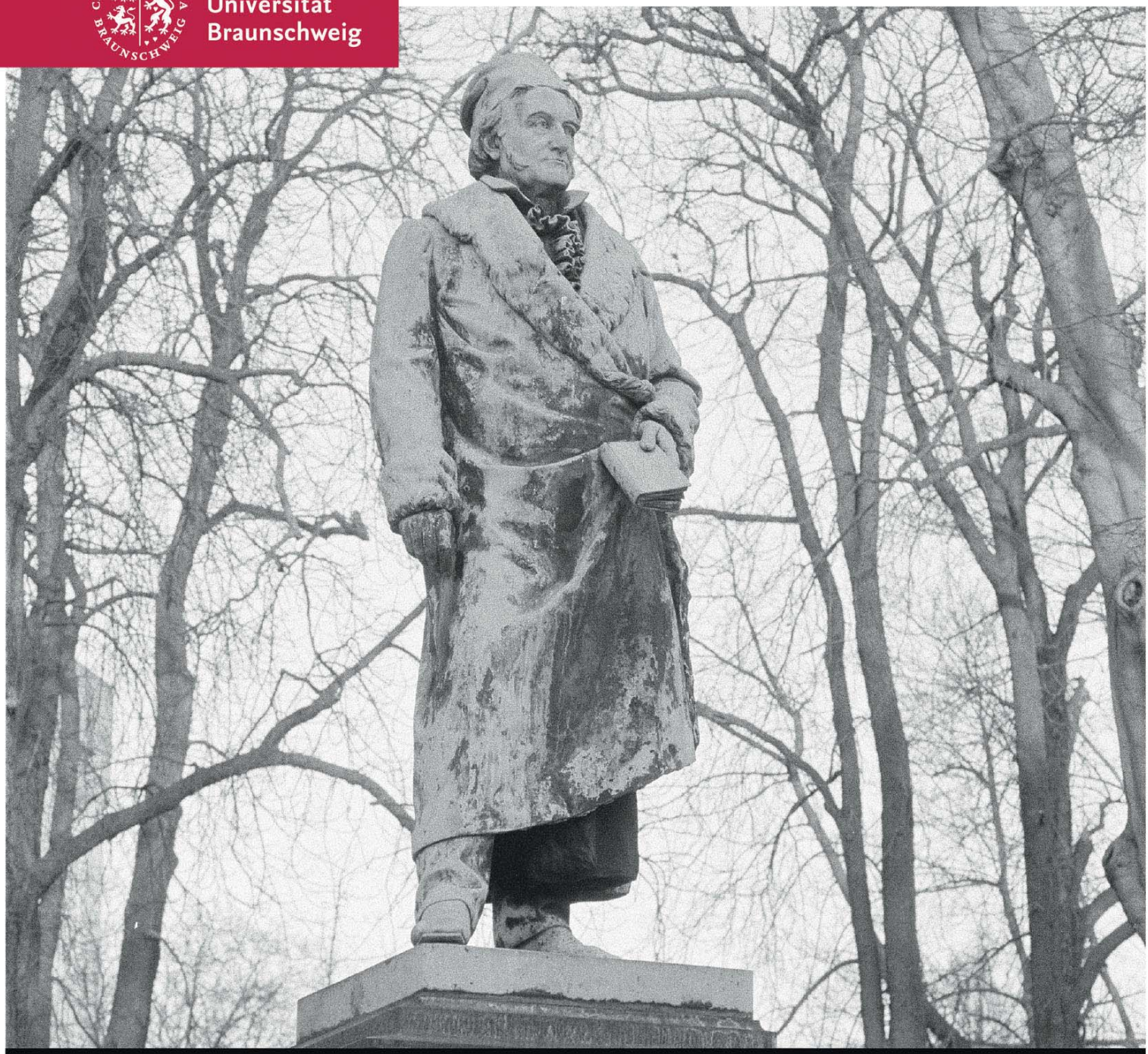
Technische Universität Braunschweig



Spectral-density-driven Bootstrap and Time Series Modeling on Dynamic Networks

Jonas Krampe

Cuvillier Verlag Göttingen
Internationaler wissenschaftlicher Fachverlag

Spectral-density-driven Bootstrap and
Time Series Modeling on Dynamic Networks

Spectral-density-driven Bootstrap and Time Series Modeling on Dynamic Networks

Von der
Carl-Friedrich-Gauß-Fakultät
der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigte Dissertation

von

Jonas Krampe
geboren am 6. Januar 1990 in Hameln

Eingereicht am: 19. April 2018
Disputation am: 18. Mai 2018

Referent: Prof. Dr. Jens-Peter Kreiß, TU Braunschweig
Koreferent: Prof. Dr. Efstathios Paparoditis, University of Cyprus

(2018)

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte bibliographische Daten sind im Internet
über http://dnb.d-nb.de abrufbar.
1. Aufl. - Göttingen: Cuvillier, 2018
Zugl.: (TU) Braunschweig, Univ., Diss., 2018

Nonnenstieg 8, 37075 Göttingen
Telefon: 0551-54724-0
Telefax: 0551-54724-21
www.cuvillier.de

1. Auflage, 2018
Gedruckt auf umweltfreundlichem, säurefreiem Papier aus nachhaltiger Forstwirtschaft.

ISBN 978-3-7369-9819-3
eISBN 978-3-7369-8819-4

Abstract

This thesis focuses on time series data. Its contribution is the development of new methods for time series analysis. In the first part we mainly consider univariate time series. Then, the second-order dependence structure of purely nondeterministic stationary process is described by the coefficients of the well-known Wold representation. These coefficients can be obtained by factorizing the spectral density of the process. This relation together with some spectral density estimator is used in order to obtain consistent estimators of these coefficients. A spectral-density-driven bootstrap for time series is then developed which uses the entire sequence of estimated moving average coefficients together with appropriately generated pseudo innovations in order to obtain a bootstrap pseudo time series. It is shown that if the underlying process is linear and if the pseudo innovations are generated by means of an i.i.d. wild bootstrap which mimics, to the necessary extent, the moment structure of the true innovations, this bootstrap proposal asymptotically works for a wide range of statistics. The relations of the proposed bootstrap procedure to some other bootstrap procedures, including the autoregressive-sieve bootstrap, are discussed. It is shown that the latter is a special case of the spectral-density-driven bootstrap, if a parametric autoregressive spectral density estimator is used. Simulations investigate the performance of the new bootstrap procedure in finite sample situations. Furthermore, a real-life data example is presented.

In the second part we consider multivariate time series on dynamic networks with a fixed number of vertices. Each component of the time series is assigned to a vertex of the underlying network. The dependency of the various components of the time series is modeled dynamically by means of the edges of the underlying network. We make use of a multivariate doubly stochastic time series framework, that is we assume linear processes for which the coefficient matrices are stochastic processes themselves. We explicitly allow for dependence in the dynamics of the coefficient matrices, including of course an i.i.d. structure as is typically assumed in random coefficients models. In this work asymptotic normality of simple statistics like the sample mean is investigated. Furthermore, autoregressive moving average models are defined in this framework. Different parameterizations of autoregressive models are considered. Some are more flexible, whereas others have only few parameters and are able to handle high-dimensional cases. Estimators of the parameters are discussed for various parameterizations of such network autoregressive models and how this can be used for forecast in this purposes. Interesting features of these processes are shown in simulations and the finite sample behavior of the forecast procedure is investigated.

Zusammenfassung

Zeitreihen und die Entwicklung neuer Methoden zur Zeitreihenanalyse stehen im Fokus dieser Arbeit. Der erste Teil konzentriert sich auf univariate Zeitreihen. Für diese gilt, dass die zweite Momentstruktur eines rein nichtdeterministischen Prozesses durch die Koeffizienten der Wold-Darstellung gegeben ist. Diese Koeffizienten lassen sich durch eine Faktorisierung der Spektraldichte erhalten. Dieser Zusammenhang zusammen mit einem Spektraldichteschätzer ermöglicht eine konsistente Schätzung der Wold-Koeffizienten. Auf Grundlage der geschätzten Wold-Koeffizienten kann ein Bootstrapverfahren entwickelt werden, welches allein durch den zugrunde liegenden Spektraldichteschätzer gesteuert werden kann, aber dennoch Pseudobeobachtungen im Zeitbereich erzeugt. Es wird im Folgenden als SDDB bezeichnet. Hierfür werden Pseudoinnovationen benötigt, welche sich durch einen u.i.v. Wildbootstrapansatz generieren lassen. Es wird gezeigt, dass, wenn der zugrunde liegende Prozess linear ist und die Pseudoinnovationen so erzeugt sind, dass sie die Momentenstruktur der wahren Innovationen ausreichend approximieren, das SDDB für eine Vielzahl an Statistiken konsistent ist. Weiter wird der Zusammenhang zu anderen Bootstrapverfahren erläutert und unter anderem wird gezeigt, dass das Autoregressivesievebootstrap ein Spezialfall des SDDB bei Verwendung eines parametrischen autoregressiven Spetrakldichteschätzers ist. Weiter wird in Simulationen die Leistungsfähigkeit des SDDB im endlichen Stichprobenfall untersucht. Außerdem wird die Anwendung des SDDB auf einen Realdatensatzes gezeigt.

Im zweiten Teil dieser Arbeit stehen multivariate Zeitreihen im Vordergrund, bzw. genauer, multivariate Zeitreihen auf dynamischen Netzwerken mit einer festen Anzahl an Knoten. Das heißt, dass jede Komponente der Zeitreihe einem Knoten des zugrundeliegenden Netzwerks zugeordnet werden kann. Die Abhängigkeitsstruktur der Komponenten der Zeitreihe wird hierbei durch Veränderungen in den Kanten dynamisch beeinflusst. Zur Modellierung solch eines Prozesses wird ein multivariater zweifachstochastischer Zeitreihenansatz verwendet. Dies bedeutet, dass für einen linearen Prozess die Koeffizientenmatrizen selbst stochastische Prozesse sind. Hierbei wird ausdrücklich auch Abhängigkeit für das zugrundeliegende Netzwerk zugelassen. Für ein u.i.v. Netzwerk ergibt sich der Spezialfall eines multivariaten Random Coefficient Models. Für autoregressive Zeitreihenmodelle auf Netzwerken werden für unterschiedliche Parametrisierungen konsistente Schätzmethoden präsentiert. Diese Schätzer können zur Vorhersage von solchen Zeitreihen genutzt werden. Die Leistungsfähigkeit dieser Vorhersagemethoden wird in Simulationen untersucht und die Anwendbarkeit auf einen Realdatensatz dargelegt.

Acknowledgment

First and foremost, I would like to express my deep gratitude to my thesis adviser Jens-Peter Kreiß for the true research spirit that he showed to me, his clear guidance, his patience as well as the freedom he gave me during the research process of my current studies. In addition, my extended appreciation goes towards his unwavering support. He always took the time to listen to my wishes and problems.

A special thanks goes to my co-advisor Efstathios Paparoditis whose constant encouragement and guidance helped me in the completion of this dissertation. He was always there to meet, talk about ideas and ask me critical questions to help me clearly think through my problems. Furthermore, I like to thank him for his generous hospitality during my stays in Cyprus.

My friend and co-worker Alexander Braumann deserves my profound gratitude for the many fruitful discussions and constant encouragement.

Furthermore, I am also grateful to all current and former members of the Institute for Mathematical Stochastics in Braunschweig for the remarkably pleasant working atmosphere in our offices as well as for the pleasant time we spent together during after-work hours.

I also thank Rudolf Suppes for his wide consultations about the English language. His friendship, knowledge, and hospitality have enlightened me in myriad ways.

I would like to express my heartfelt thanks to my invaluable, generous and loving friends. Especially, I like to thank my friends in Wolverines for not letting me forget my roots.

A special thanks goes to my brother David. His encouragement, advice and attention have been invaluable throughout my entire life.

I would like to express my sincere thanks to my family for providing great support and constant encouragement. Most of all, I like to thank my parents for raising me with the spirit to always pursue my interests and for blindly trusting in my ways.
Thank you!

Contents

List of Figures

List of Tables

1 Introduction

It has been proclaimed that 'data are the resource of the 21st century' [1]. Looking back in history, most people can agree on that oil was the most important natural resource in the 20th century. However, if we take a closer look at oil, we see that it is not very useful on its own. Surely, oil is flammable, but things like cars, planes, or chemical industry plants are needed to make oil the useful resource we know and therefore, to make it an important resource. Hence, tools and equipments are needed to create a benefit using the natural resource. The same situation applies to data. Without adequate tools data is nothing more than ones and zeroes. However, with the right methods data can be analyzed to extract useful information to gain new insights. Like diesel needs a different consumption engine than gasoline some type of data requires different methods than other types. This work focuses on time series data. In almost all everyday life situations time series data can occur. For instance, sensors monitoring the weather and collecting the air and ground temperature or the amount of rainfall, sensors monitoring vital parameters like the heart rate of a person, or smartphones counting a person's steps, all results in time series data. Time series data is also given by financial markets, e.g. stock market prices, or by social-economical data, e.g. unemployment rates or gross national products.

An important characteristic of time series data is that different data points are usually not independent from each other. That is why such a type of data requires special treatment. An example for this situation is the bootstrap method, see section 1.2 for an introduction. If this method is applied in the same manner as it is applied for data which consists of independent data points, the bootstrap method would give in general not the answer it should. Or one may even say, it could give a wrong answer. This problem is tackled in chapter 2.

An important question in the time series setting is the question of forecast; given data up to today what can be predicted for the future. For a specific class of time series chapter 3 gives insights how a good prediction can be achieved.

1.1 Time Series Fundamentals

A stochastic process is a family of random variables $\{X_t, t \in T\}$, where T is some index set, defined on a probability space $(\Omega, \mathcal{F}, P)$, c.f. (Brockwell and Davis, 1991, Chapter 1). In time series analysis

[1] Angela Merkel (Chancellor of Germany), Hanover, 2016: http://www.cebit.de/de/news-trends/news/bundeskanzlerin-merkel-daten-sind-die-rohstoffe-des-21-jahrhunderts-1190

the index set is usually given by $\mathbb{Z}$, hence $\{X_t, t \in \mathbb{Z}\}$. We distinguish here between univariate time series, where the stochastic process is $\mathbb{C}$-valued, and multivariate time series, where a vector-valued process is considered, $\mathbb{C}^d, d \geq 1$. chapter 2 deals mainly with univariate time series, whereas multivariate time series are in focus of chapter 3. Throughout this work we concentrate on stationary time series. A time series $X = \{X_t, t \in \mathbb{Z}\}$ is said to be stationary if $E|X_t^2| < \infty, EX_t = \mu$ for all $t \in \mathbb{Z}$, and for all $t, h \in \mathbb{Z}$ we have

$$\mathrm{Cov}(X_{t+h}, X_t) = E[(X_{t+h} - EX_{t+h})(X_t - E(X_t))^\top] = E[(X_h - EX_h)(X_0 - E(X_0))^\top] =: \gamma(h). \tag{1.1.1}$$

The function $\gamma(h), h \in \mathbb{Z}$, defined by (1.1.1) is called the autocovariance function of time series X. Furthermore, a time series $\{X_t, t \in \mathbb{Z}\}$ is said to be strictly stationary if $P^{X_{t_1},\ldots,X_{t_k}} = P^{X_{t_1+h},\ldots,X_{t_k+h}}$ for all $t_1, \ldots, t_k \in \mathbb{Z}$ and $h \in \mathbb{Z}$, where $P^{X_{t_1},\ldots,X_{t_k}}$ denotes the joint distribution of $X_{t_1}, \ldots, X_{t_k}$.

A simple example for a stationary time series is white noise. It is given by an uncorrelated time series $\{\varepsilon_t, t \in \mathbb{Z}\}$ with $E\varepsilon_t = 0$ and $\mathrm{Var}\varepsilon_t = \Sigma_\varepsilon^2 < \infty$ for all $t \in \mathbb{Z}$. Furthermore, we introduce here two important time series models: moving average (MA) models and autoregressive (AR) models. Based on some white noise $\{\varepsilon_t, t \in \mathbb{Z}\}$ a moving average process $\{X_t, t \in \mathbb{Z}\}$ of order q is defined by

$$X_t = \sum_{j=0}^{q} B_j \varepsilon_{t-j}, t \in \mathbb{Z}, \tag{1.1.2}$$

where $B_0, \ldots, B_q \in \mathbb{C}^{d\times d}$, B_0 is usually normalized to the identity matrix and $B_q \neq 0$. An autoregressive process $\{X_t, t \in \mathbb{Z}\}$ of order p is defined by

$$X_t = \sum_{j=1}^{p} A_j X_{t-j} + \varepsilon_t, t \in \mathbb{Z}, \tag{1.1.3}$$

where $A_1, \ldots, A_p \in \mathbb{C}^{d\times d}$ and $A_p \neq 0$. Both models are special cases of autoregressive moving average (ARMA) models of order (p, q) which are given by

$$X_t - \sum_{j=1}^{p} A_j X_{t-j} = \sum_{j=1}^{q} B_j \varepsilon_{t-j} + \varepsilon_t, t \in \mathbb{Z}, \tag{1.1.4}$$

where $A_1, \ldots, A_p, B_1, \ldots, B_q \in \mathbb{C}^{d\times d}$ and $A_p, B_q \neq 0$.

A stationary time series and its properties can be expressed either in the time domain or in the frequency domain. The frequency domain is only used in chapter 2, hence in the univariate case. In order to simplify notation it is described here for the univariate case. However, the results given here can be transfered to the multivariate case. The autocovariance function describes the second-

order properties of a time series in time domain. Herglotz's Theorem, c.f. (Brockwell and Davis, 1991, Theorem 4.3.1) gives the corresponding representation in the frequency domain: A function $\gamma : \mathbb{Z} \to \mathbb{C}$ is non-negative definite (hence, an autocovariance function) if and only if

$$\gamma(h) = \int_{(-\pi,\pi]} \exp(ih\nu) dF(\nu), \text{ for all } h \in \mathbb{Z},$$

where $F(\cdot)$ is a right-continuous, non-decreasing, bounded function on $[-\pi, \pi]$ and $F(-\pi) = 0$. The function F is called the spectral distribution function of γ and if $F(\lambda) = \int_{-\pi}^{\lambda} f(\nu) d\nu, -\pi \leq \lambda \leq \pi$, then f is called a spectral density of γ. Furthermore, we have, c.f. (Brockwell and Davis, 1991, Theorem 4.3.2), that if $\sum_{h\in\mathbb{Z}} |\gamma(h)| < \infty$ then

$$\gamma(h) = \int_{-\pi}^{\pi} \exp(ih\nu) f(\nu) d\nu, h \in \mathbb{Z},$$

where

$$f(\lambda) = \frac{1}{2\pi} \sum_{h\in\mathbb{Z}} \exp(-ih\lambda)\gamma(h).$$

Throughout this work, we denote by second-order properties of a time series the properties defined by the entire autocovariance function in time domain or by the spectral density in frequency domain, respectively.

For a stationary ARMA model given by (1.1.4), the spectral density can be directly derived by using the corresponding AR and MA polynomials. The following theorem, c.f. (Brockwell and Davis, 1991, Theorem 4.4.2), gives insight: Let $X = \{X_t, t \in Z\}$ be an ARMA(p, q) process satisfying $A(L)X_t = B(L)\varepsilon_t, \{\varepsilon_t, t \in \mathbb{Z}\}$ is some white noise with variance σ^2, L is the lag-operator, and $A(z) = 1 - \sum_{j=1}^{p} a_j z^j, B(z) = 1 + \sum_{j=1}^{q} b_j$. If the polynomials $A(z)$ and $B(z)$ have no common zeroes and $A(z) \neq 0$ for $|z| = 1$, then X has spectral density

$$f(\lambda) = \frac{\sigma^2}{2\pi} \frac{A\left(\exp(-i\lambda)\right)}{B\left(\exp(-i\lambda)\right)}, -\pi \leq \lambda \leq \pi.$$

Besides the autocovariance, the time series itself can be expressed in frequency domain by using an orthogonal increment process. Since it is not used in this work, we are not going into detail here. It is more important that both domains contain the same amount of information. The only difference is the way this information is given. This different point of view can be enlightening for some applications, see section 1.5 in Brillinger (2001) for applications of the frequency domain. The autocovariance as well as the spectral density can be estimated with some observations $X_1, \ldots, X_n$. An estimator for the autocovariance is the sample autocovariance given by

$$\hat{\gamma}_n(h) = \frac{1}{n} \sum_{t=1}^{n-h} \left(X_{t+h} - \frac{1}{n} \sum_{s=1}^{n} X_s \right) \left(X_t - \frac{1}{n} \sum_{s=1}^{n} X_s \right), 0 \leq h \leq n-1, \tag{1.1.5}$$

$\hat{\gamma}_n(h) = 0, h \geq n, \hat{\gamma}_n(h) = \hat{\gamma}_n(-h)$. The corresponding spectral density is given by

$$I_n(\lambda) = \frac{1}{2\pi} \sum_{h \in \mathbb{Z}} \hat{\gamma}_n(h) \exp(-ih\lambda) = \frac{1}{2\pi n} \left| \sum_{t=1}^{n} \left(X_t - \frac{1}{n} \sum_{s=1}^{n} X_s \right) \exp(-it\lambda) \right|^2, -\pi \leq \lambda \leq \pi,$$

and I_n is called the Peridogram. Point-wise consistency of $\hat{\gamma}_n$, hence, for a given h, can be established, c.f. (Brockwell and Davis, 1991, Section 7.2). However, the periodogram I_n is inconsistent, c.f. (Kreiss and Neuhaus, 2006, Satz 12.7), which also implies that the absolute error for all sample autocovariance do not vanish, hence $\sum_{h=0}^{\infty} |\gamma(h) - \hat{\gamma}_n(h)| \neq o_P(1)$. This also implies that $\hat{\Sigma}_n = [\hat{\gamma}_n(i-j)]_{i,j=1,\dots,n}$ is not a consistent estimator of the autocovariance matrix $\Sigma_n = [\gamma(i-j)]_{i,j=1,\dots,n}$, c.f. McMurry and Politis (2010). Additional smoothing is required to get consistent estimators of the second-order properties. This can be achieved by using a truncated autocovariance estimator such as $\tilde{\gamma}_n(h) = k(h/M(n))\hat{\gamma}_n(h)$, where k is some kernel with support $[-1,1]$ and $M(n) < n$ such that $\tilde{\gamma}_n(h) = 0$ for $h > M(n)$. The resulting spectral density estimators $\hat{f}(\lambda) = 1/(2\pi) \sum_{h \in \mathbb{Z}} \tilde{\gamma}_n(h) \exp(-ih\lambda)$ are denoted as lag-window estimators and give consistent results, see Jentsch and Subba Rao (2015) as well as section 2.2.3 for details. Since the spectral density and the autocovariance describe the same information, only in different domains, such a truncation leads also to consistent estimators $\tilde{\Sigma}_n = [\tilde{\gamma}_n(i-j)]_{i,j=1,\dots,n}$ for the autocovariance matrix Σ_n, see Wu and Pourahmadi (2009) and McMurry and Politis (2010) for details. The spectral density plays a major role in chapter 2.

1.2 Overview of Bootstrap Methods for Time Series

In statistics when a certain quantity is estimated with a given statistic often the questions occurs how precise the estimation is and what deviation can be expected in x out of 100 cases. To answer such questions it is helpful to derive the distribution of the statistic. However, it is usually the case that it is not possible to derive the exact distribution. Instead, a consistent approximation is used. Bootstrap methods can be used to estimate the distribution of a given statistic. In its basic form the bootstrap method was introduced by Efron (1979). For a given statistic T the idea is as follows; Based on a sample $X = (X_1, \dots, X_n)$ new samples $(X_1^{*,j}, \dots, X_n^{*,j}), j = 1, \dots, N$ are created by using the empirical distribution function given by the sample X. Then the statistic is evaluated for each new sample, hence, we obtain $T_1^* = T(X_1^{*,1}, \dots, X_n^{*,1}), \dots, T_N^*$. The empirical distribution function of $T_1^*, \dots, T_N^*$ is then used as an approximation of the distribution function of T. N is the number of bootstrap samples and is similar to the number of trials in a Monte Carlo simulation. However, nothing is said about the performance of this approximation. We say that a bootstrap method is valid if $c_n(T_n - ET_n)$ and $c_n(T_n^* - E^*T_n^*)$ have the same limiting distribution, where c_n is such that $c_n(T_n - ET_n)$ converges to a non-degenerate distribution. Or more precisely, (Kreiss and Paparodi-

tis, 2017, Definition 1.41), let $(\Omega_n, \mathcal{A}_n, P_n), n \in \mathbb{N}$, be a sequence of statistical experiments and L_n a sequence of random variables on Ω_n. Given $x_n \in \Omega_n$, let $(\Omega_n^*, \mathcal{A}_n^*, P_n^*), n \in \mathbb{N}$, be a corresponding bootstrap statistical experiment and L_n^* bootstrap random variables. Denote by $\mathcal{L}_n$ the distribution of L_n and by $\mathcal{L}_n^*(x_n)$ the distribution of L_n^* given x_n. We denote the bootstrap proposal L_n^* as valid (or consistent, respectively) for L_n if and only if $\lim_{n\to\infty} d(\mathcal{L}_n, \mathcal{L}_n^*(x_n)) = 0$, in P_n-probability, where d is some distance measure between distributions. For features in probability see section 1.3.1 and especially Definition 1.7 in Kreiss and Paparoditis (2017). Possible distance measures are the Kolmogorov's distance, c.f. section 1.4.2 in Kreiss and Paparoditis (2017), and the Mallow's distance, c.f. section 1.4.3 in Kreiss and Paparoditis (2017). In this work the Mallow's distance is mainly considered. If the data consists of independent and identically distributed data points the bootstrap proposal of Efron (1979) is valid for most statistics and settings. However, time series data is considered here, hence the data points are dependent. In this case, the classical bootstrap proposal is not even valid for the sample mean $1/n \sum_{t=1}^{n} X_t$. That is why several new bootstrap ideas have been proposed to overcome this shortcoming of the classical bootstrap proposal. These ideas can be grouped and in the following only the basic concepts of the three most important groups are presented. The review paper by Kreiss and Paparoditis (2011) is recommended for a more exhaustive overview of the several bootstrap ideas. Further details can be found in Kreiss and Paparoditis (2017) and Lahiri (2003).

An intuitive extension of the classical proposal is the block bootstrap. In the classical proposal new samples are generated by drawing with replacement from the original sample. However, this destroys the dependent structure. In order to retain the dependent structure, the idea is to generate new samples by drawing with replacement from blocks of data points. Hence, within such a block a fraction of the dependence structure of the data is kept. In order to fully capture the dependence structure of the underlying process it is necessary that the block length increases to infinity as the sample size increases to infinity. For a valid approximation it is also necessary that the number of blocks increases as well. Many authors have adapted this idea. Some work with non-overlapping blocks has been done by Carlstein (1986) or Hall (1985), with overlapping blocks by Künsch (1989) or even overlapping blocks with random block length by Politis and Romano (1994). Furthermore, it is possible to taper the block-ends to get a smoother transition between blocks, c.f. Paparoditis and Politis (2001). The block bootstrap idea does not require that the underlying process follows some parametric structure. However, all block bootstrap variations have in common that they are in general very sensitive regarding the choice of the block length.

The setting of the residual bootstrap is that the underlying process $X = \{X_t, t \in \mathbb{Z}\}$ possesses some structure which can be expressed by $X_t = f(\varepsilon_t, \dots)$, where f is some unknown function and $\{\varepsilon_t, t \in \mathbb{Z}\}$ is a process which is less dependent than X. The ε's are denoted as the residuals.

The residuals are often uncorrelated, in some cases they are even independent. Based on a sample $X_1, \dots, X_n$ the idea of the residual bootstrap is to estimate f and the residuals. Afterwards the classical bootstrap approach on the residuals is being used. Hence, a new bootstrap observation of X_t is given by $X_t^* = \hat{f}(\varepsilon_t^*, \dots)$, where (ε_t^*) is sampled by the empirical distribution function given by $\hat{\varepsilon}_t, t = 1, \dots, n$. A classical example here is the case when X_t is an AR(p) process. Hence, $X_t = \sum_{j=1}^{p} a_j X_{t-j} + \varepsilon_t$. However, this bootstrap idea is not restricted to finite models. The AR-sieve bootstrap has the idea of approximating the dependence structure with AR-models of increasing order, see Kreiss (1992), Bühlmann (1997), Paparoditis and Streitberg (1991), and Kreiss et al. (2011). The linear process bootstrap by McMurry and Politis (2010) is another bootstrap proposal which does not require a specific finite model. This method is described in more detail in section 2.8.

A special form of the residual bootstrap is the frequency domain bootstrap, c.f. Franke and Hardle (1992), Hurvich and Zeger (1987) or Dahlhaus et al. (1996). For a time series $X_t = \sum_{j\in\mathbb{Z}} \phi_j \varepsilon_{t-j}, t \in \mathbb{Z}$, such bootstrap methods use the following approximation of the periodogram for linear processes at Fourier frequencies λ_j, $I_n(\lambda_j) \approx f(\lambda_j) I_{n,\varepsilon}(\lambda_j)$, where f is the spectral density of X and $I_{n,\varepsilon} = (2\pi n)^{-1} |\sum_{t=1}^{n} \varepsilon_t \exp(it\lambda_j)|^2$ is the periodogram of the residuals ε_t. Furthermore, we have under some conditions that the periodogram is asymptotically independent for different Fourier frequencies, c.f. (Brillinger, 2001, Theorem 5.2.6) or (Brockwell and Davis, 1991, Theorem 10.3.2). Hence, given some spectral density estimator $\hat{f}_n$ residuals $\tilde{\varepsilon}_k$ can be obtained by $\tilde{\varepsilon}_k = I_n(\lambda_k)/\hat{f}_n(\lambda_k)$. After normalization, those residuals can be resampled i.i.d. to obtain bootstrap values for the periodogram. Statistics as the sample autocovariance, sample autocorrelation, or spectral density estimators can be expressed by the integrated periodogram given by $\int_0^{2\pi} W(\lambda) I_n(\lambda) d\lambda)$, for some function $W : [0, 2\pi] \to \mathbb{R}$, see section 12.7 in Kreiss and Neuhaus (2006) for details. Therefore, the frequency domain bootstrap can be applied to those statistics. This bootstrap scheme creates new samples in the frequency domain. Some authors, c.f. Jentsch and Kreiss (2010) or Kirch et al. (2011), extended the idea of the frequency domain bootstrap to create also samples in the time domain.

The residuals used within such a residual bootstrap procedure can be bootstrapped wild. Hence, instead of using the estimated residuals some predefined distribution is used to sample residuals. Usually the residuals are sampled i.i.d., however it is possible to give these residuals also a predefined dependent structure.

1.3 Network Fundamentals

In its most general form a network denotes simply a collection of interconnected things, see (Kolaczyk, 2009, Chapter 1). Network data occur in many different fields such as social sciences, biology, physics or logistics. For instance, a social network of friendships between 34 members of a karate club, Zachary (1977), a network representing the topology of the western states power grid of the

United States, Watts and Strogatz (1998), or a network of human contact which could help to understand epidemics, Rocha et al. (2011).

A graph structure is used to describe this mathematically. A graph $G = (V, E)$ is a mathematical structure consisting of a set V of vertices and a set E of edges. In this work the vertices are labeled by $1, \dots, n$ such that $V = \{1, \dots, n\}$. Here, we consider directed edges. Consequently, E consists of ordered pairs $\{u, v\}, u, v \in V$. In the undirected case there is no distinction between $\{u, v\}$ and $\{v, u\}$. An edge $\{u, u\}$ is denoted as a loop and it is also possible that an edge $\{u, v\}$ is contained multiple times in E. Such edges are denoted as multi-edges. Graphs with directed edges and multi-edges are also denoted as multi-digraphs, see (Kolaczyk, 2009, Chapter 2). The connectivity of a graph G can be captured in an $n \times n$ matrix A with entries $A_{ij} = |\{e \in E : e = (i, j)\}|$. The matrix A is called the adjacency matrix and entry i, j gives the number of edges from vertex i to vertex j. The row sum $d_i^{\text{out}} = A_{i+} = \sum_{j=1}^{n} A_{ij}$ gives the number of edges which are going out from vertex i and d_i^{out} is denoted as the out-degree. The number of edges going into vertex i is given by the column sum $d_i^{\text{in}} = A_{+i} = \sum_{j=1}^{n} A_{ji}$ and is denoted as the in-degree. A graph with no multi-edges can contain at most n^2 edges. Hence, the density of a graph with no multiple edges can be defined by $\text{den}(G) = |E|/(n^2)$. We denoted a network as sparse if $|E| = O(n)$ and dense if $|E| = O(n^2)$.

In the example of the karate club, Zachary (1977), a vertex represents a person and an edge between two vertices represents friendship between the corresponding persons.

In this work a dynamic network is given by a family of graphs $\{G_t = (V_t, E_t), t \in \mathbb{Z}\}$ and a static network is given by a single graph G. That is why often the terms 'graph' and 'network' are used inter-changeably. If a static number of vertices is considered, then a dynamic network can be described by a time-dependent adjacency matrix $\mathbf{Ad} = \{Ad_t, t \in \mathbb{Z}\}$.

Several statistical models have been developed to describe such network data. An important model class is the exponential random graph model (ERGM), see section 6.5 in Kolaczyk (2009). We denote that a random vector Z belongs to an exponential family if its probability function can be expressed in the form $P_\Theta(Z = z) = \exp(\Theta^\top g(z) - \phi(\Theta))$, where $\Theta \in \mathbb{R}^p$ is a vector of parameters, g is a p-dimensional function of z, and $\phi(\Theta)$ is a normalization term, c.f. equation (6.23) in Kolaczyk (2009) or section 4.4 in Mood (1970). Let $Y_{ij}, i, j = 1, \dots, n$ be a binary random variable indicating the presence or absence of an edge from vertex i to vertex j. Then, an exponential random graph model is a model for which the joint distribution of elements in Y is specified in exponential family form. A special case of the ERGMs is the Bernoulli random graph model. For these models, it is considered that the edges are independent to each other, hence, $Y_{i,j}$ is independent to $Y_{s,k}$ for any $i, j \neq s, k$. Furthermore, $Y_{i,j}, i, j = 1 \dots, n$ is Bernoulli distributed and often it is further simplified that all edges share one common parameter. An ERGM describes a static network, however, Hanneke and Xing (2007) have extended these models to dynamic networks. In the dynamic setting, a common

assumption is that the network possesses some form of Markov property, c.f. Crane (2015). Hence, for a dynamic network with a static number of vertices this means that $\mathbf{Ad} = \{Ad_t, t \in \mathbb{Z}\}$ is a Markov process.

Bibliography

Brillinger, D. (2001). *Time Series: Data Analysis and Theory*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics.

Brockwell, P. and Davis, R. A. (1991). *Time Series: Theory and Methods (2nd edition)*. Springer, New York.

Bühlmann, P. (1997). Sieve bootstrap for time series. *Bernoulli*, 3(2):123–148.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics*, pages 1171–1179.

Crane, H. (2015). Time-varying network models. *Bernoulli*, 21(3):1670–1696.

Dahlhaus, R., Janas, D., et al. (1996). A frequency domain bootstrap for ratio statistics in time series analysis. *The Annals of Statistics*, 24(5):1934–1963.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26.

Franke, J. and Hardle, W. (1992). On bootstrapping kernel spectral estimates. *The Annals of Statistics*, pages 121–145.

Hall, P. (1985). Resampling a coverage pattern. *Stochastic processes and their applications*, 20(2):231–246.

Hanneke, S. and Xing, E. (2007). Discrete temporal models of social networks. *Statistical Network Analysis: Models, Issues, and New Directions Edited by:Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, Anna Goldenberg, Eric P. Xing, Alice X. Zheng*.

Hurvich, C. M. and Zeger, S. (1987). *Frequency domain bootstrap methods for time series*. New York University, Graduate School of Business Administration.

Jentsch, C. and Kreiss, J.-P. (2010). The multiple hybrid bootstrap—resampling multivariate linear processes. *Journal of Multivariate Analysis*, 101(10):2320–2345.

Jentsch, C. and Subba Rao, S. (2015). A test for second order stationarity of a multivariate time series. *Journal of Econometrics*, 185(1):124–161.

Kirch, C., Politis, D. N., et al. (2011). Tft-bootstrap: Resampling time series in the frequency domain to obtain replicates in the time domain. *The Annals of Statistics*, 39(3):1427–1470.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Publishing Company, Incorporated, 1st edition.

Kreiss, J.-P. (1992). *Bootstrap procedures for AR(∞)-processes*, volume 376. Springer, In Bootstrapping and Related Techniques of *Lecture Notes in Economics and Mathematical Systems*. 107–113.

Kreiss, J.-P. and Neuhaus, G. (2006). *Einführung in die Zeitreihenanalyse*. Statistik und ihre Anwendungen. Springer.

Kreiss, J.-P. and Paparoditis, E. (2011). Bootstrap methods for dependent data: A review. *Journal of the Korean Statistical Society*, 40(4):357–378.

Kreiss, J.-P. and Paparoditis, E. (2017). *Bootstrap for Time Series: Theory and Applications*. to appear.

Kreiss, J.-P., Paparoditis, E., and Politis, D. N. (2011). On the range of validity of the autoregressive sieve bootstrap. *Ann. Statist.*, 39(4):2103–2130.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, 17(3):1217–1241.

Lahiri, S. (2003). *Resampling Methods for Dependent Data*. Springer Series in Statistics. Springer.

McMurry, T. L. and Politis, D. N. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *Journal of Time Series Analysis*, 31(6):471–482.

Mood, A. M. (1970). Introduction to the theory of statistics. Third edition.

Paparoditis, E. and Politis, D. N. (2001). Tapered block bootstrap. *Biometrika*, 88(4):1105–1119.

Paparoditis, E. and Streitberg, B. (1991). Order identification statistics in stationary autoregressive moving-average models: Vector autocorrelations and the bootstrap. *Journal of Time Series Analysis*, 13(5):415–434.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313.

Rocha, L. E., Liljeros, F., and Holme, P. (2011). Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS computational biology*, 7(3):e1001109.

Shao, X. (2010). The dependent wild bootstrap. *Journal of the American Statistical Association*, 105(489):218–235.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-worldnetworks. *nature*, 393(6684):440.

Wu, W. B. and Pourahmadi, M. (2009). Banding sample autocovariance matrices of stationary processes. *Statistica Sinica*, pages 1755–1768.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473.

2 Estimated Wold Representation and Spectral-Density-Driven Bootstrap for Time Series

Based on: Krampe, J., Kreiss, J.-P. and Paparoditis, E.: Estimated Wold representation and spectral-density-driven bootstrap for time series. J. R. Stat. Soc. B. (2018)

2.1 Introduction

The spectral density, if it exists, plays an important role as a quantity which completely describes the so-called second-order properties of stationary time series. A broad literature exists on spectral density estimators, among them parametric (e.g. autoregressive) estimators, nonparametric (e.g. lag window or smoothed periodogram) estimators or semiparametric estimators as a mixture of both. Time series analysts typically are rather skilled in estimating spectral densities and they know, depending on the required application, the pros and cons of the various estimators. This work intends to bring together several bootstrap procedures under the umbrella of spectral density estimation.

Recall that for a purely nondeterministic and stationary stochastic process $X = (X_t, t \in \mathbb{Z})$ with spectral density f, Szegö's factorization expresses f as a power series. The coefficients of this factorization, appropriately normalized, coincide with the coefficients of the well-known Wold representation of X. Recursive formulas, which make use of the Fourier coefficients of $\log(f)$ - the so-called cepstral-coefficients -, to calculate the coefficients of the Wold representation of the process have been developed; cf. Pourahmadi (1983). Moreover, if f is strictly positive then X also obeys an autoregressive (AR) representation and similar recursive formulas to compute the coefficients of this representation have also been derived; see again Pourahmadi (1983). Using these recursions we suggest a procedure to estimate the coefficients of both the moving average (MA) and the autoregressive representations based on an estimator $\hat{f}_n$ of the spectral density f. In particular, we show that under certain conditions on f and on the used estimator $\hat{f}_n$, the sequence of coefficients of the Wold and of the autoregressive representation of the process can be consistently estimated. Furthermore, under additional smoothness conditions the pointwise consistency of the estimators can

be extended to uniform consistency for the entire sequence of coefficients. It should be noted that the factorization of the spectral density has been considered in the literature also for implementing and investigating the so-called Wiener-Kolmogorov predictor in linear prediction (cf. Jones (1964), Bhansali (1974, 1977) and Pourahmadi (1983)).

The availability of estimates of the moving average coefficients of the Wold representation enables the development of a general spectral-density-driven bootstrap (SDDB) procedure for time series. In particular, a pseudo time series can be generated by using the estimated sequence of moving average coefficients and an appropriately chosen sequence of pseudo innovations. The resulting bootstrap procedure is then fully determined by the particularly chosen spectral density estimator $\hat{f}_n$ and the stochastic properties of the generated pseudo innovations. The estimated Wold representation used should mainly be regarded as a means to an end to generate a pseudo time series which exactly has the chosen spectral density estimator as its spectrum.

For instance, choosing a parametric autoregressive spectral density estimator, the coefficients of the estimated Wold representation coincide with the coefficients of the inverted estimated autoregressive polynomial and therefore, the autoregressive model can just as well be used to generate the bootstrap data. In other words, using a parametric autoregressive spectral density estimator will lead to the well-known AR-sieve bootstrap for time series (cf. Kreiss (1992), Bühlmann (1997) and Kreiss et al. (2011)). However, a parametric autoregressive spectral density estimator often is not the first choice. Let us consider a nonparametric competitor, for instance, a lag window estimator of f with truncation lag M_n. As we will see, this will lead us essentially to a moving average process of finite order M_n which can be used to generate the pseudo time series. Therefore, the spectral-density-driven bootstrap proposed in this work, is a general notion of bootstrap for time series which allows for a variety of possibilities to generate the pseudo time series. These possibilities are determined by the particular spectral density estimator $\hat{f}_n$ used to obtain the estimates of the coefficients of the Wold representation. Notice that although the spectral-density-driven bootstrap generates bootstrap pseudo time series in the time domain, the second-order dependence structure of the underlying process is entirely mimicked in the frequency domain by means of the selected spectral density estimator used. Thus, various well-known and flexible methods for spectral density estimation can be used in our bootstrap method. As a consequence, we formulate the assumptions, which are needed for our theoretical developments, in terms of the spectral density and its estimator, only. This allows us to restrict the class of admissible spectral density estimators as little as possible.

Fed by independent and identically distributed (i.i.d.) pseudo innovations the proposed spectral-density-driven bootstrap generates pseudo time series stemming from a linear process. For such a choice of pseudo innovations we compare our bootstrap proposal to some other linear bootstrap

procedures, like the AR-sieve bootstrap (cf. Kreiss (1992) and Bühlmann (1997)) and the linear process bootstrap, cf. McMurry and Politis (2010). As already indicated, it is shown that the AR-sieve bootstrap is a special case of the spectral-density-driven bootstrap which is obtained if a parametric autoregressive spectral density estimator $\hat{f}_n$ is used. Furthermore, we show that the linear process bootstrap essentially generates pseudo observations by factorizing banded autocovariance matrices. This technique is related to the factorization of spectral densities which is used in this work. However, in finite samples the two approaches differ from each other.

It is worth mentioning that pseudo innovations generated in a different way than i.i.d. could also be used in the proposed spectral-density-driven bootstrap procedure. For instance, pseudo innovations generated by means of a block bootstrap applied to appropriately defined residuals may be used. Although such a proposal would most likely extend the range of validity of the spectral-density-driven bootstrap to nonlinear time series, we do not consider such an approach in this work, i.e., we restrict ourselves to the linear process set-up. We show that if the pseudo innovations are generated by means of an i.i.d. wild bootstrap that appropriately mimics the first, the second, and the fourth moment structure of the true innovations, then the proposed spectral-density-driven bootstrap is asymptotically valid for a wide range of statistics commonly used in time series analysis. Besides the sample mean, statistics described by the so-called class of generalized autocovariances are also considered. Note that this class includes sample autocovariances, sample autocorrelations and lag window spectral density estimators as special cases; see section 2.3 for details. We demonstrate by means of simulations that our asymptotic findings coincide with a good finite sample behavior of the proposed bootstrap procedure. Furthermore, the performance of the new bootstrap method is compared with that of the asymptotic normal approximations and of some other bootstrap competitors, like the linear process, the AR-sieve and the tapered block bootstrap. An *R*-code to generate pseudo time series with the spectral-density-driven bootstrap is available at `www.tu-bs.de/Medien-DB/stochastik/code-snippet_sddb.txt`.

The chapter is organized as follows: Section 2.2 briefly describes the Wold and the AR representation of a stationary time series and discusses the method used to estimate the entire sequence of coefficients in both representations. Local and global consistency properties of the estimators are established. Section 2.3 introduces the spectral-density-driven bootstrap procedure for time series and establishes, for linear processes and for relevant classes of statistics, its asymptotic validity. A comparison with the AR-sieve and with the linear process bootstrap is also given in this section. Section 2.4 presents some numerical simulations investigating the finite sample behavior of the proposed bootstrap method and compares its performance with that of other bootstrap methods and of asymptotic normal approximations. A real-life data example demonstrates the applicability of the suggested bootstrap procedure. Auxiliary results as well as proofs of the main results are

deferred to section 2.6. Answer on how the spectral density factorization can be used to estimate a MA(q) model is shown in section 2.7. A detailed comparison with the linear process bootstrap is given in section 2.8. Finally, further simulation results and proofs are given in section 2.9 and section 2.10.

2.2 Estimated Wold Representation

2.2.1 Moving Average and Autoregressive Representation

Stationary processes are commonly classified using the concept of linear prediction; see for example Brockwell and Davis (1991, Section 5.7) or Pourahmadi (2001, Section 5.5). To elaborate, let $X = \{X_t, t \in \mathbb{Z}\}$ be a stationary stochastic process and define by $\mathcal{M}_n(X) = \overline{\text{span}\{X_t, -\infty < t \leq n\}}$ and $\mathcal{M}_{-\infty}(X) = \bigcap_{n=-\infty}^{\infty} \mathcal{M}_n$ the closed linear subspaces of the Hilbert space $\mathcal{M}(X) = \overline{\text{span}\{X_t, t \in \mathbb{Z}\}}$. Note that an overlined set denotes its closure. Let $P_{\mathcal{M}_n(X)}X_{n+1}$ be the projection of X_{n+1} onto $\mathcal{M}_n(X)$ and define by $\sigma^2 = E|X_{n+1} - P_{\mathcal{M}_n(X)}X_{n+1}|^2$ the mean square error of the best (in the mean square sense) one-step, linear predictor. The process X is called deterministic if and only if $X_t \in \mathcal{M}_{-\infty}(X)$ or equivalently if and only if $\sigma^2 = 0$. It is called nondeterministic if $X_{n+1} \notin \mathcal{M}_n(X)$ and consequently $\sigma^2 > 0$. Furthermore, it is called purely-nondeterministic if it is nondeterministic and $\mathcal{M}_{-\infty}(X) = \{0\}$.

If the process X possesses a spectral density f, which is the case if $\sum_{h\in\mathbb{Z}} |\gamma_h| < \infty$, with $\gamma_h = Cov(X_t, X_{t+h})$, then it holds true that X is nondeterministic if and only if

$$\int_{(-\pi,\pi]} \log f(\lambda) d\lambda > -\infty, \tag{2.2.1}$$

see Pourahmadi (2001, Theorem VII).

Wold's decomposition, see Pourahmadi (2001, Theorem 5.11), guarantees that any nondeterministic process can be divided into a deterministic and a purely-nondeterministic part. Furthermore, the purely-nondeterministic part of the process has a unique one-sided moving average (MA) representation given by

$$X_t = \sum_{k=0}^{\infty} c_k \varepsilon_{t-k}, \quad t \in \mathbb{Z}, \tag{2.2.2}$$

where $\sum_k |c_k|^2 < \infty$ and $\{\varepsilon_t, t \in \mathbb{Z}\}$ is a white noise process defined by $\varepsilon_{n+1} = X_{n+1} - P_{\mathcal{M}_n(X)}X_{n+1}$, $n \in \mathbb{Z}$, called the innovation process. Here, white noise refers to an uncorrelated time series. Notice that even if X is a linear process driven by i.i.d. innovations, the white noise process appearing in the corresponding one-sided moving average representation (2.2.2) might not be i.i.d.. To give an example consider the linear, first order moving average process, $X_t = e_t + \theta e_{t-1}$ where $\{e_t, t \in \mathbb{Z}\}$ is an i.i.d. process and $\theta > 1$. The Wold representation of this process is given by $X_t = \varepsilon_t + \theta^{-1}\varepsilon_{t-1}$

where $\varepsilon_t = e_t + (1-\theta^2)\sum_{j=1}^{\infty}(-\theta)^{-j}e_{t-j}$ is white noise with variance θ^2. Obviously, the innovations ε_t are not independent.

Another interesting one-sided representation of the process X is the so-called autoregressive representation which appears if the spectral density f is bounded away from zero, i.e., if $\inf_{\lambda\in[0,\pi]} f(\lambda) = C > 0$. Instead of using the full history of the innovation process $\{\varepsilon_t, t \in \mathbb{Z}\}$, in this case the full history of the process X itself is used to express the value X_t at any time point t. X_t can then be written as

$$X_t = \sum_{k=1}^{\infty} b_k X_{t-k} + \varepsilon_t, \quad t \in \mathbb{Z}, \tag{2.2.3}$$

where $\sum_k |b_k|^2 < \infty$ and $\{\varepsilon_t\}$ is the same white noise innovation process as in (2.2.2); see Pourahmadi (2001, Section 6.2.1). Expression (2.2.3) is called the autoregressive representation of the process X and should not be confused with that of a linear, infinite order autoregressive process driven by i.i.d. innovations. To demonstrate this, consider again the previous example of the linear, noninvertible moving average process $X_t = e_t + \theta e_{t-1}$ with $\theta > 1$. This process has the autoregressive representation $X_t = -\sum_{j=1}^{\infty}(-\theta)^{-j}X_{t-j} + \varepsilon_t$ where $\{\varepsilon_t, t \in \mathbb{Z}\}$ is the uncorrelated but not independent white noise processes appearing in the Wold representation of X_t.

To derive recursive formulas for the coefficients in the moving average representation (2.2.2) and the autoregressive representation (2.2.3), we start with some basic factorization properties of the spectral density f. Notice first that f can be expressed as $f(\cdot) = (2\pi)^{-1}|V(\exp(-i\cdot))|^2$ for a power series $V(z) = \sum_{k=0}^{\infty} v_k z^k$ and that such a factorization exists if and only if condition (2.2.1) is fulfilled; see Szegö (1921). The above factorization of the spectral density is not unique. However, if we restrict ourselves to power series which have no roots inside the unit disk and appropriately normalize the coefficients, a unique representation occurs. The coefficients of this unique power series coincide with the coefficients c_k of the Wold representation (2.2.2), if additionally, the power series $V(z)$ is appropriately normalized, i.e., if $\tilde{V} = V/v_0$ is used. We denote this unique and normalized power series by $C(z) = \sum_{k=0}^{\infty} c_k z^k$. Notice that (2.2.1) ensures, that the Fourier coefficients of $\log f$ are well defined. Furthermore, since $C(z)$ has no zeros inside the open unit disc, $\log(C(z))$ is analytic inside the same region and we have for $|z| < 1$ that

$$\sigma(2\pi)^{-1/2}\sum_{k=0}^{\infty} c_k z^k = \exp\left(a_0/2 + \sum_{k=1}^{\infty} a_k z^k\right), \tag{2.2.4}$$

where a_k is the k-th Fourier coefficients of $\log f$,

$$a_k = \int_{-\pi}^{\pi} \log f(\lambda)\exp(-ik\lambda)d\lambda/(2\pi). \tag{2.2.5}$$

Differentiation of equation (2.2.4) together with comparison of coefficients leads to a recursive formula to calculate the coefficients $\{c_k\}$ of this power series by using the Fourier coefficients of $\log f$, see Pourahmadi (1983, 1984). In particular, setting $c_0 = 1$, the following recursive formula can be used to obtain the coefficients $\{c_k\}$,

$$c_{k+1} = \sum_{j=0}^{k} \left(1 - \frac{j}{k+1}\right) a_{k+1-j} c_j, \quad k = 0, 1, 2, \ldots. \tag{2.2.6}$$

Furthermore, $\sigma^2 = 2\pi \exp(a_0)$. If the process X possesses also the autoregressive representation (2.2.3), then the coefficients $(b_k)_{k\in\mathbb{N}}$ of this representation can be calculated using the relation $C(z)^{-1} = B(z) = \sum_{k=0}^{\infty}(-b_k)z^k$. Setting $b_0 = -1$ the corresponding recursive formula to obtain the b_k's is given by

$$b_{k+1} = -\sum_{j=0}^{k} \left(1 - \frac{j}{k+1}\right) a_{k+1-j} b_j, \quad k = 0, 1, 2, \ldots. \tag{2.2.7}$$

A proof of (2.2.4) can be found in the section 2.10, Lemma 2.10.2. As we see from the proof of (2.2.4), this approach cannot be transferred directly to the multivariate case. Matrix multiplication is not commutative and therefore the exponential laws do not apply for matrices. However, these properties are essential for the proof of (2.2.4). Moreover, there are examples where (2.2.4) is not valid in the multivariate case. Consequently, the recursive formulae (2.2.6) and (2.2.7) cannot be directly applied to multivariate time series.

2.2.2 Estimating the Coefficients of the Wold Representation

Our next goal is to estimate the coefficients $\{c_k, k \in \mathbb{N}\}$ of the Wold representation (2.2.2). The basic idea is to use an estimator $\hat{f}_n$ of the spectral density f to get estimates of the Fourier coefficients of $\log(f)$ and to plug in these estimates into the recursive formula (2.2.6). Notice that estimates of the coefficients $\{b_k, k \in \mathbb{N}\}$ of the autoregressive representation (2.2.3) can be obtained by using formula (2.2.7) and the estimates of the a_k's.

Let $\hat{a}_{k,n} = (2\pi)^{-1} \int_{-\pi}^{\pi} \log(\hat{f}_n(\lambda)) \exp(-ik\lambda) d\lambda$ be the estimator of the k-th Fourier coefficient of $\log(f)$ and denote by $\{\hat{c}_{k,n}, k \in \mathbb{N}\}$, the estimators of the coefficients of the Wold representation obtained using formula (2.2.6), e.g.

$$\hat{c}_{0,n} = 1, \hat{c}_{k+1,n} \sum_{j=0}^{k} (1 - j/(k+1))\, \hat{a}_{k+1-j,n} \hat{c}_{j,n}, \quad k = 0, 1, 2, \ldots. \tag{2.2.8}$$

Let $\{\hat{b}_{k,n}\}$ be the corresponding estimators of the coefficients of $\{b_k\}$ using formula (2.2.7).

The calculation of Fourier coefficients of $\log f$ can be done efficiently by using the Fast Fourier Transform; see for instance Markel (1971). Since this algorithm is especially fast on girds identi-

cal to powers of 2, the sample sizes considered in the simulation study are chosen to be of that order. When approximating the Fourier coefficients of $\log \hat{f}_n$ by a sum over $m < \infty$ (not necessarily Fourier) frequencies to compute the moving average or autoregressive coefficients, respectively, one faces two sources of approximation errors. On the one hand, the approximated Fourier coefficients become periodic, i.e., it is not recommended to use the recursive formula for the computation of the coefficients beyond periodicity one. On the other hand, we get an approximation error which decreases with the smoothness of $\log \hat{f}_n$, see Epstein (2005) for details. Using the recursive formula this approximation error is transferred to the Wold's coefficients. If we assume that $\log \hat{f}_n$ is continuously differentiable then the k-th Wold's coefficient possesses an approximation error of order $O(1/(mk))$. The number of frequencies used depends on the computing power available and apart from numerical issues, the more frequencies are used the better. For spectral densities bounded away from zero, summability properties of the autocovariance function such as $\sum_h (1+|h|)^r |\gamma(h)| < \infty$, transfer to the Wold coefficients, hence, $\sum_{k=0}^{\infty}(1+k)^r|c_k| < \infty$. This follows by (2.4) and the Wiener-Levy-Theorem, see for instance Bhatt and Dedania (2003). Consequently, the coefficients usually decay rapidly and the approximation error quickly vanishes. To give an example, consider Model II used in the simulation study; see section 2.4. This model possesses the slowest decaying autocovariance of all three models considered. Nevertheless, using 1024 instead of 8192 Fourier frequencies to compute Wold's coefficients gives an overall squared error of less than 10^{-5}.

It is clear that the properties of the estimators $\hat{c}_{k,n}$ and $\hat{b}_{k,n}$ depend heavily on the properties of the estimator $\hat{f}_n$. To obtain consistency, the following condition suffices which essentially requires that $\hat{f}_n$ is a uniformly consistent estimator of f. For lag window estimators such a uniform consistency has been established by Jentsch and Subba Rao (2015, Lemma A.2), and for autoregressive spectral density estimators by Bühlmann (1995, Theorem 3.2).

Assumption 1 The estimator $\hat{f}_n$ satisfies $\int_{(-\pi,\pi]} \log(\hat{f}_n(\lambda))d\lambda > -\infty$. Furthermore,

$$\sup_{\lambda\in[0,\pi]} |\hat{f}_n(\lambda) - f(\lambda)| \xrightarrow{P} 0, \text{ as } n \to \infty. \tag{2.2.9}$$

Then, the following result can be established.

Theorem 2.2.1. *Suppose that f satisfies (2.2.1) and that Assumptions 1 holds true. Then, as $n \to \infty$,*

a) $\sup_{k\in\mathbb{N}} |\hat{a}_{k,n} - a_k| \xrightarrow{P} 0$,

and for every fixed $k \in \mathbb{N}$,

b) $\hat{c}_{k,n} \xrightarrow{P} c_k$,

c) $\hat{b}_{k,n} \xrightarrow{P} b_k$.

By the above theorem, for an M-dependent process, we have $\sum_{k=0}^{M} |c_k - \hat{c}_{k,n}| = o_P(1)$. Imposing more conditions on f and its estimator $\hat{f}_n$, the consistency properties of the estimators $\hat{a}_{k,n}$ and $\hat{c}_{k,n}$ can be refined and inequalities, similar to the well-known Baxter inequality for the AR-coefficients, Baxter (1962), can be established. Such inequalities are useful since they control the overall estimation error that occurs when the estimated spectral density $\hat{f}_n$ instead of the true spectral density f is used in order to obtain the estimates of interest.

Assumption 2 The estimator $\hat{f}_n$ fulfills the following conditions.

(i) There exists constants $0 < C_1 < C_2 < \infty$ such that $C_1 \leq \hat{f}_n(\lambda) \leq C_2$ for all $\lambda \in [0, \pi]$ and all $n \in \mathbb{N}$.

(ii) The first derivative of $\hat{f}_n$ with respect to λ exists, is continuous and integrable. Furthermore,

$$\sup_{\lambda \in [-\pi,\pi]} \left| \frac{\mathrm{d}}{\mathrm{d}\lambda} \hat{f}_n(\lambda) - \frac{\mathrm{d}}{\mathrm{d}\lambda} f(\lambda) \right| \xrightarrow{P} 0, \text{ as } n \to \infty. \tag{2.2.10}$$

Condition (ii) can be verified for lag window estimators by using similar arguments as in the proof of Lemma A.2 in Jentsch and Subba Rao (2015) under the same cumulant conditions and a slightly faster decay of the autocovariance function. However, in the case of the derivate the rate of convergence of the estimation error is slightly slower. For the autoregressive spectral density estimators the same condition can be verified by using arguments similar to those used in the proof of Theorem 3.2 in Bühlmann (1995). Notice that boundedness of the spectral density is ensured by an absolute summable autocovariance function, which is a common assumption for bootstrap procedures for time series. Furthermore, the assumption regarding the existence of derivatives of the spectral density can be transferred to assumptions on the summability of the autocovariance function. However, since the bootstrap approach proposed in this work is spectral-density-driven, we prefer to formulate the conditions needed as assumptions for the spectral density of the underlying process. The following theorem summarizes the properties of the estimators $\{\hat{a}_{k,n}, n \in \mathbb{N}\}$ and $\{\hat{c}_{k,n}, n \in \mathbb{N}\}$.

Theorem 2.2.2. *Let the spectral density f be strictly positive and bounded with continuous and integrable first derivative. Then, as $n \to \infty$,*

(a) If $\hat{f}_n$ satisfies Assumption 1 and Assumption 2(i) then

$$\sum_{k=-\infty}^{\infty} |\hat{a}_{k,n} - a_k|^2 = \int_0^{2\pi} |\log f(\lambda) - \log \hat{f}_n(\lambda)|^2 \mathrm{d}\lambda/(2\pi) \xrightarrow{P} 0 \tag{2.2.11}$$

and

$$\sum_{k=0}^{\infty} |\hat{c}_{k,n} - c_k|^2 \xrightarrow{P} 0. \tag{2.2.12}$$

(b) If $\hat{f}_n$ satisfies Assumption 1 and Assumption 2, then

$$\sum_{k=-\infty}^{\infty} k^2|\hat{a}_{k,n} - a_k|^2 \xrightarrow{P} 0 \text{ and } \sum_{k=1}^{\infty} |\hat{a}_{k,n} - a_k| \xrightarrow{P} 0. \tag{2.2.13}$$

Furthermore,

$$\sum_{k=0}^{\infty} k^2|\hat{c}_{k,n} - c_k|^2 \xrightarrow{P} 0 \text{ and } \sum_{k=0}^{\infty} |\hat{c}_{k,n} - c_k| \xrightarrow{P} 0. \tag{2.2.14}$$

Relation (2.2.4) plays a key role in the proofs of assertions (2.2.12) and (2.2.14). Notice that since $C(z)^{-1} = B(z)$, similar relations for $\{b_k, k \in \mathbb{N}\}$ can be derived. Furthermore, the results of Theorem 2.2.2 can be straightforwardly extended to the sequence of estimation errors $\{(\hat{b}_{k,n} - b_k), k \in N\}$.

There are some alternative approaches to estimate the coefficients c_k and b_k which have been proposed in the literature. In particular and for estimating the coefficients c_k, one option is the innovation-algorithm which works by fitting MA(q) models where the order q increases to infinity as the sample size n increases to infinity; see (Brockwell and Davis, 1991, Proposition 5.2.2). For estimating the coefficients b_k, commonly an AR(p) model is fitted to the time series at hand by means of Yule-Walker estimators, where, the order p is also allowed to increase to infinity with sample size; see (Brockwell and Davis, 1991, Section 8.1). Under certain conditions, both approaches are consistent; see (Pourahmadi, 2001, Theorem 7.14). However, the basic idea behind these approaches differs from ours and so do the estimators obtained via spectral density factorization. In the above mentioned approaches, the estimated autocovariance matrix is used to fit a finite moving average or a finite autoregressive model. Consistency of the corresponding estimators is then obtained by allowing the order of the fitted model to increase to infinity at an appropriate rate as the sample size n increases to infinity. These approaches face, therefore, two sources of errors. The first is the estimation error which is caused by the fact that estimated autocovariances are used instead of true ones. The second is the approximation error which is due to the fact that a finite order model is used to approximate the underlying infinite order structure. Although the estimation error cannot be avoided, our approach does not necessarily use a finite order model since it is based on a (uniformly) consistent estimator $\hat{f}_n$ of the spectral density f. So the approximation error caused by our estimation procedure is different. This error depends on the quality of the spectral density estimator $\hat{f}_n$ used to approximate the true spectral density f, where $\hat{f}_n$ is selected from a wide range of possible estimators and not only from those obtained by using finite order autoregressive or moving average parametric models. The innovation-algorithm is similar to the factorization of autocovariance matrices which is used in the linear process bootstrap. In section 2.8 a simple example is discussed to point out the differences between factorizing autocovariance matrices and spectral densities.

2.2.3 Spectral Density Estimators

Since our estimation procedure relies on a spectral density estimator $\hat{f}_n$, we briefly discuss the variety of such estimators that can be used and their impact on the estimators $\{\hat{c}_{k,n}\}$ or $\{\hat{b}_{k,n}\}$ obtained.

As already mentioned, spectral densities can be estimated using a parametric approach, that is, by fitting a parametric model to the time series at hand and using the spectral density of the fitted model as an estimator of the spectral density of the process. Since autoregressive models are easy to fit, they are commonly used for such a purpose; see Akaike (1969), Shibata (1981) and Brockwell and Davis (1991, Section 10.6). In this context, parameter estimators, like Yule-Walker estimators, are popular because they ensure invertibility of the corresponding estimated AR-polynomial; see Brockwell and Davis (1991, Section 8.1). Now, if an autoregressive spectral density estimator is used in the spectral factorization procedure, then the estimated coefficients $\{\hat{c}_{k,n}\}$ obtained are identical to those appearing in the power series expansion of the inverted estimated autoregressive polynomial. Furthermore, the corresponding sequence of estimated coefficients $\{\hat{b}_{k,n}\}$ is finite and the $\hat{b}_{k,n}$'s, $k \in \{1, 2 \ldots, p\}$, coincide with the estimated autoregressive parameters.

Using nonparametric methods like lag window or kernel smoothed periodogram estimators is another popular approach to estimate the spectral density; cf. Brockwell and Davis (1991, Section 10.4). Lag window estimators truncate the estimated autocovariances at a given lag controlled by a truncation parameter. Such estimators of the spectral density can be interpreted as obtained by (implicitly) fitting a finite order moving average model to the time series at hand; see also Brockwell and Davis (1991, Prop. 3.2.1). The sequence of estimated coefficients $\{\hat{c}_{k,n}\}$ of the Wold representation obtained by using such a spectral density estimator is finite with $\hat{c}_{k,n} = 0$ for values of k larger than the truncation parameter. Due to the asymptotic equivalence between lag window and smoothed periodogram estimators, similar remarks can be made also for spectral density estimators obtained by smoothing the periodogram. Furthermore, as mentioned in section 2.2, lag window estimators as well as autoregressive estimators satisfy Assumptions 1 and 2, see (Jentsch and Subba Rao, 2015, Lemma A.2) and (Bühlmann, 1995, Theorem 3.2) respectively.

A different nonparametric approach to estimate the spectral density is to truncate the Fourier series of $\log(f)$ which presumes an exponential model for the spectral density; see Bloomfield (1973). Such a model is given by $f(\lambda) = (2\pi)^{-1}\sigma^2 \exp\{2\sum_{j=1}^{r} \theta_j \cos(\lambda j)\}$. Unlike truncating the autocovariance function, non-negative definiteness of the spectral density f is ensured for all possible values of the parameters $\theta_j, j = 1, \ldots, r$. As Bloomfield (1973) pointed out, the autocovariance function of such an exponential model cannot, in general, be described by a finite autoregressive or a finite moving average model. Thus, using such an estimator of the spectral density in the factorization algorithm leads to an infinite sequence of estimators $(\hat{c}_{k,n})$ or $(\hat{b}_{k,n})$ respectively. Notice that the Fourier coefficients of $\log(f)$ are also known as the cepstral coefficients or vocariances and

they have been widely used in the signal processing literature to estimate the spectral density; see Stoica and Sandgren (2006) and Kaderli and Kayhan (2000). However, Stoica and Sandgren (2006) define the cepstral coefficients as the finite approximation over N Fourier frequencies and the integral definition, as we have used in (2.2.5), is called 'theoretical cepstrum'; see Stoica and Sandgren (2006, Eq. (8)). The finite approximated cepstral coefficients cannot be linked directly without error to Wold's coefficients.

An interesting combination of nonparametric and parametric approaches for spectral density estimation is offered by the so-called pre-whitening approach; see Blackman and Tukey (1958). The idea is to use a parametric model to filter the time series and then apply a nonparametric estimator to the time series of residuals. Using an AR-model for pre-whitening (filtering) and a lag window estimator for estimating the spectral density of the residuals, can be interpreted as (implicitly) fitting an ARMA-model to the time series at hand. The idea is that the parametric AR-model fit is able to represent the peaks of the spectral density quite well while the lag window estimator applied to the residuals can capture features of the spectral density that are not covered by the parametric fit. Notice that for the pre-whitening approach consistency of the lag window estimator is obtained even in the case, where the parametric fit does not improve the estimation. However, since only $n - p$ instead of n observation are used, the rate of converge is slightly slower. Consequently even for n-dependent p, as long as $n - p(n) \to \infty$ as $n \to \infty$ the pre-whitening approach is consistent and satisfy Assumption 1 and 2. Using such a spectral density estimator for the factorization algorithm the coefficients $\{\hat{c}_{k,n}\}$ and $\{\hat{b}_{k,n}\}$ obtained will be those of the infinite order moving average representation and infinite order autoregressive representation of the (implicitly) fitted ARMA model, respectively. However, to reduce numerical errors, the use of the ARMA representation is recommend, the moving average coefficients are obtained by the factorization of the pre-whitened spectral density and the autoregressive coefficients are those of the fitted AR-model.

2.3 Spectral-Density-Driven Bootrstrap

2.3.1 The Spectral-Density-Driven Bootrstrap Procedure

In the previous section we have dealt with the coefficients $\{c_k, k \in \mathbb{N}\}$ of the moving average and $\{b_k, k \in \mathbb{N}\}$ of the autoregressive representation of the process. For the coefficients in both representations, consistent estimators have been developed. Consequently, both representations can be used in principle to develop a bootstrap procedure to generate pseudo time series $X_1^*, X_2^*, \ldots, X_n^*$. We focus in this work on the moving average representation, since it exists for every spectral density. Clearly, such a bootstrap procedure will be determined by the spectral density estimator $\hat{f}_n$ used to obtain the coefficients $\{\hat{c}_{k,n}\}$ and by the generated series of pseudo innovations $\{\epsilon_t^*\}$ (cf. Step 3 below). Thus, the tuning parameters of this bootstrap procedure coincide with those used

for the spectral density estimation. Consequently, one can follow data-driven methods proposed in the literature to choose these parameters. Now, given an estimator $\hat{f}_n$ of the spectral density f, the spectral-density-driven bootstrap algorithm consists of the following steps.

Step 1. Compute the Fourier coefficients of $\log(\hat{f}_n)$ given by
$\hat{a}_{k,n} = 1/(2\pi) \int_0^{2\pi} \log(\hat{f}_n(\lambda)) \exp(-ik\lambda) d\lambda$ for $k = 1, 2, \ldots$.

Step 2. Let $\hat{\sigma}_n^2 = 2\pi \exp(\hat{a}_{0,n})$ and compute the coefficients $\hat{c}_{k,n}, k = 1, 2, \ldots$ using the formula $\hat{c}_{k+1,n} = \sum_{j=0}^{k} (1 - j/(k+1)) \hat{a}_{k+1-j,n} \hat{c}_{j,n}, k = 0, 1, 2, \ldots$, and the starting value $\hat{c}_{0,n} = 1$.

Step 3. Generate i.i.d. pseudo innovations $\{\varepsilon_t^*, t \in \mathbb{Z}\}$ with mean zero and variance $\hat{\sigma}_n^2$.

Step 4. The pseudo time series $X_1^*, X_2^*, \ldots, X_n^*$ is then obtained as $X_t^* = \sum_{j=0}^{\infty} \hat{c}_{j,n} \varepsilon_{t-j}^* + \bar{X}_n$, $t = 1, 2, \ldots, n$, where $\bar{X}_n = n^{-1} \sum_{t=1}^{n} X_t$ is the sample mean.

It should be stressed that the above bootstrap algorithm with i.i.d. pseudo innovations represents a general procedure to generate a pseudo time series stemming from a linear process. Regarding the particular generation of the i.i.d. innovations in Step 3, different possibilities can be considered depending on the stochastic properties of the time series at hand which should be mimicked by the pseudo time series $X_1^*, X_2^*, \ldots, X_n^*$. In particular, suppose that $X_1, X_2, \ldots, X_n$ stems from a linear process and that a statistic $T_n = T(X_1, X_2, \ldots, X_n)$ is considered, the distribution of which should be approximated by the bootstrap. We then propose to generate the i.i.d. innovations in a way which asymptotically matches the first, the second and the fourth moment structure of the true innovations ε_t. Matching also the fourth moment structure of ε_t turns out to be important for some statistics T_n; we refer to section 2.3.3 for examples.

One possibility to achieve this requirement is, to generate the ε_t^*'s as i.i.d. random variables with the following discrete distribution: $P\left(\varepsilon_t^* = \hat{\sigma}_n \sqrt{\tilde{\kappa}_4}\right) = P\left(\varepsilon_t^* = -\hat{\sigma}_n \sqrt{\tilde{\kappa}_4}\right) = 1/(2\tilde{\kappa}_4)$ and $P(\varepsilon_t^* = 0) = 1 - 1/\tilde{\kappa}_4$. Here $\tilde{\kappa}_4 = \tilde{\kappa}_{4,n}/\hat{\sigma}_n^4 > 0$ and $\tilde{\kappa}_{4,n}$ denotes a consistent estimator of the fourth moment $E(\varepsilon_1^4)$ of the innovations ε_t. Consistent, nonparametric estimators of κ_4 have been proposed in Kreiss and Paparoditis (2012) and Fragkeskou and Paparoditis (2015).

In the above bootstrap algorithm, the pseudo time series $X_1^*, X_2^*, \ldots, X_n^*$ is generated using the estimated coefficients of the moving average representation. Modifying the algorithm appropriately, the pseudo time series can be also generated using the estimated autoregressive representation of the process. For this, we set $\hat{\sigma}_n^2 = 2\pi \exp(\hat{a}_{0,n})$ and calculate the coefficients $\hat{b}_{k,n}$, $k = 0, 1, 2, \ldots$ using the recursive formula starting with $\hat{b}_{0,n} = -1$ and $\hat{b}_{k+1,n} = -\sum_{j=0}^{k} (1 - j/(k+1)) \hat{a}_{k+1-j,n} \hat{b}_{j,n}$, for $k = 0, 1, 2, \ldots$. Using these estimates of the coefficients of the autoregressive representation, the pseudo time series is then obtained as $X_t^* = \sum_{j=1}^{\infty} \hat{b}_{j,n} (X_{t-j}^* - \bar{X}_n) + \varepsilon_t^* + \bar{X}_n$.

Here, we stress the fact that the spectral-density-driven bootstrap should not be considered as an MA-sieve bootstrap procedure, where the order of the moving average model is allowed to increase

to infinity as the sample size increases to infinity. The spectral-density-driven bootstrap procedure is rather governed by the spectral density estimator $\hat{f}_n$ used, which appropriately describes the entire autocovariance structure of the underlying process. The moving average representation in this bootstrap procedure is solely used as a device to generate a time series with a second-order structure characterized by the spectral density estimator $\hat{f}_n$ used. Notice however, that some spectral density estimators can implicitly lead to an MA-sieve type bootstrap.

The spectral-density-driven bootstrap can be easily used in *R*, R Core Team (2016). Apart form the recursive formulas all parts are already implemented. An *R*-code example to generate pseudo bootstrap time series with the spectral-density-driven bootstrap can be found at `www.tu-bs.de/Medien-DB/stochastik/code-snippet_sddb.txt`

2.3.2 Comparison with other Linear Bootstrap Procedures

The idea of the AR-sieve bootstrap is to fit a p-th order autoregressive model to the time series at hand and to use the estimated model structure together with i.i.d. pseudo innovations generated according to the empirical distribution function of the centered residuals. In order to fully cover the second-order dependence structure of the underlying process X, the order p of the fitted AR-model is allowed to increase to infinity (at an appropriate rate) as the sample size increases to infinity; see Kreiss (1992), Paparoditis and Streitberg (1991), and Bühlmann (1997). The range of validity of this bootstrap procedure has been investigated in Kreiss et al. (2011). As already mentioned, the AR-sieve bootstrap is a special case of the spectral-density-driven bootstrap described in section 2.3.1 when $\hat{f}_n$ is chosen to be a parametric AR(p) spectral density estimator and the innovations $\{\varepsilon_t^*\}$ are generated through i.i.d. resampling from the centered residuals of the autoregressive fit. Using the estimated AR-parametric spectral density, the factorization algorithm leads to a sequence $\{\hat{c}_{k,n}\}$ of estimated moving average coefficients that correspond to the MA(∞) representation obtained by inverting the estimated autoregressive polynomial. However, and as already mentioned, the spectral-density-driven bootstrap is a much more general procedure since it is not restricted to describing the dependence structure of the time series at hand by means of a finite order parametric autoregressive model. Notice that both bootstrap approaches work under similar conditions, see Assumptions 1 and 2. However, if a lag window spectral density estimator is used, there are situations where the spectral-density-driven bootstrap is valid, whereas validity of the AR-sieve is not clear; see section 2.3.3 for details.

The linear process bootstrap, established by McMurry and Politis (2010) is also related to the spectral-density-driven bootstrap. It uses the factorization of banded autocovariance matrices instead of the spectral density itself to generate the pseudo observations. A factorization of autocovariance matrices is similar to the innovation algorithm, see Brockwell and Davis (1991, Proposition 5.2.2). As pointed out at the end of section 2.2.2 this leads in finite sample situations to different

results. Furthermore, the linear process bootstrap aims to generate a data vector with a given covariance structure, while the spectral-density-driven bootstrap generates a stationary time series. A more detailed discussion can be found in the section 2.8.

2.3.3 Bootstrap Validity

In this section we prove validity of the proposed spectral-density-driven bootstrap procedure for the sample mean and under quite general dependence assumptions on the underlying process which go far beyond linearity. Furthermore, we show that if the underlying process is linear, the same bootstrap procedure driven by i.i.d. pseudo innovations is valid for the class of so-called generalized autocovariance statistics. We first focus on this general class of statistics which appears to be more involved than that of the mean.

Definition 2.3.1. *Let $\{d_p(n), n \in \mathbb{Z}\}$ be a sequence of real numbers such that $\sum_{h\in\mathbb{Z}} |d_p(h)| < \infty$, where $p \in \{1, 2, \ldots, P\}$. Let further $g : \mathbb{R}^P \to \mathbb{R}$ be a differentiable function. Then, the generalized autocovariance statistic is defined as*

$$\hat{T}_n = g(\hat{T}_{n,1}, \ldots, \hat{T}_{n,P}), \text{ where for } p \in \{1, \ldots, P\}, \tag{2.3.1}$$

$$\hat{T}_{n,p} = 1/n \sum_{t=1}^{n} \sum_{h=1-t}^{n-t} d_p(h)(X_t - \bar{X}_n)(X_{t+h} - \bar{X}_n) \text{ and } \bar{X}_n = 1/n \sum_{t=1}^{n} X_t.$$

The above class of statistics contains, among others, sample autocovariances, sample autocorrelations and lag window spectral density estimators. To elaborate, let $h \in \{0, \ldots, n-1\}$ and set $d_1(h) = 1$ and $d_1(x) = 0$ for $x \neq h$. We then have that $\hat{T}_{n,1} = 1/n \sum_{t=1}^{n} \sum_{s=1}^{n} \mathbb{1}_{\{t-s=h\}}(X_t - \bar{X}_n)(X_s - \bar{X}_n) = \hat{\gamma}_n(h)$. Similarly for $d_2(0) = 1$ and $d_2(x) = 0$ for $x \neq 0$, we get that $\hat{T}_{n,2} = \hat{\gamma}_n(0)$. Furthermore, the sample autocorrelation at lag h is obtained by choosing $g(x, y) = x/y$. Lag window spectral density estimators are also included in the above class. For this, one chooses $d(h) = \mathbb{1}_{\{h\leq M\}} 1/(2\pi) K(h/M) (\mathbb{1}_{\{h=0\}} + \mathbb{1}_{\{h>0\}}(\exp(-ih\omega) + \exp(ih\omega)))$, where K is some appropriate smoothing kernel.

Assumption 3: $\{X_t, t \in \mathbb{Z}\}$ is a linear process given by $X_t = \sum_{j\in\mathbb{Z}} \varphi_j \varepsilon_{t-j} + \mu, \mu \in \mathbb{R}$ with i.i.d. innovations $\{\varepsilon_t, t \in \mathbb{Z}\}$, where $E\varepsilon_t = 0$, $E\varepsilon_t^2 = \sigma_\varepsilon^2$, $E\varepsilon_t^4 = \kappa_4$ and $E\varepsilon_t^8 < \infty$. We write for short $\varepsilon_t \sim IID(0, \sigma_\varepsilon^2, \kappa_4)$. The coefficients in the moving average representation fulfill the summability condition $\sum_{j\in\mathbb{Z}} |j\varphi_j| < \infty$.

As the following theorem shows, the proposed spectral-density-driven bootstrap procedure is valid for approximating the distribution of statistics belonging to the class of generalized autocovariances. Here and in the sequel, for two random variables X, and Y, $d_2(X, Y)$ denotes Mallow's distance, i.e., $d_2(X, Y) = \{\int_0^1 \left(F_X^{-1}(x) - F_Y^{-1}(x)\right)^2 dx\}^{1/2}$, where F_X and F_Y denote the cumulative distribution functions of X and Y, respectively.

Theorem 2.3.1. *Let* $\hat{T}_n^* = g(\hat{T}_{n,1}^*, \ldots, \hat{T}_{n,P}^*)$, *where*

$$\hat{T}_{n,p}^* = 1/n \sum_{t=1}^{n} \sum_{h=1-t}^{n-t} d_p(h)(X_t^* - \bar{X}_n^*)(X_{t+h}^* - \bar{X}_n^*), \text{ for } p = 1, \ldots, P,$$

and $(d_p(h))_{h\in\mathbb{Z}}$ *is a sequence of real numbers as in Definition 2.3.1. Furthermore,* $X_1^*, X_2^*, \ldots, X_n^*$ *is a pseudo time series generated using the spectral-density-driven bootstrap procedure with a pseudo innovation process* $\{\varepsilon_t^*, t \in \mathbb{Z}\}$ *satisfying* $\varepsilon_t^* \sim IID(0, \hat{\sigma}_n^2, \hat{\kappa}_{4,n})$ *with* $\hat{\kappa}_{4,n} = E^*(\varepsilon_t^*)^4$, *a consistent estimator of* κ_4 *which also fulfills* $\sup_{n\in\mathbb{N}} \hat{\kappa}_{4,n} \leq C$ *for some constant* $C < \infty$ *which does not depend on* n. *Finally, assume that the estimated Wold coefficients fulfill* $\sum_{k\in\mathbb{N}} |c_k - \hat{c}_{k,n}| = o_P(1)$ *and* $\sum_{k\in\mathbb{N}} |k\hat{c}_{k,n}| \leq C$. *Then under Assumption 3 and as* $n \to \infty$,

$$d_2(\sqrt{n}(\hat{T}_n^* - E^*\hat{T}_n^*), \sqrt{n}(\hat{T}_n - E\hat{T}_n)) \to 0, \text{ in probability.}$$

The assumptions $\sup_{n\in\mathbb{N}} \hat{\kappa}_{4,n} \leq C < \infty$ and $\sum_{k\in\mathbb{N}} |k\hat{c}_{k,n}| \leq C$ are of rather technical nature and can be satisfied by using appropriate estimators of $\hat{\kappa}_4$ and $\hat{f}_n$. If the spectral density estimator $\hat{f}_n$ fulfills $\sup_{\lambda\in(-\pi,\pi]} \frac{\mathrm{d}^3}{\mathrm{d}\lambda^3} \log \hat{f}_n(\lambda) \leq C$ then the requirement $\sum_{k\in\mathbb{N}} |k\hat{c}_{k,n}| \leq C$ of the above theorem is satisfied. Notice that sufficiently smooth kernels guarantee the required differentiability of $\log \hat{f}_n$. Furthermore, by using an appropriate truncation, boundedness of $\hat{\kappa}_{4,n}$ and $\hat{f}_n$ can also be guaranteed.

In section 2.2 we gave conditions under which $\sum_{k\in\mathbb{N}} |c_k - \hat{c}_{k,n}| = o_P(1)$ holds, see Theorem 2.2.1 and 2.2.2. Moreover, there are settings in which it is not clear whether the AR-sieve bootstrap is valid while the spectral-density-driven bootstrap in connection with a lag window spectral density estimator can lead to a valid approximation. For instance, the spectral-density-driven bootstrap remains valid for statistics $\hat{T}_n$ as in (2.3.1) when the time series is generated by finite moving average processes with unit roots, like for instance by the process $X_t = \varepsilon_t - \varepsilon_{t-1}$ or even by nonlinear continuous transformations of M-dependent stationary processes.

The following theorem establishes validity of the spectral-density-driven bootstrap for the case of the sample mean, which is not covered by the class of general covariance statistics T_n given in (2.3.1). Notice, that for this case, it suffices that the pseudo innovations $\{\varepsilon_t^*\}$ mimic asymptotically correct only the first and the second moment of the true innovations ε_t. Furthermore, no linearity assumptions of the underlying processes X are needed. What is needed is that $\sqrt{n}(\bar{X}_n - \mu)$ converges to a normal distribution with variance $2\pi f(0)$, which, however, is fulfilled for a huge class of stationary processes. For instance, appropriate mixing or weak dependence conditions are sufficient for this statistic to satisfy the required asymptotic normality of $\sqrt{n}(\bar{X}_n - \mu)$. Furthermore, regarding the spectral-density-driven bootstrap, the spectral density f and its estimator $\hat{f}_n$ need to fulfill less restrictive conditions. In particular, for a lag window spectral density estimator $\hat{f}_n$, the assumptions $|\gamma(h)| \leq C/|h|^{2+\varepsilon}$ and $\sup_t \sum_{t_1,t_2,t_3} |\mathrm{cum}(X_t, X_{t_1}, X_{t_2}, X_{t_3})| < \infty$, see Jentsch and Subba Rao (2015,

Lemma A.2), suffice to ensure uniform consistency of $\hat{f}_n$. Under the same cumulant condition and the absolute summability of autocovariance function, consistency of block bootstrap approaches can be established for the sample mean; see Künsch (1989), Politis and Romano (1994), and Nordman (2009) for details. Thus, the spectral-density-driven bootstrap is applicable in similar settings as the block-related bootstrap approaches.

Theorem 2.3.2. *Assume that $\{X_t : t \in \mathbb{Z}\}$ is a purely nondeterministic stationary process with mean μ, spectral density f, and autocovariance γ with $\sum_h |\gamma(h)| < \infty$ and assume that $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} \mathcal{N}(0, 2\pi f(0))$, as $n \to \infty$. Denote by $\hat{f}_n$ a uniformly consistent and bounded estimator of f fulfilling Assumptions 1 and $\sum_{k=0}^{\infty} |\hat{c}_{k,n}| < C$, where C does not depend on n. Assume that $X_1^*, X_2^*, \ldots, X_n^*$ is generated by using the spectral-density-driven bootstrap procedure with an i.i.d. innovation process $\{\varepsilon_t^*, t \in \mathbb{Z}\}$, where $E^*(\varepsilon_t^*) = 0$, $E^*(\varepsilon_t^*)^2 = \hat{\sigma}_n^2$, and $E^*(\varepsilon_t^*)^4 < C < \infty$. Then, as $n \to \infty$,*

$$d_2(\sqrt{n}(\bar{X}_n^* - \bar{X}_n), \sqrt{n}(\bar{X}_n - \mu)) \to 0, \text{ in probability.}$$

The assumption $\sum_{k=0}^{\infty} |\hat{c}_{k,n}| < C$ is satisfied if a strictly positive, differentiable, and bounded spectral density estimator $\hat{f}_n$ is used.

Notice that validity of block bootstrap approaches is often established for so-called generalized mean statistics, see Künsch (1989, Example 2.2). For a time series $X_1, \ldots, X_n$, this class of statistics is given by

$$T_n = h\left(1/(n-m+1)\sum_{t=1}^{n-m+1} Y_t\right), \text{ where } h : \mathbb{R}^k \to \mathbb{R}^s, s \leq k,$$

and

$$Y_t = g(X_t, X_{t+1}, \ldots, X_{t+m-1}), t = 1, \ldots, n-m+1, g : \mathbb{R}^m \to \mathbb{R}^k, k \leq m < n$$

and m is fixed. Let $\tilde{n} = n - m$. The validity of the spectral-density-driven bootstrap for this class can be derived by applying the results of Theorem 2.3.2. The stated cumulant and autocovariance conditions have to be fulfilled by the process $\{Y_t, t \in \mathbb{Z}\}$.

Corollary 2.3.1. *Let $Y = \{Y_t : t \in \mathbb{Z}\}$ fulfill the assumptions of Theorem 2.3.2 and denote the mean by $\mu_Y = EY_1$. Furthermore, assume that h is differentiable at μ_Y and $Y_1^*, \ldots, Y_{\tilde{n}}^*$ is generated using the spectral-density-driven bootstrap procedure with an i.i.d. innovation process $\{\varepsilon_t^*, t \in \mathbb{Z}\}$, where $E^*(\varepsilon_t^*) = 0$, $E^*(\varepsilon_t^*)^2 = \hat{\sigma}_n^2$, and $E^*(\varepsilon_t^*)^4 < C < \infty$. Then, as $\tilde{n} \to \infty$,*

$$d_2(\sqrt{\tilde{n}}(h(\bar{Y}_{\tilde{n}}^*) - h(E^*Y*)), \sqrt{\tilde{n}}(h(\bar{Y}_{\tilde{n}}) - h(\mu_Y)) \to 0, \text{ in probability.}$$

An improved finite sample performance of bootstrap approximations is often achieved by applying the bootstrap to studentized statistics, see for instance Lahiri (2003, Chapter 6); Götze and Künsch (1996); Romano and Wolf (2006). A studentized form is obtained by normalizing the statis-

tic of interest with a consistent estimator of the asymptotic standard deviation. Since in Theorem 2.3.2 the asymptotic variance is given by $2\pi f(0)$ and this quantity can be consistently estimated, we get $\sqrt{n}(\bar{X}_n - \mu)/(2\pi\tilde{f}_n(0))^{-1/2}$ as a studentized statistic where $\hat{f}_n$ is a consistent estimator of f. A bootstrap approximation of this studentized statistic is then given by $\sqrt{n}(\bar{X}_n^* - \bar{X}_n)/(2\pi\tilde{f}_n^*(0))^{-1/2}$, where $\tilde{f}_n^*$ is the same spectral density estimator as $\tilde{f}_n$ obtained using the pseudo observations $X_1^*, \ldots, X_n^*$.

Corollary 2.3.2. *Let $f(0) > 0$ and $\tilde{f}_n(0), \tilde{f}_n^*(0)$ be consistent estimators of $f(0)$ which are bounded from below by $\delta > 0$. Under the assumption of Theorem 2.3.2 and if the spectral density estimator used for the spectral-density-driven bootstrap is two times differentiable with a second derivative of bounded variation independent from n, then, as $n \to \infty$,*

$$d_2(\sqrt{n}(\bar{X}_n^* - \bar{X}_n)/(2\pi\tilde{f}_n^*(0))^{1/2}, \sqrt{n}(\bar{X}_n - \mu)/(2\pi\tilde{f}_n(0))^{1/2}) \to 0, \text{ in probability.}$$

The asymptotic variance of the generalized autocovariance statistic depends on the spectral density and it may also depend on the fourth moment κ_4 of the underlying innovations of the linear process. This fourth moment can be estimated consistently, by say $\hat{\kappa}_4$; see Fragkeskou and Paparoditis (2015). Since the pseudo time series $\{X_t^*\}$ is driven by i.i.d. innovations, the fourth moment of $\{\varepsilon_t^*\}$ can be estimated using the same estimator as for κ_4. Consequently, an asymptotically valid approximation of the spectral-density-driven bootstrap for studentized generalized autocovariance statistics can be established. This is done in the following corollary, where, and in order to simplify notation, only the case $P = 1$ is considered. In this case the statistic of interest is given by $\hat{T}_n = 1/n \sum_{t=1}^{n} \sum_{h=1-t}^{n-t} d(h)(X_t - \bar{X}_n)(X_{t+h} - \bar{X}_n)$ and the asymptotic variance by

$$\tau^2 = (\kappa_4/\sigma^4 - 3)(\int_0^{2\pi} f(\lambda) \sum_{h\in\mathbb{Z}} d(h) \exp(ih\lambda) \mathrm{d}\lambda)^2 + 4\pi \int_0^{2\pi} |f(\lambda) \sum_{h\in\mathbb{Z}} d(h) \exp(ih\lambda)|^2 \mathrm{d}\lambda.$$

Corollary 2.3.3. *Let $\tau^2 > \delta > 0$ and let $\tilde{f}_n, \tilde{f}_n^*$ be consistent spectral density estimators which are bounded from below by $\delta > 0$. Furthermore, let $\tilde{\kappa}_{4,n}, \tilde{\kappa}_{4,n}^*$ be consistent estimators of κ_4. Under the assumptions of Theorem 2.3.1 and if $E^*(\varepsilon_1^*)^8 < C$ independent from n then, as $n \to \infty$,*

$$d_2(\sqrt{n}(\hat{T}_n^* - E^*\hat{T}_n^*)/\tilde{\tau}_n^*, \sqrt{n}(\hat{T}_n - E\hat{T}_n)/\tilde{\tau}_n) \to 0, \text{ in probability.}$$

The assumption $\tau^2 > \delta > 0$ ensures that $\hat{T}_n$ converges to a non-degenerate distribution. It is fulfilled if $\kappa_4/\sigma^4 > \tilde{\delta} > 1$ or if $f(\cdot) \sum_{h\in\mathbb{Z}} d(h) \exp(ih\cdot)$ is a non-constant function. The estimators $\tilde{\tau}_n^*$ and $\tilde{\tau}_n$ are estimators of τ based on $\tilde{f}_n^*$ and $\tilde{\kappa}_{4,n}^*$ and $\tilde{f}_n$ and $\tilde{\kappa}_{4,n}$, respectively.

2.4 Numerical Examples

2.4.1 Simulations

In this section we investigate by means of simulations the finite sample behavior of the spectral-density-driven bootstrap and compare its performance with that of two other linear bootstrap methods, the AR-sieve bootstrap and the linear process bootstrap. We also compare all three linear bootstrap methods with the tapered block bootstrap, cf. Paparoditis and Politis (2001), and the moving block bootstrap, cf. Künsch (1989). Two statistics T_n are considered, the sample mean $\overline{X}_n$ and the sample autocorrelation $\widehat{\rho}(2) = \widehat{\gamma}(2)/\widehat{\gamma}(0)$. The time series used have been generated from the following three models:

Model I: $X_t = 0.9X_{t-1} + \varepsilon_t$,

Model II: $X_t = 1.34X_{t-1} - 1.88X_{t-2} + 1.32X_{t-3} - 0.8X_{t-4} + \varepsilon_t + 0.71\varepsilon_{t-1} + 0.25\varepsilon_{t-2}$,

Model III: $X_t = \varepsilon_t + \sum_{k=1}^{10} \binom{n}{k}(-1)^k \varepsilon_{t-k}$.

In all cases the innovation process $\{\varepsilon_t\}$ consists of i.i.d. random variables having a t-student distribution with 3 degrees of freedom and variance normalized to 1. Model I is tailor made for the AR-sieve bootstrap. The spectral density in Model II has strong peak around frequency $\lambda = 1.5$ which can be estimated difficulty. Furthermore, this model possesses a slowly decaying autocovariance function which oscillates with two frequencies; one for the odd lags and one for the even lags. Model III is an moving average process with a unit root; the spectral density is zero at frequency zero. Consequently, the sample mean converges to a degenerated distribution making a studentization inappropriate. In order to investigate the finite sample performance of the different bootstrap methods, empirical coverage probabilities of two-sided confidence intervals obtained for the levels $\alpha = 0.2, 0.1$ and 0.05 are presented. The empirical coverage probabilities are based on $2,000$ realizations of each process and $B = 1,000$ bootstrap repetitions. Here, we present the results for the case $n = 128$, while results for the case $n = 512$ are given in section 2.9.

For the AR-sieve bootstrap, denoted by *ARS*, the Akaike's information criterion (AIC) is used to select the autoregressive order p, cf. Akaike (1969). The *SDDB* is applied using an AR-pre-whitening, nonparametric estimator of the spectral density, where the order of the autoregressive part has been selected by the AIC and a smoothed periodogram is used with Gaussian kernel and of bandwidth selected by cross-validation; see Beltrão and Bloomfield (1987). Furthermore, for this bootstrap procedure, i.i.d. Gaussian innovations are used. Furthermore, the linear process bootstrap, denoted by *LPB*, has been implemented as in McMurry and Politis (2010), and the tapered block bootstrap, denoted by *TBB*, has been applied with a block length choice and a tapering window as in Paparoditis and Politis (2001). Due to the strong dependence of some of the models considered, this rule for

choosing the block length leads to unfeasible results especially for small sample sizes. For instance, even for $n = 512$ this rule delivers for Model II block lengths of around 400. For this reason, we also consider the moving block bootstrap with nonrandom block length given by $l = n^{1/3}$. This procedure is denoted by *BB*.

As mentioned in section 2.3.3, a better finite sample performance may be obtained by using bootstrap approximations of studentized statistics. Thus, we consider for the sample mean the statistic $\bar{X}_n(2\pi\hat{f}_n(0))^{-1/2}$, where $\hat{f}_n$ is the same spectral density estimator as the one used for *SDDB*. The sample autocorrelation is studentized as well, where the variance is estimated by Bartlett's formula, Brockwell and Davis (1991, Theorem 7.2.1), based on the autocorrelation function corresponding to the estimated spectral density $\hat{f}_n$. Finally, a standard normal distribution is considered as a further competitor for the studentized statistics and is denoted in the following by *ND*. For non-studentized statistics a normal distribution is used with the variance estimated by using the *SDDB* procedure. Studentization brings clear improvements for all models and all statistics considered. Hence, the focus is on the studentized case and the non-studentized tables can be found in the section 2.9.

The coverage probabilities for the studentized sample mean are displayed in Table 2.1. As it is seen from Table 2.1, none of the competitors outperforms the *SDDB* procedure. In fact, in many cases the *SDDB* performs best. Finally, and for Model III it seems that only the *SDDB* procedure gives reasonable estimates. Notice that the spectral density of Model III is not bounded away from zero, that is, it is not clear whether the *LPB* or the *ARS* are valid in this case. The coverage probabilities for the studentized sample autocorrelation are displayed in Table 2.2. For this statistic over all, the most accurate coverage probabilities are those obtained by using the *ARS* and the *SDDB* procedures.

Notice that block bootstrap methods have their strength in their general applicability, i.e., they are applicable not only to linear processes, like those considered in the simulation study, and to a broad class of statistics. Consequently, it is not surprising that these methods do not perform best for the linear processes considered.

Summarizing our numerical findings, it seems that the spectral-density-driven bootstrap performs very good in all model situations and for both statistics considered. In combination with a flexible spectral density estimator, like for instance the pre-whitening based estimator used in the simulations, the spectral-density-driven bootstrap seems to be a valuable tool for bootstrapping time series.

2.4.2 A Real-Life Data Example

We consider the time series of annual measurements of the water level, in feet, of Lake Huron; cf. Series A in the Appendix of Brockwell and Davis (1991) or in the *R*-package *datasets::LakeHuron*, R Core Team (2016). Figure 2.1 shows the results of the following five spectral density estimators applied to this time series: An AR-pre-whitened, nonparametric estimator of the spectral density,

Table 2.1: Coverage probabilities (in percent) for the mean using the studentized statistic of $\bar{X}_n(2\pi\hat{f}_n)^{-1/2}$ and for a sample size $n = 128$

	Model I			Model II			Model III		
$(1-\alpha)100$	80.0	90.0	95.0	80.0	90.0	95.0	80.0	90.0	95.0
SDDB	78.0	87.1	92.2	78.1	88.7	94.3	80.2	90.0	94.8
LPB	76.0	85.5	90.8	78.2	88.1	92.0	35.1	48.1	61.8
TBB	66.9	77.0	83.0	39.4	46.9	52.2	49.1	56.1	62.7
ND	67.8	78.2	84.6	64.1	76.4	84.2	24.2	32.4	40.0
ARS	76.6	85.9	91.1	74.2	85.5	92.3	39.5	51.8	62.4
BB	28.4	41.0	49.6	30.4	41.2	50.8	34.7	41.6	47.3

Table 2.2: Coverage probabilities (in percent) for the lag 2 autocorrelation using the studentized empirical autocorrelation at lag 2 and for a sample size $n = 128$

	Model I			Model II			Model III		
$(1-\alpha)100$	80.0	90.0	95.0	80.0	90.0	95.0	80.0	90.0	95.0
SDDB	82.5	91.5	96.0	79.4	89.3	93.5	81.0	90.3	95.3
LPB	85.5	94.2	97.2	92.0	95.9	97.2	82.8	91.9	96.7
TBB	76.3	83.9	87.7	20.6	25.1	27.6	65.4	72.6	77.2
ND	75.8	87.3	92.3	73.4	84.4	90.0	79.4	88.7	94.0
ARS	81.8	91.3	95.9	80.8	88.9	93.2	82.0	91.4	95.7
BB	32.6	45.0	57.0	21.4	31.2	42.8	27.9	40.0	52.2

denoted by *Pre-Whitening*, where the order of the autoregressive part has been selected by AIC and the truncation lag by cross-validation; a nonparametric spectral density estimator using cepstrum thresholding, denoted by *Cepstrum*, see Stoica and Sandgren (2006); a lag window estimator with a trapezoid kernel and the truncation rule as in Politis (2003), denoted by *Trapezoid* and an autoregressive parametric approach, where the order of the autoregressive part has been selected by AIC. Although, all estimators have a more or less similar overall behavior, they are different with the autoregressive based approaches possessing a stronger peak at frequency zero than the other. We next discuss the impact of these different estimators on the spectral-density-driven bootstrap. As mentioned in section 2.3.1, the spectral-density-driven bootstrap can be either used with the moving average or with the autoregressive representation of the process corresponding to the spectral density estimator applied. Table 2.3 shows for each estimator the obtained MA-coefficients and AR-coefficients, respectively. As it is seen, depending on the spectral density estimator used, the moving average or the autoregressive representation describes the structure of the process more effectively, i.e., less non-zero coefficients are needed. Clearly, the differences between the spectral density estimators used manifest themselves in the moving average or the autoregressive coefficients obtained. Notice that the oscillation of the *Trapezoid* spectral density estimator can be also

Table 2.3: The moving average $(c_{k,n})$ and autoregressive $(b_{k,n})$ coefficients , $k = 2, \ldots, 11$ for the different spectral density estimates shown in Figure 2.1

$c_{k,n}$	2	3	4	5	6	7	8	9	10	11
autoregressive	1.05	0.84	0.61	0.42	0.28	0.18	0.12	0.07	0.05	0.03
Pre-Whitening	1.07	0.85	0.61	0.42	0.28	0.18	0.12	0.07	0.05	0.03
Cepstrum	0.93	0.43	0.14	0.03	0.01	0	0	0	0	0
Trapezoid	0.33	0.23	0.16	0.12	0.11	0.09	0.08	0.09	0.12	0.09
$b_{k,n}$	2	3	4	5	6	7	8	9	10	11
autoregressive	1.05	-0.27	0	0	0	0	0	0	0	0
Pre-Whitening	1.07	-0.29	0.01	0	0	0	0	0	0	0
Cepstrum	0.93	-0.43	0.14	-0.03	0.01	0	0	0	0	0
Trapezoid	0.33	0.12	0.05	0.03	0.03	0.02	0.02	0.03	0.05	0.01

seen in the behavior of the moving average coefficients and that this estimator implicitly fits an moving average model to the time series at hand.

As this example demonstrates, the broad literature to spectral density estimation offers a variety of techniques to estimate this function. Therefore, the model used to generate the bootstrap pseudo observations depends on the spectral density estimator used and which is preferred by the practitioner. The resulting moving average or autoregressive representation can then be applied to generate pseudo observations in order to bootstrap some statistic of interest.

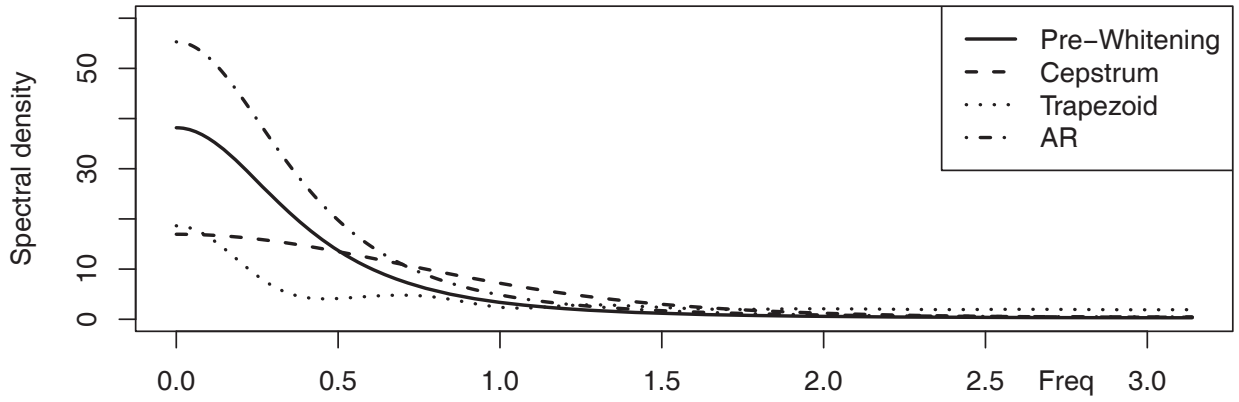


Figure 2.1: Different spectral density estimates for the Lake Huron data

2.5 Conclusions

In this work a spectral density factorization has been used to obtain consistent estimates of the entire sequence of moving average coefficients of the Wold representation of a stationary nondeterministic process. A bootstrap procedure then has been proposed which uses the estimated sequence of moving average coefficients together with a sequence of pseudo innovations to generate new pseudo time series. Apart for the choice of the pseudo innovations, this bootstrap procedure is completely driven and controlled by the spectral density estimator used. For i.i.d. pseudo innovations the new bootstrap method generalizes existing linear bootstrap methods, like for instance, the AR-sieve bootstrap. The latter is a special case of the spectral-density-driven bootstrap, which is obtained if an autoregressive spectral density estimator is used. We established asymptotic validity of the proposed bootstrap method driven by i.i.d. pseudo innovations for linear processes and for interesting classes of statistics. The good finite sample behavior of the new bootstrap method has been demonstrated by means of simulations.

2.6 Proofs

Proof of Theorem 2.2.1. a) Since $\hat{f}_n$ is a uniformly consistent estimator, it follows that some function g, with $g(n) \to \infty$ as $n \to \infty$, exists such that

$$\sup_{\lambda \in [-\pi,\pi]} |\hat{f}_n(\lambda) - f(\lambda)| = o_P(1) = O_P\left(g(n)^{-1}\right).$$

Consequently, we have that for all $\varepsilon > 0$, there exists a $\Omega_0 \in \mathcal{A}$ with $P(\Omega_0) \geq 1 - \varepsilon$ and a $n_0 \in \mathbb{N}$, such that for all $\omega \in \Omega_0$ a constant $C > 1$ exits, such that for all $n \geq n_0$ it holds true that $\sup_{\lambda \in [-\pi,\pi]} |\hat{f}_n(\lambda) - f(\lambda)| \leq Cg(n)^{-1}$. Since $\log f$ and $\log \hat{f}_n$ are integrable, and the set $\{\lambda \in [-\pi, \pi) : f(\lambda) = 0 \text{ or } \hat{f}_n(\lambda) = 0\} =: B_0^C$ is a null set, we have

$$\begin{aligned}\hat{a}_{k,n} =& \frac{1}{2\pi}\int_{-\pi}^{\pi} \log(\hat{f}_n(\lambda)) \exp(-ik\lambda) d\lambda = \int_{-\pi}^{\pi} \left(\log(\hat{f}_n(\lambda)) + \log f(\lambda) - \log f(\lambda)\right) \exp(-ik\lambda)\frac{\mathrm{d}\lambda}{2\pi} \\ =& a_k + \int_{-\pi}^{\pi} \left[\log(\hat{f}_n\lambda) - \log f(\lambda)\right] \exp(-ik\lambda)\frac{\mathrm{d}\lambda}{2\pi} = a_k + \int_{B_0} \left[\log(\hat{f}_n\lambda) - \log f(\lambda)\right] \exp(-ik\lambda)\frac{\mathrm{d}\lambda}{2\pi}.\end{aligned}$$

We get further

$$\begin{aligned}2\pi \sup_{k \in \mathbb{N}} |\hat{a}_{k,n} - a_k| &= \sup_{k \in \mathbb{N}} \left| \int_{B_0} \left(\log(\hat{f}_n(\lambda)) - \log f(\lambda)\right) \exp(-ik\lambda) \mathrm{d}\lambda \right| \leq \int_{B_0} |\log(\hat{f}_n(\lambda)) - \log(f(\lambda))| \mathrm{d}\lambda \\ &= \int_{B_0} \mathbb{1}_{\{\hat{f}_n(\lambda) > f(\lambda)\}}(\lambda) \log\left(\hat{f}_n(\lambda)/f(\lambda)\right) d\lambda + \int_{B_0} \mathbb{1}_{\{f(\lambda) > \hat{f}_n(\lambda)\}}(\lambda) \log\left(f(\lambda)/\hat{f}_n(\lambda)\right) d\lambda.\end{aligned}$$

Let $\omega \in \Omega_0$ and consider the case $\hat{f}_n(\lambda) > f(\lambda) > 0$. Then we have $\hat{f}_n(\lambda) \leq Cg(n)^{-1} + f(\lambda)$. Consequently,

$$\begin{aligned}&\int_{B_0} \mathbb{1}_{\{\hat{f}_n(\lambda) > f(\lambda) > 0\}}(\lambda) \log\left(\hat{f}_n(\lambda)/f(\lambda)\right) d\lambda \leq \int_{B_0 \cup \{\hat{f}_n > f\}} \log\left(\frac{C/g(n) + f(\lambda)}{f(\lambda)}\right) d\lambda \\ &= \int_{B_0 \cup \{\hat{f}_n > f\}} \log\left(C/(g(n)f(\lambda)) + 1\right) d\lambda. = \int_{B_0 \cup \{\hat{f}_n > f\}} \log(f(\lambda) + C/g(n)) - \log(f\lambda) d\lambda.\end{aligned}$$

Assume $g(n) > 1$. Then it holds true for all $n \in \mathbb{N}$ and all $\lambda \in B_0$ that

$$\begin{aligned}|\log\left(C(g(n)f(\lambda))^{-1} + 1\right)| &\leq |\log\left(C/f(\lambda) + 1\right)| = \log\left(\frac{C + f(\lambda)}{f(\lambda)}\right) \\ &= |\log(C + f(\lambda)) - \log(f(\lambda))| \leq |\log(C + f(\lambda))| + |\log(f(\lambda))|.\end{aligned}$$

Since $\log f$ is integrable, $\log(C + f(\cdot))$ is integrable as well, and consequently, the dominated convergence theorem can be applied. We get

$$\lim_{n \to \infty} \int_{B_0 \cup \{\hat{f}_n > f\}} \log(f(\lambda) + C/g(n)) - \log f(\lambda) d\lambda = 0.$$

Analogously, it can be shown that $\lim_{n\to\infty} \int_{B_0\cup\{\hat{f}_n<f\}} \log\left(f(\lambda)/\hat{f}_n(\lambda)\right) d\lambda = 0$. Thus, we have for all $\omega \in \Omega_0$ that $\sup_{k\in\mathbb{N}} |1/(2\pi) \int_{-\pi}^{\pi} \log(\hat{f}_n(\lambda)) - \log f(\lambda) \exp(-ik\lambda) d\lambda| \to 0$ as $n \to \infty$. This proves assertion a). For b) and c) fix a $k \in \mathbb{N}$ and observe that $\hat{c}_{k,n}$ is a continuous transformation of a finite number of $\hat{a}_{k,n}$'s. Thus, $\sup_{k\in\mathbb{N}} |\hat{a}_{k,n} - a_k| = o_P(1)$ ensures, that $\hat{c}_{k,n} \xrightarrow{P} c_k$ as $n \to \infty$. The same arguments apply to $\hat{b}_{k,n}$. □

Before proving Theorem 2.2.2, we notice the following useful lemma.

Lemma 2.6.1. *Let the condition of part b) of Theorem 2.2.2 be satisfied. Then it holds true that*

a)

$$\sup_{\lambda\in[-\pi,\pi]} |f^{1/2}(\lambda) - \hat{f}_n^{\,1/2}(\lambda)| = o_P(1),$$

b)

$$\sup_{\lambda\in[-\pi,\pi]} |\log(f(\lambda)) - \log(\hat{f}_n(\lambda))| = o_P(1),$$

c)

$$\sup_{\lambda\in[-\pi,\pi]} \left| \frac{\mathrm{d}}{\mathrm{d}\lambda} \log(f(\lambda)) - \frac{\mathrm{d}}{\mathrm{d}\lambda} \log\left(\hat{f}_n(\lambda)\right) \right| = o_P(1).$$

Proof. Since (2.2) ensures that f and $\hat{f}_n$ are bounded away from zero, *a)* and *b)* follow immediately from the mean value theorem. To see why the third assertion is true, one considers the following bound

$$\begin{aligned}
&\sup_{\lambda\in[-\pi,\pi]} \left| \frac{\mathrm{d}}{\mathrm{d}\lambda} \log(f(\lambda)) - \frac{\mathrm{d}}{\mathrm{d}\lambda} \log\left(\hat{f}_n(\lambda)\right) \right| \\
&\leq \sup_{\lambda\in[-\pi,\pi]} \left| \frac{\mathrm{d}}{\mathrm{d}\lambda} f(\lambda) - \frac{\mathrm{d}}{\mathrm{d}\lambda} \hat{f}_n(\lambda) \right| / f(\lambda) + \sup_{\lambda\in[-\pi,\pi]} \frac{\mathrm{d}}{\mathrm{d}\lambda} f(\lambda) \frac{|\hat{f}_n(\lambda) - f(\lambda)|}{f(\lambda)\hat{f}_n(\lambda)}.
\end{aligned}$$

Since f and $\hat{f}_n$ are bounded away from zero and $\frac{\mathrm{d}}{\mathrm{d}\lambda} f$ is finite, *c)* follows from (2.8) and (2.9). □

Proof of Theorem 2.2.2. Since (a_k) and $(\hat{a}_{k,n})$ are the Fourier coefficients of $\log f$ and $\log \hat{f}_n$ respectively, we have

$$\int_0^{2\pi} |\log f(\lambda) - \log \hat{f}_n(\lambda)|^2 \mathrm{d}\lambda = \int_0^{2\pi} |\sum_{k=-\infty}^{\infty} (a_k - \hat{a}_{k,n}) \exp(ik\lambda)|^2 \mathrm{d}\lambda,$$

and consequently (2.2.11) follows from Parseval's identity. (2.2.13) follows by the same argument.

Using Jensen's inequality and since $\sum_{k=1}^{\infty} k^{-2} = \pi^2/6$, we get

$$\sum_{k=1}^{\infty} |a_k - \hat{a}_{k,n}| = \frac{\pi^2}{6} \left(\frac{6}{\pi^2} \sum_{k=1}^{\infty} \frac{|a_k - \hat{a}_{k,n}| k^2}{k^2} \right)^{2/2} \leq \frac{\pi^2}{6} \left(\frac{6}{\pi^2} \sum_{k=1}^{\infty} \frac{|a_k - \hat{a}_{k,n}|^2 k^4}{k^2} \right)^{1/2}$$
$$\leq \frac{\pi}{\sqrt{12}} \left(\sum_{k=-\infty}^{\infty} |a_k - \hat{a}_{k,n}|^2 k^2 \right)^{1/2} \leq \sqrt{\pi/24} \left(\int_0^{2\pi} |\frac{\mathrm{d}}{\mathrm{d}\lambda} \log f(\lambda) - \frac{\mathrm{d}}{\mathrm{d}\lambda} \log \hat{f}_n(\lambda)|^2 \mathrm{d}\lambda \right)^{1/2}.$$

Lemma 2.6.1 implies then, that the above bound converges to zero in probability as $n \to \infty$. Let $\mathrm{sgn}(k) = \mathbb{1}_{\{k>0\}} - \mathbb{1}_{\{k<0\}}$. Since Lemma A.2 (or see Pourahmadi (1984)) ensures that for all $\lambda \in [0, 2\pi]$ it holds true that $\sigma/\sqrt{2\pi} \sum_{k=0}^{\infty} c_k \exp(ik\lambda) = \exp(a_0/2 + \sum_{k=1}^{\infty} a_k \exp(ik\lambda))$ and similarly for $\{\hat{c}_{k,n}\}$ with a_k replaced by $\hat{a}_{k,n}$, we get by Parseval's identity, the fact that $\cos(x) \geq 1 - 0.5x^2$ for all $x \in \mathbb{R}$ and $\int_0^{2\pi} \sin(k\lambda)\sin(l\lambda)\mathrm{d}\lambda = \pi \mathbb{1}_{\{k=l\}}$ for all $k, l \in \mathbb{N}$, that

$$\sum_{k=0}^{\infty} |\sigma c_k - \hat{\sigma}_n \hat{c}_{k,n}|^2 = \int_0^{2\pi} \left| \sum_{k=0}^{\infty} (\sigma c_k - \hat{\sigma}_n \hat{c}_{k,n}) \exp(ik\lambda) \right|^2 \mathrm{d}\lambda/(2\pi)$$
$$= \int_0^{2\pi} \left| \exp[a_0/2 + \sum_{k=1}^{\infty} a_k \exp(ik\lambda)] - \exp[a_{0,n}/2 + \sum_{k=1}^{\infty} \hat{a}_{k,n} \exp(ik\lambda)] \right|^2 \mathrm{d}\lambda$$
$$= \int_0^{2\pi} \left| \exp\left[1/2 \sum_{k=-\infty}^{\infty} a_k \exp(ik\lambda) + 1/2 \sum_{k=-\infty}^{\infty} \mathrm{sgn}(k) a_k \exp(ik\lambda) \right] \right.$$
$$\left. - \exp\left[1/2 \sum_{k=-\infty}^{\infty} \hat{a}_{k,n} \exp(ik\lambda) + 1/2 \sum_{k=-\infty}^{\infty} \mathrm{sgn}(k) \hat{a}_{k,n} \exp(ik\lambda) \right] \right|^2 \mathrm{d}\lambda$$
$$= \int_0^{2\pi} \left| f^{1/2}(\lambda) \exp\left[i \sum_{k=1}^{\infty} a_k \sin(k\lambda) \right] - \hat{f}_n^{\,1/2}(\lambda) \exp\left[i \sum_{k=1}^{\infty} \hat{a}_{k,n} \sin(k\lambda) \right] \right|^2 \mathrm{d}\lambda$$
$$= \int_0^{2\pi} \left| \exp\left[i \sum_{k=1}^{\infty} \hat{a}_{k,n} \sin(k\lambda) \right] \right|^2 \left| f^{1/2}(\lambda) \exp\left[i \sum_{k=1}^{\infty} (a_k - \hat{a}_{k,n}) \sin(k\lambda) \right] - \hat{f}_n^{\,1/2}(\lambda) \right|^2 \mathrm{d}\lambda$$
$$= \int_0^{2\pi} \left(f(\lambda) + \hat{f}_n(\lambda) - 2 f^{1/2}(\lambda) \hat{f}_n^{\,1/2}(\lambda) \cos\left[\sum_{k=1}^{\infty} (a_k - \hat{a}_{k,n}) \sin(k\lambda) \right] \right) \mathrm{d}\lambda$$
$$\leq \int_0^{2\pi} \left(f^{1/2}(\lambda) - \hat{f}_n^{\,1/2}(\lambda) \right)^2 \mathrm{d}\lambda + \int_0^{2\pi} (f(\lambda)\hat{f}_n(\lambda))^{1/2} \left| \sum_{k=1}^{\infty} (a_k - \hat{a}_{k,n}) \sin(k\lambda) \right|^2 \mathrm{d}\lambda$$
$$\leq \int_0^{2\pi} \left(f^{1/2}(\lambda) - \hat{f}_n^{\,1/2}(\lambda) \right)^2 \mathrm{d}\lambda + \sup_{\lambda \in [0,2\pi]} (f(\lambda)\hat{f}_n(\lambda))^{1/2} \int_0^{2\pi} (\log f(\lambda) - \log \hat{f}_n(\lambda))^2 \mathrm{d}\lambda = \mathrm{o}_P(1)$$

where the last equation follows by (2.2.11) and Assumption 2. Assertion (2.2.12) follows since $\hat{\sigma}_n \xrightarrow{P} \sigma$ and

$$\sum_{k=0}^{\infty} |c_k - \hat{c}_{k,n}|^2 \leq 2/\sigma \sum_{k=0}^{\infty} |\sigma c_k - \hat{\sigma}_n \hat{c}_{k,n}|^2 + 2/\sigma \sum_{k=0}^{\infty} |\hat{c}_{k,n}|^2 |\hat{\sigma}_n - \sigma|^2 = \mathrm{o}_P(1).$$

By Jensen's inequality, we have

$$\begin{aligned}
&\sum_{k=1}^{\infty} |\sigma c_k - \hat{\sigma}_n \hat{c}_{k,n}| \leq \frac{\pi^2}{6} \left(\frac{6}{\pi^2} \sum_{k=1}^{\infty} |\sigma c_k - \hat{\sigma}_n \hat{c}_{k,n}|^2 k^2 \right)^{1/2} \\
&= \frac{\pi^2}{6} \left(\frac{6}{\pi^2} \int_0^{2\pi} \left| \sum_{k=0}^{\infty} (\sigma c_k - \hat{\sigma}_n \hat{c}_{k,n}) k \exp(ik\lambda) \right|^2 \mathrm{d}\lambda/(2\pi) \right)^{1/2} \\
&= \frac{\pi}{\sqrt{6}} \left(\int_0^{2\pi} \left| \frac{\mathrm{d}}{\mathrm{d}\lambda} \sum_{k=0}^{\infty} (\sigma c_k - \hat{\sigma}_n \hat{c}_{k,n}) \exp(ik\lambda) \right|^2 \mathrm{d}\lambda \right)^{1/2} \\
&= \frac{\pi}{\sqrt{6}} \left(\int_0^{2\pi} \left| \frac{\mathrm{d}}{\mathrm{d}\lambda} \left(\exp\left[\frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \exp(ik\lambda) \right] - \exp\left[\frac{\hat{a}_{0,n}}{2} + \sum_{k=1}^{\infty} \hat{a}_{k,n} \exp(ik\lambda) \right] \right) \right|^2 \mathrm{d}\lambda \right)^{1/2} \\
&= \frac{\pi^2}{6} \left(\frac{6}{\pi^2} \int_0^{2\pi} \left| \frac{\mathrm{d}}{\mathrm{d}\lambda} \left(\exp[a_0/2 + \sum_{k=1}^{\infty} a_k \exp(ik\lambda)] \times \right. \right. \right. \\
&\qquad \left. \left. \left. \left[1 - \exp[a_{0,n}/2 - a_k/2 + \sum_{k=1}^{\infty} (\hat{a}_{k,n} - a_k) \exp(ik\lambda)] \right] \right) \right|^2 \mathrm{d}\lambda \right)^{1/2} \\
&= \frac{\pi}{\sqrt{6}} \left(\int_0^{2\pi} \left| \exp\left[\frac{\hat{a}_{0,n}}{2} + \sum_{k=1}^{\infty} (\hat{a}_{k,n}) \exp(ik\lambda) \right] \sum_{k=1}^{\infty} (\hat{a}_{k,n} - a_k)(ki) \exp(ik\lambda) - \right. \right. \\
&\left. \left. \sum_{k=1}^{\infty} (ki) a_k \exp(ik\lambda) \left(\exp\left[\frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \exp(ik\lambda) \right] - \exp\left[\frac{\hat{a}_{0,n}}{2} + \sum_{k=1}^{\infty} (\hat{a}_{k,n}) \exp(ik\lambda) \right] \right) \right|^2 \mathrm{d}\lambda \right)^{1/2}.
\end{aligned}$$

The term on the right hand side of the last equation can be bounded by

$$\begin{aligned}
&\frac{\pi}{\sqrt{3}} \left(\int_0^{2\pi} \left[\hat{f}_n(\lambda) \left| \sum_{k=1}^{\infty} (\hat{a}_{k,n} - a_k)(ki) \exp(ik\lambda) \right|^2 + \left| \sum_{k=1}^{\infty} (ki) a_k \exp(ik\lambda) \right|^2 \right. \right. \\
&\qquad \left. \left. \times \left| \exp\left[a_0/2 + \sum_{k=1}^{\infty} a_k \exp(ik\lambda) \right] - \exp\left[a_{0,n}/2 + \sum_{k=1}^{\infty} (\hat{a}_{k,n}) \exp(ik\lambda) \right] \right|^2 \right] \mathrm{d}\lambda \right)^{1/2}.
\end{aligned}$$

Furthermore, the second part of the last term can be bounded analogously as in the case for (2.2.12) and because of Assumption 2, we get

$$\sum_{k=1}^{\infty} |\sigma c_k - \hat{\sigma}_n \hat{c}_{k,n}| \leq \frac{C\pi}{\sqrt{3}} \sup_{\lambda\in[0,2\pi]} \left|\sum_{k=1}^{\infty} a_k k \exp(ik\lambda)\right| \left(\frac{1}{4}\int_0^{2\pi} \left| \frac{\mathrm{d}}{\mathrm{d}\lambda} \log f(\lambda) - \frac{\mathrm{d}}{\mathrm{d}\lambda} \log \hat{f}_n(\lambda) \right|^2 \mathrm{d}\lambda \right.$$
$$\left. + \sum_{k=1}^{\infty} (\hat{a}_{k,n} - a_k)^2 k^2 \int_0^{2\pi} \left| f^{1/2}(\lambda) - \hat{f}_n^{\,1/2}(\lambda) \right|^2 \mathrm{d}\lambda \right)^{1/2}$$
$$= \left(\mathcal{O}_P(1) \int_0^{2\pi} \left| \frac{\mathrm{d}}{\mathrm{d}\lambda} \log f(\lambda) - \frac{\mathrm{d}}{\mathrm{d}\lambda} \log \hat{f}_n(\lambda) \right|^2 \mathrm{d}\lambda + \mathcal{O}_P(1) \int_0^{2\pi} \left| f^{1/2}(\lambda) - \hat{f}_n^{\,1/2}(\lambda) \right|^2 \mathrm{d}\lambda \right)^{1/2}.$$

The convergence to zero in probability, as $n \to \infty$, of the term on the right hand side of the last equality follows from (2.2.13) and the fact that $\hat{\sigma}_n \xrightarrow{P} \sigma$. □

Proof of Theorem 2.3.1. For simplicity, the proof is stated for centered X_t^*'s and a single $\hat{T}_{n,p}$ and in order to simplify notation the subscript p is in the following omitted. Since g is a smooth function, the delta-method can be applied to get the asymptotic normality of $\hat{T}_n^*$. Consider the statistic

$$T_n^* = \frac{1}{n} \sum_{t=1}^{n} \sum_{h=1-t}^{n-t} d(h) X_t^* X_{t+h}^*.$$

We firstly show that, as $n \to \infty$, $\sqrt{n}(T_n^{*M} - E^* T_n^{*M}) \to \mathcal{N}(0, \sigma_M^2)$, in probability, where $T_n^{*M} = \frac{1}{n}\sum_{t=1}^{n} \sum_{h=1-t}^{M\wedge(n-t)} d(h) X_t^* X_{t+h}^*$. For this, we use a central limit theorem for triangular arrays of weakly dependent random variables established by Neumann (2013). Let

$$\sqrt{n}(T_n^{*M} - E^* T_n^{*M}) = \sum_{t=1}^{n} \frac{1}{\sqrt{n}} \sum_{h=1-t}^{M\wedge(n-t)} d(h)(X_t^* X_{t+h}^* - \widehat{\widehat{\gamma}}_n(h)) = \sum_{t=1}^{n} Z_{t,n}^*,$$

with an obvious notation for $Z_{t,n}^*$ and $\widehat{\widehat{\gamma}}_n(h) = 1/(2\pi)\int_{-\pi}^{\pi} \hat{f}_n(\lambda)\exp(-ih\lambda)\mathrm{d}\lambda$. For the mean of $Z_{t,n}^*$, we have $E^* Z_{t,n}^* = 0$ for all $t \in \mathbb{Z}$. Furthermore, since $\{X_t^*, t \in \mathbb{Z}\}$ is a linear process, its strictly stationarity can be used to show that

$$\sum_{t=1}^{n} E^*(Z_{t,n}^*)^2 = \sum_{t=1}^{n} \frac{1}{n} \sum_{h_1=1-t}^{M\wedge(n-t)} d(h_1) \sum_{h_2=1-t}^{M\wedge(n-t)} d(h_2) E^* \Big[\Big((X_t^*)^2 X_{t+h_1}^* X_{t+h_2}^* - \widehat{\widehat{\gamma}}_n(h_1)\widehat{\widehat{\gamma}}_n(h_2) \Big) \Big]$$
$$\leq \sum_{t=1}^{n} \frac{1}{n} \sum_{h_1=1-t}^{M\wedge(n-t)} |d(h_1)| \sum_{h_2=1-t}^{M\wedge(n-t)} |d(h_2)| E^* \left(2(X_t^*)^4 + 4(X_{t+h_1}^*)^4 + 4(X_{t+h_2}^*)^4 + \widehat{\widehat{\gamma}}_n(h_1)\widehat{\widehat{\gamma}}_n(h_2) \right)$$
$$= \sum_{t=1}^{n} \frac{1}{n} \sum_{h_1=1-t}^{M\wedge(n-t)} |d(h_1)| \sum_{h_2=1-t}^{M\wedge(n-t)} |d(h_2)| \left(10 E^*(X_1^*)^4 + \widehat{\widehat{\gamma}}_n(h_1)\widehat{\widehat{\gamma}}_n(h_2) \right)$$
$$\leq (E^*(X_1^*)^4 + \widehat{\widehat{\gamma}}_n(0)^2)(\sum_{h\in Z} |d(h)|)^2 \leq C < \infty,$$

where C is independent of n, since $\hat{\kappa}_{4,n}$ and $\widehat{\widehat{\gamma}}_n$ are uniformly bounded and so is $E^*(X_1^*)^4$.

Consider the weak dependence structure of $Z^*_{t,n}$. For this, let $u \in \mathbb{N}$ and consider time points $1 < s_1 < \cdots < s_u < s_u + r = t_1 \leq t_2 \leq n$, a square integrable and measurable function $f : \mathbb{R}^u \to \mathbb{R}$ and a bounded and measurable function $\tilde{f} : \mathbb{R}^u \to \mathbb{R}$. Without loss of generality, we assume, that $r > M$. Then, we have

$$|\sqrt{n}\mathrm{Cov}(f(Z^*_{s_1,n},\ldots,Z^*_{s_u,n}),Z^*_{t_1,n})| = \left|\mathrm{Cov}\left(f(Z^*_{s_1,n},\ldots,Z^*_{s_u,n}),\sum_{h=1-s_u-r}^{M\wedge(n-(s_u+r))} d(h)(X^*_{s_u+r}X^*_{s_u+r+h})\right)\right|$$
$$= \left|\mathrm{Cov}\left(f(Z^*_{s_1,n},\ldots,Z^*_{s_u,n}),\sum_{h=1-s_u-r}^{M\wedge(n-(s_u+r))} d(h)\sum_{j=0}^{\infty}\hat{c}_{j,n}\sum_{l=0}^{\infty}\hat{c}_{l,n}\varepsilon^*_{s_u+r-j}\varepsilon^*_{s_u+r+h-l}\right)\right|.$$

Since $\{X^*_t\}$ is an one-sided linear process, we have $f(Z^*_{s_1,n},\ldots,Z^*_{s_u,n}) = f(g(\varepsilon^*_{s_u+M},\varepsilon^*_{s_u+M-1},\ldots))$ for some measurable function g. Consequently, by the independence of the ε^*_t's and applying Cauchy-Schwarz's inequality, it follows for the last expression above that it can be bounded by

$$n^{-1/2}\sum_{h=1-s_u-r}^{M\wedge(n-(s_u+r))}|d(h)|\sum_{j=r-M}^{\infty}|\hat{c}_{j,n}|\sum_{l=r+h-M}^{\infty}|\hat{c}_{l,n}|(E^*f^2(Z^*_{s_1,n},\ldots,Z^*_{s_u,n}))^{1/2}\left[E^*(\varepsilon^*_{s_u+r-j}\varepsilon^*_{s_u+r+h-l})^2\right]^{1/2}$$
$$\leq (E^*f^2(Z^*_{s_1,n},\ldots,Z^*_{s_u,n}))^{1/2}\frac{1}{\sqrt{n}}\sum_{h\in\mathbb{Z}}|d(h)|\sum_{l=0}^{\infty}|\hat{c}_{l,n}|\max(\hat{\kappa}_{4,n},1)\sum_{j=r-M}^{\infty}|\hat{c}_{j,n}|$$
$$\leq (E^*f^2(Z^*_{s_1,n},\ldots,Z^*_{s_u,n}))^{1/2}\frac{1}{\sqrt{n}}C\sum_{j=r-M}^{\infty}|\hat{c}_{j,n}|,$$

where $C < \infty$ is independent of n since $\hat{\kappa}_{4,n}$ and $\sum_{j=0}^{\infty}\hat{c}_{j,n}$ are uniformly bounded. We have

$$|\mathrm{Cov}(\tilde{f}(Z^*_{s_1,n},\ldots,Z^*_{s_u,n}),Z^*_{t_1,n}Z^*_{t_2,n})| =$$
$$= \left|\sum_{h=1-s_u-r}^{M\wedge(n-(s_u+r))} d(h)\sum_{j=r-M}^{\infty}\hat{c}_{j,n}\sum_{l=r+h-M}^{\infty}\hat{c}_{l,n}\mathrm{Cov}(\tilde{f}(Z^*_{s_1,n},\ldots,Z^*_{s_u,n}),\frac{1}{\sqrt{n}}\varepsilon^*_{s_u+r-j}\varepsilon^*_{s_u+r+h-l}Z^*_{t_2,n})\right|$$
$$\leq \sum_{h=1-s_u-r}^{M\wedge(n-(s_u+r))}|d(h)|\sum_{j=r-M}^{\infty}|\hat{c}_{j,n}|\sum_{l=r+h-M}^{\infty}|\hat{c}_{l,n}|2\|\tilde{f}\|_\infty E^*(\frac{1}{\sqrt{n}}|\varepsilon^*_{s_u+r-j}\varepsilon^*_{s_u+r+h-l}Z^*_{t_2,n}|)$$
$$\leq \sum_{h=1-s_u-r}^{M\wedge(n-(s_u+r))}|d(h)|\sum_{j=r-M}^{\infty}|\hat{c}_{j,n}|\sum_{l=r+h-M}^{\infty}|\hat{c}_{l,n}|\|\tilde{f}\|_\infty\frac{2}{n}E^*\left[\left(\varepsilon^*_{s_u+r-j}\varepsilon^*_{s_u+r+h-l}\right)^2+(Z^*_{t_2,n})^2\right]$$
$$\leq \|\tilde{f}\|_\infty\frac{2}{n}(\max(\hat{\kappa}_{4,n},\hat{\sigma}_n^2)+E^*(Z^*_{t_2,n})^2)\sum_{h\in\mathbb{Z}}|d(h)|\sum_{l=0}^{\infty}|\hat{c}_{l,n}|\sum_{j=r-M}^{\infty}|\hat{c}_{j,n}| \leq \|\tilde{f}\|_\infty\frac{C}{n}\sum_{j=r-M}^{\infty}|\hat{c}_{j,n}|.$$

Consequently, the sequence $\{Z^*_{t,n}\}$ fulfills the weakly dependence condition of Neumann (2013), if $2C\sum_{j=r-M}^{\infty}|\hat{c}_{j,n}| \leq \theta_r$ for some summable $(\theta_r)_{r\in\mathbb{N}}$. Since $\sup_{\lambda\in(-\pi,\pi]}(\frac{\mathrm{d}}{\mathrm{d}\lambda})^3\log\hat{f}_n(\lambda) \leq C$ holds independently of n, it follows similarly to Lemma A.3 that $\sup_{j,n}|\hat{c}_{j,n}j^3| \leq C$. Hence, $2C\sum_{j=r-M}^{\infty}|\hat{c}_{j,n}| \leq \sum_{j=r-M}^{\infty}Cj^{-3} =: \theta_r$ for all $r > M$ and for some $C > 0$. If $r \leq M$ we set $\theta_r := C$. Then it holds

$\sum_{r=0}^{\infty}\theta_r = C(M+1+\sum_{r=M+1}^{\infty}\sum_{j=r-M}^{\infty} j^{-3}) = C(M+1+\sum_{j=1}^{\infty} j^{-2}) < \infty$. Regarding the variance of T_n^{*M}, consider firstly $T_n^{*M} = 1/\sqrt{n}\sum_{h=-n+1}^{M}\sum_{t=1\vee(1-h)}^{n\wedge(n-h)} d(h)X_t^*X_{t+h}^*$. Using the linear process structure of $\{X_t^*\}$ and additionally, let $c_j = 0$ for all $j < 0$, we get

$$
\begin{aligned}
\mathrm{Var}(T_n^{*M}) &= \sum_{h_1,h_2=-n+1}^{M} 1/n \sum_{t=1\vee(1-h_1)}^{n\wedge(n-h_1)} d(h_1) \sum_{s=1\vee(1-h_2)}^{n\wedge(n-h_2)} d(h_2)\Big(E^*(X_t^*X_{t+h_1}^*X_s^*X_{s+h_2}^*) - \widehat{\widehat{\gamma}}_n(h_1)\widehat{\widehat{\gamma}}_n(h_2)\Big)\\
&= \sum_{h_1,h_2=-n+1}^{M} 1/n \sum_{t=1\vee(1-h_1)}^{n\wedge(n-h_1)} d(h_1) \sum_{s=1\vee(1-h_2)}^{n\wedge(n-h_2)} d(h_2)\Big(\sum_{j=0}^{\infty}\hat{c}_{j,n}\hat{c}_{j+h_1,n}\hat{c}_{j+s-t,n}\hat{c}_{j+s+h_2-t,n}(\hat{\kappa}_{4,n}-3)\\
&\qquad + \widehat{\widehat{\gamma}}_n(s-t)\widehat{\widehat{\gamma}}_n(s-t+h_2-h_1) + \widehat{\widehat{\gamma}}_n(s+h_2-t)\widehat{\widehat{\gamma}}_n(s-t-h_1)\Big)\\
&= \frac{1}{n}\sum_{h_1=1}^{M}\sum_{h_2=1}^{M}\sum_{k\in\mathbb{Z}} d(h_1)d(h_2)\Big(\sum_{j=0}^{\infty}\hat{c}_{j,n}\hat{c}_{j+h_1,n}\hat{c}_{j+k,n}\hat{c}_{j+k+h_2,n}(\hat{\kappa}_{4,n}-3)\\
&\quad + \widehat{\widehat{\gamma}}_n(k)\widehat{\widehat{\gamma}}_n(k+h_2-h_1) + \widehat{\widehat{\gamma}}_n(k+h_2)\widehat{\widehat{\gamma}}_n(k-h_1)\Big)\sum_{t=1\vee(1-h_1)}^{n\wedge(n-h_1)}\sum_{s=1\vee(1-h_2)}^{n\wedge(n-h_2)}\mathbb{1}_{\{k=s-t\}}\\
&= \sum_{h_1,h_2=-n+1}^{M}\sum_{k=-(n-1)}^{n-1} d(h_1)d(h_2)\Big(\sum_{j=0}^{\infty}\hat{c}_{j,n}\hat{c}_{j+h_1,n}\hat{c}_{j+k,n}\hat{c}_{j+k+h_2,n}(\hat{\kappa}_{4,n}-3)\\
&\qquad + \widehat{\widehat{\gamma}}_n(k)\widehat{\widehat{\gamma}}_n(k+h_1-h_2) + \widehat{\widehat{\gamma}}_n(k+h_2)\widehat{\widehat{\gamma}}_n(k-h_1)\Big)\frac{\max(0,n-(|k|+|h_1-h_2|))}{n}.
\end{aligned}
$$

Since $\hat{\kappa}_{4,n}$ is a consistent estimator of κ_4 and $\sum_{j=0}^{\infty}|c_j-\hat{c}_{j,n}| = o_P(1)$ from which it follows that $\sum_{k\in\mathbb{Z}}\widehat{\widehat{\gamma}}_n(k)\widehat{\widehat{\gamma}}_n(k+x) = \sum_{k\in\mathbb{Z}}\gamma(k)\gamma(k+x) + o_P(1)$, we have that the last term is equal to

$$
\begin{aligned}
&\sum_{h_1=-n+1}^{M}\sum_{h_2=-n+1}^{M}\sum_{k=-(n-1)}^{n-1} d(h_1)d(h_2)\Big(\sum_{j=0}^{\infty}c_jc_{j+h_1}c_{j+k}c_{j+k+h_2}(\kappa_4-3)\\
&+\gamma(k)\gamma(k+h_1-h_2)+\gamma(k+h_2)\gamma(k-h_1)\Big)\frac{\max(0,n-(|k|+|h_1-h_2|))}{n} + o_P(1).
\end{aligned}
$$

We then have for the first term of the last equality above, that it equals, as $n\to\infty$, to

$$
\begin{aligned}
&\sum_{h_1=-n+1}^{M}\sum_{h_2=-n+1}^{M} d(h_1)d(h_2)\Big(\sum_{j=0}^{\infty}c_jc_{j+h_1}\sum_{k\in\mathbb{Z}}c_{j+k+h_1-h_2}c_{j+k+h_1}(\kappa_4-3)\\
&\qquad + \sum_{k\in\mathbb{Z}}\gamma(k+h_1-h_2)\gamma(k) + \gamma(k+h_1)\gamma(k-h_2)\Big)\\
&= \sum_{h_1,h_2=-n+1}^{M} d(h_1)d(h_2)\Big(\gamma(h_1)\gamma(-h_2)(\kappa_4/\sigma^4-3) + \sum_{k\in\mathbb{Z}}\gamma(k+h_1-h_2)\gamma(k) + \gamma(k+h_1)\gamma(k-h_2)\Big).
\end{aligned}
$$

Since $\sum_{j=0}^{\infty}|jc_j| < \infty$ and $\sum_{k\in\mathbb{Z}}|k\gamma(k)| < \infty$, the second term of the same equality is of order $\mathcal{O}(1/n)$. Hence, we have in probability, as $n\to\infty$,

$$
\mathrm{Var}(T_n^{*M}) \to \sum_{h_1,h_2=-\infty}^{M} d(h_1)d(h_2)\Big(\gamma(h_1)\gamma(-h_2)\Big(\frac{\kappa_4}{\sigma^4}-3\Big) + \sum_{k\in\mathbb{Z}}\gamma(k+h_1-h_2)\gamma(k) + \gamma(k+h_1)\gamma(k-h_2)\Big).
$$

Using the strictly stationarity of the process (X_t^*) and since $\sum_{k\in\mathbb{Z}}|d(h)|<\infty$ and $E(X_t^*)^4<\infty$, we can verify Lindberg's condition by means of Lebesgue's dominated convergence theorem, that is

$$\begin{aligned}
&\Big|1/n\sum_{t=1}^{n}\sum_{h_1,h_2=1-t}^{M\wedge(n-t)} d(h_1)d(h_2)E^*\Big((X_t^{*2}X_{t+h_1}^*X_{t+h_2}^*-\widehat{\widehat{\gamma}}_n(h_1)\widehat{\widehat{\gamma}}_n(h_2))\mathbb{1}_{\{|\sum_{h_3=1-t}^{M\wedge(n-t)} d(h_3)(X_t^*X_{t+h_3}^*-\widehat{\widehat{\gamma}}_n(h_3))|>\sqrt{n}\varepsilon\}}\Big)\Big| \\
&\le \frac{1}{n}\sum_{t=1}^{n}\sum_{h_1,h_2\in\mathbb{Z}}|d(h_1)d(h_2)|E^*\Big((|(X_t^*)^2X_{t+h_1}^*X_{t+h_2}^*|+|\widehat{\widehat{\gamma}}_n(h_1)\widehat{\widehat{\gamma}}_n(h_2)|)\mathbb{1}_{\{\sum_{h_3\in\mathbb{Z}}|d(h_3)|(|X_t^*X_{t+h_3}^*|+|\widehat{\widehat{\gamma}}_n(h_3)|)>\sqrt{n}\varepsilon\}}\Big) \\
&= \sum_{h_1,h_2\in\mathbb{Z}}|d(h_1)d(h_2)|E^*\Big((|(X_1^*)^2X_{1+h_1}^*X_{1+h_2}^*|+|\widehat{\widehat{\gamma}}_n(h_1)\widehat{\widehat{\gamma}}_n(h_2)|)\mathbb{1}_{\{\sum_{h_3\in\mathbb{Z}}|d(h_3)|(|X_1^*X_{1+h_3}^*|+|\widehat{\widehat{\gamma}}_n(h_3)|)>\sqrt{n}\varepsilon\}}\Big) \\
&\to 0, \text{as } n\to\infty.
\end{aligned}$$

Therefore, by the central limit theorem for triangular arrays of weakly dependent random variables given in Neumann (2013), we have that, as $n\to\infty$, $\sqrt{n}(T_n^{*M}-E^*T_n^{*M})\to\mathcal{N}(0,\sigma_M^2)$, in probability. Now, using a version in probability of Theorem 4.2 of Billingsley (1968) the proof of the theorem is concluded, since additionally to the converges for any fixed M, we have

$$\lim_{M\to\infty}\sigma_M^2=\sum_{h_1,h_2=-\infty}^{\infty}d(h_1)d(h_2)\Big(\gamma(h_1)\gamma(-h_2)(\frac{\kappa_4}{\sigma^4}-3)+\sum_{k\in\mathbb{Z}}\gamma(k+h_1-h_2)\gamma(k)+\gamma(k+h_1)\gamma(k-h_2)\Big).$$

Finally, condition (3) of Theorem 4.2 of Billingsley (1968) holds in probability since by Tchebysheff's inequality we have

$$\begin{aligned}
&\lim_{M\to\infty}\limsup_{n\to\infty}P\left(\left|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\sum_{h=1-t}^{M\wedge(n-t)}d(h)\left(X_t^*X_{t+h}^*-\widehat{\widehat{\gamma}}_n(h)\right)-\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\sum_{h=1-t}^{(n-t)}d(h)\left(X_t^*X_{t+h}^*-\widehat{\widehat{\gamma}}_n(h)\right)\right|>\varepsilon\right) \\
&\quad=\lim_{M\to\infty}\limsup_{n\to\infty}P\left(\left|\sum_{t=1}^{n}\sum_{h=M+1}^{n-t}d(h)\left(X_t^*X_{t+h}^*-\widehat{\widehat{\gamma}}_n(h)\right)\right|>\sqrt{n}\varepsilon\right) \\
&\quad\le\lim_{M\to\infty}\limsup_{n\to\infty}\mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\sum_{h=M+1}^{n-t}d(h)\left(X_t^*X_{t+h}^*\right)\right)/(\varepsilon^2) \\
&\quad=\lim_{M\to\infty}\limsup_{n\to\infty}\mathrm{Var}\left(\sum_{h=M+1}^{n}d(h)\sum_{t=1}^{n-h}\left(X_t^*X_{t+h}^*\right)\right)/(n\varepsilon^2),
\end{aligned}$$

together with similar calculations as in the evaluation of $\mathrm{Var}(T_n^{*M})$ and $\sum k\in\mathbb{Z}|\gamma(k)|<\infty$ we get $\lim_{M\to\infty}\sum_{h_1,h_2=M+1}^{\infty}d(h_1)d(h_2)O_P(1)=0$, by the fact that $\sum_{h\in\mathbb{Z}}|d(h)|<\infty$. Using this summability property and the same arguments as in the calculation of $\mathrm{Var}(T_n^{*M})$ leads to, as $n\to\infty$ and in probability,

$$\mathrm{Var}(T_n^*)\to\sum_{h_1,h_2=-\infty}^{\infty}d(h_1)d(h_2)\Big(\gamma(h_1)\gamma(-h_2)(\frac{\kappa_4}{\sigma^4}-3)+\sum_{k\in\mathbb{Z}}\gamma(k+h_1-h_2)\gamma(k)+\gamma(k+h_1)\gamma(k-h_2)\Big)$$

which can be written in frequency domain as

$$(\kappa_4/\sigma^4 - 3)(\int_0^{2\pi} f(\lambda) \sum_{h\in\mathbb{Z}} d(h) \exp(ih\lambda)\mathrm{d}\lambda)^2 + 4\pi \int_0^{2\pi} |f(\lambda) \sum_{h\in\mathbb{Z}} d(h) \exp(ih\lambda)|^2 \mathrm{d}\lambda.$$

Since the statistic in Definition 2.3.1 has for linear processes the same asymptotic distribution as the one derived above, the assertion of the theorem follows by the triangular inequality. □

Proof of Theorem 2.3.2. Without loss of generality, let X_t^* be centered. Since $\sum_{h\in\mathbb{Z}} |\gamma(h)| < \infty$, we have $\sup_{\lambda\in[-\pi,\pi]} f(\lambda) < \infty$. To proof Theorem 2.3.2 we use a central limit theorem for M-dependent sequences, see Romano and Wolf (2000) and a version in probability of Theorem 4.2 of Billingsley (1968).

Fix $M \in \mathbb{N}$ and consider the M-dependent process $X_{t,n,M}^* = \sum_{k=0}^M \hat{c}_{k,n}\varepsilon_{t-k}^*$, $t \in \mathbb{Z}$ with spectral density $\hat{f}_M(\lambda) = |\sum_{k=0}^M \hat{c}_{k,n} \exp(i\lambda k)|^2 \hat{\sigma}_n^2/(2\pi)$ and autocovariance $\hat{\gamma}_{n,M}(h) = \int_{-\pi}^{\pi} \hat{f}_M(\lambda) \exp(ih\lambda)\mathrm{d}\lambda$. Applying a central limit theorem for M-dependent sequences we show, as $n \to \infty$, that

$$\sum_{t=1}^n n^{-1/2} X_{t,n,M}^* \xrightarrow{D} \mathcal{N}(0, 2\pi f_M(0)), \text{in probability, where } f_M(0) = |\sum_{k=0}^M c_k|^2 \sigma^2/(2\pi).$$

Theorem 2.2.1 gives for the variance, let $n \geq M$,

$$n^{-1}\mathrm{Var}\left(\sum_{t=1}^n X_{t,n,M}^*\right) = \sum_{h=-M}^M (1 - |h|/n)\hat{\gamma}_M(h) = (2\pi)\hat{f}_M(0) + O_P(1/n) \to 2\pi f_M(0),$$

in probability, as $n \to \infty$. If $f_M(0) = 0$ the assertion follows. Assume that $f_M(0) > 0$. Since $\{X_{t,n,M}^*\}$ is stationary, we have for $k \geq M$ and all $a \in \mathbb{N}$

$$\mathrm{Var}\left(\sum_{t=a}^{a+k-1} n^{-1/2} X_{t,n,M}^*\right) = \frac{1}{n}\mathrm{Var}\left(\sum_{t=1}^k X_{t,n,M}^*\right) = 2\pi\hat{f}_M(0)k/n + O_P(1/n).$$

Furthermore, the process $\{X_{t,n,M}^*\}$ has i.i.d. innovations with a finite fourth moment. Thus, the fourth moment can easily be bounded; we have

$$E\left(n^{-1/2} X_{1,n,M}^*\right)^4 = \left(E((\varepsilon_1^*)^4 - 3\hat{\sigma}_n^4) \sum_{k=0}^M \hat{c}_{k,n}^4 + 3\hat{\gamma}_M(0)^2\right) n^{-2} = \mathcal{O}_P(1/n^2).$$

Consequently, the conditions of Theorem 2.1 of Romano and Wolf (2000) can be easily verified and it follows that $\sum_{t=1}^n n^{-1/2} X_{t,n,M}^* \xrightarrow{D} \mathcal{N}(0, 2\pi f_M(0))$, in probability. Furthermore, the absolute summability of $\gamma(h)$ implies, as $M \to \infty$, $f_M(0) \to f(0)$. Since $\{\hat{c}_{k,n}\}$ is absolutely summable, it can

be shown that the M-approximation used is sufficiently close. The absolute summability of $\{\hat{c}_{k,n}\}$ implies absolute summability of $\{\widehat{\gamma}_n(h) = E^* X^*_{t+h,n} X^*_{t,n}\}$. Let $\delta > 0$,

$$\begin{aligned}
\lim_{M\to\infty} \limsup_{n\to\infty} P\left(|\sum_{t=1}^{n} X^*_{t,n,M} - \sum_{t=1}^{n} X^*_{t,n}| > \sqrt{n}\delta \right) \delta^2 &= \lim_{M\to\infty} \limsup_{n\to\infty} \delta^2 P\left(|\sum_{t=1}^{n} \frac{1}{\sqrt{n}} \sum_{k=M+1}^{\infty} \hat{c}_{k,n} \varepsilon^*_{t-k}| > \delta \right) \\
&\leq \lim_{M\to\infty} \limsup_{n\to\infty} \sum_{h=-n+1}^{n-1} \sum_{k=M+1}^{\infty} (1 - |h|/n) \hat{c}_{k,n} \hat{c}_{k+h,n} \hat{\sigma}_n^2 \\
&= \lim_{M\to\infty} \limsup_{n\to\infty} \sum_{h=M+1}^{\infty} \widehat{\widehat{\gamma}}_n(h) - \sum_{h=-n+1}^{n-1} |h|/n \sum_{k=M+1}^{\infty} \hat{c}_{k,n} \hat{c}_{k+h,n} \hat{\sigma}_n^2 - \sum_{|h|\geq n} \sum_{k=M+1}^{\infty} \hat{c}_{k,n} \hat{c}_{k+h,n} \hat{\sigma}_n^2 \\
&= \lim_{M\to\infty} \sum_{h=M+1}^{\infty} \gamma(h) = 0,
\end{aligned}$$

where the last equation follows by the absolute summability of $\gamma(h)$. Thus, the assertion follows with a version in probability of Theorem 4.2 of Billingsley (1968).
Since $\hat{f}_n$ is a uniformly consistent estimator and $\{\widehat{\widehat{\gamma}}_n\}$ is absolutely summable, we have

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{\sqrt{n}} \sum_{t=1}^{n} X^*_{t,n} \right) = \sum_{h=-n+1}^{n-1} (1 - |h|/n)\, \widehat{\widehat{\gamma}}_n(h) &= \hat{f}_n(\lambda) - \sum_{|h|\geq n} \widehat{\widehat{\gamma}}_n(h) - \sum_{h=-n+1}^{n-1} |h|/n \widehat{\widehat{\gamma}}_n(h) \\
&= f(\lambda) + o_P(1) - \sum_{|h|\geq n} \widehat{\widehat{\gamma}}_n(h) - \sum_{h=-n+1}^{n-1} |h|/n \widehat{\widehat{\gamma}}_n(h) \to f(\lambda),
\end{aligned}$$

as $n \to \infty$, in probability. The assertion follows by the triangular inequality. □

Proof of Corollary 2.3.1. By applying the spectral-density-driven bootstrap to the time series $Y_1, \ldots, Y_n$, validity of the spectral-density-driven bootstrap for this statistic can be derived using the delta-method and similar arguments as those used the proof of Theorem 2.3.2. □

Proof of Corollary 2.3.2. Since $\tilde{f}_n(0)$ is a consistent estimator, we have with Slutsky's Theorem $\sqrt{n}(\bar{X}_n - \mu)/(2\pi\tilde{f}_n(0))^{1/2} \to \mathcal{N}(0,1)$. Furthermore, we have for $0 < \varepsilon_n < f(0)$, $\varepsilon_n \to 0$, as $n \to \infty$, that $E(n(\bar{X}_n - \mu)^2)/(2\pi\tilde{f}_n(0)) \to 1$, since

$$\begin{aligned}
E\left[\frac{n(\bar{X}_n - \mu)^2}{2\pi(f(0) + \tilde{f}_n(0) - f(0)} \mathbb{1}_{\{|\tilde{f}_n(0) - f(0)| < \varepsilon_n\}} + \frac{n(\bar{X}_n - \mu)^2}{(2\pi\tilde{f}_n(0)} \mathbb{1}_{\{|\tilde{f}_n(0) - f(0)| \geq \varepsilon_n\}} \right] \\
\leq E\frac{n(\bar{X}_n - \mu)^2}{2\pi f(0) + \varepsilon_n} + E\frac{n(\bar{X}_n - \mu)^2}{2\pi\delta} E\mathbb{1}_{\{|\tilde{f}_n(0) - f(0)| \geq \varepsilon_n\}} \to 1,
\end{aligned}$$

and $En(\bar{X}_n - \mu)^2/\{2\pi\tilde{f}_n(0)\} \geq En(\bar{X}_n - \mu)^2/\{2\pi\tilde{f}(0) + \varepsilon_n\}\mathbb{1}_{\{|\tilde{f}_n(0) - f(0)| < \varepsilon_n\}} \to 1$, as $n \to \infty$.
For a valid bootstrap approximation it is necessary that $\tilde{f}^*_n$ is a consistent estimator. The differentiability of $\hat{f}_n$ ensures that the corresponding autocovariance function fulfills $|\widehat{\widehat{\gamma}}_n(h)| \leq |h|^{-2+\varepsilon} C$ for some $\varepsilon > 0$ and for all $n \in \mathbb{N}$. Since $\{X^*_t, t \in \mathbb{Z}\}$ possesses a one-sided moving average representation with i.i.d. innovations, we have with the absolute summability of $\{\hat{c}_{k,n}\}$ and the bounded-

ness of fourth moment of the innovations that $\sup_t \sum_{h_1,h_2,h_3\in\mathbb{Z}} \mathrm{cum}^*(X_t^*, X_{t+h_1}^*, X_{t+h_2}^*, X_{t+h_3}^*) < C$. Thus, similarly to Lemma A.2 in Jentsch and Subba Rao (2015) it can be shown that $\sup_\lambda |\hat{f}_n^{\,*}(\lambda) - f(\lambda)| \to 0$, in probability, where $\hat{f}_n^{\,*}$ is a lag window spectral density estimator based on the pseudo observations $X_1^*, \ldots, X_n^*$ and fulfilling Assumption 2.1 of Jentsch and Subba Rao (2015). We construct our consistent estimator as $\tilde{f}_n^*(0) = \hat{f}_n(0) + \delta \mathbb{1}_{\{\hat{f}_n(0)<\delta\}}$. Using Theorem 2.3.2 and the same arguments as above we get $\sqrt{n}(\bar{X}_n^* - \bar{X}_n)/(2\pi\tilde{f}_n^*(0))^{1/2} \to \mathcal{N}(0,1)$, as $n \to \infty$, in probability. Furthermore, we have $E^* n(\bar{X}_n^* - \bar{X}_n)^2/(2\pi\tilde{f}_n^*(0)) = 1$ and the assertions follows by the triangular inequality. □

Proof of Corollary 2.3.3. The assumption $\sum_{k\in\mathbb{N}} |k\hat{c}_{k,n}| \leq C$ of Theorem 2.3.1 ensures that $\sum_{k\in\mathbb{N}} |k\hat{c}_{k,n}|^2 \leq \tilde{C}$. Furthermore, we have $E^*(\varepsilon_t^*)^8 < \infty$ independently from n. Thus the nonparametric estimator $\tilde{\kappa}_4^*$ of Fragkeskou and Paparoditis (2015) for the fourth moment of the innovation $\{\varepsilon_t^*\}$ can be applied and is consistent under this conditions. The assumption $\sum_{k\in\mathbb{N}} |k\hat{c}_{k,n}| \leq C$ ensures that the corresponding autocovariance function $|\widehat{\widehat{\gamma}}_n(h)| \leq |h|^{-2+\varepsilon}C$ for some $\varepsilon > 0$ and for all $n \in \mathbb{N}$. Thus, the consistency of a lag-window spectral density estimator given in Jentsch and Subba Rao (2015) follows by the same arguments as in the proof of Corollary 2.3.2. Since τ^2 is a continuous transformation of κ_4 and f, we construct the following consistent estimators of $\tau^2 > \delta$ where $0 < \varepsilon_n < \delta$ and $\varepsilon_n \to 0$, as $n \to \infty$,

$$\tilde{\tau}_2^* = \max(\varepsilon_n, (\frac{\tilde{\kappa}_4^*}{\sigma^4} - 3)(\int_0^{2\pi} \tilde{f}_n^*(\lambda) \sum_{h\in\mathbb{Z}} d(h)\exp(ih\lambda)\mathrm{d}\lambda)^2 + 4\pi \int_0^{2\pi} |\tilde{f}_n^*(\lambda) \sum_{h\in\mathbb{Z}} d(h)\exp(ih\lambda)|^2 \mathrm{d}\lambda)$$

which is based on $X_1*, \ldots, X_n^*$ and

$$\tilde{\tau}_2 = \max(\varepsilon_n, (\frac{\tilde{\kappa}_4}{\sigma^4} - 3)(\int_0^{2\pi} \tilde{f}_n(\lambda) \sum_{h\in\mathbb{Z}} d(h)\exp(ih\lambda)\mathrm{d}\lambda)^2 + 4\pi \int_0^{2\pi} |\tilde{f}_n(\lambda) \sum_{h\in\mathbb{Z}} d(h)\exp(ih\lambda)|^2 \mathrm{d}\lambda)$$

which is based on $X_1, \ldots, X_n$. The assertions follows with Theorem 2.3.1 and the same arguments as in the proof of Corollary 2.3.2. □

2.7 Estimation of a Moving Average Model

The result established in section 2.2.2 can be used to estimate the coefficients of a moving average model of order q. Let $\{\varepsilon_t, t \in \mathbb{Z}\}$ be some white noise with variance σ^2 and $\{X_t, t \in \mathbb{Z}\}$ a MA(q) model given by $X_t = \sum_{j=1}^q c_j \varepsilon_{t-j} + \varepsilon_t$, where $C(z) = \sum_{j=1}^q c_j z^j + 1 \neq 0$ for all $|z| \leq 1$. Hence, we are considering here only the case where all roots of the moving average polynomial are outside the unit disk. This ensures that a given spectral density possesses a unique representation by a moving average model. Furthermore, since all roots are outside the unit disk, the spectral density f fulfills $C_2 > f(\lambda) > C_1$ for all $\lambda \in [0, 2\pi]$ for some $C_1, C_2 > 0$. Denote $\min_{\lambda\in[0,2\pi]} f(\lambda) = C_1 > 0$ and $\max_{\lambda\in[0,2\pi]} f(\lambda) = C_2 < \infty$. This restriction is similar to the case given in the Yule-

Walker equations for the estimation of an autoregressive model. Using the Yule-Walker equations to estimate the coefficients of an autoregressive model leads to an autoregressive polynomial with all roots outside the unit disk. We consider the case that the order q of the moving average polynomial is known. Furthermore, consider the spectral density estimator

$$\hat{f}_n(\lambda) = \tilde{f}_n(\lambda) - \min_{\lambda\in[0,2\pi]} \tilde{f}_n(\lambda) + C_1, \text{ where } \tilde{f}_n(\lambda) 2\pi \sum_{h=-q}^{q} \hat{\gamma}_n(h) \exp(-ih\lambda), \tag{2.7.1}$$

where $\hat{\gamma}_n(h)$ is an estimator of the autocovariance function and it fulfills $|\hat{\gamma}_n(h) - \gamma(h)| = \mathcal{O}_P(\sqrt{(n-h)}/n)$. Let $\hat{c}_{1,n}, \ldots, \hat{c}_{q,n}, \sigma_n^2$ be the estimators obtained by using the spectral density factorization described in section 2.2.2 with the spectral density estimator (2.7.1). The following results can be established.

Theorem 2.7.1. *Let $X_t = \sum_{j=1}^{q} c_j \varepsilon_{t-j} + \varepsilon_t, t \in \mathbb{Z}$ be a moving average model of order q and the moving average polynomial has all its roots outside the unit disk. Furthermore, let $\hat{\gamma}_n$ be an estimator of the autocovariance function and it fulfills $|\hat{\gamma}_n(h) - \gamma(h)| = \mathcal{O}_P(\sqrt{(n-h)}/n)$. Then the estimators $\hat{c}_{1,n}, \ldots, \hat{c}_{q,n}, \sigma_n^2$ obtained by using the spectral density factorization described in section 2.2.2 with the spectral density estimator* (2.7.1) *fulfill*

1.
$$\hat{\sigma}_n^2 = \sigma^2 + \frac{1}{C_1} \mathcal{O}_P\left(\frac{q}{\sqrt{n}}\right),$$

2.
$$\sum_{k=1}^{q} |c_k - \hat{c}_{k,n}| = \sqrt{\frac{C_2}{C_1}} \mathcal{O}_P\left(q^{3/2} \frac{1}{\sqrt{n}}\right).$$

Proof. Since we are in the setting of a moving average process of order q, we have for the spectral density estimator 2.7.1 the following convergence rate

$$\begin{aligned} \sup_{\lambda\in[0,2\pi]} |\hat{f}_n(\lambda) - f(\lambda)| &\leq 2 \sup_{\lambda\in[0,2\pi]} |\frac{1}{2\pi} \sum_{h=-q}^{q} (\gamma(h) - \hat{\gamma}_n(h)) \exp(-ih\lambda)| \\ &\leq \frac{1}{\pi} \sum_{h=-q}^{q} |\gamma(h) - \hat{\gamma}_n(h)| = \sum_{h=-q}^{q} \mathcal{O}_P\left(\frac{\sqrt{n-h}}{n}\right) = \mathcal{O}_P\left(\frac{q}{\sqrt{n}}\right). \end{aligned}$$

Since $f, \hat{f}_n$ are strictly positive, this convergence rate can be also obtained for $\hat{f}_n^{1/2}$ and $\log \hat{f}_n$. Hence, we have

$$\sup_{\lambda\in[0,2\pi]} |\hat{f}_n(\lambda)^{1/2} - f(\lambda)^{1/2}| = C_1^{-1/2} \mathcal{O}_P\left(\frac{q}{\sqrt{n}}\right),$$

and

$$\sup_{\lambda\in[0,2\pi]} |\log \hat{f}_n(\lambda) - \log f(\lambda)| = C_1^{-1} \mathcal{O}_P\left(\frac{q}{\sqrt{n}}\right).$$

Theorem 2.2.2 a) gives us the following bound

$$\sum_{k=1}^{q}|c_k-\hat{c}_{k,n}|^2 \leq \int_0^{2\pi}\Big(f^{1/2}(\lambda)-\hat{f}_n^{\,1/2}(\lambda)\Big)^2 \mathrm{d}\lambda + \sup_{\lambda\in[0,2\pi]}(f(\lambda)\hat{f}_n(\lambda))^{1/2}\int_0^{2\pi}(\log f(\lambda)-\log \hat{f}_n(\lambda))^2\mathrm{d}\lambda.$$

Hence, we obtain $\sum_{k=1}^{q}|c_k-\hat{c}_{k,n}|^2 \leq \mathcal{O}_P(q^2/n)C_2/C_1$. Using Minkowski's inequality we obtain $\sum_{k=1}^{q}|c_k-\hat{c}_{k,n}| \leq \mathcal{O}_P(q^{3/2}/\sqrt{n})\sqrt{C_2/C_1}$. Since $\hat{\sigma}_n^2=\exp((2\pi)^{-1}\int_{-\pi}^{\pi}\log(\hat{f}_n(\lambda))d\lambda)$, we have with Jensen's theorem

$$\begin{aligned}|(\hat{\sigma}_n^2-\sigma^2)/\sigma^2| &= \pm(\exp((2\pi)^{-1}\int_{-\pi}^{\pi}\log(\hat{f}_n(\lambda)/f(\lambda))d\lambda)-1)\\ &\leq \begin{cases}(2\pi)^{-1}\int_{-\pi}^{\pi}(\hat{f}_n(\lambda)-f(\lambda))/f(\lambda))d\lambda) \leq 1/C_1\sup_{\lambda\in[0,2\pi]}|\hat{f}_n(\lambda)-f(\lambda)|\\ -(1+(2\pi)^{-1}\int_{-\pi}^{\pi}\log(\hat{f}_n(\lambda)/f(\lambda))d\lambda)+1 \leq \sup_{\lambda\in[0,2\pi]}|\log\hat{f}_n(\lambda)-\log f(\lambda)|\end{cases}\\ &= C_1^{-1}\mathcal{O}_P\left(\frac{q}{\sqrt{n}}\right).\end{aligned}$$

□

The spectral density estimator as well as the proof are kept simple and might be not chosen optimally regarding the obtained constants. The focus is here to obtain $\sqrt{n}$ consistency for the moving average coefficients. Hence, as can be seen in the spectral density estimator given by 2.7.1, the spectral density factorization gives an easy and $\sqrt{n}$ consistent way to fit an MA(q) model. However, in a finite sample, one challenge remains. The constant C_1 is usually unknown, that is why we recommend to use some factor such as $1/n$. This ensures that the estimator is positive. Furthermore, this factor is small enough so that this factor does not interfere with the $\sqrt{n}$-rate. Besides the strategy of lifting the whole spectrum, which is like adding independent white noise to the process, another strategy is to bound the obtained spectrum from below. Hence $\tilde{f}=\max(C,\hat{f})$. Both approaches are asymptotically negligible. However, lifting the whole spectrum keeps the inherent structure and may give better finite sample results.

2.8 Comparison with the Linear Process Bootstrap

The Linear Process Bootstrap (LPB) has been introduced by McMurry and Politis (2010). To briefly describe this procedure, we center the vector of observations $\underline{X}=(X_1,\ldots,X_n)$ and consider the empirical autocovariance matrix $\hat{\mathbf{\Gamma}}_{\mathbf{n}}=n^{-1}(\underline{X}-\overline{X}_n\mathbf{1}_n)(\underline{X}-\overline{X}_n\mathbf{1}_n)^\top$, where $\overline{X}_n$ is the sample mean and $\mathbf{1}_n=(1,1,\ldots,1)^\top$. To ensure stability, $\hat{\mathbf{\Gamma}}_n$ is tapered such that a banded matrix $\tilde{\mathbf{\Gamma}}_n$ is obtained, see McMurry and Politis (2010) for details. A key step in the LPB algorithm is the computation of the matrices $\tilde{\mathbf{\Gamma}}_n^{-1/2}$ and $\tilde{\mathbf{\Gamma}}_n^{1/2}$. Using Cholesky's decomposition gives the expression $\tilde{\mathbf{\Gamma}}_n=\mathbf{\Phi}_n\mathbf{\Sigma}_n\mathbf{\Phi}_n^\top$ with $\mathbf{\Sigma}_n$ a diagonal matrix and $\mathbf{\Phi}_n$ a unit lower triangular matrix. Let $\mathbf{B}_n=\mathbf{\Phi}_n^{-1}$. The LPB algorithm then

generates new pseudo time series $\underline{Y}_n^* = (Y_1^*, Y_2^*, \ldots, Y_n^*)^\top$ as follows. It first computes the residual vector $\underline{Z}_n = \mathbf{\Sigma}_n^{-1/2}\mathbf{B}_n\underline{X}$ and standardizes $\underline{Z}_n$, such that the residuals have mean zero and variance one. Denote the standardized vector by $\widetilde{Z}_n = (\widetilde{Z}_1, \widetilde{Z}_2, \ldots, \widetilde{Z}_n)^\top$. A vector $\underline{Z}_n^* = (Z_1^*, \ldots, Z_n^*)$ is generated by choosing for each $j = 1, 2, \ldots, n$, the pseudo random variable Z_j^* with replacement from the set $\{\widetilde{Z}_1, \ldots, \widetilde{Z}_n\}$. $\underline{Y}_n^*$ is then computed as $\underline{Y}_n^* = \mathbf{\Phi}_n\mathbf{\Sigma}_n^{1/2}\underline{Z}_n^*$.

To better understand the above LPB procedure, it is important to shed some light on the matrix $\mathbf{B}_n$; see also Pourahmadi (2001, Chapter 6 and 7). Let $\hat{X}_1 = 0$ and $\hat{X}_{l+1}$ be the projection of X_{l+1} onto $\text{span}\{X_1, \ldots, X_l\}$, i.e., $\hat{X}_{l+1} = \sum_{k=1}^{l} \beta_{k,l} X_{l+1-k}$, where the coefficients $\beta_{k,l}$ are obtained by solving the corresponding Yule-Walker equations. These projections for $l = 0, \ldots, n-1$ together with the Gram-Schmidt procedure lead to the innovations $(Z_1, \ldots, Z_n)$; see also Pourahmadi (2001, Section 7.1.2). That is, we have, $Z_{l+1} = (X_{l+1} - \sum_{k=1}^{l} \beta_{k,l} X_{l+1-k})/\sigma_l, \quad l = 1, \ldots, n-1$, where $\sigma_l = \|X_{l+1} - \sum_{k=1}^{l} a_{k,l} X_{l+1-k}\|$ and $Z_1 = X_1/\|X_1\|$. This means that in the first step, the LPB procedure uses the above equation for $l = 0, \ldots, n-1$ to get the residuals $Z_1, Z_2, \ldots, Z_n$ and the covariance matrix $\mathbf{\Sigma}_n = \text{diag}(\sigma_0^2, \ldots, \sigma_{n-1}^2)$. The unit lower triangular matrix $\mathbf{B}_n = (-\beta_{i-j,i-1})_{i=2,\ldots,n, j=1,\ldots,i-1}$ consists of the corresponding AR(l) coefficients for values of $l = 1, \ldots, n-1$. With $\mathbf{B}_n^{-1} = \mathbf{\Phi}_n = (\phi_{i-j,i-1})_{i=1,\ldots,n, j=1,\ldots,i}$, Y_{l+1}^* is then generated as $Y_{l+1}^* = \sum_{k=0}^{l} \phi_{k,l}\sigma_{l-k} Z_{l+1-k}^*, \ l = 1, \ldots, n-1$. Thus the pseudo observation $Y_1^*, \ldots, Y_n^*$ of the LPB are obtained as linear transformations of the generated pseudo innovations $Z_1^*, \ldots, Z_n^*$, where the coefficients of this linear transformations depend on the index t of Y_t^*. In fact to generate Y_l^* a MA($l-1$) model is used and the order of this moving average model changes with the index l. Furthermore, the calculations undertaken to compute the matrix $\mathbf{\Phi}_n$ are identical to those used in the innovation algorithm (Brockwell and Davis, 1991, Proposition 5.2.2) to compute the moving average coefficients. Consequently, the MA($q-1$)-model used to generate the q-th pseudo observation Y_q^*, corresponds to the $(q-1)$-th iteration of the innovation algorithm. Since $\tilde{\mathbf{\Gamma}}_\mathbf{n}$ is a banded matrix, the order of the used moving average model stabilizes. Nevertheless, the coefficients of these models can still differ slightly from iteration to iteration. As pointed out in section 2.2, the coefficient $\phi_{k,l}$ converges to the coefficient c_k of the Wold representation of the process X as $l \to \infty$ for any $k = 1, 2, \ldots$. Similarly, the coefficient $\beta_{k,l}$ of the corresponding AR(l) fit converges to the coefficient b_k of the autoregressive representation of the process, as $l \to \infty$, for any $k = 1, 2, \ldots$. Nevertheless, in finite samples, the coefficients $\beta_{k,l}$ resp. $\phi_{k,l}$ differ in general from the coefficients $\hat{c}_{k,n}$ resp. $\hat{b}_{k,n}$ obtained using the factorization of the estimated spectral density $\hat{f}_n$.

In order to clarify the differences, we examine the following example. Consider the time series $X_t = \varepsilon_t - \varepsilon_{t-1}, t \in \mathbb{Z}$, with an innovations process ε_t having mean zero and variance one. The above MA(1) representation is also the Wold representation of this process. Assume that the autocovariance is known. The LPB algorithm leads to the coefficients $b_{0,l} = 1, b_{1,l} = l/(l+1), b_{k,l} = 0$, for

$k \geq 2$ and the variance $\sigma_l^2 = 2\prod_{k=1}^{l}(1-(l+1)^{-2})$; see also example 7.1 b in Pourahmadi (2001). Let $Z_1^*, \ldots, Z_n^*$ be the standardized pseudo innovations. Then we have the following equations for generating the pseudo observations:

$$\begin{aligned} Y_1^* &= \sqrt{2}Z_1^*, \\ Y_2^* &= \sqrt{3/2}Z_2^* - \sqrt{1/2}Z_1^*, \\ Y_3^* &= \sqrt{4/3}Z_3^* - \sqrt{2/3}Z_2^* \\ &\vdots \end{aligned}$$

Notice that the pseudo observations $Y_1^*, Y_2^*, \ldots$ have the same covariance structure as the process X, but a stable *model structure* appears only asymptotically. The same holds true for the coefficients of the *moving average representation* used to generate the pseudo observations Y_l^*. As it is seen, these coefficients converge to the true coefficients as $l \to \infty$, but they differ for any finite sample.
In contrast to this, if one factorizes the spectral density, $f(\cdot) = 1/(2\pi)\sum_{h=-1}^{1}\gamma(h)\exp(ih\cdot)$, as used in the spectral-density-driven bootstrap procedure proposed, one would directly get the true coefficients in the moving average representation and could use them to generate the pseudo observations. Consequently, the LPB creates a pseudo data-vector $(Y_1^*, \ldots, Y_n^*)$ where for each index t a different MA-model is used to generate the pseudo observation Y_t^* and the MA-order used increases with the index t. This is in contrast to the spectral-density-driven bootstrap procedure, where a time series $X_1^*, X_2^*, \ldots, X_n^*$ is generated using the estimated (possible infinite) moving average representation of the process.

The linear process bootstrap procedure can be slightly modified such that a stable moving average representation is used, c.f. McMurry and Politis (2017). The idea is to create bootstrap samples by using the last row of coefficients given by $\boldsymbol{\Phi}_n$, $e_n^\top \boldsymbol{\Phi}_n$. Hence, we have $Y_t^* = \sum_{j=0}^{n-1}\phi_{j,n-1}\varepsilon_t^*, t \in \mathbb{Z}$, where $\{\varepsilon_t^*, t \in \mathbb{Z}\}$ is some pseudo innovation process. The obtained bootstrap procedure is similar to the spectral-density-driven bootstrap. The difference is that the coefficients $\phi_{j,n-1}, j = 0, \ldots, n-1$ are obtained by the factorization of an autocovariance matrix $\Gamma_n = (\gamma(i-j))_{i,j=1,\ldots,n}$, whereas the coefficients $\{c_k, k \in \mathbb{N}_0\}$ are obtained by the factorization of a spectral density $f(\cdot) = 1/(2\pi)\sum_{h\in\mathbb{Z}}\gamma(h)\exp(-ih\cdot)$. For the case $n \to \infty$, hence, $\Gamma = (\gamma(i-j))_{i,j\in\mathbb{N}}$, the autocovariance matrix contains the same information as the spectral density. Both describe the second-order properties of a given time series completely. Given the same informations, both factorization procedures lead to the same coefficients. However, in practice it is not possible to work with the infinite structure. Instead of Γ a finite Γ_m is factorized and as mentioned in section 2.2.2 the integrals to compute the coefficients (c_k) are usually approximated by a sum over a finite number M of frequencies. These approximations lead to a slightly different result for finite m and M. However, simulations

for a given autocovariance show that if m and M gets larger the differences soon becomes negligible, see Subsection 2.8 for details. The factorization of autocovariance matrices has the charming aspect that one gets 'true' zeros, whereas the factorization of spectral densities has the advantage that the numerical error is smaller in the case that a closed form of the spectral density exists for autocovariance functions γ with $|\gamma(h)| > 0$ for all $\mathbb{Z}$, such as autoregressive or exponential models. Furthermore, the fast Fourier transform used by the spectral density factorization scales better than the Cholesky factorization. Hence, despite the fact that M needs to be of order $\mathcal{O}(n^2)$ to obtain adequate approximation of $c_k, k = 1, \ldots, n$, for larger n the spectral density is way faster than a factorization of Γ_n.

Let (c_k) be the Wold coefficients, $(\hat{\phi}_{k,n})$ be estimators given by a factorization of an empirical autocovariance matrix and $(\hat{c}_{k,n})$ be estimators given by a spectral density estimator. Then we have under some assumptions, mainly that the spectral density estimator is uniformly consistent, see Theorem 2.2.2, that $\sum_{k=0}^{\infty} |\hat{c}_{k,n} - c_k| = o_P(1)$. Under some assumptions, mainly that the empirical autocovariance matrix is trimmed adequately by using a flat-top kernel, McMurry and Politis (2017) proofed $\sum_{k=0}^{\infty} |\hat{\phi}_{k,n} - c_k| = o_P(1)$. So it can be shown that both estimators are globally consistent, however, the consistency of the factorization of an empirical autocovariance matrix is (so far) only proofed for a special class of estimators of the second-order structure.

Comparison of autocovariance matrix factorization and spectral density factorization

In order to investigate the differences in finite samples between the factorization of spectral densities and autocovariance matrices a simulation study is performed. For this the second-order properties of Model I, Model II and Model III given in section 2.4 are used. The focus is not to investigate the finite sample performance of a given spectral density estimator that is why no estimator at all is used. Instead, we consider the setting that the true autocovariance is known up to order n. This is comparable to the setting that the same estimation procedure is used in both factorization methods. In order to factorize an autocovariance matrix of order $m > n$ or to factorize a spectral density based on the autocovariances up to n given at Fourier frequencies $M > n$, it is necessary to extend the autocovariance. This is done by adding zeroes up to order m or M, respectively. Positive definiteness of the extended autocovariance function is ensured by lifting the resulting spectral density above level zero. Hence, $\gamma(0)$ is increased such that the extended autocovariance function becomes positively definite.

The results for Model I, an AR(1) process given by $X_t = 0.9X_{t-1} + \varepsilon_t$, are given in Table 2.4. This process possesses a fastly decaying autocovariance function, hence, its second-order properties are almost entirely described by an autocovariance function up to lag 1024. Furthermore, its Wold coefficients decline rapidly and consequently, this model brings no difficulties. Both factorizations perform perfectly.

The results for Model II, an ARMA$(4,2)$ process given by $X_t = 1.34X_{t-1} - 1.88X_{t-2} + 1.32X_{t-3} - 0.8X_{t-4} + \varepsilon_t + 0.71\varepsilon_{t-1} + 0.25\varepsilon_{t-2}$, are given in Table 2.5. This process possesses a slowly decaying autocovariance function, even $|\rho(2^{10})| \approx 5\%$. Consequently, its autocovariance function is needed to a large lag to sufficiently describe its second-order properties. That is why, independently from the used factorization the deviation between the true Wold coefficients and their estimation only gets small if n is sufficiently large, $n \geq 4096$. The differences between $(\hat{\phi}_{k,n,m})$ and $(\hat{c}_{k,n,M})$ are rather small for larger m, M, respectively. However, for smaller n the autocovariance matrix factorization seems to perform a little bit better, whereas its vice versa for larger n.

The results for Model III, a MA(10) process given by $X_t = \varepsilon_t + \sum_{k=1}^{10} \binom{n}{k}(-1)^k \varepsilon_{t-k}$, are given in Table 2.6. Since this process is a moving average model of order 10, only autocovariances up to 10 could be different from 0. The moving average polynomial possesses a ten times unit root at $\exp(i0)$, hence, its spectral density is zero at $\lambda = 0$. This makes this setting very challenging and it can be seen that both factorizations more or less fail. The deviation between the true Wold coefficients and their estimation is rather large. Nevertheless, it is nicely visible that the deviation gets smaller the bigger m, M gets. In this setting the spectral density factorization clearly outperforms the autocovariance factorization, so the spectral density is best among worst in this setting. However, even for $M = 2^{28}$ we have a deviation of 348. Note that the deviation between the corresponding spectral densities is negligible, $\int_{-\pi}^{\pi} |\sum_{k=0}^{\infty} c_k - \hat{c}_{k,1024,2^{24}} \exp(ik\lambda)| \mathrm{d}\lambda < 10^{-5}$. Furthermore, note that $m = 2^{14}$ and $M = 2^{28}$ are the highest (integer) powers of 2 which requires less than 16 GB of RAM to compute the factorization in *R*.

Table 2.4: Comparison of autocovariance matrix factorization $((\hat{\phi}_{k,n,m}))$ and spectral density factorization $((\hat{c}_{k,n,M}))$ for Model I: $X_t = 0.9X_{t-1} + \varepsilon_t$

$n =$	$m, M^{1/2} = \min(n, \cdot)$	$\sum_{k=1}^{100} \lvert\hat{\phi}_{k,n,m} - c_k\rvert$	$\sum_{k=1}^{100} \lvert\hat{c}_{k,n,M} - c_k\rvert$	$\sum_{k=1}^{100} \lvert\hat{\phi}_{k,n,m} - \hat{c}_{k,n,M}\rvert$
1024	256	2.06e-14	1.93e-14	2.56e-14
1024	1024	2.57e-14	1.40e-14	2.11e-14
1024	4096	1.76e-14	1.43e-14	1.90e-14
2048	256	2.59e-14	1.93e-14	2.95e-14
2048	1024	2.59e-14	1.40e-14	2.55e-14
2048	4096	1.97e-14	1.42e-14	1.76e-14
4096	256	1.67e-14	1.90e-14	2.04e-14
4096	1024	1.67e-14	1.42e-14	1.43e-14
4096	4096	2.07e-14	1.43e-14	1.79e-14
8192	256	1.78e-14	1.80e-14	2.00e-14
8192	1024	1.78e-14	1.42e-14	1.66e-14
8192	4096	1.78e-14	1.44e-14	1.77e-14

Table 2.5: Comparison of autocovariance matrix factorization $((\hat{\phi}_{k,n,m}))$ and spectral density factorization $((\hat{c}_{k,n,M}))$ for Model II: $X_t = 1.34X_{t-1} - 1.88X_{t-2} + 1.32X_{t-3} - 0.8X_{t-4} + \varepsilon_t + 0.71\varepsilon_{t-1} + 0.25\varepsilon_{t-2}$

$n =$	$m, M^{1/2} = \min(n, \cdot)$	$\sum_{k=1}^{100} \lvert\hat{\phi}_{k,n,m} - c_k\rvert$	$\sum_{k=1}^{100} \lvert\hat{c}_{k,n,M} - c_k\rvert$	$\sum_{k=1}^{100} \lvert\hat{\phi}_{k,n,m} - \hat{c}_{k,n,M}\rvert$
1024	256	94.1	71.8	60.5
1024	1024	74.1	71.8	3.17
1024	4096	70.8	7.18e+01	1.70e+00
2048	256	32.1	2.56e+01	6.85e+00
2048	1024	32.1	2.56e+01	6.84e+00
2048	4096	24.1	2.56e+01	1.52e+00
4096	256	0.343	3.04e-03	3.40e-01
4096	1024	0.343	3.04e-03	3.40e-01
4096	4096	1.63e-11	3.04e-03	3.04e-03
8192	256	3.20e-07	2.08e-11	3.20e-07
8192	1024	3.20e-07	7.85e-12	3.20e-07
8192	4096	3.20e-07	1.15e-11	3.20e-07

Table 2.6: Comparison of autocovariance matrix factorization $((\hat{\phi}_{k,n,m}))$ and spectral density factorization $((\hat{c}_{k,n,M}))$ for Model III: $X_t = \varepsilon_t + \sum_{k=1}^{10} \binom{n}{k}(-1)^k \varepsilon_{t-k}$

$n =$	$m, M^{1/2} = \min(n, \cdot)$	$\sum_{k=1}^{100} \lvert\hat{\phi}_{k,n,m} - c_k\rvert$	$\sum_{k=1}^{100} \lvert\hat{c}_{k,n,M} - c_k\rvert$	$\sum_{k=1}^{100} \lvert\hat{\phi}_{k,n,m} - \hat{c}_{k,n,M}\rvert$
1024	256	624	524	111
1024	1024	585	458	135
1024	4096	548	399	151
2048	256	624	524	111
2048	1024	585	458	135
2048	4096	548	399	151
1024	$m = 2^{13}, M = 2^{26}$	530	372	158
1024	$m = 2^{14}, M = 2^{28}$	513	348	165

2.9 Additional Simulation Results

2.9.1 Sample Size $n = 128$

Tables 2.7 and 2.8 present additional results about the coverage probabilities if $n = 128$ and non-studentized statistics are used by the different bootstrap methods.

Table 2.7: Coverage probabilities (in percent) for the mean using the non-studentized statistic $\bar{X}_n$ for a sample size $n = 128$

	Model I			Model II			Model III		
$(1-\alpha)100$	80.0	90.0	95.0	80.0	90.0	95.0	80.0	90.0	95.0
SDDB	73.0	82	87.3	75.6	86.5	92.5	81.2	90.0	94.1
LPB	67.5	76.2	83.0	83.4	90.4	94.3	67.2	74.9	81.2
TBB	60.7	70.2	75.1	36.8	44.0	49.7	52.3	60.9	67.5
ND	74.1	82.3	87.1	76.2	86.6	92.5	79.4	92.3	96
ARS	72.4	81.7	87.3	67	78.4	86.4	60.6	73.2	81.3
BB	42.2	49.3	54.0	59.5	70.5	78.3	16.3	21.2	27.1

Table 2.8: Coverage probabilities (in percent) for the lag 2 autocorrelation using the non-studentized empirical autocorrelation at lag2 and for a sample size $n = 128$

	Model I			Model II			Model III		
$(1-\alpha)100$	80.0	90.0	95.0	80.0	90.0	95.0	80.0	90.0	95.0
SDDB	81.7	89.6	92.9	76.8	83.9	89.1	78.2	88.2	92.5
LPB	85.1	91.4	94.9	86.9	90.9	94.0	81.8	91.0	94.7
TBB	73.1	78.3	82.0	15.4	18.5	20.5	66.1	72.8	76.2
ND	83.4	92.1	96.2	81.4	89.3	93.4	78.9	89.4	94.5
ARS	81.5	89.5	92.8	55.4	62.5	66.8	79.5	88.5	93.3
BB	88.3	95.0	97.1	62.7	80.7	86.9	92.9	98.1	99.1

2.9.2 Sample Size $n = 512$

Table 2.10 and 2.9 present additional results about the coverage probabilities if $n = 512$ and the studentized and non-studentized mean statistic are used by the different bootstrap methods.

Table 2.9: Coverage probabilities (in percent) for the mean using studentized statistic of $\bar{X}_n(2\pi\hat{f}_n)^{-1/2}$ and for a sample size $n = 512$

	Model I			Model II			Model III		
$(1-\alpha)100$	80.0	90.0	95.0	80.0	90.0	95.0	80.0	90.0	95.0
SDDB	80.2	90.1	95.0	79.2	89.1	94.1	82.4	91.2	95.9
LPB	79.4	89.2	93.8	82.3	89.0	92.3	39.1	52.0	69.0
TBB	72.0	81.9	87.6	31.1	37.3	42.1	47.0	52.6	58.0
ND	77.0	87.0	92.8	74.9	84.9	91.4	27.1	37.1	44.9
ARS	79.2	89.4	94.1	79.8	88.9	93.7	46.7	60.7	73.4
BB	19.1	35.8	51.8	20.7	35.4	52.2	30.1	41.9	51.4

Table 2.10: Coverage probabilities (in percent) for the mean using the non-studentized statistic $\bar{X}_n$ and for a sample size $n = 512$

	Model I			Model II			Model III		
$(1-\alpha)100$	80.0	90.0	95.0	80.0	90.0	95.0	80.0	90.0	95.0
SDDB	76.2	86.9	92.8	79.5	89.8	95.0	83.1	90.9	95.2
LPB	74.0	84.1	89.3	83.0	90.8	94.5	69.4	76.4	85.4
TBB	70.0	79.6	84.8	29.0	35.7	41.3	50.5	58.0	65.4
ND	77.0	87.5	92.8	79.8	89.8	95.1	81.3	95.0	97.3
ARS	76.1	87.1	92.8	73.7	85.3	91.6	67.2	80.5	87.3
BB	59.6	62.5	64.7	92.7	96.9	98.4	50.2	61.5	64.2

Table 2.12 and 2.11 present additional results about the coverage probabilities when $n = 512$ and the studentized and non-studentized sample autocorrelation at lag 2 are used by the different bootstrap methods.

Table 2.11: Coverage probabilities (in percent) for the lag 2 autocorrelation using the studentized the empirical autocorrelation at lag 2 and for a sample size $n = 512$

	Model I			Model II			Model III		
$(1-\alpha)100$	80.0	90.0	95.0	80.0	90.0	95.0	80.0	90.0	95.0
SDDB	81.8	90.7	94.7	83.3	91.6	95.3	79.6	89.5	95.5
LPB	83.5	92.4	95.2	93.6	96.8	98.0	80.3	89.8	95.3
TBB	84.0	91.4	94.5	26.6	29.1	31.6	78.5	86.4	90.8
ND	80.1	89.6	94.0	78.1	88.8	93.2	79.7	89.1	95.1
ARS	81.3	90.3	94.8	84.7	92.6	95.8	79.8	90.4	94.7
BB	19.9	37.0	53.9	18.0	31.6	48.5	20.4	35.9	52.5

Table 2.12: Coverage probabilities (in percent) for the lag 2 autocorrelation using the non-studentized the empirical autocorrelation at lag 2 and for a sample size $n = 512$

	Model I			Model II			Model III		
$(1-\alpha)100$	80.0	90.0	95.0	80.0	90.0	95.0	80.0	90.0	95.0
SDDB	79.4	87.0	92.9	76.2	83.7	87.8	82.4	91.2	94.9
LPB	81.2	89.5	94.3	88.8	92.3	95.0	83.6	92.0	95.8
TBB	78.6	85.9	90.1	24.0	28.0	30.6	79.5	88.1	90.8
ND	80.4	89.1	95.2	80.6	89.6	93.5	82.5	91.5	95.6
ARS	79.4	87.5	92.1	63.0	70.4	74.7	82.6	91.1	94.7
BB	99.8	100	100	86.1	91.3	93.7	100	100	100

2.10 Additional proofs

In order to proof (2.4), we first proof this useful lemma.

Lemma 2.10.1. *Consider the open set* $G_\beta = \{z \in \mathbb{C} : z = 1 - \beta\tilde{z}, \tilde{z} \in \mathbb{C}, |\tilde{z}| < 1\}$. *Then it holds for* $|\beta| \leq 1$ *that* log *is analytic on* G_β *and furthermore, we have*

$$\log(z) = -\sum_{k=1}^{\infty} \frac{(\beta\tilde{z})^k}{k}, \text{ for all } z = 1 - \beta\tilde{z}, \tilde{z} \leq 1, \beta\tilde{z} \neq 1.$$

Proof. There is no periodicity of exp in G_β, thus, exp is injective on G_β. To see this, consider $z = 1 - \beta\tilde{z} \in G_\beta$ and a 2π shift of it, i.e., $z + il2\pi, l \in \mathbb{Z}, l \neq 0$. We have $z + il2\pi = 1 - \beta(\tilde{z} - il2\pi/\beta) \notin G_\beta$, since $|\tilde{z} - il2\pi/\beta| \geq ||\tilde{z}| - |2\pi|/|\beta|| = 2\pi/\beta - |\tilde{z}| \geq 2\pi - 1 > 1$. Furthermore, it holds $\exp(z) \neq 0$ for all $z \in G_\beta$.

With the Implicit Function Theorem, see Freitag and Busam (2006, Satz 5.7, p. 46), we have that $\exp(G_\beta) = \{y \in \mathbb{C} : y = \exp(z), z \in G_\beta\}$ is an open set and $\exp^{-1}(y) =: \log(y) : \exp(G_\beta) \to \mathbb{C}$ is an analytic function with derivative $1/y$.

For $|z = t\exp(i\lambda)| < 1$ it holds $\sum_{k=0}^{\infty} t^k \exp(ik\lambda) = 1/(1 - t\exp(\lambda))$. An integration over t form 0 to $r < 1$ gives us $\int_0^r \sum_{k=0}^{\infty} t^k \exp(i\lambda k)dt = \int_0^r \frac{1}{1-t\exp(i\lambda)}dt \iff -\sum_{k=1}^{\infty} \frac{1}{k} r^k \exp(i\lambda k) = \log(1 - r\exp(i\lambda))$ Due to the continuity of these expression, this can be extended to $r = 1$ for $\lambda \neq 0$. □

Wold's power series has no roots inside the unit disk, thus every linear factor can be written in the form of G_β.

Lemma 2.10.2. *Let* $f = |C(\exp(i\lambda))|^2|\sigma/(2\pi)$ *be a spectral density of a nondeterministic stationary process and assume that the power series* $C(z) = \sum_{k=0}^{\infty} c_k z^k, c_0 = 1$ *has no roots inside the unit disk. Furthermore, let* $a_k = \int_{(-\pi,\pi]} \log f(\lambda) \exp(-ik\lambda)\mathrm{d}\lambda/(2\pi), k \in \mathbb{Z}$ *be the* k*-th Fourier coefficient of* log f. *Then it holds for* $|z| < 1$

$$\sigma(2\pi)^{-1/2} \sum_{k=0}^{\infty} c_k z^k = \exp\left(a_0/2 + \sum_{k=1}^{\infty} a_k z^k\right).$$

If the spectral density is bounded, this holds for $|z| = 1$ *as well.*

Proof. Let $\beta_j, j = 1, 2, \ldots$ be the roots of C, consequently $C(z) = \sum_{k=0}^{\infty} c_k z^k = \prod_{j=1}^{\infty}(1 - \beta_j z)$. Since $C(z) \neq 0$ for all $|z| < 1$, it holds $|\beta_j| \leq 1$ for all $j = 1, 2, \ldots$. Let $\beta_0 = \sigma/\sqrt{2\pi}$, $\phi(z) = \beta_0 \prod_{j=1}^{\infty}(1 - \beta_j z)$, and $\phi^+(z) = \exp\left(\frac{1}{2\pi}\int_0^{2\pi} \frac{\exp(it)+z}{\exp(it)-z} \log|\phi(\exp(it))|dt\right)$. Since $\log f$ is integrable, $\log|\phi|$ is integrable and ϕ^+ defines an outer function and is analytic inside the unit disk, see (Hoffman, 1962, Chapter 5). Let $z = r\exp(i\lambda)$, then

$$\phi^+(r\exp(i\lambda)) = \exp\left(\frac{1}{2\pi}\int_0^{2\pi}\frac{1+r\exp(i(\lambda-t))}{1-r\exp(i(\lambda-t))}\log|\phi(\exp(it))|dt\right)$$
$$= \exp\left(\frac{1}{2\pi}\int_0^{2\pi}\left(2\sum_{j=0}^{\infty}r^j\exp(i(\lambda-t)j)-1\right)\log|\phi(\exp(it))|dt\right).$$

With $\log|\phi(\exp(it))| = \log|\beta_0| + \sum_{l=1}^{\infty}\log|1-\beta_l\exp(it)|$ we have

$$\phi^+(r\exp(i\lambda)) = |\beta_0|\exp\left(\frac{1}{2\pi}\int_0^{2\pi}\left(2\sum_{j=0}^{\infty}r^j\exp(i(\lambda-t)j)-1\right)dt\right)$$
$$\times\exp\left(-\frac{1}{2\pi}\int_0^{2\pi}\sum_{l=1}^{\infty}\log|1-\beta_l\exp(it)|dt\right)$$
$$\times\exp\left(\frac{1}{2\pi}\int_0^{2\pi}\left(\sum_{j=0}^{\infty}r^j\exp(i(\lambda-t)j)-1\right)\sum_{l=1}^{\infty}\log\left((1-\beta_j\exp(it))(1-\bar\beta_j\exp(-it))\right)dt\right)$$
$$=: (I)\times(II)\times(III),$$

with obvious notation for $(I),(II),(III)$. We get for these terms using Lemma 2.10.1

$$(III) = \exp\left(\frac{1}{2\pi}\int_0^{2\pi}\left(\sum_{j=0}^{\infty}r^j\exp(i(\lambda-t)j)-1\right)\sum_{l=1}^{\infty}\log(1-\beta_j\exp(it))+\log(1-\bar\beta_j\exp(-it))dt\right)$$
$$= \prod_{l=1}^{\infty}\exp\left(\left(\sum_{j=0}^{\infty}r^j\exp(i(\lambda)j)-1\right)\sum_{k=1}^{\infty}\frac{(-1)}{k}(\beta_l)^k\frac{1}{2\pi}\int_0^{2\pi}\exp(it(k-j))dt\right)$$
$$\cdot\exp\left(\left(\sum_{j=0}^{\infty}r^j\exp(i(\lambda)j)-1\right)\sum_{k=1}^{\infty}\frac{(-1)}{k}(\bar\beta_l)^k\frac{1}{2\pi}\int_0^{2\pi}\exp(-it(k+j))dt\right)$$
$$= \prod_{l=1}^{\infty}\exp\left(-\sum_{k=1}^{\infty}\frac{(\beta)^k}{k}r^k\exp(ik\lambda)\right) = \prod_{l=1}^{\infty}(1-\beta_l r\exp(i\lambda)),$$

$(I) = \exp\left(2\sum_{j=0}^{\infty}r^j\exp(i\lambda j)\int_{-\pi}^{\pi}\exp(-itj)dt/(2\pi)-1\right)|\beta_0| = |\beta_0|$, and

$$(II) = \exp\left(-1/2\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{l=1}^{\infty}\log(1-\beta_l\exp(it))+\log(1-\bar\beta_l\exp(-it))dt\right)\right)$$
$$= \exp\left(\sum_{l=1}^{\infty}\sum_{k=1}^{\infty}\frac{1}{2k}\left((\beta_l)^k\int_{-\pi}^{\pi}\exp(itk)\mathrm{d}\lambda+(\bar\beta_l)^k\int_{-\pi}^{\pi}\exp(-itk)\mathrm{d}\lambda\right)\right) = 1.$$

Consequently, we get

$$\phi^+(r\exp(i\lambda)) = |\beta_0|\prod_{l=1}^{\infty}(1-\beta_l r\exp(i\lambda)) = |\beta_0|\prod_{l=1}^{\infty}(1-\beta_l z) = \phi(z).$$

Furthermore, we have

$$\exp\left(\frac{1}{2\pi}\int_0^{2\pi}\left(2\sum_{j=0}^{\infty} r^j \exp(i(\lambda-t)j)-1\right)\log|\phi(\exp(it))|dt\right)=\exp\left(a_0/2+\sum_{k=1}^{\infty}a_k(r\exp(i\lambda))^k\right),$$

and with $\beta_0=|\beta_0|=\sigma/\sqrt{(2\pi)}$ the assertion follows.

A bounded spectral density ensures that $\sum_{k=0}^{\infty}c_k$ exists and is bounded. If $\lim_{n\to\infty}\sum_{k=1}^{n}a_k=-\infty$ we consider $\exp(a_0/2+\sum_{k=1}^{\infty}a_k)=0$. Because the spectral density is bounded, $\lim_{n\to\infty}\sum_{k=1}^{n}a_k=\infty$ is not possible. Thus, $\exp(a_0/2+\sum_{k=1}^{\infty}a_k)$ is also well defined and bounded. Consequently, since both expression are continuous in z, the assertion holds in the case of a bounded spectral density for $|z|=1$ as well. □

Lemma 2.10.3. *If the Fourier coefficients of* $\log f$ *fulfill* $a_k=\int_{-\pi}^{\pi}\log(f(\lambda))\exp(-ik\lambda)\mathrm{d}\lambda/(2\pi)=\mathcal{O}(k^{-3})$*, then the coefficients* $(c_j)_{j\geq 0}$ *satisfy* $c_j=\mathcal{O}(j^{-3})$ *and therefore* $\sum_{j=0}^{\infty}|jc_j|<\infty$*.*
If $\log f$ *is twice continuous differentiable and the second derivate is of bounded variation, the condition* $a_k=\mathcal{O}(k^{-3})$ *is fulfilled, see (Zygmund, 2002, Ch. 2, Theorem 4.12).*

Proof. Follows by equation (2.4). □

Notice that for a strictly positive spectral density, the decay behavior of the Fourier coefficients of $\log f$ can be derived by the decay behavior of the autocovariance due to the Wiener-Levy-Theorem, see for instance Bhatt and Dedania (2003).

Bibliography

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21(1):243–247.

Baxter, G. (1962). An asymptotic result for finite predictor. *Math. Scand.*, 10:137–144.

Beltrão, K. I. and Bloomfield, P. (1987). Determining the bandwidth of a kernel spectrum estimate. *Journal of Time Series Analysis*, 8(1):21–38.

Bhansali, J. R. (1974). Asymptotic Properties of the Wiener-Kolmogorov Predictor. I. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1):61–73.

Bhansali, J. R. (1977). Asymptotic Properties of the Wiener-Kolmogorov Predictor. II. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):66–72.

Bhatt, S. and Dedania, H. (2003). Beurling algebra analogues of the classical theorems of wiener and lévy on absolutely convergent fourier series. *Proceedings Mathematical Sciences*, 113(2):179–182.

Billingsley, P. (1968). *Convergence of probability measures*. Wiley Series in probability and Mathematical Statistics: Tracts on probability and statistics. Wiley.

Blackman, R. B. and Tukey, J. W. (1958). The measurement of power spectra from the point of view of communications engineering. *Bell System Technical Journal*, 37.

Bloomfield, P. (1973). An exponential model for the spectrum of a scalar time series. *Biometrika*, 60(2):217–226.

Brockwell, P. and Davis, R. A. (1991). *Time Series: Theory and Methods (2nd edition)*. Springer, New York.

Bühlmann, P. (1995). Moving-average representation of autoregressive approximations. *Stochastic Processes and their Applications*, 60(2):331 – 342.

Bühlmann, P. (1997). Sieve bootstrap for time series. *Bernoulli*, 3(2):123–148.

Epstein, C. L. (2005). How well does the finite fourier transform approximate the fourier transform? *Communications on Pure and Applied Mathematics*, 58(10):1421–1435.

Fragkeskou, M. and Paparoditis, E. (2015). Inference for the Fourth Order Innovation Cumulant in Linear Time Series. *Journal of Time Series Analysis*, 37(2):240–266.

Freitag, E. and Busam, R. (2006). *Funktionentheorie 1*. Funktionentheorie / Eberhard Freitag, Rolf Busam. Springer.

Götze, F. and Künsch, H. R. (1996). Second-order correctness of the blockwise bootstrap for stationary observations. *The Annals of Statistics*, 24(5):1914–1933.

Hoffman, K. (1962). *Banach Spaces of Analytic Functions*. Prentice-Hall series in modern analysis. Prentice-Hall.

Jentsch, C. and Subba Rao, S. (2015). A test for second order stationarity of a multivariate time series. *Journal of Econometrics*, 185(1):124–161.

Jones, R. H. (1964). Spectral analysis and linear prediction of meteorological time series. *Journal of Applied Meteorology*, 3(1):1964.

Kaderli, A. and Kayhan, A. (2000). Spectral estimation of ARMA processes using ARMA-cepstrum recursion. *IEEE Signal Processing Letters*, 7(9):259–261.

Kreiss, J.-P. (1992). *Bootstrap procedures for AR(∞)-processes*, volume 376. Springer, In Bootstrapping and Related Techniques of *Lecture Notes in Economics and Mathematical Systems*. 107–113.

Kreiss, J.-P. and Paparoditis, E. (2012). The Hybrid Wild Bootstrap for Time Series. *Journal of the American Statistical Association*, 107(499):1073–1084.

Kreiss, J.-P., Paparoditis, E., and Politis, D. N. (2011). On the range of validity of the autoregressive sieve bootstrap. *Ann. Statist.*, 39(4):2103–2130.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, 17(3):1217–1241.

Lahiri, S. (2003). *Resampling Methods for Dependent Data*. Springer Series in Statistics. Springer.

Markel, J. (1971). Fft pruning. *IEEE Transactions on Audio and Electroacoustics*, 19(4):305–311.

McMurry, T. L. and Politis, D. N. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *Journal of Time Series Analysis*, 31(6):471–482.

McMurry, T. L. and Politis, D. N. (2017). Estimating ma parameters through factorization of the autocovariance matrix and an ma-sieve bootstrap.

Neumann, M. H. (2013). A central limit theorem for triangular arrays of weakly dependent random variables, with applications in statistics. *ESAIM: Probability and Statistics*, 17:120–134.

Nordman, D. (2009). A note on the stationary bootstrap's variance. *The Annals of Statistics*, 37(1):359–370.

Paparoditis, E. and Politis, D. N. (2001). Tapered block bootstrap. *Biometrika*, 88(4):1105–1119.

Paparoditis, E. and Streitberg, B. (1991). Order identification statistics in stationary autoregressive moving-average models: Vector autocorrelations and the bootstrap. *Journal of Time Series Analysis*, 13(5):415–434.

Politis, D. N. (2003). Adaptive bandwidth choice. *Journal of Nonparametric Statistics*, 15(4-5):517–533.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313.

Pourahmadi, M. (1983). Exact factorization of the spectral density and its application to forecaresting and time series analysis. *Communications in Statistics-Theory and Methods*, 12(18):2085–2094.

Pourahmadi, M. (1984). Taylor expansion of $\exp(\sum_{k=0}^{\infty} a_k z^k)$ and some applications. *The American Mathematical Monthly*, 91(5):303–307.

Pourahmadi, M. (2001). *Foundations of Time Series Analysis and Prediction Theory*. Wiley, Wiley Series in Probability and Statistics.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Romano, J. P. and Wolf, M. (2000). A more general central limit theorem for m-dependent random variables with unbounded m. *Statistics & Probability Letters*, 47(2):115–124.

Romano, J. P. and Wolf, M. (2006). Improved nonparametric confidence intervals in time series regressions. *Nonparametric Statistics*, 18(2):199–214.

Shibata, R. (1981). An optimal autoregressive spectral estimate. *The Annals of Statistics*, 9(2):300–306.

Stoica, P. and Sandgren, N. (2006). Smoothed Nonparametric Spectral Estimation Via Cepstrum Thresholding. *IEEE Signal Processing Magazine*, 34(9):34–45.

Szegö, G. (1921). Über die Randwerte einer analytischen Funktion. *Mathematische Annalen*, 84(3-4):232–244.

Zygmund, A. (2002). *Trigonometric Series Number Bd. 1 in Mathematical Library*. Cambridge University Press, Cambridge.

3 Time Series Modeling on Dynamic Networks

3.1 Introduction

Consider a vertex-labeled network with d vertices $V = \{1, \dots, d\}$. The number of vertices is fixed over time, whereas, the edges are time dependent. Thus, over time edges may vanish or new ones may appear. Throughout this work, directed edges are considered and multi-edges can occur. Such a dynamic network with a fixed number of vertices can be described by a time dependent adjacency matrix, here denoted by $\mathbf{Ad} = \{Ad_t, t \in \mathbb{Z}\}$, where Ad_t is $\mathbb{N}_0^{d\times d}$- valued and $Ad_{t;ij}$ gives the number of edges at time t from vertex i to vertex j. The notation $X_{t;ij}$ is used here for the i-th entry of the j-th row of X_t. It is further possible to take the strength of the connection into account by using a weight function $w : E \to \mathbb{R}$, see Boccaletti et al. (2006). Hence, each edge is given some weight. This can be directly expressed in a process $\{\tilde{A}d_t, t \in \mathbb{Z}\}$, where $\tilde{A}d_t$ is $\mathbb{R}^{d\times d}$-valued and $\tilde{A}d_{t;ij} = w((i,j))Ad_{t;ij}$. Some results require a limited number of multi-edges which we indicate by the notation that Ad_t is $\{0, \dots, l\}-$valued, $l \in \mathbb{N}$. However, weighting does not affect any of the results given in this chapter. Hence, the process $\mathbf{Ad}$ describes here a weighted, directed network. It is considered that the network is driven by some random process, hence, the corresponding adjacency matrix process $\mathbf{Ad}$ is a stochastic process.

Such networks could describe social networks, where some actors (e.g. persons) are represented by the vertices and these actors have some form of relation (e.g. friendship, communication) which is represented by the edges, see for instance Hanneke and Xing (2007). Since these relations could change over time, the corresponding network is dynamic. Social media networks such as *Facebook* or *Twitter* are examples for dynamic networks. The actors in such networks often posses attributes or properties. These attributes can be static (e.g a person's name or birthday) or dynamic (e.g. personal income, time a person does sports or amount of alcohol a person drinks). These dynamic attributes may be affected by the attributes of other actors, especially by actors with which the considered actor is connected. We denote such an attribute a network-influenced attribute. In this work the dynamic attributes are denoted by a d-dimensional time series $\underline{\mathbf{X}} = \{\underline{X}_t, t \in \mathbb{Z}\}$, where each component of the time series is assigned to a vertex (actor) of the underlying network. In the

social-economical literature the influence of connected actors on the attributes is denoted as peer effects, see Goldsmith-Pinkham and Imbens (2013); Manski (1993).

In this work the focus is on the network-influenced attributes and not on the network itself. Consequently, this work is not about modeling a dynamic network. For modeling these dynamic networks, many models for static networks have been extended to the dynamic case as it is done by Hanneke et al. (2010); Krivitsky and Handcock (2014) for the Exponential Random Graph Models (ERGM), see Section 6.5 in Kolaczyk (2009), or by Xu (2015) for the stochastic block model (SBM), see Goldenberg et al. (2010). This work gives a framework which models the network-influenced dynamic attributes, that means modeling a time series on a dynamic network in which the edges influence the dependency of the time series. Knight et al. (2016); Zhu et al. (2017) have considered these network-influenced attributes for non-random edges, which mainly covers static networks. In the context of a static network, network-influenced properties can be considered as an *ordinary* multivariate time series with additional information and can be modeled by using vector autoregressive (VAR) models, see Lütkepohl (2007, Chapter 2). However, VAR models have many parameters which is why Knight et al. (2016); Zhu et al. (2017) focus on how to use the network structure to reduce the number of parameters so that high dimensions become feasible. In contrast, this work deals with a random network structure and consequently the process $\underline{\mathbf{X}}$ cannot be modeled appropriately by using VAR models. That is why we make use of a multivariate doubly stochastic time series framework. That is, we consider linear processes or autoregressive models in which the coefficient matrices are stochastic processes themselves. Doubly stochastic time series models were introduced in Tjstheim (1986). In this work, a slightly different notion more similar to the one of Pourahmadi (1986, 1988) is used.

This chapter is structured as follows. In section 3.2 time series on dynamic networks are defined and some basic properties are given. In section 3.3 the focus is on statistical results; for instance, a central limit theorem for the sample mean is displayed and forecasting with such models is discussed. Some of the forecasting results are underlined by a simulation study which is given in section 3.4. A real data example is given in section 3.5. Proofs can be found in section 3.7.

3.2 Time Series Modeling on Dynamic Networks

Recall that the dynamic network with d vertices is described by the $\mathbb{R}^{d\times d}$-valued stochastic process $\mathbf{Ad} := \{Ad_t, t \in \mathbb{Z}\}$. As mentioned before, the doubly stochastic framework is used to model the time series $\underline{\mathbf{X}} := \{\underline{X}_t, t \in \mathbb{Z}\}$ on the random network $\mathbf{Ad}$. That means, besides some innovation process $\underline{\varepsilon} := \{\underline{\varepsilon}_t, t \in \mathbb{Z}\}$, the time series is also driven by the stochastic process $\mathbf{Ad}$. Furthermore, it is considered that the underlying network $\mathbf{Ad}$ is strictly stationary in order to get a stationary process $\underline{\mathbf{X}}$. Since some interesting features come only into play if non-centered innovations are

considered, the innovations posses a mean $\underline{\mu} \in \mathbb{R}^d$. The induced norms are used as matrix norms, i.e., $\|A\| = \sup_{\|x\|=1} \|Ax\|$. If not stated otherwise, the L_2-norm is used. The main process structure of $\underline{\mathbf{X}}$ is defined as follows:

Definition 3.2.1. *Let* $\mathbf{Ad}$ *be a* $\mathbb{R}^{d\times d}$*-valued, strictly stationary stochastic process and let* $f_j : \mathbb{R}^{(d\times d)j} \to \mathbb{R}^{d\times d}$ *be measurable functions. Furthermore, let* $\underline{\varepsilon}$ *be an i.i.d. sequence of* $\mathbb{R}^d$*-valued random vectors with* $E\underline{\varepsilon}_1 = \underline{\mu} \in \mathbb{R}^d$, $Var\,\underline{\varepsilon}_1 = \Sigma$ *(positive definite and* $\|\Sigma\| < \infty$*), and* $\underline{\varepsilon}$ *and* $\mathbf{Ad}$ *are mutually independent. If the following* L_2*-limes exists,*

$$\underline{X}_t = \sum_{j=1}^{\infty} f_j(Ad_{t-1}, \ldots, Ad_{t-j})\underline{\varepsilon}_{t-j} + \underline{\varepsilon}_t =: \sum_{j=1}^{\infty} B_{t,j}\underline{\varepsilon}_{t-j} + \underline{\varepsilon}_t, \tag{3.2.1}$$

we denote the process given by $\underline{\mathbf{X}} = \{\underline{X}_t, t \in \mathbb{Z}\}$ *a (doubly stochastic) network linear process (DSNLP).*

Let $p, q \in \mathbb{N}$ *and* $f_j : \mathbb{R}^{(d\times d)j} \to \mathbb{R}^{d\times d}, g_s : \mathbb{R}^{(d\times d)s} \to \mathbb{R}^{d\times d}, j = 1, \ldots, p, s = 1, \ldots, q$ *be measurable functions. A process* $\underline{\mathbf{X}}$ *fulfilling equation* (3.2.2) *is denoted as a (doubly stochastic) network autoregressive moving average process of order* (p, q) *(DSNARMA*(p, q)*)*

$$\underline{X}_t = \sum_{j=1}^{p} f_j(Ad_{t-1}, \ldots, Ad_{t-j})\underline{X}_{t-j} + \sum_{s=1}^{q} g_s(Ad_{t-1}, \ldots, Ad_{t-s})\underline{\varepsilon}_{t-s} + \underline{\varepsilon}_t. \tag{3.2.2}$$

The notation $\underline{X}_t = \sum_{j=0}^{\infty} B_{t,j}\underline{\varepsilon}_{t-j}$, where $B_{\cdot,0} \equiv I_d$ and $B_{t,j} = f_j(Ad_{t-1}, \ldots, Ad_{t-j})$, is used to simplify the notation of DSNLP. Notice that $B_{\cdot,j}$ is a stochastic process and independent of $\underline{\varepsilon}$. Defining this with a stochastic process $\mathbf{B}$ not necessarily generated by a random network process leads to doubly stochastic linear processes. For such processes similar results can be established. Since this work focuses on networks, the focus is on doubly stochastic network processes.

There is no single feasible model which covers all kinds of dynamic networks. Instead, there exist several models and each of them is suitable for a specific kind of network. The intuition behind the assumption that $\underline{\varepsilon}$ and $\mathbf{Ad}$ are mutually independent is that the time series can be modeled regardless of what network is underneath. Thus, it does not matter if its a sparse or dense network or if it has properties like small-world-network. In this work, apart from a mixing condition, the dynamic network does not need to fulfill any further conditions. Hence, this assumptions gives flexibility in a way that the time series and the dynamic network can be modeled separately. One is not fixed to one specific network model as it would be the case for a jointly modeling approach. Instead, the idea is that the approach describe here is used to model the time series and one of the several models for dynamic networks can be used to model the network. However, this assumptions is more restrictive. It implies that the influence between the network and the time series $\underline{\mathbf{X}}$ is unidirectional; $\mathbf{Ad}$ can influence $\underline{\mathbf{X}}$, however, $\mathbf{Ad}$ is not influenced by $\underline{\mathbf{X}}$. Some real life examples may violate this assumption. For instance, when considering the influence of peers on obesity,

Christakis and Fowler (2007) or grades, Goldsmith-Pinkham and Imbens (2013), the influence can go both ways. But if shorter periods are considered, the influence of $\underline{\mathbf{X}}$ on **Ad** may not come into play. Furthermore, restricting peers to relatives as it is done in Christakis and Fowler (2007) it seems reasonable to assume that one's properties like obesity or grades do not influence one's relatives, hence $\underline{\mathbf{X}}$ does not influence **Ad**.

If **Ad** is a deterministic sequence the DSNARMA are closely related to time-varying ARMA models, which, for instance, are used in the locally stationary framework; see Dahlhaus et al. (1999); Wiesel et al. (2013). Furthermore, if **Ad** is i.i.d. the doubly stochastic framework reduces to the framework of random coefficient models, see for instance Nicholls and Quinn (1982) and for the multivariate setting Nicholls and Quinn (1981). However, assuming independence between different time-points for the process **Ad**, seems to be inappropriate in the framework of dynamic networks. Some form of influence of the recent history seems to be more reasonable, see Hanneke and Xing (2007). As already mentioned, the focus is not on modeling the network, which is why the dependence structure of the network is not further specified here. However, in order to derive statistical results, the dependence structure needs to be restricted and we consider α-mixing (see section 3.3 for details). Since the innovation process $\underline{\varepsilon}$ and the network process **Ad** are independent, and both are stationary, it is an arbitrary choice at which time point the network process is used to define $\underline{X}_t$. Thus, a process given by $\underline{X}_t = \sum_{j=1}^{p} f_j(Ad_t, \dots, Ad_{t+1-j})\underline{X}_{t-j} + \sum_{s=1}^{q} g_s(Ad_t, \dots, Ad_{t+1-s})\underline{\varepsilon}_{t-s} + \underline{\varepsilon}_t$ has the same properties. The only difference is the interpretation. Thus, when choosing a definition, one has to answer: 'Does the current network determine how recent effects influence the process. Or does the network, which was present when recent effects occurred, determine how recent effects influence the process?' If not stated otherwise, we follow the latter interpretation and use the corresponding Definition 3.2.2. Since directed edges are considered, two natural dependence concepts occur; the concept that the influence goes in direction with the edge and vice versa. The general definition of DSNLP and DSNARMA can handle both concepts, however, the model given by (3.2.3) as well as the models specified in section 3.3 are defined in the sense that the influence goes in edge direction. That means, if social media data such as from *Twitter* is considered, a person j could be influenced by the persons whom j follows. Thus, these persons would have a directed edge to j. It is also possible to define it the other way around, see Wasserman and Faust (1994). However, if $\underline{\mathbf{X}}$ represents flow in a network, such as traffic amount at given locations, it seems more appropriate to define the influence in direction of the flow, thus, of the edges.

Consider the following example for the functions f_j and g_j in Definition (3.2.1). The componentwise multiplication of $\mathbb{R}^{d\times d}$ matrices is denote by $\circledast$, thus, for $A, B \in \mathbb{R}^{d\times d}$, $A \circledast B = (a_{ij}b_{ij})_{i,j=1,\dots,d}$. Let $\alpha_j \in \mathbb{R}^{n\times n}, j = 1, \dots, p, \beta_j \in \mathbb{R}^{n\times n}, j = 1, \dots, q, p, q \in \mathbb{N}$. With $f_j(Ad_{t-1}, \dots, Ad_{t-j}) = (\alpha_j \circledast$

$Ad_{t-j})^\top, j = 1,\dots,p$ and $g_s(Ad_{t-1},\dots,Ad_{t-s}) = (\beta_j \circledast Ad_{t-s})^\top, j = 1,\dots,q$ we get the following (doubly stochastic) network autoregressive moving average process of order (p,q)

$$\underline{X}_t = \sum_{j=1}^{p} (\alpha_j \circledast Ad_{t-j})^\top \underline{X}_{t-j} + \sum_{j=1}^{q} (\beta_j \circledast Ad_{t-j})^\top \underline{\varepsilon}_{t-j} + \underline{\varepsilon}_t \tag{3.2.3}$$

In this model, the 'influence' between components is in direction with the edges and each edge is given a weight. Since $Ad_{t;ij} = 1$ indicates that an edge from i to j is present, we work here with $Ad_t^\top$. This model (3.2.3) is inspired by Knight et al. (2016) and it coincides with the definition of network autoregressive (moving average) process of order (p,q) with neighborhood order 1 for all lags, see Knight et al. (2016). Higher neighborhood orders can be achieved by using more than one adjacency matrix at a time.

In the following Lemma we specify conditions which ensure stationarity of DSNLPs:

Lemma 3.2.2. *Let $\underline{\mathbf{X}}$ be a doubly stochastic linear process as defined in (3.2.1). If*

i) $\sum_{s=0}^{\infty} \left(E|B_{j,s+l}\Sigma B_{0,s}^\top|\right) + \sum_{s_1=0}^{\infty}\sum_{s_2=0}^{\infty} |Cov\left(B_{j,s_1}\underline{\mu}, B_{l,s_2}\underline{\mu}\right)| < \infty$ *(component-wise) for all* $j,l \in \mathbb{N}$

ii) $\sum_{s=0}^{\infty} \left(E|B_{0,s}|\right) < \infty$ *(component-wise),*

is fulfilled, then $\underline{X}_t = \lim_{q\to\infty} \sum_{j=0}^{q} B_{t,j}\underline{\varepsilon}_{t-j}$ *converges component-wise in the L_2-Limit and the autocovariance function is given by* $\Gamma_X(h) = \Gamma_X(-h)^\top$ *and*

$$\Gamma_X(h) = \sum_{s=0}^{\infty} E\left(B_{h,s+h}\Sigma B_{0,s}^\top\right) + \sum_{j=0}^{\infty}\sum_{s=0}^{\infty} Cov\left(B_{h,j}\underline{\mu}, B_{0,s}\mu\right), h \geq 0 \tag{3.2.4}$$

and the mean function by $\underline{\mu}_x = \sum_{j=0}^{\infty} EB_{0,j}\underline{\mu}$.

The latter term of the autocovariance function, $\sum_{j=0}^{\infty}\sum_{s=0}^{\infty} \mathrm{Cov}\left(B_{h,j}\underline{\mu}, B_{0,s}\mu\right)$, comes only into play for non-centered innovations and is driven by the linear dependency structure of the network. Consequently, it can be seen that the linear dependency of the network directly influences the linear dependency of the process $\underline{\mathbf{X}}$. As a consequence, even an DSNMA(q) process may posses a nonzero autocovariance for lags higher than q. In order to better understand this, consider a small toy example with three vertices and two possible edges, $(1,3)$ and $(2,3)$, and only one is present at a time. Let $\{e_t, t \in \mathbb{Z}\}$ be i.i.d. random variables with uniform distribution on $[0,1]$, i.e., $e_1 \sim \mathcal{U}[0,1]$. Which edge is present at time t is given by the random variables (e_t) in the following way. If $Ad_{t-1;13} = 1$, then if $e_t > 0.05$, then $Ad_{t;13} = 1$ else $Ad_{t;23} = 1$. If $Ad_{t-1;13} = 0$ (that means $Ad_{t-1;23} = 1$), then if $e_t > 0.95$, then $Ad_{t;13} = 1$ else $Ad_{t;23} = 1$. Consequently, in this network we flip between the edges $(1,3)$ and $(2,3)$ and if one edge is present at time t it is more likely (with probability 0.95) that it is

present at time $t+1$ than flipping to the other edge. We have dependency between different time points as well as between edges. $\underline{\varepsilon}_1 \sim \mathcal{N}(\underline{\mu}, I_3)$, and $\underline{\mu} = (10, -10, 0)^\top$. Let $\underline{\mathbf{X}}$ be given by

$$\underline{X}_t = (Ad_{t-1})^\top \underline{\varepsilon}_{t-1} + \varepsilon_t = (Ad_{t-1})^\top \underline{X}_{t-1} + \underline{\varepsilon}_t, \text{ where } Ad_\cdot^\top = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ * & * & 0 \end{pmatrix}. \tag{3.2.5}$$

Thus, $\underline{\mathbf{X}}$ is a DSNMA(1) process and the influence goes in direction with the edges. Since no edge goes into vertex 1 or 2, $\{\underline{X}_{t;1}, t \in \mathbb{Z}\}$ and $\{\underline{X}_{t;2}, t \in \mathbb{Z}\}$ are white noise. This can be also seen in the autocovariance function which is displayed in its two parts in Figure 3.1. The left-hand-side figure displays the first part; $\sum_{s=0}^{\infty} E\left(B_{h,s+h} \Sigma B_{0,s}^\top\right)$. The dependency of the network has no influence on the first part, thus, this part would remain the same if **Ad** is replaced by its expected value. That is why this part of the autocovariance function has the structure one expects from a vector moving average (VMA) process of order 1. The right-hand-side figures display the latter part of the autocovariance function; $\sum_{j=0}^{\infty} \sum_{s=0}^{\infty} \mathrm{Cov}\left(B_{h,j}\underline{\mu}, B_{0,s}\mu\right)$. As already mentioned, this part is completely driven by the linear dependence structure of the network. For the two edges, we have the following linear dependency: $\mathrm{Cov}(Ad_{t+h;23}, Ad_{t;23}) = \mathrm{Cov}(Ad_{t+h;13}, Ad_{t;13}) = 0.9^h/4, \mathrm{Cov}(Ad_{t+h;23}, Ad_{t;13}) = \mathrm{Cov}(Ad_{t+h;13}, Ad_{t;23}) = -0.9^h/4$. This explains the geometric decay in the autocovariance function of the third component of $\underline{\mathbf{X}}$, whereas the absolute value of the autocovariance function of the third component is mainly given by the difference of the mean of the innovations of the first two components. Hence, a greater difference of the innovations mean makes it harder to identify the linear dependency between components 1 and 3, or 2 and 3 respectively. In this particular example with mean $\underline{\mu} = (10, -10, 0)^\top$, no linear dependency between the different components can be identified for moderate sample sizes. A sample autocorrelation function as well as a realization of the third component of $\underline{\mathbf{X}}$ is displayed in Figure 3.2 for a sample size $n = 500$. Instead, looking from the perspective of the classical time series analysis, the sample autocorrelation function looks like three uncorrelated components where the first two components are white noises and the third could be an AR(1) process. Hence, this examples gives two important aspects to keep in mind: Firstly, the linear dependency of the network can influence the linear dependency of the time series directly. Secondly, the problem that the autocovariance function may not suffice to identify doubly stochastic network models such as DSNAR(1).

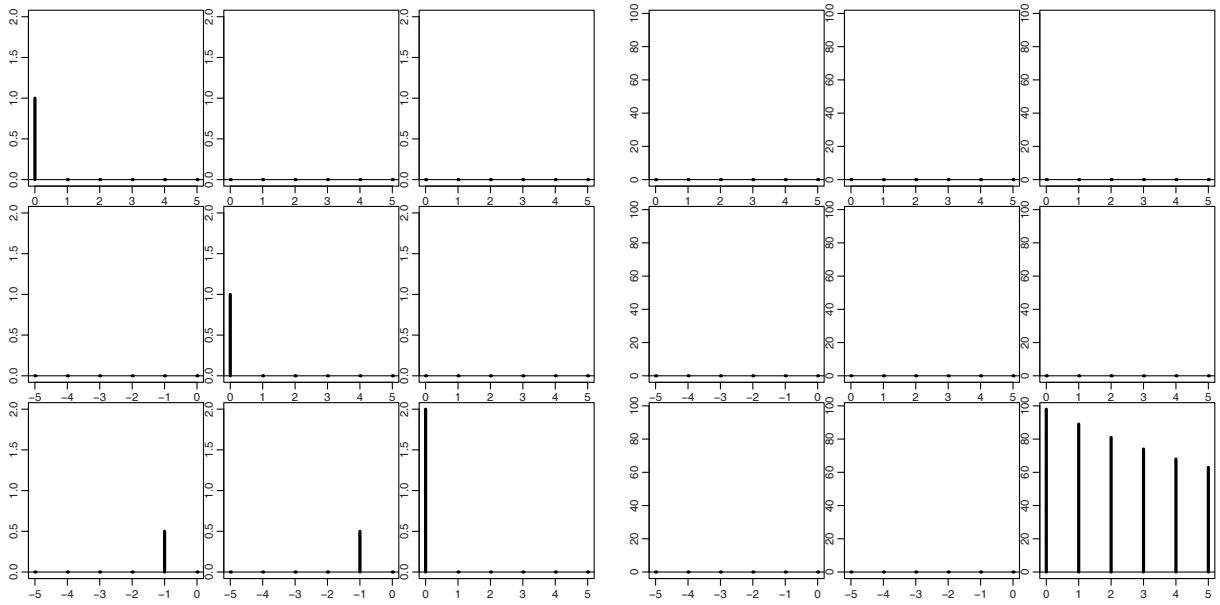


Figure 3.1: Left-hand-side $\left(\sum_{s=0}^{\infty} E(B_{h,s+h}\Sigma B_{0,s}^{\top})\right.$; left figure) and right-hand-side $\left(\sum_{j=0}^{\infty}\sum_{s=0}^{\infty} \mathrm{Cov}(B_{h,j}\mu, B_{0,s}\mu)\right.$; right figure) of the autocovariance function (3.2.4) of process (3.2.5)

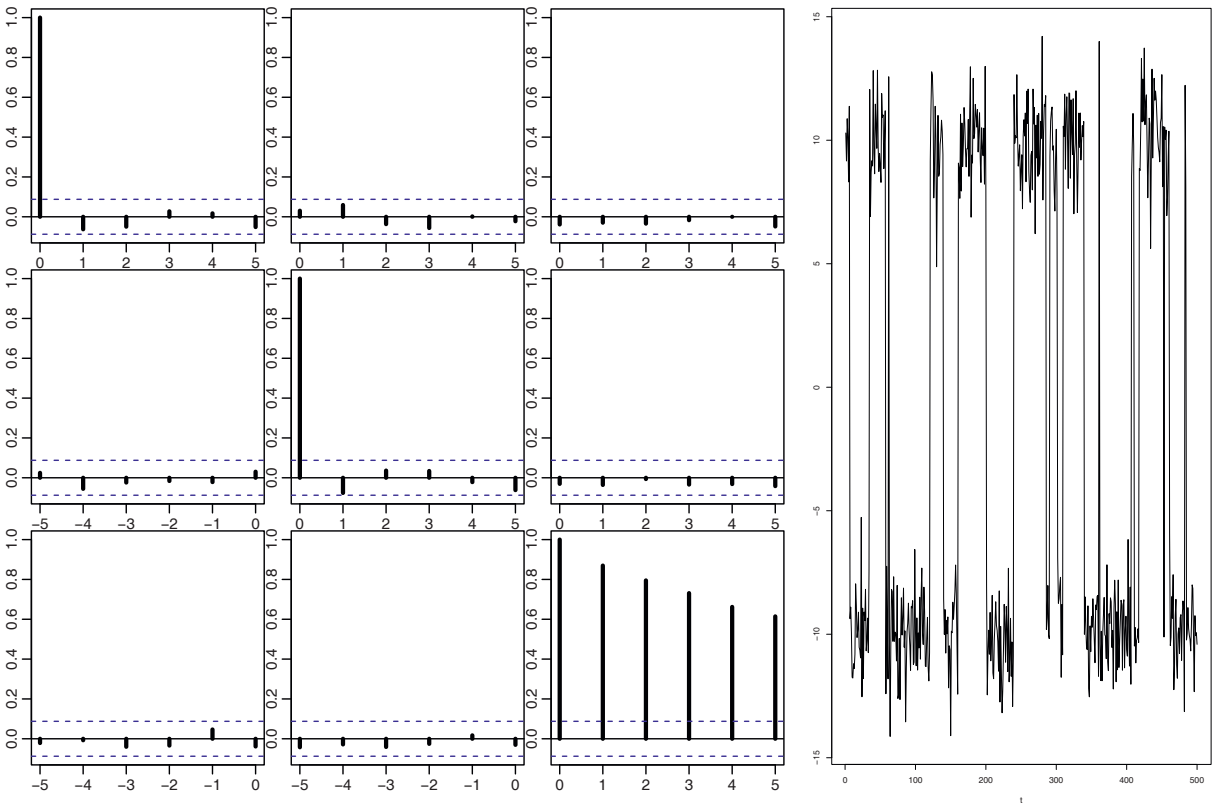


Figure 3.2: Sample autocorrelation function and realization of the third component of process (3.2.5), based on $n = 500$

In order to give conditions under which there exists a solution of (3.2.2), we firstly consider a DSARMA(1,0), a doubly stochastic autoregressive process of order 1, given by $\underline{X}_t = f(Ad_{t-1})\underline{X}_{t-1} + \underline{\varepsilon}_t$. In the univariate case Pourahmadi (1988) gives conditions for the existence of a stationary solution of such processes. We transfer his ideas to the multivariate case in the following lemma:

Lemma 3.2.3 (Multivariate Version of (Pourahmadi, 1988, Lemma. 2.1))**.** *Consider a doubly stochastic autoregressive process of order* 1*, thus, we have*

$$\underline{X}_t = f(Ad_{t-1})\underline{X}_{t-1} + \underline{\varepsilon}_t =: A_{t-1}\underline{X}_{t-1} + \underline{\varepsilon}_t, E\underline{\varepsilon}_1 = \underline{\mu}, Var\underline{\varepsilon}_1 = \Sigma. \tag{3.2.6}$$

A stationary solution of (3.2.6) *is given by*

$$\underline{X}_t = \sum_{j=0}^{\infty}\left[\prod_{s=1}^{j} A_{t-s}\right]\underline{\varepsilon}_{t-j} =: \sum_{j=0}^{\infty} B_{t,j}\underline{\varepsilon}_{t-j}, \tag{3.2.7}$$

where $B_{t,0} = \prod_{s=1}^{0} A_{t-s} := I_d$, $B_{t,j} = \prod_{s=1}^{j} A_{t-s} \in \mathbb{N}$, *if (component-wise)*

$$\sum_{j=0}^{\infty} E|B_{0,j}| = \sum_{j=0}^{\infty} E\left|\prod_{s=1}^{j} A_{-s}\right| < \infty, \tag{3.2.8}$$

$$\sum_{j=0}^{\infty} E|B_{0,j}\Sigma_d B_{0,j}^{\top}| = \sum_{j=0}^{\infty} E\left[\left|\prod_{s=1}^{j} A_{-s}\Sigma_d\left(\prod_{s=1}^{j} A_{-s}\right)^{\top}\right|\right] < \infty. \tag{3.2.9}$$

The mean function is then given by $\sum_{j=0}^{\infty} E\prod_{s=1}^{j} A_{-s}\underline{\mu} = \sum_{j=0}^{\infty} EB_{0,t}\underline{\mu}$ *and the ACF is given by*

$$\Gamma_X(h) = \sum_{j=0}^{\infty} E\left[\left(\prod_{s=1}^{j+h} A_{h-s}\right)\Sigma\left(\prod_{s_2=1}^{j} A_{-s}\right)^{\top}\right] + \sum_{j_1=0}^{\infty}\sum_{j_2=0}^{\infty} Cov\left(\prod_{s_1=1}^{j_1} A_{-s_1}\underline{\mu}, \prod_{s_2=1}^{j_2} A_{-s_2}\underline{\mu}\right), h \geq 0,$$

$\Gamma_X(h) = \Gamma_X(-h)^{\top}$. *The solution 3.2.7 fits into the framework of* (3.2.1).

The conditions (3.2.8) and (3.2.9) may not be easy to check. That is why the following Lemma gives conditions which ensure (3.2.8) and (3.2.9):

Lemma 3.2.4. *Let* $A_t = f(Ad_t)$ *and* $\mathbf{Ad}$ *is* α*-mixing. If there exists a* $q \geq 1$ *such that*

$$E\log\|\prod_{s=1}^{q} A_{-s}\| < 0, \tag{3.2.10}$$

then (3.2.8) *and* (3.2.9) *is fulfilled, hence* $\sum_{j=0}^{\infty} E\left|\prod_{s=1}^{j} A_{-s}\right| < \infty$ *and*

$$\sum_{j=0}^{\infty} E\left[\left|\prod_{s=1}^{j} A_{-s}\Sigma_d\left(\prod_{s=1}^{j} A_{-s}\right)^{\top}\right|\right] < \infty.$$

In the same manner as a VAR(p) model can be written as an extended VAR(1) model, see (Lütkepohl, 2007, p. 15), a d-dimensional DSNAR(p) model can be written as a $d \times p$-dimensional DSNAR(1) model. Consider a DSNAR(p) model given by $\underline{X}_t = \sum_{j=1}^{p} f_j(Ad_{t-1}, \ldots, Ad_{t-j})\underline{X}_{t-j}$. Then define $Y_t = (\underline{X}_t, \ldots, \underline{X}_{t-p+1})^\top, \varepsilon_t' = (\underline{\varepsilon}_t, 0, \ldots, 0), \tilde{A}d_t = (Ad_t, \ldots, Ad_{t-p+1})^\top$ and

$$g(\tilde{A}d_{t-1}) =: \begin{pmatrix} f_1(Ad_{t-1}) & f_2(Ad_{t-1}, Ad_{t-2}) & \ldots & f_p(Ad_{t-1}, \ldots, Ad_{t-p}) \\ I_d & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & I_d & 0 \end{pmatrix},$$

such that $Y_t = g(\tilde{A}d_{t-1})Y_{t-1} + \varepsilon_t'$. We denote the process (Y_t) as the stacked process. Thus, the results for DSNAR(1) models can be transfered to DSNAR(p) models. That is why the focus in this work is on DSNAR(1) models.

Some note to the condition (3.2.10). Consider the simplify setting that (Ad_t) is deterministic, thus we consider a simple VAR(p) process $X_t = \sum_{j=1}^{p} A_j X_{t-j} + \varepsilon_t$. Let the considered VAR(p) process be stable, which is given if $\det(I - \sum_{j=1}^{p} A_j z^j) \neq 0$ for all $|z| \leq 1$, see Chapter 2 in Lütkepohl (2007). Let $\tilde{A}$ be the coefficient matrix of a stacked VAR(1) process. For a stable VAR(p) process we have that all eigenvalues of $\tilde{A}$ have modulus less than 1, see Chapter 2 in Lütkepohl (2007). However, for such a stacked coefficient matrix we have that $\|A\| \not\leq 1$, see Lemma E.2 in Basu and Michailidis (2015). But, there exists a $q \geq 1$ such that $\|A^q\| < 1$. To see this, let $Q\Lambda Q^{-1} = \tilde{A}$ be the Jordan canonical form. Furthermore, we have $A^q = Q\Lambda^q Q^{-1}$ and $\|\Lambda^q\| = O(\lambda_1^q)$, where $\lambda_1 < 1$ is the greatest absolute eigenvalue of $\tilde{A}$, see Appendix A.6 in Lütkepohl (2007) for a representation of Λ^q. Since $\|A^q\| \leq \|Q\|\|Q^{-1}\|\|\Lambda^q\| = O(\lambda_1^q)$, there exists a q such that $\|A^q\| < 1$. Consequently, Lemma 3.2.4 is not limited to DSNAR(1) processes and can be also applied to stacked DSNAR(p) models.

3.3 Statistical Results for Doubly Stochastic Network Processes

Lemma 3.2.2 gives conditions for the existence of the ACF and the mean function. In the following passage we are interested in estimating these quantities based on observation $\underline{X}_1, \ldots, \underline{X}_n$. Since the dependency of $\mathbf{Ad}$ influences the dependency of process $\underline{\mathbf{X}}$, conditions for the dependency of $\mathbf{Ad}$ are required to ensure, for instance, an absolutely summable ACF. In order to include many dynamic network models, we are working with an α-mixing condition of the dependency of $\mathbf{Ad}$. This, for instance, includes Markovian dynamic networks, see (Bradley, 2007, Theorem 21.22), such as Temporal ERGMs, see Hanneke et al. (2010). Under the condition that the network process is α-mixing, consistency and asymptotic normality of the sample mean are shown in the following theorem.

Theorem 3.3.1. *Let $\underline{X}_t = \sum_{j=0}^{\infty} B_{t,j}\underline{\varepsilon}_{t-j}$ be an $\mathbb{R}^{d\times d}$-valued doubly stochastic network linear process, with the following assumptions*

1. *$B_0 = I_d$, $B_{t,j} = f_j(Ad_t, \ldots, Ad_{t-j-1})$, $f_j : \mathbb{R}^{(d\times d)^j} \to \mathbb{R}^{d\times d}$ and f_j measurable, $j \in \mathbb{N}$. **Ad** is a strictly stationary, $\mathbb{R}^{d\times d}$-valued, α-mixing process fulfilling $\sum_{n=1}^{\infty} \alpha(\mathbf{Ad}, n)n^3 < \infty$.*
2. *The innovations $\underline{\varepsilon}$ are $\mathbb{R}^d$-valued and i.i.d. with $E\underline{\varepsilon}_0 = \underline{\mu}$, $Cov(\underline{\varepsilon}_0, \underline{\varepsilon}_0) = \Sigma$, $E|\varepsilon_{0;i_1}\varepsilon_{0;i_2}\varepsilon_{0;i_3}\varepsilon_{0;i_4}| = \kappa_{4;i_1,i_2,i_3,i_4} < \infty$ for all $i_1, \ldots, i_4 = 1, \ldots, d$. $\underline{\varepsilon}$ and **Ad** are independent.*
3. *$E|B_{0,j;i_1,i_2}|^4 < \infty$ for all $j \in \mathbb{N}, i_1, i_2 = 1, \ldots, d$,*
   $\sum_{h\in\mathbb{Z}} \sum_{s_1,s_2=0}^{\infty} |Cov(B_{h,s_1}\underline{\mu}, B_{0,s_2}\underline{\mu})| + \sum_{s=0}^{\infty} |EB_{h,s+h}\Sigma_d B_{0,s}^{\top}| < \infty$ (component-wise).

Then, the autocovariance function is absolutely summable and is given by $\Gamma_X(h) = \sum_{s=0}^{\infty} E\left(B_{h,s+h}\Sigma B_{0,s}^{\top}\right) + \sum_{j=0}^{\infty}\sum_{s=0}^{\infty} Cov\left(B_{h,j}\underline{\mu}, B_{0,s}\underline{\mu}\right), h \geq 0, \Gamma(h) = \Gamma(-h)^{\top}$. The mean function is given by $\underline{\mu}_x = \sum_{j=0}^{\infty} EB_{0,j}\underline{\mu}$. Consider observations $\underline{X}_1, \ldots, \underline{X}_n$. We have, as $n \to \infty$,

$$\sqrt{n}\left(\frac{1}{n}\sum_{t=1}^{n}\underline{X}_t - \underline{\mu}_X\right) = \sqrt{N}\left(\bar{X}_n - \underline{\mu}_X\right) \xrightarrow{D} \mathcal{N}\left(\underline{0}, \sum_{h\in\mathbb{Z}}\Gamma_X(h)\right). \tag{3.3.1}$$

In the context of single stochastic linear processes, Assumption 3 is similar to the assumption of dealing with linear processes with absolutely summable coefficients. If the process **Ad** in Assumption 1 is a Markov process, then the mixing condition is fulfilled under moderate conditions, see for instance Theorem 21.22 in Bradley (2007) and notice that $\phi(\mathcal{A}, \mathcal{B}) \geq \alpha(\mathcal{A}, \mathcal{B})$ for some σ-fields $\mathcal{A}, \mathcal{B}$. Under similar conditions as in Theorem 3.3.1, $\sqrt{n}$-consistency of the sample autocovariance can be derived.

Theorem 3.3.2. *Let $\underline{X}_t = \sum_{j=0}^{\infty} B_{t,j}\underline{\varepsilon}_{t-j}$ be an $\mathbb{R}^d$-valued doubly stochastic network linear process, with the following assumptions*

1. *$B_0 = I_d$, $B_{t,j} = f_j(Ad_t, \ldots, Ad_{t-j-1})$, $f_j : \mathbb{R}^{(d\times d)^j} \to \mathbb{R}^{d\times d}$ and f_j measurable, $j \in \mathbb{N}$. **Ad** is a strictly stationary, $\mathbb{R}^{d\times d}$-valued, α-mixing process fulfilling $\sum_{n=1}^{\infty} \alpha(\mathbf{Ad}, n)^{1/5} < \infty$.*
2. *The innovations $\underline{\varepsilon}$ are $\mathbb{R}^d$-valued and i.i.d. with $E\underline{\varepsilon}_0 = \underline{\mu}$, $Cov(\underline{\varepsilon}_0, \underline{\varepsilon}_0) = \Sigma$, $E|\varepsilon_{0;i_1}\varepsilon_{0;i_2}\varepsilon_{0;i_3}\varepsilon_{0;i_4}| = \kappa_{4;i_1,i_2,i_3,i_4} < \infty$ for all $i_1, \ldots, i_4 = 1, \ldots, d$. $\underline{\varepsilon}$ and **Ad** are independent.*
3. *$E|B_{0,j;i_1,i_2}|^4 < \infty$ for all $j \in \mathbb{N}, i_1, i_2 = 1, \ldots, d$,*
   $\sum_{h\in\mathbb{Z}} \sum_{s_1,s_2=0}^{\infty} |Cov(B_{h,s_1}\underline{\mu}, B_{0,s_2}\underline{\mu})| + \sum_{s=0}^{\infty} |EB_{h,s+h}\Sigma_d B_{0,s}^{\top}| < \infty$ (component-wise).
4. *$\sum_{s=0}^{\infty}\left(E\left[e_i^{\top}(B_{h,s}\mu - EB_{0,s}\mu)\right]^5\right)^{1/5} + \sum_{s=1}^{\infty} s\left(E\left[e_i^{\top}(B_{h,s}\underline{\varepsilon}_1 - EB_{0,s}\mu)\right]^4\right)^{1/4} < \infty$ for all $i = 1, \ldots, d, h \in \mathbb{Z}$.*

Then, given observations $\underline{X}_1, \ldots \underline{X}_n$, the sample autocovariance function $\widehat{\Gamma}(h) = 1/n\sum_{t=1}^{n-h}(\underline{X}_{t+h} - \bar{X}_n)(\underline{X}_t - \bar{X}_n)^{\top}$, where $\bar{X}_n = 1/n\sum_{t=1}^{n}\underline{X}_t$, is a consistent estimator, we have $\widehat{\Gamma}(h) = \Gamma(h) + \mathcal{O}_P(n^{-1/2})$.

Theorem 3.3.2 gives a consistent estimator for the autocovariance function which helps to identify VAR models, however, as seen in the example in section 3.2, the autocovariance function is not helpful to identify DSNAR models. Nevertheless, in order to forecast doubly stochastic network processes, an estimation of DSNAR models seems helpful. That is why in the following passage the focus is on deriving consistent estimators for DSNAR(1) models. Hence, a DSNAR(1) processes as defined in (3.2.2) and given by $\underline{\mathbf{X}} = f(Ad_{t-1})\underline{X}_{t-1} + \underline{\varepsilon}_t$ is considered. Notice that even if **Ad** is Markovian a DSNAR(1) processes can generally not be written as a Hidden Markov models (HMM). This is because $\underline{\mathbf{X}}$ given **Ad** is not a sequence of conditionally independent variables and cannot be written as a noisy functional of Ad_{t-1} only, which is required by a HMM (see Bickel et al. (1998) for details to HMM). Consequently, techniques used for HMM cannot be applied here. Instead, the same setting as in Zhu et al. (2017) is considered, thus, the process $\underline{\mathbf{X}}$ as well as the network **Ad** is observed; we have observations $\underline{X}_1, \dots, \underline{X}_n$ and $Ad_1, \dots, Ad_n$. A DSNAR(1) model is given by the measurable function $f : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ and the mean of the innovations μ. We consider three different parametrization-settings for f. Ranging from seeing f as an arbitrary function to the setting for all edges common parameters. We start here with the general setting that f is an arbitrary measurable function. This may sound like an nonparametric setting, however if we consider the case that the number of multi-edges is limited then the process **Ad** is discrete and bounded. That is why $f(\mathbf{Ad})$ has only a finite number N of possible states. Let the possible states of **Ad** be denoted by $\tilde{A}d_1, \dots, \tilde{A}d_N$ and $f(\tilde{A}d_k) =: \alpha_k \in \mathbb{R}^{d\times d}, k = 1, \dots, N$, which reduces the problem to a parametric one. However, the number of parameters can be challenging — we will come back to this later. Let $R^k = \{r \in \{0, \dots, n-1\} : Ad_r = \tilde{A}d_k\}$ be the set of indices at which time points the state $\tilde{A}d_k$ is observed. Then we have $\underline{X}_{t+1;j} = \sum_{s=1}^{d} \alpha_{k;js}\underline{X}_{t;s} + \underline{\varepsilon}_{t;j} = \alpha_{k;j\cdot}\underline{X}_t + \underline{\varepsilon}_{t;j}, t \in R^k, j = 1, \dots, d$. The least squares approach is used to derive consistent estimators for $\alpha_{k;j\cdot}$ as well as $E\underline{\varepsilon}_{t;j} = \underline{\mu}_j$. Hence, we have

$$\operatorname{argmin}_{\hat{\mu}_j, \hat{\alpha}_{k;j\cdot}} \sum_{t\in R^k} (\underline{X}_{t+1;j} - \hat{\mu}_j - \hat{\alpha}_{k;j\cdot})^2, j = 1, \dots, d, k = 1, \dots, N. \tag{3.3.2}$$

This leads to the following linear system:

$$\sum_{t\in R^k} \begin{pmatrix} \underline{X}_{t+1;j}\underline{X}_t \\ \underline{X}_{t+1;j} \end{pmatrix} = \sum_{t\in R^k} \begin{pmatrix} \underline{X}_t & \underline{X}_t\underline{X}_t^\top \\ 1 & \underline{X}_t^\top \end{pmatrix} \begin{pmatrix} \hat{\mu}_j \\ \hat{\alpha}_{k;j\cdot} \end{pmatrix}.$$

By one elementary operation and denoting $|R^k|^{-1}\sum_{t\in R^k} =: \tilde{\sum}_t$, we have

$$\begin{pmatrix} \tilde{\sum}_t \underline{X}_{t+1;j}\underline{X}_t - \tilde{\sum}_t \underline{X}_{t+1;j}\tilde{\sum}_t \underline{X}_t \\ \tilde{\sum}_t \underline{X}_{t+1;j} \end{pmatrix} = \begin{pmatrix} 0 & \tilde{\sum}_t \underline{X}_t\underline{X}_t^\top - \tilde{\sum}_t \underline{X}_t \tilde{\sum}_t \underline{X}_t^\top \\ 1 & \tilde{\sum}_t \underline{X}_t^\top \end{pmatrix} \begin{pmatrix} \hat{\mu}_j \\ \hat{\alpha}_{k;j\cdot} \end{pmatrix} \tag{3.3.3}$$

As can be seen in (3.3.3), the method to derive consistent estimators for $\hat{\mu}_j, \hat{\alpha}_{k;j\cdot}$ is to estimate a somehow localized version of the autocovariance function. Define the conditional covariance as $\mathrm{Cov}(\underline{X}_{t+h}, \underline{X}_t | Ad_t = \tilde{A}d_k) = E\left[(\underline{X}_{t+h} - E(\underline{X}_{t+h}|Ad_t = \tilde{A}d_k))(\underline{X}_t - E(\underline{X}_t|Ad_t = \tilde{A}d_k))^\top | Ad_t = \tilde{A}d_k\right] =: \Gamma_{\underline{X}_t|Ad_t=\tilde{A}d_k}(h)$, where $E(X|Ad_t = \tilde{A}d_k) = E(X\mathbb{1}_{Ad_t=\tilde{A}d_k})/P(Ad_t = \tilde{A}d_k)$ if $P(Ad_t = \tilde{A}d_k) > 0$, else 0. Following similar ideas and conditions as used in Theorem 3.3.2, as $n \to \infty$ and convergence is meant in probability, we obtained

$$
\begin{aligned}
|R^k|^{-1} \sum_{t \in R^k} \underline{X}_{t+1;j} &\to E[\underline{X}_1 | Ad_0 = \tilde{A}d], \\
|R^k|^{-1} \sum_{t \in R^k} \underline{X}_{t;j} &\to E[\underline{X}_0 | Ad_0 = \tilde{A}d], \\
|R^k|^{-1} \sum_{t \in R^k} \underline{X}_t \underline{X}_t^\top - |R^k|^{-2} \sum_{t_1,t_2 \in R^k} \underline{X}_{t_1} \underline{X}_{t_2}^\top &\to \mathrm{Cov}(\underline{X}_0, \underline{X}_0 | Ad_0 = \tilde{A}d_k), \\
|R^k|^{-1} \sum_{t \in R^k} \underline{X}_{t+1;j} \underline{X}_t - |R^k|^{-2} \sum_{t_1,t_2 \in R^k} \underline{X}_{t_1+1;j} \underline{X}_{t_2} &\to \mathrm{Cov}(\underline{X}_0, \underline{X}_{1;j} | Ad_0 = \tilde{A}d_k).
\end{aligned}
$$

Consistency of the estimators $(\hat{\mu}_j, \hat{\alpha}_{k;j\cdot})$ follows by using similar ideas as in Theorem 3.3.3. However, the question about the number of parameters remains. For each state a $d \times d$ matrix needs to be estimated and the number of states can be enormous, $N = \mathcal{O}((l+1)^{d^2})$, where l denotes the number of multi-edges. Even for moderate networks this approach becomes soon infeasible. That is why in order to reduce the possible number of parameters, more structure is imposed on f. In the above setting f is an arbitrary function. That means, for a given vertex any change in the network could have a direct effect on the dependence structure of the corresponding time series. However, it may seem reasonable to limit the effects in such a way that only changes 'close' to the given vertex may have a direct effect on the dependence structure of the corresponding time series. For instance, lets say we have d persons, $P_1, \ldots, P_d$, who are in this example the vertices and each day is a time point. If two persons talk to each other at a given day, an edge between the corresponding vertices is drawn. A property at time t of persons P_j might be directly influenced by the persons with whom P_j talked at time $t-1$. However, it may not affect P_j's property at time t if two other persons talked to each other at time $t-1$. Of course, this may affect the property of these persons at time t which may affect P_j's property at time $t+1$. This leads to the following assumptions: The time series corresponding to vertex j is only influenced by j's in- and out-edges. This means that the j-th component is only influenced by the j-th row and j-th column of the adjacency matrix process which results in the following structure for f:

$$
f(X) = (g_1(X_{1\cdot}, X_{\cdot 1}), \ldots, g_d(X_{d\cdot}, X_{\cdot d}))^\top, g_j : \mathbb{R}^{2d} \to \mathbb{R}^{1\times d}. \tag{3.3.4}
$$

With these assumptions the DSNAR(1) models reads as follows

$$\underline{X}_{t;j} = g_j(Ad_{t-1;j\cdot}, Ad_{t-1;\cdot j})\underline{X}_{t-1} + \underline{\varepsilon}_{t;j}, j = 1, \dots, d, t \in \mathbb{Z}. \tag{3.3.5}$$

In order to estimate g_j, the same ideas used to estimate f under the general setting can be applied. However, for estimating g_j it is only necessary to condition on the same state of the j-th row and column of the adjacency matrix. This limits the number of possible states to which it is necessary to condition on. Let $d^*_{\text{out}}(j) = \sum_{s=1}^{d} \mathbb{1}_{\{\sup_{t\in\mathbb{Z}} |Ad_{t;js}|>0\}} - \mathbb{1}_{\{\inf_{t\in\mathbb{Z}} |Ad_{t;js}|>0\}}$ and $d^*_{\text{in}}(j) = \sum_{s=1}^{d} \mathbb{1}_{\{\sup_{t\in\mathbb{Z}} |Ad_{t;sj}|>0\}} - \mathbb{1}_{\{\inf_{t\in\mathbb{Z}} |Ad_{t;sj}|>0\}}$ be the maximal changes in in-degree and out-degree respectively. The number of parameters for component j is given by $\alpha_{j,1}, \dots, \alpha_{j,N_j} \in \mathbb{R}^d$, $N_j = \mathcal{O}(l^{d^*_{\text{out}}(j)+d^*_{\text{in}}(j)})$. For moderate-varying networks this could reduce the number of parameters dramatically. Larger networks usually come together with some form of sparsity, see for instance examples in Section 3.5 in Kolaczyk (2009). This means that a vertex has only a connection to a small fraction of the other vertices. Consequently, a further reasonable assumption could be to assume that only connections have an influence on the vertex' time series. In the example with persons $P_1, \dots, P_d$ this means that P_j's property at time t is only affected by persons to whom P_j talked at time $t-1$; but not by persons to whom P_j did not talk at time $t-1$. Thus, for state $Ad_{k;j\cdot}, Ad_{k;\cdot j}$ define $S^j_k = \{s \in \{1, \dots, d\} : \tilde{A}d_{k;js} \neq 0 \text{ or } \tilde{A}d_{k;sj} \neq 0\}$, then the assumptions reads as

$$g_{j;s}(Ad_{k;j\cdot}, Ad_{k;\cdot j}) = 0 \text{ for all } s \notin S^j_k. \tag{3.3.6}$$

Especially for sparse networks, this highly reduces the number of parameters. We have $\sum_{j=1}^{d} \sum_{k=1}^{N_j} |S^j_k| \leq \sum_{j=1}^{d} ((d_{\text{out}}(j) + d_{\text{in}}(j))^{d^*_{\text{out}}(j)+d^*_{\text{in}}(j)})$ parameters for f and d for $\underline{\mu}$. As mentioned above, consistency of the estimators can be derived in the same manner without these assumptions leading to the general setting discussed above. However, since the general settings is usually infeasible, it is not presented in a theorem here. The following theorem summarizes the results under the assumptions (3.3.4) and (3.3.6):

Theorem 3.3.3. *Let $\underline{\mathbf{X}}$ be a DSNAR(1) given by $\underline{X}_t = f(Ad_{t-1})\underline{X}_{t-1} + \underline{\varepsilon}_t, t \in \mathbb{Z}$, and $X_0, \dots, X_n$ and $Ad_0, \dots, Ad_{n-1}$ are observed. Furthermore, let $(\tilde{A}d_{k;j\cdot}, \tilde{A}d_{k;\cdot j}), j \in \{1, \dots, d\}$ be a given state of the adjacency matrix process $\mathbf{Ad}$ with $\sum_{t=0}^{n-1} \mathbb{1}_{\{Ad_{t;j\cdot}=\tilde{A}d_{k;j\cdot}, Ad_{t;\cdot j}=\tilde{A}d_{k;\cdot j}\}} > 0$. The process $\underline{\mathbf{X}}$ fulfills the following conditions:*

i) $\mathbf{Ad}$ *is a strictly stationary, $\{0, \dots, l\}^{d\times d}$-valued, $l \in \mathbb{N}$ fixed, α-mixing process fulfilling $\sum_{n=1}^{\infty} \alpha(\mathbf{Ad}, n)^{1/5} < \infty$.*

ii) The innovations $\underline{\varepsilon}$ are $\mathbb{R}^d$-valued and i.i.d. with $E\underline{\varepsilon}_0 = \underline{\mu}$, $Cov(\underline{\varepsilon}_0, \underline{\varepsilon}_0) = \Sigma_d$, $E\underline{\varepsilon}_{0;i_1}\underline{\varepsilon}_{0;i_2}\underline{\varepsilon}_{0;i_3}\underline{\varepsilon}_{0;i_4} = \kappa_{4;i_1,i_2,i_3,i_4} < \infty$ for all $i_1, \dots, i_4 = 1, \dots, d$. $\underline{\varepsilon}$ and $\mathbf{Ad}$ are independent.

iii) *Set* $B_0 = I_d$, $B_{t,j} = \prod_{s=1}^{j} f(Ad_{t-s})$. $E|B_{0,j;i_1,i_2}|^4 < \infty$ *for all* $j \in \mathbb{N}, i_1, i_2 = 1, \dots, d$, $\sum_{h\in\mathbb{Z}} \sum_{s_1,s_2=0}^{\infty} |Cov(B_{h,s_1}\underline{\mu}, B_{0,s_2}\underline{\mu})| + \sum_{s=0}^{\infty} |EB_{h,s+h}\Sigma_d B_{0,s}^{\top}| < \infty$ *(component-wise).*

iv) $\sum_{s=0}^{\infty} \left(E\left|e_i^{\top}(B_{0,s}\mu - EB_{0,s}\mu)\right|^5\right)^{1/5} + s\left(E\left[e_i^{\top}(B_{0,s}\underline{\varepsilon}_1 - EB_{0,s}\mu)\right]^4\right)^{1/4} < \infty$ *for all* $i = 1, \dots, d$

v) *The measurable function* $f : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ *fulfills* $f(X) = (g_1(X_{1\cdot}, X_{\cdot 1}), \dots, g_d(X_{d\cdot}, X_{\cdot d}))^{\top}$, *where* $g_j : \mathbb{R}^{2d} \to \mathbb{R}^{1\times d}$, $j = 1, \dots, d$.

vi) *For all* $j = 1, \dots, d$ *if the* l*-th and* $l+d$*-th components of the argument of* g_j *are zero, then the* l*-th component of* g_j *is zero, i.e.,* $g_j(x_1, \dots, x_{l-1}, 0, x_{l+1}, \dots, x_{d+l-1}, 0, x_{d+l+1}, \dots, x_{2d}) = (y_1, \dots, y_{l-1}, 0, y_{l+1}, y_d)^{\top}$

Let $g_j(\tilde{A}d_{k;j\cdot}, \tilde{A}d_{k;\cdot j}) = a_{j,k}$ *be the quantity of interest. Define* $R_k^j = \{r \in \{0, \dots, n-1\} : Ad_{r;j\cdot} = \tilde{A}d_{k;j\cdot}$ *and* $Ad_{r;\cdot j} = \tilde{A}d_{k;\cdot j}\}$ *and* $S_k^j = \{s \in \{1, \dots, d\} : \tilde{A}d_{k;js} \neq 0$ *or* $\tilde{A}d_{k;sj} \neq 0\}$. *Then the process* $\underline{\mathbf{X}}$ *fulfills the equation* $\underline{X}_{t+1;j} = \sum_{s\in S_k^j} a_{j,k;s}\underline{X}_{t;s} + \underline{\varepsilon}_{t+1;j}, t \in R_k^j$, *which results from using the least squares approach in the following linear system:*

$$\left(\sum_{r\in R^k} (\underline{X}_{r;s_1} - \frac{1}{|R^k|}\sum_{v\in R^k} \underline{X}_{v;s_1})_{s_1\in S_k^j} (\underline{X}_{r;s_2} - \frac{1}{|R^k|}\sum_{v\in R^k} \underline{X}_{v;s_2})_{s_2\in S_k^j}^{\top}\right) (\tilde{\alpha}_{j,k;s})_{s\in S_k^j}$$
$$= \left(\sum_{r\in R^k} (\underline{X}_{r+1;j} - \frac{1}{|R^k|}\sum_{v\in R^k} \underline{X}_{v+1;j})(\underline{X}_{r;s} - \frac{1}{|R^k|}\sum_{v\in R^k} \underline{X}_{v;s})_{s\in S_k^j}\right).$$

The estimator $\hat{\alpha}_{j,k} \in \mathbb{R}^d$ *is defined by* $\hat{\alpha}_{j,k;s} = \tilde{\alpha}_{j,k;s}$ *if* $s \in S_k^j$ *and* $\hat{\alpha}_{jk;s} = 0$ *if* $s \notin S_k^j$. *Under the assumptions i) to vi) we have* $\hat{\alpha}_{j,k} = \alpha_{j,k} + \mathcal{O}_P((nP(Ad_{1;j\cdot} = \tilde{A}d_{k;j\cdot}, Ad_{1;\cdot j} = \tilde{A}d_{k;\cdot j}))^{-1/2})$ *for all* $j = 1, \dots, d, k = 0, \dots, n-1$. *Furthermore,* $\underline{\hat{\mu}}_j = \frac{1}{|R_k^j|}\sum_{r\in R_k^j} \underline{X}_{r+1;j} - \tilde{\alpha}_{j,k}\underline{X}_r = \underline{\mu}_j + \mathcal{O}_P((nP(Ad_{1;j\cdot} = \tilde{A}d_{k;j\cdot}, Ad_{1;\cdot j} = \tilde{A}d_{k;\cdot j}))^{-1/2})$.

The results of Theorem 3.3.3 can be used to forecast the process $\underline{\mathbf{X}}$, where $\underline{X}_t = f(Ad_{t-1})^{\top}\underline{X}_t + \underline{\varepsilon}_t$. In order to forecast $\underline{X}_{n+1}$, it is only necessary to estimate f at the state of Ad_n, here denoted by $\tilde{A}d_n$ with $f(\tilde{A}d_n) = (\alpha_{1,n}, \dots, \alpha_{d,n})^{\top}$. If Ad_n is observed, $\underline{X}_{n+1}$ can be forecasted by $\hat{\underline{X}}_{n+1}^{(1)} = \left(\sum_{s\in S_n^j} \hat{\alpha}_{j,n;s}\underline{X}_{n;s}\right)_{j=1,\dots,d} + \underline{\hat{\mu}}$. Since the innovation process is i.i.d., $\hat{\alpha}_{j,n} = \alpha_{j,n} + \mathcal{O}_P((nP(Ad_{1;j\cdot} = \tilde{A}d_{k;j\cdot}, Ad_{1;\cdot j} = \tilde{A}d_{k;\cdot j}))^{-1/2})$, and $\underline{\hat{\mu}} = \underline{\mu} + \mathcal{O}_P((nP(Ad_{1;j\cdot} = \tilde{A}d_{k;j\cdot}, Ad_{1;\cdot j} = \tilde{A}d_{k;\cdot j}))^{-1/2})$, we have $E[(\underline{X}_{n+1} - \hat{\underline{X}}_{n+1}^{(1)})(\underline{X}_{n+1} - \hat{\underline{X}}_{n+1}^{(1)})^{\top}|(Ad_n, \underline{X}_n)] = \mathrm{Var}((\hat{\alpha}_{j,n} - \alpha_{j,n})_{j=1,\dots,d}\underline{X}_n + \underline{\varepsilon}_{n+1} - \underline{\hat{\mu}}|(Ad_n, \underline{X}_n)) = \mathcal{O}_P((nP(Ad_{1;j\cdot} = \tilde{A}d_{k;j\cdot}, Ad_{1;\cdot j} = \tilde{A}d_{k;\cdot j}))^{-1/2}) + \Sigma$. If Ad_n is not observed, Ad_n itself needs to be predicted first. For instance, if $\mathbf{Ad}$ is Markovian, a prediction using an estimated transition matrix based on $Ad_0, \dots, Ad_{n-1}$ may be possible.

Even though these two assumptions decrease the number of parameters, this approach is not feasible for large networks. The estimation error for each state of the adjacency matrix is of the order $\mathcal{O}((nP(Ad_{1;j\cdot} = \tilde{A}d_{k;j\cdot}, Ad_{1;\cdot j} = \tilde{A}d_{k;\cdot j}))^{-1/2})$. Since the number of states can grow faster than a

polynomial growth, the probability to observe a given state, $P(Ad_{1;j\cdot} = \tilde{A}d_{k;j\cdot}, Ad_{1;\cdot j} = \tilde{A}d_{k;\cdot j})$, may decrease faster than polynomially. Thus, the number of observations needed to get adequate results could be of exponential order to the number of vertices. That is why a different approach is presented in the following: The previous approach considered f as an arbitrary measurable function, whereas the following approaches will parametrize f. First, consider the setting of (3.2.3). Thus, each edge gets a fixed parameter resulting in the following representation of the DSNAR(1) model: $\underline{X}_t = (\alpha \circledast Ad_{t-1})^\top \underline{X}_{t-1} + \underline{\varepsilon}_t, t \in \mathbb{Z}$. This results in only $\sum_{j=1}^d \sum_{s=1}^d \mathbb{1}_{\{\sup_{t\in\mathbb{Z}} |Ad_{t;js}|>0\}} \leq d^2$ parameters for f. Since in sparse network the number of edges grows linearly with the number of vertices, this model can be parameterized in sparse networks with $\mathcal{O}(d)$ parameters. Again, the least squares approach is used to derive consistent estimators for α and $\underline{\mu}$. However, an important difference to the estimation in Theorem 3.3.3 is that for this model a global approach can be used. With this parameterization of f the influence of each edge does not depend on the state of the adjacency matrix. That is why in contrast to the estimation in Theorem 3.3.3, it is not necessary to condition on a given state of the adjacency matrix. Since all observations can be used, this results in a more stable estimation. Consider we have the observations $X_0, \dots, X_n$ and $Ad_0, \dots, Ad_{n-1}$ and $\underline{\mathbf{X}}$ is given by

$$\underline{X}_{t;s} = \sum_{j=1}^{d} \alpha_{js} Ad_{t-1;js} \underline{X}_{t-1;j} + \varepsilon_{t;s}, s = 1, \dots, d, t = 1, \dots, n. \tag{3.3.7}$$

Let $\tilde{S}_s = \{j \in \{1, \dots, d\} : \sup_{t\in\mathbb{Z}} Ad_{t;js} > 0\}$ and define the $|\tilde{S}_s|$-dimensional vectors $Y_t^s := (Ad_{t;js}\underline{X}_{t;j})_{j\in\tilde{S}_s}, t \in \mathbb{Z}$, $\tilde{\alpha}_{\cdot s} = (\alpha_{js})_{j\in\tilde{S}_s}$ so that $\underline{X}_{t;s} = \tilde{\alpha}_{\cdot s}^\top Y_{t-1}^s + \varepsilon_{t;s}$. Using the least squares approach to estimate $(\tilde{\alpha}_{\cdot s}, \underline{\mu}_s)$ leads to the following linear system:

$$\begin{pmatrix} \sum_{t=1}^n \underline{X}_{t;s} Y_{t-1}^s - 1/n \sum_{t_1,t_2=1}^n Y_{t_1-1}^s \underline{X}_{t_2;s} \\ \sum_{t=1}^n \underline{X}_{t;s} \end{pmatrix} = \begin{pmatrix} 0 & \sum_{t=0}^{n-1} Y_t^s (Y_t^s)^\top - 1/n \sum_{t_1,t_2=0}^{n-1} (Y_{t_1}^s)(Y_{t_2}^s)^\top \\ n & \sum_{t=0}^{n-1} (Y_t^s)^\top \end{pmatrix} \begin{pmatrix} \hat{\underline{\mu}}_s \\ \hat{\alpha}_{\cdot s} \end{pmatrix}. \tag{3.3.8}$$

Since $\alpha_{js}, j \notin S_s, s = 1, \dots, d$ does not come into play in (3.3.7) and therefore, can be chosen arbitrarily without changing the model, we set them to 0. In a finite sample where not every edge with nonzero occurrence probability is observed, one is naturally only able to estimated those α_{js} for which the corresponding edge is observed, i.e. for $j, s \in 1, \dots, d : \sum_{t=0}^{n-1} Ad_{t;js} > 0$. Consistency of this estimator is shown in Theorem 3.3.4.

Theorem 3.3.4. *Let $\underline{X}_t = (\alpha \circledast Ad_{t-1})\underline{X}_{t-1} + \underline{\varepsilon}_t$ be an $\mathbb{R}^d$-valued DSNAR(1) and it is observed $X_0, \dots, X_n$ and $Ad_0, \dots, Ad_{n-1}$. For $s = 1, \dots, d$ define $\tilde{S}_s = \{j \in \{1, \dots, d\} : \sup_{t\in\mathbb{Z}} Ad_{t;js} > 0\}$, $Y_t^s := (Ad_{t;js}\underline{X}_{t;j})_{j\in\tilde{S}_s}$, and $\tilde{\alpha}_{\cdot s} = (\alpha_{js})_{j\in\tilde{S}_s}$ so that $\underline{X}_{t;s} = \tilde{\alpha}_{\cdot s} Y_{t-1}^s + \varepsilon_{t;s}$. Furthermore, set $A_t^s := \operatorname{diag}(Ad_{t;\cdot s})$. If*

1. **Ad** *is a strictly stationary, $\mathbb{R}^{d\times d}$-valued, α-mixing process fulfilling $\sum_{n=1}^\infty \alpha(\mathbf{Ad}, n)^{1/5} < \infty$,*

2. *the innovations* $\underline{\varepsilon}$ *are* $\mathbb{R}^d$*-valued and i.i.d. with* $E\underline{\varepsilon}_0 = \underline{\mu}$, $Cov(\underline{\varepsilon}_0, \underline{\varepsilon}_0) = \Sigma_d$, $E\underline{\varepsilon}_{0;i_1}\underline{\varepsilon}_{0;i_2}\underline{\varepsilon}_{0;i_3}\underline{\varepsilon}_{0;i_4} = \kappa_{4;i_1,i_2,i_3,i_4} < \infty$ *for all* $i_1, \ldots, i_4 = 1, \ldots, d$. $\underline{\varepsilon}$ *and* **Ad** *are independent,*
3. $B_{t,0} = I_d$, $B_{t,j} = \prod_{l=1}^{j}(\alpha \circledast Ad_{t-l})$ *is set and* $E|A_0^s B_{0,j;i_1,i_2}|^4 < \infty$ *for all* $j \in \mathbb{N}, s, i_1, i_2 = 1, \ldots, d$, $\sum_{h\in\mathbb{Z}}\sum_{s_1,s_2=0}^{\infty}|Cov(A_h^s B_{h,s_1}\underline{\mu}, A_0^s B_{0,s_2}\underline{\mu})| + \sum_{l=0}^{\infty}|EA_h^s B_{h,l+h}\Sigma_d B_{0,l}^\top (A_0^s)^\top| < \infty$ *(component-wise),*
4. *and* $\sum_{l=0}^{\infty}\left(E\left|e_i^\top(A_h^s B_{h,l}\mu - EA_0^s B_{0,l}\mu)\right|^5\right)^{1/5} + l\left(E\left[e_i^\top(A_h^s B_{h,l}\underline{\varepsilon}_1 - EA_0^s B_{0,l}\mu)\right]^4\right)^{1/4} < \infty$ *for all* $s, i = 1, \ldots, d, h = 0, 1,$

then the estimator given by (3.3.7) *is consistent. We have, as* $n \to \infty$,

$$\sqrt{n}\begin{pmatrix}(\hat{\underline{\mu}}_s - \underline{\mu}_s)\\ (\hat{\alpha}_{\cdot s}^\top - \alpha_{\cdot s}^\top)\end{pmatrix} \xrightarrow{D} \mathcal{N}\left(\underline{0}, \Sigma_{ss}\begin{pmatrix}(1 + E(Y_1^s)^\top\Gamma_{Y^s}(0)^{-1}EY_1^s) & E(Y_1^s)^\top\Gamma_{Y^s}(0)^{-1}\\ \Gamma_{Y^s}(0)^{-1}EY_1^s & \Gamma_{Y^s}(0)^{-1}\end{pmatrix}\right). \tag{3.3.9}$$

Furthermore, the following covariances are obtained for $k \in \{1, \ldots, d\}$, *as* $n \to \infty$*:*

$$Cov(\sqrt{n}(\hat{\alpha}_{\cdot s} - \alpha_{\cdot s}), \sqrt{n}(\hat{\alpha}_{\cdot k} - \alpha_{\cdot k})) \to \Sigma_{sk}\Gamma_{Y^s}(0)^{-1}\Gamma_{Y^sY^k}(0)\Gamma_{Y^k}(0)^{-1},$$

$$Cov(\sqrt{n}(\hat{\alpha}_{\cdot s} - \alpha_{\cdot s}), \sqrt{n}(\hat{\underline{\mu}}_k - \underline{\mu}_k)) \to \Sigma_{sk}\Gamma_{Y_s}(0)^{-1}\Gamma_{Y^sY^k}(0)\Gamma_{Y^k}(0)^{-1}EY_1^k,$$

and

$$Cov(\sqrt{n}(\hat{\underline{\mu}}_s - \underline{\mu}_s), \sqrt{n}(\hat{\underline{\mu}}_k - \underline{\mu}_k)) \to \Sigma_{sk}(1 + E(Y_1^k)^\top\Gamma_{Y_s}(0)^{-1}\Gamma_{Y^sY^k}(0)\Gamma_{Y^k}(0)^{-1}EY_1^k).$$

This can be used to forecast the process $\underline{\mathbf{X}}$, where $\underline{X}_t = (\alpha \circledast Ad_{t-1})^\top \underline{X}_{t-1} + \underline{\varepsilon}_t$. If Ad_n is observed, $\underline{X}_{n+1}$ can be forecasted by $\hat{\underline{X}}_{n+1}^{(1)} = (\hat{\alpha} \circledast Ad_n)^\top \underline{X}_t$. Since the innovation process is i.i.d. and $\hat{\alpha} = \alpha + \mathcal{O}_{\mathcal{P}}(1/\sqrt{n})$, we have $\mathrm{Var}(\underline{X}_{n+1} - \hat{\underline{X}}_{n+1}^{(1)}|(Ad_n, \underline{X}_n)) = \mathrm{Var}(((\hat{\alpha} - \alpha) \circledast Ad_n)\underline{X}_n + \underline{\varepsilon}_{n+1}|(Ad_n, \underline{X}_n)) = \mathcal{O}(1/n) + \Sigma$.

Even though the parameterization used in Theorem 3.3.4 reduces the number of parameter to $\mathcal{O}(d^2)$ or $\mathcal{O}(d)$ for sparse networks, respectively, this may be too large to tackle very large networks such as social media networks as *Twitter* or *Facebook*. Those networks often contain more than millions of vertices, whereas the number of observed time points is considerably small. Consequently, a more radical approach needs to be applied here in order to reduce the number of parameters. Here we adapt the idea of model (2.1) in Zhu et al. (2017). The model reads as follows:

$$\underline{X}_t = \alpha\underline{X}_{t-1} + \beta h(Ad_{t-1})\underline{X}_{t-1} + \mu + \underline{\varepsilon}_t, \tag{3.3.10}$$

where $\alpha, \beta, \mu \in \mathbb{R}$ and $\{\underline{\varepsilon}_t, t \in \mathbb{Z}\}$ is an i.i.d. innovations process with $E\underline{\varepsilon}_1 = 0$ and $\mathrm{Var}\underline{\varepsilon}_1 = \Sigma$. The function $h : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ is assumed to be known. Thus, some prior knowledge is put into the model. For instance, if the vertices are considered as cities then h can be in such a way that each

edge is assigned to some distance between these cities, see Knight et al. (2016, Section 3.1). Another example for h is given by choosing h as follows: $h(X)_{j\cdot} = (\sum_{s=1}^{d} X_{sj})^{-1} X_{\cdot j}$ which is closely related to Linear-In-Means models for peer effects, see Manski (1993). This means that if two edges go into vertex j both have an impact of 0.5β and if we have four edges, each one has an impact of 0.25β. In order to better visualize the parameters, here the common mean of the innovation is written directly in the equation for $\underline{X}_t$. Since each vertex shares the same parameters, the least square approach can be used in the following way:

$$\begin{aligned} &\operatorname{argmin}_{\hat{\alpha},\hat{\beta},\hat{\mu}} \sum_{t=1}^{n} \|\underline{X}_t - \alpha \underline{X}_{t-1} - \beta h(Ad_{t-1})\underline{X}_{t-1} - \mu \mathbb{1}\|_2^2 \\ =&\operatorname{argmin}_{\hat{\alpha},\hat{\beta},\hat{\mu}} \sum_{t=1}^{n} \sum_{s=1}^{d} \left(\underline{X}_{t;s} - \alpha \underline{X}_{t-1;s} - \beta h(Ad_{t-1})_{s\cdot}\underline{X}_{t-1} - \mu\right)^2. \end{aligned} \tag{3.3.11}$$

Define $Y_{t-1;s} = h(Ad_{t-1})_{s\cdot}\underline{X}_{t-1}$. This results in the following linear system:

$$\sum_{t=1}^{n}\sum_{s=1}^{d} \begin{pmatrix} \underline{X}_{t-1;s}\underline{X}_{t;s} \\ Y_{t-1;s}\underline{X}_{t;s} \\ \underline{X}_{t;s} \end{pmatrix} = \sum_{t=1}^{n}\sum_{s=1}^{d} \begin{pmatrix} \underline{X}_{t-1;s}^2 & \underline{X}_{t-1;s}Y_{t-1;s} & \underline{X}_{t-1;s} \\ \underline{X}_{t-1;s}Y_{t-1;s} & Y_{t-1;s}^2 & Y_{t-1;s} \\ \underline{X}_{t-1;s} & Y_{t-1;s} & 1 \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \\ \hat{\mu} \end{pmatrix}. \tag{3.3.12}$$

The main difference to the previous models is evident in this linear system. The number of parameters is fixed to 3 and is independent from the number of vertices. Furthermore, it can be seen that a more accurate estimation can be achieved by an increasing number of time points as well as by an increasing number of vertices (as long as there is not a perfect correlation between the components of $\underline{X}_t$). The benefit of a larger network is higher the less the components of the time series are correlated. The correlation of the components is mainly influenced by the function h. If this function has similar properties as the function f in Assumption v) in Theorem 3.3.3 than sparsity of the network could result in less correlated components. For this model it is reasonable to consider $d \to \infty$. However, the setting $d \to \infty$ requires another definition of stationarity. This should not be the scope of this work (for this refer to section 2.3 in Zhu et al. (2017)). Since an increasing number of vertices is only appropriate for such reduced models as given in (3.3.10) and not for the previous discussed models, this work keeps the setting of a fixed number of vertices and only the number of time points increases. Note further that the model can be written as: $\underline{X}_t = (\alpha I_d + \beta h(Ad_{t-1}))\underline{X}_{t-1} + \mu + \underline{\varepsilon}_t$. Under some conditions we get the following stationary solution:

$$\underline{X}_t = \sum_{j=0}^{\infty} \left[\prod_{s=1}^{j} (\alpha I_d + \beta h(Ad_{t-s}))\right](\mu + \underline{\varepsilon}_t). \tag{3.3.13}$$

Hence, for the mean we have $E\underline{X}_t = \sum_{j=0}^{\infty} E\left[\prod_{s=1}^{j}(\alpha I_d + \beta h(Ad_{t-s}))\right]\mu$. The mean as well as the autocovariance structure depends on the underlying network. So, despite all components of the time series sharing the same parameters, the mean of the components can differ. The linear system (3.3.12) gives consistent estimators for $t \to \infty$ which is shown in Theorem 3.3.5.

Theorem 3.3.5. *Let $\underline{X}_t = \alpha\underline{X}_{t-1} + \beta h(Ad_{t-1})\underline{X}_{t-1} + \mu + \underline{\varepsilon}_t$ be a $\mathbb{R}^d$-valued DSNAR(1). It is observed $X_0, \ldots, X_n$ and $Ad_0, \ldots, Ad_n$ and h is assumed to be known. Furthermore, define $Y_t = h(Ad_t)\underline{X}_t$. If*

1. **Ad** *is a strictly stationary, $\mathbb{R}^{d\times d}$-valued, α-mixing process fulfilling $\sum_{n=1}^{\infty}\alpha(\mathbf{Ad}, n)^{1/5} < \infty$,*

2. *the innovations $\underline{\varepsilon}$ are $\mathbb{R}^d$-valued and i.i.d. with $E\underline{\varepsilon}_0 = \underline{\mu}$, $Cov(\underline{\varepsilon}_0, \underline{\varepsilon}_0) = \Sigma_d$, $E\underline{\varepsilon}_{0;i_1}\underline{\varepsilon}_{0;i_2}\underline{\varepsilon}_{0;i_3}\underline{\varepsilon}_{0;i_4} = \kappa_{4;i_1,i_2,i_3,i_4} < \infty$ for all $i_1, \ldots, i_4 = 1, \ldots, d$. $\underline{\varepsilon}$ and **Ad** are independent,*

3. *$B_{t,0} = I_d$, $B_{t,j} = \prod_{s=1}^{j}(\alpha I_d + \beta h(Ad_{t-s}))$, and $E|B_{0,j;i_1,i_2}|^4 < \infty$ for all $j \in \mathbb{N}, i_1, i_2 = 1, \ldots, d$,*
$\sum_{h\in\mathbb{Z}}\sum_{s_1,s_2=0}^{\infty}|Cov(B_{h,s_1}\underline{\mu}, B_{0,s_2}\underline{\mu})| + \sum_{s=0}^{\infty}|EB_{h,s+h}\Sigma_d B_{0,s}^{\top}| < \infty$ (component-wise),
$\sum_{s=0}^{\infty}\left(E\left|e_i^{\top}(B_{h,s}\mu - EB_{0,s}\mu)\right|^5\right)^{1/5} + \sum_{s=1}^{\infty} s\left(E\left[e_i^{\top}(B_{h,s}\underline{\varepsilon}_1 - EB_{0,s}\mu)\right]^4\right)^{1/4} < \infty$ for all $i = 1, \ldots, d, h = 0, 1$ and

4. *$\tilde{B}_{t,0} = \beta h(Ad_t)$, $\tilde{B}_{t,j} = \beta h(Ad_t)\prod_{s=1}^{j}(\alpha I_d + \beta h(Ad_{t-s}))$, $E|\tilde{B}_{0,j;i_1,i_2}|^4 < \infty$ for all $j \in \mathbb{N}, i_1, i_2 = 1, \ldots, d$,*
$\sum_{h\in\mathbb{Z}}\sum_{s_1,s_2=0}^{\infty}|Cov(\tilde{B}_{h,s_1}\underline{\mu}, \tilde{B}_{0,s_2}\underline{\mu})| + \sum_{s=0}^{\infty}|E\tilde{B}_{h,s+h}\Sigma_d\tilde{B}_{0,s}^{\top}| < \infty$ (component-wise),
$\sum_{s=0}^{\infty}\left(E\left|e_i^{\top}(\tilde{B}_{h,s}\mu - E\tilde{B}_{0,s}\mu)\right|^5\right)^{1/5} + \sum_{s=1}^{\infty} s\left(E\left[e_i^{\top}(\tilde{B}_{h,s}\underline{\varepsilon}_1 - E\tilde{B}_{0,s}\mu)\right]^4\right)^{1/4} < \infty$ for all $i = 1, \ldots, d, h = 0, 1$,

then the estimator given by (3.3.12) is consistent. We have, as $n \to \infty$,

$$\sqrt{n}\begin{pmatrix}\hat{\alpha}-\alpha\\ \hat{\beta}-\beta\\ \hat{\underline{\mu}}-\underline{\mu}\end{pmatrix} \xrightarrow{D} \mathcal{N}\left(\underline{0}, \begin{pmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13}\\ \sigma_{12} & \sigma_{22} & \sigma_{23}\\ \sigma_{13} & \sigma_{23} & \sigma_{33}\end{pmatrix}\right), \tag{3.3.14}$$

where, denoting $\tilde{\sum}_s := 1/d\sum_{s=1}^{d}$ and $\tilde{\sum}_{s_1,s_2} := 1/d^2\sum_{s_1,s_2=1}^{d}$,

$$\sigma_{11} = \left(\tilde{\sum_s} E\underline{X}_{1;s}^2 - (\tilde{\sum_s} E\underline{X}_{1;s})^2 - (\tilde{\sum_s} EY_{1;s}^2 - (\tilde{\sum_s} EY_{1;s})^2)^{-1}(\tilde{\sum_s} EY_{1;s}\underline{X}_{1;s} - (\tilde{\sum_s} E\underline{X}_{1;s})(\tilde{\sum_s} EY_{1;s}))\right)^{-2}$$
$$\times\left(\tilde{\sum_{s_1,s_2}}\Sigma_{s_1,s_2}[E(\underline{X}_{1;s_1}\underline{X}_{1;s_2}) + (\tilde{\sum_s} EY_{1;s}^2 - (\tilde{\sum_s} EY_{1;s})^2)^{-1}E(Y_{1;s_1}\underline{X}_{1;s_2})\right.$$
$$\left.+(\tilde{\sum_s} EY_{1;s}^2 - (\tilde{\sum_s} EY_{1;s})^2)^{-2}E(Y_{1;s_1}Y_{1;s_2})]\right),$$

$$\sigma_{22} = \left(\tilde{\sum_s} E\underline{X}_{1;s}^2 - (\tilde{\sum_s} E\underline{X}_{1;s})^2 - (\tilde{\sum_s} E\underline{X}_{1;s}^2 - (\tilde{\sum_s} E\underline{X}_{1;s})^2)^{-1} (\tilde{\sum_s} EY_{1;s}\underline{X}_{1;s} - (\tilde{\sum_s} E\underline{X}_{1;s})(\tilde{\sum_s} EY_{1;s})) \right)^{-2}$$
$$\times \left(\tilde{\sum_{s_1,s_2}} \Sigma_{s_1,s_2} [E(Y_{1;s_1} Y_{1;s_2}) + (\tilde{\sum_s} E\underline{X}_{1;s}^2 - (\tilde{\sum_s} E\underline{X}_{1;s})^2)^{-1} E(Y_{1;s_1} \underline{X}_{1;s_2}) \right.$$
$$\left. + (\tilde{\sum_s} E\underline{X}_{1;s}^2 - (\tilde{\sum_s} E\underline{X}_{1;s})^2)^{-2} E(\underline{X}_{1;s_1} \underline{X}_{1;s_2})] \right),$$

$$\sigma_{12} = \left(\tilde{\sum_s} E\underline{X}_{1;s}^2 - (\tilde{\sum_s} E\underline{X}_{1;s})^2 - (\tilde{\sum_s} E\underline{X}_{1;s}^2 - (\tilde{\sum_s} E\underline{X}_{1;s})^2)^{-1} (\tilde{\sum_s} EY_{1;s}\underline{X}_{1;s} - (\tilde{\sum_s} E\underline{X}_{1;s})(\tilde{\sum_s} EY_{1;s})) \right)^{-1}$$
$$\times \left(\tilde{\sum_s} E\underline{X}_{1;s}^2 - (\tilde{\sum_s} E\underline{X}_{1;s})^2 - (\tilde{\sum_s} EY_{1;s}^2 - (\tilde{\sum_s} EY_{1;s})^2)^{-1} (\tilde{\sum_s} EY_{1;s}\underline{X}_{1;s} - (\tilde{\sum_s} E\underline{X}_{1;s})(\tilde{\sum_s} EY_{1;s})) \right)^{-1}$$
$$\times \left(\tilde{\sum_{s_1,s_2}} \Sigma_{s_1,s_2} E(Y_{1;s_1} \underline{X}_{1;s_2}) (1 + (\tilde{\sum_s} E\underline{X}_{1;s}^2 - (\tilde{\sum_s} E\underline{X}_{1;s})^2)^{-1} (\tilde{\sum_s} EY_{1;s}^2 - (\tilde{\sum_s} EY_{1;s})^2)^{-1}) \right.$$
$$\left. - (\tilde{\sum_s} E\underline{X}_{1;s}^2 - (\tilde{\sum_s} E\underline{X}_{1;s})^2)^{-1} E(\underline{X}_{1;s_1} \underline{X}_{1;s_2}) - (\tilde{\sum_s} EY_{1;s}^2 - (\tilde{\sum_s} EY_{1;s})^2)^{-1} E(Y_{1;s_1} Y_{1;s_2}) \right),$$

$$\sigma_{33} = 1/d^2 \sum_{s_1,s_2}^{d} \Sigma_{s_1,s_2},$$

$$\sigma_{13} = \left(\tilde{\sum_s} E\underline{X}_{1;s}^2 - (\tilde{\sum_s} E\underline{X}_{1;s})^2 - (\tilde{\sum_s} EY_{1;s}^2 - (\tilde{\sum_s} EY_{1;s})^2)^{-1} (\tilde{\sum_s} EY_{1;s}\underline{X}_{1;s} - (\tilde{\sum_s} E\underline{X}_{1;s})(\tilde{\sum_s} EY_{1;s})) \right)^{-1}$$
$$\times \left(\tilde{\sum_{s_1,s_2}} \Sigma_{s_1,s_2} [E\underline{X}_{1;s_1} + (\tilde{\sum_s} EY_{1;s}^2 - (\tilde{\sum_s} EY_{1;s})^2)^{-1} EY_{1;s_1}] \right),$$

$$\sigma_{23} = \left(\tilde{\sum_s} E\underline{X}_{1;s}^2 - (\tilde{\sum_s} E\underline{X}_{1;s})^2 - (\tilde{\sum_s} E\underline{X}_{1;s}^2 - (\tilde{\sum_s} E\underline{X}_{1;s})^2)^{-1} (\tilde{\sum_s} EY_{1;s}\underline{X}_{1;s} - (\tilde{\sum_s} E\underline{X}_{1;s})(\tilde{\sum_s} EY_{1;s})) \right)^{-1}$$
$$\times \left(\tilde{\sum_{s_1,s_2}} \Sigma_{s_1,s_2} [EY_{1;s_1} + (\tilde{\sum_s} E\underline{X}_{1;s}^2 - (\tilde{\sum_s} E\underline{X}_{1;s})^2)^{-1} E\underline{X}_{1;s_1}] \right).$$

Note that if the components of $\underline{X}_t$ are not fully linear dependent, then terms of the asymptotic variance such as $1/d^2 \sum_{s_1,s_2}^{d} \Sigma_{s_1,s_2}$ decrease with higher d. Hence, as mentioned previously, this approach benefits — as long as the components are not fully correlated— from an increasing dimension in the sense of a decreasing variance. Under some conditions the variance could decrease with rate $\mathcal{O}(1/d)$.

Lemma 3.3.6 gives an easy-to-check criteria, whether a given DSNAR(1) model fulfills the moment conditions of the Theorem 3.3.1 to 3.3.5. However, this condition is more restrictive since it implies an exponential decay of $(\prod_{s=1}^{n} f(Ad_{-s}))_n$. The models used in the numerical examples fulfill this criteria.

Lemma 3.3.6. *Let $\underline{X}_t = f(Ad_{t-1})\underline{X}_{t-1} + \underline{\varepsilon}_t$ be a DSNAR(1) process as in 3.2.2. If there exists a $q \geq 1$ so that $E \log \| \prod_{j=1}^{q} f(Ad_1) \| < 0$ and $\mathbf{Ad}$ is α-mixing with $\sum_{n=0}^{\infty} \alpha(\mathbf{Ad}, n)^{1/2} < \infty$, then the moment conditions iii), iv) of Theorem 3.3.3, 3), 4) of Theorem 3.3.4 and 3), 4) of Theorem 3.3.5 hold.*

In practice we face the situation that we have observation of some process but a priori it is usually unknown which models fits best to the process. We have that model (3.3.7) is more general than model (3.3.10) and model (3.3.5) is even more general. Hence, the more general models apply to more processes. However, as mentioned above the number of parameters of these more general models can be large, especially for model (3.3.5). Thus, the usual bias-variance dilemma occur, see section 7.2 in Friedman et al. (2017)). Hence, the bias of the more general models may be smaller but the variance can increase significantly. If the forecasting performance is of interest then cross-validation, see section 7.10 in Friedman et al. (2017), can be used to identify which of these three model approaches gives the best forecasting performance regarding some metric, e.g. the mean-squared-error.

Notice that all these approaches are based on observations of the process $\underline{\mathbf{X}}$ as well as observations of the network $\mathbf{Ad}$. Hence, if only observations of $\underline{\mathbf{X}}$ are available these methods cannot be applied. However, as seen in the example in section 3.2, the autocovariance function of $\underline{\mathbf{X}}$ cannot be used in general to identify a DSNAR(1) model. Furthermore, as already mentioned, a DSNAR(1) model does not fit into the framework of Hidden Markov models. Thus, the corresponding techniques cannot be applied here either. It remains to consider $\underline{\mathbf{X}}$ as a *standard* multivariate time series, which may be tackled by VAR-models. However, VAR-models cannot benefit from the additional structure. In section 3.4 we investigate the finite sample performance of these forecasting methods under the precondition that the network is observed.

3.4 Numerical Examples

In this Section the one-step-forecasting error for $\underline{X}_{n+1}$ of the methods presented in Section 3 are compared based on observations $\underline{X}_1, \ldots, \underline{X}_n$ and $Ad_1, \ldots, Ad_n$. In the low-dimensional examples, the methods of Theorem 3.3.3, here denoted as *NP.NAR*, as well as of Theorem 3.3.4, here denoted as *FIX.NAR*, are compared with the approach using standard VAR models. The standard VAR model is not able to use the observations of $\mathbf{Ad}$, which makes it in some sense an unfair competition. However, the aim is here to see what the benefit is of using this additional structure. Some of these methods presented in Section 3 have many parameters. Nevertheless, under appropriate conditions they should clearly outperform the VAR model. Since the method of Theorem 3.3.5, here denoted as *RAD.NAR*, uses a priori knowledge, it is only used in the last example.

We begin with the example given in section 3.2 by (3.2.5). Hence, $\underline{\mathbf{X}}$ is a 3-dimensional time series, where the first two components are whites noise and the third component is either influenced by

the first or by the second component, see Section 2 for details. *NP.NAR* as well as *FIX.NAR* are valid. The one-step-forecasting error based on $\underline{X}_1, \dots, \underline{X}_n$ and $Ad_1, \dots, Ad_n$ is compared with the forecasting error using VAR and is displayed in Table 3.1 for various sample sizes n. It can be seen that in this example there is not much of a difference between *NP.NAR* and *FIX.NAR*. Hence, the disadvantage of the additional parameters in *NP.NAR* can be handled well in this low-dimension example. However, *NP.NAR* as well as *FIX.NAR* has their difficulties for $n = 100$. For this sample size these methods are not able to reduce the forecasting error to the innovations variance. Nevertheless, both clearly benefit from the additional structure and are able to give a more accurate forecast for the third component than the *VAR* approach. Hence, using only the information given by the autocovariance function does not give a good forecast for this process.

Table 3.1: Mean squared one-step forecasting error for $\hat{\underline{X}}_{n+1}$ based on $\underline{X}_1, \dots, \underline{X}_n$ and $Ad_1, \dots, Ad_n$ of process (3.2.5).

n	100			200			400		
Component	1	2	3	1	2	3	1	2	3
VAR	1.1	1.0	1391.0	1.0	1.0	1129.0	1.0	1.0	1055.4
FIX.NAR	1.0	1.0	34.6	1.0	1.0	1.0	1.0	1.0	1.0
NP.NAR	1.1	1.0	23.6	1.0	1.0	1.0	1.0	1.0	1.0

In the second example, a network with 4 vertices is considered. The adjacency matrix process **Ad** is a Markovian process and the edges are independent from each other. The process **Ad** is given by

$$\begin{aligned}
\big(P(Ad_{t;ij} = 1 | Ad_{t-1;ij} = 1)\big)_{i,j=1,\dots,d} &= \begin{pmatrix} 0.95 & 0.70 & 0.99 & 0 \\ 0 & 0.95 & 0.70 & 0 \\ 0.99 & 0.50 & 0.95 & 0.95 \\ 0.30 & 0 & 0 & 0.95 \end{pmatrix}, \\
\big(P(Ad_{t;ij} = 1 | Ad_{t-1;ij} = 0)\big)_{i,j=1,\dots,d} &= \begin{pmatrix} 0.05 & 0.10 & 0.01 & 0 \\ 0 & 0.05 & 0.30 & 0 \\ 0.01 & 0.50 & 0.05 & 0.05 \\ 0.30 & 0 & 0 & 0.05 \end{pmatrix}. \qquad (3.4.1)
\end{aligned}$$

The edges have fixed weights $\alpha = \begin{pmatrix} 0.25 & 0.75 & 0 & 0 \\ 0 & 0.25 & 0.75 & 0 \\ 0 & 0 & 0.25 & 0.75 \\ 0.75 & 0 & 0 & 0.25 \end{pmatrix}$ and the time series $\underline{\mathbf{X}}$ is an DSNAR(1)

process given by

$$\underline{X}_t = (\alpha \circledast Ad_{t-1})\,\underline{X}_{t-1} + \underline{\varepsilon}_t, \underline{\varepsilon}_1 \sim \mathcal{N}\left((-1,4,-9,16)^\top, I_4\right). \tag{3.4.2}$$

A realization of the network as well as of the time series is displayed in Figure 3.4. Furthermore, the sample autocovariance function is displayed in Figure 3.4, which indicates that $\underline{\mathbf{X}}$ possesses a lot of structure on which the forecasting can rely on. The edges $(3,1)$ and $(1,3)$ have a weight of 0, hence, whether they are present or not, they do not influence the time series $\underline{\mathbf{X}}$. However, the number of possible states is increased which may decrease the performance of the *NP.NAR* approach. Due to the relative large mean of the innovations of the 4th component regarding the 1st component, $\underline{\mu}_4 = 16$ versus $\underline{\mu}_1 = 1$, the presence of an edge $(4,1)$ at $t-1$ has a strong influence on $\underline{X}_{t;1}$. That is why this component has the largest variance of the four components. *NP.NAR* as well as *FIX.NAR* are valid and a one-step-forecast is performed. The one-step-forecasting error is displayed in Figure 3.3 for $n = 500$ and the squared forecasting error for each component is given by Table 3.2.

Table 3.2: Mean squared one-step forecasting error for $\hat{\underline{X}}_{n+1}$ based on $\underline{X}_1, \ldots, \underline{X}_n$ and $Ad_1, \ldots, Ad_n$ of process (3.4.2).

n	250				500				1000			
Component	1	2	3	4	1	2	3	4	1	2	3	4
VAR	31.1	5.5	6.0	6.0	29.8	5.2	6.2	5.8	29.6	4.6	5.9	5.3
FIX.NAR	1.1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
NP.NAR	10.9	2.6	32.8	1.1	3.3	1.1	3.7	1.0	1.3	1.1	1.7	1.0

As can be seen in Table 3.2 as well as in Figure 3.3, *FIX.NAR* performs best and this method is able to reduce the forecasting error to the variance of the innovations. *NP.NAR* has a much larger variance for component 1 and 3 especially. Figure 3.3 gives some insight, the forecast based on *NP.NAR* has many outliers for components 1 and 3 , whereas the 50% area is almost as tight as it is for component 2 and 4. The reason for this is that additional zero-weighted edges $(3,1)$ and $(1,3)$ can occur. On the one side, they increase the number of states. Whereas there are 8 different states for component 2 and 4 which results in 12 parameters, components 1 and 3 have 32 different state which results in 72 parameters. On the other side, the edges $(3,1)$ and $(1,3)$ change their current state only with low probability and consequently it is possible that the state Ad_n is not often observed. This could result in a poor forecast especially for smaller sample sizes as seen in Table 3.2. Nevertheless, *NP.NAR* is able to benefit from the additional structure and can give a more accurate forecast than *VAR*.

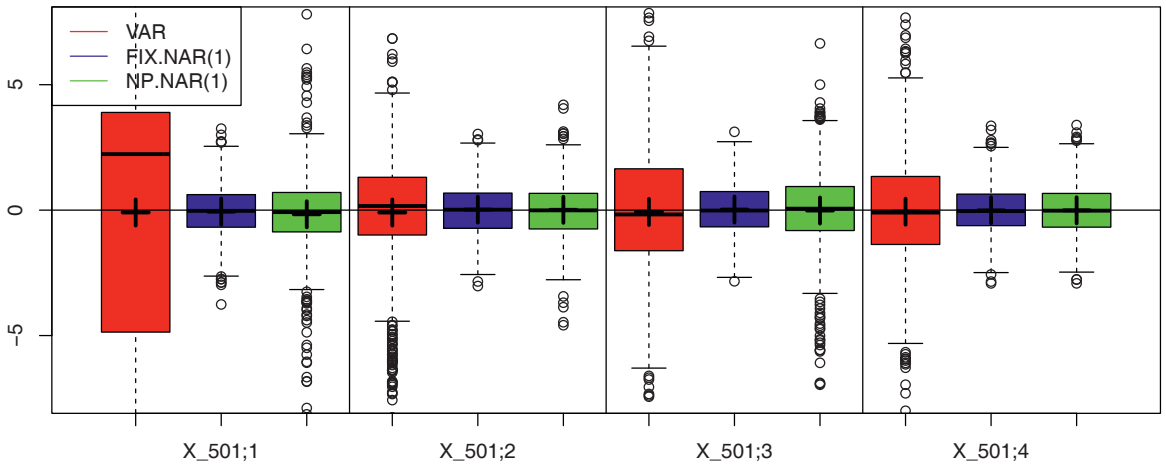


Figure 3.3: One-step-forecasting error for $\hat{\underline{X}}_{501}$ based on $\underline{X}_1, \ldots, \underline{X}_{500}$ and $Ad_1, \ldots, Ad_{500}$ of process (3.4.2). The crosses display the mean forecasting error.

In the next example a DSNAR(1) process $\underline{\mathbf{X}}$ is investigated where f is not given by fixed edge weights. A network with 6 vertices is considered; the edges are independent from each other and $\mathbf{Ad}$ is a Markovian process given by

$$\big(P(Ad_{t;ij} = 1 | Ad_{t-1;ij} = 1)\big)_{i,j=1,\ldots,d} = \begin{pmatrix} 1 & 0 & 0.25 & 0 & 0 & 0 \\ 0 & 1 & 0.80 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.80 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.75 & 0.75 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\big(P(Ad_{t;ij} = 1 | Ad_{t-1;ij} = 0)\big)_{i,j=1,\ldots,d} = \begin{pmatrix} 1 & 0 & 0.25 & 0 & 0 & 0 \\ 0 & 1 & 0.20 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.80 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.10 & 0.05 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{3.4.3}$$

The components of the innovations process are independent and $\underline{\varepsilon}_{1;1} \sim \exp(1/5), \underline{\varepsilon}_{1;2} \sim -\exp(1/5)$ and $\underline{\varepsilon}_{1;j} \sim \mathcal{N}(0,1), j = 3, \ldots, 6$. The time series $\underline{\mathbf{X}}$ is given by

$$\underline{X}_{t;j} = f(Ad_{t-1;ij})\underline{X}_{t-1;i} + \underline{\varepsilon}_{t;j}, \text{ where } f(X) = \begin{pmatrix} g_1(X_{1\cdot}, X_{\cdot 1}) \\ \vdots \\ g_d(X_{d\cdot}, X_{\cdot d}) \end{pmatrix} \tag{3.4.4}$$

and the j-th row of f is given by

$$g_j : \mathbb{R}^{2d} \to \mathbb{R}^{1\times d}, g_j(X_{j\cdot}, X_{\cdot j}) = \begin{cases} X_{\cdot j} - e_j & , \sum_{s\neq j} X_{js} > 0, \\ X_{\cdot j} - 0.05 e_j & , \text{ else,} \end{cases} \quad j = 1, \dots, d.$$

A realization of the time series $\underline{\mathbf{X}}$ as well as of the network is shown in Figure 3.6. The function f works in the following way: As long there is no edge to another vertex, the corresponding time series charges up load, component 1 positively and component 2 negatively and the other components keep their charge (or more precisely 95% of it plus some noise). If there is now an edge to another vertex present, the load is transferred to this vertex. These edges are directed and the load flows in the direction of the edges. Hence, the load of components 1 and 2 flows through 3 and 4 to the end vertices 5 and 6. The function f fulfills the requirements of *NP.NAR* but not the ones of *FIX.NAR*. Nevertheless, a forecasting of $\underline{X}_{n+1}$ is performed with *NP.NAR*, *FIX.NAR* and *VAR* based on $\underline{X}_1, \dots, \underline{X}_n$ and $Ad_1, \dots, Ad_n$. The one-step-forecasting error is displayed in Figure 3.5 and the mean squared forecasting error for each component is given by Table 3.3. Notice that the innovations of the first two components are exponentially distributed with $\mathrm{Var}(\underline{\varepsilon}_{1;1}) = \mathrm{Var}(\underline{\varepsilon}_{1;2}) = 25$. That is why the median forecasting error for these components is not near 0, only the mean forecasting error is; in Figure 3.5 the mean is displayed by a cross. The valid method *NP.NAR* performs best and is able to reduce the forecasting error near the order of the innovations variance. Components 3 and 4 have more possible states than the other components which explains the higher forecast error for those components. Components 5 and 6 fit in the framework of *FIX.NAR* and for these components the method is able to reduce the forecasting variance to the variance of the innovations.

Table 3.3: Mean squared one-step forecasting error for $\hat{\underline{X}}_{n+1}$ based on $\underline{X}_1, \dots, \underline{X}_n$ and $Ad_1, \dots, Ad_n$ of process (3.4.4). Note that the innovations variance is $(25, 25, 1, 1, 1, 1)$.

n	250			500			1000		
	VAR	*FIX.NAR*	*NP.NAR*	*VAR*	*FIX.NAR*	*NP.NAR*	*VAR*	*FIX.NAR*	*NP.NAR*
1	125	122	27	120	118	27	111	110	26
2	101	99	28	103	103	26	107	107	27
3	212	34	4	226	33	1	212	32	1
4	112	74	14	119	81	1	109	76	1
5	63	1	1	71	1	1	64	1	1
6	41	1	1	39	1	1	41	1	1

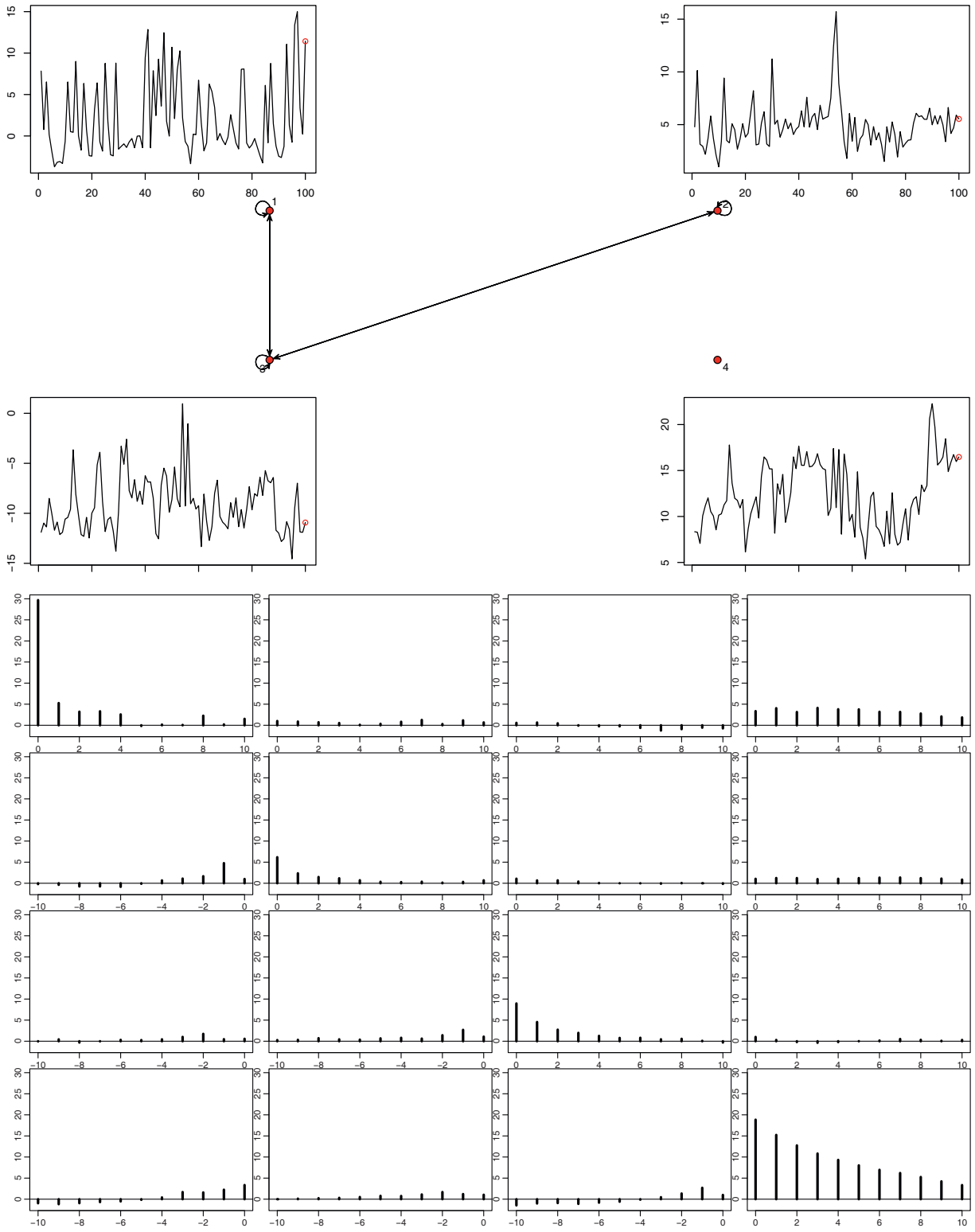


Figure 3.4: The upper figure presents a realization of the network of the example given by (3.4.1) and a realization of the time series X given by (3.4.2). Red dots indicate the current time point. This figure contains animation only visible on screen. The lower graphic presents the sample autocovariance function of X.

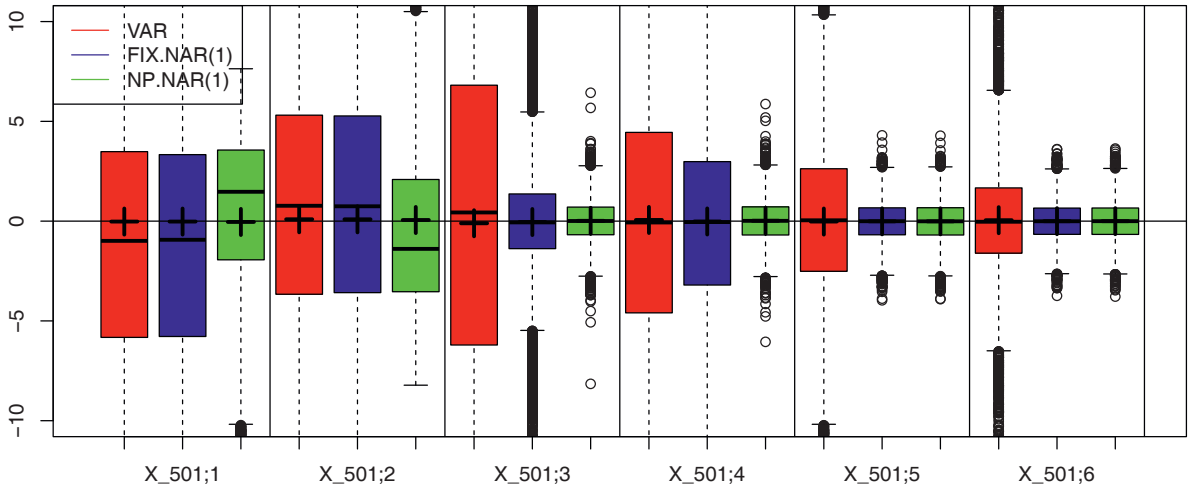


Figure 3.5: One-step-forecasting error for $\hat{\underline{X}}_{501}$ based on $\underline{X}_1, \ldots, \underline{X}_{500}$ and $Ad_1, \ldots, Ad_{500}$ of process (3.4.3). The crosses display the mean forecasting error. Note that the innovations of component 1 and 2 posses an exponential distribution.

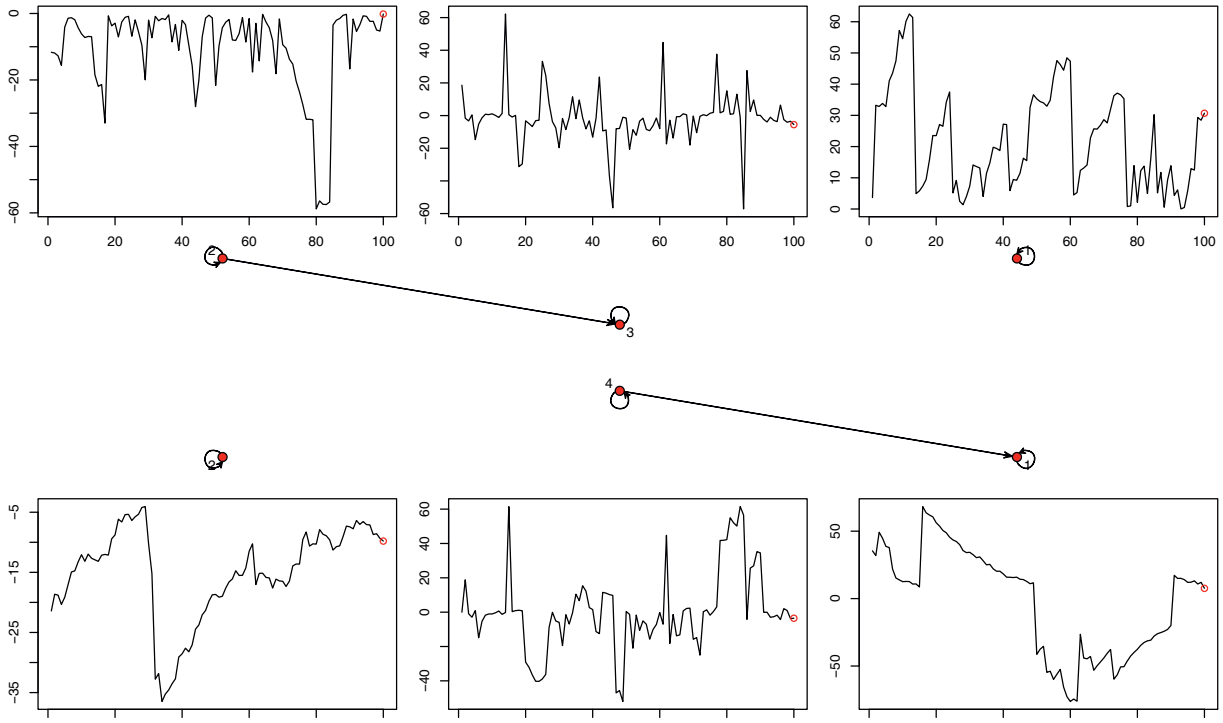


Figure 3.6: Realization of the network of the example given by (3.4.3); realization of the time series X given by (3.4.4). Red dots indicate the current time point. This figure contains animation only visible on screen.

In the next example a Separable Temporal Exponential Random Graph Model (STERGM), see Krivitsky and Handcock (2014) and also Krivitsky and Handcock (2016) for the used *R* package *tergm*,

with $d = 1000$ vertices is considered. Two types of networks are considered: a slow-varying network (dissolution-coefficient 8, formation-coefficient -13.3) and a fast-varying network (dissolution-coefficient 4, formation-coefficient -9.3). Both networks have a mean density of 0.005 which results in around 5000 edges. The slow-varying network has about 350 edge changes from $t = 1$ to $t = 100$, whereas the fast-varying network has about 8500 edge changes from $t = 1$ to $t = 100$. These two networks differ mainly in their dynamics, whereas their inner structure is similar as can be seen, for instance, in the out-degree distribution given in Figure 3.7. The out-degree distribution can be approximated by a normal distribution with a mean of 5 and a standard variation of 2.2. The in-degree distributions has a similar structure. Hence, no vertex takes a special role, which is why a homogeneous model seems appropriate. Thus, every component of the time series has the same parameters. The time series is given by

$$\underline{X}_t = 0.15 A d_{t-1}^{\top} \underline{X}_{t-1} + 5 + \underline{\varepsilon}_t, \text{ where } \underline{\varepsilon}_1 \sim \mathcal{N}(\underline{0}, I_{1000}). \tag{3.4.5}$$

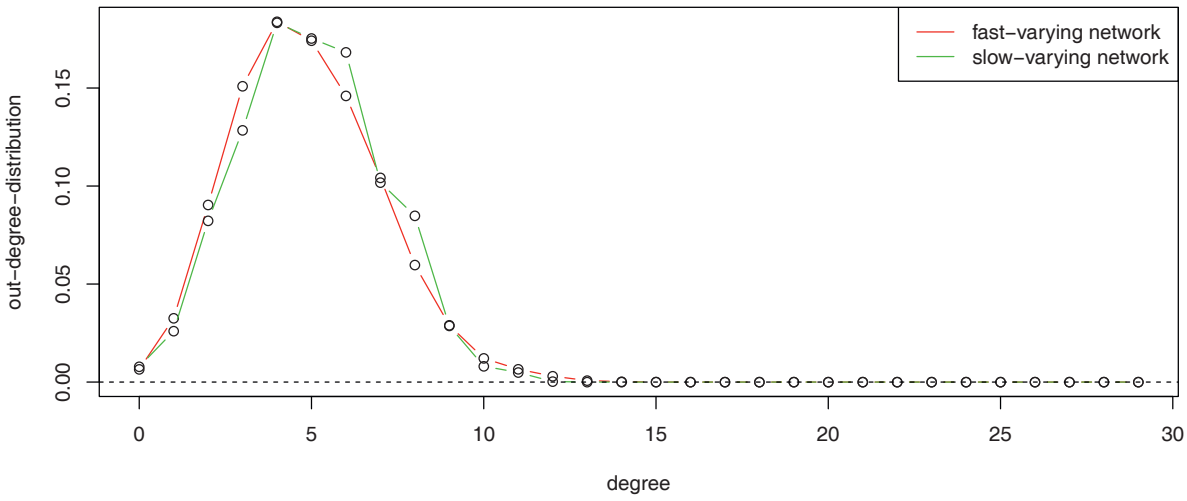


Figure 3.7: Out-degree-distributions of the slow-varying and fast-varying STERGMs with $d = 1000$ and density 0.005.

This setting is suited for *RAD.NAR* and it is used here with $g(X) = X^{\top}$. Furthermore, the method *FIX.NAR* is applied for forecasting. Due to the high-dimensional setting ($d = 1000$ regarding $n = 100$) a *VAR* approach cannot be applied. Instead, a reduced *VAR* approach is used. The model structure (3.4.5) implies that the components of non-connected vertices are independent or more precisely only components $i \in \{s \in \{1, \dots, d\} : \sup_{k \le t} A d_{k;sj} > 0\}$ can influence $\underline{X}_{\cdot;j}$. Thus, to perform a forecast of $\underline{X}_{n;j}, j = 1, \dots, d$ based on $\underline{X}_1, \dots, \underline{X}_n$ and $A d_1, \dots, A d_n$, we only consider

$(\underline{X}_{1;s},\dots,\underline{X}_{n;s})_{s\in S_j}$, where $S_j := \{s \in \{1,\dots,d\} : \max_t Ad_{t;sj} > 0\}$. Hence, this *VAR* approach uses the observed network to reduce the number of parameters. However the network dynamics cannot be appropriately captured by this VAR approach. Since the components are homogeneous, the average error over all components is considered. The average squared forecasting error for $\underline{X}_{101}$ and $\underline{X}_{201}$ is displayed in Table 3.4. Since in fast-varying networks more edges occur than in slow-varying networks, in the fast-varying network setting *FIX.NAR* as well as reduced *VAR* have more non-zero parameters. That is why these approaches perform considerably worse in fast-varying networks. Besides that, since this setting is tailor-made for the *RAD.NAR*, it performs best and this approach is able to reduce the forecasting error of size of the innovation error. Notice further that *RAD.NAR* is the only presented model which benefits from the high number of vertices. *FIX.NAR* is consistent in this setting. However, *FIX.NAR* has many more parameters than *RAD.NAR*, which is why it performs worse for this small sample size. Notice that in this network every edge can occur at some time point. Thus, if a longer time period is observed, the number of adjacent vertices to a given vertex j increases; $S_{t;j} = \sum_{s=1}^{d} \mathbb{1}_{\{\sup_{k\le t} Ad_{k;sj}>0\}}$ is monotonic in t and converges to $d-1$ (if self-loops are not possible). In the fast-varying network 14 adjacent vertices are observed on average over the time period $t = 1,\dots,100$, whereas 23 adjacent vertices are observed over the time period $t = 1,\dots,200$. Hence, a larger time period may increase the number of non-zero parameters for *FIX.NAR* and reduced *VAR*. That could explain why *FIX.NAR* does not benefit from the doubled sample size.

Table 3.4: Average one-step-ahead forecasting error, $1/d\sum_{j=1}^{d} E(\hat{\underline{X}}_{n+1;j} - \underline{X}_{n+1;j})^2$, for $\hat{\underline{X}}_{n+1}$ based on $\underline{X}_1,\dots,\underline{X}_n$ and $Ad_1,\dots,Ad_n$ of process (3.4.5).

network	fast-varying network		slow-varying network	
n	100	200	100	200
reduced *VAR*	5.2	5.1	1.4	1.6
FIX.NAR	3.5	3.9	1.2	1.2
RAD.NAR	1.0	1.0	1.0	1.0

3.5 Real Data Example

Here We consider data given by a play of the German card game *Doppelkopf*.[1] It is played by four players and the main focus here is on the overall score. Hence, we are not going into detail regarding the rules of the game and how to play it, especially, because the rules differ from region to region. The important aspect is that it is played in teams. The teams are chosen by the cards and therefore the teams are chosen randomly. A team wins or loses together and each member of the team gets

[1] For more details on *Doppelkopf* refer to `http://www.doko-verband.de/Regeln__Ordnungen.html`, `https://de.wikipedia.org/wiki/Doppelkopf` and `https://en.wikipedia.org/wiki/Doppelkopf`

the same score (displayed in the game column) which is added by winning and subtracted by losing. Thus, the scoreboard displays also the information of who played with whom. Notice that it is possible that one player plays versus three others.

Table 3.5: Scoreboard of a play of the German card game 'Doppelkopf'

#	Hinnerk	Maddy	Jonas	Annika	game
1	2	-2	2	-2	2
2	8	-8	8	-8	6
3	18	-38	18	2	10
4	14	-42	14	14	4
5	4	-12	4	4	10
6	0	-8	8	0	4
7	-12	-20	20	12	12
8	-20	-28	28	20	8
9	-27	-21	35	13	7
10	-29	-23	37	15	2
11	-25	-27	41	11	4
12	-27	-25	43	9	2
13	-3	-33	35	1	8
14	-7	-29	39	-3	4
15	-1	-23	45	-21	6
16	0	-24	44	-20	1
17	-16	-40	60	-4	16
18	-20	-36	56	0	4
19	-22	-34	58	-2	2
20	-21	-33	57	-3	1
21	-25	-29	61	-7	4

Here the score given by Table 3.5 is considered as a multivariate time series $(\underline{X}_t)$ and the aim is to predict the score. Figure 3.8 presents the process in the usual way of time series and network. Hence, this figure shows all the given observations of $(\underline{X}_t)$ and (Ad_t). In order to be a valid score, the score of all player has to be sum up to zero, hence $\sum_{s=1}^{d} \underline{X}_{t;s} = 0$ for all t. Thus, even though we observe a 4 dimensional time series, it is only of 3 dimensions. For modeling the score with an DSNAR(1), denoted as *NAR*, we use this relation and set $\underline{X}_{t;4} = -\sum_{s=1}^{3} \underline{X}_{t;s}$. The other components, $\underline{X}_{t;j}, j = 1,2,3$, are given by

$$\underline{X}_{t;j} = \underline{X}_{t-1;j} + \sum_{s \neq j} \alpha_{js} \tilde{A}d_{t,js} \underline{X}_{t-1;j} + \mu_j + \underline{\varepsilon}_{t;j}, \tag{3.5.1}$$

where $\tilde{A}d_{t,ij} =$ #team members (usually 2) if player i and player j are on the same team for game t and $\tilde{A}d_{t,ij} = -$#opponents if they are opponents for game t. This model has $9+3$ parameters. In the same manner we consider the following *VAR* model given by $X_{t;4} = -\sum_{j=1}^{3} X_{t;j}$ and

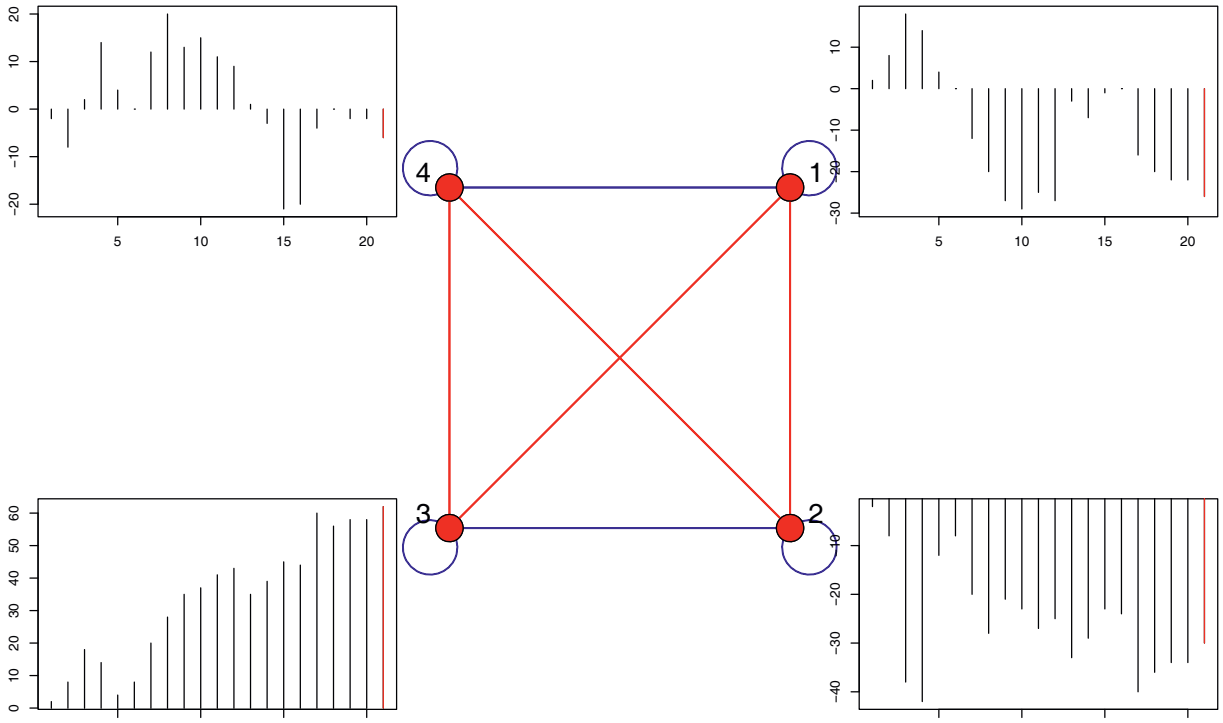


Figure 3.8: Doppelkopf data, 3.6, in time series representation. Blue edges indicate that the connected vertices are on the same team, whereas red edges indicate that the connected vertices are opponents. This figure contains animation only visible on screen.

$$\underline{X}_{t;j} = X_{t-1;j} + \sum_{s \neq j} a_{js} \underline{X}_{t;s} + \mu_s + \varepsilon_{t;j}, j = 1,2,3. \tag{3.5.2}$$

This VAR model also has 12 parameters and as in the DSNAR model the coefficient determining the influence of $X_{t-1;j}$ on $X_{t,j}$ is set to 1 for all j. Furthermore, a structural DSNAR(1) is considered, which is given by

$$\underline{X}_t = \underline{X}_{t-1} + \begin{pmatrix} Ad_{t;12} & Ad_{t;13} & Ad_{t;14} \\ Ad_{t;21} & -Ad_{t;24} & -Ad_{t;23} \\ -Ad_{t;34} & Ad_{t;31} & -Ad_{t;32} \\ -Ad_{t;43} & -Ad_{t;42} & Ad_{t;41} \end{pmatrix} \begin{pmatrix} e_{t;1} \\ e_{t;2} \\ e_{t;3} \end{pmatrix} =: \underline{X}_{t-1} + f(Ad_t)e_t, \tag{3.5.3}$$

where $Ad_{t,ij} = 1$ if player i and player j are on the same team and else $Ad_{t,ij} = 0$. This model can be written as $Z_t = \underline{X}_t - \underline{X}_{t-1} = f(Ad_t)e_t$. Hence, this model can be seen as DSNMA(0) model and the parameters of this model are the innovation's mean. This model is denoted as *NMA*. As a further benchmark we consider a forecast by two simple approaches, $\hat{\underline{X}}_{t+1} = \underline{X}_t$, denoted as *NAIV*, and

$\hat{\underline{X}}_{t+1} = \underline{X}_t + 1/t\underline{X}_t$, denoted as *NAIV2*. The forecast results are given in Table 3.6 and the forecast error is given in Table 3.7. The network time series models, NAR and NMA, give on average for these 3 time points considered the best forecast. The NAIV approach is also upfront, whereas VAR performs worse. However, the sample size is considerably small so that a reliable statement cannot be made. Note that the prediction $\hat{\underline{X}}_{t+1}$ given by the NAR and NMA uses the network structure at $t+1$. Prediction may not be the most interesting question to answer for this setting. Of interest is also the question who plays well with whom. Such questions can be easily answered by interpreting the parameters of the network time series models. For NMA, (3.5.3), we have $Ee_{t;1} =: \mu_{(1,2),-(3,4)}$ giving the playing performance of player 1 and 2, whereas, due to the symmetry of the game score, $-Ee_{t;1} = -\mu_{(1,2),-(3,4)}$ represents the playing performance of player 3 and 4. Similarly, $\mu_{(1,3),-(2,4)}$ represents the playing performance of player 1 and 3 and with minus sign for players 2 and 4. Based on the given data we obtain $\hat{\mu}_{(1,2),-(3,4)} = -9.5, \hat{\mu}_{(1,3),-(2,4)} = 2, \mu_{(1,4),-(2,3)} = -2.75$.

Table 3.6: Predicted scores of a play of the German card game 'Doppelkopf' at $t = 19, 20, 21$.

	$t=19$				$t=20$				$t=21$			
Player	1	2	3	4	1	2	3	4	1	2	3	4
$\underline{X}_t$	-26	-30	62	-6	-22	-34	58	-2	-22	-34	58	-2
NAIV	-22	-34	58	-2	-22	-34	58	-2	-20	-36	56	0
NAIV2	-23	-36	61	-2	-23	-36	61	-2	-21	-38	59	0
VAR	-22	-35	60	-3	-22	-36	60	-3	-20	-35	57	-1
NAR	-23	-33	60	-3	-24	-33	60	-4	-22	-35	59	-2
NMA	-25	-31	61	-5	-25	-31	61	-5	-23	-33	59	-3

Table 3.7: The forecast error of a play of the German card game 'Doppelkopf' at $t = 19, 20, 21.$, $\|\underline{X}_t - \hat{\underline{X}}_t\|_2$, is given in the lower table.

	NAIV	NAR	NAIV2	VAR	NMA
19	8	5	8	8	3
20	0	3	4	3	6
21	4	1	5	3	2

3.6 Conclusions

In this chapter the doubly stochastic framework has been used to model a multivariate time series on a dynamic network with static vertices. In this framework network linear processes and network autoregressive processes have been defined. Independence of the time series' innovations and the network enables the possibility to model time series and network separately. This gives flexibility in the sense that one is not limited to a specific network model. By restricting to α-mixing networks this framework becomes feasible and statistical results can be derived. For instance, based on observations of the time series and the network consistency of estimators for the parameters of a network AR(1) model is shown. These estimators can be used to do forecasting and, as can be seen in the numerical examples, the benefit of using the additional structure can be quite large. It is further possible to interpret the parameters to gain new insight as can be seen in the real data example.

3.7 Proofs

Proof of Lemma 3.2.2. (i) and (ii) gives the existence of the L_2-Limit of X_t, so that it can be written as $\underline{X}_t = \sum_{j=0}^{\infty} B_{t,j}\underline{\varepsilon}_{t-j}$. We have $B_{t,j} = f_j(Ad_{t-1}, \dots, Ad_{t-j})$ and $\{\underline{\varepsilon}_t, t \in \mathbb{Z}\}$ is i.i.d and independent to the stationary process $\{Ad_t, t \in \mathbb{Z}\}$. Thus, $\{\underline{\varepsilon}_t, t \in \mathbb{Z}\}$ and $(\mathrm{vec}(B_{t,j}, j \in \mathbb{N}))_{t\in\mathbb{Z}}$ are independent. We have $\underline{\mu}_x = \sum_{j=0}^{\infty} EB_{0,j}\underline{\mu}$ and for the autocovariance function

$$\begin{aligned}\Gamma_X(h) &= \sum_{j=0}^{\infty}\sum_{s=0}^{\infty}\left(E\left(B_{t+h,j}\underline{\varepsilon}_{t+h-j}\underline{\varepsilon}_{t-s}^{\top}B_{t,s}^{\top}\right) - E\left(B_{t+h,j}\underline{\mu}\underline{\mu}^{\top}B_{t,s}\right) + E\left(B_{t+h,j}\underline{\mu}\underline{\mu}^{\top}B_{t,s}\right) - E(B_{t+h,j})\underline{\mu}\underline{\mu}^{\top}E(B_{t,s}^{\top})\right)\\ &= \sum_{s=0}^{\infty} E\left(B_{h,s+h}\Sigma B_{0,s}^{\top}\right) + \sum_{j=0}^{\infty}\sum_{s=0}^{\infty}\mathrm{Cov}\left(B_{h,j}\underline{\mu}, B_{0,s}\mu\right), h \geq 0.\end{aligned}$$

□

Proof of Lemma 3.2.3. Since (3.2.7) defines a doubly stochastic linear process and due to (3.2.8) and (3.2.9) the assertion follows by Lemma 3.2.2. □

Proof of Lemma 3.2.4. Let $\tilde{A}_t = \prod_{s=(t-1)q}^{tq} A_{-s}$. Hence, $\prod_{s=1}^{j} A_{-s} = (\prod_{s=1}^{\tilde{j}} \tilde{A}_s)\tilde{A}'_{\tilde{j}}$, where $\tilde{j} = \lfloor j/q \rfloor$ and $\tilde{A}'_{\tilde{j}} = \prod_{\tilde{j}}^{j} A_s$ denotes the remaining A_{-s}'s which do not make a full $\tilde{A}_s$. Let $E\log\|\tilde{A}_1\| < 0$. Then there exists a $\rho > 1$ so that $\log\rho + E\log\|\tilde{A}_1\| < 0$. Since **Ad** is α-mixing, we have that $\{\log\|\tilde{A}_t\|, t \in \mathbb{Z}\}$ is α-mixing as well. Consequently, $\{\log\|\tilde{A}_t\|, t \in \mathbb{Z}\}$ is ergodic, see Bradley (2007, Proposition 2.8, 2.6). Hence, as $\tilde{j} \to \infty$, $1/\tilde{j}\sum_{s=1}^{\tilde{j}} \log\rho + \log\|\tilde{A}_{-s}\| \to \log\rho + E\log\|\tilde{A}_1\| < 0$ a.s.. Thus, as $\tilde{j} \to \infty$, $\prod_{s=1}^{\tilde{j}} \rho\|\tilde{A}_{-s}\| = \exp(\sum_{s=1}^{\tilde{j}} \log\rho + \log\|\tilde{A}_{-s}\|) \to 0$ a.s.. Since

$$\|\sum_{j=0}^{\infty}|\prod_{s=1}^{j} A_{t-s}| = \|\sum_{j=0}^{\infty}|(\prod_{s=1}^{\tilde{j}}\tilde{A}_{t-s})\tilde{A}'_{\tilde{j}}|\| \leq \sum_{j=0}^{\infty}\|\tilde{A}'_{\tilde{j}}\|\prod_{s=1}^{\tilde{j}}\|\tilde{A}_{t-s}\| = \sum_{j=0}^{\infty}\|\tilde{A}'_{\tilde{j}}\|\rho^{-\tilde{j}}\rho^{\tilde{j}}\prod_{s=1}^{\tilde{j}}\|\tilde{A}_{t-s}\| < \infty \text{ a.s.},$$

we have (3.2.8). Since $\|\sum_{j=0}^{\infty} E|\prod_{s=1}^{j} A_{-s}\Sigma_d(\prod_{s=1}^{j} A_{-s})^{\top}|\| \leq \|\Sigma_d\|\sum_{j=0}^{\infty}\|\tilde{A}'_{\tilde{j}}\|^2 E\prod_{s=1}^{\tilde{j}}\|\tilde{A}_{-s}\|^2$, (3.2.9) follows by $E\log\|\tilde{A}_1\|^2 < 0$ in the same steps as above. □

Proof of Theorem 3.3.1. For a fixed $M \in \mathbb{N}$ we consider the approximation given by $\underline{X}_{t,M} = \sum_{j=0}^{M} B_{t,j}\underline{\varepsilon}_{t-j}$. We show the asymptotic normality for $1/\sqrt{n}\sum_{t=1}^{n}\underline{X}_{t,M}$. If the approximation is sufficiently close, the assertion follows by Theorem 4.2 of Billingsley (1968).

Since $B_t = f_j(Ad_t, \dots, Ad_{t-j-1})$, for some measurable functions $g, \tilde{g}$ we have $\underline{X}_{t,M} = g(B_{t,0}, \dots, B_{t,M}, \underline{\varepsilon}_t, \dots, \underline{\varepsilon}_{t-M}) = \tilde{g}(Ad_t, \dots, Ad_{t-M}, \underline{\varepsilon}_t, \dots, \underline{\varepsilon}_{t-M})$. Thus, we have $\sigma(\underline{X}_{t,M}) \subseteq \sigma(\sigma(Ad_t, \dots, Ad_{t-M}) \cup \sigma(\underline{\varepsilon}_t, \dots, \underline{\varepsilon}_{t-m}))$. This gives us $\sigma(\underline{X}_{k,M}, k \leq t) \subseteq \sigma(\sigma(Ad_k, k \leq t) \cup \sigma(\underline{\varepsilon}_k, k \leq t))$ and $\sigma(\underline{X}_{k,M}, k \geq t) \subseteq \sigma(\sigma(Ad_k, k \geq t - M) \cup \sigma(\underline{\varepsilon}_k, k \geq t - M))$. Since **Ad** and $\underline{\varepsilon}$ are independent with Theorem 6.1 of Bradley (2007) we have $\alpha((\underline{X}_{t,M}), n) \leq \alpha(\mathbf{Ad}, n - M) + \alpha(\underline{\varepsilon}, n - M)$ and due to the i.i.d structure of $\underline{\varepsilon}$ we have $\alpha((\underline{X}_{t,M}), n) \leq \alpha(\mathbf{Ad}, n - M)$ for $n > M$. Hence, the strong

mixing conditions of **Ad** transfer to $(\underline{X}_{t,M})$, due to Assumption 1 we have $\sum_{n=1}^{\infty}\alpha((\underline{X}_{t,M}),n) \leq MC + \sum_{n=M}^{\infty}\alpha(\mathbf{Ad}, n-M)n^3 < \infty$. The autocovariance of $(\underline{X}_{t,M})$ is given by

$$\Gamma_{x,M}(h) = \sum_{s=0}^{M-|h|} EB_{h,s+h}\Sigma_d B_{0,s}^{\top} + \sum_{j_1=0}^{M}\sum_{j_2=0}^{M}\mathrm{Cov}(B_{h,j}\mu, B_{0,j_2}\mu).$$

We use the Cramér-Wold-device to show the asymptotic normality of $1/\sqrt{n}\sum_{t=1}^{n}\underline{X}_{t,M}$. Thus, we consider $c \in \mathbb{R}^d$ arbitrary and show that

$$\sqrt{N}(c^{\top}(\bar{X}_{n,M} - E\underline{X}_{0,M})) \xrightarrow{D} \mathcal{N}\left(0, \sum_{h\in\mathbb{Z}} c^{\top}\Gamma_{X,M}(h)c\right), \text{ as } N \to \infty.$$

For this we use Corollary 10.22 of Bradley (2007). We have, as $n \to \infty$,

$$E(\sum_{t=1}^{n} c^{\top}(\underline{X}_{t,M} - E(\underline{X}_{t,M})))^2 = \sum_{h=-n+1}^{n-1}(n-|h|)c^{\top}\Gamma_{x,M}(h)c \to \infty,$$

and

$$E(1/\sqrt{n}\sum_{t=1}^{n} c^{\top}(\underline{X}_{t,M} - E(\underline{X}_{t,M})))^2 \to \sum_{h\in\mathbb{Z}} c^{\top}\Gamma_{x,M}(h)c.$$

Since $(c^{\top}\underline{X}_{t,M})$ fulfills the required strong mixing condition, it remains to show that $E|c^{\top}(\underline{X}_0 - E\underline{X}_0)|^4 < \infty$. To see this, we have with Assumption 2 and 3 and since Ad and $(\underline{\varepsilon}_t)$ are independent

$$\begin{aligned}
E|c^{\top}(\underline{X}_0 - E\underline{X}_0)|^4 = &\sum_{j_1,\ldots,j_4=0}^{M}\sum_{i_1,\ldots,i_4=1}^{d}\sum_{s_1,\ldots,s_4=1}^{d} |c_{i_1}c_{i_2}c_{i_3}c_{i_4} \\
&\Big[E\underline{\varepsilon}_{-j_1;s_1}\underline{\varepsilon}_{-j_2;s_2}\underline{\varepsilon}_{-j_3;s_3}\underline{\varepsilon}_{-j_4;s_4}\mathrm{Cov}(B_{0,j_1;i_1s_1}B_{0,j_2;i_2s_2}, B_{0,j_3;i_3s_3}B_{0,j_4;i_4s_4}) + \\
&\mathrm{Cov}(\underline{\varepsilon}_{-j_1;s_1}\underline{\varepsilon}_{-j_2;s_2}, \underline{\varepsilon}_{-j_3;s_3}\underline{\varepsilon}_{-j_4;s_4})E(B_{0,j_1;i_1s_1}B_{0,j_2;i_2s_2})E(B_{0,j_3;i_3s_3}B_{0,j_4;i_4s_4})\Big]| \\
&\leq \sum_{j_1,\ldots,j_4=0}^{M}\sum_{i_1,\ldots,i_4=1}^{d}\sum_{s_1,\ldots,s_4=1}^{d} |c_{i_1}c_{i_2}c_{i_3}c_{i_4}| \\
&\Big((E\underline{\varepsilon}_{0;s_1}^4)(E\underline{\varepsilon}_{0;s_2}^4)(E\underline{\varepsilon}_{0;s_3}^4)(E\underline{\varepsilon}_{0;s_4}^4)(EB_{0,j_1;i_1s_1}^4)(EB_{0,j_2;i_2s_2}^4)(EB_{0,j_3;i_3s_3}^4)(EB_{0,j_4;i_4s_4}^4)\Big)^{1/4} \\
&= M^4C < \infty.
\end{aligned}$$

Thus, we have the asymptotic normality of $\sqrt{n}\bar{X}_{n,M}$. Since $\sum_{h\in\mathbb{Z}}\sum_{s=0}^{\infty}|EB_{h,s+h}\Sigma_d B_{0,s}^{\top}| < \infty$ and $\sum_{h\in\mathbb{Z}}\sum_{j_1=0}^{\infty}\sum_{j_2=0}^{\infty}|\mathrm{Cov}(B_{h,j}\mu, B_{0,j_2}\mu)| < \infty$, see Assumption 3, we have $\sum_{h\in\mathbb{Z}}\Gamma_{X,M}(h) \to \sum_{h\in\mathbb{Z}}\Gamma_X(h)$, as $M \to \infty$. Hence, the asymptotic variance of $\sqrt{n}\bar{X}_{n,M}$ converges to the asymptotic variance of $\sqrt{n}\bar{X}_n$. It remains to show that the approximation is sufficiently close. For $\delta > 0$ we have

$$
\begin{aligned}
&\lim_{M\to\infty}\limsup_{n\to\infty} P\left(|c^\top \frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\underline{X}_t - E\underline{X}_0 - (\underline{X}_{t-M} - E\underline{X}_{t,M}))| > \delta\right)\\
&\leq \lim_{M\to\infty}\limsup_{n\to\infty}\sum_{h=-n+1}^{n-1}\frac{n-|h|}{n}\sum_{j_1,j_2=M+1}^{\infty} c^\top \mathrm{Cov}(B_{0,j_1}\underline{\varepsilon}_{-j_2}, B_{h,j_2}\underline{\varepsilon}_{h-j_2})c/\delta^2\\
&= \lim_{M\to\infty}\limsup_{n\to\infty}\sum_{h=-n+1}^{n-1}\frac{n-|h|}{n}c^\top\left(\sum_{s=M+1}^{\infty} EB_{0,s}\Sigma_d B_{h,s+|h|}^\top + \sum_{j_1,j_2=M+1}^{\infty}\mathrm{Cov}(B_{0,s}\underline{\mu}, B_{h,j}\underline{\mu})\right)c/\delta^2\\
&\leq \lim_{M\to\infty}\sum_{s=M+1}^{\infty}\sum_{h\in\mathbb{Z}} c^\top|EB_{0,s}\Sigma_d B_{h,s+h}^\top|c/\delta^2 + \sum_{j_1,j_2=M+1}^{\infty}\sum_{h\in\mathbb{Z}} c^\top|\mathrm{Cov}(B_{0,s}\underline{\mu}, B_{h,j}\underline{\mu})|c/\delta^2 = 0,
\end{aligned}
$$

due to Assumption 3. □

Proof of Theorem 3.3.2. In order to simplify notation, let $h \geq 0$. Let $\tilde{\Gamma}(h) = 1/n\sum_{t=1}^{n-h}(\underline{X}_{t+h} - \underline{\mu}_x)(\underline{X}_t - \underline{\mu}_x)^\top$. Since Assumptions 1 to 3 ensure that Theorem 3.3.1 gives $1/n\sum_{t=1}^{n}\underline{X}_t = \bar{X}_n = \underline{\mu}_x + \mathcal{O}_P(n^{-1/2})$ and since we have

$$
\begin{aligned}
\widehat{\Gamma}(h) &= \frac{1}{n}\sum_{t=1}^{n-h}(\underline{X}_{t+h} - \underline{\mu}_x + \underline{\mu}_x - \bar{X}_n)(\underline{X}_t - \underline{\mu}_x + \underline{\mu}_x - \bar{X}_n)^\top\\
&= \frac{1}{n}\sum_{t=1}^{n-h}(\underline{X}_{t+h} - \underline{\mu}_x)(\underline{X}_t - \underline{\mu}_x)^\top + (\underline{\mu}_x - \bar{X}_n)(\underline{X}_t - \bar{\underline{X}}_n)^\top + (\underline{X}_{t+h} - \underline{\mu}_x)(\underline{\mu}_x - \bar{X}_n)^\top,
\end{aligned}
$$

$\widehat{\Gamma}(h) = \tilde{\Gamma}(h) + \mathcal{O}_P(h/n^{-1/2})$ follows immediately. Furthermore, we have $E\tilde{\Gamma}(h) = \Gamma(h) + O(h/n)$. In the following we show that the variance of $\tilde{\Gamma}(h)$ is of order $\mathcal{O}(1/n)$ and consequently $\tilde{\Gamma}(h)$ as well as $\widehat{\Gamma}(h)$ are consistent estimators for $\Gamma(h)$:

$$
\begin{aligned}
\mathrm{Cov}(\tilde{\Gamma}(h)_{j_1,j_2}, \tilde{\Gamma}(h)_{j_3,j_4}) &= \mathrm{Cov}\left(e_{j_1}^\top\frac{1}{n}\sum_{t=1}^{n-h}(\underline{X}_{t+h} - \underline{\mu}_x)(\underline{X}_t - \underline{\mu}_x)^\top e_{j_2}, e_{j_3}^\top\frac{1}{n}\sum_{t=1}^{n-h}(\underline{X}_{t+h} - \underline{\mu}_x)(\underline{X}_t - \underline{\mu}_x)^\top e_{j_4}\right)\\
&= \sum_{t_1,t_2=1}^{n-h}\mathrm{Cov}\left(e_{j_1}^\top\frac{1}{n}(\underline{X}_{t_1+h} - \underline{\mu}_x)(\underline{X}_{t_1} - \underline{\mu}_x)^\top e_{j_2}, e_{j_3}^\top\frac{1}{n}(\underline{X}_{t_2+h} - \underline{\mu}_x)(\underline{X}_{t_2} - \underline{\mu}_x)^\top e_{j_4}\right)\\
&= \frac{1}{n^2}\sum_{t_1,t_2=1}^{n-h}\sum_{s_1,s_2,s_3,s_4=0}^{\infty}\mathrm{Cov}\Big(e_{j_1}^\top(B_{t_1+h,s_1}\underline{\varepsilon}_{t_1+h-s_1} - EB_{0,s_1}\mu)(B_{t_1,s_2}\underline{\varepsilon}_{t_1-s_2} - EB_{0,s_2}\mu)^\top e_{j_2},\\
&\qquad e_{j_3}^\top(B_{t_2+h,s_3}\underline{\varepsilon}_{t_2+h-s_3} - EB_{0,s_3}\mu)(B_{t_2,s_4}\underline{\varepsilon}_{t_2-s_4} - EB_{0,s_4}\mu)^\top e_{j_4}\Big).
\end{aligned}
$$

The innovations $\underline{\varepsilon}_t$ are i.i.d. and therefore we divide the last term on the right hand side into five terms. For each moment structure of the innovations these are: all indices are equal, 3 indices are equal, 2 different pairs, 2 indices are equal, and all indices are different. We show that each case is of order $\mathcal{O}(1/n)$. These terms can be bounded by applying the Cauchy-Schwarz-inequality and

the boundedness follows by moment and mixing conditions given by Assumptions 1 to 3. We begin with the case that all indices are equal. We have

$$\begin{aligned}
&\frac{1}{n^2}\sum_{t_1,t_2=1}^{n-h}\Bigg[\sum_{s=0}^{\infty}\mathrm{Cov}\Big(e_{j_1}^\top(B_{t_1+h,s}\underline{\varepsilon}_{t_1+h-s}-EB_{0,s}\mu)(B_{t_1,s-h}\underline{\varepsilon}_{t_1-s+h}-EB_{0,s-h}\mu)^\top e_{j_2},\\
&\quad e_{j_3}^\top(B_{t_2+h,t_2-t_1+s}\underline{\varepsilon}_{t_1+h-s}-EB_{0,t_2-t_1+s}\mu)(B_{t_2,t_2-t_1+s-h}\underline{\varepsilon}_{t_1+h-s}-EB_{0,t_2-t_1+s-h}\mu)^\top e_{j_4}\Big)\Bigg]\\
&=\frac{1}{n}\sum_{l=-n+h+1}^{n-h-1}\frac{n-|l|}{n}\Bigg[\sum_{s=0}^{\infty}\mathrm{Cov}\Big(e_{j_1}^\top(B_{h,s}\underline{\varepsilon}_0-EB_{0,s}\mu)(B_{0,s-h}\underline{\varepsilon}_0-EB_{0,s-h}\mu)^\top e_{j_2},\\
&\quad e_{j_3}^\top(B_{l+h,l+s}\underline{\varepsilon}_0-EB_{0,l+s}\mu)(B_{l,l+s-h}\underline{\varepsilon}_0-EB_{0,l+s-h}\mu)^\top e_{j_4}\Big)\Bigg]\\
&\le\frac{1}{n}\sum_{l=-n+h+1}^{n-h-1}\frac{n-|l|}{n}\sum_{s=0}^{\infty}\Big(E\Big[e_{j_1}^\top(B_{h,s}\underline{\varepsilon}_0-EB_{0,s}\mu)(B_{0,s-h}\underline{\varepsilon}_0-EB_{0,s-h}\mu)^\top e_{j_2}\Big]^2\Big)^{1/2}\\
&\quad\Big(E\Big[e_{j_3}^\top(B_{l+h,l+s}\underline{\varepsilon}_0-EB_{0,l+s}\mu)(B_{l,l+s-h}\underline{\varepsilon}_0-EB_{0,l+s-h}\mu)^\top e_{j_4}\Big]^2\Big)^{1/2}\\
&\le\frac{1}{n}\sum_{s=0}^{\infty}\Big(E\Big[e_{j_1}^\top(B_{h,s}\underline{\varepsilon}_0-EB_{0,s}\mu)(B_{0,s-h}\underline{\varepsilon}_0-EB_{0,s-h}\mu)^\top e_{j_2}\Big]^2\Big)^{1/2}\\
&\quad\sum_{l=0}^{\infty}\Big(E\Big[e_{j_3}^\top(B_{h,l}\underline{\varepsilon}_0-EB_{0,l}\mu)(B_{0,l-h}\underline{\varepsilon}_0-EB_{0,l-h}\mu)^\top e_{j_4}\Big]^2\Big)^{1/2}=\mathcal{O}(1/n).
\end{aligned}$$

In the following we consider the case that 3 indices are equal and the fourth index is different from the others. We have

$$\begin{aligned}
&\frac{1}{n^2}\sum_{t_1,t_2=1}^{n-h}\Bigg[\sum_{s_1,s_2=0,s_1\neq s_2}^{\infty}\mathrm{Cov}\Big(e_{j_1}^\top(B_{t_1+h,s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,s_1}\mu)(B_{t_1,s_1-h}\underline{\varepsilon}_{t_1-s_1+h}-EB_{0,s_1-h}\mu)^\top e_{j_2},\\
&\qquad e_{j_3}^\top(B_{t_2+h,t_2-t_1+s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,t_2-t_1+s_1}\mu)(B_{t_2,s_2}\underline{\varepsilon}_{t_2-s_2}-EB_{0,s_2}\mu)^\top e_{j_4}\Big)\\
&\quad+\sum_{s_1,s_2=0,s_1\neq s_2}^{\infty}\mathrm{Cov}\Big(e_{j_1}^\top(B_{t_1+h,s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,s_1}\mu)(B_{t_1,s_1-h}\underline{\varepsilon}_{t_1-s_1+h}-EB_{0,s_1-h}\mu)^\top e_{j_2},\\
&\qquad e_{j_3}^\top(B_{t_2+h,s_2}\underline{\varepsilon}_{t_2+h-s_2}-EB_{0,s_2}\mu)(B_{t_2,t_2-t_1+s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,t_2-t_1+s_1}\mu)^\top e_{j_4}\Big)\\
&\quad+\sum_{s_1,s_2=0,s_1\neq s_2}^{\infty}\mathrm{Cov}\Big(e_{j_3}^\top(B_{t_1+h,s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,s_1}\mu)(B_{t_1,s_1-h}\underline{\varepsilon}_{t_1-s_1+h}-EB_{0,s_1-h}\mu)^\top e_{j_4},\\
&\qquad e_{j_1}^\top(B_{t_2+h,t_2-t_1+s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,t_2-t_1+s_1}\mu)(B_{t_2,s_2}\underline{\varepsilon}_{t_2+h-s_2}-EB_{0,s_2}\mu)^\top e_{j_2}\Big)\\
&\quad+\sum_{s_1,s_2=0,s_1\neq s_2}^{\infty}\mathrm{Cov}\Big(e_{j_3}^\top(B_{t_1+h,s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,s_1}\mu)(B_{t_1,s_1-h}\underline{\varepsilon}_{t_1-s_1+h}-EB_{0,s_1-h}\mu)^\top e_{j_4},\\
&\qquad e_{j_1}^\top(B_{t_2+h,s_2}\underline{\varepsilon}_{t_2+h-s_2}-EB_{0,s_2}\mu)(B_{t_2,t_2-t_1+s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,t_2-t_1+s_1}\mu)^\top e_{j_2}\Big)\Bigg],
\end{aligned}$$

which is equal to

$$\frac{1}{n}\sum_{l=-n+h+1}^{n-h-1}\frac{n-|l|}{n}\Bigg[\sum_{s_1,s_2=0,s_1\neq s_2}^{\infty}\operatorname{Cov}\Big(e_{j_1}^\top(B_{h,s_1}\underline{\varepsilon}_0-EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_0-EB_{0,s_1-h}\mu)^\top e_{j_2},$$
$$e_{j_3}^\top(B_{l+h,l+s_1}\underline{\varepsilon}_0-EB_{0,l+s_1}\mu)(B_{l,s_2}\underline{\varepsilon}_1-EB_{0,s_2}\mu)^\top e_{j_4}\Big)$$
$$+\sum_{s_1,s_2=0,s_1\neq s_2}^{\infty}\operatorname{Cov}\Big(e_{j_1}^\top(B_{h,s_1}\underline{\varepsilon}_0-EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_0-EB_{0,s_1-h}\mu)^\top e_{j_2},$$
$$e_{j_3}^\top(B_{l+h,s_2}\underline{\varepsilon}_1-EB_{0,s_2}\mu)(B_{l,l+s_1}\underline{\varepsilon}_0-EB_{0,l+s_1}\mu)^\top e_{j_4}\Big)$$
$$+\sum_{s_1,s_2=0,s_1\neq s_2}^{\infty}\operatorname{Cov}\Big(e_{j_3}^\top(B_{h,s_1}\underline{\varepsilon}_0-EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_0-EB_{0,s_1-h}\mu)^\top e_{j_4},$$
$$e_{j_1}^\top(B_{l+h,l+s_1}\underline{\varepsilon}_0-EB_{0,l+s_1}\mu)(B_{l,s_2}\underline{\varepsilon}_1-EB_{0,s_2}\mu)^\top e_{j_2}\Big)$$
$$+\sum_{s_1,s_2=0,s_1\neq s_2}^{\infty}\operatorname{Cov}\Big(e_{j_3}^\top(B_{h,s_1}\underline{\varepsilon}_0-EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_0-EB_{0,s_1-h}\mu)^\top e_{j_4},$$
$$e_{j_1}^\top(B_{l+h,s_2}\underline{\varepsilon}_1-EB_{0,s_2}\mu)(B_{l,l+s_1}\underline{\varepsilon}_0-EB_{0,l+s_1}\mu)^\top e_{j_2}\Big)\Bigg]=O(1/n).$$

To see this we take a closer look at the first part. The same arguments can also be applied to the other parts. Using the Cauchy-Schwarz-inequality and due to $E(B_{l,s}\underline{\varepsilon}_1-EB_{0,s}\mu)=0$ for all s,l, we get

$$\frac{1}{n}\sum_{l=-n+h+1}^{n-h-1}\frac{n-|l|}{n}\Bigg[\sum_{s_1,s_2=0}^{\infty}\operatorname{Cov}\Big(e_{j_1}^\top(B_{h,s_1}\underline{\varepsilon}_0-EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_0-EB_{0,s_1-h}\mu)^\top e_{j_2},$$
$$e_{j_3}^\top(B_{l+h,l+s_1}\underline{\varepsilon}_0-EB_{0,l+s_1}\mu)(B_{l,s_2}\underline{\varepsilon}_1-EB_{0,s_2}\mu)^\top e_{j_4}\Big)\Bigg]$$
$$\leq\frac{1}{n}\sum_{l=0}^{\infty}\sum_{s_1,s_2=0}^{\infty}\Bigg[\Big(E\Big[e_{j_1}^\top(B_{h,s_1}\underline{\varepsilon}_0-EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_0-EB_{0,s_1-h}\mu)^\top e_{j_2}\Big]^2\Big)^{1/2}$$
$$\times\Big(E\Big[e_{j_3}^\top(B_{0,l}\underline{\varepsilon}_0-EB_{0,l}\mu)\Big]^4\Big)^{1/4}\Big(E\Big[e_{j_4}^\top(B_{0,s_2}\underline{\varepsilon}_1-EB_{0,s_2}\mu)\Big]^4\Big)^{1/4}$$
$$+E\Big|e_{j_1}^\top(B_{h,s_1}\underline{\varepsilon}_0-EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_0-EB_{0,s_1-h}\mu)^\top e_{j_2}\Big|$$
$$\times\Big(E\Big[e_{j_3}^\top(B_{0,l}\underline{\varepsilon}_0-EB_{0,l}\mu)\Big]^2\Big)^{1/2}\Big(E\Big[e_{j_4}^\top(B_{0,s_2}\underline{\varepsilon}_1-EB_{0,s_2}\mu)\Big]^2\Big)^{1/2}\Bigg]=O(1/n),$$

since $\sum_{l=0}^{\infty}\Big(E\Big[e_i^\top(B_{0,l}\underline{\varepsilon}_0-EB_{0,l}\mu)\Big]^4\Big)^{1/4}<\infty$ for all $i=1,\ldots,d$. In the next step we consider the case that we have 2 pairs of indices and the 2 pairs are not equal. We have

$$
\begin{aligned}
&\frac{1}{n^2}\sum_{t_1,t_2=1}^{n-h}\Bigg[\sum_{s_1,s_2=0,s_1\neq s_2}^{\infty}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{t_1+h,s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,s_1}\mu)(B_{t_1,s_1-h}\underline{\varepsilon}_{t_1-s_1+h}-EB_{0,s_1-h}\mu)^{\top}e_{j_2},\\
&\qquad e_{j_3}^{\top}(B_{t_2+h,s_2}\underline{\varepsilon}_{t_2+h-s_2}-EB_{0,s_2}\mu)(B_{t_2,s_2-h}\underline{\varepsilon}_{t_2+h-s_2}-EB_{0,s_2}\mu)^{\top}e_{j_4}\Big)\\
&\quad+\sum_{s_1,s_2=0,s_1\neq s_2}^{\infty}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{t_1+h,s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,s_1}\mu)(B_{t_1,s_2}\underline{\varepsilon}_{t_1-s_2}-EB_{0,s_2}\mu)^{\top}e_{j_2},\\
&\qquad e_{j_3}^{\top}(B_{t_2+h,t_2-t_1+s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,t_2-t_1+s_1}\mu)(B_{t_2,t_2-t_1+s_2}\underline{\varepsilon}_{t_1-s_2}-EB_{0,t_2-t_1+s_2}\mu)^{\top}e_{j_4}\Big)\\
&\quad+\sum_{s_1,s_2=0,s_1\neq s_2}^{\infty}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{t_1+h,s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,s_1}\mu)(B_{t_1,s_2}\underline{\varepsilon}_{t_1-s_2}-EB_{0,s_2}\mu)^{\top}e_{j_4},\\
&\qquad e_{j_3}^{\top}(B_{t_2+h,t_2-t_1+h+s_2}\underline{\varepsilon}_{t_1-s_2}-EB_{0,t_2-t_1+h+s_2}\mu)(B_{t_2,t_2-t_1-h+s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,t_2-t_1-h+s_1}\mu)^{\top}e_{j_2}\Big)\Bigg]\\
=&\frac{1}{n}\sum_{l=-n+h+1}^{n-h-1}\frac{n-|l|}{n}\Bigg[\sum_{s_1,s_2=0,s_1\neq s_2}^{\infty}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{h,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1-EB_{0,s_1-h}\mu)^{\top}e_{j_2},\\
&\qquad e_{j_3}^{\top}(B_{l+h,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)(B_{l,s_2-h}\underline{\varepsilon}_2-EB_{0,s_2}\mu)^{\top}e_{j_4}\Big)\\
&\quad+\sum_{s_1,s_2=0,s_1\neq s_2}^{\infty}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{h,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)(B_{0,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)^{\top}e_{j_2},\\
&\qquad e_{j_3}^{\top}(B_{l+h,l+s_1}\underline{\varepsilon}_1-EB_{0,l+s_1}\mu)(B_{l,l+s_2}\underline{\varepsilon}_2-EB_{0,l+s_2}\mu)^{\top}e_{j_4}\Big)\\
&\quad+\sum_{s_1,s_2=0,s_1\neq s_2}^{\infty}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{h,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)(B_{0,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)^{\top}e_{j_2},\\
&\qquad e_{j_3}^{\top}(B_{l+h,l+h+s_2}\underline{\varepsilon}_2-EB_{0,l+h+s_2}\mu)(B_{l,l-h+s_1}\underline{\varepsilon}_1-EB_{0,l-h+s_1}\mu)^{\top}e_{j_4}\Big)\Bigg]\\
=&\mathcal{O}(1/n).
\end{aligned}
$$

To see this, we take a closer look at each part. The first part of the right-hand-side of the last equation can be bounded in the following way by using Corollary 10.16 in Bradley (2007) and the Cauchy-Schwarz inequality. Note that $B_{t,j}=f_j(Ad_t,\ldots,Ad_{t-j})$, hence, for some function $g,\tilde{g}$ we have

$$
e_{j_1}^{\top}(B_{h,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1-EB_{0,s_1-h}\mu)^{\top}e_{j_2}=g(\varepsilon_1,Ad_h,\ldots,Ad_{h-s_1}),\tag{3.7.1}
$$

and

$$
e_{j_3}^{\top}(B_{l+h,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)(B_{l,s_2-h}\underline{\varepsilon}_2-EB_{0,s_2}\mu)^{\top}e_{j_4}=\tilde{g}(\varepsilon_2,Ad_{l+h},\ldots,Ad_{l+h-s_2}.\tag{3.7.2}
$$

Thus, (3.7.2) is at least $l - s_2$ time points ahead (3.7.1) and we get

$$\begin{aligned}
&\frac{2}{n}\sum_{l=0}^{n-h-1}\frac{n-l}{n}\Bigg[\sum_{s_1=0}^{\infty}\sum_{s_2=0}^{l}\operatorname{Cov}\Big(e_{j_1}^\top(B_{h,s_1}\underline{\varepsilon}_1 - EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1 - EB_{0,s_1-h}\mu)^\top e_{j_2},\\
&\qquad e_{j_3}^\top(B_{l+h,s_2}\underline{\varepsilon}_2 - EB_{0,s_2}\mu)(B_{l,s_2-h}\underline{\varepsilon}_2 - EB_{0,s_2}\mu)^\top e_{j_4}\Big)\\
&\qquad + \sum_{s_1=0}^{\infty}\sum_{s_2=l+1}^{\infty}\operatorname{Cov}\Big(e_{j_1}^\top(B_{h,s_1}\underline{\varepsilon}_1 - EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1 - EB_{0,s_1-h}\mu)^\top e_{j_2},\\
&\qquad e_{j_3}^\top(B_{l+h,s_2}\underline{\varepsilon}_2 - EB_{0,s_2}\mu)(B_{l,s_2-h}\underline{\varepsilon}_2 - EB_{0,s_2}\mu)^\top e_{j_4}\Big)\Bigg]\\
&\le \frac{2}{n}\Bigg[\sum_{l=0}^{n-h-1}\frac{n-l}{n}\sum_{s_1=0}^{\infty}\sum_{s_2=0}^{l}\Big(E\Big[e_{j_1}^\top(B_{h,s_1}\underline{\varepsilon}_1 - EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1 - EB_{0,s_1-h}\mu)^\top e_{j_2}\Big]^4\Big)^{1/4}\\
&\qquad \times\Big(E\Big[e_{j_3}^\top(B_{h,s_2}\underline{\varepsilon}_2 - EB_{0,s_2}\mu)(B_{0,s_2-h}\underline{\varepsilon}_2 - EB_{0,s_2}\mu)^\top e_{j_4}\Big]^4\Big)^{1/4}\alpha(Ad, l-s_2)^{1/2}\\
&\qquad + \sum_{l=0}^{n-h-1}\frac{n-l}{n}\sum_{s_1=0}^{\infty}\sum_{s_2=l+1}^{\infty}\Big(E\Big[e_{j_1}^\top(B_{h,s_1}\underline{\varepsilon}_1 - EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1 - EB_{0,s_1-h}\mu)^\top e_{j_2}\Big]^2\Big)^{1/2}\\
&\qquad \times\Big(E\Big[e_{j_3}^\top(B_{h,s_2}\underline{\varepsilon}_2 - EB_{0,s_2}\mu)(B_{0,s_2-h}\underline{\varepsilon}_2 - EB_{0,s_2}\mu)^\top e_{j_4}\Big]^2\Big)^{1/2}\Bigg]\\
&\le \frac{2}{n}\Bigg[\sum_{s_1=0}^{\infty}\sum_{s_2=0}^{\infty}\sum_{l=0}^{\infty}\Big(E\Big[e_{j_1}^\top(B_{h,s_1}\underline{\varepsilon}_1 - EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1 - EB_{0,s_1-h}\mu)^\top e_{j_2}\Big]^4\Big)^{1/4}\\
&\qquad \times\Big(E\Big[e_{j_3}^\top(B_{h,s_2}\underline{\varepsilon}_2 - EB_{0,s_2}\mu)(B_{0,s_2-h}\underline{\varepsilon}_2 - EB_{0,s_2}\mu)^\top e_{j_4}\Big]^4\Big)^{1/4}\alpha(Ad, l)^{1/2}\\
&\qquad + \sum_{s_1=0}^{\infty}\sum_{s_2=0}^{\infty}s_2\Big(E\Big[e_{j_1}^\top(B_{h,s_1}\underline{\varepsilon}_1 - EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1 - EB_{0,s_1-h}\mu)^\top e_{j_2}\Big]^2\Big)^{1/2}\\
&\qquad \times\Big(E\Big[e_{j_3}^\top(B_{h,s_2}\underline{\varepsilon}_2 - EB_{0,s_2}\mu)(B_{0,s_2-h}\underline{\varepsilon}_2 - EB_{0,s_2}\mu)^\top e_{j_4}\Big]^2\Big)^{1/2}\Bigg] = \mathcal{O}(1/n).
\end{aligned}$$

Due to Assumption 1 and 4. The second part can be bounded by applying again the Cauchy-Schwarz-inequality. Hence, we have

$$
\begin{aligned}
&\frac{1}{n}\sum_{l=-n+h+1}^{n-h-1}\frac{n-|l|}{n}\sum_{s_1,s_2=0}^{\infty}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{h,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)(B_{0,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)^{\top}e_{j_2},\\
&\qquad e_{j_3}^{\top}(B_{l+h,l+s_1}\underline{\varepsilon}_1-EB_{0,l+s_1}\mu)(B_{l,l+s_2}\underline{\varepsilon}_2-EB_{0,l+s_2}\mu)^{\top}e_{j_4}\Big)\\
&\le\frac{1}{n}\sum_{l=-n+h+1}^{n-h-1}\frac{n-|l|}{n}\sum_{s_1,s_2=0}^{\infty}\Big(E\Big[e_{j_1}^{\top}(B_{0,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)\Big]^4\Big)^{1/4}\Big(E\Big[(B_{0,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)^{\top}e_{j_2}\Big]^4\Big)^{1/4}\\
&\qquad\times\Big(E\Big[e_{j_3}^{\top}(B_{0,l+s_1}\underline{\varepsilon}_1-EB_{0,l+s_1}\mu)\Big]^4\Big)^{1/4}\Big(E\Big[(B_{0,l+s_2}\underline{\varepsilon}_2-EB_{0,l+s_2}\mu)^{\top}e_{j_4}\Big]^4\Big)^{1/4}\\
&\le\frac{C}{n}\sum_{s_1,s_2,l=0}^{\infty}\Big(E\Big[e_{j_1}^{\top}(B_{0,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)\Big]^4E\Big[(B_{0,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)^{\top}e_{j_2}\Big]^4E\Big[e_{j_3}^{\top}(B_{0,l}\underline{\varepsilon}_1-EB_{0,l}\mu)\Big]^4\Big)^{1/4}\\
&=\mathcal{O}(1/n).
\end{aligned}
$$

Similar arguments can be applied to the third part. In the next step, we consider the case that 2 indices are equal and the other indices are different from each other. We have

$$
\begin{aligned}
&\frac{1}{n^2}\sum_{t_1,t_2=1}^{n-h}\Bigg[\sum_{s_1,s_2,s_3=0,s_1\neq s_2\neq s_3}^{\infty}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{t_1+h,s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,s_1}\mu)(B_{t_1,s_1-h}\underline{\varepsilon}_{t_1-s_1+h}-EB_{0,s_1-h}\mu)^{\top}e_{j_2},\\
&\qquad e_{j_3}^{\top}(B_{t_2+h,s_2}\underline{\varepsilon}_{t_2+h-s_2}-EB_{0,s_2}\mu)(B_{t_2,s_3}\underline{\varepsilon}_{t_2-s_3}-EB_{0,s_3}\mu)^{\top}e_{j_4}\Big)\\
&\quad+\sum_{s_1,s_2,s_3=0,s_1\neq s_2\neq s_3}^{\infty}\operatorname{Cov}\Big(e_{j_3}^{\top}(B_{t_1+h,s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,s_1}\mu)(B_{t_1,s_1-h}\underline{\varepsilon}_{t_1-s_1+h}-EB_{0,s_1-h}\mu)^{\top}e_{j_4},\\
&\qquad e_{j_1}^{\top}(B_{t_2+h,s_2}\underline{\varepsilon}_{t_2+h-s_2}-EB_{0,s_2}\mu)(B_{t_2,s_3}\underline{\varepsilon}_{t_2-s_3}-EB_{0,s_3}\mu)^{\top}e_{j_2}\Big)\\
&\quad+\sum_{s_1,s_2,s_3=0,s_1\neq s_2\neq s_3}^{\infty}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{t_1+h,s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,s_1}\mu)(B_{t_1,s_2}\underline{\varepsilon}_{t_1-s_2}-EB_{0,s_2}\mu)^{\top}e_{j_2},\\
&\qquad e_{j_3}^{\top}(B_{t_2+h,t_2-t_1+s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,t_2-t_1+s_1}\mu)(B_{t_2,s_3}\underline{\varepsilon}_{t_2-s_3}-EB_{0,s_3}\mu)^{\top}e_{j_4}\Big)\\
&\quad+\sum_{s_1,s_2,s_3=0,s_1\neq s_2\neq s_3}^{\infty}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{t_1+h,s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,s_1}\mu)(B_{t_1,s_2}\underline{\varepsilon}_{t_1-s_2}-EB_{0,s_2}\mu)^{\top}e_{j_2},\\
&\qquad e_{j_3}^{\top}(B_{t_2+h,s_3}\underline{\varepsilon}_{t_2+h-s_3}-EB_{0,s_3}\mu)(B_{t_2,t_2-t_1+s_2}\underline{\varepsilon}_{t_1-s_2}-EB_{0,t_2-t_1+s_2}\mu)^{\top}e_{j_4}\Big)\\
&\quad+\sum_{s_1,s_2,s_3=0,s_1\neq s_2\neq s_3}^{\infty}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{t_1+h,s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,s_1}\mu)(B_{t_1,s_2}\underline{\varepsilon}_{t_1-s_2}-EB_{0,s_2}\mu)^{\top}e_{j_2},\\
&\qquad e_{j_3}^{\top}(B_{t_2+h,t_2-t_1+h+s_2}\underline{\varepsilon}_{t_1-s_2}-EB_{0,t_2-t_1+h+s_2}\mu)(B_{t_2,s_3}\underline{\varepsilon}_{t_2-s_3}-EB_{0,s_3}\mu)^{\top}e_{j_4}\Big)\\
&\quad+\sum_{s_1,s_2,s_3=0,s_1\neq s_2\neq s_3}^{\infty}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{t_1+h,s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,s_1}\mu)(B_{t_1,s_2}\underline{\varepsilon}_{t_1-s_2}-EB_{0,s_2}\mu)^{\top}e_{j_2},\\
&\qquad e_{j_3}^{\top}(B_{t_2+h,s_3}\underline{\varepsilon}_{t_2+h-s_3}-EB_{0,s_3}\mu)(B_{t_2,t_2-t_1-h+s_1}\underline{\varepsilon}_{t_1+h-s_1}-EB_{0,t_2-t_1-h+s_1}\mu)^{\top}e_{j_4}\Big)\Bigg],
\end{aligned}
$$

which is equal to

$$\begin{aligned}
&\frac{1}{n}\sum_{l=-n+h-1}^{n-h-1}\frac{n-|l|}{n}\Bigg[\sum_{s_1,s_2,s_3=0,s_1\neq s_2\neq s_3}^{\infty}\operatorname{Cov}\Big(e_{j_1}^\top(B_{h,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1-EB_{0,s_1-h}\mu)^\top e_{j_2},\\
&\qquad e_{j_3}^\top(B_{l+h,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)(B_{l,s_3}\underline{\varepsilon}_3-EB_{0,s_3}\mu)^\top e_{j_4}\Big)\\
&\quad+\sum_{s_1,s_2,s_3=0,s_1\neq s_2\neq s_3}^{\infty}\operatorname{Cov}\Big(e_{j_3}^\top(B_{h,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1-EB_{0,s_1-h}\mu)^\top e_{j_4},\\
&\qquad e_{j_1}^\top(B_{l+h,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)(B_{l,s_3}\underline{\varepsilon}_3-EB_{0,s_3}\mu)^\top e_{j_2}\Big)\\
&\quad+\sum_{s_1,s_2,s_3=0,s_1\neq s_2\neq s_3}^{\infty}\operatorname{Cov}\Big(e_{j_1}^\top(B_{h,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)(B_{0,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)^\top e_{j_2},\\
&\qquad e_{j_3}^\top(B_{l+h,l+s_1}\underline{\varepsilon}_1-EB_{0,l+s_1}\mu)(B_{l,s_3}\underline{\varepsilon}_3-EB_{0,s_3}\mu)^\top e_{j_4}\Big)\\
&\quad+\sum_{s_1,s_2,s_3=0,s_1\neq s_2\neq s_3}^{\infty}\operatorname{Cov}\Big(e_{j_1}^\top(B_{h,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)(B_{0,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)^\top e_{j_2},\\
&\qquad e_{j_3}^\top(B_{l+h,s_3}\underline{\varepsilon}_3-EB_{0,s_3}\mu)(B_{l,l+s_2}\underline{\varepsilon}_2-EB_{0,l+s_2}\mu)^\top e_{j_4}\Big)\\
&\quad+\sum_{s_1,s_2,s_3=0,s_1\neq s_2\neq s_3}^{\infty}\operatorname{Cov}\Big(e_{j_1}^\top(B_{h,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)(B_{0,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)^\top e_{j_2},\\
&\qquad e_{j_3}^\top(B_{l+h,l+h+s_2}\underline{\varepsilon}_2-EB_{0,l+h+s_2}\mu)(B_{l,s_3}\underline{\varepsilon}_3-EB_{0,s_3}\mu)^\top e_{j_4}\Big)\\
&\quad+\sum_{s_1,s_2,s_3=0,s_1\neq s_2\neq s_3}^{\infty}\operatorname{Cov}\Big(e_{j_1}^\top(B_{h,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)(B_{0,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)^\top e_{j_2},\\
&\qquad e_{j_3}^\top(B_{l+h,s_3}\underline{\varepsilon}_3-EB_{0,s_3}\mu)(B_{l,l-h+s_1}\underline{\varepsilon}_1-EB_{0,l-h+s_1}\mu)^\top e_{j_4}\Big)\Bigg]\\
&=\mathcal{O}(1/n).
\end{aligned}$$

To see this result notice the first and the second part as well as the third to the sixth part of the last term can be bounded by using similar arguments. The first part can be bounded by using Corollary 10.16 in Bradley (2007) and the Cauchy-Schwarz inequality. We have

$$
\begin{aligned}
&\frac{1}{n}\sum_{l=-n+h-1}^{n-h-1}\frac{n-|l|}{n}\Bigg[\sum_{s_1,s_2,s_3=0}^{\infty}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{h,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1-EB_{0,s_1-h}\mu)^{\top}e_{j_2},\\
&\quad e_{j_3}^{\top}(B_{l+h,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)(B_{l,s_3}\underline{\varepsilon}_3-EB_{0,s_3}\mu)^{\top}e_{j_4}\Big)\Bigg]\\
\le&\frac{2}{n}\sum_{l=0}^{n-h-1}\frac{n-|l|}{n}\Bigg[\sum_{s_1=0}^{\infty}\sum_{s_2,s_3=0}^{l}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{h,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1-EB_{0,s_1-h}\mu)^{\top}e_{j_2},\\
&\quad e_{j_3}^{\top}(B_{l+h,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)(B_{l,s_3}\underline{\varepsilon}_3-EB_{0,s_3}\mu)^{\top}e_{j_4}\Big)\\
&\quad+\sum_{s_1=0}^{\infty}\sum_{s_2=0}^{\infty}\sum_{s_3=l+1}^{\infty}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{h,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1-EB_{0,s_1-h}\mu)^{\top}e_{j_2},\\
&\quad e_{j_3}^{\top}(B_{l+h,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)(B_{l,s_3}\underline{\varepsilon}_3-EB_{0,s_3}\mu)^{\top}e_{j_4}\Big)\\
&\quad+\sum_{s_1=0}^{\infty}\sum_{s_2=l+1}^{\infty}\sum_{s_3=0}^{l}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{h,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1-EB_{0,s_1-h}\mu)^{\top}e_{j_2},\\
&\quad e_{j_3}^{\top}(B_{l+h,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)(B_{l,s_3}\underline{\varepsilon}_3-EB_{0,s_3}\mu)^{\top}e_{j_4}\Big)\Bigg]\\
\le&\frac{2}{n}\Bigg[\sum_{s_1=0}^{\infty}\sum_{l=0}^{n-h-1}\sum_{s_2,s_3=0}^{l}\Big(E\Big[e_{j_1}^{\top}(B_{h,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1-EB_{0,s_1-h}\mu)^{\top}e_{j_2}\Big]^4\Big)^{1/4}\\
&\quad\times\Big(E\Big[e_{j_3}^{\top}(B_{0,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)\Big]^4\Big)^{1/4}\Big(E\Big[e_{j_4}^{\top}(B_{0,s_3}\underline{\varepsilon}_3-EB_{0,s_3}\mu)\Big]^4\Big)^{1/4}\alpha(Ad,l-\max(s_2,s_3))^{1/4}\\
&\quad+\sum_{l=0}^{n-h-1}\sum_{s_1=0}^{\infty}\sum_{s_2=0}^{\infty}\sum_{s_3=l+1}^{\infty}\Big(E\Big[e_{j_1}^{\top}(B_{h,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1-EB_{0,s_1-h}\mu)^{\top}e_{j_2}\Big]^2\Big)^{1/2}\\
&\quad\times\Big(E\Big[e_{j_3}^{\top}(B_{0,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)\Big]^4\Big)^{1/4}\Big(E\Big[e_{j_4}^{\top}(B_{0,s_3}\underline{\varepsilon}_3-EB_{0,s_3}\mu)\Big]^4\Big)^{1/4}\\
&\quad+\sum_{l=0}^{n-h-1}\sum_{s_1=0}^{\infty}\sum_{s_2=l+1}^{\infty}\sum_{s_3=0}^{l}\Big(E\Big[e_{j_1}^{\top}(B_{h,s_1}\underline{\varepsilon}_1-EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1-EB_{0,s_1-h}\mu)^{\top}e_{j_2}\Big]^2\Big)^{1/2}\\
&\quad\times\Big(E\Big[e_{j_3}^{\top}(B_{0,s_2}\underline{\varepsilon}_2-EB_{0,s_2}\mu)\Big]^4\Big)^{1/4}\Big(E\Big[e_{j_4}^{\top}(B_{0,s_3}\underline{\varepsilon}_3-EB_{0,s_3}\mu)\Big]^4\Big)^{1/4}\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
\leq \frac{2}{n}\Bigg[& \sum_{s_1=0}^{\infty}\sum_{s_2,s_3=0}^{\infty}\sum_{l=0}^{\infty}\left(E\left[e_{j_1}^{\top}(B_{h,s_1}\underline{\varepsilon}_1 - EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1 - EB_{0,s_1-h}\mu)^{\top}e_{j_2}\right]^4\right)^{1/4} \\
& \times\left(E\left[e_{j_3}^{\top}(B_{0,s_2}\underline{\varepsilon}_2 - EB_{0,s_2}\mu)\right]^4\right)^{1/4}\left(E\left[e_{j_4}^{\top}(B_{0,s_3}\underline{\varepsilon}_3 - EB_{0,s_3}\mu)\right]^4\right)^{1/4}\alpha(Ad,l)^{1/4} \\
& + \sum_{s_1=0}^{\infty}\sum_{s_2=0}^{\infty}\sum_{s_3=0}^{\infty} s_3\left(E\left[e_{j_1}^{\top}(B_{h,s_1}\underline{\varepsilon}_1 - EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1 - EB_{0,s_1-h}\mu)^{\top}e_{j_2}\right]^2\right)^{1/2} \\
& \times\left(E\left[e_{j_3}^{\top}(B_{0,s_2}\underline{\varepsilon}_2 - EB_{0,s_2}\mu)\right]^4\right)^{1/4}\left(E\left[e_{j_4}^{\top}(B_{0,s_3}\underline{\varepsilon}_3 - EB_{0,s_3}\mu)\right]^4\right)^{1/4} \\
& + \sum_{s_1=0}^{\infty}\sum_{s_2=0}^{\infty} s_2\sum_{s_3=0}^{\infty}\left(E\left[e_{j_1}^{\top}(B_{h,s_1}\underline{\varepsilon}_1 - EB_{0,s_1}\mu)(B_{0,s_1-h}\underline{\varepsilon}_1 - EB_{0,s_1-h}\mu)^{\top}e_{j_2}\right]^2\right)^{1/2} \\
& \times\left(E\left[e_{j_3}^{\top}(B_{0,s_2}\underline{\varepsilon}_2 - EB_{0,s_2}\mu)\right]^4\right)^{1/4}\left(E\left[e_{j_4}^{\top}(B_{0,s_3}\underline{\varepsilon}_3 - EB_{0,s_3}\mu)\right]^4\right)^{1/4}\Bigg] \\
= {} & \mathcal{O}(1/n),
\end{aligned}
$$

since $\sum_{s_2=0}^{\infty} s_2\left(E\left[e_{j_i}^{\top}(B_{0,s_2}\underline{\varepsilon}_2 - EB_{0,s_2}\mu)\right]^4\right)^{1/4} < \infty$ for all $i = 1,\ldots,d$. The third term can be bounded by using the Cauchy-Schwarz inequality. Hence, we have

$$
\begin{aligned}
& \frac{1}{n}\sum_{l=-n+h-1}^{n-h-1}\frac{n-|l|}{n}\sum_{s_1,s_2,s_3=0}^{\infty}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{h,s_1}\underline{\varepsilon}_1 - EB_{0,s_1}\mu)(B_{0,s_2}\underline{\varepsilon}_2 - EB_{0,s_2}\mu)^{\top}e_{j_2}, \\
& \qquad e_{j_3}^{\top}(B_{l+h,l+s_1}\underline{\varepsilon}_1 - EB_{0,l+s_1}\mu)(B_{l,s_3}\underline{\varepsilon}_3 - EB_{0,s_3}\mu)^{\top}e_{j_4}\Big) \\
& \leq \frac{1}{n}\sum_{l=0}^{\infty}\sum_{s_1,s_2,s_3=0}^{\infty}\left(E\left[(e_{j_1}^{\top}(B_{h,s_1}\underline{\varepsilon}_1 - EB_{0,s_1}\mu)(B_{0,s_2}\underline{\varepsilon}_2 - EB_{0,s_2}\mu)^{\top}e_{j_2}\right]^2\right)^{1/2} \\
& \qquad \left(E\left[e_{j_3}^{\top}(B_{0,l}\underline{\varepsilon}_1 - EB_{0,l}\mu)\right]^4\right)^{1/4}\left(E\left[e_{j_4}^{\top}(B_{0,s_3}\underline{\varepsilon}_3 - EB_{0,s_3}\mu)\right]^4\right)^{1/4} \\
& = \mathcal{O}(1/n).
\end{aligned}
$$

It remains to consider the last case in which all indices are different from each other. We apply Corollary 10.16 in Bradley (2007) and the Cauchy-Schwarz inequality, and similarly to (3.7.1),(3.7.2).

We obtain

$$
\begin{aligned}
&\frac{1}{n^2}\sum_{t_1,t_2=1}^{n-h}\Bigg[\sum_{s_1,s_2,s_3,s_4=0,s_1\neq s_2\neq s_3\neq s_4}^{\infty}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{t_1+h,s_1}\varepsilon_{t_1+h-s_1}-EB_{0,s_1}\mu)(B_{t_1,s_2}\varepsilon_{t_1-s_2}-EB_{0,s_2}\mu)^{\top}e_{j_2},\\
&\qquad e_{j_3}^{\top}(B_{t_2+h,s_3}\varepsilon_{t_2+h-s_3}-EB_{0,s_3}\mu)(B_{t_2,s_4}\varepsilon_{t_2-s_4}-EB_{0,s_4}\mu)^{\top}e_{j_4}\Big)\Bigg]\\
\le&\frac{2}{n}\sum_{l=0}^{n-h-1}\frac{n-l}{n}\Bigg[\sum_{s_1,s_2,s_3,s_4=0}^{\infty}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{h,s_1}\mu-EB_{0,s_1}\mu)(B_{0,s_2}\mu-EB_{0,s_2}\mu)^{\top}e_{j_2},\\
&\qquad e_{j_3}^{\top}(B_{l+h,s_3}\mu-EB_{0,s_3}\mu)(B_{l,s_4}\mu-EB_{0,s_4}\mu)^{\top}e_{j_4}\Big)\Bigg]\\
=&\frac{2}{n}\sum_{l=0}^{n-h-1}\frac{n-l}{n}\Bigg[\sum_{s_1,s_2=0}^{\infty}\sum_{s_3,s_4=0}^{l}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{h,s_1}\mu-EB_{0,s_1}\mu)(B_{0,s_2}\mu-EB_{0,s_2}\mu)^{\top}e_{j_2},\\
&\qquad e_{j_3}^{\top}(B_{l+h,s_3}\mu-EB_{0,s_3}\mu)(B_{l,s_4}\mu-EB_{0,s_4}\mu)^{\top}e_{j_4}\Big)\\
&+\sum_{s_1,s_2=0}^{\infty}\sum_{s_3=0}^{\infty}\sum_{s_4=l+1}^{\infty}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{h,s_1}\mu-EB_{0,s_1}\mu)(B_{0,s_2}\mu-EB_{0,s_2}\mu)^{\top}e_{j_2},\\
&\qquad e_{j_3}^{\top}(B_{l+h,s_3}\mu-EB_{0,s_3}\mu)(B_{l,s_4}\mu-EB_{0,s_4}\mu)^{\top}e_{j_4}\Big)\\
&+\sum_{s_1,s_2=0}^{\infty}\sum_{s_3=l+1}^{\infty}\sum_{s_4=0}^{l}\operatorname{Cov}\Big(e_{j_1}^{\top}(B_{h,s_1}\mu-EB_{0,s_1}\mu)(B_{0,s_2}\mu-EB_{0,s_2}\mu)^{\top}e_{j_2},\\
&\qquad e_{j_3}^{\top}(B_{l+h,s_3}\mu-EB_{0,s_3}\mu)(B_{l,s_4}\mu-EB_{0,s_4}\mu)^{\top}e_{j_4}\Big)\Bigg]\\
\le&\frac{2}{n}\Bigg[\sum_{s_1,s_2=0}^{\infty}\sum_{s_3,s_4=0}^{\infty}\sum_{l=0}^{\infty}\Big(E\Big[e_{j_1}^{\top}(B_{h,s_1}\mu-EB_{0,s_1}\mu)\Big]^5\Big)^{1/5}\Big(E\Big[e_{j_2}^{\top}(B_{0,s_2}\mu-EB_{0,s_2}\mu)\Big]^5\Big)^{1/5}\\
&\qquad\Big(E\Big[e_{j_3}^{\top}(B_{0,s_3}\mu-EB_{0,s_3}\mu)\Big]^5\Big)^{1/5}\Big(E\Big[e_{j_4}^{\top}(B_{0,s_4}\mu-EB_{0,s_4}\mu)\Big]^5\Big)^{1/5}\alpha(Ad,l)^{1/5}\\
&+\sum_{s_1,s_2=0}^{\infty}\sum_{s_3=0}^{\infty}\sum_{s_4=0}^{\infty}s_4\Big(E\Big[e_{j_1}^{\top}(B_{h,s_1}\mu-EB_{0,s_1}\mu)\Big]^4\Big)^{1/4}\Big(E\Big[e_{j_2}^{\top}(B_{0,s_2}\mu-EB_{0,s_2}\mu)\Big]^4\Big)^{1/4}\\
&\qquad\Big(E\Big[e_{j_3}^{\top}(B_{0,s_3}\mu-EB_{0,s_3}\mu)\Big]^4\Big)^{1/4}\Big(E\Big[e_{j_4}^{\top}(B_{0,s_4}\mu-EB_{0,s_4}\mu)\Big]^4\Big)^{1/4}\\
&+\sum_{s_1,s_2=0}^{\infty}\sum_{s_3=0}^{\infty}s_3\sum_{s_4=0}^{\infty}\Big(E\Big[e_{j_1}^{\top}(B_{h,s_1}\mu-EB_{0,s_1}\mu)\Big]^4\Big)^{1/4}\Big(E\Big[e_{j_2}^{\top}(B_{0,s_2}\mu-EB_{0,s_2}\mu)\Big]^4\Big)^{1/4}\\
&\qquad\Big(E\Big[e_{j_3}^{\top}(B_{0,s_3}\mu-EB_{0,s_3}\mu)\Big]^4\Big)^{1/4}\Big(E\Big[e_{j_4}^{\top}(B_{0,s_4}\mu-EB_{0,s_4}\mu)\Big]^4\Big)^{1/4}\Bigg]\\
&=\mathcal{O}(1/n),
\end{aligned}
$$

since $\sum_{s_1=0}^{\infty}\Big(E\Big[e_{j_i}^{\top}(B_{h,s_1}\mu-EB_{0,s_1}\mu)\Big]^5\Big)^{1/5}<\infty$ for all $i=1,\ldots,d$. □

Proof of Theorem 3.3.3. Let $P(Ad_{1;j\cdot} = \tilde{A}d_{k;j\cdot}, Ad_{1;\cdot j} = \tilde{A}d_{k;\cdot j}) =: P_k^j$. In order to simplify the notation, in this proof we write $\underline{X}_t := (\underline{X}_{t;s})_{s\in S}$ and $\underline{\varepsilon}_t := (\underline{\varepsilon}_{t;s})_{s\in S}$ which is as (without loss of generality) $S = \{1,\dots,d\}$. Furthermore, define the random variable $\phi_k^j(t) = \{\omega \in \Omega : Ad_{t;j\cdot}(\omega) = \tilde{A}d_{k;j\cdot}, Ad_{t;\cdot j}(\omega) = \tilde{A}d_{t;\cdot j}\}$ which is an indicator that Ad_t coincides in the j-th row and column with the considered state $\tilde{A}d_k$. For $r \in R_j^k$ we have $\underline{X}_{r+1;j} = a_{jk}\underline{X}_r + \underline{\varepsilon}_{r;j}$. We have, as $n \to \infty$, $|R_k^j|/n = 1/n\sum_{r=0}^{n-1}\{\phi_k^j(r)\} \xrightarrow{P} P_k^j$ since $E|R_k^j|/n = E\{\phi_k^j(1)\} = P_k^j$ and since **Ad** is α-mixing, we have

$$\mathrm{Var}(|R_k^j|/n) = \sum_{h=-n+1}^{n-1}\frac{n-|h|}{n^2}\mathrm{Cov}(\phi_k^j(0),\phi_k^j(h)) \le C\sum_{h=-n+1}^{n-1}\frac{n-|h|}{n^2}\alpha(Ad,|h|) = \mathcal{O}(1/n).$$

Furthermore, we have $E(nP_k^j)^{-1}\sum_{r\in R_j^k}\underline{X}_r = (P_k^j)^{-1}E\underline{X}_1\phi_k^j(1) = E\left[\underline{X}_1|\phi_k^j(1)\right]$ and for $i_1,i_2 = 1,\dots,d$ we have with Assumption i) to iv)

$$\begin{aligned}
&e_{i_1}^\top \mathrm{Var}\left((nP_k^j)^{-1}\sum_{r\in R_j^k}\underline{X}_r\right)e_{i_2} = (P_k^j)^{-2}\sum_{h=-n+1}^{n-1}\frac{n-|h|}{n^2}\mathrm{Cov}(e_{i_1}^\top\underline{X}_0\phi_k^j(0), e_{i_2}^\top\underline{X}_h\phi_k^j(h))\\
&= (P_k^j)^{-2}\sum_{h=-n+1}^{n-1}\frac{n-|h|}{n^2}\sum_{s_1,s_2=0}^{\infty}\mathrm{Cov}(e_{i_1}^\top B_{0,s_1}\underline{\varepsilon}_{-s_1}\phi_k^j(0), e_{i_2}^\top B_{h,s_2}\underline{\varepsilon}_{h-s_1}\phi_k^j(h))\\
&\le (P_k^j)^{-2}2/n\sum_{h=0}^{\infty}\Bigg[\sum_{s=0}^{\infty}E(e_{i_1}^\top B_{0,s}\Sigma B_{h,h+s}^\top e_{i_2}\phi_k^j(0)\phi_k^j(h))\\
&+\sum_{s_1=0}^{\infty}\Bigg(\sum_{s_2=0}^{|h|}\mathrm{Cov}(e_{i_1}^\top B_{0,s_1}\underline{\mu}\phi_k^j(0), e_{i_2}^\top B_{h,s_2}\underline{\mu}\phi_k^j(h)) + \sum_{s_2=|h|+1}^{\infty}\mathrm{Cov}(e_{i_1}^\top B_{0,s_1}\underline{\mu}\phi_k^j(0), e_{i_2}^\top B_{h,s_2}\underline{\mu}\phi_k^j(h))\Bigg)\Bigg]\\
&\le (P_k^j)^{-2}2/n\Bigg[\sum_{h=0}^{\infty}\sum_{s=0}^{\infty}\|\Sigma\|_\infty\left((E(e_{i_1}^\top B_{0,s}\mathbb{1})^2E(e_{i_2}B_{0,h}\mathbb{1})^2\right)^{1/2}\\
&+\sum_{h=0}^{\infty}\sum_{s_1,s_2=0}^{\infty}\left(E(e_{i_1}^\top B_{0,s_1}\underline{\mu})^4E(e_{i_2}^\top B_{0,s_2}\underline{\mu})^4\right)^{1/4}\alpha(Ad,h)^{1/2} + \sum_{s_1,s_2=0}^{\infty}s_2\left((E(e_{i_1}^\top B_{0,s}\underline{\mu})^2E(e_{i_2}B_{0,h}\underline{\mu})^2\right)^{1/2}\Bigg]\\
&= \mathcal{O}\left(\frac{1}{n}(P_k^j)^{-2}\right).
\end{aligned}$$

Hence, $(nP_k^j)^{-1}\sum_{r\in\mathbb{R}}\underline{X}_r \xrightarrow{P} E\left[\underline{X}_1|\phi_k^j(1)\right]$, as $n\to\infty$. Similarly, we have, as $n\to\infty$,

$$(nP_k^j)^{-1}\sum_{r\in\mathbb{R}}\underline{X}_{r+1;j} \xrightarrow{P} E\left[\underline{X}_{2;j}|\phi_k^j(1)\right] = E\left[\alpha_{jk}\underline{X}_1 + \underline{\mu}|\phi_k^j(1)\right].$$

Notice that $E(\phi_k^j(1)) = P_k^j$ is independent from n and could be dropped in the $\mathcal{O}$-notation. However, this probability could be very small and to keep that in mind, we keep this constant. This gives us

$$\left((nP_k^j)^{-1}\sum_{r\in R_k^j}(\underline{X}_r-|R_k^j|^{-1}\sum_{v\in R_k^j}\underline{X}_v)(\underline{X}_r-|R_k^j|^{-1}\sum_{v\in R_k^j}\underline{X}_v)^\top\right)$$
$$=\left((nP_k^j)^{-1}\sum_{r\in R_k^j}\left(\underline{X}_r-E\left[\underline{X}_1|\phi_k^j(1)\right]\right)\left(\underline{X}_r-E\left[\underline{X}_1|\phi_k^j(1)\right]\right)^\top\right)+\mathcal{O}(n^{-1/2}(P_k^j)^{-1})$$

This matrix is very similar to $\hat{\Gamma}(0)$ and the same arguments can be applied. For the mean of the matrix we have

$$E\left((nP_k^j)^{-1}\sum_{r\in R_k^j}\left(\underline{X}_r-E\left[\underline{X}_1|\phi_k^j(1)\right]\right)\left(\underline{X}_r-E\left[\underline{X}_1|\phi_k^j(1)\right]\right)^\top\right)$$
$$=E\left((nP_k^j)^{-1}\sum_{r=0}^{n-1}\left(\underline{X}_r-E\left[\underline{X}_1|\phi_k^j(1)\right]\right)\left(\underline{X}_r-E\left[\underline{X}_1|\phi_k^j(1)\right]\right)^\top\phi_k^j(r)\right)=\mathrm{Var}\left(\underline{X}_1|\phi_k^j(1)\right),$$

which is positive definite since $P_k^j>0$ and Σ is positive definite. The variance can be bounded by using the same arguments used to bound the variance of $(nP_k^j)^{-1}\sum_{r\in R_j^k}\underline{X}_r$ and the variance of the sample autocovariance, see proof of Theorem 3.3.2. We get that the variance is of order $\mathcal{O}((n(P_k^j)^2)^{-1})$. Consequently, we have

$$\left((nP_k^j)^{-1}\sum_{r\in R_k^j}(\underline{X}_r-|R_k^j|^{-1}\sum_{v\in R_k^j}\underline{X}_v)(\underline{X}_r-|R_k^j|^{-1}\sum_{v\in R_k^j}\underline{X}_v)^\top\right)=\mathrm{Var}\left(\underline{X}_1|\phi_k^j(1)\right)+\mathcal{O}_P(1/\sqrt{n}(P_k^j)^{-1}).$$

Thus, if n is large enough, we have a matrix that is invertible with high probability and we can consider the case that this matrix is invertible. Due to Assumption v) we have $\underline{X}_{r+1;j}=a_{jk}\underline{X}_r+\underline{\varepsilon}_{r;j}, r\in R_k^j$. Hence,

$$
\begin{aligned}
&\left((nP_k^j)^{-1}\sum_{r\in R_k^j}(\underline{X}_r-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{X}_v)(\underline{X}_r-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{X}_v)^\top\right)\tilde{\alpha}_{jk}\\
&=\left((nP_k^j)^{-1}\sum_{r\in R_k^j}(\underline{X}_{r+1;j}-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{X}_{v+1;j})(\underline{X}_r-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{X}_v)\right)\\
&=\left((nP_k^j)^{-1}\sum_{r\in R_k^j}\left(\alpha_{jk}^\top\underline{X}_r+\underline{\varepsilon}_{r+1;j}-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}(\alpha_{jk}^\top\underline{X}_v+\underline{\varepsilon}_{v+1;j})\right)(\underline{X}_r-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{X}_v)\right)\\
&=\left((nP_k^j)^{-1}\sum_{r\in R_k^j}(\underline{X}_r-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{X}_v)(\underline{X}_r-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{X}_v)^\top\alpha_{jk}+\right.\\
&\qquad\left.(nP_k^j)^{-1}\sum_{r\in R_k^j}(\underline{X}_r-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{X}_v)(\underline{\varepsilon}_{r+1;j}-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{\varepsilon}_{v+1;j})\right)\\
&\Rightarrow\tilde{\alpha}_{jk}=\alpha_{jk}+\left((nP_k^j)^{-1}\sum_{r\in R_k^j}(\underline{X}_r-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{X}_v)(\underline{X}_r-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{X}_v)^\top\right)^{-1}\\
&\qquad\left((nP_k^j)^{-1}\sum_{r\in R_k^j}(\underline{X}_r-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{X}_v)(\underline{\varepsilon}_{r+1;j}-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{\varepsilon}_{v+1;j})\right).
\end{aligned}
$$

In the next step we show that, as $n\to\infty$,

$$
\left((nP_k^j)^{-1}\sum_{r\in R_k^j}(\underline{X}_r-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{X}_v)(\underline{\varepsilon}_{r+1;j}-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{\varepsilon}_{v+1;j})\right)\xrightarrow{P}\underline{0}.
$$

Since the innovation process $\underline{\varepsilon}$ is i.i.d. and independent from **Ad**, we have $\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{\varepsilon}_{r+1;j}\xrightarrow{P}\underline{\mu}_j$, as $n\to\infty$. Furthermore, since $\underline{X}_t=\sum_{l=0}^{\infty}B_{0,l}\underline{\varepsilon}_{t-l}$, we have

$$
\begin{aligned}
&E\left((nP_k^j)^{-1}\sum_{r\in R_k^j}(\underline{X}_r-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{X}_v)(\underline{\varepsilon}_{r+1;j}-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{\varepsilon}_{v+1;j})\right)\\
=&E\left((nP_k^j)^{-1}\left(\sum_{r\in R_k^j}\sum_{l=0}^{\infty}B_{r,l}\underline{\varepsilon}_{r-l}\underline{\varepsilon}_{r+1;j}-\frac{1}{|R_k^j|}\sum_{r\in R_k^j}\sum_{v\in R_k^j}\sum_{l=0}^{\infty}B_{v,l}\underline{\varepsilon}_{v-l}\underline{\varepsilon}_{r+1;j}\right.\right.\\
&\left.\left.-\frac{1}{|R_k^j|}\sum_{r\in R_k^j}\sum_{v\in R_k^j}\sum_{l=0}^{\infty}B_{r,l}\underline{\varepsilon}_{r-l}\underline{\varepsilon}_{v+1;j}+\frac{1}{|R_k^j|}\sum_{v_1,v_2\in R_k^j}\sum_{l=0}^{\infty}B_{v_1,l}\underline{\varepsilon}_{v_1-l}\underline{\varepsilon}_{v_2+1;j}\right)\right)\\
=&E\left((nP_k^j)^{-1}\left(\sum_{r\in R_k^j}\sum_{l=0}^{\infty}B_{r,l}\underline{\mu}\underline{\mu}_j-\frac{1}{|R_k^j|}\sum_{r\in R_k^j}\sum_{v\in R_k^j}\sum_{l=0}^{\infty}B_{v,l}\underline{\mu}\underline{\mu}_j-\frac{1}{|R_k^j|}\sum_{r\in R_k^j}\sum_{\substack{v\in R_k^j\\ v=r+l+1}}\sum_{l=0}^{\infty}B_{v,l}\Sigma_{j\cdot}\right)\right)\\
=&E\left((nP_k^j)^{-1}\left(-\frac{1}{|R_k^j|}\sum_{r\in R_k^j}\sum_{\substack{v\in R_k^j\\ v=r+l+1}}\sum_{l=0}^{\infty}B_{v,l}\Sigma_{j\cdot}\right)\right)=\mathcal{O}((nP_k^j)^{-1}).
\end{aligned}
$$

Furthermore, we have

$$
\begin{aligned}
&\left((nP_k^j)^{-1}\sum_{r\in R_k^j}(\underline{X}_r-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{X}_v)(\underline{\varepsilon}_{r+1;j}-\frac{1}{|R_k^j|}\sum_{v\in R_k^j}\underline{\varepsilon}_{v+1;j})\right)\\
=&\left((nP_k^j)^{-1}\sum_{r\in R_k^j}\left(\underline{X}_r-E\left[\underline{X}_1|\phi_k^j(1)\right]\right)\left(\underline{\varepsilon}_{r+1;j}-\underline{\mu}_j\right)\right)+o_P(1).
\end{aligned}
$$

Hence, a bound for the variance of the latter term is sufficient. We have

$$
\begin{aligned}
&\mathrm{Var}\left(\left((nP_k^j)^{-1}\sum_{r\in R_k^j}\left(\underline{X}_r-E\left[\underline{X}_1|\phi_k^j(1)\right]\right)\left(\underline{\varepsilon}_{r+1;j}-\underline{\mu}_j\right)\right)\right)\\
=&E\left((nP_k^j)^{-2}\sum_{r_1=0}^{n-1}\sum_{r_2=0}^{n-1}(\underline{\varepsilon}_{r_1+1;j}-\underline{\mu}_j)(\underline{\varepsilon}_{r_2+1;j}-\underline{\mu}_j)(\underline{X}_{r_1}-E[\underline{X}_1|\phi_k^j(1)])\left(\underline{X}_{r_2}-E[\underline{X}_1|\phi_k^j(1)]\right)^{\top}\phi_k^j(r_1)\phi_k^j(r_2)\right)\\
=&\Sigma_{jj}E\left((nP_k^j)^{-2}\sum_{r=0}^{n-1}(\underline{X}_r-E[\underline{X}_1|\phi_k^j(1)])(\underline{X}_r-E[\underline{X}_1|\phi_k^j(1)])^{\top}\phi_k^j(r)\right)\\
=&\Sigma_{jj}\frac{1}{n}(P_k^j)^{-1}\mathrm{Var}\left(\underline{X}_1|\phi_k^j(1)\right)=\mathcal{O}(1/n(P_j^k)^{-1}).
\end{aligned}
$$

Thus, $\tilde{\alpha}_{jk}$ is asymptotically unbiased and consistent with variance (without error terms of minor order) $\Sigma_{jj}(nP_k^j)^{-1}\left(\mathrm{Var}\left(\underline{X}_1|\phi_k^j(1)\right)\right)^{-1}$. As already mentioned, since P_k^j can be relatively small, we write $\tilde{\alpha}_{jk}=\alpha_{jk}+\mathcal{O}((nP(Ad_{1;j\cdot}=\tilde{A}d_{k;j\cdot},Ad_{1;\cdot j}=\tilde{A}d_{k;\cdot j}))^{-1/2})$. Furthermore, we have

$$\hat{\underline{\mu}}_j = \frac{1}{|R_k^j|}\sum_{r\in R_k^j} \underline{X}_{r+1;j} - \tilde{\alpha}_{jk}\underline{X}_r = \frac{1}{|R_k^j|}\sum_{r\in R_k^j} (\alpha_{jk}-\tilde{\alpha}_{jk})\underline{X}_r + \underline{\varepsilon}_{r+1;j} = \underline{\mu}_j + \mathcal{O}((nP_k^j))^{-1/2})$$

and the assertion follows. □

Proof of Theorem 3.3.4. Let $s = 1,\dots,d$. In order to simplify notation, it is assumed without loss of generality that $S_s = \{1,\dots,d\}$. Thus, we have $\underline{X}_{t;s} = \tilde{\alpha}_{\cdot s} Y_{t-1}^s + \varepsilon_{t;s}$, where $Y_t^s := Ad_{t;\cdot s} \circledast \underline{X}_t = \sum_{j=0}^{\infty} g(Ad_t,\dots,Ad_{t-j})\underline{\varepsilon}_{t-j} = \sum_{j=0}^{\infty} \tilde{B}_{t,j}\underline{\varepsilon}_{t-j}$ for some function g. Thus, the process $Y^s := \{Y_t^s : t \in \mathbb{Z}\}$ fits in the framework of a doubly stochastic network linear process. The only difference is that the first coefficient is not normalized to the identity matrix. Due to the assumptions, Lemma 3.2.2 implies that Y^s is stationary and possesses an absolute summable ACF. Firstly, the consistency for $\hat{\alpha}_{\cdot s}$ given by the linear system (3.3.7) is shown. For the left-hand-side term of the corresponding linear system we have

$$\frac{1}{n}\sum_{t=1}^{n} \underline{X}_{t;s} Y_{t-1}^s - (1/n)^2 \sum_{t_1,t_2=1}^{n} Y_{t_1-1}^s \underline{X}_{t_2;s} =$$
$$\left(\frac{1}{n}\sum_{t=0}^{n-1} Y_t^s (Y_t^s)^\top - (\frac{1}{n}\sum_{t=0}^{n-1} Y_t^s)(\frac{1}{n}\sum_{t=0}^{n-1} Y_t^s)^\top\right)\alpha_{s\cdot}^\top + \left(\frac{1}{n}\sum_{t=1}^{n}\underline{\varepsilon}_{t;s} Y_{t-1}^s - (\frac{1}{n}\sum_{t=0}^{n-1} Y_t^s)(\frac{1}{n}\sum_{t=1}^{n}\underline{\varepsilon}_{t;s})\right).$$

Since $\underline{\varepsilon}$ is i.i.d., we have $(\frac{1}{n}\sum_{t=1}^{n}\underline{\varepsilon}_{t;j}) \xrightarrow{P} \underline{\mu}_j$, as $n\to\infty$. Theorem 3.3.1 implies $\frac{1}{n}\sum_{t=0}^{n-1} Y_t^s = EY_1^s + \mathcal{O}_P(n^{-1/2})$. Since Y^s is one-side and $\underline{\varepsilon}$ is i.i.d., we have $E\left(\frac{1}{n}\sum_{t=1}^{n}\underline{\varepsilon}_{t;j} Y_{t-1}^s - (\frac{1}{n}\sum_{t=0}^{n-1} Y_t^s)(\frac{1}{n}\sum_{t=1}^{n}\underline{\varepsilon}_{t;j})\right) = 0$. Before having a look at the variance, first consider that

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n}(\underline{\varepsilon}_{t;s} - \frac{1}{n}\sum_{l=1}^{n}\underline{\varepsilon}_{l;s})(Y_{t-1}^s - \frac{1}{n}\sum_{l=0}^{n-1} Y_l^s) &= \frac{1}{n}\sum_{t=1}^{n}(\underline{\varepsilon}_{t;s}-\underline{\mu}_s)(Y_{t-1}^s - EY_1^s) + (\frac{1}{n}\sum_{l=0}^{n-1} Y_l^s - EY_1^s)(\frac{1}{n}\sum_{t=1}^{n}\underline{\varepsilon}_{t;s} - \underline{\mu}_s)\\ &= \frac{1}{n}\sum_{t=1}^{n}(\underline{\varepsilon}_{t;s}-\underline{\mu}_s)(Y_{t-1}^s - EY_1^s) + \mathcal{O}_P(n^{-1}).\end{aligned}$$

With this we have

$$\begin{aligned}\mathrm{Var}(\frac{1}{n}\sum_{t=1}^{n}(\underline{\varepsilon}_{t;s}-\underline{\mu}_s)(Y_{t-1}^s - EY_1^s)) &= \frac{1}{n^2}\sum_{r_1=0}^{n}\sum_{r_2=0}^{n} E[(\underline{\varepsilon}_{r_1;s}-\underline{\mu}_s)(Y_{r_1-1}^s - EY_1^s)(\underline{\varepsilon}_{r_2;s}-\underline{\mu}_s)(Y_{r_2-1}^s - EY_1^s)^\top]\\ &= \frac{1}{n^2}\sum_{r=0}^{n} E[(\underline{\varepsilon}_{r;s}-\underline{\mu}_s)(\underline{\varepsilon}_{r;s}-\underline{\mu}_s)(Y_{r-1}^s - EY_1^s)(Y_{r-1}^s - EY_1^s)^\top] = \Sigma_{ss}\frac{1}{n}\Gamma_{Y^s}(0). \qquad (3.7.3)\end{aligned}$$

Theorem 3.3.2 implies $\left(\frac{1}{n}\sum_{t=0}^{n-1} Y_t^s (Y_t^s)^\top - (\frac{1}{n}\sum_{t=0}^{n-1} Y_t^s)(\frac{1}{n}\sum_{t=0}^{n-1} Y_t^s)^\top\right) = \Gamma_{Y^s}(0) + \mathcal{O}_P(n^{-1/2})$. Thus, $\left(\frac{1}{n}\sum_{t=0}^{n-1} Y_t^s (Y_t^s)^\top - (\frac{1}{n}\sum_{t=0}^{n-1} Y_t^s)(\frac{1}{n}\sum_{t=0}^{n-1} Y_t^s)^\top\right)$ is invertible with high probability for n large enough and and we have

$$\hat{\alpha}_{s\cdot}^{\top} = \left(\frac{1}{n}\sum_{t=0}^{n-1} Y_t^s (Y_t^s)^{\top} - \frac{1}{n^2}\sum_{t_1,t_2=0}^{n-1} Y_{t_1}^s (Y_{t_2}^s)^{\top}\right)^{-1} \left(\frac{1}{n}\sum_{t=1}^{n} \underline{X}_{t;s} Y_{t-1}^s - \frac{1}{n^2}\sum_{t_1,t_2=1}^{n} Y_{t_1-1}^s \underline{X}_{t_2;s}\right)$$
$$= \alpha_{s\cdot} + \mathcal{O}_P(n^{-1/2}).$$

Furthermore, we have

$$\hat{\underline{\mu}}_s - \underline{\mu}_s = \frac{1}{n}\sum_{t=1}^{n} \underline{X}_{t;s} - \hat{\alpha}_{s\cdot} 1/n \sum_{t=0}^{n-1} Y_t^s - \underline{\mu}_s = (\alpha_{s\cdot} - \hat{\alpha}_{s\cdot})\frac{1}{n}\sum_{t=1}^{n} Y_{t-1}^s + \frac{1}{n}\sum_{t=1}^{n}(\underline{\varepsilon}_{t;s} - \underline{\mu}_s)$$
$$= (\alpha_{s\cdot} - \hat{\alpha}_{s\cdot}) E Y_1^s + \frac{1}{n}\sum_{t=1}^{n}(\underline{\varepsilon}_{t;s} - \underline{\mu}_s) + \mathcal{O}(n^{-1})$$
$$= \frac{1}{n}\sum_{t=1}^{n}(\underline{\varepsilon}_{t;s} - \underline{\mu}_s)(1 + (Y_{t-1}^s - E Y_1^s)^{\top} \Gamma_{Y^s}(0)^{-1} E Y_1^s) + \mathcal{O}(n^{-1}).$$

This is centered and due to the independence of $\underline{\varepsilon}_t$ and Y_{t-1}^s the variance is

$$\mathrm{Var}\left(\frac{1}{n}\sum_{t=1}^{n}(\underline{\varepsilon}_{t;s} - \underline{\mu}_s)(1 + (Y_{t-1}^s - E Y_1^s)^{\top} \Gamma_{Y^s}(0)^{-1} E Y_1^s)\right) = \frac{1}{n}\Sigma_{ss} E[(1 + (Y_1^s - E Y_1^s)^{\top} \Gamma_{Y^s}(0)^{-1} E Y_1^s)^2]$$
$$= \frac{1}{n}\Sigma_{ss}(1 + E(Y_1^s)^{\top} \Gamma_{Y^s}(0)^{-1} E(Y_1^s)^{\top}).$$

Using the M-approximation $(Y_t^s)^M = \sum_{j=0}^{M} \tilde{B}t, j\underline{\varepsilon}_{t-j}$ gives an α-mixing process with $\sum_{n=0}^{\infty} \alpha((Y_t^s)^M, n)^{1/5} < \infty$. This α-mixing property is obtained for $(Y_{t-1}^s)^M \underline{\varepsilon}_t$. Thus, the same ideas used in the proof of Theorem 3.3.1 and (3.7.3) lead to, as $n \to \infty$,

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\underline{\varepsilon}_{t;s} - \underline{\mu}_s)(Y_{t-1}^s - E Y_1^s) \xrightarrow{D} \mathcal{N}(0, \Sigma_{ss}\Gamma_{Y^s}(0)).$$

Hence, we get

$$\sqrt{n}(\hat{\alpha}_{s\cdot} - \alpha_{s\cdot}) \xrightarrow{D} \mathcal{N}(0, \Sigma_{ss}\Gamma_{Y^s}(0)^{-1}).$$

With similar arguments we get that

$$\sqrt{n}(\hat{\underline{\mu}}_s - \underline{\mu}_s) \xrightarrow{D} \mathcal{N}(0, \Sigma_{ss}(1 + E(Y_1^s)^{\top} \Gamma_{Y^s}(0)^{-1} E(Y_1^s)^{\top})).$$

Furthermore, by using the same arguments as in the variance calculation we obtain, as $n \to \infty, k \in \{1, \ldots, d\}$:

$$\mathrm{Cov}(\sqrt{n}(\hat{\alpha}_{s\cdot} - \alpha_{s\cdot}), \sqrt{n}(\hat{\underline{\mu}}_s - \underline{\mu}_s)) \to \Sigma_{ss}\Gamma_{Y^s}(0)^{-1} E Y_1^s,$$

$$\mathrm{Cov}(\sqrt{n}(\hat{\alpha}_{s\cdot} - \alpha_{s\cdot}), (\hat{\alpha}_{k\cdot} - \alpha_{k\cdot})) \to \Sigma_{sk}\Gamma_{Y^s}(0)^{-1}\Gamma_{Y^sY^k}(0)\Gamma_{Y^k}(0)^{-1},$$

$$\mathrm{Cov}(\sqrt{n}(\hat{\alpha}_{s\cdot} - \alpha_{s\cdot}), \sqrt{n}(\hat{\underline{\mu}}_k - \underline{\mu}_k)) \to \Sigma_{sk}\Gamma_{Y_s}(0)^{-1}\Gamma_{Y^sY^k}(0)\Gamma_{Y^k}(0)^{-1} E Y_1^k,$$

and

$$\mathrm{Cov}(\sqrt{n}(\hat{\underline{\mu}}_s - \underline{\mu}_s), \sqrt{n}(\hat{\underline{\mu}}_k - \underline{\mu}_k)) \to \Sigma_{sk}(1 + E(Y_1^k)^\top \Gamma_{Y_s}(0)^{-1}\Gamma_{Y^sY^k}(0)\Gamma_{Y^k}(0)^{-1}EY_1^k).$$

With this, the assertion follows. □

Proof of Theorem 3.3.5 . We have

$$\underline{X}_t = \alpha\underline{X}_{t-1} + \beta h(Ad_{t-1})\underline{X}_{t-1} + \mu + \underline{\varepsilon}_t = \alpha\underline{X}_{t-1} + \beta Y_{t-1} + \mu + \underline{\varepsilon}_t,$$

where $Y_t = h(Ad_{t-1})\underline{X}_{t-1}$. To shorten the notation let $\tilde{\sum}_{t,s} = \frac{1}{nd}\sum_{t=1}^{n}\sum_{s=1}^{d}$. The linear system (3.3.12) can be written as

$$\begin{pmatrix} \tilde{\sum}_{t,s}\underline{X}_{t-1;s}\underline{X}_{t;s} - (\tilde{\sum}_{t,s}\underline{X}_{t;s})^2 \\ \tilde{\sum}_{t,s}Y_{t-1;s}\underline{X}_{t;s} - \tilde{\sum}_{t,s}Y_{t-1;s}\tilde{\sum}_{t,s}\underline{X}_{t;s} \\ \underline{X}_{t;s} \end{pmatrix} =$$
$$\begin{pmatrix} \tilde{\sum}_{t,s}\underline{X}_{t-1;s}^2 - \tilde{\sum}_{t,s}\underline{X}_{t-1;s} & \tilde{\sum}_{t,s}\underline{X}_{t-1;s}Y_{t-1;s} - \tilde{\sum}_{t,s}Y_{t-1;s}\tilde{\sum}_{t,s}\underline{X}_{t-1;s} & 0 \\ \underline{X}_{t-1;s}Y_{t-1;s} - \tilde{\sum}_{t,s}Y_{t-1;s}\tilde{\sum}_{t,s}\underline{X}_{t-1;s} & Y_{t-1;s}^2 - (\tilde{\sum}_{t,s}Y_{t-1;s})^2 & 0 \\ \underline{X}_{t-1;s} & Y_{t-1;s} & 1 \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \\ \hat{\mu} \end{pmatrix}.$$

Hence, with one additional step the linear sytem gives the following linear equations:

$$\hat{\alpha} = \alpha + \left(\tilde{\sum_{t,s}}(\underline{X}_{t-1;s} - \tilde{\sum_{t,s}}\underline{X}_{t-1;s})^2 - (\tilde{\sum_{t,s}}(Y_{t-1;s} - \tilde{\sum_{t,s}}Y_{t-1;s})^2)^{-1}\tilde{\sum_{t,s}}(\underline{X}_{t-1;s} - \tilde{\sum_{t,s}}\underline{X}_{t-1;s})(Y_{t-1;s} - \tilde{\sum_{t,s}}Y_{t-1;s})\right)^{-1}$$
$$\left(\tilde{\sum_{t,s}}(\underline{X}_{t-1;s} - \tilde{\sum_{t,s}}\underline{X}_{t-1;s})(\underline{\varepsilon}_{t;s} - \tilde{\sum_{t,s}}\underline{\varepsilon}_{t;s}) - (\tilde{\sum_{t,s}}(Y_{t-1;s} - \tilde{\sum_{t,s}}Y_{t-1;s})^2)^{-1}\tilde{\sum}_{t,s}(Y_{t-1;s} - \tilde{\sum_{t,s}}Y_{t-1;s})(\underline{\varepsilon}_{t;s} - \tilde{\sum_{t,s}}\underline{\varepsilon}_{t;s})\right)$$

$$\hat{\beta} = \beta + \left(\tilde{\sum_{t,s}}(Y_{t-1;s} - \tilde{\sum_{t,s}}Y_{t-1;s})^2 - (\tilde{\sum_{t,s}}(\underline{X}_{t-1;s} - \tilde{\sum_{t,s}}\underline{X}_{t-1;s})^2)^{-1}\tilde{\sum_{t,s}}(\underline{X}_{t-1;s} - \tilde{\sum_{t,s}}\underline{X}_{t-1;s})(Y_{t-1;s} - \tilde{\sum_{t,s}}Y_{t-1;s})\right)^{-1}$$
$$\left(\tilde{\sum_{t,s}}(Y_{t-1;s} - \tilde{\sum_{t,s}}Y_{t-1;s})(\underline{\varepsilon}_{t;s} - \tilde{\sum_{t,s}}\underline{\varepsilon}_{t;s}) - (\tilde{\sum_{t,s}}(\underline{X}_{t-1;s} - \tilde{\sum_{t,s}}\underline{X}_{t-1;s})^2)^{-1}\tilde{\sum_{t,s}}(\underline{X}_{t-1;s} - \tilde{\sum_{t,s}}\underline{X}_{t-1;s})(\underline{\varepsilon}_{t;s} - \tilde{\sum_{t,s}}\underline{\varepsilon}_{t;s})\right)$$

$$\hat{\mu} = \mu + \tilde{\sum_{t,s}}\underline{X}_{t-1;s}(\hat{\alpha} - \alpha) + \tilde{\sum_{t,s}}Y_{t-1;s}(\hat{\beta} - \beta) + \tilde{\sum_{t,s}}\underline{\varepsilon}_{t;s}.$$

Furthermore, we have $E\tilde{\sum}_{t,s}\underline{X}_{t-1;s} = 1/d\sum_{s=1}^{d}E\underline{X}_{1;s}$ and

$$\mathrm{Var}\tilde{\sum_{t,s}}\underline{X}_{t-1;s} = \frac{1}{n}\sum_{h=-n+1}^{n-1}\frac{n-|h|}{n}\frac{1}{d^2}\mathbb{1}\Gamma_X(h)\mathbb{1} = \mathcal{O}_P(1/n).$$

Note that $\mathbb{1}\Gamma_X(h)\mathbb{1}/d^2$ depends on the linear dependence between the components. Under some moderate assumptions on the cross dependence structure this could be of order $\mathcal{O}(1/d)$. Since d is fixed and nothing is assumed for the cross dependence structure we drop d in the $\mathcal{O}$-notation. We have

$$\tilde{\sum}_{t,s}(\underline{X}_{t-1;s} - \tilde{\sum}_{t,s}\underline{X}_{t-1;s})^2 = \tilde{\sum}_{t,s}\underline{X}^2_{t-1;s} - (\tilde{\sum}_{t,s}\underline{X}_{t-1;s})^2$$

and

$$\mathrm{Var}(\tilde{\sum}_{t,s}\underline{X}^2_{t,s}) = \frac{1}{n^2d^2}\sum_{s_1,s_2=1}^{d}\sum_{t_1,t_2=1}^{n}\underline{X}^2_{t_1,s_1}\underline{X}^2_{t_2,s_2} = \frac{1}{nd^2}\sum_{s_1,s_2=1}^{d}\sum_{h=-n+1}^{n-1}\sum_{n-|h|} n\underline{X}^2_{0,s_1}\underline{X}^2_{h,s_2} = \mathcal{O}(1/n),$$

by applying the same arguments used to bound the variance of $\hat{\Gamma}_X(h)$ in the proof of Theorem 3.3.2 and using the Assumptions 1, 2, and 3. Hence, we get

$$\tilde{\sum}_{t,s}(\underline{X}_{t-1;s} - \tilde{\sum}_{t,s}\underline{X}_{t-1;s})^2 = 1/d\sum_{s=1}^{d}E\underline{X}^2_{1;s} - (1/d\sum_{s=1}^{d}E\underline{X}_{1;s})^2 + \mathcal{O}_P(1/\sqrt{n}).$$

Similarly, with Assumptions 1, 2, and 4 we get

$$\tilde{\sum}_{t,s}(Y_{t-1;s} - \tilde{\sum}_{t,s}Y_{t-1;s})^2 = 1/d\sum_{s=1}^{d}EY^2_{1;s} - (1/d\sum_{s=1}^{d}EY_{1;s})^2 + \mathcal{O}_P(1/\sqrt{n})$$

and

$$\tilde{\sum}_{t,s}(\underline{X}_{t-1;s} - \tilde{\sum}_{t,s}\underline{X}_{t-1;s})(Y_{t-1;s} - \tilde{\sum}_{t,s}Y_{t-1;s}) = \frac{1}{d}\sum_{s=1}^{d}EY_{1,s}\underline{X}_{1;s} - (\frac{1}{d}\sum_{s=1}^{d}EY_{1;s})(\frac{1}{d}\sum_{s=1}^{d}E\underline{X}_{1;s}) + \mathcal{O}_P(\frac{1}{\sqrt{n}}).$$

We have $\tilde{\sum}_{t,s}(\underline{X}_{t-1;s} - \tilde{\sum}_{t,s}\underline{X}_{t-1;s})(\underline{\varepsilon}_{t;s} - \tilde{\sum}_{t,s}\underline{\varepsilon}_{t;s}) = \tilde{\sum}_{t,s}\underline{X}_{t-1;s}\underline{\varepsilon}_{t;s} - (\tilde{\sum}_{t,s}\underline{X}_{t;s})(\tilde{\sum}_{t,s}\underline{\varepsilon}_{t;s})$. Note that $\underline{\varepsilon}$ is centered and since $\underline{X}_{t-1}$ and $\underline{\varepsilon}_t$ are independent, the mean is 0. Since $\underline{\varepsilon}$ is i.i.d. the variance of $\tilde{\sum}_{t,s}\underline{X}_{t-1;s}\underline{\varepsilon}_{t;s}$ is

$$\mathrm{Var}(\tilde{\sum}_{t,s}\underline{X}_{t-1;s}\underline{\varepsilon}_{t;s}) = \frac{1}{n^2d^2}\sum_{t_1,t_2=1}^{n}\sum_{s_1,s_2=1}^{d}E(\underline{X}_{t_1-1;s_1}\underline{X}_{t_2-1;s_2}\underline{\varepsilon}_{t_1;s_1}\underline{\varepsilon}_{t_2:s_2}) = \frac{1}{nd^2}\sum_{s_1,s_2=1}^{d}E(\underline{X}_{1,s_1}\underline{X}_{1,s_2})\Sigma_{s_1,s_2}.$$

Furthermore, due to Assumptions 2, we have $\mathrm{Var}\left((\tilde{\sum}_{t,s}\underline{X}_{t;s})(\tilde{\sum}_{t,s}\underline{\varepsilon}_{t;s})\right) = \mathcal{O}(1/n^2)$. Thus,

$$\mathrm{Var}(\tilde{\sum}_{t,s}(\underline{X}_{t-1;s} - \tilde{\sum}_{t,s}\underline{X}_{t-1;s})(\underline{\varepsilon}_{t;s} - \tilde{\sum}_{t,s}\underline{\varepsilon}_{t;s})) = \frac{1}{nd^2}\sum_{s_1,s_2=1}^{d}E(\underline{X}_{1,s_1}\underline{X}_{1,s_2})\Sigma_{s_1,s_2} + \mathcal{O}(n^{-3/2}).$$

Similarly, we get

$$\mathrm{Var}(\tilde{\sum}_{t,s}(Y_{t-1;s} - \tilde{\sum}_{t,s}Y_{t-1;s})(\underline{\varepsilon}_{t;s} - \tilde{\sum}_{t,s}\underline{\varepsilon}_{t;s})) = \frac{1}{nd^2}\sum_{s_1,s_2=1}^{d}E(Y_{1,s_1}Y_{1,s_2})\Sigma_{s_1,s_2} + \mathcal{O}(n^{-3/2})$$

and

$$\mathrm{Cov}(\tilde{\sum}_{t,s}(Y_{t-1;s}-\tilde{\sum}_{t,s}Y_{t-1;s})(\underline{\varepsilon}_{t;s}-\tilde{\sum}_{t,s}\underline{\varepsilon}_{t;s}),\tilde{\sum}_{t,s}(\underline{X}_{t-1;s}-\tilde{\sum}_{t,s}\underline{X}_{t-1;s})(\underline{\varepsilon}_{t;s}-\tilde{\sum}_{t,s}\underline{\varepsilon}_{t;s}))=\frac{1}{nd^2}\sum_{s_1,s_2=1}^{d}E(Y_{1,s_1}\underline{X}_{1,s_2})\Sigma_{s_1,s_2}+\mathcal{O}(n^{-3/2}).$$

Denote $\tilde{\sum}_s := 1/d\sum_{s=1}^{d}$ and $\tilde{\sum}_{s_1,s_2} := 1/d^2\sum_{s_1,s_2=1}^{d}$. Thus, we have, as $n\to\infty$, $E\sqrt{n}(\hat{\alpha}-\alpha)\to 0, E\sqrt{n}(\hat{\beta}-\beta)\to 0, E\sqrt{n}(\hat{\mu}-\mu)=0$. Furthermore, denote $\bar{\bar{Y}}^2 := (\tilde{\sum}_s EY_{1;s}^2-(\tilde{\sum}_s EY_{1,s})^2)$, $\bar{\bar{\underline{X}}}^2 := (\tilde{\sum}_s E\underline{X}_{1;s}^2-(\tilde{\sum}_s E\underline{X}_{1,s})^2)$. We get

$$\mathrm{Var}(\sqrt{n}(\hat{\alpha}-\alpha))\to\left(\tilde{\sum}_s E\underline{X}_{1;s}^2-(\tilde{\sum}_s E\underline{X}_{1,s})^2-\bar{\bar{Y}}^{2^{-1}}(EY_{1;s}\underline{X}_{1;s}-(\tilde{\sum}_s E\underline{X}_{1,s})(\tilde{\sum}_s EY_{1,s}))\right)^{-2}$$
$$\times\left(\tilde{\sum}_{s_1,s_2}\Sigma_{s_1,s_2}[E(\underline{X}_{1,s_1}\underline{X}_{1,s_2})+\bar{\bar{Y}}^{2^{-1}}E(Y_{1,s_1}\underline{X}_{1,s_2})+\bar{\bar{Y}}^{2^{-2}}E(Y_{1,s_1}Y_{1,s_2})]\right),$$

$$\mathrm{Var}(\sqrt{n}(\hat{\beta}-\beta))\to\left(\tilde{\sum}_s E\underline{X}_{1;s}^2-(\tilde{\sum}_s E\underline{X}_{1,s})^2-\bar{\bar{\underline{X}}}^{2^{-1}}(\tilde{\sum}_s EY_{1;s}\underline{X}_{1;s}-(\tilde{\sum}_s E\underline{X}_{1,s})(\tilde{\sum}_s EY_{1,s}))\right)^{-2}$$
$$\times\left(\tilde{\sum}_{s_1,s_2}\Sigma_{s_1,s_2}[E(Y_{1,s_1}Y_{1,s_2})+\bar{\bar{\underline{X}}}^{2^{-1}}E(Y_{1,s_1}\underline{X}_{1,s_2})+\bar{\bar{\underline{X}}}^{2^{-2}}E(\underline{X}_{1,s_1}\underline{X}_{1,s_2})]\right),$$

$$\mathrm{Cov}(\sqrt{n}(\hat{\alpha}-\alpha),\sqrt{n}(\hat{\beta}-\beta))\to\left(\tilde{\sum}_s E\underline{X}_{1;s}^2-(\tilde{\sum}_s E\underline{X}_{1,s})^2-\bar{\bar{\underline{X}}}^{2^{-1}}(\tilde{\sum}_s EY_{1;s}\underline{X}_{1;s}-(\tilde{\sum}_s E\underline{X}_{1,s})(\tilde{\sum}_s EY_{1,s}))\right)^{-1}$$
$$\times\left(\tilde{\sum}_s E\underline{X}_{1;s}^2-(\tilde{\sum}_s E\underline{X}_{1,s})^2-\bar{\bar{Y}}^{2^{-1}}(\tilde{\sum}_s EY_{1;s}\underline{X}_{1;s}-(\tilde{\sum}_s E\underline{X}_{1,s})(\tilde{\sum}_s EY_{1,s}))\right)^{-1}$$
$$\times\left(\tilde{\sum}_{s_1,s_2}\Sigma_{s_1,s_2}E(Y_{1,s_1}\underline{X}_{1,s_2})(1+\bar{\bar{\underline{X}}}^{2^{-1}}\bar{\bar{Y}}^{2^{-1}})\right.$$
$$\left.-\bar{\bar{\underline{X}}}^{2^{-1}}E(\underline{X}_{1,s_1}\underline{X}_{1,s_2})-\bar{\bar{Y}}^{2^{-1}}E(Y_{1,s_1}Y_{1,s_2})\right).$$

Furthermore, we have $\sqrt{n}(\hat{\mu}-\mu)=\tilde{\sum}_{t,s}\underline{X}_{t-1;s}\sqrt{n}(\hat{\alpha}-\alpha)+\tilde{\sum}_{t,s}Y_{t-1;s}\sqrt{n}(\hat{\beta}-\beta)+\sqrt{n}\tilde{\sum}_{t,s}\underline{\varepsilon}_{t-1;s}$. Note that $\mathrm{Var}(\tilde{\sum}_{t,s}\underline{X}_{t-1;s}\sqrt{n}(\hat{\alpha}-\alpha))\to 0$ since $\mathrm{Var}(\tilde{\sum}_{t,s}\underline{X}_{t-1;s})=O(1/n)$ and $\mathrm{Var}(\sqrt{n}\hat{\alpha}-\alpha)=O(1)$. Similar arguments apply to the following parts. Thus, as $n\to\infty$, $\mathrm{Var}(\sqrt{n}(\hat{\mu}-\mu))\to 1/d^2\sum_{s_1,s_2}^{d}\Sigma_{s_1,s_2}$. Furthermore, note that due to $\{\underline{\varepsilon}_t, t\in\mathbb{Z}\}$ being i.i.d. we have

$$\mathrm{Cov}(\sqrt{n}\tilde{\sum}_{t,s}\underline{X}_{t-1,s}\underline{\varepsilon}_{t,s},\sqrt{n}\tilde{\sum}_{t,s}\underline{\varepsilon}_{t,s})=1/d^2\sum_{s_1,s_2}^{d}E\underline{X}_{1,s_1}\Sigma_{s_1,s_2}$$

and $\mathrm{Cov}(\sqrt{n}\tilde{\sum}_{t,s}\underline{X}_{t-1,s}\tilde{\sum}_{t,s}\underline{\varepsilon}_{t,s}, \tilde{\sum}_{t,s}\underline{\varepsilon}_{t,s}) = O(n^{-1/2})$. Hence,

$$\mathrm{Cov}(\sqrt{n}(\hat{\mu}-\mu), \sqrt{n}(\hat{\alpha}-\alpha)) \to \left(\tilde{\sum_s} E\underline{X}_{1;s}^2 - (\tilde{\sum_s} E\underline{X}_{1,s})^2 - \bar{\bar{Y}}^{2^{-1}}(\tilde{\sum_s} EY_{1;s}\underline{X}_{1;s} - (\tilde{\sum_s} E\underline{X}_{1,s})(\tilde{\sum_s} EY_{1,s}))\right)^{-1}$$
$$\times \left(\sum_{s_1,s_2}\Sigma_{s_1,s_2}[E\underline{X}_{1;s_1} + (\tilde{\sum_s} EY_{1;s}^2 - (\tilde{\sum_s} EY_{1;s})^2)^{-1}EY_{1;s_1}]\right),$$

and

$$\mathrm{Cov}(\sqrt{n}(\hat{\mu}-\mu), \sqrt{n}(\hat{\beta}-\beta)) \to \left(\tilde{\sum_s} E\underline{X}_{1;s}^2 - (\tilde{\sum_s} E\underline{X}_{1,s})^2 - \bar{\underline{X}}^{2^{-1}}(\tilde{\sum_s} EY_{1;s}\underline{X}_{1;s} - (\tilde{\sum_s} E\underline{X}_{1,s})(\tilde{\sum_s} EY_{1,s}))\right)^{-1}$$
$$\times \left(\sum_{s_1,s_2}\Sigma_{s_1,s_2}[EY_{1;s_1} + (\tilde{\sum_s} E\underline{X}_{1;s}^2 - (\tilde{\sum_s} E\underline{X}_{1;s})^2)^{-1}E\underline{X}_{1;s_1}]\right),$$

With this, the asymptotic normality follows by using the same arguments as in the proof of Theorem 3.3.1 and 3.3.4. □

Proof of Lemma 3.3.6. To simplify notation we consider the case $q = 1$. The proof can be transferred with the same notation as used in the proof of Lemma 3.2.4 to $q \geq 1$. As given by the proof of Lemma 3.2.4, $E\log f(Ad_1) < 0$ and **Ad** α-mixing implies an almost surely exponential decay with rate $\rho \in (0,1)$ for $\prod_{s=1}^{j}\|f(Ad_{-s})\|$. Consequently, for finite p_1, p_2 we have $\sum_{s=0}^{\infty} s^{p_1} E\|\prod_{i=1}^{s} f(Ad_{-i})\|^{p_2} < \infty$. Furthermore, regarding Theorem 3.3.4 and 3.3.5, for some measurable and bounded function $\tilde{f}$ we have

$$\sum_{j=0}^{\infty} E|e_i\tilde{f}(Ad_h)\prod_{s=1}^{j} f(Ad_{l-s})\mu|^p \leq \sum_{j=0}^{\infty}(E\|\tilde{f}(Ad_0)\|^{2p})^{1/2}(E\prod_{s=1}^{j}\|f(Ad_{-s}\|^{2p})^{1/2}\|\mu\|$$
$$\leq C\sum_{j=0}^{\infty}(E\prod_{s=1}^{j}\|f(Ad_s)\|^{2p})^{1/2} < \infty,$$

due to the a.s. exponential decay. It remains

$$\sum_{h\in\mathbb{Z}}\sum_{s_1,s_2=0}^{\infty}|\mathrm{Cov}\left(\prod_{i_1=1}^{s_1} f(Ad_{h-i_1})\underline{\mu}, \prod_{i_2=1}^{s_2} f(Ad_{-i_2}\underline{\mu})\right)| < \infty.$$

For this, consider

$$\prod_{i_1=1}^{s_1} f(Ad_{h-i_1})\underline{\mu} = g(Ad_{h-1}, \ldots, Ad_{h-s_1})$$

and

$$\prod_{i_2=1}^{s_2} f(Ad_{-i_2}\underline{\mu}) = \tilde{g}(Ad_{-1}, \ldots, Ad_{-s_2})$$

for some measurable function $g, \tilde{g}$. Hence, for $h \geq 0$ and by applying (Bradley, 2007, Corollary 10.16) we have

$$\begin{aligned}
&\sum_{h=0}^{\infty} \sum_{s_1,s_2=0}^{\infty} |\mathrm{Cov}\left(\prod_{i_1=1}^{s_1} f(Ad_{h-i_1})\underline{\mu}, \prod_{i_2=1}^{s_2} f(Ad_{-i_2}\underline{\mu})\right)| \\
&\leq \sum_{s_1,s_2=0}^{\infty} \sum_{h=0}^{s_2} \left(E\left(\prod_{i=1}^{s_1} f(Ad_{-i})\underline{\mu}\right)^4 E\left(\prod_{i=1}^{s_2} f(Ad_{-i})\underline{\mu}\right)^4\right)^{1/4} 4\alpha(\mathbf{Ad}, \max(0, h-s-2+1))^{1/2} \\
&+ \sum_{s_1,s_2=0}^{\infty} \sum_{h=s_2+1}^{\infty} \left(E\left(\prod_{i=1}^{s_1} f(Ad_{-i})\underline{\mu}\right)^4 E\left(\prod_{i=1}^{s_2} f(Ad_{-i})\underline{\mu}\right)^4\right)^{1/4} 4\alpha(\mathbf{Ad}, \max(0, h-s-2+1))^{1/2} \\
&\leq C \sum_{h=0}^{\infty} \alpha(Ad, h)^{1/2} < \infty.
\end{aligned}$$

The sum for $h < 0$ can be bounded with the same arguments. Hence, the assertion follows. □

Bibliography

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567.

Bickel, P. J., Ritov, Y., Ryden, T., et al. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden markov models. *The Annals of Statistics*, 26(4):1614–1635.

Billingsley, P. (1968). *Convergence of probability measures*. Wiley Series in probability and Mathematical Statistics: Tracts on probability and statistics. Wiley.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308.

Bradley, R. (2007). *Introduction to Strong Mixing Conditions*. Kendrick Press, ISBN 0-9740427-9-X.

Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *n engl j med*, 2007(357):370–379.

Dahlhaus, R., Neumann, M. H., Von Sachs, R., et al. (1999). Nonlinear wavelet estimation of time-varying autoregressive processes. *Bernoulli*, 5(5):873–906.

Friedman, J., Hastie, T., and Tibshirani, R. (2017). *The elements of statistical learning*, volume 12. Springer series in statistics New York.

Goldenberg, A., Zheng, A. X., Fienberg, S. E., Airoldi, E. M., et al. (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233.

Goldsmith-Pinkham, P. and Imbens, G. W. (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31(3):253–264.

Hanneke, S., Fu, W., Xing, E. P., et al. (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4:585–605.

Hanneke, S. and Xing, E. (2007). Discrete temporal models of social networks. *Statistical Network Analysis: Models, Issues, and New Directions Edited by:Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, Anna Goldenberg, Eric P. Xing, Alice X. Zheng*.

Knight, M., Nunes, M., and Nason, G. (2016). Modelling, detrending and decorrelation of network time series. *arXiv preprint arXiv:1603.03221*.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Publishing Company, Incorporated, 1st edition.

Krivitsky, P. N. and Handcock, M. S. (2014). A separable model for dynamic networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):29–46.

Krivitsky, P. N. and Handcock, M. S. (2016). *tergm: Fit, Simulate and Diagnose Models for Network Evolution Based on Exponential-Family Random Graph Models*. The Statnet Project. R package version 3.4.0.

Lütkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*. Springer Berlin Heidelberg.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542.

Nicholls, D. and Quinn, B. (1981). Multiple autoregressive models with random coefficients. *Journal of Multivariate Analysis*, 11(2):185 – 198.

Nicholls, D. F. and Quinn, B. G. (1982). *Random Coefficient Autoregressive Models: An Introduction*, volume 1. Springer Science & Business Media.

Pourahmadi, M. (1986). On stationarity of the solution of a doubly stochastic model. *Journal of Time Series Analysis*, 7(2):123–131.

Pourahmadi, M. (1988). Stationarity of the solution of $x_t = a_t x_{t-1} + \varepsilon_t$ and analysis of non-gaussian dependent random variables. *Journal of Time Series Analysis*, 9(3):225–239.

Tjstheim, D. (1986). Some doubly stochastic time series models. *Journal of Time Series Analysis*, 7(1):51–72.

Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

Wiesel, A., Bibi, O., and Globerson, A. (2013). Time varying autoregressive moving average models for covariance estimation. *IEEE Trans. Signal Processing*, 61(11):2791–2801.

Xu, K. (2015). Stochastic block transition models for dynamic networks. In *Artificial Intelligence and Statistics*, pages 1079–1087.

Zhu, X., Pan, R., Li, G., Liu, Y., Wang, H., et al. (2017). Network vector autoregression. *The Annals of Statistics*, 45(3):1096–1123.