left\langle f_T^0(\omega), \mathrm{Id}\right\rangle \\ &\leq \sqrt{\left\|f_T^0(\omega)\right\|^2 \left\|\mathrm{Id}\right\|^2} \\ &\leq \left\|f_T^0(\omega)\right\| \end{aligned} \tag{A.2}$$

Now, using (A.1) and (A.2), we show that $\alpha_T^2(\omega)$ is bounded, too:

$$\begin{aligned}\alpha_T^2(\omega) &= \left\|f_T^0(\omega) - \mu_T(\omega)\,\mathrm{Id}\right\|^2 \\ &\leq \left\|f_T^0(\omega)\right\|^2 + 2\mu_T(\omega)\left\|f_T^0(\omega)\right\| + \mu_T^2(\omega)\end{aligned}$$

It remains to show that $\beta_T^2(\omega)$ and $\delta_T^2(\omega)$ are bounded. Now, as the Pythagorean relationship (2.10),

$$\alpha_T^2(\omega) + \beta_T^2(\omega) = \delta_T^2(\omega)$$

holds true, it suffices to show that $\beta_T^2(\omega)$ is bounded. Since $\beta_T^2(\omega) = \mathrm{E}\left\|\hat{f}_T^0(\omega) - f_T^0(\omega)\right\|^2$, this follows from theorem 2.3, which completes the proof. □

A.1.2 Proof of theorem 2.3

We make use of Theorem 7.3.2. in Brillinger (1975), according to which

$$\lim_{T\to\infty} \sum_{i,j=1}^{p} \mathrm{Var}\, \hat{f}_T^{0(ij)}(\omega) = \frac{1}{m_T}\left(\mathrm{tr}(f_T(\omega)\right)^2. \tag{A.3}$$

Our assumption 2.5 implies the assumptions made on the covariance structure of the time series made there. We obtain, using (A.3):

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

$$
\begin{aligned}
&= \lim_{T\to\infty} \frac{1}{p_T}\frac{1}{m_T}\left((\operatorname{tr} f_T(\omega) + \operatorname{tr} f_T^0(\omega))(\operatorname{tr} f_T(\omega) - \operatorname{tr} f_T^0(\omega))\right) \\
&= \lim_{T\to\infty} \frac{1}{p_T}\frac{1}{m_T}\left((\operatorname{tr} f_T(\omega) + \operatorname{tr} f_T^0(\omega))\,\mathrm{O}\left(\frac{m_T p_T}{T}\right)\right) = \mathrm{O}\left(\frac{p_T}{T}\right)
\end{aligned}
$$

□

A.1.3 Proof of theorem 2.4

This can be shown by a simple geometrical argument: The risk

$$\mathcal{R}(\hat{\varphi}_T^*(\omega), f_T^0(\omega))$$

is the distance between the two respective points in the Hilbert space of Hermitian p-dimensional random matrices. In figure A.1.3 we see that the two triangles $f_T^0(\omega) \longleftrightarrow \mu_T(\omega)\,\mathrm{Id} \longleftrightarrow \hat{f}_T^0(\omega)$ and $\hat{\varphi}_T^*(\omega) \longleftrightarrow f_T^0(\omega) \longleftrightarrow \mu_T(\omega)\,\mathrm{Id}$ are similar. This means that 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 $\alpha_T^2(\omega) \times \frac{\beta_T^2(\omega)}{\delta_T^2(\omega)}$.

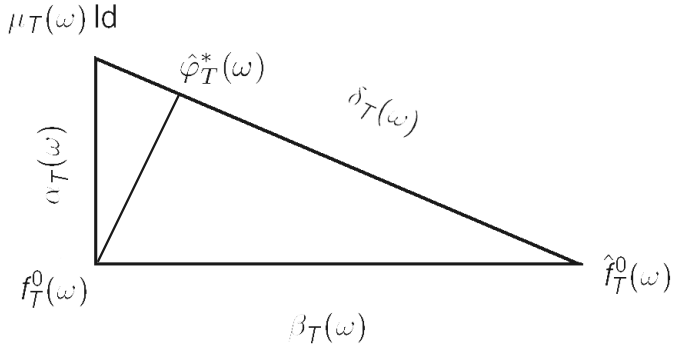

Figure A.1: Geometrical derivation of the risk of the oracle. The risk $\mathcal{R}(\hat{\varphi}_T^*(\omega), f_T^0(\omega))$ is the distance between the two respective points.

□

A.1.4 Proof of lemma 2.5

We will divide the proof into separate paragraphs, each of which gives the proof of one of the formulae (2.18) to (2.21).

A.1.4.1 Proof of (2.18)

We first show that $\mathrm{E}(\hat{\mu}_T(\omega) - \mu_T(\omega))^4 \to 0$: We start by remarking that $\mathrm{E}\,\hat{\mu}_T(\omega) = \mu_T(\omega)$ as

$$\mathrm{E}\,\hat{\mu}_T(\omega) = \mathrm{E}\left\langle \hat{f}_T^0(\omega), \mathrm{Id} \right\rangle = \left\langle f_T^0(\omega), \mathrm{Id} \right\rangle = \mu_T(\omega).$$

Using the ys defined in (2.5), we can expand $\hat{\mu}_T(\omega)$ and $\mu_T(\omega)$:

$$\begin{aligned}
\hat{\mu}_T(\omega) &= \frac{1}{m_T} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \frac{1}{p_T} \sum_{i=1}^{p_T} I_{ii}(\tilde{\omega}_k) \\
&= \frac{1}{m_T} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \frac{1}{p_T} \sum_{i=1}^{p_T} |y_i(\tilde{\omega}_k)|^2
\end{aligned}$$

$$\begin{aligned}
\mu_T(\omega) &= \frac{1}{m_T} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \frac{1}{p_T} \sum_{i=1}^{p_T} \mathrm{E}\, I_{ii}(\tilde{\omega}_k) \\
&= \frac{1}{m_T} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \frac{1}{p_T} \sum_{i=1}^{p_T} \mathrm{E}\, |y_i(\tilde{\omega}_k)|^2
\end{aligned}$$

This enables us to expand the quartic mean as

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

$$\times \ldots \times \left(\frac{1}{p_T} \sum_{i=1}^{p_T} (|y_i(\tilde{\omega}_{k_4})|^2 - \mathrm{E}\,|y_i(\tilde{\omega}_{k_4})|^2) \right) \Bigg] \tag{A.4}$$

Here, we must distinguish two cases: The first is that all the $k_{(\cdot)}$ are distinct. In this case, we use lemma B.17 and lemma B.7 to obtain that (A.4) is $\mathrm{O}\left(\frac{p_T^4}{T^2}\right)$, which is sufficient due to assumptions 2.3 and 2.6.
The second case is that at least two of the $k_{(\cdot)}$ are equal. There are six symmetric conditions, and making extensive use of B.14 and assumption 2.4, 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:

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

where $\tilde{\omega}_m$ denotes the $\tilde{\omega}_{(\cdot)}$ for which the fourth moment above is maximal. Using a binomial expansion and assumption 2.4, we proceed:

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

$$
\begin{aligned}
&\leq \frac{6}{m_T}\sum_{q=0}^{4}\binom{4}{q}\left(\mathrm{E}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_m)|^2\right)^4\right)^{q/4}\\
&\quad\times\left(\mathrm{E}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_m)|^2\right)^4\right)^{(4-q)/4}\\
&\leq \frac{96}{m_T}\,\mathrm{E}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_m)|^2\right)^4\\
&\leq \frac{96}{m_T}\frac{1}{p_T}\sum_{i=1}^{p_T}\mathrm{E}\,|y_i(\tilde{\omega}_m)|^8\\
&\leq \frac{96K_2}{m_T}
\end{aligned}
$$

Overall, we have thus shown so far that

$$
\begin{aligned}
\mathrm{E}(\hat{\mu}_T(\omega)-\mu_T(\omega))^4 &\leq \frac{96}{m_T}\,\mathrm{E}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_m)|^2\right)^4+\mathrm{O}\left(\frac{p_T^4}{T^2}\right)\\
&\leq \frac{96}{m_T}\,\mathrm{E}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_m)|^8\right)+\mathrm{O}\left(\frac{p_T^4}{T^2}\right)\\
&\leq \frac{96K_2}{m_T}+\mathrm{O}\left(\frac{p_T^4}{T^2}\right)
\end{aligned}
$$

which completes the proof of the quartic mean convergence of $\hat{\mu}_T(\omega)$.

A.1.4.2 Proof of (2.21)

We proceed with the quadratic convergence of $\hat{\delta}_T^2(\omega)$ to its expectation. First, we decompose the difference:

$$
\begin{aligned}
&\hat{\delta}_T^2(\omega)-\delta_T^2(\omega)\\
&=\left(\left\|\hat{f}_T^0(\omega)-\hat{\mu}_T(\omega)\,\mathrm{Id}\right\|^2-\left\|\hat{f}_T^0(\omega)-\mu_T(\omega)\,\mathrm{Id}\right\|^2\right)\\
&\quad+\left(\left\|\hat{f}_T^0(\omega)-\mu_T(\omega)\,\mathrm{Id}\right\|^2-\mathrm{E}\left\|\hat{f}_T^0(\omega)-\mu_T(\omega)\,\mathrm{Id}\right\|^2\right) \qquad \text{(A.5)}
\end{aligned}
$$

It suffices to show that both summands on the right side of (A.5) converge to zero in quadratic mean. The first summand does so because of

the quartic mean convergence of $\hat{\mu}_T(\omega)$:

$$\begin{aligned}
& \left\| \hat{f}_T^0(\omega) - \hat{\mu}_T(\omega)\,\mathrm{Id} \right\|^2 - \left\| \hat{f}_T^0(\omega) - \mu_T(\omega)\,\mathrm{Id} \right\|^2 \\
= \; & \left\| \hat{f}_T^0(\omega) \right\|^2 - \hat{\mu}_T(\omega) \left\langle \hat{f}_T^0(\omega), \mathrm{Id} \right\rangle - \hat{\mu}_T(\omega) \left\langle \mathrm{Id}, \hat{f}_T^0(\omega) \right\rangle + \|\hat{\mu}_T(\omega)\|^2 \\
& - \left\| \hat{f}_T^0(\omega) \right\|^2 + \mu_T(\omega) \left\langle \hat{f}_T^0(\omega), \mathrm{Id} \right\rangle + \mu_T(\omega) \left\langle \mathrm{Id}, \hat{f}_T^0(\omega) \right\rangle - \|\mu_T(\omega)\|^2 \\
= \; & \|\hat{\mu}_T(\omega)\|^2 - 2\hat{\mu}_T(\omega)\mu_T(\omega) + \|\mu_T(\omega)\|^2 \\
= \; & (\mu_T(\omega) - \hat{\mu}_T(\omega))^2
\end{aligned}$$

The second summand on the right side of (A.5) consists of a deterministic part and a stochastic part. It suffices to treat the stochastic part:

$$\left\| \hat{f}_T^0(\omega) - \mu_T(\omega)\,\mathrm{Id} \right\|^2 = \mu_T^2(\omega) - 2\mu_T(\omega)\hat{\mu}_T(\omega) + \left\| \hat{f}_T^0(\omega) \right\|^2 \quad \text{(A.6)}$$

Again, we treat the summands one by one. $\mu_T^2(\omega)$ is deterministic and $2\mu_T(\omega)\hat{\mu}_T(\omega)$ converges to $2\mu_T^2(\omega)$ in quadratic mean according to (2.18). We have thus reduced the proof of (2.21) to showing htat $\left\| \hat{f}_T^0(\omega) \right\|^2$ converges in quadratic mean to its expectation, which we will do now.

$$\begin{aligned}
\left\| \hat{f}_T^0(\omega) \right\|^2 &= \frac{1}{p_T} \sum_{i,j=1}^{p_T} \left| \frac{1}{m_T} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} y_i(\tilde{\omega}_k) y_j^*(\tilde{\omega}_k) \right|^2 \\
&\leq \frac{1}{p_T} \sum_{i,j=1}^{p_T} \left(\frac{1}{m_T} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} |y_i(\tilde{\omega}_k) y_j(\tilde{\omega}_k)| \right)^2 \\
&= \frac{1}{p_T} \sum_{i,j=1}^{p_T} \frac{1}{m_T^2} \sum_{k,l=-(m_T-1)/2}^{(m_T-1)/2} |y_i(\tilde{\omega}_k) y_j(\tilde{\omega}_k) y_i(\tilde{\omega}_l) y_j(\tilde{\omega}_l)| \\
&= \frac{p_T}{m_T^2} \sum_{k,l=-(m_T-1)/2}^{(m_T-1)/2} \left(\frac{1}{p_T} \sum_{i=1}^{p_T} |y_i(\tilde{\omega}_k) y_i(\tilde{\omega}_l)| \right)^2 \\
&= \underbrace{\frac{p_T}{m_T^2} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \left(\frac{1}{p_T} \sum_{i=1}^{p_T} |y_i(\tilde{\omega}_k)|^2 \right)^2}_{[3]}
\end{aligned}$$

$$+ \underbrace{\frac{p_T}{m_T^2} \sum_{\substack{k,l=-(m_T-1)/2 \\ k\neq l}}^{(m_T-1)/2} \left(\frac{1}{p_T} \sum_{i=1}^{p_T} |y_i(\tilde{\omega}_k) y_i(\tilde{\omega}_l)| \right)^2}_{[4]}$$

We show that the variance of both [3] and [4] vanishes asymptotically, where we make essentially use of assumptions 2.3 and 2.4 and the asymptotic behavior of cumulants of the $y(\cdot)$'s at different Fourier frequencies (lemmata B.15,B.16 and B.17). We start with [3]:

$$\begin{aligned} \mathrm{Var}[3] &= \frac{p_T^2}{m_T^4} \mathrm{Var} \left(\sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \left(\frac{1}{p_T} \sum_{i=1}^{p_T} |y_i(\tilde{\omega}_k)|^2 \right)^2 \right) \\ &= \underbrace{\frac{p_T^2}{m_T^4} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \mathrm{Var} \left(\frac{1}{p_T} \sum_{i=1}^{p_T} |y_i(\tilde{\omega}_k)|^2 \right)^2}_{[5]} \end{aligned}$$

$$+ \underbrace{\frac{p_T^2}{m_T^4} \sum_{\substack{k,l=-(m_T-1)/2 \\ k\neq l}}^{(m_T-1)/2} \mathrm{Cov} \left(\left(\frac{1}{p_T} \sum_{i=1}^{p_T} |y_i(\tilde{\omega}_k)|^2 \right)^2, \left(\frac{1}{p_T} \sum_{i=1}^{p_T} |y_i(\tilde{\omega}_l)|^2 \right)^2 \right)}_{[6]} \tag{A.7}$$

First, we treat [5]. The order of magnitude of [5] 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:

$$\begin{aligned} [5] &\approx \frac{p_T^2}{m_T^3} \mathrm{Var} \left(\frac{1}{p_T} \sum_{i=1}^{p_T} |y_i(\omega)|^2 \right)^2 \\ &\leq \frac{p_T^2}{m_T^3} \mathrm{E} \left(\frac{1}{p_T} \sum_{i=1}^{p_T} |y_i(\omega)|^2 \right)^4 \\ &\leq \frac{1}{m_T} \left(\frac{p_T}{m_T} \right)^2 \left(\frac{1}{p_T} \sum_{i=1}^{p_T} \mathrm{E} \, |y_i(\omega)|^8 \right) \end{aligned}$$

$$\begin{aligned} &\leq \frac{K_1^2 K_2}{m_T} \\ &\to 0 \end{aligned}$$

where we have used assumptions 2.3 and 2.4. To treat A.7, we first expand:

$$\begin{aligned} &\operatorname{Cov}\left(\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_k)|^2\right)^2, \left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_l)|^2\right)^2\right) \\ = \;& \frac{1}{p_T^4}\operatorname{Cov}\left(\left(\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_k)|^2\right)^2, \left(\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_l)|^2\right)^2\right) \\ = \;& \frac{1}{p_T^4}\sum_{i_1,i_2,i_3,i_4=1}^{p_T}\operatorname{Cov}\left(|y_{i_1}(\tilde{\omega}_k)|^2|y_{i_2}(\tilde{\omega}_k)|^2, |y_{i_3}(\tilde{\omega}_l)|^2|y_{i_4}(\tilde{\omega}_l)|^2\right) \end{aligned} \tag{A.8}$$

Now we use Lemma B.15. We must distinguish three cases: firstly, if the covariance in (A.8) 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.8) is decomposed into terms of the kind

$$\frac{1}{p_T^4}\sum_{i_1,i_2,i_3,i_4=1}^{p_T}\operatorname{Var}(y_{i_1}(\tilde{\omega}_k))\operatorname{Var}(y_{i_3}(\tilde{\omega}_l))\left(\operatorname{Cov}(y_{i_2}(\tilde{\omega}_k), y_{i_4}(\tilde{\omega}_l))\right)^2. \tag{A.9}$$

Now, using Assumption 2.4, we have that (A.9) is $\mathrm{O}\left(\frac{p_T^2}{T^2}\right)$, which is also the convergence rate of the whole term A.7.
It remains to be shown that the variance of [4] converges to zero, too.

$$\operatorname{Var}[4] = \frac{p_T^2}{m_T^4}\sum_{\substack{k_1,k_2,k_3,k_4=-(m_T-1)/2 \\ k_1\neq k_2, k_3\neq k_4}}^{(m_T-1)/2} C(k_1,k_2,k_3,k_4) \tag{A.10}$$

where $C(k_1,k_2,k_3,k_4)$ is an abbreviation of

$$\operatorname{Cov}\left\{\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_2})|\right)^2, \left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_3})y_i(\tilde{\omega}_{k_4})|\right)^2\right\}$$

The order of the covariances on the right side of (A.10) only depends on the cardinal of the intersection of the two sets $\{k_1, k_2\}$ and $\{k_3, k_4\}$, which lies between zero and two. We study the three cases separately:

A.1.4.2.1 $\#(\{k_1, k_2\} \cap \{k_3, k_4\}) = 0$ In this case, we have

$$\frac{p_T^2}{m_T^4} \sum_{\substack{k_1,k_2,k_3,k_4=-(m_T-1)/2 \\ k_i \neq k_j}}^{(m_T-1)/2} C(k_1, k_2, k_3, k_4)$$

The sum and the prefactor $\frac{1}{m_T^4}$ eliminating one another, we continue:

$$\begin{aligned}
&p_T^2 \operatorname{Cov}\left(\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_2})|\right)^2, \left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_3})y_i(\tilde{\omega}_{k_4})|\right)^2\right) \\
&= \frac{1}{p_T^2}\operatorname{Cov}\left(\left(\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_2})|\right)^2, \left(\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_3})y_i(\tilde{\omega}_{k_4})|\right)^2\right) \\
&= \frac{1}{p_T^2}\sum_{i_1,i_2,i_3,i_4=1}^{p_T} \operatorname{Cov}\left(|y_{i_1}(\tilde{\omega}_{k_1})y_{i_2}(\tilde{\omega}_{k_1})y_{i_1}(\tilde{\omega}_{k_2})y_{i_2}(\tilde{\omega}_{k_2})|,\right. \qquad \text{(A.11)} \\
&\qquad\qquad \left. |y_{i_3}(\tilde{\omega}_{k_3})y_{4_1}(\tilde{\omega}_{k_3})y_{i_3}(\tilde{\omega}_{k_4})y_{i_4}(\tilde{\omega}_{k_4})|\right)
\end{aligned}$$

Using Lemma B.15, the covariance of order eight can be decomposed to a product of four covariances of order two. In each of the covariances, either the dimension indexes or the frequencies are different. Using lemma B.2 and lemma B.4, we obtain

$$(\text{A.11}) = \mathrm{O}\left(\frac{p_T^6}{T^4}\right)$$

which converges to zero according to assumption 2.6.

A.1.4.2.2 $\#(\{k_1, k_2\} \cap \{k_3, k_4\}) = 1$ This happens

$$4m_T(m_T - 1)(m_T - 2)$$

times in the summation in (A.10). The covariance that is summed over is now shown to have an upper and a lower bound that converge to zero fast enough, thus the absolute value of the covariance converges to zero at a sufficiently fast rate, too.

A.1.4.2.2.1 Upper bound of(A.10) We will make use of the abbreviation

$$
\begin{aligned}
& y_a(b) = y_{i_a}(\tilde{\omega}_{k_b}) \\
& \operatorname{Cov}\left(\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_2})|\right)^2, \left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_3})y_i(\tilde{\omega}_{k_4})|\right)^2\right) \\
& = \operatorname{Cov}\left(\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_2})|\right)^2, \left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_3})|\right)^2\right) \\
\leq\ & \frac{1}{p_T^4}\sum_{i_1,i_2,i_3,i_4=1}^{p_T}\operatorname{E}|y_1(1)y_1(2)y_2(1)y_2(2)y_3(1)y_3(3)y_4(1)y_4(3)| \\
=\ & \frac{1}{p_T^4}\sum_{i_1,i_2,i_3,i_4=1}^{p_T}\operatorname{Cov}\left(|y_1(1)y_2(1)y_3(1)y_4(1)|, |y_1(2)y_2(2)y_3(3)y_4(3)|\right) \\
& + \frac{1}{p_T^4}\sum_{i_1,i_2,i_3,i_4=1}^{p_T}\operatorname{E}|y_1(1)y_2(1)y_3(1)y_4(1)|\operatorname{E}|y_1(2)y_2(2)y_3(3)y_4(3)| \\
=\ & \underbrace{\frac{1}{p_T^4}\sum_{i_1,i_2,i_3,i_4=1}^{p_T}\operatorname{Cov}\left(|y_1(1)y_2(1)y_3(1)y_4(1)|, |y_1(2)y_2(2)y_3(3)y_4(3)|\right)}_{[7]} \\
+\ & \underbrace{\frac{1}{p_T^4}\sum_{i_1,i_2,i_3,i_4=1}^{p_T}\operatorname{E}|y_1(1)y_2(1)y_3(1)y_4(1)|\operatorname{Cov}(|y_1(2)y_2(2), y_3(3)y_4(3)|)}_{[8]} \\
+\ & \underbrace{\frac{1}{p_T^4}\sum_{i_1,i_2,i_3,i_4=1}^{p_T}\operatorname{E}|y_1(1)y_2(1)y_3(1)y_4(1)|\operatorname{E}|y_1(2)y_2(2)|\operatorname{E}|y_3(3)y_4(3)|}_{[9]}
\end{aligned}
$$

Now, using lemma B.15, [7] is $\mathrm{O}\left(\frac{m_T^4}{T^4}\right)$, and using Lemma B.15 and Assumption 2, [8] is $\mathrm{O}\left(\frac{m_T^2}{T^2}\right)$, so we can focus on [9]. For an appropriate choice of $\bar{\omega}$ near ω, we have:

$$
[9] \leq \frac{1}{p_T^4}\sum_{i,j=1}^{p_T}\operatorname{E}|y_i(\tilde{\omega}_{k_1})y_j(\tilde{\omega}_{k_1})|^2\operatorname{E}|y_i(\tilde{\omega}_{k_2})|^2\operatorname{E}|y_j(\tilde{\omega}_{k_3})|^2
$$

$$
\begin{aligned}
&\leq \frac{1}{p_T^4}\sum_{i,j=1}^{p_T}\sqrt{\mathrm{E}\,|y_i(\tilde{\omega}_{k_1})|^4}\sqrt{\mathrm{E}\,|y_j(\tilde{\omega}_{k_1})|^4}\,\mathrm{E}\,|y_i(\tilde{\omega}_{k_2})|^2\,\mathrm{E}\,|y_j(\tilde{\omega}_{k_3})|^2\\
&\leq \frac{1}{p_T^2}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}\sqrt{\mathrm{E}\,|y_i(\bar{\omega})|^4}\,\mathrm{E}\,|y_i(\bar{\omega})|^2\right)^2\\
&\leq \frac{1}{p_T^2}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}\mathrm{E}\,|y_i(\bar{\omega})|^8\right)\\
&\leq \frac{K_2^2}{p_T^2}
\end{aligned}
$$

A.1.4.2.2.2 Lower bound of(A.10) Now, we will continue with the lower bound of (A.10) in the case $\#\left(\{k_1,k_2\}\cap\{k_3,k_4\}\right)=1$ where, again for a $\bar{\omega}$ appropriately chosen, we have:

$$
\begin{aligned}
&-\mathrm{Cov}\left(\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_2})|\right)^2,\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_3})y_i(\tilde{\omega}_{k_4})|\right)^2\right)\\
&=-\mathrm{Cov}\left(\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_2})|\right)^2,\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_3})|\right)^2\right)\\
&\leq \mathrm{E}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_2})|\right)^2\mathrm{E}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_3})|\right)^2\\
&\leq \left(\mathrm{E}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_2})|\right)^2\right)^2 \qquad \text{(A.12)}
\end{aligned}
$$

where we assume w.l.o.g. that

$$
\mathrm{E}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_2})|\right)^2\geq\mathrm{E}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_3})|\right)^2
$$

(otherwise we just exchange the indexes k_2 and k_3). We continue:

$$
(\text{A.12})\leq\left(\frac{1}{p_T^2}\sum_{i,j=1}^{p_T}\mathrm{E}\,|y_i(\tilde{\omega}_{k_1})y_j(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_2})y_j(\tilde{\omega}_{k_2})|\right)^2
$$

$$
\begin{aligned}
&= \left(\frac{1}{p_T^2} \sum_{i,j=1}^{p_T} \Big(\mathrm{E}\,|y_i(\tilde{\omega}_{k_1}) y_j(\tilde{\omega}_{k_1})| \,\mathrm{E}\,|y_i(\tilde{\omega}_{k_2}) y_j(\tilde{\omega}_{k_2})| \right. \\
&\quad \left. + \operatorname{Cov}(|y_i(\tilde{\omega}_{k_1}) y_j(\tilde{\omega}_{k_1})|, |y_i(\tilde{\omega}_{k_2}) y_j(\tilde{\omega}_{k_2})|) \Big) \right)^2 \\
&= \left(\frac{1}{p_T^2} \sum_{i,j=1}^{p_T} \left(\mathrm{E}\,|y_i(\tilde{\omega}_{k_1}) y_j(\tilde{\omega}_{k_1})| \,\mathrm{E}\,|y_i(\tilde{\omega}_{k_2}) y_j(\tilde{\omega}_{k_2})| + \mathrm{O}\left(\frac{p_T}{T}\right) \right) \right)^2 \\
&\approx \left(\frac{1}{p_T^2} \sum_{i,j=1}^{p_T} \left(\mathrm{E}\,|y_i(\tilde{\omega}_{k_1}) y_j(\tilde{\omega}_{k_1})| \,\mathrm{E}\,|y_i(\tilde{\omega}_{k_2}) y_j(\tilde{\omega}_{k_2})| \right) \right)^2 \\
&\leq \left(\frac{1}{p_T^2} \sum_{i,j=1}^{p_T} \left(\mathrm{E}\,|y_i(\tilde{\omega}_{k_1}) y_j(\tilde{\omega}_{k_1})| \right)^2 \right)^2 \qquad \text{(A.13)}
\end{aligned}
$$

where again we assume without loss of generality that

$$\mathrm{E}\,|y_i(\tilde{\omega}_{k_1}) y_j(\tilde{\omega}_{k_1})| \geq \mathrm{E}\,|y_i(\tilde{\omega}_{k_2}) y_j(\tilde{\omega}_{k_2})|.$$

We continue:

$$
\begin{aligned}
\text{(A.13)} &= \left(\frac{1}{p_T^2} \sum_{i=1}^{p_T} \left(\mathrm{E}\,|y_i(\tilde{\omega}_{k_1})|^2 \right)^2 + \frac{1}{p_T^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{p_T} \left(\mathrm{E}\,|y_i(\tilde{\omega}_{k_1}) y_j(\tilde{\omega}_{k_1})| \right)^2 \right)^2 \\
&\leq \left(\frac{1}{p_T^2} \sum_{i=1}^{p_T} \mathrm{E}\,|y_i(\tilde{\omega}_{k_1})|^4 + \mathrm{O}\left(\frac{p_T^2}{T^2} \right) \right)^2 \\
&\approx \left(\frac{1}{p_T^2} \sum_{i=1}^{p_T} \mathrm{E}\,|y_i(\tilde{\omega}_{k_1})|^4 \right)^2 \\
&= \frac{1}{p_T^2} \left(\frac{1}{p_T} \sum_{i=1}^{p_T} \mathrm{E}\,|y_i(\tilde{\omega}_{k_1})|^4 \right)^2 \\
&\leq \frac{1}{p_T^2} \left(\frac{1}{p_T} \sum_{i=1}^{p_T} \mathrm{E}\,|y_i(\tilde{\omega}_{k_1})|^8 \right) \\
&\leq \frac{1}{p_T^2} K_1
\end{aligned}
$$

Summarizing, when the cardinal of the intersection is one, (A.10) is $\frac{p_T^2}{m_T^4}$ times a sum of order m_T^3, and each of the summands is $\mathrm{O}\left(\frac{1}{p_T^2}\right)$, thus we have shown that in this case, (A.10) is $\mathrm{O}\left(\frac{1}{m_T}\right)$. This completes the part where the cardinal of the intersection is one.

A.1.4.2.3 $\#\left(\{k_1,k_2\}\cap\{k_3,k_4\}\right)=2$ In $2m_T(m_T-1)$ cases, the cardinal of the intersection is two. This means that in (A.10), the sum and the prefactor $\frac{p_T^2}{m_T^4}$ combine to $\frac{p_T^2}{m_T^2}$. Due to assumption 2.3, it thus suffices to show that each of the summands $C(k_1,k_2,k_3,k_4)$ in (A.10) converges to zero. Again using the abbreviation

$$y_a(b)=y_{i_a}(\tilde{\omega}_{k_b}),$$

we now treat these summands:

$$\begin{aligned}
&\left|\mathrm{Cov}\left(\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_2})|\right)^2,\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_3})y_i(\tilde{\omega}_{k_4})|\right)^2\right)\right|\\
&=\left|\mathrm{Cov}\left(\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_2})|\right)^2,\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_2})|\right)^2\right)\right|\\
&\leq\frac{1}{p_T^4}\sum_{i_1,i_2,i_3,i_4=1}^{p_T}\left|\mathrm{Cov}\left(|y_1(1)y_2(1)y_1(2)y_2(2)|,|y_3(1)y_4(1)y_3(2)y_4(2)|\right)\right|
\end{aligned}\tag{A.14}$$

We now decompose the set of quadruples of integers between 1 and p_T into two disjoint subsets: the set Q that contains the quadruples consisting of four distinct integers, and the set R that contains the rest. We obtain

$$\begin{aligned}
&\left|\mathrm{Cov}\left(\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_1})y_i(\tilde{\omega}_{k_2})|\right)^2,\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\tilde{\omega}_{k_3})y_i(\tilde{\omega}_{k_4})|\right)^2\right)\right|\\
\leq\;&\frac{1}{p_T^4}\sum_{Q}\left|\mathrm{Cov}\left(|y_1(1)y_2(1)y_1(2)y_2(2)|,|y_3(1)y_4(1)y_3(2)y_4(2)|\right)\right|\\
&+\frac{1}{p_T^4}\sum_{R}\left|\mathrm{Cov}\left(|y_1(1)y_2(1)y_1(2)y_2(2)|,|y_3(1)y_4(1)y_3(2)y_4(2)|\right)\right|
\end{aligned}\tag{A.15}$$

The cardinal of Q is of order $p(p-1)(p-2)(p-3)$. We now look directly at the elements of Q

$$
\begin{aligned}
& |\mathrm{Cov}(|y_1(1)y_2(1)y_1(2)y_2(2)|, |y_3(1)y_4(1)y_3(2)y_4(2)|)| \\
= \; & |\mathrm{E}\,|y_1(1)y_2(1)y_1(2)y_2(2)||y_3(1)y_4(1)y_3(2)y_4(2)| \\
& - \mathrm{E}\,|y_1(1)y_2(1)y_1(2)y_2(2)|\,\mathrm{E}\,|y_3(1)y_4(1)y_3(2)y_4(2)|| \\
= \; & |\mathrm{E}\,|y_1(1)y_2(1)y_3(1)y_4(1)|\,\mathrm{E}\,|y_1(2)y_2(2)y_3(2)y_4(2)| \\
& - \mathrm{Cov}(|y_1(1)y_2(1)y_3(1)y_4(1)|, |y_1(2)y_2(2)y_3(2)y_4(2)|) \\
& - (\mathrm{E}\,|y_1(1)y_2(1)|\,\mathrm{E}\,|y_1(2)y_2(2)| - \mathrm{Cov}(|y_1(1)y_2(1)|, |y_1(2)y_2(2)|)) \\
& \times (\mathrm{E}\,|y_3(1)y_4(1)|\,\mathrm{E}\,|y_3(2)y_4(2)| - \mathrm{Cov}(|y_3(1)y_4(1)|, |y_2(2)y_4(2)|))| \\
= \; & \Bigg|\mathrm{E}\,|y_1(1)y_2(1)y_3(1)y_4(1)|\,\mathrm{E}\,|y_1(2)y_2(2)y_3(2)y_4(2)| - \mathrm{O}\left(\frac{p_T^4}{T^4}\right) \\
& - \left(\mathrm{E}\,|y_1(1)y_2(1)|\,\mathrm{E}\,|y_1(2)y_2(2)| - \mathrm{O}\left(\frac{p_T^2}{T^2}\right)\right) \\
& \times \left(\mathrm{E}\,|y_3(1)y_4(1)|\,\mathrm{E}\,|y_3(2)y_4(2)| - \mathrm{O}\left(\frac{p_T^2}{T^2}\right)\right)\Bigg| \\
\approx \; & |\mathrm{E}\,|y_1(1)y_2(1)y_3(1)y_4(1)|\,\mathrm{E}\,|y_1(2)y_2(2)y_3(2)y_4(2)| \qquad \text{(A.16)} \\
& - (\mathrm{E}\,|y_1(1)y_2(1)|\,\mathrm{E}\,|y_1(2)y_2(2)|)\,(\mathrm{E}\,|y_3(1)y_4(1)|\,\mathrm{E}\,|y_3(2)y_4(2)|)|
\end{aligned}
$$

where we have applied lemma B.16 to obtain the convergence rates. Using lemma B.2, we get rid of the small terms, and again, like above, assuming w.l.o.g. that $\mathrm{E}\,|y_1(1)y_2(1)y_3(1)y_4(1)| \geq \mathrm{E}\,|y_1(2)y_2(2)y_3(2)y_4(2)|$, we continue:

$$
\begin{aligned}
& \text{(A.16)} \\
\approx \; & (E|y_1(1)y_2(1)y_3(1)y_4(1)|)^2 \\
= \; & (\mathrm{Cov}(|y_1(1)y_2(1)|, |y_3(1)y_4(1)|) + \mathrm{E}\,|y_1(1)y_2(1)| + \mathrm{E}\,|y_3(1)y_4(1)|)^2 \\
= \; & \left(\mathrm{Cov}(|y_1(1)y_2(1)|, |y_3(1)y_4(1)|) + \mathrm{O}\left(\frac{p_T}{T}\right) + \mathrm{O}\left(\frac{p_T}{T}\right)\right)^2 \\
\approx \; & (\mathrm{Cov}(|y_1(1)y_2(1)|, |y_3(1)y_4(1)|))^2
\end{aligned}
$$

Now, using lemma B.13, we have that

$$
(\mathrm{Cov}(|y_1(1)y_2(1)|, |y_3(1)y_4(1)|))^2 = \mathrm{O}\left(\frac{p_T^2}{T^2}\right)
$$

and so the sum over Q does converge to zero. It remains to show that the sum over R converges to zero, too. The elements of R are the quadruples for which one of the conditions $i_1 = i_2$, $i_1 = i_3$, $i_1 = i_4$, $i_2 = i_3$, $i_2 = i_4$, $i_3 = i_4$ is fulfilled. Although in general it does make a difference if the link is between elements on the same side of the covariance (e.g. $i_1 = i_2$) or between elements on different sides of the covariance, in the following calculations this will not be the case; the difference is eliminated by the use lemma B.14. Due to the link, the quadruple sum in (A.14) will reduce to a triple sum. We will pick $i_1 = i_2$ as an example, the five other cases are analoguous:

$$\begin{aligned}
&\frac{1}{p_T^4}\sum_R |\mathrm{Cov}(|y_1(1)y_2(1)y_1(2)y_2(2)|, |y_3(1)y_4(1)y_3(2)y_4(2)|)| \\
&\leq \frac{1}{p_T^4}\sum_{i_1,i_3,i_4} |\mathrm{Cov}(|y_1(1)y_1(2)y_1(1)y_1(2)|, |y_3(1)y_3(2)y_4(1)y_4(2)|)| \\
&\leq \frac{1}{p_T^4}\sum_{i_1,i_3,i_4} \left\{\mathrm{E}\,|y_1(1)|^4|y_1(2)|^4\,\mathrm{E}\,|y_3(1)|^2|y_3(2)|^2|y_4(1)|^2|y_4(2)|^2\right\}^{1/2} \\
&\leq \frac{1}{p_T^4}\sum_{i_1,i_3,i_4} \left\{\mathrm{E}\,|y_1(1)|^4\,\mathrm{E}\,|y_1(2)|^4 + \mathrm{Cov}(|y_1(1)|^4, |y_1(2)|^4)\right\}^{1/2} \\
&\qquad \times \left\{\mathrm{E}\,|y_3(1)|^2|y_4(1)|^2\,\mathrm{E}\,|y_3(2)|^2|y_4(2)|^2\right. \\
&\qquad\qquad \left. + \mathrm{Cov}(|y_3(1)|^2|y_4(1)|^2, |y_3(2)|^2|y_4(2)|^2)\right\}^{1/2}
\end{aligned} \tag{A.17}$$

Now, with the same argument as frequently used throughout this proof, and using Lemma B.16 , we continue:

$$\begin{aligned}
&\text{(A.17)} \\
\leq\ & \frac{1}{p^4}\sum_{i_1,i_3,i_4} \sqrt{\mathrm{E}\,|y_1(1)|^4\,\mathrm{E}\,|y_1(2)|^4 + \mathrm{O}\left(\frac{p_T^4}{T^4}\right)} \\
&\qquad \times\sqrt{\mathrm{E}\,|y_3(1)|^2|y_4(1)|^2\,\mathrm{E}\,|y_3(2)|^2|y_4(2)|^2 + \mathrm{O}\left(\frac{p_T^4}{T^4}\right)} \\
\approx\ & \frac{1}{p_T^4}\sum_{i_1,i_3,i_4} \sqrt{\mathrm{E}\,|y_1(1)|^4\,\mathrm{E}\,|y_1(2)|^4\,\mathrm{E}\,|y_3(1)|^2|y_4(1)|^2\,\mathrm{E}\,|y_3(2)|^2|y_4(2)|^2} \\
\leq\ & \frac{1}{p_T^4}\sum_{i_1,i_3,i_4} \mathrm{E}\,|y_1(1)|^4\,\mathrm{E}\,|y_3(1)|^2|y_4(1)|^2
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{p_T^4} \sum_{i_1,i_3,i_4} \mathrm{E}\,|y_1(1)|^4 \sqrt{\mathrm{E}\,|y_3(1)|^4} \sqrt{|y_4(1)|^4} \\
&\leq \frac{1}{p_T} \left(\frac{1}{p_T} \sum_{i=1}^{p_T} \mathrm{E}\,|y_i(\tilde{\omega}_{k_1})|^4 \right) \left(\frac{1}{p_T} \sum_{i=1}^{p_T} \sqrt{\mathrm{E}\,|y_i(\tilde{\omega}_{k_1})|^4} \right)^2 \\
&\leq \frac{1}{p_T} \left(\frac{1}{p_T} \sum_{i=1}^{p_T} \mathrm{E}\,|y_i(\tilde{\omega}_{k_1})|^4 \right)^2 \\
&\leq \frac{1}{p_T} \left(\frac{1}{p_T} \sum_{i=1}^{p_T} \mathrm{E}\,|y_i(\tilde{\omega}_{k_1})|^8 \right) \\
&\leq \frac{1}{p_T} K_2
\end{aligned}
$$

This proves that the sum over R converges to zero, too. We have thus shown that equation (2.21) holds true. Now we will continue:

A.1.4.3 Proof of (2.20)

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:

$$
\begin{aligned}
&\bar{\beta}_T^2(\omega) - \beta_T^2(\omega) \\
&= \Bigg[\underbrace{\frac{1}{m_T^2} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \left\| I_T(\tilde{\omega}_k) - f_T^0(\omega) \right\|^2}_{[10]} - \underbrace{\mathrm{E} \left\| \hat{f}_T^0(\omega) - f_T^0(\omega) \right\|^2}_{[11]} \Bigg] \\
&\quad + \underbrace{\Bigg[\frac{1}{m_T^2} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \left\| I_T(\tilde{\omega}_k) - \hat{f}_T^0(\omega) \right\|^2}_{[12]} \\
&\qquad \underbrace{- \frac{1}{m_T^2} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \left\| I_T(\tilde{\omega}_k) - f_T^0(\omega) \right\|^2 \Bigg]}_{[13]}
\end{aligned}
$$

Now we have to show that both ([10] − [11]) and ([12] − [13]) converge to zero in quadratic mean. We start with ([10] − [11]). First we show

that the asymptotic expectation of [10] equals the asymptotic value of [11]. The expectation of [10] is:

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

[11] can be expanded as

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

Thus, the expectation of ([10] − [11]) is asymptotically zero. We now show that the variance of ([10] − [11]) vanishes asymptotically, too. As $\mathrm{E}\left\|\hat{f}_T^0(\omega)-f_T^0(\omega)\right\|^2$ is deterministic, we restrict ourselves to investigating the variance of [10]:

$$\begin{aligned}
& \mathrm{Var}\left(\frac{1}{m_T^2}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\left\|I_T(\tilde{\omega}_k)-f_T^0(\omega)\right\|^2\right) \\
= & \frac{1}{m_T^4}\,\mathrm{Var}\left(\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\left\|I_T(\tilde{\omega}_k)-f_T^0(\omega)\right\|^2\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{m_T^4} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \operatorname{Var} \left\| I_T(\tilde{\omega}_k) - f_T^0(\omega) \right\|^2 \\
&+ \frac{1}{m_T^4} \sum_{\substack{k,l=-(m_T-1)/2 \\ k \neq l}}^{(m_T-1)/2} \operatorname{Cov} \left(\left\| I_T(\tilde{\omega}_k) - f_T^0(\omega) \right\|^2, \left\| I_T(\tilde{\omega}_l) - f_T^0(\omega) \right\|^2 \right) \\
&\approx \frac{1}{m_T^3} \operatorname{Var} \left(\left\| I_T(\omega) - f_T^0(\omega) \right\|^2 \right) \\
&= \frac{1}{m_T^3} \operatorname{Var} \left(\left\| y_T(\omega) y_T^*(\omega) - \operatorname{E} y_T(\omega) y_T^*(\omega) + \operatorname{O}\left(\frac{p_T}{m_T^2} \right) \right\|^2 \right) \\
&\approx \frac{1}{m_T^3} \operatorname{Var} \left(\left\| y_T(\omega) y_T^*(\omega) - \operatorname{E} y_T(\omega) y_T^*(\omega) \right\|^2 \right) \\
&= \frac{1}{m_T^3} \frac{1}{p_T^2} \sum_{i_1,i_2,i_3,i_4=1}^{p_T} \operatorname{Cov} \left(y_{i1}(\omega) y_{i2}^*(\omega), y_{i3}(\omega) y_{i4}^*(\omega) \right) \\
&= \frac{1}{m_T^3} \frac{1}{p_T^2} \sum_{i_1,i_2,i_3,i_4=1}^{p_T} \left(\operatorname{Cov}(y_{i1}(\omega), y_{i3}(\omega)) \operatorname{Cov}(y_{i2}^*(\omega), y_{i4}^*(\omega)) \right. \\
&\qquad \left. + \operatorname{Cov}(y_{i1}(\omega), y_{i4}^*(\omega)) \operatorname{Cov}(y_{i2}^*(\omega), y_{i3}(\omega)) \right) \\
= \; & \frac{p_T^2}{m_T^3} \operatorname{O}\left(\frac{p_T^2}{T^2} \right)
\end{aligned}
$$

where we have used lemma B.13 and lemma B.2. We can now move on to the difference ([12] − [13]). Some elementary transformations result in

$$
([12] - [13]) = \frac{1}{m_T} \left\| \hat{f}_T^0(\omega) - f_T^0(\omega) \right\|^2 .
$$

To show mean square convergence of ([12] − [13]), it is sufficient to show that its second moment vanishes:

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

Now, $\mathrm{E}\,\|\hat{f}_T^0(\omega)-\mu_T(\omega)\,\mathrm{Id}\,\|^4$ is bounded, which we have shown above, see (A.6), and $\|f_T^0(\omega)-\mu_T(\omega)\,\mathrm{Id}\,\|^4=\delta^4(\omega)$ is bounded, too, so ([12]$-$[13]) 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 $\hat{\beta}_T^2(\omega)$.

A.1.4.4 Proof of (2.19)

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. □

A.1.5 Proof of theorem 2.6

We will first show that

$$\mathrm{E}\left\|\hat{f}_T^*(\omega)-\hat{\varphi}_T^*(\omega)\right\|^2\to 0. \tag{A.18}$$

We decompose

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

$$\leq \quad (\hat{\mu}_T(\omega) - \mu_T(\omega))^2 + \frac{(\hat{\alpha}_T^2(\omega)\delta_T^2(\omega) - \alpha_T^2(\omega)\hat{\delta}_T^2(\omega))^2}{\hat{\delta}_T^2(\omega)\delta_T^4(\omega)}$$

Now, $\mathrm{E}(\hat{\mu}_T(\omega) - \mu_T(\omega))^2$ converges to zero by lemma 2.5. Furthermore, using that $\alpha_T^2(\omega) \leq \delta_T^2(\omega)$ and $\hat{\alpha}_T^2(\omega) \leq \hat{\delta}_T^2(\omega)$ and lemma B.11, we can show that the expectation of the second term converges to zero too. The requirements of B.11 are fulfilled as, first

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

and second as $\mathrm{E}(\hat{\alpha}_T^2(\omega)\delta_T^2(\omega) - \alpha_T^2(\omega)\hat{\delta}_T^2(\omega))^2 \to 0$ because

$$\begin{aligned} & \hat{\alpha}_T^2(\omega)\delta_T^2(\omega) - \alpha_T^2(\omega)\hat{\delta}_T^2(\omega) \\ = \quad & (\hat{\alpha}_T^2(\omega) - \alpha_T^2(\omega))\delta_T^2(\omega) - (\hat{\delta}_T^2(\omega) - \delta_T^2(\omega))\alpha_T^2(\omega) \end{aligned} \tag{A.19}$$

and according to lemma 2.5 and lemma 2.2, (A.19) converges to zero in quadratic mean.
Thus, the first statement of theorem 2.6 is shown. The second statement follows immediately from the first as

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

which completes the proof.

□

A.1.6 Proof of theorem 2.7

We will first introduce a new parameter $\dot{\alpha}_T^2(\omega)$ with the aid of which we will then give an explicit formula for the theoretically optimal choice of random shrinkage coefficients. We will then show that the DDSSE $\hat{f}_T^*(\omega)$ converges to $\hat{\varphi}_T^{**}(\omega)$ in quadratic mean.
Let us define

$$\dot{\alpha}_T^2(\omega) = \left\langle f_T^0(\omega), \hat{f}_T^0(\omega)\right\rangle - \mu_T(\omega)\hat{\mu}_T(\omega)$$

The expectation of $\dot{\alpha}_T^2(\omega)$ is $\alpha_T^2(\omega)$:

$$\begin{aligned}
\mathrm{E}\,\dot{\alpha}_T^2(\omega) &= \left\langle f_T^0(\omega), f_T^0(\omega)\right\rangle - \mu_T^2(\omega) \\
&= \left\| f_T^0(\omega)\right\|^2 - \mu_T^2(\omega) \\
&= \left\| f_T^0(\omega) - \mu_T(\omega)\,\mathrm{Id}\right\|^2 \\
&= \alpha_T^2(\omega)
\end{aligned} \tag{A.20}$$

For the absolute value of $\dot{\alpha}_T^2(\omega)$ we have

$$\begin{aligned}
\left|\dot{\alpha}_T^2(\omega)\right| &= \left|\left\langle f_T^0(\omega), \hat{f}_T^0(\omega)\right\rangle - \mu_T(\omega)\hat{\mu}_T(\omega)\right| \\
&= \left|\left\langle f_T^0(\omega) - \mu_T(\omega)\,\mathrm{Id}, \hat{f}_T^0(\omega) - \hat{\mu}_T(\omega)\,\mathrm{Id}\right\rangle\right| \\
&\leq \sqrt{\left\| f_T^0(\omega) - \mu_T(\omega)\,\mathrm{Id}\right\|^2}\sqrt{\left\| \hat{f}_T^0(\omega) - \hat{\mu}_T(\omega)\,\mathrm{Id}\right\|^2} \\
&\leq \delta_T(\omega)\hat{\delta}_T(\omega)
\end{aligned}$$

Where $\delta_T(\omega) = \sqrt{\delta_T^2(\omega)}$ and $\hat{\delta}_T(\omega) = \sqrt{\hat{\delta}_T^2(\omega)}$. We will now prove mean square convergence of $\dot{\alpha}_T^2(\omega)$ to $\hat{\alpha}_T^2(\omega)$. Due to (A.20), it suffices to show that $\mathrm{Var}(\dot{\alpha}_T^2(\omega)) \to 0$:

$$\begin{aligned}
\mathrm{Var}(\dot{\alpha}_T^2(\omega)) \quad &\leq 2\,\mathrm{Var}\left(\left\langle f_T^0(\omega), \hat{f}_T^0(\omega)\right\rangle\right) + 2\,\mathrm{Var}\left(\mu_T(\omega)\hat{\mu}_T(\omega)\right) \\
&= \underbrace{2\,\mathrm{Var}\left(\left\langle f_T^0(\omega), \hat{f}_T^0(\omega)\right\rangle\right)}_{[14]} + \underbrace{2\mu_T^2(\omega)\,\mathrm{Var}(\mu_T(\omega))}_{[15]}
\end{aligned}$$

Now,

$$([15]) \to 0$$

due to lemma 2.2 and lemma 2.5. To prove null convergence of ([14]), we first examine the term $\left\langle f_T^0(\omega), \hat{f}_T^0(\omega)\right\rangle$:

$$\begin{aligned}
&\left\langle f_T^0(\omega), \hat{f}_T^0(\omega)\right\rangle \\
=&\frac{1}{p_T}\,\mathrm{tr}\left[\left(f_T^0(\omega)\right)\left(\hat{f}_T^0(\omega)\right)^*\right] \\
=&\frac{1}{p_T}\,\mathrm{tr}\left[\left(\frac{1}{m_T}\sum_k \mathrm{E}\,I_T(\tilde{\omega}_k)\right)\left(\frac{1}{m_T}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2} I_T^*(\tilde{\omega}_k)\right)\right]
\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{p_T}\operatorname{tr}\left[\left(\frac{1}{m_T}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\mathrm{E}\, y_T(\tilde{\omega}_k)y_T^*(\tilde{\omega}_k)\right)\right.\\
&\qquad\left.\times\left(\frac{1}{m_T}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}y_T(\tilde{\omega}_k)y_T^*(\tilde{\omega}_k)\right)\right]\\
=&\frac{1}{p_T}\sum_{i,j=1}^{p_T}\left[\left(\frac{1}{m_T}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\mathrm{E}\, y_i(\tilde{\omega}_k)y_j^*(\tilde{\omega}_k)\right)\right.\\
&\qquad\left.\times\left(\frac{1}{m_T}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}y_i(\tilde{\omega}_k)y_j^*(\tilde{\omega}_k)\right)\right]\\
=&\frac{1}{p_T}\sum_{i,j=1}^{p_T}\left[\left(\frac{1}{m_T}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\left(\lambda_{ij}(\tilde{\omega}_k)+\mathrm{O}\left(\frac{p_T}{T}\right)\right)\right)\right.\\
&\qquad\left.\times\left(\frac{1}{m_T}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}y_i(\tilde{\omega}_k)y_j^*(\tilde{\omega}_k)\right)\right]\\
=&\frac{1}{p_T}\sum_{i=1}^{p_T}\left[\left(\frac{1}{m_T}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\left(\lambda_{ii}(\tilde{\omega}_k)+\mathrm{O}\left(\frac{p_T}{T}\right)\right)\right)\right.\\
&\qquad\left.\times\left(\frac{1}{m_T}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}y_i(\tilde{\omega}_k)y_j^*(\tilde{\omega}_k)\right)\right]\\
&+\frac{1}{p_T}\sum_{\substack{i,j=1\\ i\neq j}}^{p_T}\left[\left(\frac{1}{m_T}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\mathrm{O}\left(\frac{p_T}{T}\right)\right)\right.\\
&\qquad\left.\times\left(\frac{1}{m_T}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}y_i(\tilde{\omega}_k)y_j^*(\tilde{\omega}_k)\right)\right]
\end{aligned}
$$

Now, the variance of the term on the last two lines vanishes asymptotically due to assumptions 2.4 and 2.6. We thus focus on the term on the two lines before, the variance of which we will discuss using the

abbreviation

$$\bar{\lambda_i}(\omega) = \frac{1}{m_T} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \lambda_{ii}(\tilde{\omega}_k)$$

Omitting the $\mathrm{O}\left(\frac{p_T}{T}\right)$, that term is

$$= \frac{1}{p_T} \sum_{i=1}^{p_T} \bar{\lambda_i}(\omega) \left(\frac{1}{m_T} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} y_i(\tilde{\omega}_k) y_i^*(\tilde{\omega}_k) \right)$$

and we can proceed to examining its variance:

$$\begin{aligned}
& \mathrm{Var}\left\{ \frac{1}{p_T} \sum_{i=1}^{p_T} \bar{\lambda_i}(\omega) \left(\frac{1}{m_T} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} y_i(\tilde{\omega}_k) y_i^*(\tilde{\omega}_k) \right) \right\} \\
= \quad & \mathrm{Var}\left\{ \frac{1}{m_T} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \left(\frac{1}{p_T} \sum_{i=1}^{p_T} \bar{\lambda_i}(\omega) |y_i(\tilde{\omega}_k)|^2 \right) \right\} \\
= \quad & \underbrace{\frac{1}{m_T^2} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \mathrm{Var}\left\{ \frac{1}{p_T} \sum_{i=1}^{p_T} \bar{\lambda_i}(\omega) |y_i(\tilde{\omega}_k)|^2 \right\}}_{[16]} \\
& + \frac{1}{m_T^2} \sum_{\substack{k,l=-(m_T-1)/2 \\ k \neq l}}^{(m_T-1)/2} \mathrm{Cov}\left\{ \left(\frac{1}{p_T} \sum_{i=1}^{p_T} \bar{\lambda_i}(\omega) |y_i(\tilde{\omega}_k)|^2 \right), \right. \\
& \qquad \left. \left(\frac{1}{p_T} \sum_{i=1}^{p_T} \bar{\lambda_i}(\omega) |y_i(\tilde{\omega}_l)|^2 \right) \right\}
\end{aligned}$$

Now, the term on the last two lines converges to zero because

$$\mathrm{Cov}(|y_i(\tilde{\omega}_k)|^2, |y_i(\tilde{\omega}_l)|^2) = \mathrm{O}\left(\frac{p_T}{T}\right)$$

and assumption 2.4 guarantees that the weights $\bar{\lambda_i}(\omega)$ do not exhibit a too irregular behavior. We can thus focus on [16] and show its null convergence using lemma B.14 and assumption 2.4:

$$[16] \approx \frac{1}{m_T} \mathrm{Var}\left\{ \frac{1}{p_T} \sum_{i=1}^{p_T} \bar{\lambda_i}(\omega) |y_i(\omega)|^2 \right\}$$

$$
\begin{aligned}
&\leq \frac{1}{m_T}\,\mathrm{E}\left\{\frac{1}{p_T}\sum_{i=1}^{p_T}\bar{\lambda}_i(\omega)|y_i(\omega)|^2\right\}^2 \\
&= \frac{1}{m_T}\,\mathrm{E}\left\{\left(\frac{1}{p_T}\sum_{i=1}^{p_T}\left(\mathrm{E}\,|y_i(\omega)|^2+\mathrm{O}\left(\frac{p_T}{T}\right)\right)|y_i(\omega)|^2\right)\right\}^2 \\
&\approx \frac{1}{m_T}\,\mathrm{E}\left\{\frac{1}{p_T}\sum_{i=1}^{p_T}\left(\left(\mathrm{E}\,|y_i(\omega)|^2\right)|y_i(\omega)|^2\right)\right\}^2 \\
&\leq \frac{1}{m_T}\,\mathrm{E}\left\{\left(\frac{1}{p_T}\sum_{i=1}^{p_T}\left(\mathrm{E}\,|y_i(\tilde{\omega}_k)|^2\right)^2\right)\left(\frac{1}{p_T}\sum_{i=1}^{p_T}|y_i(\omega)|^4\right)\right\} \\
&\leq \frac{1}{m_T}\left(\frac{1}{p_T}\sum_{i=1}^{p_T}\mathrm{E}\,|y_i(\omega)|^4\right)^2 \\
&\leq \frac{1}{m_T}K_2
\end{aligned}
$$

Summarizing what we have shown so far: we have constructed a new parameter $\dot{\alpha}_T^2(\omega)$ that converges in mean square to $\alpha_T^2(\omega)$. Now, we will give an explicit formula for the optimal random coefficients $\dot{r_T}(\omega)$ and $\dot{s_T}(\omega)$. These are given by the orthogonal projection according to the scalar product $\langle\cdot,\cdot\rangle$ of the expected averaged periodogram $f_T^0(\omega)$ onto the plane spanned by the identity matrix and the averaged periodogram $\hat{f}_T^0(\omega)$. We need an orthonormal basis for this plane. As the identity matrix has norm 1, it will be chosen as the first basis vector. We can easily find a second:

$$
\begin{aligned}
\left\langle \hat{f}_T^0(\omega)-\hat{\mu}_T(\omega)\,\mathrm{Id},\mathrm{Id}\right\rangle &= \left\langle \hat{f}_T^0(\omega),\mathrm{Id}\right\rangle - \langle\hat{\mu}_T(\omega)\,\mathrm{Id},\mathrm{Id}\rangle \\
&= \hat{\mu}_T(\omega)-\hat{\mu}_T(\omega)\,\langle\mathrm{Id},\mathrm{Id}\rangle \\
&= 0
\end{aligned}
$$

Thus, the orthonormal basis is given by $\left\{\mathrm{Id},\frac{\hat{f}_T^0(\omega)-\hat{\mu}_T(\omega)\,\mathrm{Id}}{\left\|\hat{f}_T^0(\omega)-\hat{\mu}_T(\omega)\,\mathrm{Id}\right\|}\right\}$. We obtain the following formula for the random oracle estimator:

$$
\begin{aligned}
\hat{\varphi}_T^{**}(\omega) &= \left\langle f_T^0(\omega),\mathrm{Id}\right\rangle\mathrm{Id} \\
&\quad + \left\langle f_T^0(\omega),\frac{\hat{f}_T^0(\omega)-\hat{\mu}_T(\omega)\,\mathrm{Id}}{\left\|\hat{f}_T^0(\omega)-\hat{\mu}_T(\omega)\,\mathrm{Id}\right\|}\right\rangle\frac{\hat{f}_T^0(\omega)-\hat{\mu}_T(\omega)\,\mathrm{Id}}{\left\|\hat{f}_T^0(\omega)-\hat{\mu}_T(\omega)\,\mathrm{Id}\right\|}
\end{aligned}
$$

$$
\begin{aligned}
&= \mu_T(\omega)\,\mathrm{Id} \\
&\quad + \frac{\left\langle f_T^0(\omega), \hat{f}_T^0(\omega) \right\rangle - \mu_T(\omega)\hat{\mu}_T(\omega)}{\left\| \hat{f}_T^0(\omega) - \hat{\mu}_T(\omega)\,\mathrm{Id} \right\|^2} \left(\hat{f}_T^0(\omega) - \hat{\mu}_T(\omega)\,\mathrm{Id} \right) \\
&= \mu_T(\omega)\,\mathrm{Id} + \frac{\dot{\alpha}_T^2(\omega)}{\hat{\delta}_T^2(\omega)} \left(\hat{f}_T^0(\omega) - \hat{\mu}_T(\omega)\,\mathrm{Id} \right)
\end{aligned}
$$

which leads to the following random shrinkage parameters

$$
\begin{aligned}
\dot{r_T}(\omega) &= \left(\mu_T(\omega) - \frac{\dot{\alpha}_T^2(\omega)}{\hat{\delta}_T^2(\omega)} \hat{\mu}_T(\omega) \right) \\
\dot{s_T}(\omega) &= \frac{\dot{\alpha}_T^2(\omega)}{\hat{\delta}_T^2(\omega)}
\end{aligned}
$$

It remains to be shown that

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

This is shown by plugging in $\hat{\varphi}_T^{**}(\omega)$ for $\hat{\varphi}_T^*(\omega)$ in equation (A.18) and repeating exactly the same steps as in the proof of theorem 2.6, which we will not repeat here.

□

A.2 Proofs for chapter 3

A.2.1 Proof of theorem 3.2

We want to derive the optimal shrinkage intensity $\zeta_T(\omega)$, which is the solution of the optimization problem (3.16). According to (3.18), any local extremum of the function $R(z)$ is a minimum. Thus, $\zeta_T^*(\omega)$ is the value obtained for z by setting the first derivative equal to zero:

$$
\begin{aligned}
& 0 = R'(\zeta_T^*(\omega)) \\
\Leftrightarrow \quad & 0 = 2 \sum_{i,j=1}^{p} \Big\{ \zeta_T^*(\omega) \operatorname{Var} \hat{f}_{ij}^1(\omega) - (1 - \zeta_T^*(\omega)) \operatorname{Var} \hat{f}_{ij}^0(\omega) \\
& \quad + (1 - 2\zeta_T^*(\omega)) \Re \left(\operatorname{Cov} \left(\hat{f}_{ij}^1(\omega), \hat{f}_{ij}^0(\omega) \right) \right)
\end{aligned}
$$

$$
\begin{aligned}
& +\zeta_T^*(\omega)\left|f_{ij}^1(\omega)-f_{ij}^0(\omega)\right|^2\Big\} \\
\Leftrightarrow\quad & 2\zeta_T^*(\omega)\sum_{i,j=1}^{p}\Big\{\operatorname{Var}\hat{f}_{ij}^1(\omega)+\operatorname{Var}\hat{f}_{ij}^0(\omega) \\
& -2\Re\left(\operatorname{Cov}\left(\hat{f}_{ij}^1(\omega),\hat{f}_{ij}^0(\omega)\right)\right)+\left|f_{ij}^1(\omega)-f_{ij}^0(\omega)\right|^2\Big\} \\
& =2\sum_{i,j=1}^{p}\left\{\operatorname{Var}\hat{f}_{ij}^0(\omega)-2\Re\left(\operatorname{Cov}\left(\hat{f}_{ij}^1(\omega)-\hat{f}_{ij}^0(\omega)\right)\right)\right\} \\
\Leftrightarrow\quad & 2\zeta_T^*(\omega)\sum_{i,j=1}^{p}\left\{\operatorname{Var}\left(\hat{f}_{ij}^1(\omega)-\hat{f}_{ij}^0(\omega)\right)+\left|f_{ij}^1(\omega)-f_{ij}^0(\omega)\right|^2\right\} \\
& =2\sum_{i,j=1}^{p}\left\{\operatorname{Var}\hat{f}_{ij}^0(\omega)-2\Re\left(\operatorname{Cov}\left(\hat{f}_{ij}^1(\omega)-\hat{f}_{ij}^0(\omega)\right)\right)\right\} \\
\Rightarrow\quad & \zeta_T^*(\omega)=\frac{\sum_{i,j=1}^{p}\left(\operatorname{Var}\hat{f}_{ij}^0(\omega)-2\Re\operatorname{Cov}\left(\hat{f}_{ij}^1(\omega),\hat{f}_{ij}^0(\omega)\right)\right)}{\sum_{i,j=1}^{p}\left(\operatorname{Var}\left(\hat{f}_{ij}^1(\omega)-\hat{f}_{ij}^0(\omega)\right)+\left|f_{ij}^1(\omega)-f_{ij}^0(\omega)\right|^2\right)}
\end{aligned}
$$

□

A.2.2 Proof of theorem 3.3

Theorem 3.3 is proven using two technical lemmata which we will give immediately after the proof, which we give first:
If we multiply (3.19) by m_T, we obtain

$$
\begin{aligned}
& m_T\zeta_T^*(\omega)= \\
& \frac{\sum_{i,j}\left(\operatorname{Var}\left(\sqrt{m_T}\hat{f}_{ij}^0(\omega)\right)-2\Re\left(\operatorname{Cov}\left(\sqrt{m_T}\hat{f}_{ij}^1(\omega),\sqrt{m_T}\hat{f}_{ij}^0(\omega)\right)\right)\right)}{\sum_{i,j=1}^{p}\left(\operatorname{Var}(\hat{f}_{ij}^1(\omega)-\hat{f}_{ij}^0(\omega))+\left|f_{ij}^1(\omega)-f_{ij}^0(\omega)\right|^2\right)}
\end{aligned}
\tag{A.21}
$$

$\hat{f}_{ij}^0(\omega)$ and $\hat{f}_{ij}^1(\omega)$ are consistent estimators of $f_{ij}^0(\omega)$ and $f_{ij}^1(\omega)$, respectively. This means that

$$
\operatorname{Var}\left(\hat{f}_{ij}^1(\omega)-\hat{f}_{ij}^0(\omega)\right)=\mathrm{o}\,(1) \tag{A.22}
$$

Using assumption 3.1 and (A.22), we obtain that the denominator of the right side of (A.21) is $\mathrm{O}\,(1)$. The numerator of the right side of (A.21)

is $\pi(\omega) - 2\Re(\rho(\omega)) + \mathrm{O}\left(\frac{m_T}{T}\right)$ according to lemmata A.1 and A.2. This yields

$$m_T \zeta_T^*(\omega) = \frac{\pi(\omega) + \rho(\omega) + \mathrm{O}\left(\frac{m_T}{T}\right)}{\gamma(\omega)} \tag{A.23}$$

or, equivalently,

$$\zeta_T^*(\omega) = \frac{1}{m_T} \frac{\pi(\omega) + \rho(\omega)}{\gamma(\omega)} + \mathrm{O}\left(\frac{1}{T}\right) \tag{A.24}$$

□

A.2.2.1 Lemmata needed for A.2.2

Lemma A.1.

$$\mathrm{Var}\left(\sqrt{m_T}\hat{f}_{ij}^0(\omega)\right) = \pi_{ij}(\omega) + \mathrm{O}\left(\frac{m_T}{T}\right) \tag{A.25}$$

Proof.

$$\begin{aligned}
\mathrm{Var}(\sqrt{m_T}\hat{f}_{ij}^0(\omega)) &= \mathrm{Var}\left(\frac{1}{\sqrt{m_T}} \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} I_{ij}(\tilde{\omega}_k)\right) \\
&= \frac{1}{m_T} \sum_{k,l=-(m_T-1)/2}^{(m_T-1)/2} \mathrm{Var}(I_{ij}(\tilde{\omega}_k)) \\
&\quad + \frac{1}{m_T} \sum_{\substack{k,l=-(m_T-1)/2 \\ k\neq l}}^{(m_T-1)/2} \mathrm{Cov}(I_{ij}(\tilde{\omega}_k) I_{ij}(\tilde{\omega}_l)) \\
&= |f_{ij}(\omega)|^2 + \mathrm{O}\left(\frac{1}{T}\right) + \frac{1}{m_T}\,\mathrm{O}\left(\frac{m_T^2}{T}\right) \\
&= |f_{ij}(\omega)|^2 + \mathrm{O}\left(\frac{m_T}{T}\right)
\end{aligned}$$

This proves equation (A.25) and yields that

$$\pi_{ij}(\omega) = |f_{ij}(\omega)|^2 \tag{A.26}$$

□

Lemma A.2.

$$\operatorname{Cov}\left(\sqrt{m_T}\hat{f}_{ij}^{1}(\omega), \sqrt{m_T}\hat{f}_{ij}^{0}(\omega)\right) = \rho_{ij}(\omega) + \mathrm{O}\left(\frac{m_T}{T}\right) \tag{A.27}$$

Proof.

$$\begin{aligned}
&\operatorname{Cov}\left(\sqrt{m_T}\hat{f}_{ij}^{1}(\omega), \sqrt{m_T}\hat{f}_{ij}^{0}(\omega)\right) \\
=&\operatorname{Cov}\left(\frac{1}{\sqrt{m_T}}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2} b_i b_j I_{00}(\tilde{\omega}_k), \frac{1}{\sqrt{m_T}}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2} I_{ij}(\tilde{\omega}_k)\right) \\
=&\frac{1}{m_T} b_i b_j \left\{ \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \operatorname{Cov}\left(I_{00}(\tilde{\omega}_k), I_{ij}(\tilde{\omega}_k)\right) \right. \\
&\left. + \sum_{\substack{k,l=-(m_T-1)/2 \\ k\neq l}}^{(m_T-1)/2} \operatorname{Cov}\left(I_{00}(\tilde{\omega}_k), I_{ij}(\tilde{\omega}_l)\right) \right\} \\
=&\frac{1}{m_T} b_i b_j \sum_{k=-(m_T-1)/2}^{(m_T-1)/2} \operatorname{Cov}(I_{00}(\tilde{\omega}_k), I_{ij}(\tilde{\omega}_k)) + \mathrm{O}\left(\frac{m_T}{T}\right)
\end{aligned} \tag{A.28}$$

We will now focus on the covariance term in (A.28) and decompose it using lemma B.13 and lemma B.1 :

$$\begin{aligned}
&\operatorname{Cov}(I_{00}(\omega), I_{ij}(\omega)) \\
=&\operatorname{Cov}\left(d_0(\omega)\overline{d_0(\omega)}, d_i(\omega)\overline{d_j(\omega)}\right) \\
=&\operatorname{Cov}\left(d_0(\omega), d_i(\omega)\right) \operatorname{Cov}\left(\overline{d_0(\omega)}, \overline{d_j(\omega)}\right) \\
&+ \operatorname{Cov}\left(d_0(\omega), \overline{d_j(\omega)}\right) \operatorname{Cov}\left(\overline{d_0(\omega)}, d_i(\omega)\right) \\
=&\mathrm{E}\left(d_0(\omega), \overline{d_i(\omega)}\right) \mathrm{E}\left(\overline{d_0(\omega)}, d_j(\omega)\right) + \mathrm{O}\left(\frac{1}{T}\right) \mathrm{O}\left(\frac{1}{T}\right) \\
=&f_{0i}(\omega) f_{j0}(\omega) + \mathrm{O}\left(\frac{1}{T}\right) + \mathrm{O}\left(\frac{1}{T^2}\right) \\
=&f_{0i}(\omega) f_{j0}(\omega) + \mathrm{O}\left(\frac{1}{T}\right)
\end{aligned} \tag{A.29}$$

Now, according to (3.12)

$$b_i = \beta_i + \mathrm{O}\left(\frac{1}{T}\right)$$

Combining this with (A.28) and (A.29) yields thus

$$\mathrm{Cov}\left(\sqrt{m_T}\hat{f}^1_{ij}(\omega), \sqrt{m_T}\hat{f}^0_{ij}(\omega)\right) = \beta_i\beta_j f_{0i}(\omega)f_{j0}(\omega) + \mathrm{O}\left(\frac{m_T}{T}\right) \quad \text{(A.30)}$$

which proves (A.27) and at the same time yields

$$\rho_{ij}(\omega) = \beta_i\beta_j f_{0i}(\omega)f_{j0}(\omega) \quad \text{(A.31)}$$

□

A.2.3 Proof of lemma 3.4

Proof. According to (Brockwell & Davis 1987, theorem 10.3.2), we have

$$\mathrm{Var}\, I_{ij}(\tilde{\omega}_k) = |f_{ij}(\tilde{\omega}_k)|^2 + \mathrm{O}\left(1/\sqrt{T}\right) \quad \text{(A.32)}$$

and

$$\mathrm{Cov}(I_{ij}(\tilde{\omega}_k), I_{ij}(\tilde{\omega}_l)) = \mathrm{O}\left(1/T\right) \quad \text{(A.33)}$$

for $0 < \tilde{\omega}_k \neq \tilde{\omega}_l < \pi$. Furthermore, $\hat{f}^0_{ij}(\omega) = \mathrm{E}\, I_{ij}(\tilde{\omega}_k) + \mathrm{O}_p\left(1/m_T\right)$ for all $\tilde{\omega}_k$ that are smoothed over. This allows us to write

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

$$+\frac{1}{m_T}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}\mathrm{O}_p\left(\frac{1}{m_T}\right)|I_{ij}(\tilde{\omega}_k)-\mathrm{E}\,I_{ij}(\tilde{\omega}_k)|^2$$

Now, according to (A.32), $|I_{ij}(\tilde{\omega}_k)-\mathrm{E}\,I_{ij}(\tilde{\omega}_k)|^2$ is $\mathrm{O}_p(1)$, which permits us to drop the last two lines in the above calculations. We retain

$$\frac{1}{m_T}\sum_{k=-(m_T-1)/2}^{(m_T-1)/2}|I_{ij}(\tilde{\omega}_k)-\mathrm{E}\,I_{ij}(\tilde{\omega}_k)|^2$$

which, because of (A.32) in conjunction with (A.33), converges to $|f_{ij}(\omega)|^2=\pi_{ij}(\omega)$ in probability. □

A.2.4 Proof of lemma 3.5

Proof. b_i, b_j, $\hat{f}^0_{0i}(\omega)$ and $\hat{f}^0_{j0}(\omega)$ are consistent estimators of $\beta_i, \beta_j, f_{0i}(\omega)$ and $f_{j0}(\omega)$. This yields, in conjunction with (A.31), the result. □

APPENDIX B

Technical tools

This appendix provides us with some necessary technical tools and lemmata which are needed for the proofs of the theorem in the main part of the thesis.

B.1 Properties of Fourier transforms

Lemma B.1. For $i \neq j$, we have that

$$\mathrm{E}\, d_i(\omega_1) d_j(\omega_2) = \mathrm{O}\left(\frac{1}{T}\right) \tag{B.1}$$

where the null convergence is uniform in $\{\omega_1, \omega_2\} \in (0, 2\pi] \times (0, 2\pi]$.

Proof. The proof of this can be found in Shumway & Stoffer (2000, p. 275ff). □

Lemma B.2.

$$\mathrm{E}\, |y_i(\omega)|^2 = \lambda_i(\omega) + \mathrm{O}\left(\frac{p_T}{T}\right) \tag{B.2}$$

and, for $i \neq j$,

$$\mathrm{E}\, |y_i(\omega_1) y_j(\omega_2)| = \mathrm{O}\left(\frac{p_T}{T}\right). \tag{B.3}$$

Proof. Equation (B.2) follows from lemma B.5 in conjunction with lemma B.7. Equation (B.3) follows from lemma B.1 in conjunction with lemma B.7. □

Lemma B.3. For $\tilde{\omega}_k \neq \pm\tilde{\omega}_l \pmod{2\pi}$, we have that

$$\mathrm{E}\, d_i(\tilde{\omega}_k) d_i(\tilde{\omega}_l) = \mathrm{O}\left(\frac{1}{T}\right) \tag{B.4}$$

where the null convergence is uniform in $\omega \in (0, 2\pi]$.

Proof. The proof of this can be found in Shumway & Stoffer (2000, p. 275ff). □

Lemma B.4.

$$\mathrm{E}\, y_i(\tilde{\omega}_k) y_i(\tilde{\omega}_l) = \mathrm{O}\left(\frac{p_T}{T}\right)$$

Proof. This follows from lemma B.3 in conjunction with lemma B.7. □

Lemma B.5. Let $(X_t)_{t\in\mathbb{Z}}$ be a stationary time series with autocovariance function $\gamma(\cdot)$. If the regularity condition

$$\sum_{h\in\mathbb{Z}} |h|\gamma(h) < \infty$$

is fulfilled, then for $\omega \neq 0 (\mod 2\pi)$

$$\mathrm{E}\, I_T(\omega) = f(\omega) + \mathrm{O}\left(\frac{1}{T}\right)$$

where the $\mathrm{O}\left(\frac{1}{T}\right)$ is uniform in ω.

Proof. The proof is given in the book of Brillinger (1975, p.123). □

B.2 Properties of rotated Fourier transforms

By $y_T(\omega)$ we denote the discrete Fourier transform at frequency ω that is rotated to the basis given by the eigenvectors of the true spectrum at frequency ω:

$$y_T(\omega) = \Gamma_T^*(\omega) d_T(\omega)$$

The components of $y_T(\omega)$ are, omitting the index T, denoted $y_i(\omega), y_j(\omega)$. This section will give insights on the behaviour of the ys under Kolmogorov asymptotics that goes beyond the insights gained in the previous section.

Lemma B.6. 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

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

Proof. The proof of this is found in virtually any good book on matrix algebra, e.g. in Horn & Johnson (1985). □

Lemma B.7 (change of convergence rate under double asymptotics). Suppose that we have a random quadratic $p_T \times p_T$ matrix

$$A_T = [a_{ij}]_{i,j\in 1,\ldots,p_T},$$

the dimension p_T of which grows as a function of a parameter T. Suppose furthermore that the elements a_{ij}, $i, j \in 1, \ldots, p_T$ are uniformly $\mathrm{O}\left(\frac{1}{T}\right)$. We are interested in the convergence rate of the elements of a $p_T \times p_T$ matrix $B_T = [b_{ij}]_{i,j\in 1,\ldots,p_T}$ that is obtained by the orthonormal transformation:

$$B_T = \Gamma_T^* A_T \Gamma_T$$

where $\Gamma_T = [\gamma_{ij}]_{i,j\in 1,\ldots,p_T}$ is a random orthogonal $p_T \times p_T$ matrix. Then

$$b_{ij} = \mathrm{O}\left(\frac{p_T}{T}\right) \tag{B.5}$$

where again the rate of convergence is uniform over $p_T \times p_T$.

Proof. The elements of B_T are obtained from A_T by

$$\begin{aligned} b_{ij} &= \sum_{s=1}^{p_T} \overline{\gamma_{si}} \sum_{u=1}^{p_T} a_{su}\gamma_{uj} \\ &\leq \sum_{s=1}^{p_T} |\gamma_{si}| \sum_{u=1}^{p_T} |a_{su}||\gamma_{uj}| \end{aligned} \tag{B.6}$$

Now, the fact that Γ_T is orthonormal means that

$$\sum_{i=1}^{p_T} |\gamma_{ij}|^2 = 1 \qquad \forall j \in 1, \ldots, p_T$$

and

$$\sum_{j=1}^{p_T} |\gamma_{ij}|^2 = 1 \qquad \forall i \in 1, \ldots, p_T$$

Applying B.14 yields

$$\begin{aligned} \sum_{i=1}^{p_T} |\gamma_{ij}| &\leq \sqrt{\sum_{i=1}^{p_T} |\gamma_{ij}|^2} \sqrt{p_T} \\ &= \sqrt{p_T} \end{aligned}$$

Inserting this into (B.6) yields the result. □

Lemma B.8.

$$\|I_T(\omega)\|^2 = \|y_T(\omega) y_T^*(\omega)\|^2 \tag{B.7}$$

Proof. Using the invariance of the Hilbert-Schmidt norm under orthonormal rotations (lemma B.6), we obtain

$$\begin{aligned} \|I_T(\omega)\|^2 &= \|\Gamma_T^*(\omega) I_T(\omega) \Gamma_T(\omega)\|^2 \\ &= \|\Gamma_T^*(\omega) d_T(\omega) d_T^*(\omega) \Gamma_T(\omega)\|^2 \\ &= \|y_T(\omega) y_T^*(\omega)\|^2 \end{aligned}$$

□

Lemma B.9 (Null-convergence of bias of periodogram in norm). Under Kolmogorov asymptotics, the periodogram matrix is still an asymptotically unbiased estimator of the spectral matrix. However, the convergence rate changes due to the fact that the dimension is a function of T. We obtain

$$\|\mathrm{E}\, I_T(\omega) - f_T(\omega)\|^2 = \mathrm{O}\left(\frac{p_T}{T^2}\right) \tag{B.8}$$

for the unrotated periodogram. If we rotate to the eigensystem, the convergence rate is the same:

$$\|\mathrm{E}\, y_T(\omega) y_T^*(\omega) - \Lambda_T(\omega)\|^2 = \mathrm{O}\left(\frac{p_T}{T^2}\right) \tag{B.9}$$

Proof. According to lemma B.5, each of the components of the periodogram fulfills the condition

$$\mathrm{E}\, I_{ij}(\omega) = f_{ij}(\omega) + \mathrm{O}\left(\frac{1}{T}\right)$$

Unlike under standard asymptotics, the number of elements in the matrices $I_T(\omega)$ and $f_T(\omega)$ grows dynamically. We obtain

$$\begin{aligned} \|\mathrm{E}\, I_T(\omega) - f_T(\omega)\|^2 &= \frac{1}{p_T} \sum_{i,j=1}^{p_T} |\mathrm{E}\, I_{ij}(\omega) - f_{ij}(\omega)|^2 \\ &= \frac{1}{p_T} \sum_{i,j=1}^{p_T} \mathrm{O}\left(\frac{1}{T^2}\right) \\ &= \frac{1}{p_T} \mathrm{O}\left(\frac{p_T^2}{T^2}\right) \\ &= \mathrm{O}\left(\frac{p_T}{T^2}\right) \end{aligned}$$

which proofs (B.8). Applying lemma B.6 yields (B.9). □

Lemma B.10.

$$\mathrm{E}\, y_T(\omega) y_T^*(\omega) = \Lambda(\omega) + \tilde{R}(\omega) \tag{B.10}$$

where

$$\left\| \tilde{R}(\omega) \right\|^2 = \mathrm{O}\left(\frac{p_T}{T^2}\right)$$

.

Proof. This follows immediately from lemma B.9, equation (B.9). □

B.3 Miscellaneous lemmata

Lemma B.11. 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 $\frac{u_T^2}{\hat{\delta}_T^2(\omega)^{\tau_1} \delta_T^2(\omega)^{\tau_2}} \leq 2(\hat{\delta}_T^2(\omega) + \delta_T^2(\omega))$ a.s., then

$$\mathrm{E}\, \frac{u_T^2}{\hat{\delta}_T^{\tau_1}(\omega) \delta_T^{\tau_2}(\omega)} \to 0. \tag{B.11}$$

Proof. For fixed $\epsilon > 0$, let $\mathcal{D}_\epsilon$ denote the set of T such that $\delta_T^2(\omega) < \epsilon/8$, and $\overline{\mathcal{D}_\epsilon}$ its complement. According to (2.21), $\hat{\delta}_T^2(\omega) - \delta_T^2(\omega) \to 0$ in quadratic mean. Thus, $\exists T_1$ such that for all $T \geq T_1$ we have

$$\mathrm{E}\left|\delta_T^2(\omega) - \hat{\delta}_T^2(\omega)\right| \leq \frac{\epsilon}{4}$$

Now, for every $T \geq T_1$ that lies within the set $\mathcal{D}_\epsilon$, we obtain

$$\begin{aligned}
\mathrm{E}\,\frac{u_T^2}{\hat{\delta}_T^{\tau_1}(\omega)\delta_T^{\tau_2}(\omega)} &\leq 2\left(\mathrm{E}\,\hat{\delta}_T^2(\omega) + \delta_T^2(\omega)\right) \\
&\leq 2\left(\mathrm{E}\,|\hat{\delta}_T^2(\omega) - \delta_T^2(\omega)| + 2\delta_T^2(\omega)\right) \\
&\leq 2\left(\frac{\epsilon}{4} + 2\frac{\epsilon}{8}\right) \\
&= \epsilon.
\end{aligned}$$

It remains to prove that (B.11) holds true for the elements of $\overline{\mathcal{D}_\epsilon}$, too. As $\mathrm{E}\,u_T^2 \to 0$, $\exists T_2$ such that, for all $T \geq T_2$,

$$\mathrm{E}\,u_T^2 \leq \frac{\epsilon^{\tau_1+\tau_2+1}}{2^{4\tau_1+3\tau_2+1}}$$

Let $\dot{K}$ denote an upper bound of $\delta_T^2(\omega)$. Such a bound exists due to lemma 2.2. The null convergence of $\delta_T^2(\omega) - \hat{\delta}_T^2(\omega)$ in quadratic mean implies null convergence in probability. Thus, $\exists T_3$ such that for all $T \geq T_3$

$$\mathbb{P}\left(|\delta_T^2(\omega) - \hat{\delta}_T^2(\omega)| \geq \frac{\epsilon}{16}\right) \leq \frac{4\epsilon}{16\dot{K} + \epsilon}$$

Now, for each $T \geq \max(T_2, T_3)$ that lies in $\overline{\mathcal{D}_\epsilon}$, we obtain

$$\begin{aligned}
&\mathrm{E}\,\frac{u_T^2}{\hat{\delta}_T^{\tau_1}(\omega)\delta_T^{\tau_2}(\omega)} \\
= \;& \mathrm{E}\left(\frac{u_T^2}{\hat{\delta}_T^{\tau_1}(\omega)\delta_T^{\tau_2}(\omega)}\right) 1_{\{\hat{\delta}_T^2(\omega) \leq \epsilon/16\}} \\
&+ \mathrm{E}\left(\frac{u_T^2}{\hat{\delta}_T^{\tau_1}(\omega)\delta_T^{\tau_2}(\omega)}\right) 1_{\{\hat{\delta}_T^2(\omega) > \epsilon/16\}} \\
\leq \;& \mathrm{E}\left(2(\delta_T^2(\omega) + \hat{\delta}_T^2(\omega))\right) 1_{\{\hat{\delta}_T^2(\omega) \leq \epsilon/16\}}
\end{aligned}$$

$$
\begin{aligned}
&+\left(\frac{16}{\epsilon}\right)^{\tau_1}\left(\frac{8}{\epsilon}\right)^{\tau_2}\mathrm{E}\left(u_T^2 1_{\{\hat{\delta}_T^2(\omega)>\epsilon/16\}}\right)\\
\leq\ & \mathrm{E}\left(2(\dot{K}+\frac{\epsilon}{16})\right)1_{\{\hat{\delta}_T^2(\omega)\leq\epsilon/16\}}\\
&+\left(\frac{16}{\epsilon}\right)^{\tau_1}\left(\frac{8}{\epsilon}\right)^{\tau_2}\mathrm{E}\,u_T^2\\
=\ & \left(2(\dot{K}+\frac{\epsilon}{16})\right)\mathbb{P}(\hat{\delta}_T^2(\omega)\leq\epsilon/16)\\
&+\left(\frac{16}{\epsilon}\right)^{\tau_1}\left(\frac{8}{\epsilon}\right)^{\tau_2}\mathrm{E}\,u_T^2\\
\leq\ & \left(2(\dot{K}+\frac{\epsilon}{16})\right)\mathbb{P}\left(|\hat{\delta}_T^2(\omega)-\delta_T^2(\omega)|\geq\frac{\epsilon}{16}\right)\\
&+\left(\frac{16}{\epsilon}\right)^{\tau_1}\left(\frac{8}{\epsilon}\right)^{\tau_2}\mathrm{E}\,u_T^2\\
\leq\ & \left(2(\dot{K}+\frac{\epsilon}{16})\right)\frac{4\epsilon}{16\dot{K}+\epsilon}\\
&+\left(\frac{16}{\epsilon}\right)^{\tau_1}\left(\frac{8}{\epsilon}\right)^{\tau_2}\frac{\epsilon^{\tau_1+\tau_2+1}}{2^{4\tau_1+3\tau_2+1}}\\
\leq\ & \epsilon
\end{aligned}
$$

Summarizing, we have shown that for every $\epsilon > 0 \exists T_1, T_2, T_3$ such that for all $T \geq \max(T_1, T_2, T_3)$

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

□

Lemma B.12. For any finite sequence $(a_1, \ldots, a_m)$ of non-negative a_k, the following inequality holds true:

$$\left(\sum_{k=1}^{m} a_k\right)^2 \leq m\sum_{k=1}^{m} a_k^2$$

Proof. We will make use of the relationship

$$2bc \leq b^2 + c^2$$

which is true because $(b-c)^2 \geq 0 \forall b, c \in \mathbb{R}$. Using this, we obtain

$$\left(\sum_{k=1}^{m} a_k\right)^2 = \sum_{k=1}^{m} a_k^2 + 2\sum_{k=1}^{m}\sum_{l=k+1}^{m}(a_k a_l)$$

$$\begin{aligned} &\leq \sum_{k=1}^{m} a_k^2 + \sum_{k=1}^{m} \sum_{l=k+1}^{m} (a_k^2 + a_l^2) \\ &= \sum_{k=1}^{m} a_k^2 + (m-1) \sum_{k=1}^{m} a_k^2 \\ &= m \sum_{k=1}^{m} a_k^2 \end{aligned}$$

□

Lemma B.13. Let (X_1, X_2, X_3, X_4) be a 4-variate normal random variable. Then we have

$$\begin{aligned} &\operatorname{Cov}(X_1 X_2, X_3 X_4) \\ = \; &\operatorname{Cov}(X_1, X_3) \operatorname{Cov}(X_2, X_4) + \operatorname{Cov}(X_1, X_4) \operatorname{Cov}(X_2, X_3) \end{aligned}$$

Proof. The proof is found in Brillinger (1975, p.21). □

Lemma B.14 (Cauchy-Schwarz inequality). Let $\langle \cdot, \cdot \rangle$ be a scalar product on a vector space V. Then, for all $x, y \in V$,

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle$$

Special cases of the Cauchy-Schwarz inequality that are frequently used in this thesis are

$$\operatorname{Cov}(A, B) \leq \sqrt{\operatorname{Var} A} \sqrt{\operatorname{Var} B}$$

and

$$\frac{1}{p} \sum_{i=1}^{p} a_i \leq \sqrt{\frac{1}{p} \sum_{i=1}^{p} |a_i|^2}$$

Proof. The proof can be found in Horn & Johnson (1985). □

Lemma B.15 (Cumulant lemma for eigth moments). Let

$$y_{ab}, a \in \{1, 2\}, b \in \{1, \ldots, 4\}$$

be any of the rotated fourier transforms at any frequency $\omega \in (0, 2\pi)$. Define

$$Y_1 \;:=\; y_{11} y_{12} y_{13} y_{14}$$

$$Y_2 \quad := \quad y_{21}y_{22}y_{23}y_{24}$$

Then

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

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 in Brillinger (1975), where we use that the ys are centered and Gaussian, due to which the cumulants of order 1 and those of order ≥ 3 are zero. □

Lemma B.16. Let $\lambda_1, \ldots, \lambda_4 \neq 0$ be Fourier frequencies such that, for $i \neq j$,

$$\lambda_i \pm \lambda_j \neq 0(\mod 2\pi). \tag{B.12}$$

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

$$\mathrm{cum}(I_T(\lambda_1), I_T(\lambda_2), I_T(\lambda_3), I_T(\lambda_4)) = \mathrm{O}\left(\frac{1}{T^4}\right) \tag{B.13}$$

Proof. Define $\lambda_{i1} := \lambda_1, \lambda_{i2} := -\lambda_i$. The cumulant in (B.13) can be written as

$$\frac{1}{T^4} \mathrm{cum}\left(d_T(\lambda_{11})d_T(\lambda_{12}), \ldots, d_T(\lambda_{41})d_T(\lambda_{42})\right).$$

According to Brillinger (1975, theorem 2.3.2), this can be rewritten as

$$\frac{1}{T^4} \sum_{\nu \in N} \left(\mathrm{cum}(\lambda_{ij}; ij \in \nu_1) \times \cdots \times \mathrm{cum}(\lambda_{ij}; ij \in \nu_p)\right) \tag{B.14}$$

where N is the set of all indecomposable partitions $\nu = \nu_1 \cup \ldots \cup \nu_p$ of the following table:

$$\begin{array}{cc} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \\ \lambda_{41} & \lambda_{42} \end{array}$$

Now, if the summability condition (4.3.10) from Brillinger (1975) is fulfilled, which we assume throughout the paper, we have, according to Brillinger (1975, theorem 4.3.2), that

$$\begin{aligned}&\operatorname{cum}(d_T(\mu_1), \ldots, d_T(\mu_k)) \\ &= \frac{1}{(2\pi)^{k-1}} \Delta_T \left(\sum_j \mu_j \right) f^{(k)}(\mu_1, \ldots, \mu_{k-1}) + \mathrm{O}(1)\end{aligned} \tag{B.15}$$

Here, $\Delta_T(\lambda) := \sum_{t=0}^{T-1} \exp(-i\lambda t)$, which equals zero for any Fourier frequency $\lambda \neq 0$. Furthermore, $f^{(k)}(\mu_1, \ldots, \mu_{k-1})$ is the spectrum of order k of the time series. For Gaussian time series, it is well known that

$$f^{(k)}(\mu_1, \ldots, \mu_{k-1}) = 0 \quad \text{if} \quad k > 2. \tag{B.16}$$

Now, let us look at all possible cumulants that occur as factors in (B.14). We must consider cumulants of order 1 to 8. Cumulants of order 1 will be of order $\mathrm{O}(1)$ as the $\Delta_T(\cdot)$ in (B.15) is zero. Cumulants of order 3 to 8 will be of order $\mathrm{O}(1)$, too, due to (B.16). Cumulants of order 2 are zero if the frequencies do not sum up to zero, which is, because of (B.12), only not the case if the two frequencies are on the same row in the table above. But this case does not occur because then the partition would be decomposable. This completes the proof. □

Lemma B.17. Let again $\lambda_1, \ldots, \lambda_4 \neq 0$ be Fourier frequencies such that, for $i \neq j$,

$$\lambda_i \pm \lambda_j \neq 0(\mod 2\pi). \tag{B.17}$$

Then, if the underlying time series is Gaussian, we have

$$\mathrm{E}\left(\dot{I}_T(\lambda_1) \dot{I}_T(\lambda_2) \dot{I}_T(\lambda_3) \dot{I}_T(\lambda_4) \right) = \mathrm{O}\left(\frac{1}{T^2} \right), \tag{B.18}$$

where $\dot{I}_T(\lambda_1) := I_T(\lambda_1) - \mathrm{E}\, I_T(\lambda_1)$.

Proof. We use the definition of the cumulant,

$$\operatorname{cum}(X_1, \ldots, X_r)$$

$$= \sum_{\nu \in N} (-1)^{p-1}(p-1)! \left(\mathrm{E} \prod_{j \in \nu_1} X_j \right) \cdots \left(\mathrm{E} \prod_{j \in \nu_p} X_j \right),$$

and the fact that $\mathrm{E}\, \dot{I}_T(\lambda) = 0$ to make the following decomposition:

$$\begin{aligned} & \operatorname{cum}(\dot{I}_T(\lambda_1), \dot{I}_T(\lambda_2), \dot{I}_T(\lambda_3), \dot{I}_T(\lambda_4)) \\ = \; & \mathrm{E}\, \dot{I}_T(\lambda_1)\dot{I}_T(\lambda_2)\dot{I}_T(\lambda_3)\dot{I}_T(\lambda_4) \\ & - \mathrm{E}\, \dot{I}_T(\lambda_1)\dot{I}_T(\lambda_2)\, \mathrm{E}\, \dot{I}_T(\lambda_3)\dot{I}_T(\lambda_4) \\ & - \mathrm{E}\, \dot{I}_T(\lambda_1)\dot{I}_T(\lambda_3)\, \mathrm{E}\, \dot{I}_T(\lambda_2)\dot{I}_T(\lambda_4) \\ & - \mathrm{E}\, \dot{I}_T(\lambda_1)\dot{I}_T(\lambda_4)\, \mathrm{E}\, \dot{I}_T(\lambda_2)\dot{I}_T(\lambda_3) \end{aligned}$$

The last three lines are each $\mathrm{O}\left(\frac{1}{T^2}\right)$. Making use of the shift invariance of cumulants Brillinger (1975, theorem 2.3.1), we obtain that

$$\begin{aligned} & \mathrm{E}\left(\dot{I}_T(\lambda_1)\dot{I}_T(\lambda_2)\dot{I}_T(\lambda_3)\dot{I}_T(\lambda_4) \right) \\ = \; & \operatorname{cum}(I_T(\lambda_1), I_T(\lambda_2), I_T(\lambda_3), I_T(\lambda_4)) + \mathrm{O}\left(\frac{1}{T^2} \right) \end{aligned}$$

Using Lemma B.16, we obtain the desired result. □

Bibliography

Bai, J. & Ng, S. (2002), 'Determining the number of factors in approximate factor models', *Econometrica* **70**(1), 191–221.

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

Bock, M. (1975), 'Minimax estimators of the mean of a multivariate normal distribution', *The Annals of Statistics* **3**(1), 209–218.

Böhm, H. & von Sachs, R. (2007), 'Shrinkage estimation in the frequency domain of multivariate time series', Discussion paper 0706, Institut de Statistique, Université Catholique de Louvain.

Brillinger, D. R. (1975), *Time Series. Data Analysis and Theory*, Holt, Reinhart and Winston.

Brillinger, D. R. (2002), 'John W. Tukey's work on time series and spectrum analysis', *The Annals of Statistics* **30**(6), 1595–1618.

Brockwell, P. & Davis, R. (1987), *Time Series: Theory and Methods*, Springer.

Brown, L. (1966), 'On the admissibility of invariant estimators of one or more location parameters', *Annals of Mathematical Statistics* **37**(5), 1087–1136.

Chernoff, H. (1952), 'A measure of asymptotic efficiency for tests of a hypothesis based on the sum of the observations', *Annals of Mathematical Statistics* **25**, 573–578.

Dahlhaus, R. (1983), 'Spectral analysis with tapered data', *Journal of Time Series Analysis* **4**, 163–175.

Dahlhaus, R. (1988), 'Small sample effects in time series analysis: a new asymptotic theory and a new estimate', *The Annals of Statistics* **16**(2), 808–841.

Deonier, R., Tavaré, S. & Waterman, M. (2005), *Computational Genome Analysis*, Springer.

Der, Z., Shumway, R. & Herrin, E. (2002), *Monitoring the Comprehensive Nuclear-Test-Ban Treaty: Data Processing and Infrasound*, Birkhauser.

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

Dey, D. & Srinivasan, C. (1986), 'Trimmed minimax estimators of a covariance matrix', *Annals of the Institute of Statistical Mathematics* **38**, 101–108.

Efron, B. & Morris, C. (1973), 'Stein's estimation rule and its competitors - an empirical bayes approach', *Journal of the American Statistical Association* **68**(341), 117–130.

Efron, B. & Morris, C. (1976), 'Multivariate empirical bayes and estimation of covariance matrices', *The Annals of Statistics* **7**(1), 22–32.

Elton, E. & Gruber, M. (1973), 'Estimating the dependence structure of share prices - implications for portfolio selection', *Journal of Finance* **28**(5), 1203–1232.

Evans, S. & Stark, P. (1996), 'Shrinkage estimator, Skorokhod's problem and stochastic integration by parts', *The Annals of Statistics* **24**(2), 809–815.

Fama, E. (1971), 'Risk, return, and equilibrium', *Journal of Political Economy* **79**(1), 30–55.

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

Frahm, G. & Jaekel, U. (2007), 'Tyler's M-estimator, random matrix theory, and generalized elliptical distributions with applications to finance', Preprint.

Guo, W., Ombao, H., Raz, J. & von Sachs, R. (2002), 'The SLEX model of a non-stationary random process', *Annals of the Institute of Statistical Mathematics* **54**(1), 171–200.

Guo, W., Ombao, H. & von Sachs, R. (2005), 'SLEX analysis of multivariate nonstationary time series', *Journal of the American Statistical Association* **100**(470), 519–531.

Haff, L. (1977), 'Minimax estimators for a multinormal precision matrix', *Journal of Multivariate Analysis* **7**, 374–385.

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

Haff, L. (1980), 'Empirical bayes estimation of the multivariate normal covariance matrix', *The Annals of Statistics* **8**, 586–597.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning*, Springer.

Hoerl, A. & Kennard, R. (1970*a*), 'Ridge regression: applications to nonorthogonal problems', *Technometrics* **12**(1), 69–82.

Hoerl, A. & Kennard, R. (1970*b*), 'Ridge regression: biased estimation for nonorthogonal problems', *Technometrics* **12**(1), 55–67.

Horn, R. & Johnson, C. (1985), *Matrix analysis*, Cambridge University Press.

James, G. & Sugar, C. (2003), 'Clustering for sparsely sampled functional data', *Journal of the American Statistical Association* **98**(462), 397–408.

James, W. & Stein, C. (1961), Estimation with quadratic loss, *in* 'Proc. Fourth Berkeley Symp. Math. Statist. Prob.', Vol. 1, pp. 361–380.

Johnstone, I. M. & Lu, A. Y. (2004), Sparse principal components analysis.

Jolliffe, I. (2002), *Principal Component Analysis*, Springer.

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

Kreiß, J. & Neuhaus, G. (2006), *Einführung in die Zeitreihenanalyse*, Springer.

Kubokawa, T. & Konno, Y. (1990), 'Estimating the covariance matrix and the generalized variance under a symmetric loss', *Annals of the Institute of Statistical Mathematics* **42**(2), 331–343.

Kubokawa, T. & Srivastava, M. (1999), 'Robust improvement in estimation of a covariance matrix in an elliptically contoured distribution', *The Annals of Statistics* **27**(2), 600–609.

Kullback, S. & Leibler, R. (1952), 'On information and sufficiency', *Annals of Mathematical Statistics* **22**(1), 79–86.

Ledoit, O. & Wolf, M. (2003), 'Improved estimation of the covariance matrix of stock returns with an application to portfolio selection', *Journal of Empirical Finance* **10**, 603–621.

Ledoit, O. & Wolf, M. (2004), 'A well-conditioned estimator for large-dimensional covariance matrices', *Journal of Multivariate Analysis* **88**, 365–411.

Lehmann, E. & Casella, G. (1998), *Theory of point estimation*, 2 edn, Springer.

Loh (1991), 'Estimating covariance matrices', *The Annals of Statistics* **19**(1), 283–296.

Marčenko, V. & Pastur, L. (1967), 'Distribution of the eigenvalues for some sets of random matrices', *Mathematics of the USSR Sbornik* **72**, 457–483.

Markowitz, H. (1952), 'Portfolio selection', *Journal of Finance* **7**, 77–91.

Neter, J. (1978), *Applied statistics*, Allyn and Bacon.

Ombao, H., Raz, J., von Sachs, R. & Strawderman, R. (2001), 'A simple GCV method of span selection for periodogram smoothing', *Biometrika* **88**.

Priestley, M. (1981), *Spectral Analysis and Time Series*, London: Academic Press.

Renyi, A. (1961), On measures of entropy and information, *in* 'Proc. Fourth Berkeley Symp. Math. Statist. Prob.', Vol. 1, pp. 547–561.

Sachs, L. & Hedderich, J. (2006), *Angewandte Statistik*, Springer.

Sharpe, W. (1963), 'A simplified model for portfolio analysis', *Management Science* **9**, 277–293.

Sheena, Y. (2002), 'On minimaxity of some orthogonally invariant estimators of bivariate normal dispersion matrix', *Journal of the Japanese Statistical Society* **32**(2), 193–207.

Shumway, R. H. & Stoffer, D. S. (2000), *Time Series Analysis and its Applications*, Springer.

Stein, C. (1956), Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *in* 'Proc. Third Berkeley Symp. Math. Statist. Prob.', Vol. 1, pp. 297–206.

Stein, C. (1975), 'Rietz lecture', 39th annual meeting IMS.

Taniguchi, M. & Kakizawa, Y. (2000), *Asymptotic theory of statistical inference for time series*, Springer.

Tikhonov, A. (1963), 'Solution of incorrectly formulated problems and regularization method', *Soviet math. Dokl.* **4**, 1053–1083.

Yin, Y. (1986), 'Limiting spectral distribution for a class of random matrices', *Journal of Multivariate Analysis* **20**, 50–68.

Zou, H., Hastie, T. & Tibshirani, R. (2004), 'Sparse principal component analysis', *Journal of Computational and Graphical Statistics* **15**(2), 265–286.