Time Series Models for multivariate binary and categorical data

Inauguraldissertation
zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften der Universität Mannheim

vorgelegt von

M.Sc. Lena Christine Reichmann
aus Waiblingen

Mannheim 2021

Dekan: Dr. Bernd Lübcke, Universität Mannheim
Referent: Prof. Dr. Carsten Jentsch, TU Dortmund
Korreferent: Prof. Dr. Claudia Strauch, Aarhus University
Korreferent: Prof. Dr. Christian H. Weiß, Helmut-Schmidt Universität, Hamburg

Tag der mündlichen Prüfung: 07. Juni 2021

Abstract

In the following thesis, we investigate the modeling of time series data with multivariate discrete and especially binary structure. A model for categorical time series data with a nice interpretability which, in addition, is parsimonious, is the New Discrete AutoRegressive Moving Average (NDARMA) model of Jacobs and Lewis (1983). However, this model only can capture positive autocorrelation as well as positive parameters.

In the first part of the thesis, we propose an extension of the NDARMA model class for the special case of binary data, that allows for negative model parameters, and, hence, autocorrelations leading to the considerably larger and more flexible model class of generalized binary ARMA (gbARMA) processes. For this class of processes, we infer statistical properties and compare it in a simulation study with the benchmark model, the Markov Processes and other time series models.

In the second part, we adopt the approach of the first part and propose a vector-valued extension of gbAR processes, that enable the joint modeling of serial and cross-sectional dependence of multivariate binary data. The resulting class of generalized binary vector Auto-Regressive (gbVAR) models is parsimonious, nicely interpretable and allows also to model negative dependence. We further extend the gbVAR model to include a moving average part, resulting in turn in the gbVARMA model.

In the third and final part we pursue a further extension to vector-valued categorical time series data. For the proposed gbVARMA and NDVARMA models, we provide stationarity conditions and state the stationary solution. Stochastic properties, e.g. Yule-Walker- type equations and classical Yule-Walker equations for the pure autoregressive case are derived. We show φ- and ψ- mixing properties of the gbVAR and NDVARMA model by proving the strict positivity of the transition probabilities. For the NDVARMA model, we discuss the identification of the distribution of the vector-valued innovation process. In simulation studies the performance of the estimators is illustrated.

All three model classes are applied to real data examples.

Zusammenfassung

In der folgenden Arbeit wird die Modellierung von Zeitreihendaten mit multivariater diskreter und insbesondere binärer Struktur untersucht. Ein Modell für kategoriale Zeitreihendaten mit guter Interpretierbarkeit, das zudem noch parameterarm ist, ist das neue diskrete autoregressive gleitendes Mittelwertmodell (NDARMA) von Jacobs und Lewis (1983). Allerdings kann dieses Modell nur positive Autokorrelation sowie positive Parameter erfassen.

Im ersten Teil der Dissertation schlagen wir eine Erweiterung der NDARMA- Modellklasse für den Spezialfall binärer Daten vor, die negative Modellparameter und damit Autokorrelationen zulässt, was zu der wesentlich größeren und flexibleren Modellklasse der generalisierten binären ARMA-Prozesse (gbARMA) führt. Für diese Klasse von Prozessen leiten wir statistische Eigenschaften ab und vergleichen diese in einer Simulationsstudie mit dem Benchmark-Modell, den Markov-Prozessen und anderen Zeitreihenmodellen.

Im zweiten Teil übernehmen wir den Ansatz des ersten Teils und schlagen eine vektorwertige Erweiterung von gbAR-Prozessen vor, die die gemeinsame Modellierung von serieller und Querschnittsabhängigkeit von multivariaten binären Daten ermöglicht. Die resultierende Klasse von verallgemeinerten binären vektoriellen Auto - Regressionsmodellen (gbVAR) ist parameterarm, gut interpretierbar und erlaubt auch die Modellierung negativer Abhängigkeit. Ferner erweitern wir das gbVAR Modell um einen gleitenden Durchschnitt, was wiederum das gbVARMA-Modell ergibt.

Im dritten und letzten Teil verfolgen wir eine zusätzliche Erweiterung auf vektorwertige kategorielle Zeitreihendaten. Für die vorgeschlagenen gbVARMA- und NDVARMA-Modelle stellen wir Stationaritätsbedingungen auf und geben die stationäre Lösung an. Stochastische Eigenschaften, wie beispielsweise Gleichungen vom Yule-Walker-Typ und klassische Yule-Walker-Gleichungen für den reinen autoregressiven Fall werden abgeleitet. Wir zeigen φ- und ψ-Mischeigenschaften des gbVAR- und NDVARMA-Modells, indem die strikte Positivität der Übergangswahrscheinlichkeiten beweisen wird. Für das NDVARMA-Modell diskutieren wir die Identifizierbarkeit der Verteilung des vektorwertigen Innovationsprozesses. In Simulationsstudien werden die Güte der Schätzer gezeigt.

Alle drei Modellklassen werden auf reale Datenbeispiele angewendet.

Acknowledgements

First and foremost, my gratitude goes to my advisor Prof. Dr. Carsten Jentsch, offering me the position and opportunity to prepare this thesis, always believing in me and my abilities. His persistent support, advise and long discussions on my research have crucially improved this work. He gave me the opportunity to be a member of the Chair of Statistics, Department of Economics, at the University of Mannheim and Chair of Business and Social Statistics, Department of Statistic, at the TU Dortmund.

Furthermore, I'm grateful to Prof. Dr. Claudia Strauch for her advice, continuing support and encouragement during the research time. To be a part of her chair at the University of Mannheim allowed me to continue my work.

I would also like to thank Prof. Dr. Christian Weiß for his expertise and discussions on discrete valued time series, inviting me for a research visit at his Chair at the Helmut-Schmidt University, Hamburg.

Additionally, I want to thank Dr. Peter Parczewski for his encouragement and offering me a position, which allowed me to complete the final part of this thesis.

Beside, I have to thank the Baden Württemberg Stiftung, the SFB 884 (Project B6) and the Research Training Group (RTG) 1953 "Statistical Modeling of Complexe Systems and Processes" for their financial support during my PhD, giving me the chance to visit several conferences, workshops and research visit.

During my time at the university, I had the chance to meet several people and colleagues, who made the time to an unforgettable one, with many discussions, support and conversations. Above all, my gratitude goes to Alexander Kreiß for proof reading parts of the thesis, his supporting words and discussions at every time of the thesis. I also want to thank Mehmet Madensoy for his support, encouraging words and joint coffee breaks with helpful discussions. My thanks also go to Cathrine Aeckerle - Willems for all joint lunches, coffee breaks and nice conversations. By working together on our own projects, she supported me during the Corona time with encouraging words and motivation.

Last but not least, I deeply have to thank my husband, Christian Kübler, and my family for their every time support, encouraging me and keeping me grounded on my way.

Contents

1. Introduction

The analysis of time dependent data is a field in statistics that has been widely investigated over the last decades. The collection of data over time arises in diverse areas of science and application, for e.g. in economics the quarterly number of unemployed in a country, in biology and medicine the daily number of virus infected people or in the financial sector the daily closing prices of a stock, to name just a few. In all the previous examples, the data depend on each other over the time, which is the main challenge of time series analysis as the classical approaches for independent or at least uncorrelated data are not applicable. The main task is to establish models for describing the relationship between the observations, and to impose conditions that are strong enough to derive insightful results, but which are also as general as possible, so as to be able to apply such a model on real data.

In the last few years, however, categorical time series has received renewed attention, whereas in most text books, e.g. Brockwell and Davis (1991), Wei (1989) or Lütkepohl (2005), the models concentrate on continuous - valued time series data. Yet, categorical data form a special type of data, to which the common approaches of continuous data cannot be applied. A categorical variable represents a state, e.g. in medicine categories occur as nucleobases such as the adenine (A), cytosine (C), guanine (G) and thymine (T) bases as components of a DNA row. Recession, slowdown or recovery of the market cycle are three states that are examples of categorical structure which appear in an economic analysis or even, for instance, a yes/no expression is contained in the class of categorical data. Transferring these examples to the natural numbers $\mathbb{N}_0$ for its statistical analysis, the categories already have a nominal structure, meaning that no further information can be derived from the value of the category. Whereas for example the distance between two real valued data points transfers additional information about the position to each other, the distance between two categorical variables has no meaning. Due to this special structure of non ordinal ranking within the data, categorical time series can not use the common models for continuous valued data. The additional reason is, that these models do not ensure the time series stay within the permissible categorical values. However, for inference on such data, the tools and approaches of the continuous valued time series models are very helpful also in the categorical case.

To overcome the issue, Jacobs and Lewis (1978a) introduced a model to analyze such data, where the outcome is again assured to be categorical. In Kedem (1980), Davis et al. (2016) and Weiß (2009a) additional models are introduced, which we will take a closer look at and discuss later.

In addition, the development of computers and the internet has facilitated a great amount of data collection. So it is not only the time dependence occurring in the data that we might consider, but also that the observations are multidimensional and it is interesting to investigate the dependence between these dimensions. Furthermore, multivariate methods are needed for a proper analysis of the data.

1. Introduction

In this thesis, a new model for the special case of only two categories, and thus binary data, is investigated. We introduce the model for both univariate and multivariate data, as nowadays the data sets typically have more than one dimensions. With the extension of the models for multivariate time series data, we have to consider not just that the methods need to be adapted to more than one dimensions, but also that multiple dimensions present different problems to those that we face with with univariate models. Nevertheless, we would like to shed more light on the categorical data first.

Categorical and especially binary time series data appear as univariate and multivariate versions in many fields of application. Davis et al. (2016) mostly analyze univariate categorical and binary time series data, as some examples being the time series of recession states of the US, the three state time series of sleep states of the AgeWise study from the University of Pittsburgh and the filtered mortality data of Los Angeles to a binary time series.

In addition, as discussed by Kedem (1980), a binarized real valued time series data still keeps the rhythmic and random mechanism of the data. Therefore, it may also be interesting to look at the binarized version of a continuous valued time series, which also contains the main information but is cheaper to store. This procedure extends the application of binary time series to many more data sets, which initially do not have a binary structure. In Chapter 2, we examine the binarized eruption duration of the Old Faithful Geyser, as it is only of interest if a certain period is exceeded and thus whether or not a long eruption occurs. Thereby the investigation is not aimed at the eruption duration itself. For further discussion on this example, see Section 2.1.

Over last few years, the amount of collected data has been growing due to the availability of cheap storage. Interest is increasingly focused on the investigation of the dependence structure over the dimensions within the data sets. In time series analysis for example, the aim is to investigate the relationship of the dimensions on each other, therefore the cross sectional dependence structure. One example is the spread of recession states over e.g. the G7 countries. The examination of the countries together can shed light on the dependency between them and in particular their economies.

Another example of multivariate data are the binarized air pollution data of measuring stations in a city or metropolitan area to examine if they exceed a threshold or not. A detailed discussion is given in Section 3.1.

For the data described so far, variables $X_{t,k}$ are monitored, where $k = 1, \dots, K$ and all variables are taken according to a time point t. In practice, the observations are mostly observed at discrete time t and therefore we assume in this thesis $t \in \mathbb{Z}$. The variables are generated by a stochastic process $(X_{t,k}, t \in \mathbb{Z})$ and are therefore random variables for each k and t. A more convenient notation is to compose all variables to a vector-valued process $X_t := [X_{t,1}, \dots, X_{t,k}]'$, where $X_{t,k}$ is then the k-th component variable and K is the dimension of the time series. In the following, we assume the process $(X_{t,k}, t \in \mathbb{Z})$ to take values in a categorical state space $\mathcal{V} = \{0, 1, \dots, m\}$ with $m \in \mathbb{N}$ and thus the K - dimensional process $(X_t, t \in \mathbb{Z})$ takes values in the state space $\mathcal{V}^K = \{0, 1, \dots, m\}^K$.

When directly dealing with these data, we face several issues. Whereas for continuous data a large number of distributions are available, there are a lower number of distributions available for categorical data. Furthermore, the usual mathematical operations are not applicable. This can be directly illustrated by the example of nucleobases given previously. Adding the base adenine A and cytosine C, e.g. $A + C$,

we may ask what the meaning of this operation is. The issues of the meaning and informative value still remain in the context of time series analysis. For example, whilst the definitions of trend and seasonality are clear for continuous valued time series, it is not obvious how this terminology is defined for categorical valued time series. The main difficulty for categorical time series comes from the nominal structure of the data, such that the commonly used dependency measures, as the autocovariance or autocorrelation, have no meaning. For further discussion see e.g. Weiß (2009a).

Nevertheless, the special case of only two categories allows us to adopt some parts of the commonly used theory. Thus, in the binary case, the addressed measures can still provide information about existing dependencies within the data, as we only have two possibilities, as e.g. $\mathcal{V} = \{0, 1\}$. In Chapter 2 and 3 we examine the special case of only two categories, where the analysis is based on the commonly used autocorrelation.

In addition to the above examples of categorical time series, over the last few years a new class of data objects has been gaining increasing interest in theory and application. The analysis of network data has been rising as these data are available in various fields, e.g. computer networks, biological networks or social networks. Typically, networks are represented by two sets, the vertex set and the edge set. The first set consists of the vertices or nodes corresponding to the objects, e.g. computers, people or members of a social network. The second set contains the edges corresponding to the connections between the objects in the vertex set, transferring the information on whether or not two nodes are in contact to each other. An edge can contain information e.g. if two computers have a direct connection via the internet, two people had contact and the virus was spread or two members are friends. More examples from various disciplines can be found in Kolaczyk (2009), Newman (2010) and Jackson (2008).

The notation of networks make use of the graph theory, where a graph $G = (V, E)$ with V is the set of vertices $\{v_1, \ldots, v_K\}$ and the interactions between these vertices, called the set of edges $E = \{e_1, \ldots, e_n\}$ with $e_l = \{v_i, v_j\}$ for $l = 1, \ldots, n$ and $i, j \in \{1, \ldots, K\}$. Another representation of a graph G, containing the same information of the presence or absence of an edge between two nodes is the so called *adjacency matrix* A. It is of dimension $K \times K$ with $K = |V|$ the number of nodes and each row and column belongs to one vertex v_i, $i = 1, \ldots, K$. Then, it is

$$A = (a_{ij})_{i,j=1,\ldots,K}$$

where $a_{ij} \in \{0, 1\}$, such that $a_{ij} = 1$ means that there is a connection between the vertices v_i and v_j. Whenever $a_{ij} = 0$, the edge between v_i and v_j is not contained in the edge set E. The resulting network representation is of binary nature and therefore also a part of the categorical data, which we want to assume in this thesis.

Let's assume that we are observing a network of fixed dimension K not just at one time point t, but at several time points $t = 1, \ldots, T$, so that we face a sequence of adjacency matrices $A_1, \ldots, A_T$. The main interest is to investigate the dependence of the absence and presence of edges evolving over time. Examples of evolving networks of fixed number of vertices are friendship networks, where each adjacency matrix consists the information on whether two friends in the network were in contact with each other at that time or not.

Bai et al. (2019), for example, considered the dynamic face-to-face interaction network between participants playing the Resistance game. An edge between the nodes v and u is detected, whenever a participant v looks at the other participant u at time t. The interaction is detected at each 1/3 second. Whereas Paranjape et al. (2016)

analyzed the email - communication of a big European research institute, where an edge is detected if an email is sent from person v to person u at time t. This data set is observed over 803 days, such that a sequence of adjacency matrices is given over nearly two years.

1.1. Models for categorical time series

The widely investigated time series models for continuous data often do not adequately specify the categorical type of the data. However, there are many different models that can describe the special data structure.

1.1.1. Markov Processes

One prominent and well known model with serial dependence are Markov models. They form a powerful model class, having the characteristic of fixed-dependence memory and therefore the current past has an impact on the new outcome. This property means, that the future probabilistic evolution of the process only depends on the *immediate* past of the process and can be summarized in the following Definition.

Definition 1.1 (Markov Process)
A categorical process $(X_t, t \in \mathbb{N}_0)$ *with range* $\mathcal{V}$ *is said to be a* p^{th}***- order Markov Process*** *with* $p \in \mathbb{N}$*, if*

$$P\left(X_t = s_t | X_{t-1} = s_{t-1}, \dots X_0 = s_0\right) = P\left(X_t = s_t | X_{t-1} = s_{t-1}, \dots X_{t-p} = s_{t-p}\right)$$

for all $t \geq p$ *and all* $s_i \in \mathcal{V}$ *with* $i \leq t$.
If the ***transition probabilities***

$$P\left(X_t = s_t | X_{t-1} = s_{t-1}, \dots X_{t-p} = s_{t-p}\right) = p_{s_0|i_{-1},\dots s_{-p}}$$

are time invariant, then the process is said to be a ***homogeneous*** p^{th}*- order Markov Process.*

The whole dependence structure of a Markov Process can be described, when all transition probabilities from all possible states to the others in the state space $\mathcal{V}$ are composed to a transition probability matrix P.

The model class of Markov processes has been extensively studied in classical literature, where a variety of different properties are investigated for such models. Statistical Theory on Markov processes are discussed in e.g. Billingsley (1961), Bartlett (1951) and Seneta (1981).

An additional advantage of the model class is the possibility to model a wide range of dependence structure within the data, which in turn makes the model attractive for categorical and in particular binary time series data.

The next definition is based on Weiß (2009a)[Definition 9.1.1.2]. It covers some important properties of Markov processes for the following thesis. Further discussions on additional properties of finite state space Markov processes are given in Weiß (2009a).

Definition 1.2 (Properties of a Markov Process)
Let $(X_t, t \in \mathbb{N}_0)$ *be a homogeneous Markov Process of order* $p = 1$ *with range* $\mathcal{V}$ *and transition matrix* P*. Then* $(X_t, t \in \mathbb{N}_0)$ *and* P *are said to be*

i) ***irreducible*** *if for any $i, j \in \mathcal{V}$ there exists some $n \in \mathbb{N}$ such that $p_{i|j}(n) > 0$.*

ii) ***periodic*** *of period d if it is irreducible, and if for an $i \in \mathcal{V}$ (and therefore for all $i \in \mathcal{V}$) the greatest common divisor of those $n \in \mathbb{N}$, for which $p_{i|i}(n) > 0$, is equal to d.*

iii) ***primitive*** *if there exists an $n \in \mathbb{N}$ such that all $p_{i|j}(n) > 0$.*

iv) ***ergodic*** *if it is irreducible with period 1.*

With these properties of Markov process, its asymptotic behavior can be stated. Thus the property of primitivity of a Markov process indicates that it has a unique stationary distribution. Besides, a Markov process is ergodic if and only if it is primitive, see e.g. Seneta (1981)[Theorem 1.4]. From the long term behavior of an ergodic Markov process, additional meaningful properties can be derived. So it follows for the marginal distribution of an ergodic Markov process that it converges in addition to its stationary marginal distribution, whereas this stationary distribution is independent of the initial distribution of the process. Further essential conclusions from the properties mentioned above are summarized by Weiß (2009a)[Tables 5 and 6, Chapter 9] and will be applied later in the thesis.

With the assumption that the range is finite countable and a stationary transition distribution, a p^{th}- order Markov process can be fitted to the univariate process by estimating the necessary $(m+1)^p \cdot m$ transition probabilities. Despite the advantages, the number of parameters may pose a challenge and leads to an over-parametrization, see e.g. McKenzie (2003). Therefore, the associated issue of identifiability of the process parameters has to be considered and the issue potentizes when the considered process is K- dimensional. This curse of dimensionality, first introduced by Bellman (2016), indicates an exponential growth of the parameter with the dimension, which is also the case when a Markov process is adapted to a K-dimensional process.

Two models, addressing the drawback of too many parameters of Markov processes, are considered in Bühlmann and Wyner (1999) the *variable length Markov model* and in Raftery (1985) the *mixture transition distribution model*. Both reduce the parameters of the process, though the variable length Markov model has the challenge of the model choice and are further discussed in Weiß (2009a)[Chapter 9.2]. The second model, the mixture transition distribution model, is an extension of the standard homogeneous Markov process, which was introduced to approximate a high-order Markov process with far fewer parameters. The model suffers from $m(m-1)+p-1$ parameter and only one parameter for one additional lag is added, see also Berchtold and Raftery (2002). Even if this is similar to an autoregressive (AR) model, the modeling is based on a transition probability matrix, which dimension would increase in size for multivariate data.

1.1.2. Discrete autoregressive moving average model

To overcome the problem of an over-parametrization of Markov models and also to investigate the non-Markovian structure of data, Jacobs and Lewis (1978a), Jacobs and Lewis (1978b), and Jacobs and Lewis (1978c) introduced, in a series of papers, a mixed AutoRegressive Moving Average (ARMA) model for discrete data, defined as DARMA process. The process was constructed to model non-Markovian structure with a specific marginal distribution and correlation structure.

Definition 1.3 (Discrete ARMA Model)
Let $(X_t, t \in \mathbb{Z})$, $(Z_t, t \in \mathbb{Z})$ *and* $(e_t, t \in \mathbb{Z})$ *be categorical processes with range* $\mathcal{V} = \{0, 1, \ldots, m\}$. *Let* $(e_t, t \in \mathbb{Z})$ *be an independent and identically distributed (i.i.d.) discrete valued process with marginal distribution* $P(e_t = j) = p_j > 0$ *for* $j \in \{0, 1, \ldots, m\}$. *The innovation process* e_t *is assumed to be independent of the processes* X_s *and* Z_s *for* $s < t$.
Let $\mathcal{P} := [\alpha^{(q+1)}, \beta^{(0)}, \ldots, \beta^{(q)}]$ *and* $\widetilde{\mathcal{P}} := [\widetilde{\alpha}^{(1)}, \ldots, \widetilde{\alpha}^{(p)}, \widetilde{\beta}^{(0)}]$ *be the parameter vectors with* $\mathcal{P}\mathbb{1}_{q+2} = 1$ *and* $\widetilde{\mathcal{P}}\mathbb{1}_{p+1} = 1$ *where* $\mathbb{1}_l$ *is the one vector of length* l. *And further let*

$$P_t := (a_t^{(q+1)}, b_t^{(0)}, \ldots, b_t^{(q)}) \sim Mult(1; \mathcal{P})$$
$$\widetilde{P}_t := (\tilde{a}_t^{(1)}, \ldots, \tilde{a}_t^{(p)}, \tilde{b}_t^{(0)}) \sim Mult(1; \widetilde{\mathcal{P}}) \quad , t \in \mathbb{Z}$$

be i.i.d. random vectors, which are independent of (X_s), (Z_s) *for* $s < t$ *and* $(e_t, t \in \mathbb{Z})$. *Then, the process* $(Z_t, t \in \mathbb{Z})$ *is called a* ***discrete AR (DAR) process*** *of order* $p \in \mathbb{N}_0$, *if it follows the recursion*

$$Z_t = \tilde{a}_t^{(1)} Z_{t-1} + \ldots + \tilde{a}_t^{(p)} Z_{t-p} + \tilde{b}_t^{(0)} e_t.$$

The process $(X_t, t \in \mathbb{Z})$ *is called a* ***discrete ARMA (DARMA) process*** *of order* $p, q + 1\mathbb{N}$, *if it follows the recursion*

$$X_t = a_t^{(q+1)} Z_{t-q-1} + b_t^{(0)} e_t + \ldots + b_t^{(q)} e_{t-q}.$$

In contrast to the Markov models described above, the DARMA model recursion is similar to common ARMA recursion and thus its interpretation is possible directly via the model parameters. The benefit of the multinomial selection with model parameters $\mathcal{P}$ and $\widetilde{\mathcal{P}}$, is that the process always takes values within the discrete state space $\mathcal{V}$, because only one entry of the random vector P_t and $\widetilde{P}_t$ is equal to one, whereas all other entries are zero. Consequently, the multinomial selection implies that the model parameters are probabilities and take values in the unit interval. Nevertheless, it is not intuitive why the DARMA(p,q+1) model indirectly depends on the autoregression and consequently only on the predecessors of the process Z_s with $t - s > q$.

To avoid such a complex recursion definition, Jacobs and Lewis (1983) refine the DARMA model to a *new discrete ARMA model*, where the process follows only one recursion.

Definition 1.4 (New Discrete ARMA Model)
Let $(X_t, t \in \mathbb{Z})$ *be a stationary process taking values in* $\mathcal{V}$. *Let* $(e_t, t \in \mathbb{Z})$ *be an i.i.d. discrete valued innovation process, such that* e_t *is independent of* $(X_s)_{s<t}$ *with mean* $\mu_e = E(e_t)$ *and variance* $\sigma_e^2 > 0$. *Let* $\mathcal{P} := \left[\alpha^{(1)}, \ldots, \alpha^{(p)}, \beta^{(0)}, \ldots, \beta^{(q)}\right]$ *be the parameter vector such that* $\mathcal{P}\mathbb{1}_{p+q+1} = 1$ *where* $\mathbb{1}_{p+q+1}$ *the one vector of length* $p+q+1$. *Further, let*

$$P_t := \left(a_t^{(1)}, \ldots, a_t^{(p)}, b_t^{(0)}, \ldots, b_t^{(q)}\right) \sim Mult\left(1; \mathcal{P}\right), \quad t \in \mathbb{Z},$$

be i.i.d. random vectors, which are independent of $(e_t, t \in \mathbb{Z})$ *and* $(X_s, s < t)$. *Then, the process* $(X_t, t \in \mathbb{Z})$ *is said to be a* ***New Discrete ARMA (NDARMA) process*** *of order* $p, q \in \mathbb{N}_0$, *if it follows the recursion*

$$X_t = \sum_{i=1}^{p} a_t^{(i)} X_{t-i} + \sum_{j=0}^{q} b_t^{(j)} e_{t-j}.$$

In the case of $q = 0$, the process is said to be a NDAR(p) process and for $p = 0$ it is called a NDMA(q) process.

This model has a nicely interpretable autoregressive moving average structure without any indirect modeling. The multinomial selection of the random vector P_t ensures that the process always takes values within the state space $\mathcal{V}$. This is where the essential difference to common ARMA models for continuous data comes into account, because the model parameters $\mathcal{P}$ are now probabilities which are inserted into the selection mechanism.

Consequently, the time series X_t is a random mixture of the past observations or innovation processes. Therefore, the new value of the process takes either a predecessor value $X_{t-1}, \ldots, X_{t-p}$ with probability $\sum_{i=1}^{p} \alpha^{(i)}$ or an innovation term $e_t, \ldots, e_{t-q}$ with probability $1 - \sum_{i=1}^{p} \alpha^{(i)}$. The random vector P_t selects for every $t \in \mathbb{Z}$ the new value of the categorical process.

Jacobs and Lewis (1983) derived a few first properties of the model as e.g. Yule - Walker equations and autocorrelation structure. Due to its nice structure, Weiß and Göb (2008), Weiß (2009a), Weiß (2011a), Weiß (2011b) and Weiß (2013) extend the analysis of the model to further stochastic properties and dependency measures, e.g. signed and unsigned serial dependence measures as Cohen's κ and Cramer's ν. Since the model parameters can take only positive values, it is possible to capture exclusively positive dependence structure. Another characteristic of NDARMA models are the generated long runs of the same value. This results from the random selection mechanism, as it either chooses a predecessor or innovation with a certain probability. However, very few application data sets show this property of repeating the same category over a period of time and thus these properties restrict the application of the model.

Despite the limitation that only positive autocorrelations can be modeled, the NDARMA model has several advantages, which makes this model extendable to multivariate data. Especially the number of model parameter are parsimonious and therefore suitable for an extension.

Möller and Weiß (2020) used the idea of the NDARMA model, to expand it for multivariate categorical time series data but also added a data-specific variation operator to obtain a great flexibility and prevent long runs of the same value over time. Here, the variation operator is assumed to be mean preserving and also maintain the ARMA-like autocorrelation structure but whilst allowing more variation within the sample paths.

Definition 1.5 (Generalized Discrete ARMA model)
Let $(X_t, t \in \mathbb{Z})$ be a stationary K - dimensional process taking values in $\mathcal{V}^K = \{0, 1, \ldots, m\}^K$. Let $(e_t, t \in \mathbb{Z})$ be an i.i.d. discrete valued K - dimensional innovation process taking values in $\mathcal{V}^K$, such that e_t is independent of $(X_s)_{s<t}$ with mean $\mu_e = E(e_t)$ and variance $\Sigma_e^2 > 0$. Let $\mathcal{P} := \left[\alpha^{(1)}, \ldots, \alpha^{(p)}, \beta^{(0)}, \ldots, \beta^{(q)}\right]$ be the parameter vector such that $\mathcal{P}\mathbb{1}_{p+q+1} = 1$ with $\mathbb{1}_{p+q+1}$ the one vector of length $p + q + 1$. Further, let

$$P_t := \left(a_t^{(1)}, \ldots, a_t^{(p)}, b_t^{(0)}, \ldots, b_t^{(q)}\right) \sim Mult\left(1; \mathcal{P}\right), \quad t \in \mathbb{Z},$$

be i.i.d. random vectors, which are independent of $(e_t, t \in \mathbb{Z})$ and $(X_s, s < t)$.

*Then, the process $(X_t, t \in \mathbb{Z})$ is said to be a **generalized discrete ARMA (GDARMA) process** of order $p, q \in \mathbb{N}_0$ with variation operators $\mathcal{S}_{\cdot,\cdot} : \mathcal{V} \to \mathcal{V}$, if it follows the*

recursion

$$X_t = \sum_{i=1}^{p} a_t^{(i)} \mathcal{S}_{t,i}(X_{t-i}) + b_t^{(0)} e_t + \sum_{j=1}^{q} b_t^{(j)} \mathcal{S}_{t,j}(e_{t-j}).$$

The model is still parsimonious like the NDARMA model, even though multivariate data can be handled. Instead of taking scalar process variables X_t into the model equation, the idea of the NDARMA model equation is applied on a K- dimensional process X_t, therefore selecting a predecessor vector $X_{t-1}, \ldots, X_{t-p}$ or an K-dimensional innovation $e_t, \ldots, e_{t-q}$. So the model suffers again only from $p+q+1$ parameters to specify the multinomial selection.

Besides, the operator $\mathcal{S}$ randomly varies the inserted processes to reach a more flexible model. It is assumed to be mean preserving, such that the ARMA - like autocorrelation structure is maintained. This means, that the variation operator satisfies $E[\mathcal{S}(X)|X] = X$ and hence $E[\mathcal{S}(X)] = E[X]$. However, even Möller and Weiß (2020) discussed several operators, the functions are still random within the state space $\mathcal{V}^K$ and a systematic modeling of cross-sectional dependence is therefore not possible. Nevertheless, it is conceivable that in data setups after an occurring category $m_1 \in \mathcal{V}$ a systematic change to another category $m_2 \in \mathcal{V}\backslash\{m_1\}$ arises. In the NDARMA - like models, it is only possible to get a change of the category when 1) an innovation takes the value m_2 and is in addition chosen by the selection mechanism or 2) a predecessor is equal to the value m_2 and is selected. But this change is still random and not systematically driven by the model directly.

In addition, both the GDARMA and NDARMA models have the disadvantage, that the process can only handle non-negative autocorrelation structure as the model parameters have to take values in the unit interval. These two properties restrict the application of the NDARMA model to univariate categorical data and the GDARMA model multivariate categorical data, as the estimation of the model parameter and autocorrelation in most cases show invalid values, e.g. non negative parameters although negative autocorrelation can also occur by fitting the NDARMA - type model. Besides, as mentioned by Möller and Weiß (2020) for K- dimensional time series data, the multivariate extension of the GDARMA model selects the observations and innovations vector-wise, but an individual selection of just one entry of a certain observation or innovation vector should be also mentioned to allow for cross - sectional dependence.

In the following chapters, we want to address these two issues and investigate two extensions of the NDARMA model. The first question is how the model can be extended to allow also potential negative parameters and autocorrelation too, where the open question is how negative autocorrelation should be interpreted in the case of categorical data. In the special case of only two categories, such that $\mathcal{V} = \{0, 1\}$, negative autocorrelation is interpretable. The negative dependence indicates a change of the category from one step to the next and so it is directly clear to take the opposite value of the state space. However, this concept is not straightforward to transfer to a state space with more than two categories, because it is not clear which of the remaining m categories should be chosen as the opposite value. The first part of the thesis considers a new model for a binary state space $\mathcal{V} = \{0, 1\}$ based on the NDARMA idea, by now allowing for negative parameters indicating a negative dependence.

The models introduced previously are defined for univariate time series data. Only the GDARMA model suggests a way to handle multivariate data but without including the cross-sectional dependency. The extension to multivariate data is introduced in the

second and third part of the thesis (Chapter 3 and Chapter 4), where we generalize firstly the new introduced model of Chapter 2 and then the NDARMA model for multivariate data.

1.1.3. Binary time series model

For negative dependence structure within binary sequences, Kanter (1975) introduced the binary AR model and Weiß (2009b) added a moving average part, resulting in the binary ARMA model. Instead of using an additive operator within the ARMA model equation, a modulo 2 ($\oplus$) operation is applied to ensure a binary outcome each time.

Definition 1.6 (Binary ARMA model)
Let $(X_t, t \in \mathbb{Z})$ be a stationary binary process taking values in $\mathcal{V} = \{0, 1\}$. Let $(e_t, t \in \mathbb{Z})$ be an i.i.d. binary valued innovation process, such that e_t is independent of $(X_s)_{s<t}$ with mean $\mu_e = E(e_t)$ and variance $\sigma_e^2 > 0$. Let $\mathcal{P} := \left[\alpha^{(1)}, \ldots, \alpha^{(p)}, \beta^{(0)}, \ldots, \beta^{(q)}\right]$ be the parameter vector such that $\mathcal{P}\mathbb{1}_{p+q+1} = 1$ where $\mathbb{1}_{p+q+1}$ the one vector of length $p + q + 1$. Further, let

$$P_t := \left(a_t^{(1)}, \ldots, a_t^{(p)}, b_t^{(0)}, \ldots, b_t^{(q)}\right) \sim Mult\left(1; \mathcal{P}\right), \quad t \in \mathbb{Z},$$

*be i.i.d. random vectors, which are independent of $(e_t, t \in \mathbb{Z})$ and $(X_t, s < t)$. Then, the process $(X_t, t \in \mathbb{Z})$ is said to be a **Binary ARMA (BinARMA) process** of order $p, q \in \mathbb{N}_0$, if it follows the recursion*

$$X_t = \oplus_{i=1}^{p} a_t^{(i)} X_{t-i} \oplus e_t \oplus_{j=0}^{q} b_t^{(j)} e_{t-j}.$$

In the case of $q = 0$, the process is said to be a BinAR(p) process and for $p = 0$ it is called a BinMA(q) process.

The random vector P_t is defined as in the previous NDARMA model, nevertheless the process takes the value e_t with probability β_0 thus all entries within P_t are zero except $b_t^{(0)}$. The modulo operation comes into account, when either a predecessor $X_{t-1}, \ldots, X_{t-p}$ or an innovation term $e_{t-1}, \ldots, e_{t-q}$ is chosen by the random vector P_t and thus X_t takes either the value $X_{t-i} \oplus e_t$ for $i = 1, \ldots, p$ or $e_t \oplus e_{t-j}$ for $j = 1, \ldots, q$. This allows a direct modeling of negative autocorrelation within the data by a well known operation. However, the model recursion is not very intuitive to interpret and it might not even be clear how the process behaves. Nonetheless, the idea of using a modulo 2 operation enables now to model a small amount of the desired negative autocorrelation structure, as shown in Weiß (2009b). The range of possible negative autocorrelation, which can be modeled by a BinAR(2) process is also discussed in 2.3.1 together with the models mentioned above. However, we will take up the idea of using a modulo 2 operation for systematical modeling of negative dependence structure.

Especially for the modulo 2 operation, a combination of the process variable X_t with a value of one results in a more common and direct way of

$$(X_t \oplus 1) = (1 - X_t) \tag{1.1}$$

for $X_t \in \{0, 1\}$. Mostly, the right hand side of equation (1.1) is more understandable and leads in a direct representation of the systematic switching of a binary variable to its opposite state. This approach of the modulo operation (1.1) can be used for modeling negative dependence structure and thus a systematic change of a binary variable. The helpful simplification of the modulo operator in the binary case is used in Chapter 2 to extend the NDARMA model to a negative dependency structure.

1.2. Outlook

The thesis is organized as follows: In Chapter 2, we introduce the generalized binary ARMA model (gbARMA) starting with a discussion on possibly negative parameters and, hence, autocorrelations leading to the considerably larger and more flexible model class. We provide stationary conditions, give the stationary solution and derive stochastic properties of gbARMA processes as its transition probabilities, marginal distribution and common probability. For the purely autoregressive case, classical Yule - Walker equations hold that facilitate parameter estimation of gbAR models. Yule - Walker - type equations are also derived for gbARMA processes. In a simulation study, we compare several models by simulating a second order binary process to show the different ranges of manageable dependence structure.

In Chapter 3 we extend the gbARMA model that is presented in Chapter 2 to a vector-valued time series model, called the gbVARMA model. First, we discuss the pure autoregressive case in Chapter 3.2, that enables the joint modeling of serial and cross-sectional dependence of multivariate binary data. We provide stationarity conditions, state moving-average-type representations and give general stochastic properties of gbVAR processes, including formulas for transition probabilities and mixing properties. In particular, classical Yule-Walker equations hold that facilitate parameter estimation in gbVAR models. In simulations, we investigate the estimation performance and propose a parametric bootstrap procedure for statistical inference. For illustration, we apply gbVAR models to particulate matter (PM_{10}, 'fine dust') alarm data observed at six monitoring stations in Stuttgart, Germany.

We add a moving average part to the gbVAR model and discuss the extension of the resulting gbVARMA processes in Chapter 3.6. We extend the stochastic properties of the gbVAR process to the additional moving average part.

In Chapter 4, the NDARMA model is extended to vector-valued time series data taking the cross- sectional dependence into account as well. For the resulting NDVARMA model, we derive stationarity conditions and state the stationary solution. Stochastic properties are provided, e.g. Yule - Walker type equations and for the pure autoregressive case, classical Yule - Walker equations for the parameter estimation. We show φ - and ψ - mixing conditions of the NDVARMA model and discuss the distribution identification of the vector-valued innovation process. Finally, the business cycle clock data is applied to the model for illustration.

2. Generalized Binary Time Series Models - the univariate Model

Based on: Jentsch, C. and Reichmann, L.
Generalized Binary Time Series Models *- Econometrics 2019, 7 (4), 47*

2.1. Introduction

Categorical time series data are collected in many fields of applications and the statistical research focusing on such data structures evolved considerably over the last years. As an important special case, binary time series that correspond to categorical data with two categories, occur in many different contexts. Often, binary time series are obtained from binarization of observed real-valued data. Such processes are considered, e.g., in Kedem and Fokianos (2002). In Figure 2.1, we show three real data examples of binary time series from different fields of research. For example, in Figure 2.1a, the eruption duration of the Old Faithful Geyser in the Yellowstone National Park is binarized using a threshold. It is coded with a value of one if an eruption lasts for longer than three minutes and zero if it is shorter. In economics, the two states of recessions and economic growth are of interest, as discussed, e.g., in Bellégo and Ferrara (2009). One example of a recession/no-recession time series is shown in Figure 2.1b, where for every quarter it is shown if Italy is in a recession, indicated by zero, or not, indicated by one. Recently, there is great interest in the air pollution in European cities, where an exceedance of the threshold of 50 $\mu g/m^3$ PM_{10} (fine dust) causes a fine dust alarm. The resulting sequence of states of no exceedance corresponding to zero and exceedance corresponding to one is shown in Figure 2.1c. Further examples can be found, e.g., in geography, where sequences with the two states of dry and wet days are considered, e.g., in Buishand (1978). In biomedical studies, binary time series occur in the case, where the participants keep daily diaries of their disease. For example, in clinical trials, as in Fitzmaurice and Lipsitz (1995), the binary self assessment of participants of their arthritis is collected, where poor is indicated by zero and good by one. In natural language processing, the occurrence of vowels as a sequence can be of interest, as considered in Weiß (2009b), where a text is binarized by detecting a consonant or no consonant/vowel as the two states. The binarization of a time series by a threshold, as, e.g., in the PM_{10} example, or by categorizing the time series into two states, as, e.g., in dry and wet days, indeed simplifies the real valued time series to a binary version. As mentioned in Kedem (1980), nevertheless, the transformation keeps the random mechanism from which the data are generated. For the example of PM_{10} data, it might often be of more interest, whether a certain threshold is crossed (or not) instead of the actual amount. In general, the rhythm within the binarized time series contains a great amount of information of the original data.

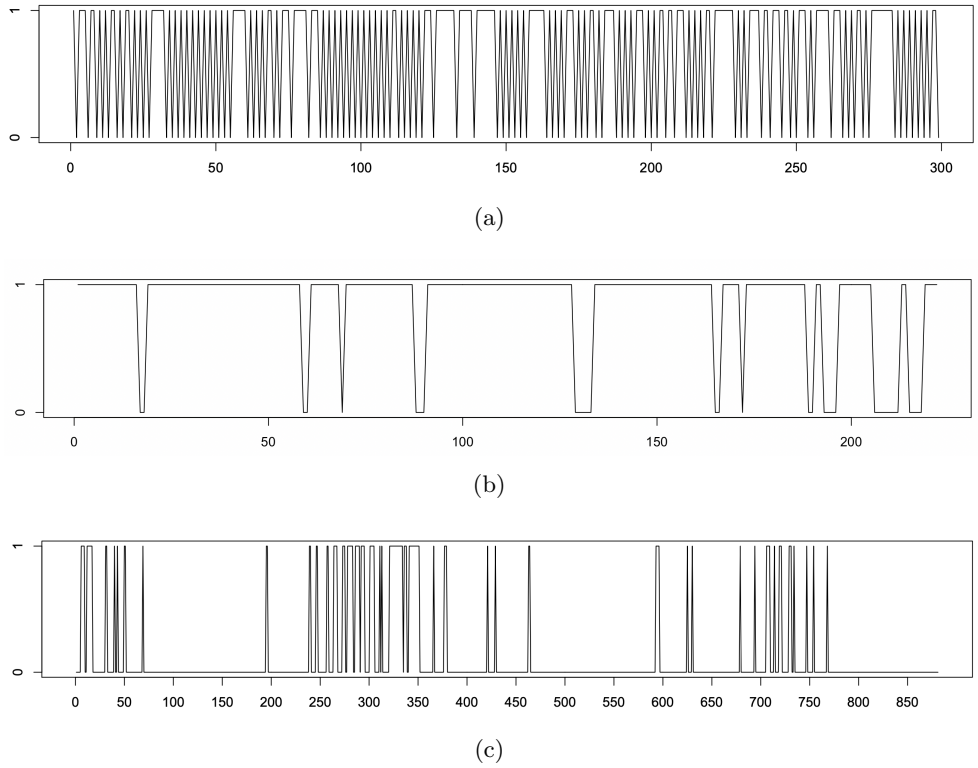


Figure 2.1.: Three real data examples of binary time series: (**a**) binarized eruption duration of the Old Faithful Geyser over 299 eruptions; (**b**) quarterly detected binarized recession/no-recession time series of Italy from Quarter 1 in 1960 to Quarter 1 in 2017 (229 time points); and (**c**) binarized fine dust (PM_{10}) data from Stuttgart, Germany recorded daily from 3 March 2016 to 31 July 2018 over 881 days.

As discussed in Kedem (1980), binary Markov chains are typically used for modeling the dependence structure due to their great flexibility. However, the number of parameters to estimate from the data grows exponentially with the order of the Markov model leading to over-parametrization (see, e.g. McKenzie (2003)).

To avoid the estimation of a large number of parameters, Jacobs and Lewis (1983) proposed the class of (new) discrete autoregressive moving-average (NDARMA) models for categorical time series. More precisely, for processes with discrete and finite state space, a parsimonious model is suggested. The idea is to choose the current value for X_t randomly either from the past values of the time series $X_{t-1}, \ldots, X_{t-p}$ or from one of the innovations $e_t, e_{t-1}, \ldots, e_{t-q}$ with certain probabilities, respectively. This random selection mechanism is described by independent and identically distributed (i.i.d) random vectors $(P_t, t \in \mathbb{Z})$ with

$$P_t := \left[a_t^{(1)}, \ldots, a_t^{(p)}, b_t^{(0)}, \ldots, b_t^{(q)}\right] \sim Mult\,(1; \mathcal{P})\,,$$

where $Mult\,(1; \mathcal{P})$ denotes the multinomial distribution with parameter 1 and probability vector $\mathcal{P} := \left[\alpha^{(1)}, \ldots, \alpha^{(p)}, \beta^{(0)}, \ldots, \beta^{(q)}\right]$ with $\alpha^{(1)}, \ldots, \alpha^{(p)} \in [0, 1)$, $\beta^{(0)} \in (0, 1]$ and $\beta^{(1)}, \ldots, \beta^{(q)} \in [0, 1)$ such that $\sum_{i=1}^{p} \alpha^{(i)} + \sum_{j=0}^{q} \beta^{(j)} = 1$. Then, the NDARMA(p,q) model equation is given by

$$X_t = \sum_{i=1}^{p} a_t^{(i)} X_{t-i} + \sum_{j=0}^{q} b_t^{(j)} e_{t-j}, \quad t \in \mathbb{Z}, \tag{2.1}$$

where $(e_t, t \in \mathbb{Z})$ is an i.i.d. process taking values in a discrete and finite state space $\mathcal{V}$. Since for each time point t only one entry in the random vector P_t is realized to be one while all others become zero, the value of X_t takes either one of the values of X_s for $s \in \{t-1, \ldots, t-p\}$ or one of the error terms e_s for $s \in \{t, \ldots, t-q\}$. This sampling mechanism assures that the time series takes values in the state space $\mathcal{V}$, such that, e.g., for a binary time series with $\mathcal{V} = \{0, 1\}$, the process stays binary. In contrast to the real-valued ARMA model, the lagged time series values and errors are not weighted according to the model coefficients and summed-up since only one of them is actually multiplied with one and all the others with zero based on the realization of P_t.

The model parameters are the probabilities of the multinomial distribution, summarized in the parameter vector $\mathcal{P}$, where all entries of $\mathcal{P}$ lie in the unit interval and sum-up to one. In comparison to Markov Chains, NDARMA models maintain the nice interpretable ARMA-type structure and have a parsimonious parametrization. Furthermore, NDARMA models fulfill certain Yule–Walker-type equations, as shown in Weiß and Göb (2008).

In Figure 2.2, one realization of an NDARMA(1,0) process, denoted by NDAR(1),

$$X_t = a_t X_{t-1} + b_t e_t, \quad [a_t, b_t] \sim Mult\left(1; \alpha, \beta\right), \quad \beta = 1 - \alpha \tag{2.2}$$

with binary state space is shown. NDAR(1) models are probably the simplest members of the NDARMA class, but Figure 2.2 nicely illustrates the limited flexibility of the whole NDARMA class. The sampling mechanism of choosing the predecessor with some probability α tends to generate long runs of the same value in particular when the parameter $\alpha \in (0, 1)$ is large. A switching from one state to the other, e.g., from $X_{t-1} = 0$ to $X_t = 1$, can only occur, e.g., if the error term e_t is selected (with probability $1 - \alpha$) and the error term takes the value $e_t = 1$. Hence, the NDARMA class does not allow *systematically* selecting the opposite value of X_{t-1} for X_t.

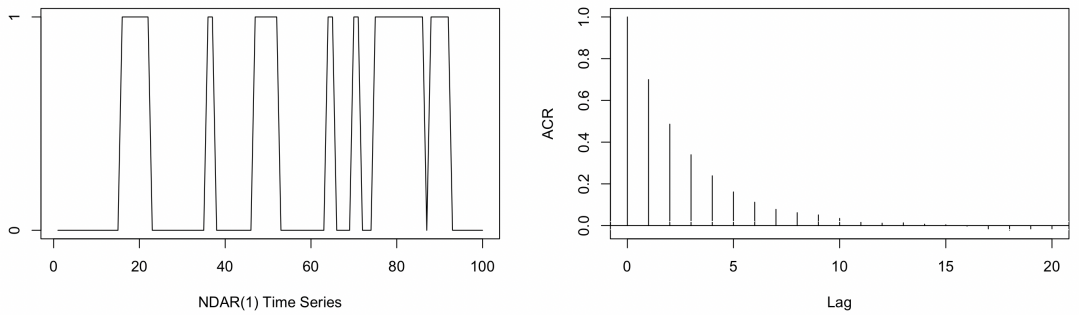


Figure 2.2.: Realization of an NDAR(1) process (Equation (2.2)) with parameter vector $\mathcal{P} = [0.7, 0.3]$ and error distribution $P\left(e_t = 1\right) = 0.5$ and corresponding autocorrelation function (ACF).

As for the NDARMA class all model parameters are restricted to be non-negative, which explains in particular why the NDARMA class can model exclusively non-negative autocorrelations in the data. For the example of a NDAR(1) process, the autocorrelation at lag one is equal to $\alpha \in [0, 1)$, such that any alternating pattern that

corresponds to negative model parameters as, e.g., observed in Figure 2.1a, cannot be captured. For a more detailed discussion of the properties of NDARMA models, we refer also to Jacobs and Lewis (1983) or Weiß (2009a). To increase its flexibility, Gouveia et al. (2018) proposed an extension of the NDARMA model class by using a variation function, but the resulting models do also not allow for negative model parameters and, hence, no negative dependence structure. Hence, whenever negative dependence structure is present in binary time series data, the NDARMA model is not suitable. In fact, in all three data examples of Figure 2.1, a straightforward estimation based on Yule–Walker estimators leads to at least some negative coefficients, such that NDAR models turn out to be not applicable.

To address this lacking flexibility of the NDARMA model class, we propose a simple and straightforward extension of the original idea of Jacobs and Lewis (1983) that allows also negative serial dependence. The resulting generalized binary ARMA (gbARMA) model class maintains the nicely interpretable model structure. Furthermore, no additional parameters are required to handle the negative dependence, preserving the parsimonious parameterization as well. In Figure 2.3, a realization of a gbARMA(1,0) process, denoted as gbAR(1), is shown. As a straightforward extension of an NDAR(1) model in Figure 2.2, gbAR(1) models allow for negative serial dependence. In fact, the range of the autocorrelation at lag one is extended from $[0, 1)$ for NDAR(1) to $(-1, 1)$ for gbAR(1) models.

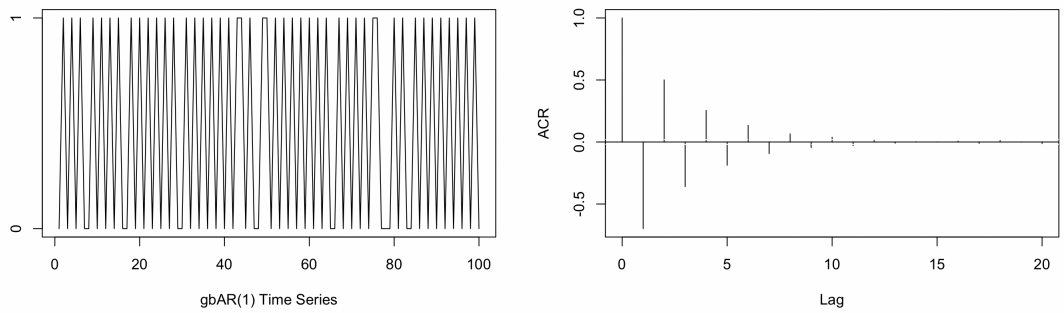


Figure 2.3.: Realization of a gbAR(1) process (Equation (2.4)) with parameter vector $\mathcal{P} = [-0.7, 0.3]$ and error distribution $P(e_t = 1) = 0.5$ and the corresponding autocorrelation function (ACF).

To allow for negative autocorrelation up to some limited extend, Kanter (1975) proposed the binary ARMA model class, where he applied the modulo 2 operator in an ARMA-type model equation. Using the modulo operation assures to stay in the binary state space, but the nice interpretability of the dependence structure in the model is lost since the past values of the time series are summed up prior to the modulo operation, see also McKenzie (1981). We follow a different path in this thesis and propose a much simpler operation that enables modeling a systematic change of the state from one time point to the other.

The idea of allowing for negative serial dependence resulting in the gbARMA class is as follows: a negative model parameter $\alpha \in (-1, 0)$ (and hence a negative autocorrelation $\alpha \in (-1, 0)$) in binary time series data corresponds to the time series *systematically* changing from one state to the other over time. Hence, the natural idea to incorporate negative serial dependence in the *binary* NDAR(1) Model (Equation

(2.2)) is to replace X_{t-1} by $1 - X_{t-1}$ as

$$(1 - X_{t-1}) = \begin{cases} 1 & \text{for } X_{t-1} = 0 \\ 0 & \text{for } X_{t-1} = 1 \end{cases} \tag{2.3}$$

holds. This leads to the model equation

$$X_t = a_t (1 - X_{t-1}) + b_t e_t, \quad [a_t, b_t] \sim Mult(1; |\alpha|, \beta).$$

This process has negative autocorrelation α at lag one. Note that, in comparison to Equation (2.2), as $\alpha \in (-1, 0)$ here, we have to use its absolute value $|\alpha|$ as the probability to select the $1 - X_{t-1}$. Altogether, for $\alpha \in (-1, 1)$, we can define the generalized binary AR(1) (gbAR(1)) process by the model equation

$$X_t = \begin{cases} a_t X_{t-1} + b_t e_t, & [a_t, b_t] \sim Mult(1; \alpha, \beta), & \alpha \in [0, 1) \\ a_t (1 - X_{t-1}) + b_t e_t, & [a_t, b_t] \sim Mult(1; |\alpha|, \beta), & \alpha \in (-1, 0) \end{cases}. \tag{2.4}$$

Note that Equation (2.4) extends the parameter space from $\alpha \in [0, 1)$ for NDAR(1) models to $\alpha \in (-1, 1)$ for gbAR(1) models. Further, note that, for identification of the model, we have to assume $\beta^{(0)} = \beta \in (0, 1]$. Using indicator variables, Equation (2.4) can be compactly written as

$$X_t = a_t \left(\mathbb{1}_{\{\alpha \geq 0\}} X_{t-1} + \mathbb{1}_{\{\alpha < 0\}} (1 - X_{t-1}) \right) + b_t e_t \tag{2.5}$$

$$= \left[a_t^{(+)} X_{t-1} + a_t^{(-)} \right] + b_t e_t \tag{2.6}$$

with $[a_t, b_t] \sim Mult(1; |\alpha|, \beta)$, $\beta = 1 - |\alpha|$, $a_t^{(+)} := a_t \left(\mathbb{1}_{\{\alpha \geq 0\}} - \mathbb{1}_{\{\alpha < 0\}} \right)$ and $a_t^{(-)} := a_t \mathbb{1}_{\{\alpha < 0\}}$.

In Figure 2.3, a realization of a gbAR(1) process with negative parameter $\alpha = -0.7$ is shown, where the time series tends to take systematically the opposite state of the predecessor. The corresponding autocorrelation plot reflects the negative serial dependence leading to an alternating pattern. Runs of the same state can only occur, when the error term e_t is selected (with probability $1 - |\alpha|$) and the error term e_t takes the same value as X_{t-1}, that is, $e_t = X_{t-1}$. The empirical autocorrelations for the Old Faithful Geyser data can be found in Figure 2.4a, where the pronounced alternating behavior clearly indicates negative linear dependence to be present in the data.

The idea of allowing for a negative model coefficient by replacing X_{t-1} by $1 - X_{t-1}$ in gbAR(1) processes (Equation (2.5)) can be also employed for each parameter in pth order gbAR processes, where each X_{t-i}, $i = 1, \ldots, p$ may be replaced by $1 - X_{t-i}$ in the model equation.

The chapter is organized as follows. In Section 2.2, generalized binary AR processes of order $p \in \mathbb{N}$ are defined, where we also give stationarity conditions and state the stationary solution. Further, stochastic properties are derived that include formulas for the transition probabilities, the marginal distribution, and Yule–Walker equations. As a real data example, we illustrate the applicability of our model class to the geyser eruption data in Section 2.1. In Section 2.3, we present several simulation experiments. First, in Section 2.3.1, for the example of a gbAR(2) model, we illustrate the generality of the resulting gbAR model class in comparison to natural competitors including AR, NDAR, and Markov models of order two, respectively. In Section 2.3.2, we examine

the estimation performance of Yule–Walker estimators in the gbAR models in Section 2.3.2. In Section 2.3.2, we investigate the benefit of using the parsimonious gbAR models in comparison to Markov models and their robustness in cases where the model is mis-specified. By adding a moving-average part to gbAR models in Section 2.4, ARMA-type extensions of gbAR models leading to gbARMA processes are discussed. We conclude in Section 2.5. All proofs are deferred to Section 2.6.

2.2. The generalized binary Autoregressive (gbAR) Model Class

We define now generalized binary AR(p) (gbAR(p)) models for binary data based on the notation of NDAR(p) models by adopting the idea of replacing X_{t-1} by $1-X_{t-1}$ for a negative parameter α as in Equations (2.5) and (2.6) separately for all or some of the lagged values $X_{t-1}, \ldots, X_{t-p}$. To be most flexible, each parameter $\alpha^{(i)}$ corresponding to the lagged value X_{t-i}, $i = 1, \ldots, p$ is allowed to be either positive or negative, that is, $\alpha^{(i)} \in (-1, 1)$, respectively.

2.2.1. gbAR Models

The parameter vector $\mathcal{P} := \left[\alpha^{(1)}, \ldots, \alpha^{(p)}, \beta^{(0)}\right]$ contains the probabilities of the multinomial distribution that controls the selection mechanism of NDAR models. As we allow now for $\alpha^{(i)} \in (-1, 1)$, $i = 1, \ldots, p$, i.e. the parameters can be negative, $\mathcal{P}$ has to be modified to serve again as a parameter vector of probabilities. This is achieved by taking entry-wise absolute values and we define

$$\mathcal{P}_{|\cdot|} := \left[|\alpha^{(1)}|, \ldots, |\alpha^{(p)}|, \beta^{(0)}\right], \tag{2.7}$$

where $\beta^{(0)} \in (0, 1]$ such that $\sum_{i=1}^{p} |\alpha^{(i)}| + \beta^{(0)} = 1$. This enables us to give the definition of the generalized binary AR model of arbitrary order $p \in \mathbb{N}$.

Definition 2.1 (Generalized binary AR processes)
Let $(X_t, t \in \mathbb{Z})$ be a stationary process taking values in $\{0, 1\}$. Let $(e_t, t \in \mathbb{Z})$ be a binary error process such that e_t is independent of $(X_s, s < t)$ with mean $\mu_e = E(e_t) = P(e_t = 1)$ and variance $\sigma_e^2 = Var(e_t) = P(e_t = 1)(1 - P(e_t = 1)) > 0$. Let $\mathcal{P} := \left[\alpha^{(1)}, \ldots, \alpha^{(p)}, \beta^{(0)}\right]$ be the parameter vector with $\mathcal{P}_{|\cdot|}$ as in Equation (2.7) such that $\mathcal{P}_{|\cdot|}\mathbb{1}_{p+1} = 1$ with $\mathbb{1}_{p+1}$ the one vector of length $p+1$. Further, let

$$P_t := \left(a_t^{(1)}, \ldots, a_t^{(p)}, b_t^{(0)}\right) \sim Mult\left(1; \mathcal{P}_{|\cdot|}\right), \quad t \in \mathbb{Z},$$

be i.i.d. random vectors, which are independent of $(e_t, t \in \mathbb{Z})$ and $(X_t, s < t)$. Then, the process $(X_t, t \in \mathbb{Z})$ is said to be a generalized binary AR process of order p (gbAR(p)), if it follows the recursion

$$X_t = \sum_{i=1}^{p} \left[a_t^{(+,i)} X_{t-i} + a_t^{(-,i)}\right] + b_t^{(0)} e_t \tag{2.8}$$

with $a_t^{(+,i)} := a_t^{(i)} \left(\mathbb{1}_{\{\alpha^{(i)} \geq 0\}} - \mathbb{1}_{\{\alpha^{(i)} < 0\}}\right)$ and $a_t^{(-,i)} := a_t^{(i)} \mathbb{1}_{\{\alpha^{(i)} < 0\}}$ for $i = 1, \ldots, p$.

By rewriting the random variables $a_t^{(\cdot,i)}$, $\cdot \in \{-,+\}$ in the defining model (Equation (2.8)), the model can be represented in the spirit of Equation (2.5). However, the benefit of the representation in Equation (2.8) is that only one random variable is multiplied with the lagged value X_{t-i}, whereas $a_t^{(-,i)}$ is an additional random variable that accounts for the switching that leads to negative model coefficients.

2.2.2. Stochastic Properties of gbAR Models

Before calculating moments of the binary time series process $(X_t, t \in \mathbb{Z})$ itself, we first consider the expectation of the random variables related to the multinomial selection mechanism. Noting that $E(a_t^{(i)}) = |\alpha^{(i)}|$, we have

$$E\left(a_t^{(-,i)}\right) = |\alpha^{(i)}| \mathbb{1}_{\{\alpha^{(i)}<0\}} =: \alpha^{(-,i)},$$
$$E\left(a_t^{(+,i)}\right) = \alpha^{(i)}.$$

This enables us to compute the stationary mean $\mu_X = E(X_t)$ of the process, directly leading to

$$\mu_X = \frac{\sum_{i=1}^p \alpha^{(-,i)} + \beta^{(0)}\mu_e}{1 - \sum_{i=1}^p \alpha^{(i)}}. \tag{2.9}$$

If all parameters $\alpha^{(1)}, \ldots, \alpha^{(p)}$ are non-negative, the above formula becomes $\mu_X = \mu_e$ due to $\sum_{i=1}^p \alpha^{(-,i)} = 0$ and $1 - \sum_{i=1}^p \alpha^{(i)} = \beta^{(0)}$, leading then to the well-known formula for the mean of NDAR(p) models. Otherwise, we have $\mu_X \neq \mu_e$ for gbAR(p) models in contrast to NDAR(p) models (see, e.g., Weiß (2009a)).

For the familiar stationary condition imposed on the model parameters $\alpha^{(1)}, \ldots, \alpha^{(p)}$ that all roots of the characteristic polynomial lie outside the unit circle, i.e. if

$$\left(1 - \alpha^{(1)}z - \ldots - \alpha^{(p)}z^p\right) \neq 0 \quad \forall z \leq 1 \tag{2.10}$$

holds, the stationary solution of the gbAR(p) model can be derived. Note that the condition in (2.10) is equivalent to $\sum_{i=1}^p |\alpha^{(i)}| < 1$, such that the error has to be selected with strictly positive probability $\beta^{(0)} > 0$ by the multinomial distribution. If the stationarity condition in Equation (2.10) holds, a moving-average representation of the gbAR(p) process can be derived.

For constructing the stationary solution of the gbAR time series, we follow the common approach based on a multivariate representation of the model, as in Lütkepohl (2005)[Chap. 11.3.2]. Precisely, the gbAR(p) model can be written as a p-dimensional gbVAR(1) process $(Y_t, t \in \mathbb{Z})$ with the following matrices and vectors, such that the first entry of $(Y_t, t \in \mathbb{Z})$ is equal to the gbAR(p) process. We define

$$Y_t := \begin{pmatrix} X_t \\ \vdots \\ X_{t-p+1} \end{pmatrix} \quad (p \times 1) \qquad \text{and} \quad U_t := \begin{pmatrix} e_t \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad (p \times 1)\,.$$

To obtain a vector autoregressive representation for Y_t, we have to define several matrices that contain the random variables of the multinomial distribution. Precisely, for $\cdot = \{-,+\}$, let

$$\widetilde{A}_t^{(\cdot)} := \begin{pmatrix} a_t^{(\cdot,1)} & \dots & a_t^{(\cdot,p-1)} & a_t^{(\cdot,p)} \\ 1 & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & \dots & 1 & 0 \end{pmatrix} \quad \text{and} \quad \widetilde{B}_t^{(1)} := \begin{pmatrix} b_t^{(0)} & 0 & \dots & 0 \\ 0_{p-1\times 1} & 0_{p-1\times 1} & \dots & 0_{p-1\times 1} \end{pmatrix}$$

be $p\times p$ matrices, where $0_{r\times s}$ denotes the $(r\times s)$-dimensional zero matrix. Based on the notation introduced above, gbVAR(p) processes can be represented as a vector-valued gbAR model of first order (gbVAR(1)) as follows

$$Y_t = \widetilde{A}_t^{(+)} Y_{t-1} + \widetilde{A}_t^{(-)} \mathbb{1}_p + \widetilde{B}_t^{(1)} U_t, \tag{2.11}$$

where $\mathbb{1}_p$ is the one vector of length p. The above notation enables us to state a moving-average representation of gbAR(p) processes as follows.

Theorem 2.2 (Moving-average representation of gbAR processes)
Let $(X_t, t\in\mathbb{Z})$ be a stationary gbAR(p) process, that is $(X_t, t\in\mathbb{Z})$ fulfills Equation (2.10)*. Then, we have*

(i) For $p = 1$, the gbAR(1) model has a gbMA(∞)-type representation (in L_2-sense), that is,

$$X_t = \sum_{i=0}^{\infty} \zeta_i a_{t-i}^{(-)} + \sum_{i=0}^{\infty} \zeta_i b_{t-i}^{(0)} e_{t-i}, \quad t\in\mathbb{Z}, \tag{2.12}$$

where $\zeta_0 := I_K$ and $\zeta_i := \prod_{j=0}^{i-1} a_{t-j}^{(+)}$ since $\lim_{k\to\infty} \prod_{i=0}^{k-1} a_{t-i}^{(+)} = 0$ in L_2.

(ii) For $p\in\mathbb{N}$, the gbAR(p) model has a gbMA(∞)-type representation (in L_1-sense), that is,

$$X_t = e_1^T \left(\sum_{i=0}^{\infty} \prod_{j=0}^{i-1} \widetilde{A}_{t-j}^{(+)} \widetilde{A}_{t-i}^{(-)} \mathbb{1}_p + \sum_{i=0}^{\infty} \prod_{j=0}^{i-1} \widetilde{A}_{t-j}^{(+)} \widetilde{B}_{t-i}^{(1)} U_{t-i} \right), \quad t\in\mathbb{Z}, \tag{2.13}$$

since $\lim_{k\to\infty} \prod_{i=0}^{k} \widetilde{A}_{t-i}^{(+)} = 0_{p\times p}$ in L_1. Here, e_1 is the first unit vector and $\mathbb{1}_p$ is the one vector of length p. The notation used here is obtained from that used in Section 2.4.2 for the special case of $q = 0$.

Hence, the process can be represented as an infinite weighted sum of the error terms. However, in comparison to classical AR or NDAR processes, an additional term appears that takes control of potential negative parameters and, consequently, allows for negative dependence to be modeled. This term vanishes if all parameters $\alpha_1, \dots, \alpha_p$ are positive.

The second-order dependence structure of gbAR processes coincides with that of AR or NDAR processes in the sense that the same Yule–Walker equations for $h \neq 0$ hold. However, note again that the parameter space for gbAR models is considerably larger than for NDAR models allowing for negative parameters leading to more flexibility. The Yule–Walker equations link the model parameters to the autocovariances of the process. Hence, they can be used for estimating the model parameters by the same well-known Yule–Walker estimators. A link between the autocovariances, the model coefficients, and the mean and variance of the error terms is established by the Yule–Walker equation for $h = 0$, respectively.

Theorem 2.3 (Yule Walker Equations)
Let $(X_t, t \in \mathbb{Z})$ be a stationary gbAR(p) process.

(i) For all $h \in \mathbb{N}$, we have

$$\gamma(h) = \sum_{i=1}^{p} \alpha^{(i)} \gamma(|h-i|). \tag{2.14}$$

(ii) For $h = 0$, we have

$$\gamma(0) = \sigma_e^2 + \frac{(1-2\mu_X)\sum_{i=1}^{p} \alpha^{(-,i)} + \left(\sum_{i=1}^{p} |\alpha^{(i)}| - 1\right)\mu_X^2 + \beta^{(0)}\mu_e^2}{\left(1 - \sum_{i=1}^{p} |\alpha^{(i)}|\right)}. \tag{2.15}$$

The next Lemma states some basic properties of the marginal distribution of gbAR processes and their transition probabilities. Since the time series has a binary state space, these conditional probabilities allow quantifying the probability to reach a certain state from the past values. In their derivation, the multinomial selection mechanism plays a crucial role and, in the stated formulas, the Kronecker delta δ_{ij} indicates if a past value has actually impact on the outcome of the time series or not.

Lemma 2.4 (Marginal, joint and transition probabilities of gbAR processes)
Let $(X_t, t \in \mathbb{Z})$ be a stationary gbAR(p) model and set $p_i := P(e_t = i)$. Then, the following properties hold:

(i) $P(X_t = i_0 | X_{t-1} = i_1, \ldots, e_t = j_0)$
$= \sum_{l=1}^{p} |\alpha^{(l)}| \left[\mathbb{1}_{\{\alpha^{(l)} \geq 0\}} \delta_{i_0 i_l} + \mathbb{1}_{\{\alpha^{(l)} < 0\}} \delta_{i_0 (1-i_l)}\right] + \beta^{(0)} \delta_{i_0 j_0}$

(ii) $P(X_t = i_0 | X_{t-1} = i_1, \ldots)$
$= \sum_{l=1}^{p} |\alpha^{(l)}| \left[\mathbb{1}_{\{\alpha^{(l)} \geq 0\}} \delta_{i_0 i_l} + \mathbb{1}_{\{\alpha^{(l)} < 0\}} \delta_{i_0 (1-i_l)}\right] + \beta^{(0)} p_{i_0}$

(iii) $P(X_t = j) = \frac{\beta^{(0)}}{\left(1 - \sum_{i=1}^{p} \alpha^{(i)}\right)} p_j + \frac{\sum_{i=1}^{p} |\alpha^{(i)}| \mathbb{1}_{\{\alpha^{(i)} < 0\}}}{\left(1 - \left[\sum_{i=1}^{p} \alpha^{(i)}\right]\right)}$

(iv) $P(X_t = i_0, e_t = j_0) = \beta^{(0)} \delta_{i_0 j_0}$
$+ \sum_{i=1}^{p} |\alpha^{(i)}| \mathbb{1}_{\{\alpha^{(i)} < 0\}} \left(1 - p_{j_0} + \frac{p_{j_0}}{1 - \sum_{i=1}^{p} |\alpha^{(i)}| \left[\mathbb{1}_{\{\alpha^{(i)} \geq 0\}} - \mathbb{1}_{\{\alpha^{(i)} < 0\}}\right]}\right)$

Comparing the results in Lemma 2.4 with Weiß (2009a)[Lemma 11.2.1.3] established for NDARMA(p,q) processes, the main difference is in Part (iii). The marginal distribution of the NDARMA process is equal to the marginal distribution of the error term process, but this does not hold for gbAR processes. Instead, the marginal distribution of gbAR processes depends on an additional term that results from the absolute values of the negative parameters.

In the following example, let us conclude this section with a more detailed look at the gbAR(1) model and a real data example.

Example 2.5 (gbAR(1) process)
Let $(X_t, t \in \mathbb{Z})$ be a gbAR(1) process with parameter vector $\mathcal{P} := \left[\alpha^{(1)}, \beta^{(0)}\right]$, $\alpha^{(1)} \in (-1, 1)$ and $\beta^{(0)} = 1 - |\alpha^{(1)}|$. The i.i.d. error term process $(e_t, t \in \mathbb{Z})$ follows the

distribution $P(e_t = 1) = p_1 \in (0,1)$ *such that* $\mu_e = p_1$ *and* $\sigma_e^2 = p_1(1-p_1) > 0$. *Then, the model equation equals*

$$X_t = a_t^{(1)} \left[\mathbb{1}_{\{\alpha^{(1)} \geq 0\}} X_{t-1} + \mathbb{1}_{\{\alpha^{(1)} < 0\}} (1 - X_{t-1}) \right] + b_t^{(0)} e_t,$$
$$\left[a_t^{(1)}, b_t^{(0)} \right] \sim Mult\left(1; \left[|\alpha^{(1)}|, \beta^{(0)} \right]\right).$$

At each time point t, *if* $\alpha^{(1)} \geq 0$, *either the predecessor* X_{t-1} *with probability* $\alpha^{(1)}$ *or the error term* e_t *with probability* $\beta^{(0)}$ *is selected by a multinomial distributed random variable to determine* X_t. *In the case of* $\alpha^{(1)} < 0$, *either* $1 - X_{t-1}$ *with probability* $|\alpha^{(1)}|$ *or the error term* e_t *with probability* $\beta^{(0)}$ *is selected. That is, as for each* t, *either* a_t *or* b_t *is equal to one and the other is zero, it holds*

$$X_t = \begin{cases} X_{t-1} & \text{if} \quad a_t^{(1)} = 1 \quad \text{for} \quad \alpha^{(1)} \geq 0 \\ 1 - X_{t-1} & \text{if} \quad a_t^{(1)} = 1 \quad \text{for} \quad \alpha^{(1)} < 0 \\ e_t & \text{if} \quad b_t^{(0)} = 1 \end{cases}.$$

For positive values of $\alpha^{(1)}$, *the gbAR(1) model coincides with the NDAR(1) model. A corresponding realization is shown in Figure 2.2, where for large values of* $\alpha^{(1)}$ *mainly the predecessor* X_{t-1} *is chosen and long runs of the same value occur. Figure 2.3 shows one realization of a gbAR(1) process with negative value of* $\alpha^{(1)}$. *The time series switches its states from zero to one and vice versa at most time points.*

The transition probability to move from state i_1 *at time* $t-1$ *to state* i_0 *at time* t *is given by*

$$P(X_t = i_0 | X_{t-1} = i_1) = |\alpha^{(1)}| \left[\mathbb{1}_{\{\alpha^{(1)} \geq 0\}} \delta_{i_0 i_1} + \mathbb{1}_{\{\alpha^{(1)} < 0\}} \delta_{i_0 (1-i_1)} \right] + \left(1 - |\alpha^{(1)}|\right) p_{i_0}. \tag{2.16}$$

The probability of the process taking the value $i_0 = 1$ *depends on two terms. First, the probability of choosing the error term is multiplied by the probability of the error term taking the same value as* X_t, *e.g.,* $P(e_t = i_0) = p_{i_0}$ *with* $i_0 = 1$. *If the probability of choosing the predecessor is added, it depends on its value and the sign of* α. *If, for example* $\alpha < 0$, *then the probability of choosing* X_{t-1} *is just added if its value is the contrary of* i_0, *such that the Kronecker delta is equal to one. This leads to the representation of Equation* (2.16) *as*

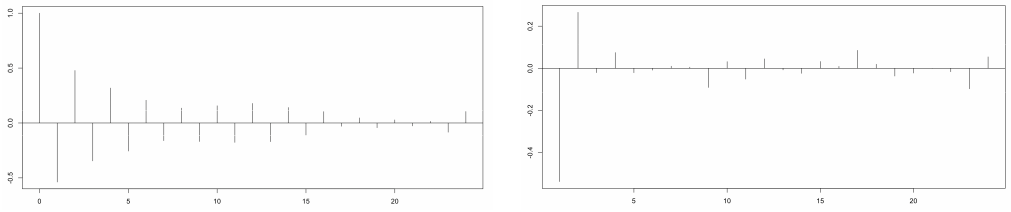
$$P(X_t = i_0 | X_{t-1} = i_1) = \begin{cases} |\alpha^{(1)}| \delta_{i_0 i_1} + \left(1 - |\alpha^{(1)}|\right) p_{i_0} & \text{if} \quad \alpha \geq 0 \\ |\alpha^{(1)}| \delta_{i_0 (1-i_1)} + \left(1 - |\alpha^{(1)}|\right) p_{i_0} & \text{if} \quad \alpha < 0 \end{cases}.$$

Example 2.6 (Eruption duration of the Old Faithful Geyser)
The binarized eruption duration of the Old Faithful Geyser is illustrated in Figure 2.1a. Its empirical autocorrelation, as shown in Figure 2.4a, clearly indicates that there is negative serial dependence present in the data such that a gbAR(p) process appears to be appropriate. The order selection using the AIC criterion leads to a model of order $p = 2$ *with* $AIC = 159.83$. *This selection is confirmed by an inspection of the partial autocorrelation in Figure 2.4b. Parameter estimation is based on the Yule–Walker Equation* (2.14) *leading to the estimated parameter vector* $\widehat{\mathcal{P}} = [-0.3949, 0.2659, 0.3393]$ *and*

the fitted model

$$X_t = \sum_{i=1}^{2} \left[a_t^{(+,i)} X_{t-i} + a_t^{(-,i)} \right] + b_t^{(0)} e_t,$$
$$P_t \sim Mult\left(1; [|-0.3949|, 0.2659, 0.3393]\right).$$

The sample mean of the binary time series is equal to $\widehat{\mu}_X = 0.6488$ *since long eruptions of the geyser arise more often than short eruption duration. The first parameter* $\widehat{\alpha}_1$ *is indeed estimated to be negative and the second one* $\widehat{\alpha}_2$ *to be positive. From* $\widehat{\beta}^{(0)} = 1 - \sum_{i=1}^{2} |\widehat{\alpha}^{(i)}|$*, an error term is chosen with probability* $\widehat{\beta}^{(0)} = 0.3393$*. In Figure 2.1a, a change from zero to one or vice versa can be observed in many time steps, whereas the run of ones in the time series correspond in most cases to choosing an error term. The error term distribution is calculated by Equation* (2.9) *with* $\widehat{\mu}_e = P(e_t = 1) = 0.9953$*.*



(a) ACF (b) pACF

Figure 2.4.: Autocorrelation (ACF) and partial autocorrelation (pACF) of the Old Faithful Geyser data

To measure the predictive power of the estimated model, we use ROC curves and the corresponding area under the curve (AUC). The ROC concept indicates a good predictive performance whenever the resulting curve is "far away" above the diagonal leading to an AUC larger than 0.5. Note that the diagonal corresponds to the case of independent observations, where no prediction based on past values is meaningful. For the one step ahead prediction, the transition probability of Lemma 2.4 (ii) is used by plugging in the estimated probabilities.

Comparing the predictor to the realized values in the sample leads to the ROC curve shown in Figure 2.5, where the corresponding AUC becomes 0.8317*. Hence, as the ROC curve is "far away" above the diagonal and the AUC is larger than 0.5, the prediction performance of the gbAR model turns out to be considerably better than that of a model that relies on independent observations. By allowing for negative model parameters, gbAR models appear to be suitable for this real data example that shows negative serial dependence.*

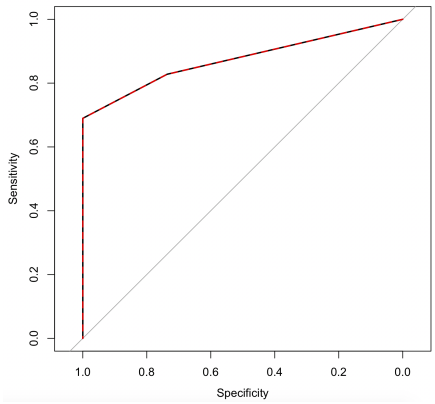


Figure 2.5.: ROC curves based on Yule–Walker estimation (black) and MLE (red) of a gbAR(2) model fitted to the binarized Old Faithful Geyser eruption duration leading to an $AUC = 0.8317$ in both cases.

For further improvement, the Yule–Walker estimates might serve as starting values for a maximum likelihood estimation (MLE) based on the conditional log-likelihood function

$$\ell\left(\theta|x_p,\ldots,x_1\right) = \sum_{t=p+1}^{T} \log p\left(x_t|x_{t-1},\ldots,x_{t-p}\right), \tag{2.17}$$

where $p\left(x_t|x_{t-1},\ldots,x_{t-p}\right) := P\left(X_t = x_t|X_{t-1} = x_{t-1},\ldots,X_{t-p} = x_{t-p}\right)$ *(see also Weiß (2018) [(B.6)]). However, the resulting parameter estimates* $\widehat{P}_{MLE} = [-0.3935, 0.2711, 0.3353]$ *differ only slightly from the Yule–Walker estimates* $\widehat{P} = [-0.3949, 0.2659, 0.3393]$*, leading to virtually the same ROC and AUC.*

(a) Yule–Walker

T	$\widehat{\boldsymbol{\alpha}}^{(1)}$	$\widehat{\boldsymbol{\alpha}}^{(2)}$	$\widehat{\boldsymbol{\beta}}^{(0)}$
50	−0.5819	0.1444	0.2738
100	−0.4610	0.2675	0.2715
150	−0.3748	0.3382	0.2871
200	−0.3738	0.3440	0.2822
250	−0.4048	0.2625	0.3328
299	−0.3949	0.2659	0.3393

(b) MLE

T	$\widehat{\boldsymbol{\alpha}}^{(1)}$	$\widehat{\boldsymbol{\alpha}}^{(2)}$	$\widehat{\boldsymbol{\beta}}^{(0)}$
50	−0.5556	0.1812	0.2632
100	−0.4546	0.2822	0.2632
150	−0.3658	0.3511	0.2830
200	−0.3706	0.3514	0.2780
250	−0.4004	0.2723	0.3259
299	−0.3935	0.2711	0.3353

Table 2.1.: Comparison of (a) Yule–Walker and (b) MLE parameter estimates based on subsamples of length $T = 50, 100, 150, 200, 250$ of the binarized Old Faithful Geyser data.

To shed some light on the potential improvement of MLE in comparison to Yule–Walker estimation, we fit a gbAR(2) model to subsamples of length $T = \{50, 100, 150, 200, 250\}$ *of the binarized Old Faithful Geyser data. The parameter estimates for Yule–Walker estimates and MLE are shown in Table 2.1. The results differ only slightly and decrease with increasing subsample sizes.*

2.3. Generality of the gbAR Model Class and Estimation Performance

In this section, we investigate the generality of the gbAR model class in comparison to obvious competitors in Section 2.3.1 and address the estimation performance in different setups and in comparison to parameter-intensive Markovian models in Section 2.3.2.

2.3.1. Illustration of the Generality of gbAR Models

By construction of the gbAR model and in contrast to NDAR models, negative parameters $\alpha^{(i)} \in (-1,1), i = 1,\dots,p$ are allowed such that negative autocorrelation is possible. Hence, the proposed gbAR model class clearly generalizes the NDAR model class. In this section, we aim to shed some light on the question how much more general the gbAR model actually is in comparison to other AR-type models such as AR, NDAR, and binAR models, as well as Markov models. For this purpose, we consider such models of order $p = 2$ and study their generality. That is, we compare the parameter ranges of these four model classes as well as the possible ranges of pairs of autocorrelation $(\rho(1), \rho(2))$. Precisely, we compare the flexibility of gbAR(2), NDAR(2), binAR(2), and AR(2) processes (even if they model continuous data) and second-order Markov chains. For all four autoregressive-type models, the autocorrelations depend on the model parameters as follows

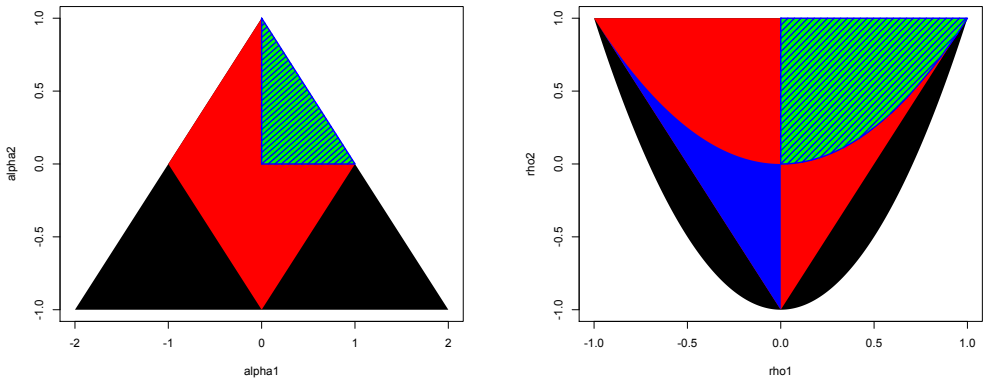
$$\rho(1) = \frac{(1-2\mu)\alpha^{(1)}}{1-(1-2\mu)\alpha^{(2)}},$$
$$\rho(2) = (1-2\mu)\frac{(1-2\mu)(\alpha^{(1)})^2}{1-(1-2\mu)\alpha^{(2)}} + (1-2\mu)\alpha^{(2)}.$$

For the gbAR(2), AR(2), and NDAR(2) processes, it holds $\mu = 0$.

For a stationary AR(2) process, the range of possible coefficients is restricted to $\alpha^{(1)} \in (-2,2)$ and $\alpha^{(2)} \in [-1,1]$ such that $\alpha^{(1)} + \alpha^{(2)} < 1$ and $\alpha^{(2)} - \alpha^{(1)} < 1$. For a stationary NDAR(2) and binAR(2) process, the parameter range is restricted by $\alpha^{(1)}, \alpha^{(2)} \in [0,1]$ with $\alpha^{(1)} + \alpha^{(2)} < 1$ and in the binAR(2) process $P(e_t = 1) = \mu \in (0,1)$ (for further details, see Weiß (2009b)). For a gbAR(2) model, the restrictions read $|\alpha^{(1)}| + |\alpha^{(2)}| < 1$ with $\alpha^{(1)}, \alpha^{(2)} \in (-1,1)$.

The parameter ranges of AR, NDAR, binAR and gbAR models of order two, respectively, are illustrated and compared in Figure 2.6a and the corresponding range of pairs of autocorrelations $(\rho(1), \rho(2))$ is shown in Figure 2.6b.

The parameter range as well as the range of autocorrelation pairs for the gbAR(2) model is considerably larger than those of an NDAR(2) model. The range of the classical AR model is again larger, but this is an unfair comparison as the AR model has been proposed for continuous data and is not suitable for binary data at all. In Figure 2.6b, the areas of AR(2), NDAR(2), and binAR(2) models are hyperboloid-shaped and, as shown in Jacobs and Lewis (1983), the autocorrelations of the NDAR(2) model take just positive values. In contrast to NDAR(2) processes, the binAR(2) process captures an additional area that corresponds to negative serial dependence. The range of autocorrelation pairs for the gbAR(2) is not hyperboloid-shaped, but forms a triangle. The range of this triangle comes actually close to the range of the AR(2) model, although the comparison with the AR(2) model is indeed unfair as



(a) Parameter ranges (b) Pairs of Autocorrelations

Figure 2.6.: Comparison of (**a**) parameter ranges of $(\alpha^{(1)}, \alpha^{(2)})$ and (**b**) pairs of autocorrelations $(\rho(1), \rho(2))$ for AR(2) (black), gbAR(2) (red), binAR(2) (blue), and NDAR(2) (green) models.

the latter has been proposed for continuous data and the gbAR(2) for binary data. Compared to NDAR(2) processes, the extension allowing also for negative parameters leads to a much larger range of possible autocorrelation pairs than just the mirrored half parable. This is explained by the four times larger possible range for the model parameters of gbAR(2) processes in comparison to NDAR(2) processes, as shown in Figure 2.6a. In summary, by allowing for negative model parameters in gbAR models, we can get a considerably more flexible model class in comparison to NDAR models that is suitable to capture a wider range of dependence structures of binary time series data.

In Figure 2.7, the possible range of autocorrelation pairs of gbAR(2) processes and Markov chains of order two are shown together. Recall that the gbAR(2) model is a parsimonious member of the class of Markov chains of order two and hence less flexible. Interestingly, with respect to pairs of autocorrelations of lags one and two, the possible range for the gbAR(2) model is only slightly smaller than that of a Markov chain of order two. Moreover, the largest range shown for the (continuous) AR(2) models in Figure 2.6b (in black) cannot be attained by Markov chains of order two. Hence, gbAR(2) models can to cover a large portion of the possible range of autocorrelation pairs of lags one and two of second-order Markov chains. However, keep in mind that for a pth-order Markov chain, 2^p parameter have to be estimated. For example, $2^2 = 4$ parameters need to be specified for a second-order Markov chain, whereas gbAR(2) processes only require three parameters.

2.3.2. Simulations

In this simulation study, we addressed two things. First, as shown in Section 2.3.2, we investigated the estimation performance of Yule–Walker estimators for gbAR models of different orders and sample sizes. Second, as shown in Section 2.3.2, we studied the flexibility of the gbAR model class and compared the prediction performance to Markovian models in the case where the estimated model was correctly specified as a gbAR model and in the case where the underlying model was a Markovian model that

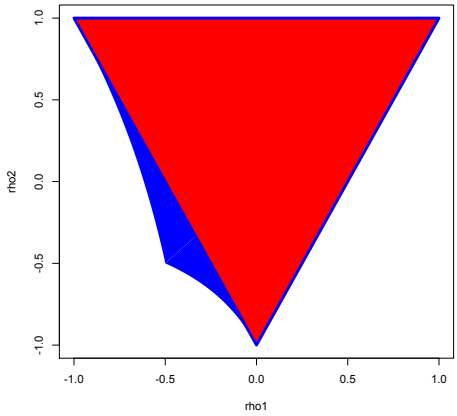


Figure 2.7.: Autocorrelation pairs of the gbAR(2) process (red) compared to the pairs of autocorrelation of a second-order Markov chain (blue).

does not belong to the class of gbAR models.

Estimation Performance

To study the estimation performance of Yule–Walker estimators in gbAR models, we considered three different specifications of gbAR(p) processes with $p = 1, 2, 3$ and sample sizes $T = 100, 200, 500, 1000$. Precisely, we considered the following gbAR data generating processes (DGPs):

(DGP1) gbAR(1) with $\alpha^{(1)} = -0.85$, $\mu_e = 0.3$, $\beta^{(0)} = 0.15$ and $\mu_X = 0.48378$.

(DGP2) gbAR(2) with $\alpha^{(1)} = 0.42$, $\alpha^{(2)} = -0.38$, $\mu_e = 0.3$, $\beta^{(0)} = 0.2$ and $\mu_X = 0.45833$.

(DGP3) gbAR(3) with $\alpha^{(1)} = 0. - 0.294$, $\alpha^{(2)} = 0.382$, $\alpha^{(3)} = -0.2393$, $\mu_e = 0.67$, $\beta^{(0)} = 0.0847$ and $\mu_X = 0.52140$.

The model parameters summarized in $\mathcal{P}$ were estimated based on Yule–Walker Equation (2.14) and the error term distribution using Equation (2.9). Note that, in all setups, we considered gbAR models that make use of the extended parameter space by including negative parameters $\alpha^{(i)}$ in the model. For each DGP, we simulated 1000 replications to calculate the mean squared error (MSE) to measure the estimation performance.

Table 2.2 summarizes the simulation results for all DGPs and all considered sample sizes. The estimation performance is generally good, as confirmed by rather small MSEs. It turns out that, as expected, in all considered setups, the estimation performance improves with increasing the sample size. It is interesting to note that, relative to the estimation of the other quantities, the estimation of the mean of the error terms μ_e is generally less precise. This can be explained by the fact that the error terms e_t do enter the time series only in the case when it is actually selected which happens only with probabilities $\beta^{(0)}_{DGP1} = 0.15$, $\beta^{(0)}_{DGP2} = 0.2$ and $\beta^{(0)}_{DGP3} = 0.0874$ for the three DGPs, respectively. A comparison of the estimation performance of the gbAR models of different orders shows that the estimation performance declines with increasing

order, which is of course plausible as the number of parameters gets larger, leading to more estimation uncertainty.

	T	**MSE of $\widehat{\alpha}^{(1)}$**	**MSE of $\widehat{\alpha}^{(2)}$**	**MSE of $\widehat{\alpha}^{(3)}$**	**MSE of $\widehat{\mu}_X$**
DGP1	100	0.00271			0.00022
	200	0.00133			0.00011
	500	0.00051			0.00004
	1000	0.00025			0.00002
DGP2	100	0.00684	0.00588		0.00191
	200	0.00338	0.00314		0.00107
	500	0.00136	0.00124		0.00044
	1000	0.00068	0.00064		0.00020
DGP3	100	0.01079	0.01253	0.00900	0.00416
	200	0.00502	0.00525	0.00386	0.00200
	500	0.00194	0.00198	0.00143	0.00082
	1000	0.00097	0.00087	0.00065	0.00040
	T	**MSE of $\widehat{\mu}_e$**	**MSE of $\widehat{\beta}^{(0)}$**		
DGP1	100	0.03140	0.00271		
	200	0.01464	0.00133		
	500	0.00543	0.00051		
	1000	0.00262	0.00025		
DGP2	100	0.04877	0.01007		
	200	0.02689	0.00521		
	500	0.01169	0.00219		
	1000	0.00555	0.00108		
DGP3	100	0.10187	0.02212		
	200	0.08931	0.01044		
	500	0.05859	0.00374		
	1000	0.03913	0.000178		

Table 2.2.: Estimation performance for several Yule–Walker parameter estimates with respect to mean squared errors for three different DGPs over 1000 Monte Carlo replications.

Robustness of gbAR Model Class

The class of gbAR(p) models form a parsimoniously parametrized subclass in the class of Markovian models of order p. To study the benefit of this newly proposed class of binary models, we wanted to compare the gbAR(p) model to Markov chains which are mostly used for binary data.

First, let us consider the case of an underlying gbAR model. Since gbAR(p) processes have a Markov chain representation, a comparison in terms of the transition probabilities becomes suitable. From Lemma 2.4 *(ii)*, the transition probabilities of

gbAR models compute to

$$P\left(X_t = i_0 | X_{t-1} = i_1, \ldots, X_{t-p} = i_p\right) = \sum_{l=1}^{p} |\alpha^{(l)}| \left[\mathbb{1}_{\{\alpha^{(l)} \geq 0\}} \delta_{i_0 i_l} + \mathbb{1}_{\{\alpha^{(l)} < 0\}} \delta_{i_0 (1-i_l)}\right] + \beta^{(0)} P\left(e_t = i_0\right). \tag{2.18}$$

First, for $p = 1, 2, 3$, we simulated from the gbAR(p) models defined for DGP1-3 in Section 2.3.2 realizations of different sample sizes to estimate the transition probabilities of: (a) a pth-order Markov chain and (b) a gbAR(p) process. For gbAR models, the true transition probabilities are given by Equation (2.18) and can be estimated by replacing the model parameters by the corresponding estimators. Then, the MSE is calculated model-wise and over all transition probabilities. For all three DGPs with model orders $p = 1, 2, 3$, the simulation results are stated in Table 2.3. By "MSE gbAR(p)", we denote the mean squared error by evaluating the difference between the estimated transition probability and the truth over all possible transition probabilities from a gbAR(p) process. Equivalently, "MSE MC" denotes the corresponding difference between the estimated transition probability and the truth of a Markov model over all possible transition probabilities.

The MSEs are calculated over 1000 replications and show clearly that the estimated gbAR(p) transition probabilities have smaller MSEs for all sample sizes and orders in comparison to the MSEs of the Markov chain fits. This indicates that, in the case of an underlying gbAR(p) process, fitting the more parsimonious model to the data leads to better estimation performance than fitting a Markov chain.

Next, we considered the situation where the underlying model is a Markov chain of order p that does *not* belong to the subclass of gbAR(p) models. In general, the $2^p \times 2$-dimensional transition probability matrix Q of a Markov chain of order p is defined by

$$Q = \left(p^{i_0}_{i_1 i_2 \ldots i_p}\right)_{i_j = 0,1, j = 0, \ldots, p}, \tag{2.19}$$

where

$$P\left(X_t = i_0 | X_{t-1} = i_1, \ldots, X_{t-p} = i_p\right) =: p^{i_0}_{i_1 i_2 \ldots i_p}. \tag{2.20}$$

For the simulations, we had to make sure that the used specifications of Q are such that the resulting model is not a member of the gbAR class. For a set of specified transition probabilities, it is actually easy to check whether the resulting model is a gbAR model, by checking whether Equation (2.18) hold true.

It turns out that the class of gbAR(1) models and the class of binary Markov chains of order 1 coincide. Hence, for the simulations study, we chose transition probabilities such that Equation (2.18) does not hold for $p = 2, 3$. Precisely, we set

$$Q = \begin{pmatrix} 0.46 & 0.54 \\ 0.56 & 0.44 \end{pmatrix}, \quad Q = \begin{pmatrix} p^0_{00} & p^1_{00} \\ p^0_{01} & p^1_{01} \\ p^0_{10} & p^1_{10} \\ p^0_{11} & p^1_{11} \end{pmatrix} = \begin{pmatrix} 0.21 & 0.79 \\ 0.69 & 0.31 \\ 0.32 & 0.68 \\ 0.89 & 0.11 \end{pmatrix},$$

$$Q = \begin{pmatrix} p^0_{000} & p^1_{000} \\ p^0_{001} & p^1_{001} \\ p^0_{010} & p^1_{010} \\ p^0_{100} & p^1_{100} \\ p^0_{011} & p^1_{011} \\ p^0_{101} & p^1_{101} \\ p^0_{110} & p^1_{110} \\ p^0_{111} & p^1_{111} \end{pmatrix} = \begin{pmatrix} 0.16 & 0.84 \\ 0.26 & 0.74 \\ 0.42 & 0.58 \\ 0.21 & 0.79 \\ 0.75 & 0.25 \\ 0.64 & 0.36 \\ 0.57 & 0.43 \\ 0.94 & 0.06 \end{pmatrix}.$$

Using such transition probabilities summarized in Q, binary time series were generated. Again, a gbAR(p) process and a pth order Markov chain were fitted. In Table 2.4, the MSE estimation performance for the different DGPs is summarized. Interestingly, although the corresponding gbAR fits (for $p = 2, 3$) do actually estimate the wrong models, with respect to MSE over all transition probabilities, their estimation performances for small sample sizes are superior to those of Markov chains that estimate the correct models. However, for large sample sizes, the estimated Markov models outperform the misspecified gbAR model fits. As it is estimating the true model, this pattern was expected. In summary, for time series with small sample size, where the true underlying DGP is indeed a Markov chain and not a gbAR(p) process, the parsimonious gbAR model might be a good approximation, leading potentially to more precise estimates of the transition probabilities although the model is misspecified.

T	**MSE gbAR(1)**	**MSE MC $p = 1$**	**MSE gbAR(2)**	**MSE MC $p = 2$**	**MSE gbAR(3)**	**MSE MC $p = 3$**
100	0.01887	0.02690	0.02647	0.02674	0.03302	0.06942
200	0.01288	0.02399	0.01889	0.01904	0.02333	0.04935
500	0.00788	0.02256	0.01125	0.01191	0.01536	0.02940
1000	0.00529	0.02168	0.00775	0.00852	0.01074	0.02024

Table 2.3.: Comparison of the estimation performance of gbAR(p) model fits and Markov chain fits of order p for $p = 1, 2, 3$ for three different gbAR-DGP1-3 with respect to the mean squared difference of estimated transition probabilities to the truth over 1000 Monte Carlo replications.

T	**MSE gbAR(1)**	**MSE MC $p = 1$**	**MSE gbAR(2)**	**MSE MC $p = 2$**	**MSE gbAR(3)**	**MSE MC $p = 3$**
100	0.03183	0.03310	0.02563	0.02784	0.02771	0.03216
200	0.02219	0.02365	0.01894	0.01951	0.02276	0.02202
500	0.01420	0.01610	0.01371	0.01242	0.01960	0.01380
1000	0.00980	0.01216	0.01132	0.00882	0.01838	0.00977

Table 2.4.: Comparison of a gbAR Model and Markov chain by its difference of the transition probabilities to the truth from an underlying Markov Process.

2.4. Further Extension: The Generalized Binary ARMA Class

In this section, we extend the gbAR model class and give a definition of generalized binary ARMA (gbARMA) models that additionally contain a moving average part in their model equations. In the spirit of the gbAR model as an extension of the NDAR model class, we allow also for negative parameters in the moving-average part of the model.

First, we provide the definition of the gbARMA(p,q) model, derive its stationary solution, and state some basic properties of marginal, joint, and transition probabilities of gbARMA(p,q) processes. We conclude this section with an example of a gbARMA(1,1) process.

2.4.1. gbARMA Models

To be most flexible, the gbARMA model class allows additionally for negative parameters to capture negative dependence structure also in the moving average part. As before, we assume $\beta^{(0)} \in (0,1]$ for identification reasons. In the gbARMA(p,q) model class, the parameters $\alpha^{(i)}$ and $\beta^{(j)}$ are allowed to be either positive or negative, e.g., $\alpha^{(i)}, \beta^{(j)} \in (-1,1)$ for $i = 1,\dots,p$ and $j = 1,\dots,q$. To modify the parameter vector $\mathcal{P} := \left[\alpha^{(1)},\dots,\alpha^{(p)},\beta^{(0)},\dots,\beta^{(q)}\right]$, again such that it contains the probabilities, we define

$$\mathcal{P}_{|\cdot|} := \left[|\alpha^{(1)}|,\dots,|\alpha^{(p)}|,\beta^{(0)},|\beta^{(1)}|\dots,|\beta^{(q)}|\right]. \tag{2.21}$$

Definition 2.7 (Generalized binary ARMA processes)
Let $(X_t, t \in \mathbb{Z})$ be a stationary process which takes values in $\{0,1\}$. Let $(e_t, t \in \mathbb{Z})$ be a binary error process such that e_t is independent of $(X_s, s < t)$ with mean μ_e and variance $\sigma_e^2 = Var(e_t) > 0$. Let $\mathcal{P} := \left[\alpha^{(1)},\dots,\alpha^{(p)},\beta^{(0)},\dots,\beta^{(q)}\right]$ be the parameter vector with $\mathcal{P}_{|\cdot|}$ as in Equation (2.21) such that $\mathcal{P}_{|\cdot|}\mathbb{1}_{p+q+1} = 1$. Further, let

$$P_t := \left(a_t^{(1)},\dots,a_t^{(p)},b_t^{(0)},\dots,b_t^{(q)}\right) \sim Mult\left(1;\mathcal{P}_{|\cdot|}\right), \qquad t \in \mathbb{Z}$$

be i.i.d. random vectors, which are independent of $(e_t, t \in \mathbb{Z})$ and $(X_t, s < t)$. Then, the process $(X_t, t \in \mathbb{Z})$ is said to be a generalized binary ARMA(p,q) process, if it follows the recursion

$$X_t = \sum_{i=1}^{p}\left[a_t^{(+,i)}X_{t-i} + a_t^{(-,i)}\right] + b_t^{(0)}e_t + \sum_{j=1}^{q}\left[b_t^{(+,j)}e_{t-j} + b_t^{(-,j)}\right] \tag{2.22}$$

with $a_t^{(+,i)} := a_t^{(i)}\left(\mathbb{1}_{\{\alpha^{(i)}\geq 0\}} - \mathbb{1}_{\{\alpha^{(i)}<0\}}\right)$, $a_t^{(-,i)} := a_t^{(i)}\mathbb{1}_{\{\alpha^{(i)}<0\}}$ and analogous definitions for $b_t^{(+,\cdot)}$ and $b_t^{(-,\cdot)}$.

The model parameters are contained in the vector P with entries $\alpha^{(i)} \in (-1,1)$ for $i = 1,\dots,p$, $\beta^{(0)} \in (0,1]$ and $\beta^{(j)} \in (-1,1)$ for $j = 1,\dots,q$. Note that, as $\beta^{(0)} > 0$ holds, no random variable $b_t^{(-,0)}$ is contained in the model equation.

With probability $\sum_{i=1}^{p}|\alpha^{(i)}|$, a predecessor $X_s,\ s \in \{t-1,\dots,t-p\}$ is chosen, whereas, with probability $\sum_{j=0}^{q}|\beta^{(j)}|$, the process takes the value of an error term $e_s,\ s \in \{t,\dots,t-q\}$, where it follows that $\sum_{i=1}^{p}|\alpha^{(i)}| + \sum_{j=0}^{q}|\beta^{(j)}| = 1$.

2.4.2. Stochastic Properties of gbARMA Models

When dealing with possibly negative parameters also in the moving-average part of gbARMA models, the idea of Equation (2.4) is employed also for the lagged error terms. Hence, this allows modeling negative dependence in the moving average part as well. In the multinomial distribution, all values of the parameter vector $\mathcal{P}$ have to be considered in absolute value, thus we have to use $\mathcal{P}_{|\cdot|}$ as defined in Equation (2.21). For the expectation of gbARMA processes, two additional sums show up in comparison to the NDARMA case. Precisely, we have

$$\mu_X = \frac{\sum_{i=1}^{p} \alpha^{(-,i)} + \sum_{j=1}^{q} \beta^{(-,j)} + \sum_{j=0}^{q} \beta^{(j)} \mu_e}{1 - \sum_{i=1}^{p} \alpha^{(i)}}.$$

The construction of the stationary solution of the gbARMA time series is similar to the construction of the gbAR(p) process introduced in Section 2.2.1 and Lütkepohl (2005)[Chap. 11.3.2]. The vector representation of the process $(Y_t, t \in \mathbb{Z})$ is equipped with a moving average part and thus the dimension of the corresponding random matrices becomes $p+q \times p+q$. Precisely, the gbARMA(p,q) model can be written as a $(p+q)$-dimensional gbVAR(1) process $(Y_t, t \in \mathbb{Z})$ with the following matrices and vectors, such that the first entry of $(Y_t, t \in \mathbb{Z})$ is equal to the gbARMA(p,q) process. We define

$$Y_t := \begin{pmatrix} X_t \\ \vdots \\ X_{t-p+1} \\ e_t \\ \vdots \\ e_{t-q+1} \end{pmatrix} \quad ((p+q) \times 1) \qquad \text{and} \quad U_t := \begin{pmatrix} e_t \\ 0 \\ \vdots \\ 0 \\ e_t \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad ((p+q) \times 1).$$

To obtain a vector autoregressive representation for Y_t, we define directly matrices that contain the random variables of the multinomial distribution. Precisely, for $\cdot = \{-, +\}$, let

$$\widetilde{A}_t^{(\cdot)} := \begin{pmatrix} A_{t,11}^{(\cdot)} & A_{t,12}^{(\cdot)} \\ A_{t,21}^{(\cdot)} & A_{t,22}^{(\cdot)} \end{pmatrix} \text{ and } \widetilde{B}_t^{(1)} := \begin{pmatrix} b_t^{(0)} & 0 & \dots & 0 \\ 0_{p-1\times 1} & 0_{p-1\times 1} & \dots & 0_{p-1\times 1} \\ 1 & 0 & \ddots & \vdots \\ 0_{q-1\times 1} & \dots & & 0_{q-1\times 1} \end{pmatrix}$$

be $(p+q) \times (p+q)$ matrices, where

$$A_{t,11}^{(\cdot)} := \begin{pmatrix} a_t^{(\cdot,1)} & \dots & a_t^{(\cdot,p-1)} & a_t^{(\cdot,p)} \\ 1 & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & \dots & 1 & 0 \end{pmatrix}, A_{t,12}^{(\cdot)} := \begin{pmatrix} b_t^{(\cdot,1)} & \dots & b_t^{(\cdot,q)} \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{pmatrix},$$

$$A_{t,22}^{(\cdot)} := \begin{pmatrix} 0 & \dots & 0 & 0 \\ 1 & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & \dots & 1 & 0 \end{pmatrix}$$

are $p \times p$, $p \times q$ and $q \times q$ matrices, respectively, and $A_{t,21} := 0_{q \times p}$. Based on the notation introduced above, gbARMA(p,q) processes can be represented as a vector-valued gbAR model of first order (gbVAR(1)) as follows

$$Y_t = \widetilde{A}_t^{(+)} Y_{t-1} + \widetilde{A}_t^{(-)} \mathbb{1}_{p+q} + \widetilde{B}_t^{(1)} U_t \tag{2.23}$$

with $\mathbb{1}_{p+q}$ being the one vector of length $p+q$.

To derive a suitable stationarity condition for the process, we know from Lütkepohl (2005) that it corresponds to the characteristic polynomial of the parameter matrix $\widetilde{\mathcal{A}} := E\left(\widetilde{A}_t^{(+)}\right)$.

$$det\left(I_{K(p+q)} - \widetilde{\mathcal{A}} z\right) \neq 0 \quad \forall \ |z| \leq 1.$$

From the block structure of $\widetilde{\mathcal{A}}$, the polynomial can be reduced to the determinant of the block matrices $\mathcal{A}_{11}^{(+)} := E\left(A_{t,11}^{(+)}\right)$ and $\mathcal{A}_{22}^{(+)} := E\left(A_{t,22}^{(+)}\right)$. Hence, a gbARMA(p,q) process is stationary if the roots of the characteristic polynomial of the autoregressive part lie outside the unit circle, that is, if

$$\left(1 - \alpha^{(1)} z - \ldots - \alpha^{(p)} z^p\right) \neq 0 \ \ \forall \ \ |z| \leq 1$$

holds. The assumption is fulfilled whenever an error term has a positive probability, such that there exists a $|\beta^{(j)}| > 0$ for some $j \in \{0, \ldots, q\}$. Therefore, the sum over all probabilities of choosing a predecessor fulfills $\sum_{i=1}^{p} |\alpha^{(i)}| < 1$. Without any restriction, we assume that $\beta^{(0)}$ is strictly positive for a stationary gbARMA process, i.e. $\beta^{(0)} \in (0, 1]$.

For a stationary gbARMA(p,q) process, a moving average representation can be derived using the above defined vectors and matrices.

Theorem 2.8 (Moving Average representation of gbARMA processes)
Let $(X_t, t \in \mathbb{Z})$ be a stationary gbARMA(p,q) process with gbVAR(1) representation (Equation (2.23)). Then, it follows that

$$X_t = e_1^T \left(\sum_{i=0}^{\infty} \prod_{j=0}^{i-1} \widetilde{A}_{t-j}^{(+)} \widetilde{A}_{t-i}^{(-)} \mathbb{1}_{p+q} + \sum_{i=0}^{\infty} \prod_{j=0}^{i-1} \widetilde{A}_{t-j}^{(+)} \widetilde{B}_{t-i}^{(1)} U_{t-i} \right),$$

where $\lim_{k \to \infty} \prod_{i=0}^{k} \widetilde{A}_{t-i}^{(+)} = 0_{(p+q) \times (p+q)}$ in L_1 and e_1 is the first unit vector.

The univariate moving average representation is obtained from the multivariate formula by multiplying it with the first unit vector because of $X_t = e_1^T Y_t$.

Considering the autocorrelation structure, Jacobs and Lewis (1983) and Weiß (2011a) showed that the NDARMA(p,q) model fulfills a set of Yule–Walker type equations which was also derived by Möller and Weiß (2020) for the GDARMA class of categorical processes. The following result shows that this property is maintained for the gbARMA class.

Theorem 2.9 (Yule–Walker-type equations)
Let $(X_t, t \in \mathbb{Z})$ be a stationary gbARMA(p,q) process. Set $\beta^{(k)} := 0$ for $k > q$. Define coefficients $(\phi_k)_{k \in \mathbb{Z}}$ recursively by

$$\phi_k = 0 \quad \text{for } k < 0, \quad \phi_0 = \beta^{(0)}, \quad \phi_k = \sum_{i=1}^{p} \alpha^{(i)} \phi_{k-i} + \beta^{(k)} \quad \text{for } k > 0.$$

Then, the autocovariance function for lags $k > 0$ is obtained by

$$\gamma(k) - \sum_{i=1}^{p} \alpha^{(i)} \gamma(|k-i|) = \sigma_e^2 \sum_{j=k}^{q} \beta^{(j)} \phi_{j-k}$$

The autocovariances of the NDARMA and GDARMA processes can only be positive, whereas the Yule–Walker type equations of gbARMA processes allow for possibly negative model parameters $\alpha^{(i)}, \beta^{(j)} \in (-1,1)$ for $i = 1,\dots,p$ and $j = 1,\dots,q$.

For the generalized binary ARMA model, formulas for the marginal, joint and transition probabilities can be calculated, extending the results from Lemma 2.4.

Lemma 2.10 (Marginal, joint, and transition probability of gbARMA processes)
Let $(X_t, t \in \mathbb{Z})$ be a stationary gbARMA(p,q) process. Then, the following properties hold:

(i) $P(X_t = i_0 | X_{t-1} = i_1, \dots, e_t = j_0 \dots)$
$= \sum_{l=1}^{p} |\alpha^{(l)}| \left[\mathbb{1}_{\{\alpha^{(l)} \geq 0\}} \delta_{i_0 i_l} + \mathbb{1}_{\{\alpha^{(l)} < 0\}} \delta_{i_0 (1-i_l)} \right] + \beta^{(0)} \delta_{i_0 j_0}$
$+ \sum_{k=1}^{q} |\beta^{(k)}| \left[\mathbb{1}_{\{\beta^{(k)} \geq 0\}} \delta_{i_0 j_k} + \mathbb{1}_{\{\beta^{(k)} < 0\}} \delta_{i_0 (1-j_k)} \right]$

(ii) *Defining $p_i := P(e_t = i)$ then it follows*
$P(X_t = i_0 | X_{t-1} = i_1, \dots, e_{t-1} = j_1 \dots)$
$= \sum_{l=1}^{p} |\alpha^{(l)}| \left[\mathbb{1}_{\{\alpha^{(l)} \geq 0\}} \delta_{i_0 i_l} + \mathbb{1}_{\{\alpha^{(l)} < 0\}} \delta_{i_0 (1-i_l)} \right]$
$+ \sum_{k=1}^{q} |\beta^{(k)}| \left[\mathbb{1}_{\{\beta^{(k)} \geq 0\}} \delta_{i_0 j_k} + \mathbb{1}_{\{\beta^{(k)} < 0\}} \delta_{i_0 (1-j_k)} \right] + \beta^{(0)} p_{i_0}$

(iii) $P(X_t = j) = \frac{\left[\sum_{l=1}^{q} |\beta^{(l)}| \left(\mathbb{1}_{\{\beta^{(l)} \geq 0\}} - \mathbb{1}_{\{\beta^{(l)} < 0\}} \right) \right] + \beta^{(0)}}{\left(1 - \sum_{i=1}^{p} \alpha^{(i)} \right)} p_j$
$+ \frac{\sum_{i=1}^{p} |\alpha^{(i)}| \mathbb{1}_{\{\alpha^{(i)} < 0\}} + \sum_{l=1}^{q} |\beta^{(l)}| \mathbb{1}_{\{\beta^{(l)} < 0\}}}{\left(1 - \sum_{i=1}^{p} \alpha^{(i)} \right)}$

(iv) $P(X_t = i_0, e_t = j_0) =$

$$p_{i_0} p_{j_0} \left[\frac{\sum_{l=1}^{q} |\beta^{(l)}| \mathbb{1}_{\{\beta^{(l)} < 0\}}}{1 - \sum_{i=1}^{p} |\alpha^{(i)}| \left[\mathbb{1}_{\{\alpha^{(i)} \geq 0\}} - \mathbb{1}_{\{\alpha^{(i)} < 0\}} \right]} - \sum_{l=1}^{q} |\beta^{(l)}| \mathbb{1}_{\{\beta < 0\}} \right] + \beta^{(0)} \delta_{i_0 j_0}$$
$$+ \sum_{l=1}^{q} |\beta^{(l)}| \mathbb{1}_{\{\beta^{(l)} < 0\}} \left(1 - p_{j_0} + \frac{p_{j_0}}{1 - \sum_{i=1}^{p} |\alpha^{(i)}| \left[\mathbb{1}_{\{\alpha^{(i)} \geq 0\}} - \mathbb{1}_{\{\alpha^{(i)} < 0\}} \right]} \right)$$
$$+ \sum_{i=1}^{p} |\alpha^{(i)}| \mathbb{1}_{\{\alpha^{(i)} < 0\}} \left(1 - p_{j_0} + \frac{p_{j_0}}{1 - \sum_{i=1}^{p} |\alpha^{(i)}| \left[\mathbb{1}_{\{\alpha^{(i)} \geq 0\}} - \mathbb{1}_{\{\alpha^{(i)} < 0\}} \right]} \right)$$

The flexibility of gbARMA models obtained by allowing for negative parameters shows also in the transition probabilities and in the joint and marginal distributions. Hence, more complex structures can be captured since systematic changes in the error terms are allowed as well.

We conclude this section with an example of a gbARMA(1,1) model.

Example 2.11 (gbARMA(1,1) process)
Let $(X_t, t \in \mathbb{Z})$ be a stationary gbARMA(1,1) process. Then, the process follows the recursion

$$X_t = a_t^{(+,1)} X_{t-1} + a_t^{(-,1)} + b_t^{(0)} e_t + b_t^{(+,1)} e_{t-1} + b_t^{(-,1)}$$

Four sign combinations of parameter pairs are possible and the corresponding model equations are given as follows:

$$X_t = \begin{cases} a_t^{(1)} X_{t-1} + b_t^{(0)} e_t + b_t^{(1)} e_{t-1} & for \quad \alpha^{(1)} \geq 0, \beta^{(1)} \geq 0 \\ a_t^{(1)} \left(1 - X_{t-1}\right) + b_t^{(0)} e_t + b_t^{(1)} e_{t-1} & for \quad \alpha^{(1)} < 0, \beta^{(1)} \geq 0 \\ a_t^{(1)} X_{t-1} + b_t^{(0)} e_t + b_t^{(1)} \left(1 - e_{t-1}\right) & for \quad \alpha^{(1)} \geq 0, \beta^{(1)} < 0 \\ a_t^{(1)} \left(1 - X_{t-1}\right) + b_t^{(0)} e_t + b_t^{(1)} \left(1 - e_{t-1}\right) & for \quad \alpha^{(1)} < 0, \beta^{(1)} < 0 \end{cases}.$$

Whereas for identification purposes $\beta^{(0)}$ only takes positive values, the predecessors X_{t-1} and e_{t-1} are systematically switched if the corresponding model parameters are negative, respectively.

For a stationary gbARMA(1,1) process, the moving average representation fulfills the following equation:

$$\begin{aligned} X_t = &\sum_{j=0}^{\infty} \prod_{i=0}^{j-1} a_{t-i}^{(+,1)} a_{t-j}^{(-,1)} + \sum_{j=0}^{\infty} \prod_{i=0}^{j-1} a_{t-i}^{(+,1)} b_{t-j}^{(0)} e_{t-j} \\ &+ \sum_{j=0}^{\infty} \prod_{i=0}^{j-1} a_{t-i}^{(+,1)} \left[b_{t-j}^{(+,1)} e_{t-(j+1)} + b_{t-j}^{(-,1)} \right]. \end{aligned}$$

From the stationarity assumption, we have $|\alpha^{(1)}| < 1$, $\beta^{(0)} \in (0, 1]$ and $|\beta^{(1)}| \in [0, 1)$.

The moving average representation consists of three parts. There first is a sum over all terms $a_{t-j}^{(-,1)}$ for the potential case of $\alpha^{(1)} < 0$. This part accounts for the choosing a predecessor and its switching. Since $\beta^{(0)}$ is strictly positive, the second is a sum over all error terms without any modification occurs. In the third sum, the random variable $b_t^{(-,1)}$ appears for controlling the case of $\beta^{(1)} < 0$.

2.5. Conclusions

By extending the NDARMA model class of Jacobs and Lewis (1983) to allow for negative parameters in the binary state space, the generalized binary ARMA model remains parsimonious, but it becomes more flexible to allow for negative model parameters and, hence, negative dependence structure in the data. The extension of the model to a more general parameter space enables the application to real data without having that many restrictions as in the NDARMA model class. Although the extension leads to additional terms in the model equation, the Yule–Walker equations still provide a direct way to estimate the model parameters.

We discuss stationarity conditions for gbARMA models and derive the stationary solution. The resulting moving average representation shows an additional term, compared to most MA(∞)-type representations. These additional terms control for the switching of the states.

An illustration of autocorrelation pairs $(\rho(1), \rho(2))$ of four different models of order 2 shows a comparison of the captured dependence structure of the time series models.

It reveals that the proposed gbARMA model can capture a wide range of negative and positive dependence structures. A second-order Markov chain is shown to capture only a slightly larger range of negative dependence structure that gbAR(2) models. Hence, by allowing for negative parameters, the proposed extension of the NDARMA model class leads to a new model class that allows capturing a wide range of dependence structures in binary time series data, while maintaining a parsimonious parametrization. Moreover, in small sample sizes, parsimonious gbAR models might turn out to be beneficial in cases where the model is actually misspecified as they may provide a sufficient approximation to the true model.

2.6. Proofs of Chapter 2

2.6.1. Proof of Theorem 2.2

Proof.
(i) By recursively inserting the model equation, the process can be expressed as

$$X_t = \prod_{i=0}^{k-1} a_{t-i}^{(+)} X_{t-k} + \sum_{i=0}^{k-1} \prod_{j=0}^{i-1} a_{t-j}^{(+)} a_{t-i}^{(-)} + \sum_{i=0}^{k-1} \prod_{j=0}^{i-1} a_{t-j}^{(+)} b_{t-i}^{(0)} e_{t-i}.$$

Since the random variables $a_t^{(+)}$ take values in $\{-1, 0, 1\}$ and the process is also binary with mean $\mu_X \in (0, 1)$, the convergence of the first part follows directly in quadratic mean (in L_2 sense), that is,

$$\begin{aligned} E\left(|\prod_{i=0}^{k-1} \left(a_{t-i}^{(+)}\right) X_{t-k}|^2 \right) &= E\left(\prod_{i=0}^{k-1} |a_{t-i}|^2 |X_{t-k}|^2 \right) = E\left(\prod_{i=0}^{k-1} a_{t-i} X_{t-k} \right) \\ &= \alpha^k E\left(X_{t-k}\right) \longrightarrow_{k\to\infty} 0 \quad \text{for } \alpha \in (-1, 1)\,. \end{aligned}$$

Part (ii) follows from Theorem 2.8 by setting $q = 0$. Its proof can be found in Section 2.6.2. □

2.6.2. Proof of Theorem 2.8

Proof.
The convergence is shown by using the p-norm $\|A\|_p$ for a matrix A, which is induced by the vector norm, that is,

$$\|A\|_p = \left(\sum_{j=1}^{K} \sum_{i=1}^{K} |a_{ij}|^p \right)^{\frac{1}{p}}.$$

For $p = 1$, we get

$$\begin{aligned}
&\lim_{k\to\infty} E\left\|\prod_{l=0}^{k-1}\widetilde{A}_{t-l}^{(+)}\right\|_1 = \lim_{k\to\infty} E\left(\sum_{i=1}^{p+q}\sum_{j=1}^{p+q}|\left(\prod_{l=0}^{k-1}\widetilde{A}_{t-l}^{(+)}\right)|_{ij}\right)\\
&= \lim_{k\to\infty} E\left[\begin{pmatrix}1 & \dots & 1\end{pmatrix}\left(|\prod_{l=0}^{k-1}\widetilde{A}_{t-l}^{(+)}|\right)\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}\right] = \lim_{k\to\infty}\begin{pmatrix}1 & \dots & 1\end{pmatrix}\prod_{l=0}^{k-1}E\left[|\widetilde{A}_{t-l}^{(+)}|\right]\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}\\
&= \lim_{k\to\infty}\begin{pmatrix}1 & \dots & 1\end{pmatrix}\prod_{l=0}^{k-1}E\left[\widetilde{A}_{t-i}\right]\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix} = \lim_{k\to\infty}\begin{pmatrix}1 & \dots & 1\end{pmatrix}\prod_{l=0}^{k-1}\widetilde{\mathcal{A}}_{|\cdot|}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}\\
&= \lim_{k\to\infty}\begin{pmatrix}1 & \dots & 1\end{pmatrix}\widetilde{\mathcal{A}}_{|\cdot|}^{k}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix} \overset{k\to\infty}{\longrightarrow} 0_{(p+q)\times(p+q)} \quad for \;\; |\alpha_{ij}| \in [0,1)\,.
\end{aligned}$$

Since the entries of $\widetilde{A}_{t-l}^{(+)}$ lie in $\{-1,0,1\}$, it follows that $|\widetilde{A}_{t-l}^{(+)}| := \widetilde{A}_{t-l} \in \{0,1\}$ and the expectation is given by $E\left(\widetilde{A}_{t-l}\right) =: \widetilde{\mathcal{A}}_{|\cdot|}$ with only positive values in $(0,1)$. Consequently, the term vanishes for $k \to \infty$ and as result we get the gbVMA(∞) representation. □

2.6.3. Proof of Theorem 2.3 and 2.9

Proof.
For a stationary gbARMA process, it follows

$$\begin{aligned}
&\gamma(k) = Cov(X_t, X_{t-k})\\
&= Cov\Bigg(\sum_{i=1}^{p} a_t^{(i)}\left[\mathbb{1}_{\{\alpha^{(i)}\geq 0\}}X_{t-i} + \mathbb{1}_{\{\alpha^{(i)}<0\}}(1 - X_{t-i})\right]\\
&+ \sum_{j=0}^{q} b_t^{(j)}\left[\mathbb{1}_{\{\beta^{(j)}\geq 0\}}e_{t-j} + \mathbb{1}_{\{\beta^{(j)}<0\}}(1 - e_{t-j})\right], X_{t-k}\Bigg)\\
&= \sum_{i=1}^{p}\alpha^{(i)}Cov(X_{t-i}, X_{t-k}) + \sum_{j=0}^{q}\beta^{(j)}Cov(e_{t-j}, X_{t-k})
\end{aligned}$$

The above equation leads in the Yule–Walker equations of Theorem 2.3 (i) since $q = 0$ and $Cov(e_t, X_{t-k}) = 0$ for $k > 0$.

We now have to consider the covariance of the error terms and the time series. Therefore, we define the variable $\varphi_k := Cov(X_t, e_{t-k})$. For $k < 0$, it follows $\varphi_k = 0$ and, for $k = 0$, we have

$$\varphi_0 = Cov(X_t, e_t) = \sum_{i=1}^{p}\alpha^{(i)}Cov(X_{t-i}, e_t) + \sum_{j=0}^{q}\beta^{(j)}Cov(e_{t-j}, e_t) = \sigma_e^2\beta^{(0)}.$$

For $k \geq 1$, we get

$$\begin{aligned}\varphi_k = Cov\left(X_t, e_{t-k}\right) &= \sum_{i=1}^{p} \alpha^{(i)} Cov\left(X_{t-i}, e_{t-k}\right) + \sum_{j=0}^{q} \beta^{(j)} Cov\left(e_{t-j}, e_{t-k}\right) \\ &= \sum_{i=1}^{p} \alpha^{(i)} \varphi_{k-i} + \beta^{(k)} \sigma_e^2.\end{aligned}$$

By defining $\phi_k := \frac{\varphi_k}{\sigma_e^2}$, the recursion of Theorem 2.9 follows.

Part (ii) of Theorem 2.3 is directly obtained by inserting the model equation and by using the property of multinomial choosing only one entry of P_t equal to one and all others to zero, such that

$$E\left((a_t^{(+,i)})^2\right) = |\alpha^{(i)}|, \quad E\left((a_t^{(-,i)})^2\right) = |\alpha^{(i)}| \mathbb{1}_{\{\alpha^{(i)}<0\}}.$$

□

2.6.4. Proof of Lemma 2.4 and 2.10

Proof.
(i) The conditional probability is an immediate consequence from the model equation and multinomial distribution of the random variables a_t and b_t.
(ii) With the independence assumption on the error terms and Part (i), the conditional probability without conditioning on the current error term is:

$$\begin{aligned}&P\left(X_t = i_0 | X_{t-1} = i_1, \ldots, e_{t-1} = j_1 \ldots\right) \\ &= \sum_{j_0=0}^{1} P\left(X_t = i_0 | X_{t-1} = i_1, \ldots, e_t = j_0, \ldots\right) \\ &\qquad\qquad P\left(e_t = j_0 | X_{t-1} = i_1, \ldots, e_{t-1} = j_1, \ldots\right) \\ &= \sum_{l=1}^{p} |\alpha^{(l)}| \left[\mathbb{1}_{\{\alpha^{(l)} \geq 0\}} \delta_{i_0 i_l} + \mathbb{1}_{\{\alpha^{(l)}<0\}} \delta_{i_0 (1-i_l)}\right] \\ &\qquad + \sum_{k=1}^{q} |\beta^{(k)}| \left[\mathbb{1}_{\{\beta^{(k)} \geq 0\}} \delta_{i_0 j_k} + \mathbb{1}_{\{\beta^{(k)}<0\}} \delta_{i_0 (1-j_k)}\right] + \beta^{(0)} p_{i_0}.\end{aligned}$$

(iii) Consider the probability that the time series is in state $j \in \{0, 1\}$ at time point t and note that $P\left(e_t = 1 - j\right) = 1 - P\left(e_t = j\right), \quad P\left(X_t = 1 - j\right) = 1 - P\left(X_t = j\right)$.

$$\begin{aligned}P\left(X_t = j\right) &= \sum_{i=1}^{p} |\alpha^{(i)}| \left[\mathbb{1}_{\{\alpha^{(i)} \geq 0\}} P\left(X_{t-i} = j\right) + \mathbb{1}_{\{\alpha^{(i)}<0\}} P\left(X_{t-i} = 1 - j\right)\right] \\ &\quad + \sum_{l=1}^{q} |\beta^{(l)}| \left[\mathbb{1}_{\{\beta^{(l)} \geq 0\}} P\left(e_{t-l} = j\right) + \mathbb{1}_{\{\beta^{(l)}<0\}} P\left(e_{t-l} = 1 - j\right)\right] \\ &\quad + \beta^{(0)} P\left(e_t = j\right) \\ &= \sum_{i=1}^{p} |\alpha^{(i)}| \left[\mathbb{1}_{\{\alpha^{(i)} \geq 0\}} P\left(X_t = j\right) + \mathbb{1}_{\{\alpha^{(i)}<0\}} P\left(X_t = 1 - j\right)\right]\end{aligned}$$

$$
\begin{aligned}
&+\sum_{l=1}^{q}|\beta^{(l)}|\left[\mathbb{1}_{\{\beta^{(l)}\geq 0\}}P\left(e_t=j\right)+\mathbb{1}_{\{\beta^{(l)}<0\}}P\left(e_t=1-j\right)\right]\\
&+\beta^{(0)}P\left(e_t=j\right)\\
&=\left(\sum_{i=1}^{p}|\alpha^{(i)}|\left[\mathbb{1}_{\{\alpha^{(i)}\geq 0\}}-\mathbb{1}_{\{\alpha^{(i)}<0\}}\right]\right)P\left(X_t=j\right)+\sum_{l=1}^{q}|\beta^{(l)}|\mathbb{1}_{\{\beta^{(l)}<0\}}\\
&\quad+\left(\sum_{l=1}^{q}|\beta^{(l)}|\left[\mathbb{1}_{\{\beta^{(l)}\geq 0\}}-\mathbb{1}_{\{\beta^{(l)}<0\}}\right]\right)p_j+\beta^{(0)}p_j+\sum_{i=1}^{p}|\alpha^{(i)}|\mathbb{1}_{\{\alpha^{(i)}<0\}}.
\end{aligned}
$$

Then, by rearranging the terms on the last right side, we get

$$
\begin{aligned}
P\left(X_t=j\right)=&\frac{\left[\sum_{l=1}^{q}|\beta^{(l)}|\left(\mathbb{1}_{\{\beta^{(l)}\geq 0\}}-\mathbb{1}_{\{\beta^{(l)}<0\}}\right)\right]+\beta^{(0)}}{\left(1-\sum_{i=1}^{p}\alpha^{(i)}\right)}p_j\\
&+\frac{\sum_{i=1}^{p}|\alpha^{(i)}|\mathbb{1}_{\{\alpha^{(i)}<0\}}+\sum_{l=0}^{q}|\beta^{(l)}|\mathbb{1}_{\{\beta^{(l)}<0\}}}{\left(1-\sum_{i=1}^{p}\alpha^{(i)}\right)}.
\end{aligned}
$$

(iv) Consider the joint probability of the error term and time series at time point t. We get

$$
\begin{aligned}
&P\left(X_t=i_0,e_t=j_0\right)\\
&=\sum_{i=1}^{p}|\alpha^{(i)}|\mathbb{1}_{\{\alpha^{(i)}\geq 0\}}P\left(X_{t-i}=i_0,e_t=j_0\right)+\sum_{i=1}^{p}|\alpha^{(i)}|\mathbb{1}_{\{\alpha^{(i)}<0\}}P\left(X_{t-i}=1-i_0,e_t=j_0\right)\\
&\quad+\sum_{l=1}^{q}|\beta^{(l)}|\mathbb{1}_{\{\beta^{(l)}\geq 0\}}P\left(e_{t-l}=i_0,e_t=j_0\right)+\sum_{l=1}^{q}|\beta^{(l)}|\mathbb{1}_{\{\beta^{(l)}<0\}}P\left(e_{t-l}=i_0,e_t=j_0\right)\\
&\quad+\beta^{(0)}P\left(e_t=i_0,e_t=j_0\right)\\
&=\sum_{i=1}^{p}|\alpha^{(i)}|\mathbb{1}_{\{\alpha^{(i)}\geq 0\}}P\left(X_{t-i}=i_0\right)p_{j_0}+\sum_{i=1}^{p}|\alpha^{(i)}|\mathbb{1}_{\{\alpha^{(i)}<0\}}P\left(X_{t-i}=1-i_0\right)p_{j_0}\\
&\qquad+\sum_{l=1}^{q}|\beta^{(l)}|\mathbb{1}_{\{\beta^{(l)}\geq 0\}}P\left(e_{t-l}=i_0\right)p_{j_0}+\sum_{l=1}^{q}|\beta^{(l)}|\mathbb{1}_{\{\beta^{(l)}<0\}}P\left(e_{t-l}=1-i_0\right)p_{j_0}\\
&\qquad+\beta^{(0)}P\left(e_t=i_0,e_t=j_0\right).
\end{aligned}
$$

By inserting Part *(iii)* into the equation above, we get

$$
\begin{aligned}
&\sum_{i=1}^{p}|\alpha^{(i)}|p_{j_0}\left[\mathbb{1}_{\{\alpha^{(i)}\geq 0\}}-\mathbb{1}_{\{\alpha^{(i)}<0\}}\right]\left(\frac{1-\sum_{i=1}^{p}|\alpha^{(i)}|-\sum_{l=0}^{q}|\beta^{(l)}|\mathbb{1}_{\{\beta^{(l)}<0\}}}{\left(1-\left[\sum_{i=1}^{p}|\alpha^{(i)}|\left(\mathbb{1}_{\{\alpha^{(i)}\geq 0\}}-\mathbb{1}_{\{\alpha^{(i)}<0\}}\right)\right]\right)}p_{i_0}\right.\\
&\qquad\left.+\frac{\sum_{i=1}^{p}|\alpha^{(i)}|\mathbb{1}_{\{\alpha^{(i)}<0\}}+\sum_{l=0}^{q}|\beta^{(l)}|\mathbb{1}_{\{\beta^{(l)}<0\}}}{\left(1-\left[\sum_{i=1}^{p}|\alpha^{(i)}|\left(\mathbb{1}_{\{\alpha^{(i)}\geq 0\}}-\mathbb{1}_{\{\alpha^{(i)}<0\}}\right)\right]\right)}\right)+\beta^{(0)}p_{j_0}\delta_{i_0,j_0}\\
&\qquad+\sum_{i=1}^{p}|\alpha^{(i)}|\mathbb{1}_{\{\alpha^{(i)}<0\}}+\sum_{l=1}^{q}|\beta^{(l)}|p_{j_0}\left[\mathbb{1}_{\{\beta^{(l)}\geq 0\}}p_{i_0}+\mathbb{1}_{\{\beta^{(l)}<0\}}\left(1-p_{i_0}\right)\right].
\end{aligned}
$$

Using the properties of the parameters, the joint distribution of the time series and error term is given by

$$
\begin{aligned}
&P\left(X_t = i_0, e_t = j_0\right) \\
&= p_{i_0} p_{j_0} \left[\frac{\sum_{l=1}^{q} |\beta^{(l)}| \mathbb{1}_{\{\beta^{(l)}<0\}}}{1 - \sum_{i=1}^{p} |\alpha^{(i)}| \left[\mathbb{1}_{\{\alpha^{(i)} \geq 0\}} - \mathbb{1}_{\{\alpha^{(i)}<0\}} \right]} - \sum_{l=1}^{q} |\beta^{(l)}| \mathbb{1}_{\{\beta<0\}} \right] + \beta^{(0)} \delta_{i_0 j_0} \\
&\quad + \sum_{l=1}^{q} |\beta^{(l)}| \mathbb{1}_{\{\beta^{(l)}<0\}} \left(1 - p_{j_0} + \frac{p_{j_0}}{1 - \sum_{i=1}^{p} |\alpha^{(i)}| \left[\mathbb{1}_{\{\alpha^{(i)} \geq 0\}} - \mathbb{1}_{\{\alpha^{(i)}<0\}} \right]} \right) \\
&\quad + \sum_{i=1}^{p} |\alpha^{(i)}| \mathbb{1}_{\{\alpha^{(i)}<0\}} \left(1 - p_{j_0} + \frac{p_{j_0}}{1 - \sum_{i=1}^{p} |\alpha^{(i)}| \left[\mathbb{1}_{\{\alpha^{(i)} \geq 0\}} - \mathbb{1}_{\{\alpha^{(i)}<0\}} \right]} \right).
\end{aligned}
$$

□

3. Generalized Binary Vector ARMA Processes

Based on: Jentsch, C. and Reichmann, L.
Generalized binary vector ARMA processes *- Working Paper*

3.1. Introduction

Categorical data is collected in many fields of applications. When such data is observed over time, serial dependence is often present that has to be taken into account e.g. for modeling purposes or for statistical inference. Hence, the statistical research focusing on such data structures evolved considerably over the last years. With the collection of huge amounts of data nowadays, this leads particularly to a growing interest in statistical methods for the analysis of multivariate categorical time series. As an important special case, multivariate binary time series, that correspond to categorical time series data with just two categories, occur in many different contexts such as agriculture, biology, economy, engineering, environmetrics, genetics, geography, geology, medical science, natural language processing or sports; see e.g. Jentsch and Reichmann (2019) for some univariate examples. For instance, recent related literature addresses the detection of dependent Bernoulli sequences in Ritzwoller and Romano (2020) or the efficient generation of high-dimensional binary data with specified correlation structures in Jiang et al. (2020). Often, binary time series are obtained from binarization of observed real-valued data, when e.g. the interest is, whether some event occurs (or not) or a certain threshold is crossed (or not) instead of the actual value. Although simplified, this transformation will generally contain a great amount of the information and the dynamics of the original data.

Multivariate binary time series obtained from a suitable thresholding procedure are for instance of much interest in economics, where periods of recession and of economic growth (no-recession) are considered; see e.g. Bellégo and Ferrara (2009) and Startz (2008) on forecasting recessions in the Euro area and in the United States, respectively. Considering jointly such recession time series of several countries, a multivariate analysis allows to study not only the serial, but also the cross-sectional dependence in the data. In turn, this allows to investigate the spillover effects between several countries, that is, how a recession in one country will affect the economy in other countries in the future. In Figure 3.1, we show quarterly time series indicating periods of recessions and of economic growth for the G7 countries Canada, France, Germany, Italy, Japan, United Kingdom and the United States from Q2/1960 - Q1/2017.

In signal processing, large numbers of nodes, i.e. inexpensive sensors, are employed to make binary decisions whenever a signal is above a certain threshold or not, see e.g. Cheng et al. (2013). Hence, multiple two-state time series with states 'detection' and 'no detection' are observed. In such applications, binarization of the original signal is particularly beneficial as binary data is inexpensive to store.

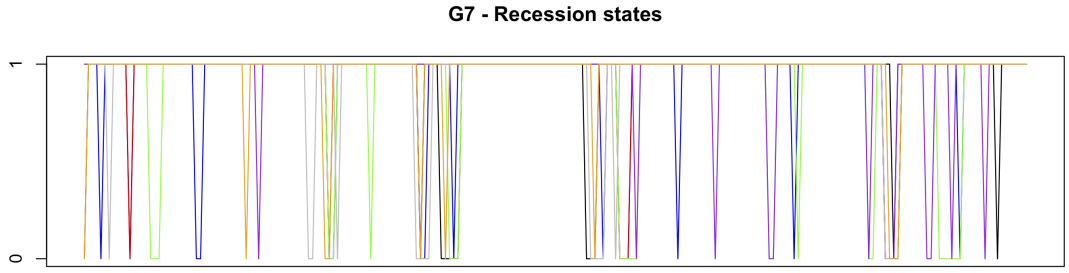


Figure 3.1.: Quarterly time series indicating periods of recessions ("0") and of economic growth ("1") for all G7 countries Canada (black), France (red), Germany (blue), Italy (green), Japan (purple), United Kingdom (grey) and United States (orange) from Q2/1960 - Q1/2017.

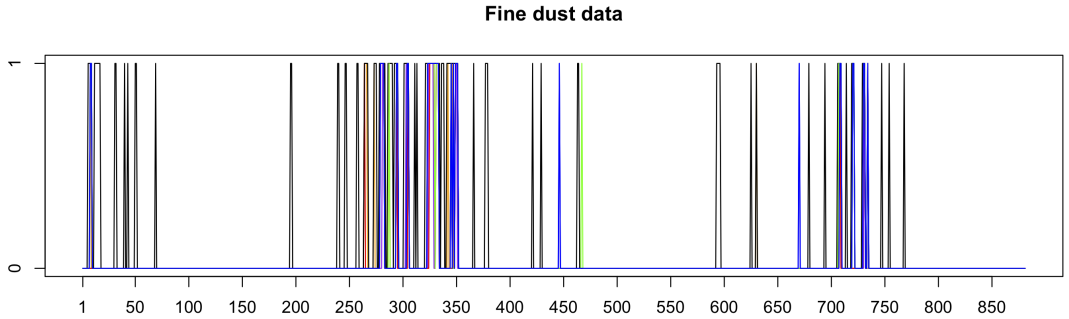


Figure 3.2.: Daily time series indicating a fine dust alarm ("1") and no fine dust alarm ("0") at six monitoring stations: Arnulf-Klett Platz (orange), Bad Cannstatt (purple), Hauptstätter Strasse (blue), Hohenheimer Strasse (green), Neckartor (black), Stadtgarten (red) in Stuttgart, Germany from 03/01/2016 - 07/31/2018.

In recent years, there is increasing interest in air pollution in European cities and metropolitan areas. The EU established the *European emission standards*, which include limits for particulates in the air. Whenever the amount of PM_{10} (coarse particles with a diameter between 2.5 and 10 micrometers, 'fine dust') exceeds the threshold of 50 $\mu g/m^3$ at a monitoring station, this causes a 'fine dust alarm'. Hence, for each monitoring station, this results in binary sequence with states 'exceedance' and 'no exceedance'. In fact, the current public discourse centers to a large extent around whether the threshold is exceeded or not, and less about the actual amount of fine dust measured. Typically, several monitoring stations in one city allow for a *joint* analysis of the fine dust pollution. In Figure 3.2, we show the recorded fine dust alarms at six monitoring stations in the urban area of Stuttgart, Germany from 03/01/2016 - 07/31/2018. The occurrences of alarms tend to cluster, but a closer inspection reveals that the station Neckartor often shows an alarm before the others. Hence, a multivariate analysis of this pattern might be helpful to allow for an improved prediction of future exceedances. In Section 3.4, we discuss this data set in more detail.

Typically, Markovian models are used to describe the dependence structure of categorical time series, see e.g. Kedem (1980). Such models are very flexible and allow to capture a broad range of serial dependence, but the number of parameters grows

exponentially with the order of the Markov model. As indicated by McKenzie (2003) already in the univariate case, this likely leads to over-parametrization. For a K-dimensional multivariate binary time series, this problem is even much more intricate as the fitting of an unrestricted Markov model of order p requires the estimation of 2^{Kp} parameters. Hence, Markov models are not feasible, when the time series dimension or the model order become large.

3.1.1. The univariate case: NDARMA vs. gbARMA

In the univariate case, to avoid the estimation of a large number of parameters, Jacobs and Lewis (1983) proposed the class of (New) Discrete AutoRegressive Moving-Average (NDARMA) models for categorical time series. To make sure that the process $(X_t, t \in \mathbb{Z})$ takes only values contained in a *discrete* state space $\mathcal{V}$, their idea is to choose X_t randomly either from the past values of the time series $X_{t-1}, \ldots, X_{t-p}$ or from one of the innovations $e_t, e_{t-1}, \ldots, e_{t-q}$ with certain probabilities, respectively. This random selection mechanism is described by independent and identically distributed (i.i.d.) random vectors $(P_t, t \in \mathbb{Z})$ with

$$P_t := \left[a_t^{(1)}, \ldots, a_t^{(p)}, b_t^{(0)}, \ldots, b_t^{(q)}\right] \sim Mult\left(1; \mathcal{P}\right), \quad t \in \mathbb{Z}, \tag{3.1}$$

where $Mult\left(1; \mathcal{P}\right)$ denotes the multinomial distribution with parameter 1 and probability vector $\mathcal{P} := \left[\alpha^{(1)}, \ldots, \alpha^{(p)}, \beta^{(0)}, \ldots, \beta^{(q)}\right]$ with $\alpha^{(1)}, \ldots, \alpha^{(p)} \in [0, 1)$, $\beta^{(0)} \in (0, 1]$ and $\beta^{(1)}, \ldots, \beta^{(q)} \in [0, 1)$ such that $\sum_{i=1}^{p} \alpha^{(i)} + \sum_{j=0}^{q} \beta^{(j)} = 1$. Then, the NDARMA(p,q) model equation is given by

$$X_t = \sum_{i=1}^{p} a_t^{(i)} X_{t-i} + \sum_{j=0}^{q} b_t^{(j)} e_{t-j}, \quad t \in \mathbb{Z}, \tag{3.2}$$

where $(e_t, t \in \mathbb{Z})$ is an i.i.d. innovation process taking values in the state space $\mathcal{V}$.

NDARMA models are contained as special cases in the broad class of Markov models, but are considerably more parsimonious and still nicely interpretable due to their ARMA-type structure. In this spirit, Weiß and Göb (2008) showed that Yule-Walker-type equations hold and Weiß (2009a) discussed the connection of NDARMA models to general Markov chains.

It is important to note that the probability vector $\mathcal{P}$ of the multinomial distribution in (3.1) contains the NDARMA model parameters, which are naturally restricted to satisfy two conditions: all entries of $\mathcal{P}$ have to lie in the unit interval and they have to sum up to one. Hence, in contrast to general Markov chains, NDARMA models are particularly restricted to capture exclusively *non-negative* serial dependence. To address this lacking flexibility of the NDARMA model class, Jentsch and Reichmann (2019) proposed a simple and straightforward extension of the original idea of Jacobs and Lewis (1983) that allows also to capture negative serial dependence in univariate binary time series. In the resulting generalized binary ARMA (gbARMA) model class, in contrast to NDARMA models, the parameters $\alpha^{(i)}$ and $\beta^{(j)}$ are allowed to be either positive or negative. Precisely, gbARMA models allow for $\alpha^{(1)}, \ldots, \alpha^{(p)} \in (-1, 1)$ and $\beta^{(1)}, \ldots, \beta^{(q)} \in (-1, 1)$, with $\beta^{(0)} \in (0, 1]$, such that $\sum_{i=1}^{p} |\alpha^{(i)}| + \beta^{(0)} + \sum_{j=1}^{q} |\beta^{(j)}| = 1$. The parameter vector $\mathcal{P} = \left[\alpha^{(1)}, \ldots, \alpha^{(p)}, \beta^{(0)}, \ldots, \beta^{(q)}\right]$ has to be modified to contain valid probabilities in $[0, 1]$ for the selection mechanism. We define

$$\mathcal{P}_{|\cdot|} := \left[|\alpha^{(1)}|, \ldots, |\alpha^{(p)}|, \beta^{(0)}, |\beta^{(1)}|, \ldots, |\beta^{(q)}|\right]. \tag{3.3}$$

As in NDARMA models, the random selection mechanism for gbARMA models is again described by i.i.d. multinomial random vectors $(P_t, t \in \mathbb{Z})$ with

$$P_t := \left(a_t^{(1)}, \ldots, a_t^{(p)}, b_t^{(0)}, \ldots, b_t^{(q)}\right) \sim Mult\left(1; \mathcal{P}_{|\cdot|}\right), \quad t \in \mathbb{Z}.$$

Then, the gbARMA(p,q) process $(X_t, t \in \mathbb{Z})$ follows the model equation

$$X_t = \sum_{i=1}^{p} \left[a_t^{(+,i)} X_{t-i} + a_t^{(-,i)}\right] + b_t^{(0)} e_t + \sum_{j=1}^{q} \left[b_t^{(+,j)} e_{t-j} + b_t^{(-,j)}\right], \quad t \in \mathbb{Z}, \tag{3.4}$$

where $(e_t, t \in \mathbb{Z})$ is an i.i.d. innovation process taking values in $\{0, 1\}$. Here, we set $a_t^{(+,i)} := a_t^{(i)} \left(\mathbb{1}_{\{\alpha^{(i)} \geq 0\}} - \mathbb{1}_{\{\alpha^{(i)} < 0\}}\right)$ and $a_t^{(-,i)} := a_t^{(i)} \mathbb{1}_{\{\alpha^{(i)} < 0\}}$, $i = 1, \ldots, p$, with analogous definitions for $b_t^{(+,j)}$ and $b_t^{(-,j)}$, $j = 1, \ldots, q$. A detailed description of the gbARMA model class and its properties can be found in Jentsch and Reichmann (2019).

3.1.2. An example: NDAR(1) vs. gbAR(1)

In a nutshell, the benefit of a gbARMA model in comparison to an NDARMA model for binary data, is that it allows to pick *systematically* the opposite value of a predecessor if the corresponding model parameter is negative. To illustrate this, let us consider the simplest case of a gbAR(1) model with $\mathcal{P} = [\alpha, \beta]$ following

$$X_t = a_t^{(+)} X_{t-1} + a_t^{(-)} + b_t e_t, \quad t \in \mathbb{Z}, \tag{3.5}$$

where $a_t^{(+)} := a_t \left(\mathbb{1}_{\{\alpha \geq 0\}} - \mathbb{1}_{\{\alpha < 0\}}\right)$ and $a_t^{(-)} := a_t \mathbb{1}_{\{\alpha < 0\}}$ such that $P_t := (a_t, b_t) \sim Mult\left(1; \mathcal{P}_{|\cdot|}\right)$ with $\mathcal{P}_{|\cdot|} = [|\alpha|, \beta]$ and $|\alpha| + \beta = 1$. Equation (3.5) can be re-written to get

$$X_t = \begin{cases} a_t X_{t-1} + b_t e_t, & [a_t, b_t] \sim Mult\left(1; \alpha, \beta\right), & \alpha \in [0, 1) \\ a_t \left(1 - X_{t-1}\right) + b_t e_t, & [a_t, b_t] \sim Mult\left(1; |\alpha|, \beta\right), & \alpha \in (-1, 0) \end{cases}. \tag{3.6}$$

Hence, depending on the sign of the parameter $\alpha \in (-1, 1)$, either the predecessor X_{t-1} or its opposite value $1 - X_{t-1}$ can be systematically picked by the random selection mechanism. Actually, the first equation in (3.6), that allows for $\alpha \in [0, 1)$, corresponds to an NDAR(1) model with a binary state space, which is contained in a gbAR(1) model as a special case. Hence, the gbAR(1) model is more flexible and allows particularly for an unrestricted first-lag autocorrelation in $(-1, 1)$.

In view of the operations $a_t X_{t-1}$ and $a_t \left(1 - X_{t-1}\right)$ in (3.6), that take the value of the time series from the time point before or its opposite value, respectively, it seems plausible to combine them and allow both at the same time. That is, we could think of a model equation of the form

$$X_t = a_{t,=} X_{t-1} + a_{t,\neq} \left(1 - X_{t-1}\right) + b_{t,=} \epsilon_t, \quad t \in \mathbb{Z}, \tag{3.7}$$

where $([a_{t,=}, a_{t,\neq}, b_{t,=}], t \in \mathbb{Z})$ are i.i.d. with $[a_{t,=}, a_{t,\neq}, b_{t,=}] \sim Mult(1; \alpha_=, \alpha_\neq, \beta_=)$ such that $\alpha_=, \alpha_\neq \in [0, 1]$, $\beta_= \in (0, 1]$ and $\alpha_= + \alpha_\neq + \beta_= = 1$ as well as $E(\epsilon_t) = \mu_\epsilon \in (0, 1)$. However, the model (3.7) is *not identified* as it is indistinguishable from the gbAR(1) model (3.6) with

$$\alpha = \alpha_= - \alpha_\neq, \quad \beta = 1 - |\alpha_= - \alpha_\neq| \tag{3.8}$$

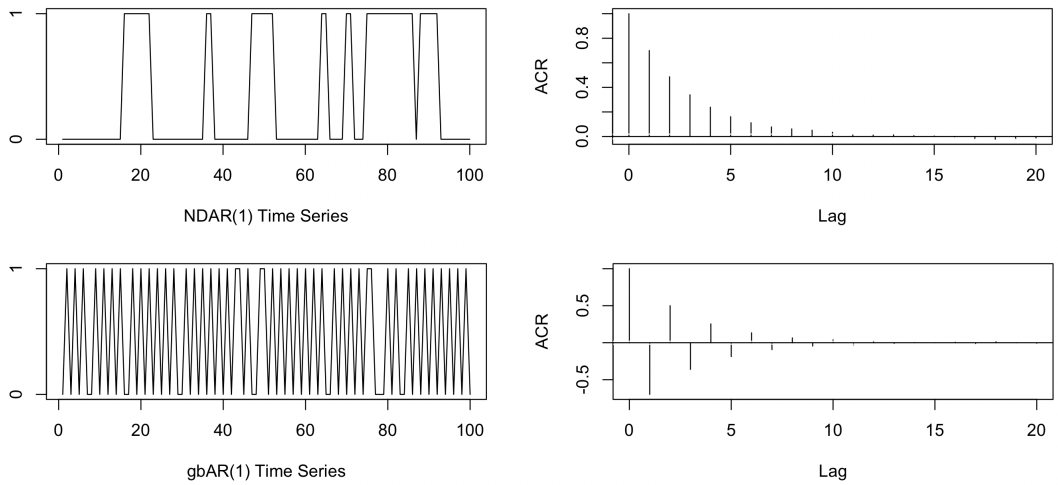


Figure 3.3.: NDAR(1) vs. gbAR(1): Realizations and ACF of an NDAR(1) process (that is, of a gbAR(1) process with positive α in the first line of (3.6)) (upper panel) with $\mathcal{P} = [0.7, 0.3]$ and of a gbAR(1) process (3.5) (lower panel) with $\mathcal{P} = [-0.7, 0.3]$. In both cases, we used $P(e_t = 1) = 0.5$.

and

$$\mu_e = \frac{\alpha_{\neq} \mathbb{1}_{\{\alpha_= - \alpha_{\neq} \geq 0\}} + \alpha_= \mathbb{1}_{\{\alpha_= - \alpha_{\neq} < 0\}} + \beta_= \mu_\epsilon}{1 - |\alpha_= - \alpha_{\neq}|}. \tag{3.9}$$

We refer to the appendix for a proof of (3.8) and (3.9).

In Figure 3.3, we show realizations and corresponding autocorrelation functions (ACFs) of a gbAR(1) process with positive parameter $\alpha \in [0, 1)$ (i.e. an NDAR(1) process) and a gbAR(1) process with negative parameter $\alpha \in (-1, 0)$. Whereas the NDAR(1) process with positive α shows long runs of the same value and a non-negative ACF, the gbAR(1) process with negative α shows an oscillating pattern and an alternating ACF.

3.1.3. Towards a multivariate analysis: a bivariate gbVAR(1)

In the case, when more than just one binary time series is observed, as e.g. for the G7 recession data in Figure 3.1 or the fine dust alarm data in Figure 3.2, a multivariate (i.e. joint) analysis is desirable.

For categorical time series data, Möller and Weiß (2020) proposed a multivariate extension of the NDARMA class with (non-negative) *scalar* model parameters that control the multinomial selection mechanism. This approach restricts the flexibility of the resulting Generalized DARMA (GDARMA) class to model joint dependence to some large extent. Instead, GDARMA models make use of a variation function applied to lagged observations and innovations to increase the entry-wise variation over time. However, the proposed variation function does not modify the past time series values in a systematical way and the resulting process is not suitable to capture negative dependence structure.

To achieve more model flexibility, let us first consider the case of two *independent* gbAR(1) processes $(X_t, t \in \mathbb{Z})$ and $(Y_t, t \in \mathbb{Z})$. By stacking them, we get a bivariate

process

$$\begin{aligned}\begin{pmatrix} X_t \\ Y_t \end{pmatrix} &= \begin{pmatrix} a_{t,X}^{(+)} X_{t-1} + a_{t,X}^{(-)} + b_{t,X} e_{t,X} \\ a_{t,Y}^{(+)} Y_{t-1} + a_{t,Y}^{(-)} + b_{t,Y} e_{t,Y} \end{pmatrix} \\ &= \begin{pmatrix} a_{t,X}^{(+)} & 0 \\ 0 & a_{t,Y}^{(+)} \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} + \begin{pmatrix} a_{t,X}^{(-)} & 0 \\ 0 & a_{t,Y}^{(-)} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} b_{t,X} & 0 \\ 0 & b_{t,Y} \end{pmatrix} \begin{pmatrix} e_{t,X} \\ e_{t,Y} \end{pmatrix}. \end{aligned} \tag{3.10}$$

However, due to the diagonal structure of the (random) coefficient matrices, such a model in (3.10) is not yet sufficient to study the cross-sectional dependence between two binary time series. Naturally, this can be achieved by allowing the off-diagonal elements of the coefficient matrices in (3.10) to be non-zero. This leads to the bivariate gbVAR(1) model

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} = \begin{pmatrix} a_{t,11}^{(+)} & a_{t,12}^{(+)} \\ a_{t,21}^{(+)} & a_{t,22}^{(+)} \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \end{pmatrix} + \begin{pmatrix} a_{t,11}^{(-)} & a_{t,12}^{(-)} \\ a_{t,21}^{(-)} & a_{t,22}^{(-)} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} b_{t,11} & 0 \\ 0 & b_{t,22} \end{pmatrix} \begin{pmatrix} e_{t,1} \\ e_{t,2} \end{pmatrix}, \tag{3.11}$$

which can be compactly written as

$$X_t = A_t^{(+)} X_{t-1} + A_t^{(-)} \mathbb{1}_2 + B_t e_t, \tag{3.12}$$

where $(e_t, t \in \mathbb{Z})$ is an i.i.d. innovation process taking values in $\{0,1\}^2$. Note that $Cov(e_t) = \Sigma_e$ is allowed to be non-diagonal, whereas B_t is imposed to be diagonal for identification reasons; compare also Remark 3.13. For a comprehensive discussion of multivariate Bernoulli distributions allowing for dependence leading to non-diagonal Σ_e, we refer to Dai et al. (2013).

The model parameters of the process (3.12) are summarized in the matrix $\mathcal{P} := [\mathcal{A}, \mathcal{B}]$ with $\mathcal{A} := (\alpha_{kl})_{k,l=1,2}$ and $\mathcal{B} := diag(\beta_{11}, \beta_{22})$ and $\mathcal{P}_{|\cdot|} := \left[\mathcal{A}_{|\cdot|}, \mathcal{B}\right]$ with $\mathcal{A}_{|\cdot|} := (|\alpha_{kl}|)_{k,l=1,2}$, where $\alpha_{kl} \in (-1,1)$, $\beta_{kk} \in (0,1]$ for all $k,l = 1,2$, such that $|\alpha_{k1}| + |\alpha_{k2}| + \beta_{kk} = 1$ for $k = 1,2$.

The natural approach is to adopt row-wise the random (multinomial) selection mechanism leading to mutually *independent* i.i.d. vector-valued processes $(P_{t,1\bullet}, t \in \mathbb{Z})$ and $(P_{t,2\bullet}, t \in \mathbb{Z})$, where

$$P_{t,1\bullet} := [a_{t,11}, a_{t,12}, b_{t,11}, 0] \sim Mult\left(1; \mathcal{P}_{|\cdot|,1\bullet}\right), \tag{3.13}$$

$$P_{t,2\bullet} := [a_{t,21}, a_{t,22}, 0, b_{t,22}] \sim Mult\left(1; \mathcal{P}_{|\cdot|,2\bullet}\right). \tag{3.14}$$

Here, $\mathcal{P}_{|\cdot|}$ is defined to contain entry-wise the absolute values of $\mathcal{P}$ with row-sums equal to one such that $\mathcal{P}_{|\cdot|,k\bullet}$ become valid arguments for the multinomial distributions in (3.13) and (3.14). Note also the desirable redundancy in the above notation obtained by including the off-diagonal zeros of B_t in $P_{t,k\bullet}$, but this allows to use the whole rows $\mathcal{P}_{|\cdot|,k\bullet}$ as arguments of the multinomial distributions.

3.1.4. Outline

In the spirit of the gbVAR(1) model (3.11) as a natural extension of the univariate gbAR(1) model in (3.5), we provide a full description and investigation of the corresponding generalized binary vector AR (gbVAR) model class in this paper. In

Section 3.2, we define generalized binary VAR processes of order $p \in \mathbb{N}$ and derive general stochastic properties including formulas for the mean, stationarity conditions, moving-average representations, Yule-Walker equations and transition probabilities. We discuss possible extensions and identification issues and prove φ - and ψ - mixing properties, that allow for a general approach to derive asymptotic theory. Further, we address parameter estimation in gbVAR models based on Yule-Walker estimators. In Section 3.3, we examine the finite sample performance of these parameter estimators by means of different criteria in simulations and we propose a parametric bootstrap method to construct confidence intervals. For illustration, we use gbVAR models to analyze a six-dimensional binary time series that indicates fine dust alarms in the urban area of Stuttgart, Germany, in Section 3.4. We conclude the extension on gbVAR models in Section 3.5 and give an extension to gbVARMA processes by adding an moving-average part in Section 3.6 All proofs, additional simulation results are deferred to Section 3.7.

3.2. The gbVAR Model Class

In Section 3.2.1, we define gbVAR processes as multivariate extensions of (univariate) gbAR models as introduced by Jentsch and Reichmann (2019). The definition naturally extends the bivariate gbVAR(1) model (3.11) to arbitrary order $p \in \mathbb{N}$ and dimension $K \in \mathbb{N}$. Stochastic properties of gbVAR models including stationarity conditions, formulas for the mean, Yule-Walker equations and transition probabilities are derived in Section 3.2.2.

3.2.1. gbVAR models

For a K-dimensional binary time series $(X_t, t \in \mathbb{Z})$, let the matrix

$$\mathcal{P} := \left[\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(p)}, \mathcal{B}\right] \tag{3.15}$$

contain the autoregressive coefficient matrices $\mathcal{A}^{(i)} = (\alpha_{kl}^{(i)})_{k,l=1,\ldots,K}$, $i = 1, \ldots, p$ and $\mathcal{B} = diag(\beta_{11}, \ldots, \beta_{KK})$ of a gbVAR(p) model. As the gbVAR model allows for $\alpha_{kl}^{(i)} \in (-1, 1)$, the entries of $\mathcal{P}$ have to satisfy $\sum_{i=1}^{p} \sum_{l=1}^{K} |\alpha_{kl}^{(i)}| + \beta_{kk} = 1$, $k = 1, \ldots, K$, and $\mathcal{P}$ has to be modified to serve as a parameter matrix containing (row-wise) valid probabilities of multinomial distributions. This is achieved by taking entry-wise absolute values in $\mathcal{P}$ and we define

$$\begin{aligned} \mathcal{P}_{|\cdot|} &:= \left[\left(|\alpha_{kl}^{(1)}|\right)_{k,l=1,\ldots,K}, \ldots, \left(|\alpha_{kl}^{(p)}|\right)_{k,l=1,\ldots,K}, diag(\beta_{11}, \ldots, \beta_{KK})\right] \\ &=: \left[\mathcal{A}_{|\cdot|}^{(1)}, \ldots, \mathcal{A}_{|\cdot|}^{(p)}, \mathcal{B}\right]. \end{aligned} \tag{3.16}$$

These prerequisites enable us to give the definition of the generalized binary vector AR model of order $p \in \mathbb{N}$.

Definition 3.1 (gbVAR(p) processes)
Let $(X_t, t \in \mathbb{Z})$ be a stationary K-dimensional process taking values in $\{0,1\}^K$. Let $(e_t, t \in \mathbb{Z})$ be an i.i.d. K-dimensional binary innovation process, such that e_t is independent of $(X_s, s < t)$ with mean vector $\mu_e = (\mu_{e,1}, \ldots, \mu_{e,K})' = E(e_t)$, where $\mu_{e,i} = P(e_{t,i} = 1)$, $i = 1, \ldots, K$, and variance-covariance matrix $\Sigma_e = (\sigma_{e,kl})_{k,l=1,\ldots,K} =$

$Cov(e_t)$, where $\sigma_{e,ii} = \mu_{e,i}(1-\mu_{e,i})$. Let $\mathcal{P}$ be the parameter matrix as in (3.15) with $\mathcal{P}_{|\cdot|}$ as in (3.16) such that $\mathcal{P}_{|\cdot|}\mathbb{1}_{K(p+1)} = \mathbb{1}_K$, where $\mathbb{1}_M = (1,\dots,1)'$ denotes the M-dimensional vector of ones, and $rk(\Sigma_e) \geq rk(\mathcal{B}\Sigma_e)$. Further, let $(P_t, t\in\mathbb{Z})$ with $P_t = [A_t^{(1)},\dots,A_t^{(p)}, B_t]$ be a $K \times K(p+1)$-dimensional i.i.d. process with mutually independent rows $(P_{t,k\bullet}, t\in\mathbb{Z})$, $k=1,\dots,K$, such that

$$P_{t,k\bullet} := \left[a_{t,k\bullet}^{(1)},\dots,a_{t,k\bullet}^{(p)}, b_{t,k\bullet}\right] \sim Mult\left(1; \mathcal{P}_{|\cdot|,k\bullet}\right),$$

which are independent of $(e_t, t\in\mathbb{Z})$ and $(X_s, s<t)$. Here, we set $A_t^{(i)} = (a_{t,kl}^{(i)})_{k,l=1,\dots,K}$ and $B_t = (b_{t,kl})_{k,l=1,\dots,K} = diag(b_{t,11},\dots,b_{t,KK})$ with $a_{t,k\bullet}^{(i)} = (a_{t,k1}^{(i)},\dots,a_{t,kK}^{(i)})$ for $i=1,\dots,p$ and $b_{t,k\bullet}$ defined similarly.

Then the process $(X_t, t\in\mathbb{Z})$ is said to be a generalized binary vector AR process of order p (gbVAR(p)), if it follows the recursion

$$X_t = \sum_{i=1}^{p}\left[A_t^{(+,i)}X_{t-i} + A_t^{(-,i)}\mathbb{1}_K\right] + B_t e_t, \quad t\in\mathbb{Z}, \tag{3.17}$$

with

$$A_t^{(+,i)} := \left\{\begin{array}{ll} a_{t,kl}^{(i)}, & \alpha_{kl}^{(i)} \geq 0 \\ -a_{t,kl}^{(i)}, & \alpha_{kl}^{(i)} < 0 \end{array}\right\}_{k,l=1,\dots,K} = \left(a_{t,kl}^{(i)}\left(\mathbb{1}_{\{\alpha_{kl}^{(i)}\geq 0\}} - \mathbb{1}_{\{\alpha_{kl}^{(i)}<0\}}\right)\right)_{k,l=1,\dots,K},$$

$$A_t^{(-,i)} := \left\{\begin{array}{ll} 0, & \alpha_{kl}^{(i)} \geq 0 \\ a_{t,kl}^{(i)}, & \alpha_{kl}^{(i)} < 0 \end{array}\right\}_{k,l=1,\dots,K} = \left(a_{t,kl}^{(i)}\mathbb{1}_{\{\alpha_{kl}^{(i)}<0\}}\right)_{k,l=1,\dots,K}$$

for $i=1,\dots,p$.

Note that it is possible to rewrite the gbVAR(p) model according to the alternative presentation of the univariate gbAR(1) model in (3.6). However, for the multivariate case, this becomes cumbersome and the main benefit of the presentation of the gbVAR model in (3.17) is the closed-form expression using an autoregressive-type model equation. Note that X_t is equal to the sum of lagged time series observations X_{t-i} multiplied in a familiar fashion with (random) matrices $A_t^{(+,\cdot)}$, i.e. $\sum_{i=1}^{p} A_t^{(+,i)}X_{t-i}$ (as in VAR-type models), plus an additional term $\sum_{i=1}^{p} A_t^{(-,i)}\mathbb{1}_K$ (related to negative coefficients) plus an innovation term $B_t e_t$.

In the following, we pick up the introductory example from Section 3.1.2 and consider a bivariate gbVAR(1) model in more detail to illustrate the class of gbVAR models.

Example 3.2 (Bivariate gbVAR(1) model)
Let $(X_t, t\in\mathbb{Z})$ follow a two-dimensional gbVAR(1) model

$$X_t = A_t^{(+)}X_{t-1} + A_t^{(-)}\mathbb{1}_K + B_t e_t, \tag{3.18}$$

with parameter matrix $\mathcal{P} = [\mathcal{A}, \mathcal{B}]$ and $\mathcal{P}_{|\cdot|} = [\mathcal{A}_{|\cdot|}, \mathcal{B}]$, where

$$\mathcal{A} = \begin{pmatrix} 0.49 & 0.35 \\ -0.43 & -0.39 \end{pmatrix}, \quad \mathcal{A}_{|\cdot|} = \begin{pmatrix} |0.49| & |0.35| \\ |-0.43| & |-0.39| \end{pmatrix} \quad \text{and} \quad \mathcal{B} = \begin{pmatrix} 0.16 & 0.00 \\ 0.00 & 0.18 \end{pmatrix}$$

such that $\mathcal{P}_{|\cdot|}\mathbb{1}_4 = \mathbb{1}_2$ holds. Hence, for the (mutually independent) multinomial selection mechanisms, we have

$$P_{t,1\bullet} = [a_{t,11}, a_{t,12}, b_{t,11}, 0] \sim Mult\big(1; (|0.49|, |0.35|, 0.16, 0)\big),$$
$$P_{t,2\bullet} = [a_{t,21}, a_{t,22}, 0, b_{t,22}] \sim Mult\big(1; (|-0.43|, |-0.39|, 0, 0.18)\big).$$

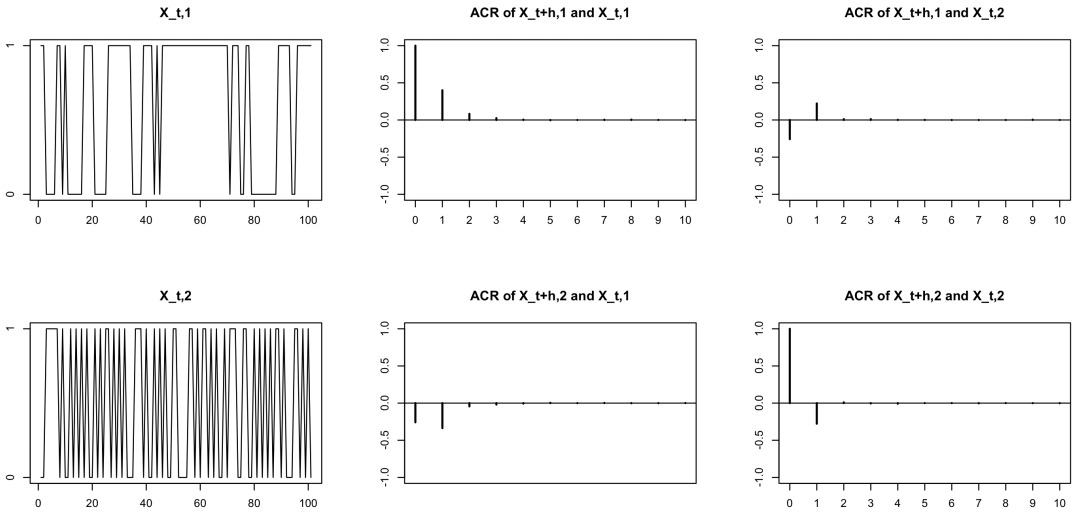


Figure 3.4.: Realization and autocorrelation structure of the bivariate gbVAR(1) process as specified in Example 3.2.

Taking the negative signs of the entries in $\mathcal{A}$ into account, the gbVAR(1) process follows the model equation

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} = \begin{pmatrix} a_{t,11} X_{t-1,1} & +a_{t,12} X_{t-1,2} & +b_{t,11} e_{t,1} & \\ a_{t,21} \left(1 - X_{t-1,1}\right) & +a_{t,22} \left(1 - X_{t-1,2}\right) & & +b_{t,22} e_{t,2} \end{pmatrix}.$$

Hence, in the second dimension the opposite values of the predecessors $X_{t-1,1}$ or $X_{t-1,2}$ are selected, whenever $a_{t,21}$ or $a_{t,22}$ become 1, respectively.

The innovation process $(e_t, t \in \mathbb{Z})$ consists of two independent Bernoulli processes $(e_{t,1}, t \in \mathbb{Z})$ and $(e_{t,2}, t \in \mathbb{Z})$ with $\mu_{e,1} = P(e_{t,1} = 1) = 0.4$ and $\mu_{e,2} = P(e_{t,2} = 1) = 0.8$ leading to $\Sigma_e = diag(0.24, 0.16)$. In Figure 3.4, we show a realization of the bivariate gbVAR(1) process with the above specification together with the corresponding serial and cross-sectional autocorrelation structure. By allowing for positive as well as negative entries in the non-diagonal parameter matrix $\mathcal{A}$, the gbVAR(1) model becomes rather flexible and allows to describe diverse serial and cross-sectional dependence structures.

3.2.2. Stochastic properties of gbVAR models

First, we consider the expectation of the random matrices $A_t^{(+,i)}$ and $A_t^{(-,i)}$, $i = 1, \ldots, p$. Note that, by construction, we have $E\big(A_t^{(i)}\big) = \mathcal{A}_{|\cdot|}^{(i)}$ and $E\big(B_t\big) = \mathcal{B}$. Hence, from the definitions of $A_t^{(+,i)}$ and $A_t^{(-,i)}$, we get

$$E\left(A_t^{(+,i)}\right) = \left[\alpha_{kl}^{(i)}\right]_{k,l=1,\ldots,K} = \mathcal{A}^{(i)}, \tag{3.19}$$

$$E\left(A_t^{(-,i)}\right) = \left[|\alpha_{kl}^{(i)}| \mathbb{1}_{\{\alpha_{kl}^{(i)} < 0\}}\right]_{k,l=1,\ldots,K} =: \mathcal{A}^{(-,i)}. \tag{3.20}$$

This enables us to compute the stationary mean vector $\mu_X := E\left(X_t\right)$ of a gbVAR(p) process.

Lemma 3.3 (Stationary mean of gbVAR processes)
Let $(X_t, t \in \mathbb{Z})$ be a stationary K-dimensional gbVAR(p) process. Then, we have,

$$\mu_X = \Big(I - \sum_{i=1}^{p} \mathcal{A}^{(i)}\Big)^{-1} \Big(\sum_{j=1}^{p} \mathcal{A}^{(-,j)} \mathbb{1}_K + \mathcal{B}\mu_e\Big). \tag{3.21}$$

The latter result reflects the relationship between the mean vector of the time series μ_X, the mean vector μ_e of the innovation process and the autoregressive parameters $\mathcal{P} = \big[\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(p)}, \mathcal{B}\big]$. In comparison to the univariate NDAR(p) process (see e.g. Weiß (2009a)), additional matrices $\mathcal{A}^{(-,\cdot)}$ appear in the formula for the mean that correspond to potentially negative model parameters. Further note that, in contrast to univariate gbAR(p) processes discussed in Jentsch and Reichmann (2019), we do not get $\mu_X = \mu_e$ in the special case when the parameter matrices in $\mathcal{P}$ contain exclusively non-negative entries, such that all $\mathcal{A}^{(-,j)}$ vanish. This is due to the diagonal structure of $\mathcal{B}$ leading to $I - \sum_{i=1}^{p} \mathcal{A}^{(i)} \neq \mathcal{B}$ in general.

In view of Definition 3.1, which supposes the gbVAR(p) process $(X_t, t \in \mathbb{Z})$ to be *stationary* and to *fulfill the gbVAR recursion* (3.17), its stationary solution can be derived in form of a moving-average-type gbVMA(∞) process. As for classical AR processes, the most simple case of $p = 1$ allows for a direct approach to construct the moving-average representation by recursively plugging-in the gbVAR(1) model equation. For all $d \in \mathbb{N}_0$, by recursively plugging-in (3.17), we get

$$\begin{aligned}
X_t &= A_t^{(+)} X_{t-1} + A_t^{(-)} \mathbb{1}_K + B_t e_t \\
&= A_t^{(+)} \Big(A_{t-1}^{(+)} X_{t-2} + A_{t-1}^{(-)} \mathbb{1}_K + B_{t-1} e_{t-1}\Big) + A_t^{(-)} \mathbb{1}_K + B_t e_t \\
&\;\;\vdots \\
&= \prod_{j=0}^{d} A_{t-j}^{(+)} X_{t-(d+1)} + \sum_{i=0}^{d} \left(\prod_{j=0}^{i-1} A_{t-j}^{(+)}\right) \Big(A_{t-i}^{(-)} \mathbb{1}_K + B_{t-i} e_{t-i}\Big) \\
&= \zeta_d X_{t-(d+1)} + \sum_{i=0}^{d} \zeta_{i-1} \eta_{t-i},
\end{aligned} \tag{3.22}$$

where $\zeta_0 := I_K$ and $\zeta_i := \prod_{j=0}^{i} A_{t-j}^{(+)}$, $i \in \mathbb{N}$ and $\eta_{t-i} := A_{t-i}^{(-)} \mathbb{1}_K + B_{t-i} e_{t-i}$. For gbVAR(p) processes of general order $p \in \mathbb{N}$, we follow the common approach described e.g. in Lütkepohl (2005)[Chap. 11.3.2] to rewrite the K-dimensional gbVAR(p) process as a Kp-dimensional gbVAR(1) process $(\widetilde{X}_t, t \in \mathbb{Z})$. Precisely, by defining the Kp-dimensional vectors

$$\widetilde{X}_t := \begin{pmatrix} X_t \\ \vdots \\ X_{t-p+1} \end{pmatrix} \quad \text{and} \quad \widetilde{e}_t := \begin{pmatrix} e_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

and the $(Kp \times Kp)$-dimensional matrices

$$\widetilde{A}_t^{(+)} := \begin{pmatrix} A_t^{(+,1)} & \dots & A_t^{(+,p-1)} & A_t^{(+,p)} \\ I_K & & 0_{K\times K} & 0_{K\times K} \\ & \ddots & & \vdots \\ 0_{K\times K} & \dots & I_K & 0_{K\times K} \end{pmatrix}, \quad \widetilde{A}_t^{(-)} := \begin{pmatrix} A_t^{(-,1)} & \dots & A_t^{(-,p-1)} & A_t^{(-,p)} \\ 0_{K\times K} & & 0_{K\times K} & 0_{K\times K} \\ & \ddots & & \vdots \\ 0_{K\times K} & \dots & 0_{K\times K} & 0_{K\times K} \end{pmatrix},$$

$$\widetilde{B}_t := \begin{pmatrix} B_t & 0_{K\times K(p-1)} \\ 0_{K(p-1)\times K} & 0_{K(p-1)\times K(p-1)} \end{pmatrix},$$

where I_K denotes the K-dimensional unity matrix and $0_{r\times s}$ the $(r\times s)$-dimensional zero matrix, we get an autoregressive representation for the process $(\widetilde{X}_t, t \in \mathbb{Z})$. That is, the K-dimensional gbVAR(p) process $(X_t, t \in \mathbb{Z})$ can be represented as a Kp-dimensional gbVAR(1) process $(\widetilde{X}_t, t \in \mathbb{Z})$ as follows

$$\widetilde{X}_t = \widetilde{A}_t^{(+)} \widetilde{X}_{t-1} + \widetilde{A}_t^{(-)} \mathbb{1}_{Kp} + \widetilde{B}_t \widetilde{e}_t, \quad t \in \mathbb{Z}, \tag{3.23}$$

where $\mathbb{1}_{Kp}$ is the one vector of length Kp. Note that the first K entries of $(\widetilde{X}_t, t \in \mathbb{Z})$ equal the gbVAR(p) process $(X_t, t \in \mathbb{Z})$. By exploiting the above representation (3.23) of $\widetilde{X}_t$ as a gbVAR(1) process, analogous to (3.22), we get

$$\widetilde{X}_t = \widetilde{\zeta}_d \widetilde{X}_{t-(d+1)} + \sum_{i=0}^{d} \widetilde{\zeta}_{i-1} \widetilde{\eta}_{t-i}, \tag{3.24}$$

with $\widetilde{\zeta}_0 := I_K$ and $\widetilde{\zeta}_i := \prod_{j=0}^{i} \widetilde{A}_{t-j}^{(+)}$, $i \in \mathbb{N}$ and $\widetilde{\eta}_t := \widetilde{A}_{t-i}^{(-)} \mathbb{1}_{Kp} + \widetilde{B}_{t-i} \widetilde{e}_{t-i}$. Now, by letting $d \to \infty$ on the right-hand sides of equations (3.22) and (3.24), respectively, this allows us to derive moving-average-type representations of gbVAR processes that make use of a stationarity conditions familiar from (causal) vector-valued autoregressive-type processes. Precisely, for $p = 1$, this condition is

$$det\left(I_K - \mathcal{A}_{|\cdot|} z\right) \neq 0 \quad \forall z \in \mathbb{C} : \quad |z| \leq 1, \tag{3.25}$$

and, for general $p \in \mathbb{N}$, it becomes

$$det\left(I_{Kp} - \widetilde{\mathcal{A}}_{|\cdot|} z\right) \neq 0 \quad \forall z \in \mathbb{C} : \quad |z| \leq 1, \tag{3.26}$$

which is equivalent to the condition that all eigenvalues of $\widetilde{\mathcal{A}}_{|\cdot|}$ have modulus smaller than one, and to the condition that all roots of the characteristic matrix polynomial lie outside the unit circle, i.e.

$$det\left(I_K - \mathcal{A}_{|\cdot|}^{(1)} z - \ldots - \mathcal{A}_{|\cdot|}^{(p)} z^p\right) \neq 0 \quad \forall z \in \mathbb{C} : \quad |z| \leq 1. \tag{3.27}$$

The following result extends Theorem 1 in Jentsch and Reichmann (2019) to the multivariate case.

Theorem 3.4 (Moving-average representation of gbVAR processes)
Let $(X_t, t \in \mathbb{Z})$ be a (stationary) K-dimensional gbVAR(p) process that fulfills (3.17).

(i) If $p = 1$ and condition (3.25) *holds, the gbVAR(1) model has a gbVMA(∞)-type representation*

$$X_t = \sum_{i=0}^{\infty} \zeta_i \eta_{t-i}, \quad t \in \mathbb{Z}, \tag{3.28}$$

converging in L_1.

(ii) If $p \in \mathbb{N}$ and condition (3.27) *holds, the gbVAR(p) model has a gbVMA(∞)-type representation*

$$X_t = J\widetilde{X}_t = J\left(\sum_{i=0}^{\infty} \widetilde{\zeta}_i \widetilde{\eta}_{t-i}\right), \quad t \in \mathbb{Z}, \tag{3.29}$$

converging in L_1, where $J := [I_K, 0_{K\times K(p-1)}]$.

In comparison to classical vector autoregressive processes, the moving-average representation of gbVAR processes contains an additional term $\widetilde{A}_{t-i}^{(-)}\mathbb{1}_{Kp}$ in $\widetilde{\eta}_{t-i} = \widetilde{A}_{t-i}^{(-)}\mathbb{1}_{Kp} + \widetilde{B}_{t-i}\widetilde{e}_{t-i}$ that takes control of potential negative parameters. If all entries in the parameter matrix $\mathcal{P} = [\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(p)}, \mathcal{B}]$ are non-negative, this additional term vanishes. It is also important to note that the stationarity conditions in (3.25) - (3.27) rely on the modified coefficient matrices $\mathcal{A}_{|\cdot|}^{(1)}, \ldots, \mathcal{A}_{|\cdot|}^{(p)}$ instead of $\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(p)}$.

The stationarity condition (3.25) for gbVAR(1) processes is illustrated in the following example.

Example 3.5 (Stationarity of gbVAR(1) models)
Let $K = 2$ and consider the bivariate gbVAR(1) process with parameter matrix $\mathcal{P} := [\mathcal{A}, \mathcal{B}]$, where

$$\mathcal{A} := \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \quad \textit{and} \quad \mathcal{B} := \begin{pmatrix} \beta_{11} & 0 \\ 0 & \beta_{22} \end{pmatrix}$$

such that $|\alpha_{11}| + |\alpha_{12}| + \beta_{11} = 1$ and $|\alpha_{21}| + |\alpha_{22}| + \beta_{22} = 1$.
Let us consider four cases:

(i) If $\beta_{11}, \beta_{22} > 0$, both innovations $e_{t,1}$ and $e_{t,2}$ are allowed to enter the gbVAR(1) model. From $|\alpha_{11}| + |\alpha_{12}| < 1$ and $|\alpha_{21}| + |\alpha_{22}| < 1$, we get that all eigenvalues of $\mathcal{A}_{|\cdot|}$ and of $\mathcal{A}$ have modulus smaller than one. Hence, condition (3.25) *holds.*

(ii) If $\beta_{11} = 0$, $\beta_{22} > 0$ and $\alpha_{12} \neq 0$, only the innovation $e_{t,2}$ is allowed to enter the gbVAR(1) model, but as $\alpha_{12} \neq 0$, it may reach $X_{t,1}$ after one time step as well. Indeed, in this case, it can be checked that all eigenvalues of $\mathcal{A}_{|\cdot|}$ and of $\mathcal{A}$ have modulus smaller than one and condition (3.25) *holds.*

(iii) If $\beta_{11} = 0$, $\beta_{22} > 0$ and $\alpha_{12} = 0$, the innovation $e_{t,2}$ is allowed to enter the gbVAR(1) model, but it will never reach $X_{t,1}$. In this case, where $|\alpha_{11}| = 1 - |\alpha_{12}| - \beta_{11} = 1$, the largest (in modulus) eigenvalues of $\mathcal{A}_{|\cdot|}$ and of $\mathcal{A}$ are equal to one and condition (3.25) *does* not *hold.*

(iv) If $\beta_{11} = \beta_{22} = 0$ and $\alpha_{ij} \neq 0$, $i, j = 1, 2$, such that

$$\mathcal{A} = \begin{pmatrix} \alpha_{11} & 1-\alpha_{11} \\ \alpha_{22} & -(1-\alpha_{21}) \end{pmatrix} \quad \textit{and} \quad \mathcal{A}_{|\cdot|} = \begin{pmatrix} \alpha_{11} & 1-\alpha_{11} \\ \alpha_{21} & 1-\alpha_{21}, \end{pmatrix}$$

no innovations at all are allowed to enter the gbVAR(1) model. In this case, the eigenvalues of $\mathcal{A}$ are (in modulus) smaller than one (e.g. for $\alpha_{11} = \alpha_{21} = 0.5$ they compute to $\lambda_{1,2} = \pm 0.7071$). Nevertheless, the largest eigenvalue (in modulus) of $\mathcal{A}_{|\cdot|}$ becomes one, such that condition (3.25) *does* not *hold.*

The latter example illustrates that $\beta_{kk} > 0$ for all $k = 1, \ldots, K$ is a sufficient, but not a necessary condition for (3.25) (and also for (3.27)) to hold.

Note that gbVAR processes are of autoregressive-type, but they are *non-linear* due to the *random* coefficient matrices. Nevertheless, we can show that the autocovariance structure of gbVAR processes coincides with that of classical VAR processes in the sense that the same Yule-Walker equations hold. For this purpose, we denote by $\Gamma_X(h) = Cov(X_{t+h}, X_t)$, $h \in \mathbb{Z}$, the corresponding autocovariance matrices of the gbVAR(p) process $(X_t, t \in \mathbb{Z})$.

Theorem 3.6 (Yule-Walker equations for gbVAR(p) models, $h > 0$)
Let $(X_t, t \in \mathbb{Z})$ be a stationary K-dimensional gbVAR(p) process. Then, for all $h \in \mathbb{N}$ (with $h \neq 0$), we have

$$\Gamma_X(h) = \sum_{i=1}^{p} \mathcal{A}^{(i)} \Gamma_X(h-i) \tag{3.30}$$

leading, in particular, to the system of Yule-Walker equations

$$[\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(p)}] \begin{pmatrix} \Gamma_X(i-j) \\ i,j = 1, \ldots, p \end{pmatrix} = [\Gamma_X(1), \ldots, \Gamma_X(p)]. \tag{3.31}$$

By replacing the autocovariances $\Gamma_X(h)$ by sample versions $\widehat{\Gamma}_X(h)$, Yule-Walker equations can be used for parameter estimation using the well-known Yule-Walker estimators. Before these will be addressed in Section 3.2.5, we will discuss possible identification issues in Section 3.2.3.

The derivation of a Yule-Walker-type equation for $h = 0$ is much more intricate. For this purpose, we will use Hadamard products denoted by "$\circ$", where $A \circ B := (a_{ij} b_{ij})_{i,j}$ for two matrices A and B of the same dimensions.

Theorem 3.7 (Yule-Walker equation for gbVAR(p) models, $h = 0$)
Let $(X_t, t \in \mathbb{Z})$ be a stationary K-dimensional gbVAR(p) process. Then, for $h = 0$, we have

$$\begin{aligned}
\Gamma_X(0) = &\sum_{i,j=1}^{p} \Big[I_K \circ \Big\{ \Big(\mathcal{A}^{(i)}_{|\cdot|} \mu_X \mathbb{1}_K' \Big) \mathbb{1}_{\{i=j\}} - \Big(\mathcal{A}^{(i)} \left(\Gamma_X(j-i) + \mu_X \mu_X' \right) \mathcal{A}^{(j)\prime} \Big) \Big\} . \\
&+ \Big(\mathcal{A}^{(i)} \left(\Gamma_X(j-i) \right) \mathcal{A}^{(j)\prime} \Big) \Big] \\
&+ \sum_{i,j=1}^{p} \Big[I_K \circ \Big\{ \Big(-\mathcal{A}^{(-,i)} \mu_X \mathbb{1}_K' \Big) \mathbb{1}_{\{i=j\}} - \Big(\mathcal{A}^{(i)} \mu_X \mathbb{1}_K' \mathcal{A}^{(-,j)\prime} \Big) \Big\} \Big] \\
&- \sum_{i=1}^{p} \Big[I_K \circ \Big\{ \Big(\mathcal{A}^{(i)} \mu_X \mu_e' \mathcal{B}' \Big) + \Big(\mathcal{A}^{(-,i)} \mathbb{1}_K \mu_e' \mathcal{B}' \Big) \Big\} \Big] \\
&+ \sum_{i,j=1}^{p} \Big[I_K \circ \Big\{ \Big(-\mathcal{A}^{(-,i)} \mu_X \mathbb{1}_K' \Big) \mathbb{1}_{\{i=j\}} - \Big(\mathcal{A}^{(-,i)} \mathbb{1}_K \mu_X' \mathcal{A}^{(j)\prime} \Big) \Big\} \Big] \\
&+ \sum_{i,j=1}^{p} \Big[I_K \circ \Big\{ \Big(\mathcal{A}^{(-,i)} \mathbb{1}_K \mathbb{1}_K' \Big) \mathbb{1}_{\{i=j\}} - \Big(\mathcal{A}^{(-,i)} \mathbb{1}_K \mathbb{1}_K' \mathcal{A}^{(-,j)\prime} \Big) \Big\} \Big] \\
&- \sum_{j=1}^{p} \Big[I_K \circ \Big\{ \Big(\mathcal{B} \mu_e \mu_X' \mathcal{A}^{(j)\prime} \Big) + \Big(\mathcal{B} \mu_e \mathbb{1}_K' \mathcal{A}^{(-,j)\prime} \Big) \Big\} \Big] \\
&+ I_K \circ \left\{ \mathcal{B} \mu_e \mathbb{1}_K' - \mathcal{B} \mu_e \mathbb{1}_K' \mathcal{B} \right\} + \mathcal{B} \Sigma_e \mathcal{B}.
\end{aligned} \tag{3.32}$$

Note that the formula derived in Theorem 3.7 contains the expression

$$\Gamma_X(0) = \sum_{i,j=1}^{p} \left(\mathcal{A}^{(i)} \left(\Gamma_X(j-i) \right) \mathcal{A}^{(j)\prime} \right) + \mathcal{B}\Sigma_e\mathcal{B},$$

which is similar to the classical formula for the Yule-Walker equation for $h = 0$ of VAR(p) processes (see e.g. Lütkepohl (2005)[Section 11.4], here only $\mathcal{B}\Sigma_e\mathcal{B}$ is replaced by Σ_e) plus some additional terms that contain "$I_K \circ \cdots$". Note that these additional terms *only* adjust the diagonal entries. They show up due to the random coefficients, that control the selection mechanism in the gbVAR model.

Alternatively, the last line in (3.32) can be rearranged to get

$$\begin{aligned} & I_K \circ \left\{ \mathcal{B}\mu_e \mathbb{1}_K' + \mathcal{B}(\Sigma_e - \mu_e \mathbb{1}_K')\mathcal{B} \right\} + (\mathbb{1}_{K\times K} - I_K) \circ \{\mathcal{B}\Sigma_e\mathcal{B}\} \\ =& I_K \circ \left\{ \mathcal{B}\mu_e \mathbb{1}_K' - \mathcal{B}\left(\mu_e\mu_e'\right)\mathcal{B} \right\} + \mathcal{B}\left\{ (\mathbb{1}_{K\times K} - I_K) \circ \Sigma_e \right\} \mathcal{B}. \end{aligned} \tag{3.33}$$

Note that the Hadamard multiplication in last term on the right-hand side sets the diagonal of Σ_e (which is already determined by μ_e) to zero and that all other expressions in (3.33) contain only $\mathcal{B}$ and μ_e, but not Σ_e. This is particularly useful to identify the off-diagonal elements of Σ_e, which can be used for estimation purposes as discussed in Section 3.2.5.

Example 3.8 (Special cases of Theorem 3.7)
For $p = 1$, the formula derived in Theorem 3.7 simplifies to become

$$\begin{aligned} \Gamma_X(0) = & I_K \circ \left\{ \left(\mathcal{A}^{(1)}_{|\cdot|} \mu_X \mathbb{1}_K' \right) - \left(\mathcal{A}^{(1)} \left(\Gamma_X(0) + \mu_X\mu_X' \right) \mathcal{A}^{(1)\prime} \right) \right\} + \left(\mathcal{A}^{(1)} \Gamma_X(0) \mathcal{A}^{(1)\prime} \right) \\ & + \left[I_K \circ \left\{ \left(-\mathcal{A}^{(-,1)} \mu_X \mathbb{1}_K' \right) - \left(\mathcal{A}^{(1)} \mu_X \mathbb{1}_K' \mathcal{A}^{(-,1)\prime} \right) \right\} \right] - \left[I_K \circ \left\{ \mathcal{A}^{(1)} \mu_X \mu_e' \mathcal{B}' \right\} \right] \\ & + \left[I_K \circ \left\{ -\mathcal{A}^{(-,1)} \mu_X \mathbb{1}_K' - \left(\mathcal{A}^{(-,1)} \mathbb{1}_K \mu_X' \mathcal{A}^{(1)\prime} \right) \right\} \right] \\ & + \left[I_K \circ \left\{ \left(\mathcal{A}^{(-,1)} \mathbb{1}_K \mathbb{1}_K' \right) - \mathcal{A}^{(-,1)} \mathbb{1}_K \mathbb{1}_K' \mathcal{A}^{(-,1)\prime} \right\} \right] \\ & - \left[I_K \circ \left\{ \mathcal{A}^{(-,1)} \mathbb{1}_K \mu_e' \mathcal{B}' \right\} \right] - \left[I_K \circ \left\{ \mathcal{B}\mu_e \mu_X' \mathcal{A}^{(1)\prime} \right\} \right] - \left[I_K \circ \left\{ \mathcal{B}\mu_e \mathbb{1}_K' \mathcal{A}^{(-,1)\prime} \right\} \right] \\ & + I_K \circ \left\{ \mathcal{B}\mu_e \mathbb{1}_K' - \mathcal{B}(\mu_e \mathbb{1}_K')\mathcal{B} \right\} + \mathcal{B}\Sigma_e\mathcal{B}. \end{aligned}$$

If additionally all entries in $\mathcal{A}^{(1)}$ are non-negative, all terms containing $\mathcal{A}^{(-,1)} = 0_{K\times K}$ vanish and we can replace $\mathcal{A}^{(1)}_{|\cdot|}$ by $\mathcal{A}^{(1)}$. This leads to

$$\begin{aligned} \Gamma_X(0) = & I_K \circ \left\{ \left(\mathcal{A}^{(1)} \mu_X \mathbb{1}_K' \right) - \left(\mathcal{A}^{(1)} \left(\Gamma_X(0) + \mu_X\mu_X' \right) \mathcal{A}^{(1)\prime} \right) \right\} + \left(\mathcal{A}^{(1)} \Gamma_X(0) \mathcal{A}^{(1)\prime} \right) \\ & - \left[I_K \circ \left\{ \mathcal{A}^{(1)} \mu_X \mu_e' \mathcal{B}' \right\} \right] - \left[I_K \circ \left\{ \mathcal{B}\mu_e \mu_X' \mathcal{A}^{(1)\prime} \right\} \right] \\ & + I_K \circ \left\{ \mathcal{B}\mu_e \mathbb{1}_K' - \mathcal{B}(\mu_e \mathbb{1}_K')\mathcal{B} \right\} + \mathcal{B}\Sigma_e\mathcal{B}. \end{aligned} \tag{3.34}$$

In the following result, we derive expressions for one step ahead transition probabilities for gbVAR processes to reach a certain state in $\{0,1\}^K$ given the past values of the time series.

Lemma 3.9 (Transition probabilities of gbVAR processes)
Let $(X_t, t \in \mathbb{Z})$ be a stationary K-dimensional gbVAR(p) process. Further, for $k = 0, 1, \ldots, p$, let $s_k, r_0 \in \{0,1\}^K$ with $s_k := (s_{k,1}, \ldots, s_{k,K})'$ and $r_0 := (r_{0,1}, \ldots, r_{0,K})'$.

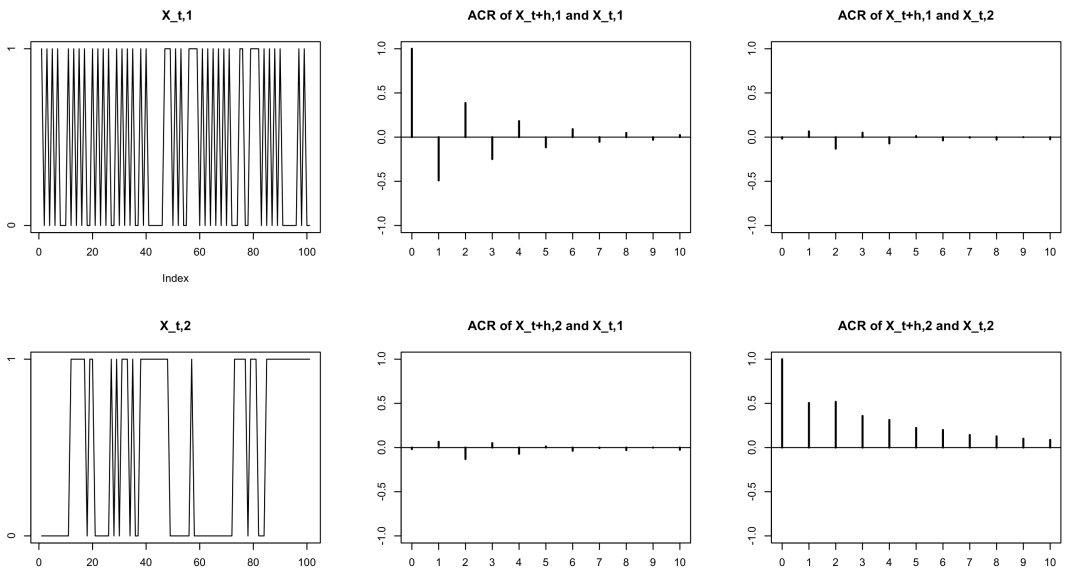


Figure 3.5.: Realization and autocorrelation structure of the bivariate gbVAR(2) process as specified in Example 3.10.

Denote by $\delta_{ij} = \mathbb{1}_{\{i=j\}}$ *the Kronecker delta and set* $p_{r_0} := P(e_t = r_0)$. *Then, the transition probability given the past values of the time series becomes*

$$P\big(X_t = s_0 | X_{t-1} = s_1, \ldots, X_{t-p} = s_p\big) = \tag{3.35}$$
$$\sum_{r_0 \in \{0,1\}^K} p_{r_0} \prod_{k=1}^{K} \left[\sum_{i=1}^{p} \sum_{l=1}^{K} |\alpha_{kl}^{(i)}| \left[\mathbb{1}_{\{\alpha_{kl}^{(i)} \geq 0\}} \delta_{s_{0,k} s_{i,l}} + \mathbb{1}_{\{\alpha_{kl}^{(i)} < 0\}} \delta_{s_{0,k}(1-s_{i,l})} \right] + \beta_{kk} \delta_{s_{0,k} r_{0,k}} \right].$$

In the following example, we illustrate the derived formula for the transition probabilities (3.35) by using a bivariate gbVAR(2) process.

Example 3.10 (Transition probabilities of a bivariate gbVAR(2) model)
Let $(X_t, t \in \mathbb{Z})$ *be a bivariate gbVAR(2) process following the model equation*

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} = \begin{pmatrix} a_{t,11}^{(+,1)} & a_{t,12}^{(+,1)} \\ a_{t,21}^{(+,1)} & a_{t,22}^{(+,1)} \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \end{pmatrix} + \begin{pmatrix} a_{t,11}^{(-,1)} & a_{t,12}^{(-,1)} \\ a_{t,21}^{(-,1)} & a_{t,22}^{(-,1)} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$
$$+ \begin{pmatrix} a_{t,11}^{(+,2)} & a_{t,12}^{(+,2)} \\ a_{t,21}^{(+,2)} & a_{t,22}^{(+,2)} \end{pmatrix} \begin{pmatrix} X_{t-2,1} \\ X_{t-2,2} \end{pmatrix} + \begin{pmatrix} a_{t,11}^{(-,2)} & a_{t,12}^{(-,2)} \\ a_{t,21}^{(-,2)} & a_{t,22}^{(-,2)} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} b_{t,11} & 0 \\ 0 & b_{t,22} \end{pmatrix} \begin{pmatrix} e_{t,1} \\ e_{t,2} \end{pmatrix}.$$

Let $s_0, s_1, s_2, r_0 \in \{0,1\}^2$. *Then, based on the model parameters summarized in* $\mathcal{P} := \big[\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \mathcal{B}\big]$, *the transition probability to observe the state vector* s_0 *at time* t *given the state vectors* s_1 *and* s_2 *have been observed at time* $t-1$ *and* $t-2$, *respectively, is*

$$P\left(X_t = s_0 | X_{t-1} = s_1, X_{t-2} = s_2\right) =$$
$$= \sum_{r_0 \in \{0,1\}^2} p_{r_0} \prod_{k=1}^{2} \left(\sum_{i=1}^{2} \sum_{l=1}^{2} \left[|\alpha_{kl}^{(i)}| \left(\mathbb{1}_{\{\alpha_{kl}^{(i)} \geq 0\}} \delta_{s_{0,k} s_{1,l}} + \mathbb{1}_{\{\alpha_{kl}^{(i)} < 0\}} \delta_{s_{0,k}(1-s_{1,l})} \right) \right] + \beta_{kk} \delta_{s_{0,k} r_{0,k}} \right).$$

Depending on the signs of the entries $\alpha_{kl}^{(i)}$, *the Kronecker deltas in the formula above control which* $|\alpha_{kl}^{(i)}|$ *enter the transition probability. In Figure 3.5, we show a realization*

of the gbVAR(2) process with parameters $\mathcal{P} = \left[\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \mathcal{B}\right]$ *specified as*

$$\mathcal{P} := \left[\begin{pmatrix} -0.37 & 0.15 \\ -0.14 & 0.33 \end{pmatrix}, \begin{pmatrix} 0.23 & -0.18 \\ -0.11 & 0.36 \end{pmatrix}, \begin{pmatrix} 0.07 & 0.00 \\ 0.00 & 0.06 \end{pmatrix} \right]$$

with innovation process $(e_t, t \in \mathbb{Z})$ *consisting of two independent Bernoulli processes* $(e_{t,1}, t \in \mathbb{Z})$ *and* $(e_{t,2}, t \in \mathbb{Z})$ *with* $\mu_{e,1} = P(e_{t,1} = 1) = 0.4$ *and* $\mu_{e,2} = P(e_{t,2} = 1) = 0.5$ *leading to* $\Sigma_e = diag(0.24, 0.25)$*. This leads to the corresponding transition probability*

$$\begin{aligned} P\left(X_t = s_0 | X_{t-1} = s_1, X_{t-2} = s_2\right) = \sum_{r_0 \in \{0,1\}^2} P\left(e_{t,1} = r_{0,1}\right) P\left(e_{t,2} = r_{0,2}\right) \cdot \\ \left(|-0.37| \delta_{s_{0,1}(1-s_{1,1})} + 0.15 \delta_{s_{0,1} s_{1,2}} + 0.23 \delta_{s_{0,1} s_{2,1}} + |-0.18| \delta_{s_{0,1}(1-s_{2,2})} + 0.07 \delta_{s_{0,1} r_{0,1}} \right) \\ \left(|-0.14| \delta_{s_{0,2}(1-s_{1,1})} + 0.33 \delta_{s_{0,2} s_{1,2}} + |-0.11| \delta_{s_{0,2}(1-s_{2,1})} + 0.36 \delta_{s_{0,2} s_{2,2}} + 0.06 \delta_{s_{0,2} r_{0,2}} \right). \end{aligned}$$

As can be seen in Figure 3.5, the first component of the gbVAR(2) process shows an alternating pattern with an alternating ACF, whereas the second component tends to show longer runs of the same value and a positive ACF.

Remark 3.11 (Dependent multinomial selection)
In view of the mutually independent multinomial selection mechanisms $P_{t,k\bullet}$*,* $k = 1, \ldots, K$*, used to define gbVAR processes in Definition 3.1, a possible extension would be to allow for dependence between the multinomial distributions. Such an extension would not affect the Yule-Walker equations for* $h > 0$ *in Theorem 3.6, but those in Theorem 3.7 for* $h = 0$*. However, we do not follow this path as it would complicate things considerably and the benefit of such an extension of the gbVAR model class would be comparatively small. Furthermore, multinomial distributions allowing for dependence seem to be less developed; see e.g. Johnson et al. (1997)[Ch. 36] for a rather restrictive attempt in this direction.*

Note that the scalar selection mechanism used for the GDARMA approach proposed in Möller and Weiß (2020) coincides with a setup of mutually dependent multinomial selection mechanisms with perfectly correlated coefficient matrices' diagonal entries.

3.2.3. Identification of gbVAR models

As discussed in Theorem 3.4, the stationarity condition (3.27) implies a gbVMA(∞)-type representation of a gbVAR(p) process. Moreover, Theorem 3.6 shows that standard Yule-Walker equations hold. However, the stationarity condition (3.27) alone is not yet sufficient to make a gbVAR(p) model identifiable even if we assume that $Var(X_{t,k}) > 0$ for all $k = 1, \ldots, K$. Such an identification problem will show whenever a stationary K-dimensional gbVAR process has a *reducible* state space $\mathcal{V}_X \subsetneq \{0,1\}^K$. That is, when the process cannot take *all* states in $\{0,1\}^K$, such that there are $s_0 \in \{0,1\}^K$ with $P(X_t = s_0) = 0$. Hence, if we assume that $Var(X_{t,k}) > 0$ for all $k = 1, \ldots, K$, for the *true* state space $\mathcal{V}_X$, we have $|\mathcal{V}_X| = 2^{\widetilde{K}}$ for some $\widetilde{K} < K$. In this case, X_t contains components $X_{t,i}$ and $X_{t,j}$ that are either perfectly positively correlated or perfectly negatively correlated, respectively, such that $\Gamma_X(0)$ has a reduced rank $rk(\Gamma_X(0)) < K$. Consequently, the Yule-Walker matrix

$$\begin{pmatrix} \Gamma_X(i-j) \\ i, j = 1, \ldots, p \end{pmatrix}$$

has a reduced rank and is not invertible. Hence, a reducible state space leads to identifiability issues as the Yule-Walker equations derived in Theorem 3.6 do not have a unique solution and corresponding Yule-Walker estimators are not applicable.

However, whenever a stationary K-dimensional binary process $(X_t, t \in \mathbb{Z})$ with $Var(X_{t,k}) > 0$ for all $k = 1, \dots, K$ has a reducible state space $\mathcal{V}_X$ with $|\mathcal{V}_X| = 2^{\widetilde{K}}$ with $\widetilde{K} < K$, it can be written as

$$X_t = M(G\mathbb{1}_{\widetilde{K}} + HY_t), \tag{3.36}$$

where $(Y_t, t \in \mathbb{Z})$ is a $\widetilde{K}$-dimensional binary process with *non-reducible* state space $\mathcal{V}_Y = \{0,1\}^{\widetilde{K}}$, M is a $K \times K$ dimensional permutation matrix, and G and H are suitable $K \times \widetilde{K}$ matrices with entries in $\{0,1\}$ and $\{-1,0,1\}$, respectively. Precisely, $(X_t, t \in \mathbb{Z})$ can be rearrange (according to M) and partitioned to get

$$X_t = M \begin{pmatrix} X_{t,pos} \\ X_{t,neg} \\ Y_t \end{pmatrix}, \tag{3.37}$$

where $(Y_t, t \in \mathbb{Z})$ is a $\widetilde{K}$-dimensional process such that $\widetilde{K}$ is the largest possible dimension with $\Gamma_Y(0) = Var(Y_t)$ has full rank and $(Y_t, t \in \mathbb{Z})$ has a *non-reducible* state space. Then, for all remaining components $X_{t,i}$ of X_t (that are not included in Y_t) there exist (only!) one component $Y_{t,j}$ in Y_t such that $X_{t,i}$ and $Y_{t,j}$ are either perfectly positively or perfectly negatively correlated. Let $X_{t,pos}$ and $X_{t,neg}$ be sub-processes of dimensions K_{pos} and K_{neg} with $K_{pos} + K_{neg} + \widetilde{K} = K$ that summarize all remaining components of X_t that are perfectly positively and perfectly negatively correlated with some component in Y_t, respectively. Furthermore, the $(K \times \widetilde{K})$ matrices G and H can be partitioned as follows

$$G = \begin{pmatrix} G_{pos} \\ G_{neg} \\ O_{\widetilde{K} \times \widetilde{K}} \end{pmatrix} \quad \text{and} \quad H = \begin{pmatrix} H_{pos} \\ H_{neg} \\ I_{\widetilde{K} \times \widetilde{K}} \end{pmatrix},$$

where

$$G_{pos} = O_{K_{pos} \times \widetilde{K}}, \quad G_{neg} = \begin{pmatrix} \mathbb{1}\{Cov(X_{t,r,neg}, Y_{t,s}) = -1)\} \\ r = 1, \dots, K_{neg} \\ s = 1, \dots, \widetilde{K} \end{pmatrix}$$

as well as

$$H_{pos} = \begin{pmatrix} \mathbb{1}\{Cov(X_{t,r,pos}, Y_{t,s}) = 1)\} \\ r = 1, \dots, K_{pos} \\ s = 1, \dots, \widetilde{K} \end{pmatrix}, \quad H_{neg} := -G_{neg},$$

are matrices of dimensions $(K_{pos} \times \widetilde{K})$ and $(K_{neg} \times \widetilde{K})$, respectively.

In the following example, to illustrate the above, we consider the concrete setup of a four-dimensional gbVAR(1) process $(X_t, t \in \mathbb{Z})$, that fulfills the stationarity condition (3.25), but has a *reducible* state space leading to identification issues as described above. We derive how this process relates via (3.36) and (3.37) to a stationary lower-dimensional gbVAR(1) process $(Y_t, t \in \mathbb{Z})$ with *non-reducible* state space.

Example 3.12 (gbVAR(1) with reducible state space)
Suppose $(X_t, t \in \mathbb{Z})$ *is a four-dimensional gbVAR(1) process* $(K = 4)$ *with*

$$\mathcal{A} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix} \quad \text{and} \quad \mathcal{B} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \beta_{44} \end{pmatrix},$$

where $\alpha_{41}, \alpha_{42}, \alpha_{43}, \alpha_{44} \geq 0$ *and* $\beta_{44} > 0$ *such that* $\alpha_{41} + \alpha_{42} + \alpha_{43} + \alpha_{44} + \beta_{44} = 1$ *and with* $E(e_t) = \mu_e$ *and* $Var(e_t) = diag(\mu_{e,k}(1-\mu_{e,k}), k = 1, \ldots, 4)$ *Then, by construction, the process* $(X_t, t \in \mathbb{Z})$ *fulfills* $X_{t,1} = -X_{t,2} = X_{t,3}$ *for all* $t \in \mathbb{Z}$ *such that the true state space becomes*

$$\mathcal{V}_X = \{(1,0,1,0), (1,0,1,1), (0,1,0,1), (0,1,0,0)\} \subsetneq \{0,1\}^4 \tag{3.38}$$

with $|\mathcal{V}_X| = 2^2 = 4$. *Hence, the process* $(X_t, t \in \mathbb{Z})$ *has a* reducible state space *as the process does not take all values in* $\{0,1\}^4$ *with* $|\{0,1\}^4| = 2^4 = 16$. *In particular,* $X_{t,1}$ *and* $X_{t,3}$ *are perfectly positively correlated and* $X_{t,2}$ *and* $X_{t,3}$ *are perfectly negatively correlated. Hence, in this case, we get*

$$X_t = M(G\mathbb{1}_2 + HY_t), \tag{3.39}$$

where $M = I_4$ *(by construction),*

$$G = \begin{pmatrix} G_{pos} \\ \hline G_{neg} \\ \hline O_{\widetilde{K}\times\widetilde{K}} \end{pmatrix} = \left(\begin{array}{cc} 0 & 0 \\ \hline 1 & 0 \\ \hline 0 & 0 \\ 0 & 0 \end{array}\right) \quad \text{and} \quad H = \begin{pmatrix} H_{pos} \\ \hline H_{neg} \\ \hline I_{\widetilde{K}\times\widetilde{K}} \end{pmatrix} = \left(\begin{array}{cc} 1 & 0 \\ \hline -1 & 0 \\ \hline 1 & 0 \\ 0 & 1 \end{array}\right).$$

The process $(Y_t, t \in \mathbb{Z})$ *follows a bivariate* $(\widetilde{K} = 2)$ *gbVAR(1) model*

$$Y_t = A_{Y,t}^{(+)} Y_{t-i} + A_{Y,t}^{(-)} \mathbb{1}_K + B_{Y,t}\epsilon_t, \quad t \in \mathbb{Z},$$

with parameters $\mathcal{A}_Y, \mathcal{A}_Y^{(-)}, \mathcal{A}_{Y,|\cdot|}$ *and* $\mathcal{B}_Y$ *as in Definition 3.1 such that*

$$\begin{aligned} E(A_{Y,t}^{(+)}) &= H^+\mathcal{A}H = \begin{pmatrix} 0 & 1 \\ \alpha_{41} - \alpha_{42} + \alpha_{43} & \alpha_{44} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ \alpha_{Y,21} & \alpha_{Y,22} \end{pmatrix} = \mathcal{A}_Y, \\ E(A_{Y,t}^{(-)}) &= \begin{pmatrix} 0 & 0 \\ |\alpha_{Y,21}|\mathbb{1}_{\{\alpha_{Y,21}<0\}} & 0 \end{pmatrix} = \mathcal{A}_Y^{(-)}, \\ E(A_{Y,t}) &= \begin{pmatrix} 0 & 1 \\ |\alpha_{Y,21}| & \alpha_{Y,22} \end{pmatrix} = \mathcal{A}_{Y,|\cdot|} \quad \text{and} \\ E(B_{Y,t}) &= \begin{pmatrix} 0 & 0 \\ 0 & 1 - |\alpha_{41} - \alpha_{42} + \alpha_{43}| - |\alpha_{44}| \end{pmatrix} = \mathcal{B}_Y. \end{aligned}$$

Here, H^+ *denotes the Moore-Penrose inverse of* H *with*

$$H^+ = \begin{pmatrix} \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Furthermore, we get $\mu_\epsilon = E(\epsilon_t)$ *with* $\mu_{\epsilon,1} := 0$ *(*$\mu_{\epsilon,1}$ *is not identified due to* $\beta_{Y,11} = 0$*) and*

$$\begin{aligned}\mu_{\epsilon,2} &= \frac{(0,1)\left(-\mathcal{A}_Y^{(-)}\mathbb{1}_2 + H^+\mathcal{A}G\mathbb{1}_2 + H^+\mathcal{A}^{(-)}\mathbb{1}_4 + H^+\mathcal{B}\mu_e - H^+G\mathbb{1}_2\right)}{\beta_{Y,22}} \\ &= \frac{(\alpha_{41}-\alpha_{42}+\alpha_{43})\mathbb{1}_{\{\alpha_{41}-\alpha_{42}+\alpha_{43}<0\}} + \alpha_{42} + \mu_{e,4}}{1-|\alpha_{41}-\alpha_{42}+\alpha_{43}|-\alpha_{44}},\end{aligned} \tag{3.40}$$

where we used that

$$\mathcal{A}^{(-)} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

by construction due to $\alpha_{41}, \alpha_{42}, \alpha_{43}, \alpha_{44} \geq 0$ *and* $\beta_{44} > 0$*. The proof of equation* (3.40) *can be found in the appendix.*

Now, we address why diagonality of $\mathcal{B}$ has to be imposed to achieve identifiability of Σ_e.

Remark 3.13 (Non-diagonal $\mathcal{B}$ matrix)
So far, we presumed a diagonal structure of the parameter matrix $\mathcal{B}$*, i.e.* $\mathcal{B} = diag(\beta_{11}, \ldots, \beta_{KK})$*. The purpose is essentially two-fold. First, the parameter matrices* $\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(p)}$ *can be easily estimated using Yule-Walker estimators as will be described in Section 3.2.5. From the requirement that row-wise the entries in* $\mathcal{P} = [\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(p)}, \mathcal{B}]$ *have to sum up to one, it is most convenient to define* $\beta_{kk} := 1 - \sum_{i=1}^p \sum_{l=1}^K |\alpha_{kl}^{(i)}|$*, for* $k = 1, \ldots, K$*. Second, if the diagonality of* $\mathcal{B}$ *is* not *enforced, it is unclear how to allocate* $1 - \sum_{i=1}^p \sum_{l=1}^K |\alpha_{kl}^{(i)}|$ *to* K *free parameters* $\beta_{k1}, \ldots, \beta_{kK}$*. Hence, allowing* $\mathcal{B}$ *to be non-diagonal leads to identification issues.*

For example, let $K = 2$ *and suppose* $\mathcal{B}$ *is potentially non-diagonal with non-negative entries. Then, the bivariate process* $\{B_t e_t, t \in \mathbb{Z}\}$ *is i.i.d. taking values* $s_0 \in \{0,1\}^2$ *and we get*

$$\begin{aligned}P\left(B_t e_t = s_0\right) &= \sum_{r_0 \in \{0,1\}^2} P\left(B_t e_t = s_0 | e_t = r_0\right) P\left(e_t = r_0\right) \\ &= \sum_{r_0 \in \{0,1\}^2} \left(\beta_{11}\delta_{r_{0,1}s_{0,1}} + \beta_{12}\delta_{r_{0,2}s_{0,1}}\right)\left(\beta_{21}\delta_{r_{0,1}s_{0,2}} + \beta_{22}\delta_{r_{0,2}s_{0,2}}\right) p_{r_0}.\end{aligned} \tag{3.41}$$

Hence, as the distribution of $B_t e_t$ *takes values in* $\{0,1\}^2$*, it is fully specified by three parameters. If* $\mathcal{B}$ *is imposed to be diagonal such that* $\beta_{12} = \beta_{21} = 0$ *and* β_{11} *and* β_{22} *are pre-determined (by* $\beta_{kk} := 1 - \sum_{i=1}^p \sum_{l=1}^K |\alpha_{kl}^{(i)}|$*), the distribution of* e_t*, which is specified by e.g.* $p_{(0,0)}$*,* $p_{(0,1)}$ *and* $p_{(1,0)}$ *(with* $p_{(1,1)} = 1 - p_{(0,0)} - p_{(0,1)} - p_{(1,0)}$*), is identified. If* $\mathcal{B}$ *is not restricted to be diagonal, this is not the case.*

3.2.4. Mixing properties of gbVAR processes

For the derivation of asymptotic theory such as e.g. central limit theorems for statistics computed from the time series data, mixing concepts are very helpful to quantify the

serial dependence structure of time series processes. For overviews of different mixing concepts, we refer e.g. to Doukhan (1994), Dedecker et al. (2007) or Bradley (2007).

A suitable Markov chain representation has been proven to be beneficial to prove mixing properties for time series processes with discrete and finite state spaces; see e.g. Jacobs and Lewis (1983) or Weiß (2009a). In this section, we adopt this approach to examine mixing properties of gbVAR processes in order to prove (geometric) ψ- and φ-mixing. As given in Billingsley (1968), when defined on a suitable probability space $(\Omega, \mathcal{A}, P)$, a process $(Z_t, t \in \mathbb{Z})$ is called ψ- mixing, if for all subsets $\mathcal{E}_1 \in \sigma(Z_t, Z_{t-1}, \ldots)$ and $\mathcal{E}_2 \in \sigma(Z_{t+h}, Z_{t+h+1}, \ldots)$ of the induced σ- fields, and a non negative sequence $(f_h, h \in \mathbb{N})$ with $f_h \to 0$ for $h \to \infty$, we have

$$|P(\mathcal{E}_1 \cap \mathcal{E}_2) - P(\mathcal{E}_1) P(\mathcal{E}_2)| \leq f_h P(\mathcal{E}_1) P(\mathcal{E}_2). \tag{3.42}$$

If the right-hand side of the inequality (3.42) is replaced by $f_h P(\mathcal{E}_1)$, we get the definition of the (weaker) φ- mixing property.

The Markov chain representation of a gbVAR(p) process $(X_t, t \in \mathbb{Z})$ is obtained by stacking the time series values X_{t-i}, $i = 0, \ldots, p-1$ and the innovation e_t to get a $K(p+1)$- dimensional random vector Z_t defined by

$$Z_t := vec([X_t, X_{t-1}, \ldots, X_{t-p+1}, e_t]), \tag{3.43}$$

which defines a $K(p+1)$- dimensional homogeneous Markov chain $(Z_t, t \in \mathbb{Z})$. Let $\mathbf{s}_0, \mathbf{s}_1 \in \{0,1\}^{K(p+1)}$ with

$$\mathbf{s}_0 := (s_0', \ldots, s_{p-1}', r_0')' \quad \text{and} \quad \mathbf{s}_1 := (\tilde{s}_1', \ldots, \tilde{s}_p', \tilde{r}_1')',$$

where $s_i, \tilde{s}_i, r_0, \tilde{r}_1 \in \{0,1\}^K$ with $s_i = (s_{i,1}, \ldots, s_{i,K})'$ etc. for $i = 0, \ldots, p$. Similar to the derivation of the transition probability stated in Lemma 3.9, we obtain the conditional probability

$$\begin{aligned} &P(X_t = s_0 | X_{t-1} = s_1, \ldots, X_{t-p} = s_p, e_t = r_0) = \\ &\prod_{k=1}^{K} \left[\sum_{i=1}^{p} \sum_{l=1}^{K} |\alpha_{kl}^{(i)}| \left[\mathbb{1}_{\{\alpha_{kl}^{(i)} \geq 0\}} \delta_{s_{0,k} s_{i,l}} + \mathbb{1}_{\{\alpha_{kl}^{(i)} < 0\}} \delta_{s_{0,k}(1-s_{i,l})} \right] + \beta_{kk}^{(0)} \delta_{s_{0,k} r_{0,k}} \right], \end{aligned} \tag{3.44}$$

which allows to express the transition probability of the Markov chain $(Z_t, t \in \mathbb{Z})$ as follows

$$\begin{aligned} &P(Z_t = \mathbf{s}_0 | Z_{t-1} = \mathbf{s}_1) = P(X_t = s_0 | X_{t-1} = \tilde{s}_1, \ldots, X_{t-p} = \tilde{s}_p, e_t = r_0) \cdot \\ &\qquad P(X_{t-1} = s_1 | X_{t-1} = \tilde{s}_1) \cdot \ldots \cdot P(X_{t-p+1} = s_{p-1} | X_{t-p+1} = \tilde{s}_{p-1}) P(e_t = r_0) \\ &= \prod_{k=1}^{K} \left[\sum_{i=1}^{p} \sum_{l=1}^{K} |\alpha_{kl}^{(i)}| \left[\mathbb{1}_{\{\alpha_{kl}^{(i)} \geq 0\}} \delta_{s_{0,k} s_{i,l}} + \mathbb{1}_{\{\alpha_{kl}^{(i)} < 0\}} \delta_{s_{0,k}(1-s_{i,l})} \right] + \beta_{kk}^{(0)} \delta_{s_{0,k} r_{0,k}} \right] \cdot \\ &\qquad \delta_{s_1 \tilde{s}_1} \cdot \ldots \delta_{s_{p-1} \tilde{s}_{p-1}} \cdot p_{r_0}, \end{aligned}$$

where $p_{r_0} = P(e_t = r_0)$. If the underlying K-dimensional gbVAR(p) process $(X_t, t \in \mathbb{Z})$ has a *non-reducible* state space, that is, we have $P(X_t = s_0) > 0$ for all $s_0 \in \{0,1\}^K$, the latter formula allows to prove *primitivity* of the Markov chain $(Z_t, t \in \mathbb{Z})$. We will address gbVAR(p) processes with *reducible state* spaces also in Theorem 3.15 below.

Lemma 3.14 (Primitivity)
Let $(Z_t, t \in \mathbb{Z})$ be the Markov Chain representation of a stationary gbVAR(p) process $(X_t, t \in \mathbb{Z})$ with non-reducible state space and $p \geq 1$. Then $(Z_t, t \in \mathbb{Z})$ is primitive, that is, for $n := \max\{p, K'\} + 1$, we have

$$p_{\mathbf{s}|\mathbf{r}}(n) := P\left(Z_t = \mathbf{s} | Z_{t-n} = \mathbf{r}\right) > 0 \quad \textit{for all} \quad \mathbf{r}, \mathbf{s} \in \{0,1\}^{K(p+1)}.$$

Following the argumentation in Weiß (2009a)[Section 11.2], we can conclude from Lemma 3.14 that the Markov process $(Z_t, t \in \mathbb{Z})$ based on a stationary gbVAR(p) process $(X_t, t \in \mathbb{Z})$ with non-reducible state space is ergodic and geometrically ψ- and φ-mixing.

As the process $(X_t, t \in \mathbb{Z})$ is contained in $(Z_t, t \in \mathbb{Z})$, we can finally conclude from the mixing properties of its Markov chain representation that the gbVAR(p) process $(X_t, t \in \mathbb{Z})$ itself is geometrically ψ- and φ- mixing if $(X_t, t \in \mathbb{Z})$ has a non-reducible state space. Otherwise, if the process $(X_t, t \in \mathbb{Z})$ has a reducible state space $\mathcal{V}_X \subsetneq \{0,1\}^K$, we consider the corresponding lower-dimensional process $(Y_t, t \in \mathbb{Z})$ with non-reducible state space as discussed in Section 3.2.3 and derive mixing properties for this process. As $(X_t, t \in \mathbb{Z})$ is obtained by a linear transformation from $(Y_t, t \in \mathbb{Z})$, the corresponding mixing properties are easily concluded also for $(X_t, t \in \mathbb{Z})$.

Theorem 3.15 (Mixing properties of gbVAR processes)
Let $(X_t, t \in \mathbb{Z})$ be a stationary K-dimensional gbVAR(p) process with $p \geq 1$ and $Var(X_{t,k}) > 0$ for all $k = 1, \ldots, K$. Then the process is ψ- and φ- mixing with exponentially decreasing weights $(f_n, n \in \mathbb{N})$ i.e. there exists an $a > 0$ and $\rho \in (0,1)$ such that $f_n = a\rho^n$.

3.2.5. Parameter estimation in gbVAR models

The joint distribution of a gbVAR process is fully determined by the marginal distribution of the i.i.d. innovations $(e_t, t \in \mathbb{Z})$ and by the model parameters in $\mathcal{P}$. In view of the second-order dependence structure of gbVAR processes, the Yule-Walker equations derived in Theorems 3.6 and 3.7 constitute an important link between the mean vector μ_X, the autocovariance function Γ_X, the gbVAR coefficients in $\mathcal{P}$, which are all population quantities of the process $(X_t, t \in \mathbb{Z})$, and innovation mean μ_e and variance Σ_e, which are population quantities of the innovation process. Here, μ_X and $\Gamma_X(h)$ are easily estimable from a gbVAR data sample $X_1, \ldots, X_n$ by their sample versions

$$\widehat{\mu}_X := \overline{X} = \frac{1}{n} \sum_{t=1}^{n} X_t, \tag{3.45}$$

$$\widehat{\Gamma}_X(h) := \begin{cases} \frac{1}{n} \sum_{t=1}^{n-h} (X_{t+h} - \overline{X})(X_t - \overline{X})', & 0 \leq h < n \\ 0, & h \geq n \end{cases} \tag{3.46}$$

and $\Gamma_X(-h) := \Gamma_X'(h)$ for $h < 0$. If the K-dimensional process $(X_t, t \in \mathbb{Z})$ has the non-reducible state space $\mathcal{V}_X = \{0,1\}^K$, we can use the Yule-Walker equation system (3.31) to construct the well-known Yule-Walker estimator $[\widehat{\mathcal{A}}^{(1)}, \ldots, \widehat{\mathcal{A}}^{(p)}]$ for $[\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(p)}]$ by replacing the ACF Γ_X by the sample ACF $\widehat{\Gamma}_X$. Otherwise, i.e. if the process has a reducible state space, we have to apply the Yule-Walker estimation to the corresponding lower-dimensional process $(Y_t, t \in \mathbb{Z})$ with non-reducible state space $\mathcal{V}_Y$ as introduced in Section 3.2.3, which is identified.

In the following, we assume without loss of generality, that $(X_t, t \in \mathbb{Z})$ has a non-reducible state space. Then, the Yule-Walker estimator (see e.g. Lütkepohl (2005)[eq. (3.3.17)]) is

$$[\widehat{\mathcal{A}}^{(1)}, \ldots, \widehat{\mathcal{A}}^{(p)}] = [\widehat{\Gamma}_X(1), \ldots, \widehat{\Gamma}_X(p)] \begin{pmatrix} \widehat{\Gamma}_X(i-j) \\ i,j = 1, \ldots, p \end{pmatrix}^{-1} \tag{3.47}$$

with $\widehat{\mathcal{A}}^{(i)} = (\widehat{\alpha}_{kl}^{(i)})_{k,l=1,\ldots,K}$. As we imposed diagonality of $\mathcal{B}$ to achieve identification of Σ_e, by using the natural restriction $\beta_{kk} = 1 - \sum_{i=1}^{p} \sum_{l=1}^{K} |\alpha_{kl}^{(i)}|$ for $k = 1, \ldots, K$, we get immediately the estimator $\widehat{\mathcal{B}}$ defined by

$$\widehat{\mathcal{B}} := I_K - diag\left([\widehat{\mathcal{A}}^{(1)}, \ldots, \widehat{\mathcal{A}}^{(p)}] \mathbb{1}_{Kp} \right). \tag{3.48}$$

For further discussions about the diagonal structure of $\mathcal{B}$, see Remark 3.13. From Lemma 3.3, we get an estimator $\widehat{\mu}_e$ for the innovation mean μ_e by rearranging equation (3.21) and plugging in the sample versions of $\mathcal{A}^{(i)}$, $\mathcal{A}^{(-,j)}$, $\mathcal{B}$ and μ_X to get

$$\widehat{\mu}_e = \widehat{\mathcal{B}}^{-1} \left(\left[I_K - \sum_{i=1}^{p} \widehat{\mathcal{A}}^{(i)} \right] \widehat{\mu}_X - \sum_{j=1}^{p} \widehat{\mathcal{A}}^{(-,j)} \mathbb{1}_K \right), \tag{3.49}$$

where $\widehat{\mathcal{A}}^{(-,i)} := (|\widehat{\alpha}_{kl}^{(i)}| \mathbb{1}_{\{\widehat{\alpha}_{kl}^{(i)} < 0\}})_{k,l=1,\ldots,K}$, $i = 1, \ldots, p$ analogous to (3.20).

In scenarios where (some) diagonal elements $\widehat{\beta}_{kk}$ equal zero such that $\widehat{\mathcal{B}}$ is no longer invertible, the corresponding $\mu_{e,k}$'s are not identified, but the remaining mean parameters are still identified via an equation similar to Lemma 3.3 of reduced dimension; see also Remark 3.16 below. The *diagonal* of Σ_e, i.e. $I_K \circ \Sigma_e$ can be estimated by

$$\widehat{I_K \circ \Sigma_e} := diag(\widehat{\mu}_{e,i}(1 - \widehat{\mu}_{e,i}), i = 1, \ldots, K). \tag{3.50}$$

Moreover, by using the Yule-Walker equation (3.32) from Theorem 3.7, it is also possible to construct an estimator for the *non-diagonal* elements of Σ_e, i.e. for $(\mathbb{1}_{K\times K} - I_K) \circ \Sigma_e$. Such an estimator is obtained by replacing the last line of (3.32) by (3.33), separating $(\mathbb{1}_{K\times K} - I_K) \circ \Sigma_e$ on one side of the equation (achieved by left- and right-multiplication with $\mathcal{B}^{-1}$) and replacing all population quantities on the other side of the equation by their sample versions proposed above.

Finally, transition probabilities derived in Lemma 3.9 can be estimated in a similar fashion by replacing population quantities by the corresponding estimators to get

$$\widehat{P}\big(X_t = s_0 | X_{t-1} = s_1, \ldots, X_{t-p} = s_p\big) = \widehat{p}_{s_0|s_1,\ldots,s_p} = \tag{3.51}$$

$$\sum_{r_0 \in \{0,1\}^K} \widehat{p}_{r_0} \prod_{k=1}^{K} \left[\sum_{i=1}^{p} \sum_{l=1}^{K} |\widehat{\alpha}_{kl}^{(i)}| \left[\mathbb{1}_{\{\widehat{\alpha}_{kl}^{(i)} \geq 0\}} \delta_{s_{0,k} s_{i,l}} + \mathbb{1}_{\{\widehat{\alpha}_{kl}^{(i)} < 0\}} \delta_{s_{0,k}\left(1 - s_{i,l}\right)} \right] + \widehat{\beta}_{kk} \delta_{s_{0,k} r_{0,k}} \right],$$

where

$$\widehat{p}_{r_0} = \left(\prod_{\substack{k=1 \\ k: r_{0,k}=1}}^{K} \widehat{\mu}_{e,k} \right) \left(\prod_{\substack{k=1 \\ k: r_{0,k}=0}}^{K} (1 - \widehat{\mu}_{e,k}) \right) \tag{3.52}$$

in the case where e_t consists of mutually independent Bernoulli random variables. In the dependent case, the estimator for off-diagonal elements of Σ_e, i.e. $(\mathbb{1}_{K\times K} - I_K) \circ \Sigma_e$, derived above can be used to allow also for dependent Bernoulli random variables to incorporate linear dependence.

Remark 3.16 (Estimation outside of the parameter space)
Estimation of the gbVAR parameters $\mathbf{A} := [\mathcal{A}^{(1)}, \dots, \mathcal{A}^{(p)}]$ *using the Yule-Walker estimator* $\widehat{\mathbf{A}} := [\widehat{\mathcal{A}}^{(1)}, \dots, \widehat{\mathcal{A}}^{(p)}]$ *can lead to invalid parameters with*

$$[\widehat{\mathcal{A}}^{(1)}_{|\cdot|}, \dots, \widehat{\mathcal{A}}^{(p)}_{|\cdot|}]\mathbb{1}_{Kp} \notin [0,1]^K \tag{3.53}$$

such that at least for one row $k_0 \in \{1, \dots, K\}$*, we have*

$$(|\widehat{\alpha}^{(1)}_{k_0,\bullet}|, \dots, |\widehat{\alpha}^{(p)}_{k_0,\bullet}|)\mathbb{1}_{Kp} = \sum_{i=1}^{p}\sum_{l=1}^{K} |\widehat{\alpha}^{(i)}_{k_0,l}| > 1. \tag{3.54}$$

Let K^* *denote the number of such rows and by* $\mathbf{K}$ *the set of all indexes* $k_0 \in \{1, \dots, K\}$ *satisfying* (3.54)*. Then, we propose to estimate a constraint model that assures valid parameters* $[\widehat{\mathcal{A}}^{(1)}_{|\cdot|}, \dots, \widehat{\mathcal{A}}^{(p)}_{|\cdot|}]\mathbb{1}_{Kp} \in [0,1]^K$*; see Lütkepohl (2005)[Section 5.2]. For this purpose, we construct a* $K^* \times (K^2p + K)$ *matrix* C *of rank* K^* *by first defining the auxiliary matrix*

$$\widetilde{c}_{k_0} = \left(\left(\mathbb{1}_{\{\mathbf{A}_{k,m} \geq 0\}} - \mathbb{1}_{\{\mathbf{A}_{k,m} < 0\}}\right)\delta_{\{k_0 = k\}}\right)_{k=1,\dots,K, \quad m=1,\dots,Kp}$$

of dimension $K \times Kp$ *for each* $k_0 \in \mathbf{K}$*. These matrices* $\widetilde{c}_{k_0}$ *have only entries of* -1 *and* 1 *in row* $k = k_0$ *and zero otherwise. Then, the matrix* C *combines the vectorized auxiliary matrices for every* $k_0 \in \mathbf{K} := \{k_{0,1}, \dots, k_{0,K^*}\}$ *by*

$$C := \begin{pmatrix} vec(\widetilde{c}_{k_{0,1}})' \\ \vdots \\ vec(\widetilde{c}_{k_{0,K^*}})' \end{pmatrix}.$$

Furthermore, we define $Z = \left[Z_{(p-1)}, \dots, Z_T\right]$ *containing the process variables* $Z_t := vec\,[X_t, \dots, X_{t-p+1}]$ *and* $\beta = vec\left(\left[\mathcal{A}^{(1)}, \dots, \mathcal{A}^{(p)}\right]\right)$*, thus* $\widehat{\beta}$ *is the vectorized estimated parameter matrix from the Yule-Walker estimator* (3.47)*. The constraint is of a row sum equal to one has influence by the vector* $c = \mathbb{1}_{K*}$*. With these components, the constraint estimator results by*

$$\widehat{\beta}_{valid} = \widehat{\beta} + \left[\left(ZZ'\right)^{-1}\right] C' \left[C\left(ZZ'\right)^{-1} C'\right]^{-1} \left(c - C\widehat{\beta}\right). \tag{3.55}$$

Re-vectorizing $\widehat{\beta}_{valid}$ *leads to the constraint parameter estimator* $\widehat{\mathbf{A}}_{valid}$ *and* $\widehat{\mathcal{B}}_{valid}$ *is calculated as described in* (3.48)*. Since* $\widehat{\mathcal{B}}_{valid}$ *contains now* K^* *zero diagonal entries, the innovation process* $(e_t, t \in \mathbb{Z})$ *is not identified for those rows. The remaining entries of* $\widehat{\mu}_e$ *can be estimated by a reduced system of dimension* $K - K^*$ *consisting of all rows with* $\widehat{\beta}_{valid,kk} \neq 0$*.*

3.3. Simulation Study

In this section, we investigate the performance of Yule-Walker-based estimators in gbVAR models as described in Section 3.2.5 by Monte-Carlo simulations. To illustrate the estimation performance in several gbVAR model setups, we consider a) the (average) mean squared error (MSE) of different parameter estimators and b) the (average) mean absolute deviation error (MADE) of transition probability estimators.

To construct confidence intervals from parameter estimators, we propose a parametric gbVAR-bootstrap method, and c) investigate its accuracy by means of (average) coverage rates.

For this purpose, we consider three different gbVAR(p) setups with orders $p = 1, 2$ and of dimensions $K = 3, 4$ for sample sizes $n = 100, 500, 1000$ to be able to judge the performance of parameter estimation in several gbVAR model specifications. Precisely, we consider data generating processes (DGPs) with the following specifications:

(DGP1) gbVAR(1) with $K = 3$, $\mu_e = (0.48, 0.52, 0.47)'$,

$$\mathcal{A}^{(1)} = \begin{pmatrix} 0.15 & -0.25 & 0.49 \\ -0.19 & 0.27 & 0.28 \\ 0.17 & -0.39 & 0.21 \end{pmatrix} \quad \text{and} \quad \mathcal{B} = diag\,(0.11, 0.26, 0.23)$$

(DGP2) gbVAR(1) with $K = 4$, $\mu_e = (0.48, 0.52, 0.47, 0.33)'$,

$$\mathcal{A}^{(1)} = \begin{pmatrix} -0.18 & 0.25 & -0.19 & -0.15 \\ 0.33 & -0.23 & 0.18 & -0.18 \\ -0.27 & -0.29 & 0.21 & -0.11 \\ 0.08 & 0.15 & -0.21 & -0.32 \end{pmatrix} \quad \text{and} \quad \mathcal{B} = diag\,(0.23, 0.08, 0.12, 0.24)$$

(DGP3) gbVAR(2) with $K = 3$, $\mu_e = (0.48, 0.52, 0.47)'$,

$$\mathcal{A}^{(1)} = \begin{pmatrix} -0.09 & 0.15 & -0.13 \\ 0.13 & -0.11 & 0.28 \\ 0.13 & -0.19 & -0.18 \end{pmatrix}, \quad \mathcal{A}^{(2)} = \begin{pmatrix} -0.18 & 0.07 & -0.19 \\ -0.09 & -0.17 & 0.15 \\ -0.17 & -0.09 & 0.14 \end{pmatrix},$$

$$\text{and} \quad \mathcal{B} = diag\,(0.19, 0.07, 0.10)$$

For all DGPs, the corresponding innovation process $(e_t, t \in \mathbb{Z})$ consists of K independent Bernoulli processes $(e_{t,k}, t \in \mathbb{Z})$, $k = 1, \ldots, K$ with $\mu_{e,k} = P\,(e_{t,k} = 1)$ leading to diagonal Σ_e matrices with diagonal entries $\mu_{e,k}(1 - \mu_{e,k})$, $k = 1, \ldots, K$. Note that we make use of positive as well as negative entries in $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(2)}$. Hence, these coefficient matrices are related to the diagonal matrix $\mathcal{B}$ via $\beta_{kk} = 1 - \sum_{i=1}^{p} \sum_{l=1}^{K} |\alpha_{kl}^{(i)}|$. In Section 3.3.2, we address also the estimation of the off-diagonal elements of a non-diagonal variance-covariance matrix Σ_e.

3.3.1. Average MSE estimation performance

To measure the estimation performance, we calculate averages of the entry-wise mean squared errors (MSE) of the estimators $\widehat{\mathcal{A}}^{(1)}$ and $\widehat{\mathcal{A}}^{(2)}$, $\widehat{\mu}_X$, $\widehat{\mu}_e$ and $\widehat{\mathcal{B}}$, respectively, based on 1000 Monte-Carlo replications for each DGP and each sample size. The simulation results are presented in Table 3.1. It can be seen that the estimation performance improves with increasing sample size for all estimators and all DGPs. In comparison, the estimation of the mean innovation vector is least precise with an average mean squared error around 10 percent. This phenomenon can be explained by formula (3.49), which requires the inversion of the diagonal matrix $\widehat{\mathcal{B}}$. Due to rather small diagonal entries of $\mathcal{B}$, already small deviations in $\widehat{\mathcal{B}}$ might lead to a less stable estimation of μ_e and to a larger MSE.

	n	MSE of $\widehat{\mathcal{A}}^{(1)}$	MSE of $\widehat{\mathcal{A}}^{(2)}$	MSE of $\widehat{\mu}_e$	MSE of $\widehat{\mu}_X$	MSE of $\widehat{\mathcal{B}}$
DGP1	100	0.0085		0.0626	0.0046	0.0214
	500	0.0017		0.0152	0.0009	0.0034
	1000	0.0008		0.0070	0.0005	0.0015
DGP2	100	0.0085		0.0794	0.0022	0.0426
	500	0.0017		0.0388	0.0004	0.0085
	1000	0.0008		0.0208	0.0002	0.0035
DGP3	100	0.0084	0.0083	0.1041	0.0015	0.0821
	500	0.0018	0.0018	0.0701	0.0003	0.0374
	1000	0.0009	0.0009	0.0502	0.0002	0.0198

Table 3.1.: Average MSE estimation performance for different parameter estimators $\widehat{\mathcal{A}}^{(1)}$, $\widehat{\mathcal{A}}^{(2)}$, $\widehat{\mu}_X$, $\widehat{\mu}_e$ and $\widehat{\mathcal{B}}$ for three different parameter specifications DGP1, DGP2 and DGP3, respectively.

3.3.2. Average MSE estimation of non-diagonal Σ_e

As discussed in Remark 3.13, the imposed diagonality of $\mathcal{B}^{(0)}$ does generally allow to identify also the non-diagonal entries of Σ_e. These can be estimated by using the Yule-Walker equation for $h = 0$ from Theorem 3.7 in conjunction with (3.33), where the corresponding estimator is obtained by replacing all population quantities by their sample analogues as described in Section 3.2.5.

For illustration, we consider a bivariate gbVAR(1) process, where we used the four entries $\alpha_{12}, \alpha_{13}, \alpha_{32}$ and α_{33} in $\mathcal{A}$ of DGP1 leading to $\mathcal{B} = diag(0.26, 0.4)$. The bivariate marginal distribution of the innovations $(e_t, t \in \mathbb{Z})$ is fully specified by $\mu_{e,1} = P(e_{t,1} = 1) = 0.260$, $\mu_{e,2} = P(e_{t,2} = 1) = 0.382$ and

$$\Sigma_e = \begin{pmatrix} 0.19240 & 0.134680 \\ 0.13468 & 0.236076 \end{pmatrix},$$

where Σ_e determines the joint probabilities $P\big((e_{t,1}, e_{t,2}) = (i, j)\big)$ for $i, j \in \{0, 1\}$. Practically, such an innovation process can be generated by first sampling the innovation $e_{t,1}$ from a Bernoulli distribution with $P(e_{t,1} = 1) = 0.26$. Then, we generate $e_{t,2}$ conditional on the outcome of $e_{t,1}$ such that $P(e_{t,2} = 1 | e_{t,1} = 1) = 0.9$ and $P(e_{t,2} = 1 | e_{t,1} = 0) = 0.2$. This leads to a marginal Bernoulli distribution of $e_{t,2}$ with $P(e_{t,2} = 1) = 0.9 \cdot 0.26 + 0.2 \cdot (1 - 0.26) = 0.382$. For a comprehensive discussion of multivariate Bernoulli distributions allowing for dependence also beyond $K = 2$, we refer to Dai et al. (2013).

In Table 3.2, we report the MSE for the diagonal and non-diagonal elements of $\widehat{\Sigma}_e$ for different sample sizes. It can be seen that the MSE decays for increasing sample size. However, for a small sample size of $n = 100$, the MSE of the off-diagonal elements $\widehat{\sigma}_{e,12}$ is huge with 16.7383. This value is caused by a mis-estimation for some few Monte-Carlo replications, where the estimated parameters already show a large MSE. Due to a matrix inversion, this leads to unstable and unreliable estimates. Nevertheless, this issue disappears for larger sample sizes such that the joint distribution of the innovation process in form of the non-diagonal entries of Σ_e can be consistently estimated. However, as $\mathcal{B}$ is imposed to be diagonal for identification reasons, the non-diagonal entries of Σ_e do not have large effects on the stochastic properties of the

n	$\widehat{\sigma}_{11}$	$\widehat{\sigma}_{12}$,$\widehat{\sigma}_{21}$	$\widehat{\sigma}_{22}$
100	0.0100	16.7383	0.0028
500	0.0025	0.0101	0.0002
1000	0.0009	0.0045	0.0001

Table 3.2.: Mean squared error of diagonal and non-diagonal elements of $\widehat{\Sigma}_e$

model. Hence, in practice, it seems to be recommendable to avoid the estimation of a non-diagonal Σ_e due to potentially unstable estimation results.

3.3.3. Average MADE estimation performance

An alternative concept to measure the estimation performance in gbVAR models is based on (average) mean absolute deviation error (MADE) of transition probability estimators. Note that a direct comparison of prediction probabilities in $[0,1]$ and outcomes in $\{0,1\}$ are not straightforward and might be misleading to judge the parameter estimation performance. Hence, we compare the (one step ahead) population transition probabilities with the corresponding estimated transition probabilities and consider

$$\begin{aligned} &|p_{s_0|s_1,\ldots,s_p} - \widehat{p}_{s_0|s_1,\ldots,s_p}| \\ &= |P\left(X_t = s_0 | X_{t-1} = s_1, \ldots, X_{t-p} = s_p\right) - \widehat{P}\left(X_t = s_0 | X_{t-1} = s_1, \ldots, X_{t-p} = s_p\right)| \end{aligned} \tag{3.56}$$

with $p_{s_0|s_1,\ldots,s_p} = P(X_t = s_0 | X_{t-1} = s_1, \ldots, X_{t-p} = s_p)$ as obtained in Theorem 3.9 and $\widehat{p}_{s_0|s_1,\ldots,s_p} = \widehat{P}(X_t = s_0 | X_{t-1} = s_1, \ldots, X_{t-p} = s_p)$ as constructed in (3.51) for the special case, where $e_{t,1}, \ldots, e_{t,K}$ such that $p_{r_0} = P(e_t = r_0) = \prod_{k=1}^{K} P(e_{t,k} = r_{0,k})$. Mainly, there are two possibilities to use (3.56) to judge the average estimation accuracy in gbVAR models. The first one, considers the absolute deviation of the transition probabilities according to their actual appearances in the Monte Carlo sample under consideration. That is, given $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$, we compute

$$\frac{1}{n-p} \sum_{t=p+1}^{n} |p_{x_t|x_{t-1},\ldots,x_{t-p}} - \widehat{p}_{x_t|x_{t-1},\ldots,x_{t-p}}|. \tag{3.57}$$

The second option is to calculate the absolute deviate of the transition probabilities over all possible states of $s_0, \ldots, s_p$ in the state space $\{0,1\}^K$ leading to

$$\frac{1}{2^{K(p+1)}} \sum_{s_0,s_1,\ldots,s_p \in \{0,1\}^K} |p_{s_0|s_1,\ldots,s_p} - \widehat{p}_{s_0|s_1,\ldots,s_p}|. \tag{3.58}$$

Based on 1000 Monte-Carlo samples, we report the estimation performance using both versions (3.57) and (3.58) of average MADE for each DGP and each sample size in Table 3.3.

3.3.4. Parametric gbVAR bootstrapping

To construct confidence intervals for the gbVAR model parameters, we propose to use a parametric gbVAR bootstrap. Autoregressive-type bootstrap methods based on resampling residuals from an AR model fit have been extensively studied in the time

	average MADE (3.57)			average MADE (3.58)		
n	DGP1	DGP2	DGP3	DGP1	DGP2	DGP3
100	0.0341	0.0358	0.0196	0.0338	0.0169	0.0179
500	0.0154	0.0151	0.0082	0.0149	0.0077	0.0076
1000	0.0108	0.0106	0.0054	0.0104	0.0054	0.0050

Table 3.3.: Estimation performance based on both versions of average MADE in (3.57) and (3.58) for DGP1, DGP2 and DGP3, respectively.

series literature; see e.g. Kreiß and Paparoditis (2011) for an overview. However, due to randomness in the coefficients of gbVAR models, it is not possible to construct meaningful residuals from gbVAR fits; see also the discussion in Jentsch and Weiß (2019) for a similar setup addressing integer-valued autoregressive (INAR) models. Nevertheless, for DGP1 - DGP3, it is possible to estimate consistently all parameters that determine the distribution of the innovation process $(e_t, t \in \mathbb{Z})$ from time series data $X_1, \ldots, X_n$ using the mean formula (3.21) which allows a straightforward parametric gbVAR bootstrap to construct confidence intervals as follows.

gbVAR bootstrap

Step 1.) Given a data sample $X_1, \ldots, X_n$ from a gbVAR(p) model, compute the (scalar) statistic of interest $T_n = T_n(X_1, \ldots, X_n)$ and estimate the model parameters $\mathcal{P} = [\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(p)}, \mathcal{B}]$ and μ_e as described in Section 3.2.5 leading to $\widehat{\mathcal{P}} = [\widehat{\mathcal{A}}^{(1)}, \ldots, \widehat{\mathcal{A}}^{(p)}, \widehat{\mathcal{B}}]$ and $\widehat{\mu}_e$ as well as $\widehat{\mathcal{P}}_{|\cdot|} = [\widehat{\mathcal{A}}^{(1)}_{|\cdot|}, \ldots, \widehat{\mathcal{A}}^{(p)}_{|\cdot|}, \widehat{\mathcal{B}}]$.

Step 2.) Generate a gbVAR(p) bootstrap sample $X_1^*, \ldots, X_n^*$ according to

$$X_t^* = \sum_{i=1}^{p} \left[A_t^{(+,i)*} X_{t-i}^* + A_t^{(-,i)*} \mathbb{1}_K \right] + B_t^* e_t^*, \quad t \in \mathbb{Z}, \tag{3.59}$$

where $\{e_{t,k}^*, t \in \mathbb{Z}\}$, $k = 1, \ldots, K$, are mutually independent i.i.d. Bernoulli variables with $P^*(e_{t,k}^* = 1) = \widehat{\mu}_{e,k}$. Further, conditional on $X_1, \ldots, X_n$, $(P_{t,k\bullet}^*, t \in \mathbb{Z})$, $k = 1, \ldots, K$ are mutually independent vector-valued i.i.d. processes where

$$P_{t,k\bullet}^* := \left[a_{t,k\bullet}^{(1)*}, \ldots, a_{t,k\bullet}^{(p)*}, b_{t,k\bullet}^* \right] \sim Mult\left(1; \widehat{\mathcal{P}}_{|\cdot|,k\bullet}\right),$$

which are independent of $(e_t^*, t \in \mathbb{Z})$ and $(X_s^*, s < t)$. Here, $A_t^{(+,i)*}$ and $A_t^{(-,i)*}$ are defined analogously to $A_t^{(+,i)}$ and $A_t^{(-,i)}$ as in Definition 3.1.

Step 3.) Compute $T_n^* = T_n(X_1^*, \ldots, X_n^*)$.

Step 4.) Repeat Steps 2.) and 3.) B times, where B is large, to get $T_{1,n}^*, \ldots, T_{B,n}^*$.

Step 5.) Let $q_{\alpha/2}^*$ and $q_{1-\alpha/2}^*$ denote the empirical $\alpha/2$ - and $1-\alpha/2$-quantiles of $T_{1,n}^* - T_n, \ldots, T_{B,n}^* - T_n$, respectively, and construct a $(1-\alpha)$-confidence interval of the form $\left[T_n - q_{1-\alpha/2}^*, T_n - q_{\alpha/2}^* \right]$.

	$T = 100$	500	1000		$T = 100$	500	1000
$\alpha_{11}^{(1)}$	0.933	0.958	0.949	$\beta_{11}^{(0)}$	0.886	0.946	0.949
$\alpha_{12}^{(1)}$	0.937	0.951	0.942	$\beta_{22}^{(0)}$	0.866	0.917	0.956
$\alpha_{13}^{(1)}$	0.922	0.946	0.943	$\beta_{33}^{(0)}$	0.899	0.927	0.926
$\alpha_{21}^{(1)}$	0.927	0.941	0.950				
$\alpha_{22}^{(1)}$	0.929	0.957	0.922				
$\alpha_{23}^{(1)}$	0.939	0.944	0.955				
$\alpha_{31}^{(1)}$	0.936	0.940	0.958	$\mu_{X,1}$	0.886	0.943	0.951
$\alpha_{32}^{(1)}$	0.931	0.939	0.947	$\mu_{X,2}$	0.910	0.947	0.957
$\alpha_{33}^{(1)}$	0.948	0.949	0.948	$\mu_{X,3}$	0.900	0.956	0.927

Table 3.4.: Coverage rates of bootstrap confidence intervals for each entry of the parameter matrices $\mathcal{A}^{(1)}$, $\mathcal{B}^{(0)}$ and μ_X for DGP1.

In Table 3.4, we report the results of a simulation study based on 1000 Monte Carlo replications and $B = 1000$ bootstrap replications for DGP1 and each sample size, where we show coverage rates of $(1 - \alpha)$ bootstrap confidence intervals for all entries of $\mathcal{A}^{(1)}$, $\mathcal{B}^{(1)}$ and μ_X for level $\alpha = 0.05$. The coverage rates for DGP2 and DGP3 show a similar behavior and can be found in Appendix 3.9.

3.4. Real Data Example: PM10 data

With increasing environmental awareness, there has been great interest in collecting and analyzing data describing the extent of the pollution of the environment and its impact on the health of the population. In recent years, there is particularly growing interest in air pollution with particulate matter in European cities and metropolitan areas. The European Union established the *European emission standards*, which include limits for particulates in the air. In 2008, the European Parliament made a policy on critical values in the rule 2008/50/EG for air pollution substances. In particular, it is by law not allowed to exceed the threshold on 35 or more days per year. For particulate matter PM_{10} (coarse particles with a diameter between 2.5 and 10 micrometers), the liability has a threshold of 50 µg / m^3. Hence, whenever the amount of PM_{10} exceeds the threshold of 50 µg/m^3 at a certain monitoring station, this will cause a '*fine dust alarm*'. Hence, for each such monitoring station, this results in a binary sequence with states 'exceedance' and 'no exceedance'. In fact, the current public discourse centers to a large extent around whether the threshold is exceeded or not, and less about the actual amount of fine dust measured.

In view of these EU regulations, Stuttgart, Germany is one poorly prominent city, where air pollution generally is a major problem. The reasons for these problems are essentially two-fold. On the one hand, they can be explained by its geographic location in a valley leading to a poor air exchange in the city area. On the other hand, the main industry such as automobile companies and suppliers as well as financial industry is located near to the city center. Due to the restricted space in a valley to expand, many people live in suburbs of Stuttgart and have to commute to their work places. The commuting traffic concentrates on few main traffic routes, which are highly frequented during rush hours. Hence, large portions of particulate matter in the air in and around

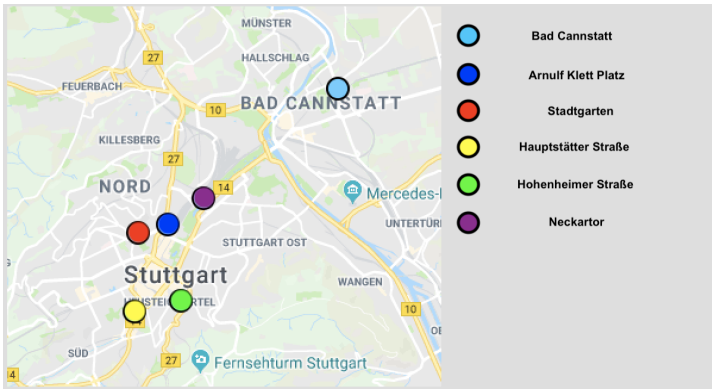


Figure 3.6.: Locations of six major PM_{10} monitoring stations in Stuttgart, Germany

Stuttgart is caused by individual mobility.

In Figure 3.2, we show the recorded fine dust alarms at six monitoring stations in Stuttgart, Germany for 886 consecutive days from 03/01/2016 - 07/31/2018. Precisely, with $Y_{t,k}$, $k = 1, \ldots, 6$, representing the day-wise mean of the PM_{10} values for each station, Figure 3.2 shows binarized time series data $(X_t = (X_{t,1}, \ldots, X_{t,6}), t \in \mathbb{Z})$, where

$$X_{t,k} = \begin{cases} 1 & if \quad Y_{t,k} \geq 50 \\ 0 & if \quad Y_{t,k} < 50 \end{cases} \quad \text{for } k = 1, \ldots, 6.$$

The locations of the six monitoring stations Arnulf-Klett Platz, Bad Cannstatt, Hauptstätter Straße, Hohenheimer Straße, Neckartor and Stadtgarten are illustrated in Figure 3.6.

A first inspection of the data in Figure 3.2 shows that fine dust alarms tend to occur in clusters indicating serial and cross-sectional dependence. However, all sequences do not show long runs of fine dust alarms, but rather long runs without any alarm. Moreover, fine dust alarms tend to show more likely in winter. This is due to the fact that the topological influence of stationary temperature inversion hinders vertical air exchange. One station (Neckartor, $k = 5$) shows considerable more exceedances in comparison to the other stations with fine dust alarms occurring in about 13% of the days in the considered time period. This is captured by the sample mean vector

$$\hat{\mu}_X = (0.0420, 0.0227, 0.0397, 0.0386, 0.1283, 0.0249)'.$$

The mean of all the other stations lie around 3 % indicating that only few fine dust alarms are detected. Overall, this is not surprising, as the station Neckartor is located at one of the most frequented roads of Stuttgart, where high buildings on one side of the road hinder the air exchange and favor air pollution. In contrast, the station Bad Cannstatt with the smallest value of 0.0227 is located at an accommodation route to the city outside of the city center.

Now, to study the serial dependence in the data, we aim to fit a gbVAR(p) model. As gbVAR processes satisfy standard Yule-Walker equations and can be estimated by Yule-Walker estimators as described in (3.47), we can make use of classical order selection criteria such as Hannan-Quinn (HQ) or BIC to determine an appropriate order p of the fitted gbVAR process. Whereas HQ selects a more parsimonious model with p=1, BIC leads to $p = 2$. To make a choice which model fits best in terms of

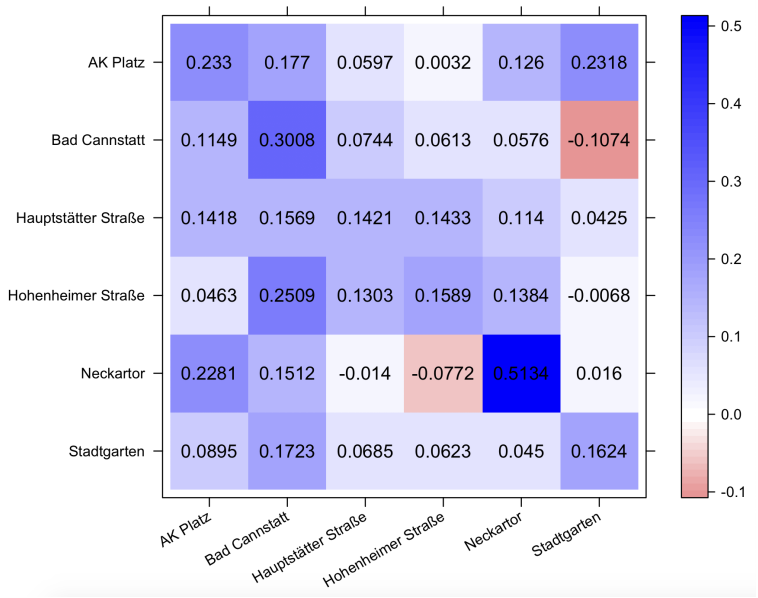


Figure 3.7.: Heatmap of the estimated parameter matrix $\widehat{\mathcal{A}}$ for a fitted gbVAR(1) process

prediction performance, we use the receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC), where an AUC near to one indicates good prediction performance. For this purpose, similar to Section 3.3.3, we make use of transition probability estimators $\widehat{P}(X_t = s_0 | X_{t-1} = s_1, \ldots, X_{t-p} = s_p)$ as constructed in (3.51) to estimate the transition probabilities for each station, which allows to compute the ROC curves and AUC values. In Table 3.5, we show the resulting component-by-component AUC values and their overall means for model orders $p \in \{1, 2\}$. Both models show a good prediction performance with AUC values near to one. However, the additional benefit of fitting a gbVAR(2) model in comparison to a more parsimonious gbVAR(1) model is minor. Hence, we make use of a gbVAR(1) model in the following to further analyze the PM_{10} data set.

	AK	BC	HS	HH	NT	SG	mean
p = 1	0.9368	0.9500	0.9337	0.9459	0.8022	0.9305	0.9164
p = 2	0.9388	0.9572	0.9352	0.9479	0.8111	0.9306	0.9201

Table 3.5.: AUC values for a fitted gbVAR(p) model for $p \in \{1, 2\}$ component- wise for each station and the overall mean

In this case, Yule-Walker estimation leads to an estimated parameter matrix having $\sum_{l=1}^{K} |\widehat{\alpha}_{kl}| = 1.1451 > 1$ for $k = 5$, which corresponds to station Neckartor. Hence, in view of Remark 3.16, we have to use constraint estimation leading to the estimated parameter matrix $\widehat{\mathcal{A}}$ as shown in Figure 3.7. The absolute eigenvalues of $\widehat{\mathcal{A}}$ compute to $\{0.7092, 0.2933, 0.2933, 0.1143, 0.1143, 0.0019\}$ such that the fitted gbVAR(1) model is stationary.

In each row of $\widehat{\mathcal{A}}$, we can see which past state at time $t-1$ of the six monitoring stations (fine dust alarm or not) does affect the state at time t. For example, with

51.33% probability, the station Neckartor takes the same value as the day before. In contrast, the largest entry in row $k = 4$ (Hohenheimer Strasse) is 0.2509 in the second column corresponding to Bad Cannstatt such that it takes its value of the day before with probability of about 25%. Note also, as can be seen in Figure 3.7, that the fitting of a gbVAR model leads to some negative coefficients in $\widehat{\mathcal{A}}$. From a modeling perspective, this naturally leads to more flexibility in comparison to models that do allow only for non-negative coefficients.

For identification purposes, we impose $\mathcal{B}$ to be diagonal; see also Remark 3.13. Since we constrained the estimation to achieve $\sum_{l=1}^{K} |\widehat{\alpha}_{5l}| = 1$ and set $\widehat{\beta}_{55} = 0$ for station Neckartor, the effective innovation process is of reduced dimension $\widetilde{K} = 6 - 1 = 5$. Hence, as described in Section 3.2.5, we get

$$diag\left(\widehat{\mathcal{B}}\right) = (0.1694, 0.2837, 0.2594, 0.2684, 0.0000, 0.3999)\,.$$

The diagonal entries of $\widehat{\mathcal{B}}$ indicate how often the corresponding innovation terms are selected. For example, at the station Stadtgarten, in about 40% of the days, the innovation term enters the gbVAR model, whereas this happens only in about 17% for Arnulf-Klett Platz.

By using formula (3.49) and (3.50), we can estimate the mean vector and the variances of the innovations. This leads to

$$\widehat{\mu}_e = (0.1176, 0.4084, 0.0693, 0.0993, 0.0000, 0.0203)'\,. \tag{3.60}$$

and

$$(\widehat{\sigma}_{e,11}, \ldots, \widehat{\sigma}_{e,KK}) = (0.1038, 0.2416, 0.0645, 0.0895, 0.000, 0.0199). \tag{3.61}$$

Hence, for PM_{10} data, (3.60) indicates that the innovation terms generally take zero values with high probability. Note that $\widehat{\mu}_{e,5}$ and $\widehat{\sigma}_{e,55}$ are not identified due to $\widehat{\beta}_{55} = 0$ and we set $\widehat{\mu}_{e,5} = \widehat{\sigma}_{e,55} = 0$ for convenience.

3.5. Conclusion

We consider vector-valued extensions of gbAR processes introduced by Jentsch and Reichmann (2019) to model multivariate binary time series data with potentially negative model parameters. We derive stationarity conditions that assure a moving-average-type representations of the stationary solution. Yule-Walker equations are derived that allow particularly for a straightforward estimation of gbVAR processes. Transition probabilities are derived and a Markov chain representation has been employed to derive ψ- and φ- mixing properties.

In a simulation study, the estimation performance of Yule-Walker estimators and related estimators is analyzed in several regards indicating good finite sample properties. For the construction of confidence intervals, we propose a parametric bootstrap resulting in coverage rates close to nominal coverage.

In a real data application, we fit gbVAR processes to binarized PM_{10} data from Stuttgart, Germany. The estimated gbVAR(1) model contains positive as well as negative coefficients to capture the serial dependence in the data and proves to yield accurate predictions.

3.6. The generalized binary VARMA Class

In this section, we extend the gbVAR model class and give a definition of generalized binary VARMA (gbVARMA) models that additionally contain a moving average part. To be most flexible, we allow for negative parameters to capture negative dependence structure also in the moving average part of gbVARMA models.

First, we provide the definition of the gbVARMA(p,q) model in Section 3.6.1. Then, we discuss the moment structure of gbVARMA models and derive its stationary solution before stating a formula for the transition probabilities in Section 3.6.2.

3.6.1. gbVARMA models

As for gbVAR processes considered in the previous section, we set $\mathcal{B}^{(0)} := \mathcal{B}$ corresponding to the current innovation e_t to be diagonal. Furthermore, to avoid identification issues, we assume that all diagonal entries $\beta_{kk}^{(0)} \in (0, 1]$ are *strictly positive* throughout this section[1]; see also the discussion in Remark 3.13.

As for the parameter matrices $\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(p)}$ of the autoregressive part, the parameter matrices $\mathcal{B}^{(1)}, \ldots, \mathcal{B}^{(q)}$ corresponding to the moving-average part are also allowed to contain negative entries. That is, similar to gbVAR models discussed in Section 3.2, we allow for $\alpha_{kl}^{(i)}, \beta_{kl}^{(j)} \in (-1, 1)$ for $k, l = 1, \ldots, K$, $i = 1, \ldots, p$ and $j = 1, \ldots, q$. Hence, the parameter matrix

$$\mathcal{P} := \left[\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(p)}, \mathcal{B}^{(0)}, \mathcal{B}^{(1)}, \ldots, \mathcal{B}^{(q)}\right] \tag{3.62}$$

has to be modified to contain (row-wise) valid probabilities of multinomial distributions. Hence, we define

$$\mathcal{P}_{|\cdot|} = \left[\mathcal{A}_{|\cdot|}^{(1)}, \ldots, \mathcal{A}_{|\cdot|}^{(p)}, \mathcal{B}^{(0)}, \mathcal{B}_{|\cdot|}^{(1)}, \ldots, \mathcal{B}_{|\cdot|}^{(q)}\right], \tag{3.63}$$

where $\mathcal{B}_{|\cdot|}^{(j)} = (|\beta_{kl}^{(j)}|)_{k,l=1,\ldots,K}$ for $j = 1, \ldots, q$, such that $\sum_{i=1}^{p}\sum_{l=1}^{K}|\alpha_{kl}^{(i)}| + \beta_{kk}^{(0)} + \sum_{j=1}^{q}\sum_{l=1}^{K}|\beta_{kl}^{(j)}| = 1$ for all $k = 1, \ldots, K$.

Definition 3.17 (gbVARMA(p, q))
Let $(X_t, t \in \mathbb{Z})$ be a stationary K-dimensional process taking values in $\{0, 1\}^K$. Let $(e_t, t \in \mathbb{Z})$ be an i.i.d. K-dimensional binary innovation process, such that e_t is independent of $(X_s, s < t)$ with mean vector $\mu_e = (\mu_{e,1}, \ldots, \mu_{e,K})' = E(e_t)$, where $\mu_{e,i} = P(e_{t,i} = 1)$, $i = 1, \ldots, K$, and variance-covariance matrix $\Sigma_e = (\sigma_{e,kl})_{k,l=1,\ldots,K} = Cov(e_t) > 0$,where $\sigma_{e,ii} = \mu_{e,i}(1 - \mu_{e,i})$. Let $\mathcal{P}$ be the parameter matrix as in (3.62) with diagonal $\mathcal{B}^{(0)}$ such that $\beta_{kk} \in (0, 1]$ for all $k = 1, \ldots, K$ and with $\mathcal{P}_{|\cdot|}$ as in (3.63) such that $\mathcal{P}_{|\cdot|}\mathbb{1}_{K(p+q+1)} = \mathbb{1}_K$.

Further, let $(P_t, t \in \mathbb{Z})$ with $P_t = [A_t^{(1)}, \ldots, A_t^{(p)}, B_t^{(0)}, B_t^{(1)}, \ldots, B_t^{(q)}]$ be a $K \times K(p + q + 1)$-dimensional i.i.d. process with mutually independent rows $(P_{t,k\bullet}, t \in \mathbb{Z})$, $k = 1, \ldots, K$, such that

$$P_{t,k\bullet} := \left[a_{t,k\bullet}^{(1)}, \ldots, a_{t,k\bullet}^{(p)}, b_{t,k\bullet}^{(0)}, b_{t,k\bullet}^{(1)}, \ldots, b_{t,k\bullet}^{(q)}\right] \sim Mult\left(1; \mathcal{P}_{|\cdot|,k\bullet}\right),$$

[1] Otherwise, if for example one diagonal entry is zero, this innovation might nevertheless enter the gbVARMA system in the first lag of the innovations, where the corresponding $\mathcal{B}$ matrix is not restricted to be diagonal which leads to identification issues.

which are independent of $(e_t, t \in \mathbb{Z})$ *and* $(X_s, s < t)$*. Here,* $A_t^{(i)}$ *and* $a_{t,k\bullet}^{(i)}$ *are as in Definition 3.1 with* $B_t^{(j)}$ *and* $b_{t,k\bullet}^{(j)}$ *defined similarly.*

Then the process $(X_t, t \in \mathbb{Z})$ *is said to be a generalized binary vector ARMA(p,q) process (gbVARMA(p,q)), if it follows the recursion*

$$X_t = \sum_{i=1}^{p} \left[A_t^{(+,i)} X_{t-i} + A_t^{(-,i)} \mathbb{1}_K \right] + B_t^{(0)} e_t + \sum_{j=1}^{q} \left[B_t^{(+,j)} e_{t-j} + B_t^{(-,j)} \mathbb{1}_K \right], \quad t \in \mathbb{Z}, \tag{3.64}$$

with $A_t^{(+,i)}$ *and* $A_t^{(-,i)}$ *as in Definition 3.1 and*

$$B_t^{(+,j)} := \left\{ \begin{array}{ll} b_{t,kl}^{(j)}, & \beta_{kl}^{(j)} \geq 0 \\ -b_{t,kl}^{(j)}, & \beta_{kl}^{(j)} < 0 \end{array} \right\}_{k,l=1,\ldots,K} = \left(b_{t,kl}^{(j)} \left(\mathbb{1}_{\{\beta_{kl}^{(j)} \geq 0\}} - \mathbb{1}_{\{\beta_{kl}^{(j)} < 0\}} \right) \right)_{k,l=1,\ldots,K}$$

and

$$B_t^{(-,j)} := \left\{ \begin{array}{ll} 0, & \beta_{kl}^{(j)} \geq 0 \\ b_{t,kl}^{(j)}, & \beta_{kl}^{(j)} < 0 \end{array} \right\}_{k,l=1,\ldots,K} = \left(b_{t,kl}^{(j)} \mathbb{1}_{\{\beta_{kl}^{(j)} < 0\}} \right)_{k,l=1,\ldots,K}$$

for $j = 1, \ldots, q$*.*

The selection mechanism of gbVARMA processes proceeds as follows: for each $k \in \{1, \ldots, K\}$, depending on the signs of $\alpha_{kl}^{(i)}$ and $\beta_{kl}^{(j)}$, the value of $X_{t,k}$ is either chosen from the entries (or their opposites) of the lagged time series $X_{t-1}, \ldots, X_{t-p}$ with probability $\sum_{i=1}^{p} \sum_{l=1}^{K} |\alpha_{kl}^{(i)}|$ or from the entries (or their opposites) of the innovations $e_t, e_{t-1}, \ldots, e_{t-p}$ with probability $\beta_{kk}^{(0)} + \sum_{j=1}^{q} \sum_{l=1}^{K} |\beta_{kl}^{(j)}|$, respectively. More precisely, with probability $|\alpha_{kl}^{(i)}|$ the predecessor $X_{t-i,l}$ is chosen for the time series value $X_{t,k}$ if $\alpha_{kl}^{(i)} \geq 0$ and its opposite value $1 - X_{t-i,l}$ if $\alpha_{kl}^{(i)} < 0$. With probability $|\beta_{kl}^{(j)}|$ the innovation $e_{t-j,l}$ is chosen if $\beta_{kl}^{(j)} \geq 0$ and its opposite value $1 - e_{t-j,l}$ if $\beta_{kl}^{(j)} < 0$.

Since the multinomial selection mechanism is executed independently for each row, the outcomes X_{t,k_1} and X_{t,k_2} for $k_1 \neq k_2$ are independent *given* the past values of the time series $X_{t-1}, \ldots, X_{t-p}$ and the innovations $e_t, e_{t-1}, \ldots, e_{t-q}$.

3.6.2. Stochastic properties of gbVARMA models

Similar to the random matrices in (3.19) and (3.20), we can compute the expectations of $B_t^{(+,j)}$ and $B_t^{(-,j)}$, $j = 1, \ldots, q$. By construction, we have $E(B_t^{(0)}) = \mathcal{B}^{(0)} = diag(\beta_{11}^{(0)}, \ldots, \beta_{KK}^{(0)})$ and $E\big(B_t^{(j)}\big) = \mathcal{B}_{|\cdot|}^{(j)}$ leading to

$$E\left(B_t^{(+,j)}\right) = \left[\beta_{kl}^{(j)}\right]_{k,l=1,\ldots,K} = \mathcal{B}^{(j)}, \tag{3.65}$$

$$E\left(B_t^{(-,j)}\right) = \left[|\beta_{kl}^{(j)}| \mathbb{1}_{\{\beta_{kl}^{(j)} < 0\}}\right]_{k,l=1,\ldots,K} =: \mathcal{B}^{(-,j)}. \tag{3.66}$$

In comparison to the purely autoregressive gbVAR(p) case in (3.21), the additional parameter matrices $\mathcal{B}^{(1)}, \ldots, \mathcal{B}^{(q)}$ show up in the formula for the mean vector $\mu_X = E(X_t)$ of a gbVARMA(p,q) process. In detail, we have:

Lemma 3.18 (Stationary mean of gbVARMA processes)
Let $(X_t, t \in \mathbb{Z})$ be a stationary K-dimensional gbVARMA(p,q) process. Then, we have

$$\mu_X = \left(I - \sum_{i=1}^{p} \mathcal{A}^{(i)} \right)^{-1} \left(\sum_{i=1}^{p} \mathcal{A}^{(-,i)} \mathbb{1}_K + \sum_{j=0}^{q} \mathcal{B}^{(j)} \mu_e + \sum_{j=1}^{q} \mathcal{B}^{(-,j)} \mathbb{1}_K \right). \tag{3.67}$$

To construct the stationary solution of a gbVARMA model in form of a gbVMA(∞)-type representation we proceed similarly to the approach used in Section 3.2.2 for gbVAR(p) models. We follow again the approach of Lütkepohl (2005)[Chapter 11.3.2] and represent the K-dimensional gbVARMA(p,q) process $(X_t, t \in \mathbb{Z})$ as a gbVAR(1) process $(\check{X}_t, t \in \mathbb{Z})$ of dimension $K(p+q)$, where the first K entries in $\check{X}_t$ correspond to X_t. For this purpose, we define the $K(p+q)$-dimensional vectors

$$\check{X}_t := \begin{pmatrix} X_t \\ \vdots \\ X_{t-p+1} \\ e_t \\ \vdots \\ e_{t-q+1} \end{pmatrix} \quad \text{and} \quad \check{e}_t := \begin{pmatrix} e_t \\ 0 \\ \vdots \\ 0 \\ e_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

and, for $\cdot \in \{+, -\}$, the $(K(p+q) \times K(p+q))$-dimensional matrices

$$\check{A}_t^{(\cdot)} := \begin{pmatrix} \check{A}_{t,11}^{(\cdot)} & \check{A}_{t,12}^{(\cdot)} \\ \check{A}_{t,21}^{(\cdot)} & \check{A}_{t,22}^{(\cdot)} \end{pmatrix} \quad \text{and} \quad \check{B}_t := \begin{pmatrix} B_t^{(0)} & 0_{K \times K(p+q-1)} \\ 0_{K(p-1) \times K} & 0_{K(p-1) \times K(p+q-1)} \\ I_K & 0_{K \times K(p+q-1)} \\ 0_{K(q-1) \times K} & 0_{K(q-1) \times K(p+q-1)} \end{pmatrix}$$

where $\check{A}_{t,21}^{(+)} = \check{A}_{t,21}^{(-)} = 0_{Kq \times Kp}$ and $\check{A}_{t,22}^{(-)} = 0_{Kq \times Kq}$. Further, $\check{A}_{t,11}^{(+)} = \widetilde{A}_t^{(+)}$ and $\check{A}_{t,11}^{(-)} = \widetilde{A}_t^{(-)}$ with $\widetilde{A}_t^{(+)}$ and $\widetilde{A}_t^{(-)}$ from Section 3.2.2 and

$$\check{A}_{t,12}^{(+)} := \begin{pmatrix} B_t^{(+,1)} & \dots & B_t^{(+,q)} \\ 0_{K \times K} & \dots & 0_{K \times K} \\ \vdots & & \vdots \\ 0_{K \times K} & \dots & 0_{K \times K} \end{pmatrix}, \check{A}_{t,22}^{(+)} := \begin{pmatrix} 0_{K \times K} & \dots & 0_{K \times K} & 0_{K \times K} \\ I_K & & 0_{K \times K} & 0_{K \times K} \\ & \ddots & & \vdots \\ 0_{K \times K} & \dots & I_K & 0_{K \times K} \end{pmatrix},$$

$$\check{A}_{t,12}^{(-)} := \begin{pmatrix} B_t^{(-,1)} & \dots & B_t^{(-,q)} \\ 0_{K \times K} & \dots & 0_{K \times K} \\ \vdots & & \vdots \\ 0_{K \times K} & \dots & 0_{K \times K} \end{pmatrix}$$

are matrices of dimension $Kp \times Kq$, $Kq \times Kq$ and $Kp \times Kq$, respectively. Then, based on the notation introduced above, the K-dimensional gbVARMA(p,q) process $(X_t, t \in \mathbb{Z})$ can be represented as a $K(p+q)$-dimensional gbVAR(1) process $(\check{X}_t, t \in \mathbb{Z})$ as follows

$$\check{X}_t = \check{A}_t^{(+)} \check{X}_{t-1} + \check{A}_t^{(-)} \mathbb{1}_{K(p+q)} + \check{B}_t \check{e}_t. \tag{3.68}$$

where $\mathbb{1}_{K(p+q)}$ is the one vector of length $K(p+q)$. Similar to equation (3.24) for gbVAR(p) derived in Section 3.2.2, we get

$$\check{X}_t = \check{\zeta}_d \check{X}_{t-(d+1)} + \sum_{i=0}^{d} \check{\zeta}_{i-1} \check{\eta}_{t-i}, \tag{3.69}$$

with $\check{\zeta}_0 := I_K$ and $\check{\zeta}_i := \prod_{j=0}^{i} \check{A}_{t-j}^{(+)}$, $i \in \mathbb{N}$ and $\check{\eta}_{t-i} := \check{A}_{t-i}^{(-)} \mathbb{1}_{K(p+q)} + \check{B}_{t-i} \check{e}_{t-i}$. Hence, the $K(p+q)$-dimensional process $(\check{X}_t, t \in \mathbb{Z})$ has a moving-average-type representation if

$$det\left(I_{K(p+q)} - \check{\mathcal{A}}_{|\cdot|} z\right) \neq 0 \quad \forall z \in \mathbb{C} : \quad |z| \leq 1, \tag{3.70}$$

where $\check{\mathcal{A}} := E(\check{A}_t^{(+)})$. The above block determinant can be reduced such that (3.70) becomes equivalent to condition (3.27); see e.g. Lütkepohl (2005)[Eq. (11.3.9)].

Theorem 3.19 (Moving-average representation of gbVARMA processes)
Let $(X_t, t \in \mathbb{Z})$ be a (stationary) K-dimensional gbVARMA(p,q) process that fulfills (3.64) *with gbVAR(1) representation* (3.68)*. Then, if condition* (3.70) *holds, the gbVARMA(p,q) model has a gbVMA(∞)-type representation*

$$X_t = J\check{X}_t = J\left(\sum_{i=0}^{\infty} \check{\zeta}_{i-1} \check{\eta}_{t-i}\right), \quad t \in \mathbb{Z}, \tag{3.71}$$

converging in L_1, where $J := [I_K, 0_{K \times K(p+q-1)}]$.

Similar to classical VARMA models, Yule-Walker-type equations can be derived also for gbVARMA models. For the NDARMA(p,q) model class in the case of univariate categorical processes, these equations were proven by Weiß (2011a).

The following result generalizes Theorem 4 in Jentsch and Reichmann (2019) established for gbARMA models to the multivariate case of gbVARMA models.

Theorem 3.20 (Yule-Walker-type equations of gbVARMA processes)
Let $(X_t, t \in \mathbb{Z})$ be a stationary gbVARMA(p,q) process that fulfills (3.70)*. Set $\mathcal{B}^{(m)} := 0_{K \times K}$ for $m > q$ and define a sequence of coefficient matrices $(\Phi_m, m \in \mathbb{Z})$ recursively by*

$$\Phi_m := 0_{K \times K} \quad \text{for } m < 0, \quad \Phi_0 := \Sigma_e \mathcal{B}^{(0)\prime}, \quad \Phi_m := \sum_{i=1}^{p} \Phi_{m-i} \mathcal{A}^{(i)\prime} + \Sigma_e \mathcal{B}^{(m)\prime} \quad \text{for } m > 0.$$

Then, for all $h \in \mathbb{N}$, we have

$$\Gamma_X(h) - \sum_{i=1}^{p} \mathcal{A}^{(i)} \Gamma_X(h-i) = \sum_{j=h}^{q} \mathcal{B}^{(j)} \Phi_{j-h}. \tag{3.72}$$

As an extension of Theorem 3.9, the one step ahead transition probability can be derived also for gbVARMA processes.

Lemma 3.21 (Transition probabilities of gbVARMA processes)
Let $(X_t, t \in \mathbb{Z})$ be a stationary K-dimensional gbVARMA(p,q) process. Further, for $k = 0, 1, \ldots, p$ and $j = 0, 1, \ldots, q$, let $s_k, r_j \in \{0,1\}^K$ with $s_k := (s_{k,1}, \ldots, s_{k,K})'$ and $r_j := (r_{j,1}, \ldots, r_{j,K})'$, denote the Kronecker delta by $\delta_{ij} = \mathbb{1}_{\{i=j\}}$ and set $p_{r_j} :=$

$P(e_t = r_j)$. *Then, the transition probability given the past values of the time series is given by*

$$P\big(X_t = s_0 | X_{t-1} = s_1, \ldots, X_{t-p} = s_p\big) =$$
$$\sum_{r_0, r_1, \ldots, r_q \in \{0,1\}^K} \prod_{m=0}^{q} p_{r_m} \prod_{k=1}^{K} \Bigg[\sum_{i=1}^{p} \sum_{l=1}^{K} |\alpha_{kl}^{(i)}| \left[\mathbb{1}_{\{\alpha_{kl}^{(i)} \geq 0\}} \delta_{s_{0,k} s_{i,l}} + \mathbb{1}_{\{\alpha_{kl}^{(i)} < 0\}} \delta_{s_{0,k}\left(1 - s_{i,l}\right)} \right]$$
$$+ \beta_{kk}^{(0)} \delta_{s_{0,k} r_{0,k}} + \sum_{j=1}^{q} \sum_{l=1}^{K} |\beta_{kl}^{(j)}| \left[\mathbb{1}_{\{\beta_{kl}^{(j)} \geq 0\}} \delta_{s_{0,k} r_{j,l}} + \mathbb{1}_{\{\beta_{kl}^{(j)} < 0\}} \delta_{s_{0,k}\left(1 - r_{j,l}\right)} \right] \Bigg].$$

The following example of a gbVARMA(1,1) model illustrates the flexibility of the gbVARMA class.

Example 3.22 (Trivariate gbVARMA(1,1) model)
Let $(X_t, t \in \mathbb{Z})$ follow a three-dimensional gbVARMA(1,1) model with parameter matrix $\mathcal{P} = [\mathcal{A}^{(1)}, \mathcal{B}^{(0)}, \mathcal{B}^{(1)}]$ and $\mathcal{P}_{|\cdot|} = [\mathcal{A}_{|\cdot|}^{(1)}, \mathcal{B}^{(0)}, \mathcal{B}_{|\cdot|}^{(1)}]$ such that $\mathcal{P}_{|\cdot|} \mathbb{1}_9 = \mathbb{1}_3$ holds. Hence, for the (mutually independent) multinomial selection mechanisms, we have

$$P_{t,1\bullet} = \left[a_{t,11}^{(1)}, a_{t,12}^{(1)}, a_{t,13}^{(1)},\ b_{t,11}^{(0)}, 0, 0,\ b_{t,11}^{(1)}, b_{t,12}^{(1)}, b_{t,13}^{(1)} \right] \sim Mult\left(1; \mathcal{P}_{|\cdot|,1\bullet}\right),$$
$$P_{t,2\bullet} = \left[a_{t,21}^{(1)}, a_{t,22}^{(1)}, a_{t,23}^{(1)},\ 0, b_{t,22}^{(0)}, 0,\ b_{t,21}^{(1)}, b_{t,22}^{(1)}, b_{t,23}^{(1)} \right] \sim Mult\left(1; \mathcal{P}_{|\cdot|,2\bullet}\right),$$
$$P_{t,3\bullet} = \left[a_{t,31}^{(1)}, a_{t,32}^{(1)}, a_{t,33}^{(1)},\ 0, 0, b_{t,33}^{(0)},\ b_{t,31}^{(1)}, b_{t,32}^{(1)}, b_{t,33}^{(1)} \right] \sim Mult\left(1; \mathcal{P}_{|\cdot|,3\bullet}\right),$$

and, in the general case, the gbVARMA(1,1) process follows the equation

$$X_t = A_t^{(+,1)} X_{t-1} + A_t^{(-,1)} \mathbb{1}_3 + B_t^{(0)} e_t + B_t^{(+,1)} e_{t-1} + B_t^{(-,1)} \mathbb{1}_3, \quad t \in \mathbb{Z}.$$

Note that, for identification reasons, we do not allow for negative entries in the diagonal $\mathcal{B}^{(0)}$, but in $\mathcal{B}^{(1)}$ leading to the additional term $B_t^{(-,1)} \mathbb{1}_3$. From (3.72), the Yule-Walker-type equation for the gbVARMA(1,1) model with $h = 1$ becomes

$$\Gamma_X(1) - \mathcal{A}^{(1)} \Gamma_X(0) = \mathcal{B}^{(1)} \Sigma_e \mathcal{B}^{(0)\prime},$$

which, similar to the classical VARMA(1,1) case, constitutes a familiar relationship between the autocovariance matrices $\Gamma_X(0)$ and $\Gamma_X(1)$, the model parameters $\mathcal{P} = [\mathcal{A}^{(1)}, \mathcal{B}^{(0)}, \mathcal{B}^{(1)}]$ and the innovations' variance-covariance matrix Σ_e.

3.7. Proofs of Chapter 3

3.7.1. Proof of Equations (3.8) and (3.9)

We can re-arrange (3.7) to get

$$\begin{aligned} X_t &= a_{t,=} X_{t-1} + a_{t,\neq} (1 - X_{t-1}) + b_{t,=} \epsilon_t \\ &= \begin{cases} (a_{t,=} - a_{t,\neq}) X_{t-1} + a_{t,\neq} + b_{t,=} \epsilon_t, & \alpha_= - \alpha_\neq \geq 0 \\ (a_{t,\neq} - a_{t,=})(1 - X_{t-1}) + a_{t,=} + b_{t,=} \epsilon_t, & \alpha_= - \alpha_\neq < 0 \end{cases} \end{aligned}$$

such that a direct comparison with (3.6) leads to

$$\alpha = \begin{cases} \alpha_= - \alpha_{\neq}, & \alpha_= - \alpha_{\neq} \geq 0, \\ -(\alpha_{\neq} - \alpha_=), & \alpha_= - \alpha_{\neq} < 0 \end{cases} = \alpha_= - \alpha_{\neq}$$

and

$$\beta = \begin{cases} 1 - (\alpha_= - \alpha_{\neq}), & \alpha_= - \alpha_{\neq} \geq 0, \\ 1 - (\alpha_{\neq} - \alpha_=), & \alpha_= - \alpha_{\neq} < 0 \end{cases} = 1 - |\alpha_= - \alpha_{\neq}|$$

as well as $\alpha_{\neq} + \beta_= \mu_\epsilon = \beta \mu_e$ if $\alpha_= - \alpha_{\neq} \geq 0$ and $\alpha_= + \beta_= \mu_\epsilon = \beta \mu_e$ if $\alpha_= - \alpha_{\neq} < 0$ leading to

$$\mu_e = \frac{\alpha_{\neq} \mathbb{1}_{\{\alpha_= - \alpha_{\neq} \geq 0\}} + \alpha_= \mathbb{1}_{\{\alpha_= - \alpha_{\neq} < 0\}} + \beta_= \mu_\epsilon}{1 - |\alpha_= - \alpha_{\neq}|}.$$

□

3.7.2. Proof of Lemma 3.3 and 3.18

By taking expectations on both sides of (3.17) for Lemma 3.3 or (3.64) for Lemma 3.18, respectively, we get immediately the formulas for the mean by using law of iterated expectations and by exploiting stationarity. □.

3.7.3. Proof of Theorem 3.4 and 3.19

We prove only Theorem 3.4 and the calculations and arguments for Theorem 3.19 are very similar. As part (i) of 3.4 follows directly as a special case, we have to prove only part (ii). Suppose all eigenvalues of $\widetilde{\mathcal{A}}_{|\cdot|}$ have modulus smaller than one. In view of equation (3.24), we have to show that $\widetilde{\zeta}_d \widetilde{X}_{t-(d+1)}$ converges to zero in L_1 sense as $d \to \infty$. Let $\|A\|_1 := E(|A|_1)$, where $|M|_1 := \sum_{r,s=1}^{Kp} |M_{rs}|$ is the 1-norm of a $(Kp \times Kp)$ matrix $M = (M_{rs})_{r,s=1\ldots,Kp}$. Then, using the bound $\|\widetilde{X}_{t-(d+1)}\| \leq Kp$, it remains to show that

$$\left\| \widetilde{\zeta}_d \right\|_1 = \left\| \prod_{j=0}^{d} \widetilde{A}_{t-j}^{(+)} \right\|_1 \to 0 \quad \text{as} \quad d \to \infty. \tag{3.73}$$

By plugging-in, a direct calculation leads to

$$\begin{aligned} \left\| \widetilde{\zeta}_d \right\|_1 &= E\left(\sum_{k,l=1}^{Kp} \left| \left(\prod_{j=0}^{d} \widetilde{A}_{t-j}^{(+)} \right)_{kl} \right| \right) \\ &= E\left(\sum_{k,l=1}^{Kp} \left(| \sum_{r_1,\ldots,r_d=1}^{Kp} \widetilde{a}_{t,kr_1}^{(+)} \widetilde{a}_{t,r_1r_2}^{(+)} \cdots \widetilde{a}_{t,r_{d-1}r_d}^{(+)} \widetilde{a}_{t,r_dl}^{(+)} | \right)_{kl} \right) \\ &\leq E\left(\sum_{k,l=1}^{Kp} \left(\sum_{r_1,\ldots,r_d=1}^{Kp} |\widetilde{a}_{t,kr_1}^{(+)}| \cdot |\widetilde{a}_{t,r_1r_2}^{(+)}| \cdots |\widetilde{a}_{t,r_{d-1}r_d}^{(+)}| \cdot |\widetilde{a}_{t,r_dl}^{(+)}| \right)_{kl} \right) \\ &= E\left(\sum_{k,l=1}^{Kp} \left(\sum_{r_1,\ldots,r_d=1}^{Kp} \widetilde{a}_{t,kr_1} \cdot \widetilde{a}_{t,r_1r_2} \cdots \widetilde{a}_{t,r_{d-1}r_d} \cdot \widetilde{a}_{t,r_dl} \right)_{kl} \right) \end{aligned}$$

$$
\begin{aligned}
&= E\left(\sum_{k,l=1}^{Kp}\left(\prod_{j=0}^{d}\widetilde{A}_{t-j}\right)_{kl}\right)\\
&= \sum_{k,l=1}^{Kp}\left(\prod_{j=0}^{d}E\left(\widetilde{A}_{t-j}\right)\right)_{kl}\\
&= \mathbb{1}'_{Kp}\widetilde{\mathcal{A}}_{|\cdot|}^{d+1}\mathbb{1}_{Kp} \underset{d\to\infty}{\to} 0,
\end{aligned}
$$

where we have used that the process $(\widetilde{A}_t^{(+)}, t \in \mathbb{Z})$ is i.i.d. and that all eigenvalues of $\widetilde{\mathcal{A}}_{|\cdot|}$ are strictly smaller than one. □

3.7.4. Proof of Theorem 3.6 and 3.20

Let $h \geq 0$. Then, by plugging-in the model equation and using the law of iterated expectations, we get for stationary gbVARMA(p,q) processes the identity

$$
\begin{aligned}
Cov(X_t, X_{t-h}) &= \sum_{i=1}^{p} Cov(A_t^{(+,i)}X_{t-i}, X_{t-h}) + \sum_{i=1}^{p} Cov(A_t^{(-,i)}\mathbb{1}_K, X_{t-h})\\
&\quad + \sum_{j=0}^{q} Cov(B_t^{(+,j)}e_{t-j}, X_{t-h}) + \sum_{j=0}^{q} Cov(B_t^{(-,j)}\mathbb{1}_K, X_{t-h})\\
&= \sum_{i=1}^{p}\mathcal{A}^{(i)}\Gamma_X(h-i) + \sum_{j=0}^{q}\mathcal{B}^{(j)}Cov\left(e_{t-j}, X_{t-h}\right),
\end{aligned}
$$

where we set $B_t^{(+,0)} := B_t^{(0)}$ for notational convenience. Next, we define the recursion for the mixed covariance terms of the innovation and time series. Let $\Phi_m := Cov(e_{t-m}, X_t)$. Then, we get $\Phi_m = 0_{K\times K}$ for $m < 0$ because $(X_s, s < t)$ and $(e_t, t \in \mathbb{Z})$ are independent due to the causal gbVMA representation (3.71). For $m = 0$, we get

$$
\begin{aligned}
\Phi_0 = Cov(e_t, X_t) &= \sum_{i=1}^{p} Cov(e_t, X_{t-j})\mathcal{A}^{(i)\prime} + \sum_{j=0}^{q} Cov(e_t, e_{t-j})\mathcal{B}^{(j)\prime}\\
&= \Sigma_e\mathcal{B}^{(0)\prime}
\end{aligned}
$$

and, for $m > 0$, we get

$$
\begin{aligned}
\Phi_m = Cov(e_{t-m}, X_t) &= \sum_{i=1}^{p} Cov(e_{t-m}, X_{t-j})\mathcal{A}^{(i)\prime} + \sum_{j=0}^{q} Cov(e_{t-m}, e_{t-j})\mathcal{B}^{(j)\prime}\\
&= \sum_{i=1}^{p}\Phi_{m-i}\mathcal{A}^{(i)\prime} + \Sigma_e\mathcal{B}^{(m)\prime}.
\end{aligned}
$$

Note $\mathcal{B}^{(m)} := 0_{K\times K}$ for $m > q$. Then the Yule-Walker type equation for gbVARMA(p,q) processes is given by

$$
\Gamma_X(h) - \sum_{i=1}^{p}\mathcal{A}^{(i)}\Gamma_X(h-i) = \sum_{j=h}^{q}\mathcal{B}^{(j)}\Phi_{j-h} \tag{3.74}
$$

due to $\Phi_m = 0_{K\times K}$ for $m < 0$. The Yule Walker equation for gbVAR(p) processes derived in Theorem 3.6 follows by setting $q = 0$. □

Lemma 3.23
For the $K \times K(p+1)$-dimensional i.i.d. process $(P_t, t \in \mathbb{Z})$ with $P_t = [A_t^{(1)}, \ldots, A_t^{(p)}, B_t]$ with $A_t^{(i)}$, $A_t^{(+,i)}$, $A_t^{(-,i)}$ and B_t as defined in Definition 3.1, we have, for all $i, j = 1, \ldots, p$ and $v, r, w, s = 1, \ldots, K$, the following:

(i) $$E(a_{t,vr}^{(i)} a_{t,ws}^{(j)}) = \begin{cases} |\alpha_{vr}^{(i)}|, & v = w, r = s, i = j \\ |\alpha_{vr}^{(i)}||\alpha_{ws}^{(j)}|, & v \neq w \\ 0, & \text{otherwise} \end{cases}$$

(ii) $$E(a_{t,vr}^{(i)} b_{t,ws}) = \begin{cases} |\alpha_{vr}^{(i)}|\beta_{ws}, & v \neq w \\ 0, & \text{otherwise} \end{cases}$$

(iii) $$E(a_{t,vr}^{(+,i)} a_{t,ws}^{(+,j)}) = \begin{cases} |\alpha_{vr}^{(i)}| & v = w, r = s, i = j \\ \alpha_{vr}^{(i)}\alpha_{ws}^{(j)}, & v \neq w \\ 0, & \text{otherwise} \end{cases}$$

(iv) $$E(a_{t,vr}^{(+,i)} a_{t,ws}^{(-,j)}) = \begin{cases} -\alpha_{vr}^{(-,i)}, & v = w, r = s, i = j \\ \alpha_{vr}^{(i)}\alpha_{ws}^{(-,j)}, & v \neq w \\ 0, & \text{otherwise} \end{cases}$$

(v) $$E(a_{t,vr}^{(-,i)} a_{t,ws}^{(-,j)}) = \begin{cases} \alpha_{vr}^{(-,i)}, & v = w, r = s, i = j \\ \alpha_{vr}^{(-,i)}\alpha_{ws}^{(-,j)}, & v \neq w \\ 0, & \text{otherwise} \end{cases}$$

(vi) $$E(a_{t,vr}^{(+,i)} b_{t,ws}) = \begin{cases} \alpha_{vr}^{(i)}\beta_{ws}, & v \neq w \\ 0, & \text{otherwise} \end{cases}$$

(vii) $$E(a_{t,vr}^{(-,i)} b_{t,ws}) = \begin{cases} \alpha_{vr}^{(-,i)}\beta_{ws}, & v \neq w \\ 0, & \text{otherwise} \end{cases}$$

(viii) $$E(b_{t,vr} b_{t,ws}) = \begin{cases} \beta_{vr}, & v = w, r = s \\ \beta_{vr}\beta_{ws}, & v \neq w \\ 0, & \text{otherwise} \end{cases}$$

Proof.
By direct calculation. □

3.7.5. Proof of Theorem 3.7

By plugging-in (3.17), we get

$$\begin{aligned}
\Gamma_X(0) &= Cov(X_t, X_t) \\
&= \sum_{i,j=1}^{p} Cov(A_t^{(+,i)} X_{t-i}, A_t^{(+,j)} X_{t-j}) + \sum_{i,j=1}^{p} Cov(A_t^{(+,i)} X_{t-i}, A_t^{(-,j)} \mathbb{1}_K) \\
&\quad + \sum_{i=1}^{p} Cov(A_t^{(+,i)} X_{t-i}, B_t e_t) + \sum_{i,j=1}^{p} Cov(A_t^{(-,i)} \mathbb{1}_K, A_t^{(+,j)} X_{t-j}) \\
&\quad + \sum_{i,j=1}^{p} Cov(A_t^{(-,i)} \mathbb{1}_K, A_t^{(-,j)} \mathbb{1}_K) + \sum_{i=1}^{p} Cov(A_t^{(-,i)} \mathbb{1}_K, B_t e_t) \\
&\quad + \sum_{j=1}^{p} Cov(B_t e_t, A_t^{(+,j)} X_{t-j}) + \sum_{j=1}^{p} Cov(B_t e_t, A_t^{(-,i)} \mathbb{1}_K) + Cov(B_t e_t, B_t e_t) \\
&=: I_1 + I_2 + \cdots + I_9
\end{aligned}$$

with an obvious notation for $I_1, \ldots, I_9$. Starting with I_1, we can write

$$Cov(A_t^{(+,i)} X_{t-i}, A_t^{(+,j)} X_{t-j}) = E(A_t^{(+,i)} X_{t-i} X_{t-j}' A_t^{(+,j)\prime}) - \mathcal{A}^{(i)} \mu_X \mu_X' \mathcal{A}^{(j)\prime}$$

and, by law of iterated expectations, we get

$$E(A_t^{(+,i)} X_{t-i} X_{t-j}' A_t^{(+,j)\prime}) = E(E(A_t^{(+,i)} X_{t-i} X_{t-j}' A_t^{(+,j)\prime} | X_{t-i}, X_{t-j})) \tag{3.75}$$

Now, we have to distinguish the cases $i = j$ and $i \neq j$. For $i = j$, we can write

$$E(E(A_t^{(+,i)} X_{t-i} X_{t-i}' A_t^{(+,i)\prime} | X_{t-i})) = \sum_{m \in \{0,1\}^K} E(A_t^{(+,i)} m m' A_t^{(+,i)\prime}) P(X_{t-i} = m) \tag{3.76}$$

By using Lemma 3.23(iii), the expectation on the last right-hand side computes to

$$\left(\sum_{r,s=1}^{K} m_r m_s E(a_{t,vr}^{(+,i)} a_{t,ws}^{(+,i)}) \right)_{v,w} = \left(\sum_{r=1}^{K} m_r^2 |\alpha_{vr}^{(i)}| \mathbb{1}_{\{v=w\}} + \sum_{r,s=1}^{K} m_r m_s \alpha_{vr}^{(i)} \alpha_{ws}^{(i)} \mathbb{1}_{\{v \neq w\}} \right)_{v,w}$$

Plugging this into (3.76) and by re-arranging the sums, this leads to

$$\begin{aligned}
&E(E(A_t^{(+,i)} X_{t-i} X_{t-i}' A_t^{(+,i)\prime} | X_{t-i})) \\
&= \left(\sum_{r=1}^{K} \left(\sum_{m \in \{0,1\}^K} m_r^2 P(X_{t-i} = m) \right) |\alpha_{vr}^{(i)}| \mathbb{1}_{\{v=w\}} \right)_{v,w} \\
&\quad + \left(\sum_{r,s=1}^{K} \alpha_{vr}^{(i)} \left(\sum_{m \in \{0,1\}^K} m_r m_s P(X_{t-i} = m) \right) \alpha_{ws}^{(i)} \mathbb{1}_{\{v \neq w\}} \right)_{v,w} \\
&= \left(\sum_{r=1}^{K} \mu_{X,r} |\alpha_{vr}^{(i)}| \mathbb{1}_{\{v=w\}} \right)_{v,w} + \left(\sum_{r,s=1}^{K} \alpha_{vr}^{(i)} \left(\Gamma_{X,rs}(0) + \mu_{X,r} \mu_{X,s} \right) \alpha_{ws}^{(i)} \mathbb{1}_{\{v \neq w\}} \right)_{v,w} \\
&= I_K \circ \left\{ \mathcal{A}_{|\cdot|}^{(i)} \mu_X \mathbb{1}_K' \right\} + (\mathbb{1}_{K \times K} - I_K) \circ \left\{ \mathcal{A}^{(i)} \left(\Gamma_X(0) + \mu_X \mu_X' \right) \mathcal{A}^{(i)} \right\}
\end{aligned}$$

due to $m_r^2 = m_r$. Similarly, for $i \neq j$, this leads to

$$\begin{aligned} &E(E(A_t^{(+,i)} X_{t-i} X_{t-j}' A_t^{(+,j)\prime} | X_{t-i}, X_{t-j})) \\ &= \sum_{m,n \in \{0,1\}^K} E(A_t^{(+,i)} m n' A_t^{(+,j)\prime}) P(X_{t-i} = m, X_{t-j} = n). \end{aligned} \tag{3.77}$$

By using again Lemma 3.23(iii), the expectation on the last right-hand side becomes

$$\left(\sum_{r,s=1}^{K} m_r n_s E(a_{t,vr}^{(+,i)} a_{t,ws}^{(+,j)}) \right)_{v,w} = \left(\sum_{r,s=1}^{K} m_r n_s \alpha_{vr}^{(i)} \alpha_{ws}^{(j)} \mathbb{1}_{\{v \neq w\}} \right)_{v,w}.$$

Plugging this into (3.77) and re-arranging the sums, this leads to

$$\begin{aligned} &E(E(A_t^{(+,i)} X_{t-i} X_{t-j}' A_t^{(+,j)\prime} | X_{t-i}, X_{t-j})) \\ &= \sum_{m,n \in \{0,1\}^K} \left(\sum_{r,s=1}^{K} m_r n_s \alpha_{vr}^{(i)} \alpha_{ws}^{(j)} \mathbb{1}_{\{v \neq w\}} \right)_{v,w} P(X_{t-i} = m, X_{t-j} = n) \\ &= \left(\sum_{r,s=1}^{K} \alpha_{vr}^{(i)} \left(\sum_{m,n \in \{0,1\}^K} m_r n_s P(X_{t-i} = m, X_{t-j} = n) \right) \alpha_{ws}^{(j)} \mathbb{1}_{\{v \neq w\}} \right)_{v,w} \\ &= \left(\sum_{r,s=1}^{K} \alpha_{vr}^{(i)} \left(\Gamma_{X,rs}(j-i) + \mu_{X,r} \mu_{X,s} \right) \alpha_{ws}^{(j)} \mathbb{1}_{\{v \neq w\}} \right)_{v,w} \\ &= (\mathbb{1}_{K \times K} - I_K) \circ \left\{ \mathcal{A}^{(i)} \left(\Gamma_X(j-i) + \mu_X \mu_X' \right) \mathcal{A}^{(j)\prime} \right\}. \end{aligned}$$

Considering both cases $i = j$ and $i \neq j$ together, gives

$$\begin{aligned} &E(A_t^{(+,i)} X_{t-i} X_{t-j}' A_t^{(+,j)\prime}) \\ &= I_K \circ \left\{ \mathcal{A}_{|\cdot|}^{(i)} \mu_X \mathbb{1}_K' \right\} \mathbb{1}_{\{i=j\}} + (\mathbb{1}_{K \times K} - I_K) \circ \left\{ \mathcal{A}^{(i)} \left(\Gamma_X(j-i) + \mu_X \mu_X' \right) \mathcal{A}^{(j)\prime} \right\}. \end{aligned}$$

leading to

$$\begin{aligned} I_1 = \sum_{i,j=1}^{p} \Big[&I_K \circ \left\{ \mathcal{A}_{|\cdot|}^{(i)} \mu_X \mathbb{1}_K' \right\} \mathbb{1}_{\{i=j\}} \\ &+ (\mathbb{1}_{K \times K} - I_K) \circ \left\{ \mathcal{A}^{(i)} \left(\Gamma_X(j-i) + \mu_X \mu_X' \right) \mathcal{A}^{(j)\prime} \right\} - \mathcal{A}^{(i)} \mu_X \mu_X' \mathcal{A}^{(i)\prime} \Big] \\ = \sum_{i,j=1}^{p} \Big[&I_K \circ \left\{ \left(\mathcal{A}_{|\cdot|}^{(i)} \mu_X \mathbb{1}_K' \right) \mathbb{1}_{\{i=j\}} - \left(\mathcal{A}^{(i)} \left(\Gamma_X(j-i) + \mu_X \mu_X' \right) \mathcal{A}^{(j)\prime} \right) \right\} \\ &+ \left(\mathcal{A}^{(i)} \left(\Gamma_X(j-i) \right) \mathcal{A}^{(j)\prime} \right) \Big] \end{aligned}$$

Similarly, continuing with I_2, we get

$$Cov(A_t^{(+,i)} X_{t-i}, A_t^{(-,j)} \mathbb{1}_K) = E(A_t^{(+,i)} X_{t-i} \mathbb{1}_K' A_t^{(-,j)\prime}) - \mathcal{A}^{(i)} \mu_X \mathbb{1}_K' \mathcal{A}^{(-,j)\prime}$$

and, by law of iterated expectations, we get

$$E(A_t^{(+,i)} X_{t-i} \mathbb{1}_K' A_t^{(-,j)\prime}) = E(E(A_t^{(+,i)} X_{t-i} \mathbb{1}_K' A_t^{(-,j)\prime} | X_{t-i})) \tag{3.78}$$
$$= \sum_{m \in \{0,1\}^K} E(A_t^{(+,i)} m \mathbb{1}_K' A_t^{(-,j)\prime}) P(X_{t-i} = m).$$

By using Lemma 3.23(iv), the expectation on the last right-hand side computes to

$$\left(\sum_{r,s=1}^{K} m_r 1 E(a_{t,vr}^{(+,i)} a_{t,ws}^{(-,i)}) \right)_{v,w} =$$
$$\left(\sum_{r=1}^{K} m_r 1 \left(-\alpha_{vr}^{(-,i)} \right) \mathbb{1}_{\{v=w, i=j\}} + \sum_{r,s=1}^{K} m_r 1 \alpha_{vr}^{(i)} \alpha_{ws}^{(-,j)} \mathbb{1}_{\{v \neq w\}} \right)_{v,w}.$$

Plugging this into (3.78) and re-arranging the sums, this leads to

$$\begin{aligned}
&E(A_t^{(+,i)} X_{t-i} \mathbb{1}_K' A_t^{(-,j)\prime}) \\
&= \left(\sum_{r=1}^{K} \left(\sum_{m \in \{0,1\}^K} m_r P(X_{t-i} = m) \right) \left(-\alpha_{vr}^{(-,i)} \right) \mathbb{1}_{\{v=w, i=j\}} \right)_{v,w} \\
&\quad + \left(\sum_{r,s=1}^{K} \alpha_{vr}^{(i)} \left(\sum_{m \in \{0,1\}^K} m_r P(X_{t-i} = m) \right) \alpha_{ws}^{(-,j)} \mathbb{1}_{\{v \neq w\}} \right)_{v,w} \\
&= \left(\sum_{r=1}^{K} \mu_{X,r} \left(-\alpha_{vr}^{(-,i)} \right) \mathbb{1}_{\{v=w, i=j\}} \right)_{v,w} + \left(\sum_{r,s=1}^{K} \alpha_{vr}^{(i)} \mu_X \mathbb{1}_K' \alpha_{ws}^{(-,j)} \mathbb{1}_{\{v \neq w\}} \right)_{v,w} \\
&= I_K \circ \left\{ -\mathcal{A}^{(-,i)} \mu_X \mathbb{1}_K' \right\} \mathbb{1}_{\{i=j\}} + (\mathbb{1}_{K \times K} - I_K) \circ \left\{ \mathcal{A}^{(i)} \mu_X \mathbb{1}_K' \mathcal{A}^{(-,j)\prime} \right\}.
\end{aligned}$$

Altogether, we get

$$\begin{aligned}
I_2 &= \sum_{i,j=1}^{p} \Big[I_K \circ \left\{ -\mathcal{A}^{(-,i)} \mu_X \mathbb{1}_K' \right\} \mathbb{1}_{\{i=j\}} \\
&\quad + (\mathbb{1}_{K \times K} - I_K) \circ \left\{ \mathcal{A}^{(i)} \mu_X \mathbb{1}_K' \mathcal{A}^{(-,j)\prime} \right\} - \mathcal{A}^{(i)} \mu_X \mathbb{1}_K' \mathcal{A}^{(-,j)\prime} \Big] \\
&= \sum_{i=1}^{p} \left[I_K \circ \left\{ -\mathcal{A}^{(-,i)} \mu_X \mathbb{1}_K' \right\} + (\mathbb{1}_{K \times K} - I_K) \circ \left\{ \mathcal{A}^{(i)} \mu_X \mathbb{1}_K' \mathcal{A}^{(-,i)\prime} \right\} - \mathcal{A}^{(i)} \mu_X \mathbb{1}_K' \mathcal{A}^{(-,i)\prime} \right] \\
&\quad + \sum_{\substack{i,j=1 \\ i \neq j}}^{p} \left[(\mathbb{1}_{K \times K} - I_K) \circ \left\{ \mathcal{A}^{(i)} \mu_X \mathbb{1}_K' \mathcal{A}^{(-,j)\prime} \right\} - \mathcal{A}^{(i)} \mu_X \mathbb{1}_K' \mathcal{A}^{(-,j)\prime} \right] \\
&= \sum_{i=1}^{p} \left[I_K \circ \left\{ -\mathcal{A}^{(-,i)} \mu_X \mathbb{1}_K' - \left(\mathcal{A}^{(i)} \mu_X \mathbb{1}_K' \mathcal{A}^{(-,i)\prime} \right) \right\} \right] - \sum_{\substack{i,j=1 \\ i \neq j}}^{p} \left[I_K \circ \left\{ \mathcal{A}^{(i)} \mu_X \mathbb{1}_K' \mathcal{A}^{(-,j)\prime} \right\} \right] \\
&= \sum_{i,j=1}^{p} \left[I_K \circ \left\{ \left(-\mathcal{A}^{(-,i)} \mu_X \mathbb{1}_K' \right) \mathbb{1}_{\{i=j\}} - \left(\mathcal{A}^{(i)} \mu_X \mathbb{1}_K' \mathcal{A}^{(-,j)\prime} \right) \right\} \right]
\end{aligned}$$

and, due to symmetry (note that the first term is diagonal)

$$I_4 = \sum_{i,j=1}^{p} \left[I_K \circ \left\{ (-\mathcal{A}^{(-,i)} \mu_X \mathbb{1}'_K) \mathbb{1}_{\{i=j\}} - \left(\mathcal{A}^{(-,i)} \mathbb{1}_K \mu'_X \mathcal{A}^{(j)\prime} \right) \right\} \right].$$

For the summands of I_3, we have

$$Cov(A_t^{(+,i)} X_{t-i}, B_t e_t) = E(A_t^{(+,i)} X_{t-i} e'_t B'_t) - \mathcal{A}^{(i)} \mu_X \mu'_e \mathcal{B}'$$

and, by using Lemma 3.23(vi),

$$\begin{aligned}
E(A_t^{(+,i)} X_{t-i} e'_t B'_t) &= E(E(A_t^{(+,i)} X_{t-i} e'_t B'_t | A_t^{(+,i)}, B_t)) \\
&= E(A_t^{(+,i)} E(X_{t-i} e'_t | A_t^{(+,i)}, B_t) B'_t) \\
&= E(A_t^{(+,i)} E(X_{t-i} e'_t) B'_t) \\
&= E(A_t^{(+,i)} \mu_X \mu'_e B'_t) \\
&= \left(\sum_{r,s=1}^{K} \mu_{X,r} \mu_{e,s} E(a_{t,vr}^{(+,i)} b_{t,ws}) \right)_{v,w} \\
&= \left(\sum_{r,s=1}^{K} \mu_{X,r} \mu_{e,s} \alpha_{vr}^{(i)} \beta_{ws} \mathbb{1}_{\{v \neq w\}} \right)_{v,w} \\
&= (\mathbb{1}_{K \times K} - I_K) \circ \left\{ \mathcal{A}^{(i)} \mu_X \mu'_e \mathcal{B}' \right\}
\end{aligned}$$

leading to

$$\begin{aligned}
I_3 &= \sum_{i=1}^{p} \left[(\mathbb{1}_{K \times K} - I_K) \circ \left\{ \mathcal{A}^{(i)} \mu_X \mu'_e \mathcal{B}' \right\} - \mathcal{A}^{(i)} \mu_X \mu'_e \mathcal{B}' \right] \\
&= -\sum_{i=1}^{p} \left[I_K \circ \left\{ \mathcal{A}^{(i)} \mu_X \mu'_e \mathcal{B}' \right\} \right]
\end{aligned}$$

and, due to symmetry,

$$I_7 = -\sum_{j=1}^{p} \left[I_K \circ \left\{ \mathcal{B} \mu_e \mu'_X \mathcal{A}^{(j)\prime} \right\} \right].$$

Continuing with I_5, we have

$$Cov(A_t^{(-,i)} \mathbb{1}_K, A_t^{(-,j)} \mathbb{1}_K) = E(A_t^{(-,i)} \mathbb{1}_K \mathbb{1}'_K A_t^{(-,j)\prime}) - \mathcal{A}^{(-,i)} \mathbb{1}_K \mathbb{1}'_K \mathcal{A}^{(-,j)\prime}$$

and (note the sign in comparison to term I_2!)

$$\begin{aligned}
&E(A_t^{(-,i)} \mathbb{1}_K \mathbb{1}'_K A_t^{(-,j)\prime}) \\
&= \left(\sum_{r,s=1}^{K} E(a_{t,vr}^{(-,i)} a_{t,ws}^{(-,j)}) \right)_{v,w} \\
&= \left(\sum_{r=1}^{K} \alpha_{vr}^{(-,i)} \mathbb{1}_{\{v=w, i=j\}} + \sum_{r,s=1}^{K} \alpha_{vr}^{(-,i)} \alpha_{ws}^{(-,j)} \mathbb{1}_{\{v \neq w\}} \right)_{v,w} \\
&= I_K \circ \left\{ \mathcal{A}^{(-,i)} \mathbb{1}_K \mathbb{1}'_K \right\} \mathbb{1}_{\{i=j\}} + (\mathbb{1}_{K \times K} - I_K) \circ \left\{ \mathcal{A}^{(-,i)} \mathbb{1}_K \mathbb{1}'_K \mathcal{A}^{(-,j)\prime} \right\}
\end{aligned}$$

resulting in

$$\begin{aligned}
I_5 &= \sum_{i,j=1}^{p} \Big[I_K \circ \left\{ \mathcal{A}^{(-,i)} \mathbb{1}_K \mathbb{1}'_K \right\} \mathbb{1}_{\{i=j\}} + (\mathbb{1}_{K\times K} - I_K) \circ \left\{ \mathcal{A}^{(-,i)} \mathbb{1}_K \mathbb{1}'_K \mathcal{A}^{(-,j)\prime} \right\} \\
&\quad - \mathcal{A}^{(-,i)} \mathbb{1}_K \mathbb{1}'_K \mathcal{A}^{(-,j)\prime} \Big] \\
&= \sum_{i,j=1}^{p} \left[I_K \circ \left\{ \mathcal{A}^{(-,i)} \mathbb{1}_K \mathbb{1}'_K \right\} \mathbb{1}_{\{i=j\}} + (-I_K) \circ \left\{ \mathcal{A}^{(-,i)} \mathbb{1}_K \mathbb{1}'_K \mathcal{A}^{(-,j)\prime} \right\} \right] \\
&= \sum_{i,j=1}^{p} \left[I_K \circ \left\{ \left(\mathcal{A}^{(-,i)} \mathbb{1}_K \mathbb{1}'_K \right) \mathbb{1}_{\{i=j\}} - \mathcal{A}^{(-,i)} \mathbb{1}_K \mathbb{1}'_K \mathcal{A}^{(-,j)\prime} \right\} \right].
\end{aligned}$$

For the summands of I_6, we have

$$Cov(A_t^{(-,i)} \mathbb{1}_K, B_t e_t) = E(A_t^{(-,i)} \mathbb{1}_K e'_t B'_t) - \mathcal{A}^{(-,i)} \mathbb{1}_K \mu'_e \mathcal{B}'$$

and, by using Lemma 3.23(vii),

$$\begin{aligned}
E(A_t^{(-,i)} \mathbb{1}_K e'_t B'_t) &= E(E(A_t^{(-,i)} \mathbb{1}_K e'_t B'_t | A_t^{(-,i)}, B_t)) \\
&= E(A_t^{(-,i)} \mathbb{1}_K E(e'_t | A_t^{(-,i)}, B_t) B'_t) \\
&= E(A_t^{(-,i)} \mathbb{1}_K E(e'_t) B'_t) \\
&= E(A_t^{(-,i)} \mathbb{1}_K \mu'_e B'_t) \\
&= \left(\sum_{r,s=1}^{K} \mu_{e,s} E(a_{t,vr}^{(-,i)} b_{t,ws}) \right)_{v,w} \\
&= \left(\sum_{r,s=1}^{K} \mu_{e,s} \alpha_{vr}^{(-,i)} \beta_{ws} \mathbb{1}_{\{v\neq w\}} \right)_{v,w} \\
&= (\mathbb{1}_{K\times K} - I_K) \circ \left\{ \mathcal{A}^{(-,i)} \mathbb{1}_K \mu'_e \mathcal{B}' \right\}
\end{aligned}$$

leading to

$$\begin{aligned}
I_6 &= \sum_{i=1}^{p} \left[(\mathbb{1}_{K\times K} - I_K) \circ \left\{ \mathcal{A}^{(-,i)} \mathbb{1}_K \mu'_e \mathcal{B}' \right\} - \mathcal{A}^{(-,i)} \mathbb{1}_K \mu'_e \mathcal{B}' \right] \\
&= -\sum_{i=1}^{p} \left[I_K \circ \left\{ \mathcal{A}^{(-,i)} \mathbb{1}_K \mu'_e \mathcal{B}' \right\} \right]
\end{aligned}$$

and, due to symmetry,

$$I_8 = -\sum_{j=1}^{p} \left[I_K \circ \left\{ \mathcal{B} \mu_e \mathbb{1}'_K \mathcal{A}^{(-,j)\prime} \right\} \right].$$

Finally, I_9 equals $Cov(B_t e_t, B_t e_t) = E(B_t e_t e_t' B_t') - \mathcal{B}\mu_e\mu_e'\mathcal{B}'$ and, by using Lemma 3.23(viii) and as $(\Sigma_e + \mu_e\mu_e')_{rr} = \mu_{e,r}(1-\mu_{e,r}) + \mu_{e,r}\mu_{e,r} = \mu_{e,r}$, we get

$$
\begin{aligned}
E(B_t e_t e_t' B_t') &= E(E(B_t e_t e_t' B_t'|B_t)) \\
&= E(B_t E(e_t e_t'|B_t) B_t') \\
&= E(B_t E(e_t e_t') B_t') \\
&= E(B_t \left(\Sigma_e + \mu_e\mu_e'\right) B_t') \\
&= \left(\sum_{r,s=1}^{K} \left(\Sigma_e + \mu_e\mu_e'\right)_{rs} E(b_{t,vr} b_{t,ws}) \right)_{v,w} \\
&= \left(\sum_{r=1}^{K} \beta_{vr} \left(\Sigma_e + \mu_e\mu_e'\right)_{rr} \mathbb{1}_{\{v=w\}} + \sum_{r,s=1}^{K} \beta_{vr} \left(\Sigma_e + \mu_e\mu_e'\right)_{rs} \beta_{ws} \mathbb{1}_{\{v\neq w\}} \right)_{v,w} \\
&= \left(\sum_{r=1}^{K} \beta_{vr}\mu_{e,r} \mathbb{1}_{\{v=w\}} + \sum_{r,s=1}^{K} \beta_{vr} \left(\Sigma_e + \mu_e\mu_e'\right)_{rs} \beta_{ws} \mathbb{1}_{\{v\neq w\}} \right)_{v,w} \\
&= I_K \circ \left\{ \mathcal{B}\mu_e \mathbb{1}_K' \right\} + (\mathbb{1}_{K\times K} - I_K) \circ \left\{ \mathcal{B} \left(\Sigma_e + \mu_e\mu_e'\right) \mathcal{B} \right\}.
\end{aligned}
$$

Putting everything together, we get

$$
\begin{aligned}
\Gamma_X(0) &= Cov(X_t, X_t) \\
&= \sum_{i,j=1}^{p} \Big[I_K \circ \Big\{ \left(\mathcal{A}^{(i)}_{|\cdot|} \mu_X \mathbb{1}_K' \right) \mathbb{1}_{\{i=j\}} - \left(\mathcal{A}^{(i)} \left(\Gamma_X(j-i) + \mu_X\mu_X' \right) \mathcal{A}^{(j)\prime} \right) \Big\}. \\
&\quad + \left(\mathcal{A}^{(i)} \left(\Gamma_X(j-i) \right) \mathcal{A}^{(j)\prime} \right) \Big] \\
&\quad + \sum_{i,j=1}^{p} \left[I_K \circ \left\{ \left(-\mathcal{A}^{(-,i)} \mu_X \mathbb{1}_K' \right) \mathbb{1}_{\{i=j\}} - \left(\mathcal{A}^{(i)} \mu_X \mathbb{1}_K' \mathcal{A}^{(-,j)\prime} \right) \right\} \right] \\
&\quad - \sum_{i=1}^{p} \left[I_K \circ \left\{ \mathcal{A}^{(i)} \mu_X \mu_e' \mathcal{B}' \right\} \right] \\
&\quad + \sum_{i,j=1}^{p} \left[I_K \circ \left\{ -\mathcal{A}^{(-,i)} \mu_X \mathbb{1}_K' \mathbb{1}_{\{i=j\}} - \left(\mathcal{A}^{(-,i)} \mathbb{1}_K \mu_X' \mathcal{A}^{(j)\prime} \right) \right\} \right] \\
&\quad + \sum_{i,j=1}^{p} \left[I_K \circ \left\{ \left(\mathcal{A}^{(-,i)} \mathbb{1}_K \mathbb{1}_K' \right) \mathbb{1}_{\{i=j\}} - \mathcal{A}^{(-,i)} \mathbb{1}_K \mathbb{1}_K' \mathcal{A}^{(-,j)\prime} \right\} \right] \\
&\quad - \sum_{i=1}^{p} \left[I_K \circ \left\{ \mathcal{A}^{(-,i)} \mathbb{1}_K \mu_e' \mathcal{B}' \right\} \right] \\
&\quad - \sum_{j=1}^{p} \left[I_K \circ \left\{ \mathcal{B}\mu_e \mu_X' \mathcal{A}^{(j)\prime} \right\} \right] \\
&\quad - \sum_{j=1}^{p} \left[I_K \circ \left\{ \mathcal{B}\mu_e \mathbb{1}_K' \mathcal{A}^{(-,j)\prime} \right\} \right] \\
&\quad + I_K \circ \left\{ \mathcal{B}\mu_e \mathbb{1}_K' - \mathcal{B}(\mu_e \mathbb{1}_K')\mathcal{B} \right\} + \mathcal{B}\Sigma_e\mathcal{B}.
\end{aligned}
$$

□

3.7.6. Proof of Lemma 3.9 and 3.21

To compute the transition probability $P\big(X_t = s_0 | X_{t-1} = s_1, \ldots, X_{t-p} = s_p\big)$, by conditioning also on the innovations $e_t, e_{t-1}, \ldots, e_{t-q}$, we get

$$\begin{aligned}
&P\big(X_t = s_0 | X_{t-1} = s_1, \ldots, X_{t-p} = s_p\big) \\
&= \sum_{r_0, r_1, \ldots, r_q \in \{0,1\}^K} P(e_t = r_0, e_{t-1} = r_1, \ldots, e_{t-q} = r_q) \\
&\qquad \times P\big(X_t = s_0 | X_{t-1} = s_1, \ldots, X_{t-p} = s_p, e_t = r_0, e_{t-1} = r_1, \ldots, e_{t-q} = r_q\big).
\end{aligned}$$

For the first probability on the last right-hand side, we have

$$P(e_t = r_0, e_{t-1} = r_1, \ldots, e_{t-q} = r_q) = \prod_{m=0}^{q} P(e_{t-m} = r_m) = \prod_{m=0}^{q} p_{r_m},$$

since $(e_t, t \in \mathbb{Z})$ are independent and identically distributed. For handling the second probability term, note that $X_t = (X_{t,1}, \ldots, X_{t,K})'$ with

$$\begin{aligned}
X_{t,k} = &\sum_{i=1}^{p} \sum_{l=1}^{K} a_{kl,t}^{(i)} \left[\mathbb{1}_{\{\alpha_{kl}^{(i)} \geq 0\}} X_{t-i,l} + \mathbb{1}_{\{\alpha_{kl}^{(i)} < 0\}} (1 - X_{t-i,l}) \right] \\
&+ b_{kk,t}^{(0)} e_{t,k} + \sum_{j=1}^{q} \sum_{l=1}^{K} b_{kl,t}^{(j)} \left[\mathbb{1}_{\{\beta_{kl}^{(j)} \geq 0\}} e_{t-j,l} + \mathbb{1}_{\{\beta_{kl}^{(j)} < 0\}} (1 - e_{t-j,l}) \right].
\end{aligned}$$

Since the matrix-valued i.i.d. process $(P_t, t \in \mathbb{Z})$ consists of mutually independent rows which are also independent of e_s, $s \leq t$ and X_s, $s < t$, this leads to

$$\begin{aligned}
&P\big(X_t = s_0 | X_{t-1} = s_1, \ldots, X_{t-p} = s_p, e_t = r_0, e_{t-1} = r_1, \ldots, e_{t-q} = r_q\big) \\
&= \prod_{k=1}^{K} P\big(X_{t,k} = s_{0,k} | X_{t-1} = s_1, \ldots, X_{t-p} = s_p, e_t = r_0, e_{t-1} = r_1, \ldots, e_{t-q} = r_q\big) \\
&= \prod_{k=1}^{K} P\Bigg(\sum_{i=1}^{p} \sum_{l=1}^{K} a_{kl,t}^{(i)} \left[\mathbb{1}_{\{\alpha_{kl}^{(i)} \geq 0\}} s_{i,l} + \mathbb{1}_{\{\alpha_{kl}^{(i)} < 0\}} (1 - s_{i,l}) \right] \\
&\qquad + b_{kk,t}^{(0)} r_{0,k} + \sum_{j=1}^{q} \sum_{l=1}^{K} b_{kl,t}^{(j)} \left[\mathbb{1}_{\{\beta_{kl}^{(j)} \geq 0\}} r_{j,l} + \mathbb{1}_{\{\beta_{kl}^{(j)} < 0\}} (1 - r_{j,l}) \right] = s_{0,k} \Bigg) \\
&= \prod_{k=1}^{K} \Bigg[\sum_{i=1}^{p} \sum_{l=1}^{K} |\alpha_{kl}^{(i)}| \left[\mathbb{1}_{\{\alpha_{kl}^{(i)} \geq 0\}} \delta_{s_{0,k} s_{i,l}} + \mathbb{1}_{\{\alpha_{kl}^{(i)} < 0\}} \delta_{s_{0,k}(1 - s_{i,l})} \right] \\
&\qquad + \beta_{kk}^{(0)} \delta_{s_{0,k} r_{0,k}} + \sum_{j=1}^{q} \sum_{l=1}^{K} |\beta_{kl}^{(j)}| \left[\mathbb{1}_{\{\beta_{kl}^{(j)} \geq 0\}} \delta_{s_{0,k} r_{j,l}} + \mathbb{1}_{\{\beta_{kl}^{(j)} < 0\}} \delta_{s_{0,k}(1 - r_{j,l})} \right] \Bigg],
\end{aligned}$$

which completes the proof of Lemma 3.21. The transition probability derived for gbVAR(p) processes in Lemma 3.9 follows directly as a special case for $q = 0$. □

3.7.7. Proof of Equation (3.40)

As $M = I_4$, plugging-in (3.39) into the gbVAR(1) model equation corresponding to the model proposed in Example 3.12, re-arranging to separate HY_t on the left-hand side of the equation and left-multiplying with H^+ leads to

$$Y_t = H^+ A_t^{(+)} H Y_{t-1} + H^+ A_t^{(+)} G \mathbb{1}_2 + H^+ A_t^{(-)} \mathbb{1}_4 + H^+ B_t e_t - H^+ G \mathbb{1}_2.$$

Taking expectations on both sides leads to

$$\mu_Y = H^+ \mathcal{A} H \mu_Y + H^+ \mathcal{A} G \mathbb{1}_2 + H^+ \mathcal{A}^{(-)} \mathbb{1}_4 + H^+ \mathcal{B} \mu_e - H^+ G \mathbb{1}_2 \tag{3.79}$$

and a comparison with a corresponding bivariate gbVAR(1) model equation (in expectation) for a process $(Y_t, t \in \mathbb{Z})$, that is,

$$\mu_Y = \mathcal{A}_Y \mu_Y + \mathcal{A}_Y^{(-)} \mathbb{1}_2 + B_Y \mu_\epsilon, \tag{3.80}$$

indicates, that we have to add $\pm \mathcal{A}_Y^{(-)} \mathbb{1}_2$ on the right-hand side of (3.79). Then, a direct comparison leads to $\mathcal{A}_Y = H^+ \mathcal{A} H$, which already fully determines $\mathcal{A}_Y^{(-)}$ and $\mathcal{B}_Y$, and we get

$$B_Y \mu_\epsilon = -\mathcal{A}_Y^{(-)} \mathbb{1}_2 + H^+ \mathcal{A} G \mathbb{1}_2 + H^+ \mathcal{A}^{(-)} \mathbb{1}_4 + H^+ \mathcal{B} \mu_e - H^+ G \mathbb{1}_2. \tag{3.81}$$

As $\beta_{Y,11} = 0$, we have to left-multiply both sides of (3.81) with $(0, 1)$ leading to (3.40). □

3.8. Proofs of Subsection 3.2.4

3.8.1. Proof of Lemma 3.14

Remark 3.24 (Transition probability for gbVAR(p) model)
For some vectors $\tilde{s}_0, \tilde{s}_1$ the $p+1$ step ahead transition probability of the Markov Chain representation based on a gbVAR(p) process calculates to

$$\begin{aligned}
p_{\tilde{s}_0|\tilde{s}_1}(p+1) &= P\left(Z_t = \tilde{s}_0 | Z_{t-(p+1)} = \tilde{s}_1\right) \\
&= \sum_{\substack{s_{\max\{p,1\}},\ldots,s_p \\ r_0,\ldots,r_p}} P\Big(X_t = s_0, \ldots, X_{t-p+1} = s_{p-1}, e_t = r_0 | \\
&\qquad\qquad X_{t-(p+1)} = s_{p+1}, \ldots, X_{t-2p} = s_{2p}, e_{t-p} = r_p\Big) \\
&= \sum_{\substack{s_{\max\{p,1\}},\ldots,s_p \\ r_0}} \prod_{n=0}^{p} \Big(P\left(X_{t-n} = s_n | X_{t-n-1} = s_{n+1}, \ldots, e_{t-n} = r_n, \ldots\right) \\
&\qquad\qquad \cdot P\left(e_{t-n} = r_n | X_{t-n-1} = s_{n+1}, \ldots, e_{t-n-1} = r_{n+1}, \ldots\right)\Big) \\
&= \sum_{\substack{s_{\max\{p,1\}},\ldots,s_{p+q} \\ r_q,\ldots,r_{q+p}}} \prod_{n=0}^{p+q} \Bigg(p_{r_n} \prod_{k=1}^{K} \Bigg[\sum_{i=1}^{p} \sum_{l=1}^{K} |\alpha_{kl}^{(i)}| \big(\mathbb{1}_{\{\alpha_{kl}^{(i)} \geq 0\}} \delta_{s_{n,k} s_{i+n,l}} \\
&\qquad\qquad + \mathbb{1}_{\{\alpha_{kl}^{(i)} < 0\}} \delta_{s_{n,k}(1 - s_{i+n,l})} \big) + \beta_{kk}^{(0)} \delta_{s_{n,k} r_{n,k}} \Bigg] \Bigg).
\end{aligned}$$

As the innovation process is i.i.d., the probability to take a value r_n is independent on the condition of the past time series values and we define $P(e_{t-n} = r_n) = p_{r_n}$.

Remark 3.25
Due to the stationarity condition of the time series $(X_t, t \in \mathbb{Z})$ *and its implication, see Example 3.5, we conclude that in every dimension an innovation term has an impact. We call the impact* directly *whenever* $\beta_{kk} \neq 0$ *in dimension* k *or* indirectly *if we face the case of Example 3.5(ii). We illustrate the indirect impact in Figure 3.25, showing a di-graph of a bivariate gbVAR - process. The innovation* e_2 *has a strictly positive*

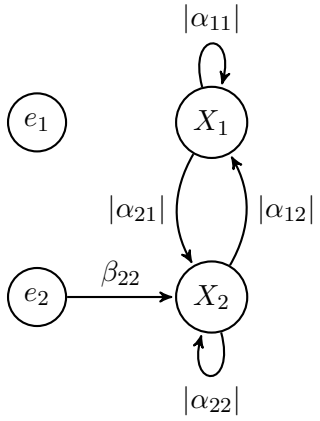


Figure 3.8.: Di - Graph of a two dimensional gbVAR(1) time series with $\beta_{11} = 0$.

probability to show up in the first dimension, as there exists a path from node e_2 *to* X_1 *via* X_2*.*

With the given Remark, we can prove the primitivity of the Markov Chain representation, when the underlying process follows a gbVAR(p) model.

Lemma 3.26
Let $(Z_t, t \in \mathbb{Z})$ *be the Markov representation of a stationary gbVAR(p) process and* $p \geq 1$*. For* $n := \max\{p, K'\}+1$ *the transition probability fulfills* $p_{s|r}(n) := P(Z_t = s | Z_{t-n} = r) > 0$ *for every* $s, r \in \{0,1\}^{(p+1)K}$*, where* K' *is the number of rows* $k \in \{1, \ldots, K\}$ *with* $\sum_{l=1}^{K} \sum_{i=1}^{p} |\alpha_{kl}^{(i)}| = 1$*.*

Proof.
Let $\mathcal{F} := \left\{k \in \{1, \ldots, K\} \middle| \beta_{kk} = 0\right\}$. For the transition probability with $\max\{p, K'\}+1$ steps with $K' = |\mathcal{F}|$ the cardinality of $\mathcal{F}$ and

$$\tilde{s}_0 := vec\,(s_0; s_1, \ldots, s_{p-1}, r_0)$$
$$\tilde{s}_1 := vec\,\left(s_{\max\{p,K'\}+1}; \ldots, s_{p+\max\{p,K'\}}, r_{\max\{p,K'\}+1}\right)$$

follows

$$P\left(Z_t = \tilde{s}_0 | Z_{t-max\{p,K'\}-1} = \tilde{s}_1\right)$$
$$= P\Big(X_t = s_0, X_{t-1} = s_1, \ldots, X_{t-p+1} = s_{p-1}, e_t = r_0 \Big| X_{t-max\{p,K'\}-1} = s_{max\{p,K'\}+1},$$
$$\ldots, X_{t-max\{p,K'\}-p} = s_{max\{p,K'\}+p}, e_{t-max\{p,K'\}-1} = r_{max\{p,K'\}+1}\Big)$$

$$
\begin{aligned}
&= \sum_{\substack{s_p,\dots,s_{max\{p,K'\}},\\ r_0,\dots,r_{max\{p,K'\}}}} \prod_{n=0}^{max\{p,K'\}} P\left(X_{t-n} = s_n | X_{t-n-1} = s_{n+1}, \dots, X_{t-n-p} = s_{n+p}, e_{t-n} = r_n\right) \\
&\qquad \cdot P\left(e_{t-n} = r_n\right) \\
&= \sum_{\substack{s_p,\dots,s_{max\{p,K'\}},\\ r_0,\dots,r_{max\{p,K'\}}}} \prod_{n=0}^{max\{p,K'\}} p_{r_n} \prod_{k=1}^{K} \sum_{l=1}^{K} \Big[\sum_{i=1}^{p} |\alpha_{kl}^{(i)}| \left(\mathbb{1}_{\{\alpha_{kl}^{(i)} \geq 0\}} \delta_{s_{n,k} s_{n+i,l}} + \mathbb{1}_{\{\alpha_{kl}^{(i)} < 0\}} \delta_{s_{n,k}\left(1 - s_{i+n,l}\right)} \right) \\
&\qquad + |\beta_{kk}| \delta_{s_{n,k} r_{0,k}} \Big] \\
&= \sum_{\substack{s_p,\dots,s_{max\{p,K'\}},\\ r_0,\dots,r_{max\{p,K'\}}}} \prod_{n=0}^{max\{p,K'\}} p_{r_n} \times \\
&\quad \underbrace{\prod_{k \in \mathcal{F}} \left(\sum_{l=1}^{K} \Big[\sum_{i=1}^{p} |\alpha_{kl}^{(i)}| \left(\mathbb{1}_{\{\alpha_{kl}^{(i)} \geq 0\}} \delta_{s_{n,k} s_{n+i,l}} + \mathbb{1}_{\{\alpha_{kl}^{(i)} < 0\}} \delta_{s_{n,k}\left(1 - s_{n+i,l}\right)} \right) \right)}_{i)} \\
&\quad \underbrace{\prod_{k \in \mathcal{F}^c} \left(\sum_{l=1}^{K} \Big[\sum_{i=1}^{p} |\alpha_{kl}^{(i)}| \left(\mathbb{1}_{\{\alpha_{kl}^{(i)} \geq 0\}} \delta_{s_{n,k} s_{n+i,l}} + \mathbb{1}_{\{\alpha_{kl}^{(i)} < 0\}} \delta_{s_{n,k}\left(1 - s_{n+i,l}\right)} \right) + \beta_{kk} \delta_{s_{n,k} r_{n,k}} \Big] \right)}_{ii)}
\end{aligned}
$$

Considering the product $ii)$ then $\forall k \in \mathcal{F}^c$ and $\forall n = 0, \dots, \max\{p, K'\}$

$$\exists r_{n,k} := s_{n,k}$$

and so for every $n = 0, \dots, \max\{p, K'\}$ there exists a positive summand and therefore the product is positive.
For the first case $i)$ it holds, since the time series $(X_t, t \in \mathbb{Z})$ is stationary and with Remark 3.25, we know that for every entry of X_t there exists a directed graph from minimal one innovation to the resulting X_t. So the probability of seeing an innovation term in an entry $X_{t,k}$ for $k \in \mathcal{F}$ is strictly positive as consequence of the stationarity assumption.
And $\forall n = 0, \dots, \max\{p, K'\}$ and $\forall k \in \mathcal{F}$

$$\exists\ s_{n,k} = \mathbb{1}_{\{\alpha_{kl_k}^{(i_k)} \geq 0\}} s_{n+i_k,l_k} + \mathbb{1}_{\{\alpha_{kl_k}^{(i_k)} < 0\}} \left(1 - s_{n+i_k,l_k}\right) > 0.$$

The above equation holds since $(X_t, t \in \mathbb{Z})$ follows a gbVAR(p) model equation. Thus the new $X_{t,k}$ value is driven by a lagged value $X_{t-n,l}$ and therefore

$$\prod_{n=0}^{max\{p,K'\}} \prod_{k \in \mathcal{F}} \left(\sum_{l=1}^{K} \Big[\sum_{i=1}^{p} |\alpha_{kl}^{(i)}| \left(\mathbb{1}_{\{\alpha_{kl}^{(i)} \geq 0\}} \delta_{s_{n,k} s_{n+i,l}} + \mathbb{1}_{\{\alpha_{kl}^{(i)} < 0\}} \delta_{s_{n,k}\left(1 - s_{n+i,l}\right)} \right) \right) > 0. \tag{3.82}$$

Overall it follows that

$$P\left(Z_t = \tilde{s}_0 | Z_{t-max\{p,K'\}-1} = \tilde{s}_1\right) > 0.$$

□

3.8.2. Proof of Theorem 3.15

Proof.
Following Weiß (2009a) define $[X_t]_a^b := (X_b, \ldots, X_a)$ for all $a, b \in \mathbb{Z}$ and with $b > a$. Let $l, k \geq \tilde{p} = \max\{1, p\}$, $a_1 \in \mathcal{V}^{k \cdot K}$ and $a_2 \in \mathcal{V}^{l \cdot K}$ and $n \in \mathbb{N}$ be arbitrarily given. Then we have to consider $|P(A \cap B) - P(A)\,P(B)|$ for the sets of X :

$$
\begin{aligned}
&|P\left([X_t]_n^{n+k-1} = a_1, [X_t]_{-l+1}^0 = a_2\right) - P\left([X_t]_n^{n+k-1} = a_1\right) P\left([X_t]_{-l+1}^0 = a_2\right)| \\
=&\Big| \sum_{b_1, b_2} \Big(P\left([X_t]_n^{n+k-1} = a_1, [e_t]_{n-q+\tilde{p}}^{n+k-1} = b_1, [X_t]_{-l+1}^0 = a_2 . [e_t]_{-l+1-q+\tilde{p}}^0 = b_2\right) \\
&- P\left([X_t]_n^{n+k-1} = a_1, [e_t]_{n-q+\tilde{p}}^{n+k-1} = b_1\right) P\left([X_t]_{-l+1}^0 = a_2, [e_t]_{-l+1-q+\tilde{p}}^0 = b_2\right)\Big)\Big| \\
=&\Big| \sum_{b_1, b_2} \Big(P\left[Z_n + k - 1 = z_{n+k-1}, \ldots, Z_{n+\tilde{p}-1} = z_{n+\tilde{p}-1}, Z_0 = z_0, \ldots, Z_{-l+\tilde{p}} = z_{-l+\tilde{p}}\right] \\
&- P\left(Z_n + k - 1 = z_{n+k-1}, \ldots, Z_{n+\tilde{p}-1} = z_{n+\tilde{p}-1}\right) P\left(Z_0 = z_0, \ldots, Z_{-l+\tilde{p}} = z_{-l+\tilde{p}}\right)\Big)\Big|
\end{aligned}
$$

Inserting the innovation process states and integrate over all appropriate vectors $b_1 \in \mathcal{V}^{(k-\tilde{p}) \cdot K}$ and $b_2 \in \mathcal{V}^{(l-\tilde{p}) \cdot K}$, the transition probabilities can be rewritten to the Markov Chain representation as given in (3.43). The values z_i are then given by the values of the vectors a_1, a_2, b_1 and b_2.

Due to the primitivity of the Markov Chain representation and Corollary 11.2.3.2 in Weiß (2009a), the Markov Chain representation of the gbVAR process is ψ - and φ - mixing with exponential decreasing weights g_n. And so it follows

$$
\begin{aligned}
&\Big|P\left([X_t]_n^{n+k-1} = a_1, [X_t]_{-l+1}^0 = a_2\right) - P\left([X_t]_n^{n+k-1} = a_1\right) P\left([X_t]_{-l+1}^0 = a_2\right)\Big| \\
\leq &\sum_{b_1, b_2} g_{n+\tilde{p}-1} P\left(Z_{n+k-1} = z_{n+k-1}, \ldots, Z_{n+\tilde{p}-1} = z_{n+\tilde{p}-1}\right) P\left(Z_{-l+1} = z_{-l+1}, \ldots, Z_0 = z_0\right) \\
= &g_{n+\tilde{p}-1} P\left([X_t]_n^{n+k-1} = a_1\right) P\left([X_t]_{-l+1}^0 = a_2\right)
\end{aligned}
$$

So for $f_n := g_{n+\tilde{p}-1}$ the gbVAR process $(X_t, t \in \mathbb{Z})$ itself is ψ - and φ mixing. In the case of $k < p$ by assembling the X_t vectors the missing $p - k$ vectors have to be added to the considered difference by summing over all possible values.

$$
\begin{aligned}
&\Big|P\left([X_t]_n^{n+k-1} = a_1, [X_t]_{-l+1}^0 = a_2\right) - P\left([X_t]_n^{n+k-1} = a_1\right) P\left([X_t]_{-l+1}^0 = a_2\right)\Big| \\
&= \Big| \sum_{a_3} P\left([X_t]_n^{n+k-1} = a_1, [X_t]_{n+k-p}^{n-1} = a_3, [X_t]_{-l+1}^0 = a_2\right) - \\
&P\left([X_t]_n^{n+k-1} = a_1, [X_t]_{n+k-p}^{n-1} = a_3\right) P\left([X_t]_{-l+1}^0 = a_2\right)\Big|
\end{aligned}
$$

Now the proof is the same as above by summing over all needed innovation terms and rewriting the vectors by using the Markov representation. For the other cases as $l < p$ and the combination of both the missing vectors have to be inserted and then the statement is proved. □

3.9. Bootstrap Confidence Intervals

	$T = 100$	500	1000		$T = 100$	500	1000
$\alpha_{11}^{(1)}$	0.935	0.959	0.953	$\alpha_{41}^{(1)}$	0.938	0.941	0.934
$\alpha_{12}^{(1)}$	0.938	0.941	0.938	$\alpha_{42}^{(1)}$	0.929	0.954	0.953
$\alpha_{13}^{(1)}$	0.938	0.943	0.937	$\alpha_{43}^{(1)}$	0.929	0.948	0.954
$\alpha_{14}^{(1)}$	0.942	0.949	0.936	$\alpha_{44}^{(1)}$	0.925	0.943	0.942
$\alpha_{21}^{(1)}$	0.898	0.915	0.947	$\beta_{11}^{(0)}$	0.869	0.932	0.948
$\alpha_{22}^{(1)}$	0.900	0.948	0.942	$\beta_{22}^{(0)}$	0.799	0.924	0.947
$\alpha_{23}^{(1)}$	0.908	0.941	0.954	$\beta_{33}^{(0)}$	0.834	0.936	0.940
$\alpha_{24}^{(1)}$	0.909	0.949	0.961	$\beta_{44}^{(0)}$	0.850	0.918	0.934
$\alpha_{31}^{(1)}$	0.921	0.929	0.937	$\mu_{X,1}$	0.881	0.941	0.943
$\alpha_{32}^{(1)}$	0.914	0.941	0.940	$\mu_{X,2}$	0.791	0.900	0.918
$\alpha_{33}^{(1)}$	0.920	0.944	0.940	$\mu_{X,3}$	0.823	0.921	0.951
$\alpha_{34}^{(1)}$	0.938	0.946	0.954	$\mu_{X,4}$	0.845	0.946	0.944

Table 3.6.: Coverage rates of bootstrap confidence intervals for each entry of the parameter matrices $\mathcal{A}^{(1)}$, $\mathcal{B}^{(0)}$ and μ_X for DGP2.

	$T = 100$	500	1000		$T = 100$	500	1000
$\alpha_{11}^{(1)}$	0.953	0.944	0.941	$\alpha_{11}^{(2)}$	0.926	0.926	0.949
$\alpha_{12}^{(1)}$	0.958	0.948	0.962	$\alpha_{12}^{(2)}$	0.949	0.952	0.950
$\alpha_{13}^{(1)}$	0.942	0.947	0.951	$\alpha_{13}^{(2)}$	0.951	0.939	0.940
$\alpha_{21}^{(1)}$	0.950	0.960	0.950	$\alpha_{21}^{(2)}$	0.954	0.947	0.957
$\alpha_{22}^{(1)}$	0.960	0.957	0.944	$\alpha_{22}^{(2)}$	0.927	0.946	0.957
$\alpha_{23}^{(1)}$	0.929	0.953	0.942	$\alpha_{23}^{(2)}$	0.943	0.965	0.955
$\alpha_{31}^{(1)}$	0.953	0.949	0.939	$\alpha_{31}^{(2)}$	0.952	0.952	0.956
$\alpha_{32}^{(1)}$	0.938	0.958	0.965	$\alpha_{32}^{(2)}$	0.948	0.932	0.955
$\alpha_{33}^{(1)}$	0.948	0.951	0.948	$\alpha_{33}^{(2)}$	0.950	0.952	0.945
$\beta_{11}^{(0)}$	0.869	0.926	0.948	$\mu_{X,1}$	0.717	0.855	0.909
$\beta_{22}^{(0)}$	0.879	0.959	0.976	$\mu_{X,2}$	0.711	0.838	0.896
$\beta_{33}^{(0)}$	0.888	0.954	0.972	$\mu_{X,3}$	0.747	0.844	0.879

Table 3.7.: Coverage rates of bootstrap confidence intervals for each entry of the parameter matrices $\mathcal{A}^{(1)}$, $\mathcal{A}^{(2)}$, $\mathcal{B}^{(0)}$ and μ_X for DGP3.

4. Vector - valued New Discrete ARMA Processes

Based on: Jentsch, C. and Reichmann, L.
***Vector - valued New Discrete ARMA Processes** - Working Paper*

4.1. Introduction

In the last years, multivariate time series analysis (Lütkepohl (2005)) gained more and more importance as multi-dimensional time series data of any kind is collected in many fields of application. When the multivariate time series is categorical, common modeling approaches for continuous-valued multivariate data are typically not suitable anymore. However, categorical valued serially dependent data arises in many fields of application, e.g. in text and speech recognition, biology as human physiology and DNA analysis as well as in computer science and economics. For instance, in biology, the DNA sequence contains plenty of information of a creature or virus. By considering the dinucleotide frequencies in the DNA sequences, Dehnert et al. (2005) show that the short range correlation serves as evolutionary fingerprints of eukaryotes, where the chromosomes form the categories of the analysis. In speech analysis, Weiß (2009b) analyzed the letters and dependence structure of the natural speech in Shakespeare's poem *Venus and Adonis*, where each letter coincides with one category. Another approach is to categorize or pre-define certain states, for instance, Stoffer et al. (1988) considered for the infant sleep data set to investigate the effects of prenatal alcohol exposure on the neurophysiological development. By using an electroencephalography (EEG), six states of the infant sleep are recorded and analyzed by Biswas and Song (2009). The economic situation and its analysis has an important role for e.g. policy makers, analysts and central bankers. Macroeconomic data sets, e.g. the gross domestic product (GDP), contain broad information of the status of an economy. To reduce complexity, Mazzi (2015) investigated the business cycle clock (BCC), which categorizes the economic situation into six different states where the main aim is to finally detect forthcoming recession periods of the countries economy. We analyze this data set for six European countries in Section 4.5.

Whereas Markov processes are clearly the most natural models of choice for describing serially dependent and discrete-valued categorical data. However, as mentioned e.g. by Moysiadis and Fokianos (2014) already for univariate data, the number of parameter increases by the selected order of the time series: $(m+1)^p$ parameter are necessary to specify a univariate p-th order Markov chain with a state space consisting of $m+1$ categories. The issue of a large number of parameters becomes even worse if the cross-sectional dimension of the data increases. Therefore the application to multivariate categorical data leads to parameter estimation problems as well as identification issues.

A more parsimonious but still Markovian model for stationary univariate categorical time series was introduced by Jacobs and Lewis (1983), the New Discrete AutoRegres-

sive Moving Average (NDARMA) class. Such processes have the typical autoregressive moving average structure and, in addition, its autocorrelation structure coincides with that of standard ARMA models. More precisely, let $(X_t, t \in \mathbb{Z})$ and $(e_t, t \in \mathbb{Z})$ be categorical univariate time dependent processes with state space $\mathcal{V} = \{0, 1, \ldots, m\}$. The innovation process $(e_t, t \in \mathbb{Z})$ is assumed to be an independent and identically distributed (i.i.d.) process with marginal distribution $P(e_t = j) = p_e(j) > 0$ for $j = 0, 1, \ldots, m$ such that e_t is independent of $(X_s)_{s<t}$. Let $(P_t, t \in \mathbb{Z})$ be i.i.d. random vectors defining a selection mechanism by

$$P_t := \left[a_t^{(1)}, \ldots, a_t^{(p)}, b_t^{(0)}, \ldots, b_t^{(q)}\right] \sim Mult\,(1; \mathcal{P}) \tag{4.1}$$

with probability vector $\mathcal{P} := \left[\alpha^{(1)}, \ldots, \alpha^{(p)}, \beta^{(0)}, \ldots, \beta^{(q)}\right]$ where $\sum_{i=1}^{p} \alpha^{(i)} + \sum_{j=0}^{q} \beta^{(j)} = 1$, which is independent of the innovation process $(e_t, t \in \mathbb{Z})$ and of $(X_s, s < t)$.

The process $(X_t, t \in \mathbb{Z})$ is called to follow an NDARMA(p,q) model, if it suffices the recursion

$$X_t = a_t^{(1)} X_{t-1} + \ldots + a_t^{(p)} X_{t-p} + b_t^{(0)} e_t + \ldots + b_t^{(q)} e_{t-q}. \tag{4.2}$$

In comparison to classical ARMA models, where X_t is a weighted average of the previous time series values and innovation terms, the multinomial sampling mechanism ensures that the process stays within the state space $\mathcal{V}$ as for each $t \in \mathbb{Z}$ only one entry of P_t is equal to one and all others are zero. That is, whenever one $a_t^{(i)},\ i = 1, \ldots, p$ takes the value one, the multinomial distribution chooses a predecessor $X_{t-1}, \ldots, X_{t-p}$ or, if $b_t^{(j)} = 1$ for some $j = 0, \ldots, q$ an innovation $e_t, \ldots, e_{t-q}$, respectively, becomes the new value X_t of the time series.

The NDARMA model maintains the typical ARMA structure and its nice interpretability, as well as the categorical characteristic of the data. For this reason, the model contains $p+q+1$ parameters. In contrast, whenever univariate categorical data is modeled using a univariate p-th order Markov Chain, $(m+1)^p$ transition probabilities are required to fully determine the model. For NDVARMA processes, Jacobs and Lewis (1983) stated Yule - Walker - type equations that allow for model identification. Signed dependence measures that are more suitable for categorical data, such as Cohen's κ and Cramer's ν, are considered in Weiß (2009a). Möller and Weiß (2020) extended the parsimonious NDARMA model class to allow vector valued categorical processes. However, the model parameters are still *scalar* for the multinomial selection mechanism, such that the vectors are selected as a whole, instead of individual entries. Nevertheless, the introduced Generalized Discrete ARMA (GDARMA) processes make additional use of a variation function applied to lagged observations and innovations to increase the entry-wise variation over time. Nevertheless, the GDARMA class does not allow for a meaningful modeling of cross-sectional dependence. Hence, we aim for a more flexible model class that allows to capture cross-sectional dependence structures, such that the selection mechanism allows to choose the new states in each dimension of X_t *independently* from the others. Hence, a row-wise applied multinomial selection mechanism defining the New Discrete Vector - valued ARMA (NDVARMA) model class, leads to more flexibility in comparison to GDARMA models.

4.1.1. A bivariate extension of the NDAR(1) process: the NDVAR(1) model

By stacking two independent NDAR(1) processes $(X_t, t \in \mathbb{Z})$ and $(Y_t, t \in \mathbb{Z})$, we get a bivariate process

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \begin{pmatrix} a_{t,X} X_{t-1} + b_{t,X} e_{t,X} \\ a_{t,Y} Y_{t-1} + b_{t,Y} e_{t,Y} \end{pmatrix} = \begin{pmatrix} a_{t,X} & 0 \\ 0 & a_{t,Y} \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \end{pmatrix} + \begin{pmatrix} b_{t,X} & 0 \\ 0 & b_{t,Y} \end{pmatrix} \begin{pmatrix} e_{t,X} \\ e_{t,Y} \end{pmatrix}.$$

However, this process is not yet suitable for studying cross-sectional dependence structures and does not require further theoretical considerations, as all parameter matrices have a diagonal structure. Of course, a natural approach is to allow for non-diagonal parameter matrices. We want to illustrate this by considering a bivariate first-order New Discrete Vector AutoRegressive process (called NDVAR(1)), where for $k = 1, 2$ the i.i.d. random vectors $P_{t,k\bullet}$ follow multinomial distributions, such that

$$\begin{aligned} P_{t,1\bullet} &:= [a_{t,11}, a_{t,12}, b_{t,11}, 0] \sim Mult\,(1; [\alpha_{11}, \alpha_{12}, \beta_{11}, 0])\,, \\ P_{t,2\bullet} &:= [a_{t,21}, a_{t,22}, 0, b_{t,22}] \sim Mult\,(1; [\alpha_{21}, \alpha_{22}, 0, \beta_{22}])\,. \end{aligned}$$

The first two entries of $P_{t,k\bullet}$ ($a_{t,k1}$ and $a_{t,k2}$) correspond to choosing one of the entries of the predecessor vector $(X_{t-1,1}, X_{t-1,2})'$ as new state of the process X_t, where $b_{t,kk}$ corresponds to the innovation process $e_{t,k}$. Analogously to the NDARMA process, the model parameter α_{k1}, α_{k2} and β_{kk} still have to fulfill the condition to sum up to one, e.g. $\alpha_{k1} + \alpha_{k2} + \beta_{kk} = 1$, as they are probabilities for the multinomial distribution.

Hence, the NDVAR(1) process follows the model equation

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} = \begin{pmatrix} a_{t,11} & a_{t,12} \\ a_{t,21} & a_{t,22} \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \end{pmatrix} + \begin{pmatrix} b_{t,11} & 0 \\ 0 & b_{t,22} \end{pmatrix} \begin{pmatrix} e_{t,1} \\ e_{t,2} \end{pmatrix} \tag{4.3}$$

where $(e_t, t \in \mathbb{Z})$ an i.i.d. bivariate categorical process taking values in $\mathcal{V}^2$.

The paper is structured as follows. First, we define the vector valued NDARMA process in Section 4.2 and illustrate it by an example. Stochastic properties of the process are stated in Section 4.3, where first the stationarity condition and stationary solutions are discussed. Yule - Walker - type equations are derived for NDVARMA models which lead to the common Yule-Walker estimators for the purely autoregressive NDVAR case. Transition probabilities are derived that allow to establish mixing properties of NDVARMA models based on a useful Markov chain representation. Further, we discuss the identification of the innovation process distribution. For illustration, NDVAR processes are employed to analyze a business cycle clock data set.

4.2. Vector new discrete ARMA processes

In this section, we extend univariate NDARMA models, introduced by Jacobs and Lewis (1983), to obtain multivariate NDVARMA processes of dimension K. To capture cross sectional dependence, the scalar model parameters $\alpha^{(1)}, \ldots, \alpha^{(p)}$ and $\beta^{(0)}, \ldots, \beta^{(q)}$ of an NDARMA process become $K \times K$ dimensional parameter matrices $\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(p)}$ and $\mathcal{B}^{(0)}, \ldots, \mathcal{B}^{(q)}$, as in the setup of classical vector ARMA models. Let the $K \times (p + q + 1)K$ dimensional parameter matrix

$$\mathcal{P} = \left[\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(p)}, \mathcal{B}^{(0)}, \ldots, \mathcal{B}^{(q)}\right] \tag{4.4}$$

summarize the autoregressive coefficients $\mathcal{A}^{(i)} = (\alpha_{kl}^{(i)})_{k,l=1,\dots,K}$ and the moving average coefficients $\mathcal{B}^{(j)} = (\beta_{kl}^{(j)})_{k,l=1,\dots,K}$. Similar to the NDARMA model setup, all entries $\alpha_{k,l}^{(i)}, \beta_{k,l}^{(j)}$ are allowed to take values in $[0,1]$. And, as a row-wise multinomial selection approach is natural, the model parameter $\mathcal{P}$ has to satisfy

$$\sum_{i=1}^{p}\sum_{l=1}^{K}\alpha_{kl}^{(i)} + \sum_{j=0}^{q}\sum_{l=1}^{K}\beta_{kl}^{(j)} = 1$$

for all $k = 1,\dots,K$. These extensions enable us to define New Discrete Vector-valued ARMA processes of orders $p \in \mathbb{N}_0$ and $q \in \mathbb{N}_0$ as follows.

Definition 4.1 (New discrete vector-valued ARMA models)
Let $(X_t, t \in \mathbb{Z})$ be a stationary K-dimensional process taking values in $\mathcal{V}^K = \{0,1,\dots,m\}^K$. Let $(e_t, t \in \mathbb{Z})$ be an i.i.d. K-dimensional discrete valued process, such that $e_t = (e_{t,1},\dots,e_{t,K})'$ is independent of $(X_s, s < t)$ with marginal distribution specified by $p_e(r) = P(e_t = r)$ for $r \in \mathcal{V}^K$ such that $P(e_{t,k} = j) > 0$ for all $k = 1,\dots,K$ and all $j \in \mathcal{V}$ with $\Sigma_e = (\sigma_{e,kl})_{k,l=1,\dots,K} = Cov(e_t) > 0$. Let $\mathcal{P}$ be the parameter matrix in (4.4) with diagonal $\mathcal{B}^{(0)} \neq 0_{K\times K}$, such that $\mathcal{P}\mathbb{1}_{K(p+q+1)} = \mathbb{1}_K$. Further, let $(P_t, t \in \mathbb{Z})$ with $P_t := \left[A_t^{(1)},\dots,A_t^{(p)},B_t^{(0)},\dots,B_t^{(q)}\right]$ be a $K \times K(p+q+1)$ dimensional i.i.d. process with mutually independent rows $(P_{t,k\bullet}, t \in \mathbb{Z})$, $k = 1,\dots,K$, such that

$$P_{t,k\bullet} := \left[a_{t,k\bullet}^{(1)},\dots,a_{t,k\bullet}^{(p)},b_{t,k\bullet}^{(0)},b_{t,k\bullet}^{(1)},\dots,b_{t,k\bullet}^{(q)}\right] \sim Mult\left(1;\mathcal{P}_{t,k\bullet}\right), \tag{4.5}$$

which are independent of $(e_t, t \in \mathbb{Z})$ and $(X_s, s < t)$. Here, we set $A_t^{(i)} = (a_{t,kl}^{(i)})_{k,l=1,\dots,K}$, $B_t^{(0)} = diag(b_{t,11}^{(0)},\dots,b_{t,KK}^{(0)})$ and $B_{t,kl}^{(j)} = (b_{t,kl}^{(j)})_{k,l=1,\dots,K}$ with $a_{t,k\bullet}^{(i)} = (a_{t,k1}^{(i)},\dots,a_{t,kK}^{(i)})$, $i = 1,\dots,p$ and $b_{t,k\bullet}^{(j)}$, $j = 1,\dots,q$ defined similarly. Then, the process $(X_t, t \in \mathbb{Z})$ is said to be a New Discrete Vector ARMA(p,q) process (NDVARMA(p,q)), if it follows the recursion

$$X_t = \sum_{i=1}^{p} A_t^{(i)} X_{t-i} + B_t^{(0)} e_t + \sum_{j=1}^{q} B_t^{(j)} e_{t-j}, \quad t \in \mathbb{Z}. \tag{4.6}$$

In the case, of $q = 0$, the process is called an NDVAR(p) process and for p=0, e.g. if $X_t = B_t^{(0)}e_t + \dots, B_t^{(q)}e_{t-q}$ it is said to be an NDVMA(q) process.

Note that $E(A_t^{(i)}) = \mathcal{A}^{(i)}$, $i = 1,\dots,p$ and $E(B_t^{(j)}) = \mathcal{B}^{(j)}$, $j = 0,\dots,q$ by construction. $\mathcal{B}^{(0)}$ is assumed to be diagonal due to identification issues otherwise, whereas we allow for Σ_e to be non-diagonal. Due to the independently applied sampling mechanism in each row, $X_{t,k}$, $k = 1,\dots,K$ are independent given the past values of the time series $X_{t-1},\dots,X_{t-p}$ and the innovations $e_t,\dots,e_{t-q}$.

The following example illustrates the above definition for the case of a bivariate NDVAR(1) process.

Example 4.2 (Bivariate NDVAR(1) process)
Consider a $K = 2$ dimensional NDVAR process $(X_t, t \in \mathbb{Z})$ of order $p = 1$ with $m = 4$ leading to $\mathcal{V} = \{0,1,2,3,4\}$. The process follows the recursion

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} = \begin{pmatrix} a_{t,11} & a_{t,12} \\ a_{t,21} & a_{t,22} \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \end{pmatrix} + \begin{pmatrix} b_{t,11} & 0 \\ 0 & b_{t,22} \end{pmatrix} \begin{pmatrix} e_{t,1} \\ e_{t,2} \end{pmatrix}$$

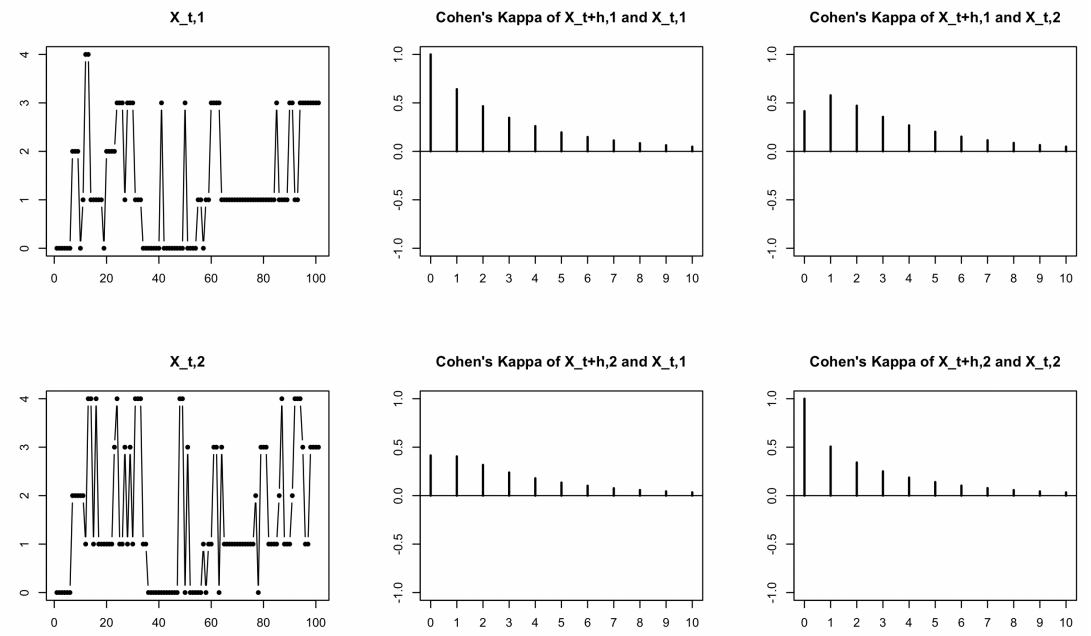


Figure 4.1.: Realization of an NDVAR(1) process of dimension $K = 2$ with $m + 1 = 5$ states of sample size $n = 100$ as specified in Example 4.2 together with corresponding Cohen's κ plots.

with parameter matrix $\mathcal{P} = [\mathcal{A}, \mathcal{B}]$*, where*

$$\mathcal{A} = \begin{pmatrix} 0.57 & 0.27 \\ 0.45 & 0.35 \end{pmatrix} \quad \textit{and} \quad \mathcal{B}^{(0)} = \begin{pmatrix} 0.16 & 0.00 \\ 0.00 & 0.20 \end{pmatrix}.$$

The processes $(e_{t,1}, t \in \mathbb{Z})$ *and* $(e_{t,2}, t \in \mathbb{Z})$ *are chosen to be independent with distributions specified by*

$$P(e_{t,1} = 0) = 0.15, \quad P(e_{t,1} = 1) = 0.35, \quad P(e_{t,1} = 2) = 0.20, \quad P(e_{t,1} = 3) = 0.20,$$
$$P(e_{t,1} = 4) = 0.10, \quad P(e_{t,2} = 0) = 0.13, \quad P(e_{t,2} = 1) = 0.15, \quad P(e_{t,2} = 2) = 0.15,$$
$$P(e_{t,2} = 3) = 0.17, \quad P(e_{t,2} = 4) = 0.40.$$

The random matrices A_t *and* B_t *are specified by the row-wise independently applied multinomial distributions*

$$P_{t,1\bullet} = [a_{t,11}, a_{t,12}, b_{t,11}, 0] \sim Mult(1; [0.57, 0.27, 0.16, 0.00]),$$
$$P_{t,2\bullet} = [a_{t,21}, a_{t,22}, 0, b_{t,22}] \sim Mult(1; [0.45, 0.35, 0.00, 0.20]).$$

One resulting realization of the described process is shown in Figure 4.1. As autocovariances are not meaningful to describe the dependence of categorical data, we present Cohen's κ *plots in Figure 4.1. Weiß (2013) used the Cohen's* κ *to measure the degree of signed serial dependence and we adapt this to the multivariate case by considering*

$$\kappa(X_{t,k}, X_{t-h,l}) = \frac{\sum_{j=0}^{m} \big[P(X_{t,k} = j, X_{t-h,l} = j) - P(X_{t,k} = j)P(X_{t-h,l} = j)\big]}{1 - \sum_{j=0}^{m} P(X_{t,k} = j)P(X_{t-h,l} = j)} \tag{4.7}$$

as a measure of serial and cross - sectional dependence in categorical time series. Nevertheless, for the univariate NDARMA model with a cardinally interpretable state space $\mathcal{V}$*, Weiß (2009a) showed that Cohen's* κ *and autocorrelations coincide, where the first measure is meaningful also for the nominal range of a state space.*

4.3. Stochastic Properties of the NDVARMA processes

4.3.1. Stationarity conditions and moving-average-type representations

First, we establish stationarity conditions and derive the stationary solution. By recursively plugging-in the NDVARMA model equation, the stationary solution can be derived having a common moving - average - type representation. For this purpose, we first illustrate the recursion for the case $p = 1$ and $q = 0$ leading to an NDVAR(1) process. By plugging-in $d \in \mathbb{N}_0$ times the NDVAR recursion (4.6), we get

$$X_t = \prod_{j=0}^{d} A_{t-j} X_{t-(d+1)} + \sum_{i=0}^{d} \left(\prod_{j=0}^{i-1} A_{t-j} \right) e_{t-i} = \zeta_d X_{t-(d+1)} + \sum_{i=0}^{d} \zeta_{i-1} B_{t-i} e_{t-i}, \quad (4.8)$$

where $\zeta_i := \prod_{j=0}^{i} A_{t-j}$ for $i \in \mathbb{N}_0$ with $\zeta_{-1} = I_K$.

Hence, whenever ζ_d vanishes for $d \to \infty$, (4.8) leads to a moving - average - type representation. For NDARMA(p,q) processes with arbitrary p and q, such an NDVMA(∞) representation can be deduced by following Lütkepohl (2005)[Chapter 11.3.2]. NDVARMA(p,q) processes $(X_t, t \in \mathbb{Z})$ can be rewritten as a $K(p+q)$ dimensional NDVAR(1) process $(\widetilde{X}_t, t \in \mathbb{Z})$. For this purpose, we define

$$\widetilde{X}_t = (X_t', \ldots, X_{t-p+1}', e_t', \ldots, e_{t-q+1}')', \quad \widetilde{e}_t = (e_t', 0, \ldots, 0, e_t', 0, \ldots, 0)'$$

and random coefficients matrices $\widetilde{A}_t$ and $\widetilde{B}_t$ containing the random matrices of the NDVARMA process $A_t^{(i)}$, $i = 1, \ldots, p$ and $B_t^{(j)}$, $j = 0, \ldots, q$. Precisely, let

$$\widetilde{A}_t := \begin{pmatrix} \widetilde{A}_{t,11} & \widetilde{A}_{t,12} \\ \widetilde{A}_{t,21} & \widetilde{A}_{t,22} \end{pmatrix} \quad \text{and} \quad \widetilde{B}_t := \begin{pmatrix} B_t^{(0)} & 0_{K \times K(p+q-1)} \\ 0_{K(p-1)\times K} & 0_{K(p-1)\times K(p+q-1)} \\ I_K & 0_{K\times K(p+q-1)} \\ 0_{K(q-1)\times K} & 0_{K(q-1)\times K(p+q-1)} \end{pmatrix}$$

be $K(p+q) \times K(p+q)$ matrices, where

$$\widetilde{A}_{t,11} := \begin{pmatrix} A_t^{(1)} & \ldots & A_t^{(p-1)} & A_t^{(p)} \\ I_K & & 0_{K\times K} & 0_{K\times K} \\ & \ddots & & \vdots \\ 0_{K\times K} & \ldots & I_K & 0_{K\times K} \end{pmatrix}, \quad \widetilde{A}_{t,12} := \begin{pmatrix} B_t^{(1)} & \ldots & B_t^{(q)} \\ 0_{K\times K} & \ldots & 0_{K\times K} \\ \vdots & & \vdots \\ 0_{K\times K} & \ldots & 0_{K\times K} \end{pmatrix},$$

$$\widetilde{A}_{t,22} := \begin{pmatrix} 0_{K\times K} & \ldots & 0_{K\times K} & 0_{K\times K} \\ I_K & & 0_{K\times K} & 0_{K\times K} \\ & \ddots & & \vdots \\ 0_{K\times K} & \ldots & I_K & 0_{K\times K} \end{pmatrix}$$

are $Kp \times Kp$, $Kp \times Kq$ and $Kq \times Kq$ matrices, respectively, and $A_{t,21} := 0_{Kq\times Kp}$. Based on the introduced notation, the NDVAR(1) process follows the recursion

$$\widetilde{X}_t = \widetilde{A}_t \widetilde{X}_{t-1} + \widetilde{B}_t \widetilde{e}_t. \quad (4.9)$$

Now, the moving - average - type representation can be derived as already shown in (4.8), such that

$$\widetilde{X}_t = \widetilde{\zeta}_d \widetilde{X}_{t-(d+1)} + \sum_{i=0}^{d} \widetilde{\zeta}_{i-1} \widetilde{B}_{t-i} \widetilde{e}_{t-i} \quad (4.10)$$

where $\widetilde{\zeta}_i := \prod_{j=0}^{i} \widetilde{A}_{t-j}$ for $i \in \mathbb{N}_0$ with $\widetilde{\zeta}_{-1} = I_K$. Setting $\widetilde{\mathcal{A}} := E(\widetilde{A}_t)$, the first term vanishes for $d \to \infty$, if and only if

$$det(I_{K(p+q)} - \widetilde{\mathcal{A}}z) \neq 0 \quad \forall z \in \mathbb{C} : |z| \leq 1 \tag{4.11}$$

and the $K(p+q)$ dimensional process $(\widetilde{X}_t, t \in \mathbb{Z})$ has a moving average representation. Due to the block structure of $\widetilde{\mathcal{A}}$, the stationarity condition (4.11) becomes equivalent to

$$det(I_K - \mathcal{A}^{(1)}z \ldots - \mathcal{A}^{(p)}z^p) \neq 0 \quad \forall z \in \mathbb{C} : |z| \leq 1, \tag{4.12}$$

see e.g.Lütkepohl (2005)[Eq. 11.3.9]. Representation (4.9) enables us to state the moving - average - type representation of $(X_t, t \in \mathbb{Z})$ as well.

Theorem 4.3 (Moving-Average representation of NDVARMA processes)
Let $(X_t, t \in \mathbb{Z})$ be a K-dimensional stationary NDVARMA(p,q) process satisfying (4.12).

(i) If $p = 1$, $q = 0$, the NDVAR(1) model has an NDVMA(∞)-type representation

$$X_t = \sum_{i=0}^{\infty} \zeta_{i-1} B_{t-i} e_{t-i}, \quad t \in \mathbb{Z}$$

converging in L_1.

(ii) If $p \in \mathbb{N}_0$, $q \in \mathbb{N}$, the NDVARMA(p,q) process fulfills the NDVAR(1) representation (4.9) *and has an NDVMA(∞)-type representation*

$$X_t = J\widetilde{X}_t = J\left(\sum_{i=0}^{\infty} \widetilde{\zeta}_{i-1} \widetilde{B}_{t-i} \widetilde{e}_{t-i}\right), \quad t \in \mathbb{Z}, \tag{4.13}$$

converging in L_1, where $J := \left[I_K, 0_{K \times K(p+q+1)}\right]$.

Example 4.4 (Cont. of Example 4.2)
For the NDVAR(1) process as specified in Example 4.2, the parameter matrix $\mathcal{A}$ has eigenvalues $\lambda \in \{0.8255, 0.0945\}$. Hence, as all eigenvalues are strictly smaller than one, the NDVAR(1) process is stationary and admits a moving - average representation.

A necessary and sufficient condition for an NDVAR(1) process to fulfill the stationarity condition (4.12) and to admit an NDVMA(∞) representation is as follows:

(a) at least one diagonal entry of the parameter matrix $\mathcal{B}^{(0)}$ is non zero, and

(b) for each univariate process $(X_{t,k}, t \in \mathbb{Z})$, $k = 1, \ldots, K$, there exists at least one innovation process $(e_{t,l}, t \in \mathbb{Z})$, $l = 1, \ldots, K$ such that $e_{t,l}$ enters $(X_{t,k}, t \in \mathbb{Z})$ with strictly positive probability after finitely many time steps.

In particular, if *all* diagonal entries of $\mathcal{B}^{(0)}$ are strictly positive, condition (b) is automatically fulfilled. If all diagonal entries of $\mathcal{B}^{(0)}$ are zero, the largest eigenvalue of $\mathcal{A}^{(1)}$ is always equal to one. We want to illustrate the situation with an example.

Remark 4.5 (Stationarity condition for NDVAR(1) process)
Consider a bivariate NDVAR(1) process with model parameter $\mathcal{P} := [\mathcal{A}, \mathcal{B}]$ *with*

$$\mathcal{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \quad \textit{and} \quad \mathcal{B} = \begin{pmatrix} \beta_{11} & 0 \\ 0 & \beta_{22} \end{pmatrix},$$

where $\alpha_{11} + \alpha_{12} + \beta_{11} = 1$ *and* $\alpha_{21} + \alpha_{22} + \beta_{22} = 1$. *Then, the process follows the recursion*

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \end{pmatrix} = \begin{pmatrix} a_{t,11} & a_{t,12} \\ a_{t,21} & a_{t,22} \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \end{pmatrix} + \begin{pmatrix} b_{t,11} & 0 \\ 0 & b_{t,22} \end{pmatrix} \begin{pmatrix} e_{t,1} \\ e_{t,2} \end{pmatrix}, \quad t \in \mathbb{Z}.$$

Figure 4.2 illustrates all possible selection paths of the NDVAR(1) process. We consider the following special cases:

1. *If* $\beta_{kk} > 0$, $k = 1, 2$, *both innovation terms have a strict positive probability to be selected. Hence, all row-sums of* $\mathcal{A}$ *are strictly smaller than one, that is,* $\alpha_{k1} + \alpha_{k2} < 1$ *for* $k = 1, 2$. *Consequently, all eigenvalues of* $\mathcal{A}$ *are also strictly smaller than one and the stationarity condition* (4.12) *is fulfilled.*

2. *If* $\beta_{22} = 0$ *and* $\beta_{11} > 0$, *only the innovation term* $e_{t,1}$ *may enter the model in* $X_{t,1}$. *To check the stationarity condition, we have to distinguish two cases:*

   a) *If* $\alpha_{21} = 0$, *we have* $\alpha_{22} = 1$ *such that* $X_{t,2}$ *only depends on its first lag* $X_{t-1,2}$. *Then, the parameter matrix* $\mathcal{A}$ *has eigenvalues* α_{11} *and* $\alpha_{22} = 1$, *such that the stationarity condition* (4.12) *is not fulfilled in this case.*

   b) *If* $\alpha_{21} > 0$, *after one extra time step the innovation* $e_{t,1}$ *may enter also the second dimension. in this case, it can be calculated, that all eigenvalues of* $\mathcal{A}$ *are strictly smaller than one such that the stationarity condition* (4.12) *holds.*

3. *If* $\beta_{kk} = 0$ *for* $k = 1, 2$, *the innovation process does not enter the time series process at all. The parameter matrix* $\mathcal{A}$ *have both row sums equal to one, such that the largest eigenvalue is equal to one and the stationarity condition* (4.12) *is not fulfilled.*

4.3.2. Yule-Walker(-type) equations and conditional probabilities

ND(V)ARMA models resemble the model recursion of *linear* (V)ARMA models, but they are *non-linear* due to the random coefficient matrices. Furthermore, although autocorrelations are generally not a meaningful measure for dependence in categorical data. However, they are very helpful for deriving Yule - Walker equations. As in classical VAR models, these can be used for parameter estimation also in NDVAR models. For the NDVARMA process, Yule - Walker - type equations can be derived as well.

The following theorem summarizes the Yule-Walker equations for the purely autoregressive case of an NDVAR(p) model and the Yule - Walker - type equations for an NDVARMA(p,q) model; see also e.g. Weiß and Göb (2008) and Weiß (2009a) for the univariate case.

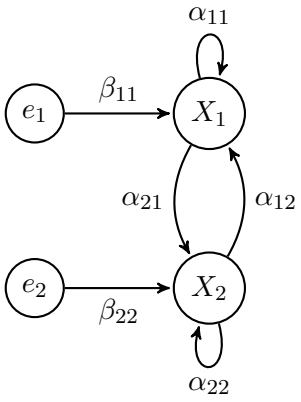


Figure 4.2.: Di - Graph of a two dimensional NDVAR(1) time series with corresponding selection probabilities.

Theorem 4.6 (Yule-Walker (-type) Equations)
Let $(X_t, t \in \mathbb{Z})$ be a stationary K-dimensional NDVARMA(p,q) process and denote by $\Gamma_X(h) = Cov(X_t, X_{t-h})$, $h \in \mathbb{N}$, the corresponding autocovariance matrices.

(i) If $p = 1$ and $q = 0$, $(X_t, t \in \mathbb{Z})$ is an NDVAR(1) process and for all $h \in \mathbb{N}$ (with $h \neq 0$), we have

$$\Gamma_X(h) = \sum_{i=1}^{p} \mathcal{A}^{(i)} \Gamma_X(h-i). \tag{4.14}$$

leading, in particular, to the system of Yule-Walker equations

$$\left[\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(p)}\right] \begin{pmatrix} \Gamma_X(i-j) \\ i,j = 1,\ldots,p \end{pmatrix} = \left[\Gamma_X(1), \ldots, \Gamma_X(p)\right]. \tag{4.15}$$

(ii) If $p \in \mathbb{N}$ and $q \in \mathbb{N}_0$ such that $(X_t, t \in \mathbb{Z})$ is a stationary NDVARMA(p,q) process. Further, set $\mathcal{B}^{(m)} := 0_{K\times K}$ for $m > q$ and define a sequence of coefficient matrices $(\Phi_m, m \in \mathbb{Z})$ recursively by

$$\Phi_m := 0_{K\times K} \quad \text{for } m < 0, \quad \Phi_0 := \Sigma_E \mathcal{B}^{(0)\prime},$$
$$\Phi_m := \sum_{i=1}^{p} \Phi_{m-i} \mathcal{A}^{(i)\prime} + \Sigma_E \mathcal{B}^{(m)\prime} \quad \text{for } m > 0.$$

Then, for all $h \in \mathbb{N}$ (with $h \neq 0$), the autocovariance function fulfills

$$\Gamma_X(h) - \sum_{i=1}^{p} \mathcal{A}^{(i)} \Gamma_X(h-i) = \sum_{j=h}^{q} \mathcal{B}^{(j)} \Phi_{j-h}. \tag{4.16}$$

Formulas (4.14) and (4.16) in Theorem 4.6 link the parameter matrices $\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(p)}$ for the NDVAR case and $\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(p)}, \mathcal{B}^{(0)}, \ldots, \mathcal{B}^{(q)}$ for the NDVARMA(p,q) case to the autocovariances $\{\Gamma_X(h),\ h \in \mathbb{Z}\}$ of the process. A similar link that includes also the process mean $\mu_X = E(X_t)$ as well as the innovation mean $\mu_e = E(e_t)$ and

covariance $\Sigma_e = Cov(e_t)$ based on the Yule-Walker equation for $h = 0$ can also be achieved as in Jentsch and Reichmann (2020)[Theorem 2.7]. However, as Σ_e is not of particular interest for categorical data, we shall investigate instead the (conditional) distributions of NDVARMA processes next.

In the following lemma, for the NDVARMA processes, we derive two types of conditional probabilities including the one step ahead transition probability, which contain the results obtained in Weiß (2009a), Weiß (2011b) and Weiß (2020) for univariate NDARMA processes as a special case.

Lemma 4.7 (Conditional probabilities of NDVARMA processes)
Suppose $(X_t, t \in \mathbb{Z})$ is a stationary NDVARMA(p,q) process of dimension K and let s_i, r_j, $i = 0, \ldots, p$ and $j = 0, \ldots, q$ denote K-dimensional vectors with values in $\mathcal{V}^K$.

(i) The conditional probability given the past values of the NDVARMA process as well as the past and present of innovation process is given by

$$P\left(X_t = s_0 | X_{t-1} = s_1, \ldots, X_{t-p} = s_p, e_t = r_0 \ldots, e_{t-q} = r_q\right) = \tag{4.17}$$
$$\prod_{k=1}^{K}\left[\sum_{i=1}^{p}\sum_{l=1}^{K}\alpha_{kl}^{(i)}\delta_{s_{0,k}s_{i,l}} + \beta_{kk}^{(0)}\delta_{s_{0,k}r_{0,k}} + \sum_{l=1}^{K}\sum_{j=1}^{q}\beta_{kl}^{(j)}\delta_{s_{0,k}r_{j,l}}\right],$$

where $\delta_{ij} = \mathbb{1}_{\{i=j\}}$ denotes the Kronecker delta.

(ii) The one step ahead transition probability is given by

$$P\left(X_t = s_0 | X_{t-1} = s_1, \ldots, X_{t-p} = s_p\right) = \tag{4.18}$$
$$\sum_{r_0, r_1, \ldots, r_q \in \mathcal{V}^K}\prod_{m=0}^{q} p_e(r_m)\prod_{k=1}^{K}\left[\sum_{i=1}^{p}\sum_{l=1}^{K}\alpha_{kl}^{(i)}\delta_{s_{0,k}s_{i,l}} + \beta_{kk}^{(0)}\delta_{s_{0,k}r_{0,k}} + \sum_{l=1}^{K}\sum_{j=1}^{q}\beta_{kl}^{(j)}\delta_{s_{0,k}r_{j,l}}\right],$$

where $p_e(r_m) := P(e_t = r_m)$.

Example 4.8 (Cont. of Example 4.2)
For the NDVAR(1) process as specified in Example 4.2, for $s_0, s_1 \in \mathcal{V}^K$, the transitions probabilities derived in (4.18) *become*

$$P(X_t = s_0 | X_{t-1} = s_1) = \sum_{r_0 \in \mathcal{V}^K} P\left(e_{t,1} = r_{0,1}, e_{t,2} = r_{0,2}\right)$$
$$\cdot \left[0.57\delta_{s_{0,1}s_{1,1}} + 0.27\delta_{s_{0,1}s_{1,2}} + 0.16\delta_{s_{0,1}r_{0,1}}\right] \cdot \left[0.45\delta_{s_{0,2}s_{1,1}} + 0.20\delta_{s_{0,2}s_{1,2}} + 0.06\delta_{s_{0,2}r_{0,2}}\right].$$

4.3.3. Identification issues of NDVARMA processes

In comparison to univariate NDARMA processes, its multivariate extension may suffer from identification issues. As discussed already in Jentsch and Reichmann (2020, Remark 2.13) for the special case of $m = 1$ leading to (multivariate) binary data, non - zero off diagonal elements in $\mathcal{B}^{(0)}$ would lead to identification issues. Hence, we imposed a diagonal structure of $\mathcal{B}^{(0)}$ in Definition 4.1.

Furthermore, we have to discuss the crucial property of non - reducibility of the state space of an NDVARMA process in the following.

We call the state space $\mathcal{V}^K$ of an NDVARMA process *non-reducible* if it takes every state $s_0 \in \mathcal{V}^K$ with strictly positive probability, i.e. $P(X_t = s_0) > 0$ for all $s_0 \in \mathcal{V}^K$.

However, this property might be not true in practice. For gbVAR(p) processes, Jentsch and Reichmann (2020)[Example 2.12] discussed this issue for a binary state space, i.e. $m = 1$. The next example illustrates this problem for general $m \in \mathbb{N}$ and state space $\mathcal{V} = \{0, 1, \ldots, m\}$.

Example 4.9 (NDVAR(1) model with reducible state space)
Consider a $K = 3$ dimensional NDVAR(1) process with parameter matrices

$$\mathcal{A}^{(1)} = \begin{pmatrix} 0 & 1 & 0 \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ 0 & 1 & 0 \end{pmatrix} \quad \textit{and} \quad \mathcal{B} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \beta_{22} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

such that $\alpha_{21} + \alpha_{22} + \alpha_{23} + \beta_{22} = 1$. The stationarity condition (4.12) holds whenever $\beta_{22} > 0$. By construction, the first and third dimensions of X_t, i.e. $X_{t,1}$ and $X_{t,3}$ always take the value $X_{t-1,2}$ for all $t \in \mathbb{Z}$. That is, the process fulfills $X_{t,1} = X_{t,3}$ for all $t \in \mathbb{Z}$, such that the process just takes states in the set $\{(m_1, m_2, m_1)$ with $m_1, m_2 \in \mathcal{V}\} \subset \mathcal{V}^K$. Hence, $(X_t, t \in \mathbb{Z})$ has a reducible state space which is completely determined by a process $(Y_t, t \in \mathbb{Z})$ of dimension $\widetilde{K} = 2 < 3 = K$ with non-reducible state space $\mathcal{V}^2$. Note that the same phenomenon occurs for an NDVMA(1) process with parameter matrices $\mathcal{B}^{(0)} := \mathcal{B}$ and $\mathcal{B}^{(1)} := \mathcal{A}$ from above.

Note that a reducible state space of the form illustrated in the above example implies perfect correlations between certain components $X_{t,i}$ and $X_{t,j}$, say, of X_t leading to a singular $\Gamma_X(0)$. Hence, even if $Var(X_{t,k}) > 0$ for all $k = 1, \ldots, K$, this leads to identification problems, because the Yule-Walker matrix $(\Gamma_X(i-j), i, j = 1, \ldots, p)$ in (4.15) does not have full rank and is therefore non-invertible. This leads in particular to issues, when Yule-Walker equations are employed for parameter estimation; see also Section 4.3.6.

To solve this practical issue, following the approach of Jentsch and Reichmann (2020) proposed for gbVAR processes having binary state spaces, we aim to find a lower-dimensional process $(Y_t, t \in \mathbb{Z})$ of dimension $\widetilde{K}$, with $\widetilde{K} < K$ completely determined by the process $(X_t, t \in \mathbb{Z})$. First, without loss of generality, the dimensions of an NDVARMA process $(X_t, t \in \mathbb{Z})$ can be rearranged to have an order like

$$X_t = \begin{pmatrix} X_t^{dep} \\ Y_t \end{pmatrix}, \tag{4.19}$$

where X_t^{dep} is of dimension $K - \widetilde{K}$ and contains all sub-processes which perfectly correlate with only one (!) component of Y_t such that Y_t is of dimension $\widetilde{K}$, where $\widetilde{K}$ is the maximal dimension leading to a non-singular $\widetilde{K} \times \widetilde{K}$ matrix $\Gamma_Y(0)$. Then, the process $(Y_t, t \in \mathbb{Z})$ has a non-reducible state space and we can write

$$X_t = HY_t, \tag{4.20}$$

where H is a $K \times \widetilde{K}$ matrix defined by

$$H := \begin{pmatrix} H^{dep} \\ I_{\widetilde{K}} \end{pmatrix} \quad \text{and} \quad H^{dep} := \begin{pmatrix} \mathbb{1}_{\{Cov(X_{t,k}^{dep}, Y_{t,l}) = 1\}} \\ k = 1, \ldots, K - \widetilde{K}, \\ l = 1 \ldots, \widetilde{K} \end{pmatrix}. \tag{4.21}$$

4.3.4. Identification of the marginal innovation distribution

For univariate NDARMA processes, Weiß (2009a) showed that the marginal distribution of the innovation process is equal to the marginal distribution of the process itself, e.g. $P(X_t = s) = P(e_t = s)$ for all $s \in \mathcal{V}$. However, for multivariate NDVARMA models, the marginal distributions of X_t and e_t will generally differ.

To relate the marginal distribution of the *unobservable* innovation process $(e_t, t \in \mathbb{Z})$ to the observable (marginal and joint) distributions of the NDVARMA process $(X_t, t \in \mathbb{Z})$, we derive a general relationship between sub-processes of $(e_t, t \in \mathbb{Z})$ and $(X_t, t \in \mathbb{Z})$ of arbitrary dimension $g \in \{1, \ldots, K\}$. For this purpose, we define

$$W_{t-1} := (X'_{t-1}, \ldots, X'_{t-p}, e'_t, \ldots, e'_{t-q})' \tag{4.22}$$

and denote by $z \in \{0,1\}^{K(p+q+1)}$ with $\|z\|_1 = \sum_{l=1}^{K(p+q+1)} |z_l| = 1$ a unit vector of length $K(p+q+1)$ as a possible realization of $P_{t,k\bullet}$ from (4.5).

Lemma 4.10 (Marginal distribution of an NDVARMA(p,q) process)
Let $(X_t, t \in \mathbb{Z})$ be a K - dimensional stationary NDVARMA(p,q) process with state space $\mathcal{V}^K = \{0, 1, \ldots, m\}^K$ for $m \in \mathbb{N}$ with K-dimensional innovation process $(e_t, t \in \mathbb{Z})$. Further, for $g \in \{1, \ldots, K\}$, let $k_1, \ldots, k_g \in \{1, \ldots, K\}$ with $k_i \neq k_j$ for $i \neq j$ and $s_1, \ldots, s_g \in \mathcal{V} = \{0, 1, 2 \ldots, m\}$. Then, we have

$$\begin{aligned} &P(X_{t,k_1} = s_1, \ldots, X_{t,k_g} = s_g) \\ &= \sum_{\substack{z_1,\ldots,z_g \in \{0,1\}^{K(p+q+1)}, \\ \|z_1\|_1 = \cdots = \|z_g\|_1 = 1}} P(z_1 W_{t-1} = s_1, \ldots, z_g W_{t-1} = s_g) \left(\prod_{r=1}^{g} P(P_{t,k_r\bullet} = z_r) \right). \end{aligned} \tag{4.23}$$

To make further use of Lemma 4.10, note that the first Kp entries of the vector W_{t-1} defined in (4.22) consists of lagged entries $(X'_{t-1}, \ldots, X'_{t-p})'$ of the process and the remaining $K(q+1)$ entries consist of lagged entries $(e'_t, \ldots, e'_{t-q})'$ of the innovations. Correspondingly, we can split each sum over z_r in (4.23) in two parts to get

$$\sum_{\substack{z_r \in \{0,1\}^{K(p+q+1)}, \\ \|z_r\|_1 = 1}} \cdots = \sum_{\substack{z_r \in \{0,1\}^{K(p+q+1)}, \\ \sum_{j=1}^{Kp} z_{r,j} = 1}} \cdots + \sum_{\substack{z_r \in \{0,1\}^{K(p+q+1)}, \\ \sum_{j=Kp+1}^{K(p+q+1)} z_{r,j} = 1}} \cdots \tag{4.24}$$

leading to 2^g many combinations of partial sums related to lagged process and lagged innovations, respectively. More precisely, the marginal distributions of $e_{t,k}$, $k = 1, \ldots, K$ can be identified using $g = 1$ as follows. In this case, Lemma 4.10 together with (4.24) leads to

$$\begin{aligned} P(X_{t,k_1} = s_1) = &\sum_{\substack{z_1 \in \{0,1\}^{K(p+q+1)}, \\ \sum_{j=1}^{Kp} z_{r,j} = 1}} P(z_1 W_{t-1} = s_1) P(P_{t,k_1\bullet} = z_1) \\ &+ \sum_{\substack{z_1 \in \{0,1\}^{K(p+q+1)}, \\ \sum_{j=Kp+1}^{K(p+q+1)} z_{r,j} = 1}} P(z_1 W_{t-1} = s_1) P(P_{t,k_1\bullet} = z_1) \\ = &\sum_{i=1}^{p} \sum_{l=1}^{K} \alpha_{k_1 l}^{(i)} P(X_{t-i,l} = s_1) + \sum_{j=0}^{q} \sum_{l=1}^{K} \beta_{k_1 l}^{(j)} P(e_{t-j,l} = s_1). \end{aligned} \tag{4.25}$$

The latter holds for all $k_1 \in \{1,\dots,K\}$ and all $s_1 \in \{0,1,\dots,m\}$ and making use of stationarity of $(X_t, t \in \mathbb{Z})$ and $(e_t, t \in \mathbb{Z})$, by defining $\mathbf{p}_e(s) := (P(e_{t,1} = s), \dots, P(e_{t,K} = s))'$ and $\mathbf{p}_X(s) := (P(X_{t,1} = s), \dots, P(X_{t,K} = s))'$ for $s \in \mathcal{V}$, we get the following relationship.

Lemma 4.11 (Marginal distributions of innovation components $e_{t,k}$)
Let $(X_t, t \in \mathbb{Z})$ be a K - dimensional stationary NDVARMA(p,q) process with state space $\mathcal{V}^K = \{0,1,\dots,m\}^K$ for $m \in \mathbb{N}$ with K-dimensional innovation process $(e_t, t \in \mathbb{Z})$. Then, we have

$$\left(I_K - \sum_{i=1}^{p} \mathcal{A}^{(i)}\right) \mathbf{p}_X(s) = \left(\sum_{j=0}^{q} \mathcal{B}^{(j)}\right) \mathbf{p}_e(s)$$

for any $s \in \mathcal{V}$. If $\sum_{j=1}^{p} \mathcal{B}^{(j)}$ is invertible, this allows to identify all marginal *distributions of $e_{t,k}$, $k = 1,\dots,K$ by writing*

$$\mathbf{p}_e(s) = \left(\sum_{j=0}^{q} \mathcal{B}^{(j)}\right)^{-1} \left(I_K - \sum_{i=1}^{p} \mathcal{A}^{(i)}\right) \mathbf{p}_X(s).$$

Note that for NDVAR(p) processes, the invertibility condition imposed on $\sum_{j=0}^{q} \mathcal{B}^{(j)}$ is fulfilled if all diagonal entries of $\mathcal{B}^{(0)}$ are strictly positive. In contrast, if some diagonal entries of $\mathcal{B}^{(0)}$ are zero, the corresponding innovation distributions are not identifiable as these innovations do not enter the NDVAR(p) process at all, which is then driven by some lower-dimensional innovation process.

Next, we illustrate how Lemma 4.10 can be employed to identify also the joint marginal innovation distribution of $e_t = (e_{t,1},\dots,e_{t,K})'$ beyond the marginal distributions of $e_{t,k}$, $k = 1,\dots,K$. By using Lemma 4.10 with $g = 2$, this leads to

$$\begin{aligned} &P(X_{t,k_1} = s_1, X_{t,k_2} = s_2) \\ &= \sum_{\substack{z_1,z_2 \in \{0,1\}^{K(p+q+1)}, \\ \|z_1\|_1 = \|z_2\|_1 = 1}} P(z_1 W_{t-1} = s_1, z_2 W_{t-1} = s_2) \left(\prod_{r=1}^{2} P(P_{t,k_r \bullet} = z_r)\right), \end{aligned} \tag{4.26}$$

which allows to identify all bivariate distributions of $(e_{t,i}, e_{t,j})$ for $i,j = 1,\dots,K$, $i \neq j$. For notational convenience, the following example illustrates this for the special case of a stationary NDVAR(1) process.

Example 4.12 (Marginal distribution of the innovation vector)
Let $(X_t, t \in \mathbb{Z})$ be a stationary K-dimensional NDVAR(1) process with non-singular $\mathcal{B}^{(0)}$. Then, using (4.26)*, decomposing both sums according to* (4.24) *and due to Theorem 4.3, we get*

$$\begin{aligned} &P(e_{t,k_1} = s_1, e_{t,k_2} = s_2) \\ &= \frac{1}{\beta_{k_1k_1}\beta_{k_2k_2}} \Bigg[p_{X,k_1k_2}(s_1,s_2) - \sum_{l_1,l_2=1}^{K} \alpha_{k_1l_1}\alpha_{k_2l_2} p_{X,l_1l_2}(s_1,s_2) \\ &\quad - \sum_{l_1=1}^{K} \alpha_{k_1l_1}\beta_{k_2k_2} p_{X,l_1}(s_1) p_{e,k_2}(s_2) - \sum_{l_2=1}^{K} \beta_{k_1k_1}\alpha_{k_2l_2} p_{e,k_1}(s_1) p_{X,l_2}(s_2) \Bigg], \end{aligned}$$

for all $k_1, k_2 \in \{1, \ldots, K\}$, $k_1 \neq k_2$, *where* $p_{X,lk}(r,s) = P(X_{t,l} = r, X_{t,k} = s)$, $p_{X,l}(s_1) = P(X_{t,l} = s_1)$ *and* $p_{e,k}(s) = P(e_{t,k} = s)$. *Note that* $p_{X,ll}(s_1, s_2) = p_{X,l}(s_1)\delta_{s_1,s_2}$ *for all* $l \in \{1, \ldots, K\}$ *and* $s_1, s_2 \in \{0, 1\}$. *As the process* $(X_t, t \in \mathbb{Z})$ *is observed and the marginal distributions of* $e_{t,k}$, $k = 1, \ldots, K$ *can be identified using Lemma 4.11, the bivariate distribution of* (e_{t,k_1}, e_{t,k_2}) *is identified.*

Analogous to the identification of bivariate innovation distributions indicated in Example 4.12, similar formulas can be derived to identify also trivariate innovation distributions etc.

4.3.5. Mixing properties of NDVARMA(p,q) processes

In this section, we aim to derive mixing properties of NDVARMA processes. Mixing concepts are in particular helpful to quantify the dependence structure of serially dependent processes. Overviews of different concepts are given e.g. in the monographs Doukhan (1994), Billingsley (1968) and Bradley (2007). For this purpose, we follow the argumentation of Jacobs and Lewis (1983), Weiß (2009a) and Weiß (2013) to prove mixing properties of NDARMA models. For the multivariate case, according to Jentsch and Reichmann (2020), who considered vector-valued gbVAR(p) models, we define a $K(p+q)$ dimensional process $(Z_t, t \in \mathbb{Z})$ with

$$Z_t := (X_t', X_{t-1}', \ldots, X_{t-p+1}', e_t', \ldots, e_{t-q+1}')', \tag{4.27}$$

which has a suitable Markov chain representation useful to derive φ- and ψ-mixing of K-dimensional NDVARMA(p,q) processes in the following. The process $(Z_t, t \in \mathbb{Z})$ is called ψ - mixing, if

$$|P(\mathcal{E}_1 \cap \mathcal{E}_2) - P(\mathcal{E}_1)P(\mathcal{E}_2)| \leq f_h P(\mathcal{E}_1)P(\mathcal{E}_2) \tag{4.28}$$

for all subsets $\mathcal{E}_1 \in \sigma(Z_t, Z_{t-1}, \ldots)$ and $\mathcal{E}_2 \in \sigma(Z_{t+h}, Z_{t+h+1}, \ldots)$ with a non-negative sequence $(f_h, h \in \mathbb{N})$ such that $f_h \to 0$ for $h \to \infty$. If the right-hand side of (4.28) is replaced by $f_h P(\mathcal{E}_1)$, the process is called φ - mixing.

To achieve the (homogeneous) Markov property (of order one) of the process $(Z_t, t \in \mathbb{Z})$, we have to establish its transition probability. This is obtained as a direct consequence of the transition probabilities derived in Lemma 4.7 for the NDVARMA(p,q) processes. Let $\mathbf{s}_0, \mathbf{s}_1 \in \mathcal{V}^{K(p+q+1)}$ with $\mathbf{s}_0 := (s_0', \ldots, s_{p-1}', r_0', \ldots, r_{q-1}')'$ and $\mathbf{s}_1 := (\tilde{s}_1', \ldots, \tilde{s}_p', \tilde{r}_1', \ldots, \tilde{r}_q')'$, where $s_i, \tilde{s}_i, r_j, \tilde{r}_j \in \mathcal{V}^K$

for $i = 0, \ldots, p$ and $j = 0, \ldots, q$ with $s_i = (s_{i,1}, \ldots, s_{i,K})'$ etc. Then, with $p_e(r_0) = P(e_t = r_0)$, we get

$$P\left(Z_t = \mathbf{s}_0 | Z_{t-1} = \mathbf{s}_1\right) = \tag{4.29}$$

$$\prod_{k=1}^{K}\left[\sum_{i=1}^{p}\sum_{l=1}^{K}\alpha_{kl}^{(i)}\delta_{s_{0,k}\tilde{s}_{i,l}} + \beta_{kk}^{(0)}\delta_{s_{0,k}r_{0,k}} + \sum_{j=1}^{q}\sum_{l=1}^{K}\beta_{kl}^{(j)}\delta_{s_{0,k}\tilde{r}_{j,l}}\right]\left(\prod_{v=1}^{p-1}\delta_{\tilde{s}_v s_v}\right)\left(\prod_{w=1}^{q-1}\delta_{\tilde{r}_w r_w}\right)p_e(r_0).$$

Therefore, the transition probability of $(Z_t, t \in \mathbb{Z})$ is determined by the transition probability of the underlying NDVARMA process and the marginal distribution of e_t. As the transition probability is time invariant, the process $(Z_t, t \in \mathbb{Z})$ is a homogeneous Markov chain of order one.

If the underlying NDVARMA process $(X_t, t \in \mathbb{Z})$ has a non-reducible state space, such that we have $P(X_t = s) > 0$ for all $s \in \mathcal{V}^K$, we can prove *primitivity* of the Markov chain $(Z_t, t \in \mathbb{Z})$.

Lemma 4.13 (Primitivity)
Let $(Z_t, t \in \mathbb{Z})$ be the Markov Chain representation of a stationary NDVARMA(p,q) process $(X_t, t \in \mathbb{Z})$ with non-reducible state space and $p+q \geq 1$. Then $(Z_t, t \in \mathbb{Z})$ is primitive, that is, for $n := \max\{p+q, K'\}+1$, we have

$$p_{\mathbf{s}|\mathbf{r}}(n) := P\left(Z_t = \mathbf{s} | Z_{t-n} = \mathbf{r}\right) > 0 \quad \textit{for all} \quad \mathbf{r}, \mathbf{s} \in \mathcal{V}^{Kp}.$$

With the primitivity of stationary NDVARMA processes with non - reducible state space and following the argumentation of Weiß (2009a)[Section 11.2], it follows that the Markov chain $(Z_t, t \in \mathbb{Z})$ with an underlying NDVARMA process is ergodic and geometrically ψ - and φ - mixing.

This allows then the conclusion, as NDVARMA processes $(X_t, t \in \mathbb{Z})$ are contained in $(Z_t, t \in \mathbb{Z})$, that stationary NDVARMA processes with non-reducible state space itself are geometrically ψ- and φ- mixing. If nevertheless, $(X_t, t \in \mathbb{Z})$ has a reducible state space, such that $\mathcal{V}_X \subsetneq \mathcal{V}^K$, we consider the reduced version $(Y_t, t \in \mathbb{Z})$ from Section 4.3.3 and derive its mixing property for this process. Due to the linear transformation of $(Y_t, t \in \mathbb{Z})$ to $(X_t, t \in \mathbb{Z})$, the property of mixing can easily be concluded also for the original process $(X_t, t \in \mathbb{Z})$.

Theorem 4.14 (Mixing of the NDVARMA process)
Let $(X_t, t \in \mathbb{Z})$ be a K - dimensional stationary NDVARMA(p,q) process with $p+q \geq 1$ and $Var(X_{t,k}) > 0$ for $k = 1, \ldots, K$. Then the process is ψ - and φ - mixing with exponentially decreasing weights $(f_h, h \in \mathbb{N})$ i.e. there exists an $a > 0$ and $\rho \in (0,1)$ such that $f_h = a\rho^h$.

4.3.6. Parameter estimation

The whole distribution of NDVARMA models is completely determined by the model parameters in $\mathcal{P}$ and by the marginal distribution of the i.i.d. innovations $(e_t, t \in \mathbb{Z})$. The Yule-Walker equations derived in Theorem 4.6 are particularly useful to construct Yule-Walker estimators $\widehat{\mathcal{A}}^{(1)}, \ldots, \widehat{\mathcal{A}}^{(p)}$ for the model parameters $\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(p)}$ of an NDVAR(p) model. Naturally, with data $X_1, \ldots, X_n$ at hand, these are obtained by replacing the population autocovariances $\Gamma_X(h)$ by their sample version

$$\widehat{\Gamma}_X(h) := \begin{cases} \frac{1}{n}\sum_{t=1}^{n-h}(X_{t+h} - \overline{X})(X_t - \overline{X})', & 0 \leq h < n \\ 0, & h \geq n \end{cases} \tag{4.30}$$

with $\Gamma_X(-h) := \Gamma_X'(h)$ for $h < 0$. If the process has a non-reducible state in the sense of Section 4.3.3, the Yule-Walker matrix is non-singular and we get

$$[\widehat{\mathcal{A}}^{(1)}, \ldots, \widehat{\mathcal{A}}^{(p)}] = [\widehat{\Gamma}_X(1), \ldots, \widehat{\Gamma}_X(p)] \begin{pmatrix} \widehat{\Gamma}_X(i-j) \\ i,j = 1, \ldots, p \end{pmatrix}^{-1}. \tag{4.31}$$

with $\widehat{\mathcal{A}}^{(i)} = (\widehat{\alpha}_{kl}^{(i)})_{k,l=1,\ldots,K}$. Otherwise, i.e. if the state space is reducible, Yule-Walker estimation has to be applied to the corresponding lower-dimensional process $(Y_t, t \in \mathbb{Z})$ with non-reducible state space $\mathcal{V}_Y$ as introduced in Section 4.3.3. Diagonality of $\mathcal{B}^{(0)}$ together with the natural restriction $\beta_{kk} = 1 - \sum_{i=1}^{p}\sum_{l=1}^{K}|\alpha_{kl}^{(i)}|$ for $k = 1, \ldots, K$ leads to the estimator $\widehat{\mathcal{B}}$ defined by

$$\widehat{\mathcal{B}} := I_K - diag\left([\widehat{\mathcal{A}}^{(1)}, \ldots, \widehat{\mathcal{A}}^{(p)}]\mathbb{1}_{Kp}\right). \tag{4.32}$$

For NDVAR(p) models, the conditional probabilities corresponding (4.17) and (4.18) as derived in Lemma 4.7 can be estimated in a similar fashion by replacing all population quantities by the corresponding estimators. To estimate the transition probabilities in (4.18), this also requires the estimation of the innovation distribution to get $p_e(r) = P(e_t = r)$ for all $r \in \mathcal{V}^K$. If $\mathcal{B}^{(0)}$ is invertible, the marginal distributions of $e_{t,k}$, $k = 1, \ldots, K$ can be estimated using the formulas derived in Lemma 4.11 by plugging-in the estimators from. This leads to

$$\widehat{p}_{e,k}(s) = \frac{1}{\widehat{\beta}^{(0)}_{kk}} \left(\widehat{p}_{X,k}(s) - \sum_{i=1}^{p} \sum_{l=1}^{K} \widehat{\alpha}^{(i)}_{kl} \widehat{p}_{X,l}(s) \right), \tag{4.33}$$

where $\widehat{p}_{X,k}(s) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{\{X_{t,k}=s\}}$ for $k \in \{1, \ldots, K\}$ and $s \in \mathcal{V}$. This fully identifies the innovation distribution under the assumption of independence of $e_{t,k}$, $k = 1, \ldots, K$. Nevertheless, similar formulas as sketched in Example 4.12 for bivariate distributions of $(e_{t,i}, e_{t,j})$, $i \neq j$, can be used to estimate also the innovation distribution beyond the marginal distributions of $e_{t,k}$, $k = 1, \ldots, K$. Note that the latter formulas rely on $\widehat{\beta}_{kk} > 0$.

However, in practice, the Yule-Walker estimators might lead to invalid model parameters with

$$\sum_{i=1}^{p} \sum_{l=1}^{K} \widehat{\alpha}^{(i)}_{kl} > 1 \tag{4.34}$$

for some $k \in \{1, \ldots, K\}$. In such cases, we propose to use an approach similar to Jentsch and Reichmann (2020, Remark 2.15) based on linearly constraint estimation to assure $\sum_{i=1}^{p} \sum_{l=1}^{K} \widehat{\alpha}^{(i)}_{kl} = 1$ and setting $\widehat{\beta}_{kk} = 0$. A similar constraint estimation approach is also applicable if certain entries are estimated negative, i.e. $\widehat{\alpha}^{(i)}_{kl} < 0$ for some $i \in \{1, \ldots, p\}$ and $k, l \in \{1, \ldots, K\}$, where we restrict these entries to zero. Note that $\widehat{\mathcal{B}}^{(0)}$ is no longer invertible, if some diagonal elements $\widehat{\beta}_{kk}$ are equal to zero and $(e_t, t \in \mathbb{Z})$ is not identified for those rows. However, the fitted NDVAR(p) model might still satisfy the stationarity condition (4.12), as long as another innovation term enters $X_{t,k}$ with strictly positive probability later; see Example 4.5.

4.4. Simulation Study

In this section, we investigate the performance of Yule-Walker-based estimators in NDVAR models by Monte-Carlo simulations. To illustrate the estimation performance in several NDVAR model setups, we consider the (average) mean squared error (MSE) of different parameter estimators.

4.4.1. The DGPs

To be able to judge the performance of parameter estimation in several NDVAR model specifications we examine three different NDVAR(p) setups with orders $p = 1, 2$, dimensions $K = 3, 4$ and $m = 3, 4$, and therefore $m + 1$ categories for sample sizes $n = 100, 500, 1000$. In all cases, we use innovation processes $(e_t, t \in \mathbb{Z})$ consisting of K independent components. We consider three data generating processes (DGPs):

v	0	1	2	3	4
$P(e_{t,1} = v)$	0.10	0.30	0.15	0.15	0.30
$P(e_{t,2} = v)$	0.30	0.20	0.40	0.05	0.05
$P(e_{t,3} = v)$	0.10	0.10	0.30	0.40	0.10
$P(e_{t,4} = v)$	0.20	0.10	0.10	0.20	0.40

Table 4.1.: Innovation distribution used for (DGP2) and (DGP3), where the $e_{t,k}$, $k = 1, \ldots, K$ are stochastically independent.

(DGP1) NDVAR(1) with $K = 3$, $m = 3$ and parameter matrices

$$\mathcal{A}^{(1)} = \begin{pmatrix} 0.15 & 0.25 & 0.49 \\ 0.19 & 0.27 & 0.28 \\ 0.17 & 0.39 & 0.21 \end{pmatrix}, \quad \mathcal{B}^{(0)} = diag\,(0.11, 0.26, 0.23)$$

and $P\,(e_{t,k} = v) = (m+1)^{-1}$ for all $k = 1, 2, 3$ and $v \in \mathcal{V}$.

(DGP2) NDVAR(1) with $K = 4$, $m = 4$ and parameter matrices

$$\mathcal{A}^{(1)} = \begin{pmatrix} 0.15 & 0.25 & 0.19 & 0.26 \\ 0.19 & 0.17 & 0.28 & 0.15 \\ 0.17 & 0.35 & 0.21 & 0.11 \\ 0.13 & 0.22 & 0.36 & 0.11 \end{pmatrix}, \quad \mathcal{B}^{(0)} = diag\,(0.15, 0.21, 0.16, 0.18)\,.$$

with innovation distribution stated in Table 4.1.

(DGP3) NDVAR(2) with $K = 3$, $m = 4$ and parameter matrices

$$\mathcal{A}^{(1)} = \begin{pmatrix} 0.15 & 0.05 & 0.23 \\ 0.09 & 0.17 & 0.28 \\ 0.17 & 0.19 & 0.21 \end{pmatrix}, \quad \mathcal{A}^{(2)} = \begin{pmatrix} 0.11 & 0.15 & 0.19 \\ 0.11 & 0.20 & 0.08 \\ 0.17 & 0.09 & 0.11 \end{pmatrix},$$
$$\text{and} \quad \mathcal{B}^{(0)} = diag\,(0.12, 0.07, 0.06)$$

with innovation distribution corresponding to dimensions $k = 1, 2, 4$ stated in Table 4.1.

4.4.2. Average MSE estimation performance

We measure the estimation performance by calculating the entry-wise MSE of each parameter estimate for 1000 Monte Carlo replications. In Table 4.2, we summarize the average MSE results for several estimators discussed in Section 4.3.6. For increasing sample size, the estimation accuracy improves for all parameter estimators and all DGPs. In comparison, the estimation of the the marginal innovation distribution(s) is least precise and the MSE turns out to be large for smaller sample size and larger dimensions. This phenomenon can be explained as the diagonal entries of $\widehat{\mathcal{B}}^{(0)}$ have to be inverted to estimate the probabilities corresponding to the innovations. Hence, when estimating small diagonal entries of $\mathcal{B}^{(0)}$, already small deviations in $\widehat{\mathcal{B}}^{(0)}$ might lead to large estimation variance leading to a larger MSE.

	n	MSE of $\widehat{\mathcal{A}}^{(1)}$	MSE of $\widehat{\mathcal{A}}^{(2)}$	MSE of $\widehat{\mathcal{B}}^{(0)}$	MSE of $\widehat{P}(e_{t,k} = s)$
DGP1	100	0.0122		0.0360	2.4750
	500	0.0023		0.0021	0.0163
	1000	0.0012		0.0009	0.0043
DGP2	100	0.0133		0.0718	0.2504
	500	0.0026		0.0038	0.0108
	1000	0.0012		0.0012	0.0044
DGP3	100	0.0065	0.0535	0.0247	0.2502
	500	0.0026	0.0190	0.0038	0.0109
	1000	0.0013	0.0100	0.00012	0.0044

Table 4.2.: Average MSE for parameter estimators $\widehat{\mathcal{A}}^{(1)}$, $\widehat{\mathcal{A}}^{(2)}$, $\widehat{\mathcal{B}}^{(0)}$ and $\widehat{P}(e_{t,k} = s)$ for the three different parameter specifications (DGP1), (DGP2) and (DGP3), respectively.

n	$\widehat{p}_{e,1}$	$\widehat{p}_{e,2}$	$\widehat{p}_{e,12}$
100	0.0291	0.0111	0.0333
500	0.0049	0.0020	0.0078
1000	0.0022	0.0010	0.0041

Table 4.3.: Average MSE for the parameter estimators $\widehat{p}_{e,k}(s)$ and $\widehat{p}_{e,kl}(s,t)$ for the DGP described in Section 4.4.3 averaged over the states $s \in \mathcal{V}$ and $(s,t) \in \mathcal{V}^2$.

4.4.3. (Average) MSE estimation of the innovation process covariance matrix

As described in Section 4.3.4, it is possible to identify not only the marginal distributions of the innovation components $e_{t,k}$, but also their joint marginal distribution. For illustration, we consider a bivariate process with three categories, i.e. $\mathcal{V} = \{0,1,2\}^2$ and model parameters $\alpha_{12}, \alpha_{13}, \alpha_{32}$ and α_{33} of (DGP1) with corresponding diagonal matrix $\mathcal{B}^{(0)} = diag(0.26, 0.40)$. The innovations e_t are generated based on the following probabilities:

$$
\begin{aligned}
&P(e_t = (0,0)') = 0.09, \quad && P(e_t = (0,1)') = 0.13, \quad && P(e_t = (0,2)') = 0.06,\\
&P(e_t = (1,0)') = 0.14, \quad && P(e_t = (1,1)') = 0.12, \quad && P(e_t = (1,2)') = 0.07,\\
&P(e_t = (2,0)') = 0.09, \quad && P(e_t = (2,1)') = 0.21, \quad && P(e_t = (2,2)') = 0.09.
\end{aligned}
$$

In Table 4.3 the simulation results based on 1000 Monte-Carlo replications are displayed. The averaged MSE clearly decreases for increasing sample size. Non-surprising, the averaged MSE for the joint marginal distribution of both innovation processes $e_{t,1}$ and $e_{t,2}$ has the biggest MSE, as the estimation make use of the previous estimated marginal distribution of each innovation process.

4.5. Real Data application: the business cycle clock

Macroeconomic data and its analysis are important for e.g. policy makers, analysts and central bankers, to understand the economic situation and adapt their behavior

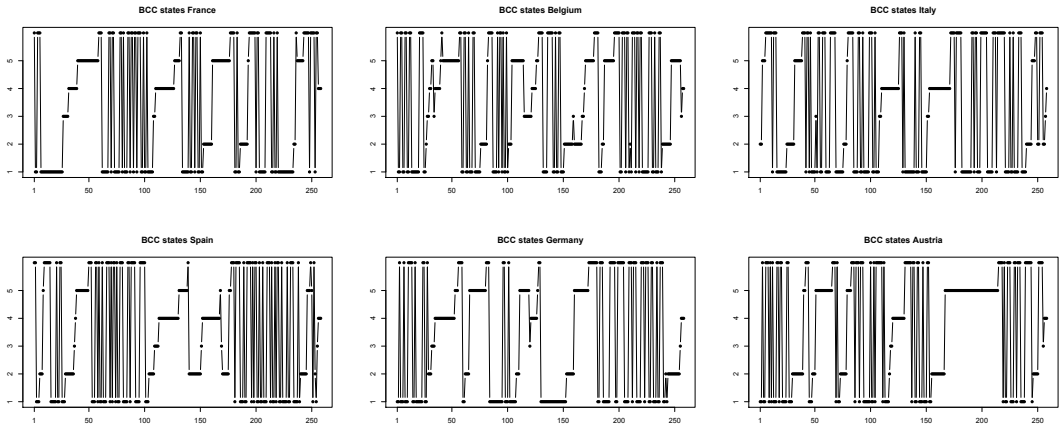


Figure 4.3.: Business Cycle Clock data of six European countries from January 1999 till June 2020

or reaction on the given and expected state, as discussed in Mazzi (2015) and based on Zarnowitz (1992). One measure of the business cycle is the gross domestic product (GDP), where the growth rate of the real domestic product indicates the economic situation of a country. As not only recession states of the countries are of interest, Mazzi (2015) developed a clock-type visualization tool, where the economic situation is sub-classified in six different more detailed states. The business cycle is divided into sections, defined as follows:

State 1: expansion of the economy with decelerating growth

State 2: slowdown

State 3: recession of an economy

State 4: recession with accelerating growth

State 5: recovery

State 6: expansion with accelerating growth

For further details see `https://ec.europa.eu/eurostat/cache/bcc/resources/locales/en/user_manual.pdf`. The expansion phase is characterized by a positive growth but with decreasing growth rate still above the trend. Whereas a slowdown (state 2) is given when the growth rate is decreasing and now below the trend. The next state, state 3, indicates a country to be in a recession period. In state 4, the growth rate increases but is still negative and below the trend, where in state 5 the growth rate is now positive. Finally, in state 6, the growth rate is increasing and above the trend.

For six European countries, the monthly Business Cycle clock (BCC) data are available from January 1999 till June 2020. All time series show an oscillating period between the states 1 and 6, thus, the growth is fluctuating between accelerating "6" and decelerating "1" growth in an expansion state. These periods occur almost for all countries at the same time such that we have cross - sectional dependence in the data. In addition, the time series takes all possible values in state space $\mathcal{V}^K = \{1, 2, 3, 4, 5, 6\}^K$. Therefore we fit a $K = 6$- dimensional NDVAR(p) model, by selecting the order p

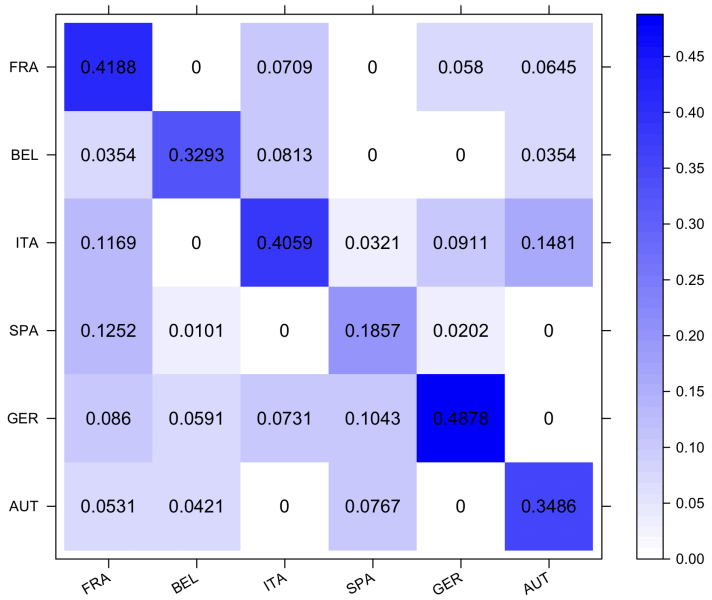


Figure 4.4.: Estimated parameter matrix $\widehat{\mathcal{A}}$ of the Business Cycle clock data of six European countries

by common order selection criteria as the Hannan - Quinn and Bayesian information criteria. Both select an order $p = 1$.

In the estimated parameter matrix $\widehat{\mathcal{A}}$, derived from the Yule - Walker equation (4.14), small negative entries occur and we apply a constraint estimation to get rid of negative parameters to receive a valid parameter estimator.

The estimated parameters $\widehat{\mathcal{A}}$ shown in Figure 4.4 indicate cross-sectional among the six countries with leading diagonal entries of $\widehat{\mathcal{A}}$. For instance, the diagonal entry of Germany, the largest social economy in Europe, takes the overall largest value in the parameter matrix $\widehat{\mathcal{A}}$. The influence of the other countries is thus relatively minor in comparison e.g. to Spain, which has the overall smallest diagonal element. Also, it is not surprising that the coefficient $\widehat{\alpha}_{41}$ is rather large, which indicates that the economy of Spain depends to some extent on the economy of its bordering country France. France is the leading economic partner of Spain with more than 60 billion Euro of annual trade, see e.g. `https://www.diplomatie.gouv.fr/en/country-files/spain/`. In comparison, the leading European economies Germany and France depend with rather large value on its own previous state and less on the economic states of the other countries. The parameters indicating a dependence on other countries are explainable by the amount of trade between these countries. For example, the parameters of Germany corresponding to countries France, Belgium, Italy and Spain, which are the most important European trading partners of Germany, sum up to about 31 percent. Nevertheless, it seems surprising that the economic status of Austria is estimated to have no direct influence on Germany in our analysis.

The parameter matrix $\widehat{\mathcal{B}}$ is estimated as discussed in Section 4.3.6 leading to

$$\widehat{\mathcal{B}} = diag(0.3878, 0.5186, 0.2059, 0.6588, 0.1896, 0.4795).$$

These parameters illustrate the findings of the parameter matrix $\widehat{\mathcal{A}}$ that strong economies tend to depend little on the innovation process, while the smaller countries as well as

Spain are highly influenced by the innovation process with values around 50 %. It might be also plausible that global shocks on the economies, coming from the innovation process, affect smaller countries first than more stable economies, as e.g. Germany and France.

This data application shows that applying the NDVAR model to the BCC time series data is suitable to reveal cross-country as well as serial dependencies of Business Cycle states.

Nevertheless, even if the estimated negative entries are small, it might be interesting to enlarge the NDVARMA process to allow also negative dependence structure. In the case of a binary state space, Jentsch and Reichmann (2019) extended the NDARMA process to negative dependence. In case of a state space with two values negative dependence means to switch to the opposite state, it is an open question what negative dependence in a state space with $m+1$ different values means.

4.6. Conclusion

In the paper, we discussed the multivariate extension of the NDARMA process, introduced by Jacobs and Lewis (1983) and further discussed by Weiß (2009a), to analyze also cross-sectional dependence structures of vector-valued categorical time dependent data.

We derive the stationarity conditions for the NDVARMA model class and state the stationary solution. Afterwards, Yule-Walker-type equations are proven that allow for a direct estimation of NDVAR models using the common Yule-Walker estimators. We derive formulas for the transition probabilities, that allow to prove mixing properties of NDVARMA processes. Potential identification issues are discussed and we show how the innovation distribution can be identified.

The estimation performance of Yule-Walker-based parameter estimators are analyzed in a simulation study. By considering business cycle clock data for six European countries, we illustrate the application of NDVAR processes to analyze the dynamic of economic states across Europe. The application reveals a reasonable amount of cross-sectional dependence that should be taken into account for the cross-country modeling of economic states.

4.7. Proofs of Chapter 4

4.7.1. Proof of Theorem 4.3

The proof works similar to the proof of Theorem 2.4 in Jentsch and Reichmann (2020). By the same argumentation, with all eigenvalues of $\widetilde{\mathcal{A}}$ lying in $[0,1)$ by assumption 4.10 and boundedness of $\widetilde{X}_{t-(d+1)}$, using 4.10, we can show

$$\left\|\widetilde{\zeta}_d\right\|_1 = \mathbb{1}'_{K(p+q)}\widetilde{\mathcal{A}}^{d+1}\mathbb{1}_{K(p+q)} \underset{d\to\infty}{\to} 0.$$

□

4.7.2. Proof of Theorem 4.6

The proof is completely analogous to the proofs of Theorems 2.6 and 2.7 in Jentsch and Reichmann (2020).

□

4.7.3. Proof of Lemma 4.7

The proof of part (ii) follows directly from the proof of Lemma 2.9 in Jentsch and Reichmann (2020) when all entries of $\mathcal{P}$ are non-negative and part (i) follows similarly. □

4.7.4. Proof of Lemma 4.10

By plugging-in component-wise the NDVARMA recursion (4.6), conditioning on the $P_{t,k\bullet}$'s and using the independence of $P_{t,k\bullet}$ and W_{t-1} and the independence of $P_{t,k_1\bullet}$ and $P_{t,k_2\bullet}$ for $k_1 \neq k_2$, we get

$$
\begin{aligned}
&P(X_{t,k_1} = s_1, \ldots, X_{t,k_g} = s_g) \\
&= \sum_{\substack{z_1,\ldots,z_g \in \{0,1\}^{K(p+q+1)}, \\ \|z_1\|_1 = \cdots = \|z_g\|_1 = 1}} P(X_{t,k_1} = s_1, \ldots, X_{t,k_g} = s_g | P_{t,k_1\bullet} = z_1, \ldots, P_{t,k_g\bullet} = z_g) \\
&\qquad \cdot P(P_{t,k_1\bullet} = z_1, \ldots, P_{t,k_g\bullet} = z_g) \\
&= \sum_{\substack{z_1,\ldots,z_g \in \{0,1\}^{K(p+q+1)}, \\ \|z_1\|_1 = \cdots = \|z_g\|_1 = 1}} P(P_{t,k_1\bullet} W_{t-1} = s_1, \ldots, P_{t,k_g\bullet} W_{t-1} = s_g | P_{t,k_1\bullet} = z_1, \ldots, P_{t,k_g\bullet} = z_g) \\
&\qquad \cdot P(P_{t,k_1\bullet} = z_1, \ldots, P_{t,k_g\bullet} = z_g) \\
&= \sum_{\substack{z_1,\ldots,z_g \in \{0,1\}^{K(p+q+1)}, \\ \|z_1\|_1 = \cdots = \|z_g\|_1 = 1}} P(z_1 W_{t-1} = s_1, \ldots, z_g W_{t-1} = s_g) \left(\prod_{r=1}^{g} P(P_{t,k_r\bullet} = z_r) \right).
\end{aligned}
$$

□

4.7.5. Proof of Lemma 4.11

The proof follows directly from Lemma 4.10 for $g = 1$ and equation (4.25). □

4.7.6. Proof of Example 4.12

For the NDVAR(p) case, decomposing both sums in (4.26) according to (4.24) leads to $2^g = 4$ terms:
Case 1: $\sum_{j=1}^{Kp} z_{rj} = 1$ for $r = 1, 2$

$$
\begin{aligned}
&\sum_{\substack{z_1,z_2 \in \{0,1\}^{K(p+1)}, \\ \sum_{j=1}^{Kp} z_{rj} = 1,\ r=1,2}} P(z_1 W_{t-1} = s_1, z_2 W_{t-1} = s_2) \left(\prod_{r=1}^{2} P(P_{t,k_r\bullet} = z_r) \right) \\
&= \sum_{i_1,i_2=1}^{p} \sum_{l_1,l_2=1}^{K} \alpha_{k_1 l_1}^{(i_1)} \alpha_{k_2 l_2}^{(i_2)} P(X_{t-i_1,l_1} = s_1, X_{t-i_2,l_2} = s_2)
\end{aligned}
$$

Case 2: $\sum_{j=1}^{Kp} z_{1j} = 1$ and $\sum_{j=Kp+1}^{K(p+1)} z_{2j} = 1$:

$$\begin{aligned}
&\sum_{\substack{z_1,z_2\in\{0,1\}^{K(p+1)},\\ \sum_{j=1}^{Kp} z_{1j}=1,\sum_{j=Kp}^{K(p+1)} z_{2j}=1}} P(z_1W_{t-1} = s_1, z_2W_{t-1} = s_2)\left(\prod_{r=1}^{2} P(P_{t,k_r\bullet} = z_r)\right)\\
&= \sum_{i_1=1}^{p}\sum_{l_1=1}^{K} \alpha_{k_1l_1}^{(i_1)}\beta_{k_2k_2}^{(0)} P(X_{t-i_1,l_1} = s_1, e_{t,k_2} = s_2)\\
&= \sum_{i_1=1}^{p}\sum_{l_1=1}^{K} \alpha_{k_1l_1}^{(i_1)}\beta_{k_2k_2}^{(0)} P(X_{t-i_1,l_1} = s_1)P(e_{t,k_2} = s_2),
\end{aligned}$$

as X_{t-i_1} and e_t are independence for all $i_1 \geq 1$.

Case 3: $\sum_{j=Kp}^{K(p+1)} z_{1j} = 1$ and $\sum_{j=1}^{Kp} z_{2j} = 1$:

$$\begin{aligned}
&\sum_{\substack{z_1,z_2\in\{0,1\}^{K(p+1)},\\ \sum_{j=Kp+1}^{K(p+1)} z_{1j}=1,\sum_{j=1}^{Kp} z_{2j}=1}} P(z_1W_{t-1} = s_1, z_2W_{t-1} = s_2)\left(\prod_{r=1}^{2} P(P_{t,k_r\bullet} = z_r)\right)\\
&= \sum_{i_2=1}^{p}\sum_{l_2=1}^{K} \beta_{k_1k_1}^{(0)}\alpha_{k_2l_2}^{(i_2)} P(e_{t,k_1} = s_1, X_{t-i_2,l_2} = s_2)\\
&= \sum_{i_2=1}^{p}\sum_{l_2=1}^{K} \beta_{k_1k_1}^{(0)}\alpha_{k_2l_2}^{(i_2)} P(e_{t,k_1} = s_1)P(X_{t-i_2,l_2} = s_2)
\end{aligned}$$

as X_{t-i_1} and e_t are independence for all $i_1 \geq 1$.

Case 4: $\sum_{j=Kp+1}^{K(p+1)} z_{rj} = 1$ for $r = 1, 2$:

$$\begin{aligned}
&\sum_{\substack{z_1,z_2\in\{0,1\}^{K(p+1)},\\ \sum_{j=Kp}^{K(p+1)} z_{rj}=1, r=1,2}} P(z_1W_{t-1} = s_1, z_2W_{t-1} = s_2)\left(\prod_{r=1}^{2} P(P_{t,k_r\bullet} = z_r)\right)\\
&= \beta_{k_1k_1}^{(0)}\beta_{k_2k_2}^{(0)} P(e_{t,k_1} = s_1, e_{t,k_2} = s_2)
\end{aligned}$$

Finally, setting $p = 1$ above leads to the formula in Example 4.12. □

4.8. Proof of results of Section 4.3.5

For proving the mixing property of NDVARMA processes, we first show the positiveness of their transition probability for afterwards concluding mixing from the consequently ergodicity.

4.8.1. Proof of Lemma 4.13

The proof follows the proof of Weiß (2009a) Lemma 11.2.3.1. for the univariate NDARMA process. We split the proof of Lemma 4.13 considering NDVMA(q), NDVAR(p) and NDVARMA(p,q) processes separately.

Remark 4.15 (Transition probability for NDVARMA(p,q) model)
For some vectors $\mathbf{s}_0, \mathbf{s}_1 \in \mathcal{V}^{K(p+q+1)}$ *the transition probability is given by*

$$
\begin{aligned}
p_{\mathbf{s}_0|\mathbf{s}_1}(p+q+1) &= P\left(Z_t = \mathbf{s}_0 | Z_{t-(p+q+1)} = \mathbf{s}_1\right) \\
&= \sum_{\substack{s_{\max\{p,1\}},\ldots,s_{p+q} \\ r_q,\ldots,r_{q+p}}} P\Big(X_t = s_0, \ldots, X_{t-p+1} = s_{p-1}, e_t = r_0, \ldots, e_{t-q+1} = r_{q-1} | \\
&\qquad\qquad X_{t-(p+q+1)} = s_{p+q+1}, \ldots, e_{t-(p+q+1)} = r_{p+q+1}, \ldots\Big) \\
&= \sum_{\substack{s_{\max\{p,1\}},\ldots,s_{p+q} \\ r_q,\ldots,r_{q+p}}} \prod_{n=0}^{p+q} \Big(P\left(X_{t-n} = s_n | X_{t-n-1} = s_{n+1}, \ldots, e_{t-n} = r_n, \ldots\right) \\
&\qquad\qquad \cdot P\left(e_{t-n} = r_n | X_{t-n-1} = s_{n+1}, \ldots, e_{t-n-1} = r_{n+1}, \ldots\right)\Big) \\
&= \sum_{\substack{s_{\max\{p,1\}},\ldots,s_{p+q} \\ r_q,\ldots,r_{q+p}}} \prod_{n=0}^{p+q} \Bigg(p_e(r_n) \prod_{k=1}^{K} \Bigg[\sum_{i=1}^{p} \sum_{l=1}^{K} \alpha_{kl}^{(i)} \delta_{s_{n,k} s_{i+n,l}} \\
&\qquad\qquad + \beta_{kk}^{(0)} \delta_{s_{n,k} r_{n,k}} + \sum_{j=1}^{q} \sum_{l=1}^{K} \beta_{kl}^{(j)} \delta_{s_{n,k} r_{n+j,l}} \Bigg] \Bigg)
\end{aligned}
$$

Since the innovation terms are independent of the past, the probability of the innovation taking the values r_n *can be defined by* $P\left(e_{t-n} = r_n\right) = p_e(r_n)$.

The first Lemma consider the NDVMA(q) process and its transition probability for the Markov chain representation. First, note, that the case of NDVMA(0) is trivial and follows directly by $X_t = B_t^{(0)} e_t$ and the innovation process has strict positive probability of $P(e_{t,k} = r) > 0$ for $r \in \mathcal{V}$ and $k = 1, \ldots, K$.

Lemma 4.16
Let $(Z_t, t \in \mathbb{Z})$ *be the Markov representation of a stationary NDVMA(q) process with non-reducible state space and* $q \geq 1$. *For some* $n := q + 1$ *the transition probability* $p_{s|r}(n) := P\left(Z_t = s | Z_{t-n} = r\right) > 0$ *for every* $s, r \in \mathcal{V}^{K(q+1)}$

Proof.
Since $P\left(e_{t,k} = r\right) > 0$ for every $k \in \{1, \ldots, K\}$ and $r \in \mathcal{V}$, the vector including all innovation terms can reach every state in $\mathcal{V}^K$. Furthermore, NDVMA processes exclusively depend on the innovation terms and have by definition of the innovation process positive probability to take every state in $\mathcal{V}^K$. It follows directly from the Markov chain representation $Z_t := (X_t', e_t', \ldots, e_{t-q+1}')'$ that at least $q+1$ steps, depending on the parameter matrices $\mathcal{B}^{(j)}$, are needed to get a strictly positive transition probability following directly from $P\left(e_{t,k} = r\right) > 0$ for every $k \in \{1, \ldots, K\}$ and $r \in \mathcal{V}$. □

The next Lemma states the primitivity of an NDVAR(p) process with corresponding Markov Chain representation. The proof is related to Jentsch and Reichmann (2020) [Lemma C.1] with state space $\mathcal{V}^{K(p+1)}$.

Lemma 4.17
Let $(Z_t, t \in \mathbb{Z})$ be the Markov representation of a stationary NDVAR(p) process and $p \geq 1$. For $n := \max\{p, K'\}+1$ the transition probability fulfills $p_{s|r}(n) := P(Z_t = s | Z_{t-n} = r) > 0$ for every $s, r \in \mathcal{V}^{K(p+1)}$, where K' is the number of rows $k \in \{1, \dots, K\}$ with $\sum_{l=1}^{K} \sum_{i=1}^{p} |\alpha_{kl}^{(i)}| = 1$.

Proof.
Define $\mathcal{F} := \left\{k \in \{1, \dots, K\} \middle| \beta_{kk}^{(0)} = 0\right\}$.

With $K' = |\mathcal{F}|$ the cardinality of $\mathcal{F}$ and

$$\tilde{s}_0 := \left(s_0'; s_1', \dots, s_{p-1}', r_0'\right)'$$
$$\tilde{s}_1 := \left(s'_{\max\{p,K'\}+1}; \dots, s'_{p+\max\{p,K'\}}, r'_{\max\{p,K'\}+1}\right)'$$

it follows for the transition probability with $\max\{p, K'\} + 1$ steps

$$P\left(Z_t = \tilde{s}_0 | Z_{t-max\{p,K'\}-1} = \tilde{s}_1\right) =$$
$$\sum_{\substack{s_p,\dots,s_{max\{p,K'\}},\\ r_0,\dots,r_{max\{p,K'\}}}} \left[\prod_{n=0}^{max\{p,K'\}} p_e(r_n) \underbrace{\prod_{k\in\mathcal{F}} \left(\sum_{l=1}^{K} \sum_{i=1}^{p} \alpha_{kl}^{(i)} \delta_{s_{n,k} s_{n+i,l}} \right)}_{i)} \right.$$
$$\left. \underbrace{\prod_{k\in\mathcal{F}^c} \left(\sum_{l=1}^{K} \left[\sum_{i=1}^{p} \alpha_{kl}^{(i)} \delta_{s_{n,k} s_{n+i,l}} + \beta_{kk} \delta_{s_{n,k} r_{n,k}} \right] \right)}_{ii)} \right]$$

Considering the product $ii)$ and summing over all states of $r_{n,k}$ then $\forall k \in \mathcal{F}^c$

$$\exists r_{n,k} = s_{n,k}$$

and so for every $n = 0, \dots, \max\{p, K'\}$ there exists a positive summand and therefore for every dimension $k \in \mathcal{F}^c$ the product $ii)$ is positive.

For the first case $i)$ it holds, since the time series $(X_t, t \in \mathbb{Z})$ is stationary and with Remark 4.5, we know that for every entry of X_t there exists a directed graph from minimal one innovation to the resulting X_t. So the probability, that minimal one innovation term enters $X_{t,k}$ for $k \in \mathcal{F}$ is strictly positive as consequence of the stationarity assumption.

And $\forall n = 0, \dots, \max\{p, K'\}$ and $\forall k \in \mathcal{F}$ there exists a tuple of (k, i_k, l_k) with $i_k \in \{1, \dots, p\}$ and $l_k \in \{1, \dots, K\}$ such that

$$\exists\ s_{n,k} = s_{n+i_k,l_k}.$$

As the process $(X_t, t \in \mathbb{Z})$ follows a NDVAR(p) model, the new $X_{t,k}$ value is driven by a lagged value $X_{t-n,l}$ and therefore

$$\prod_{n=0}^{max\{p,K'\}} \prod_{k\in\mathcal{F}} \left(\sum_{l=1}^{K} \sum_{i=1}^{p} \alpha_{kl}^{(i)} \delta_{s_{n,k} s_{n+i,l}} \right) > 0. \tag{4.35}$$

Overall as both parts $i)$ and $ii)$ yield in a positive product, it follows that

$$P\left(Z_t = \tilde{s}_0 | Z_{t-max\{p,K'\}-1} = \tilde{s}_1\right) > 0.$$

□

We can combine both statements to prove Lemma 4.13.

Proof of Lemma 4.13 .
First we define the two sets $\mathcal{F} := \left\{k, l \in \{1, \ldots, K\}, j = 0, \ldots, q | \beta_{kl}^{(j)} = 0\right\}$ and $\mathcal{F}^c$. For this, it is

$$\begin{aligned} & P\left(Z_t = \mathbf{s}_0 | Z_{t-\max\{p+q,K'\}-1} = \mathbf{s}_1\right) \\ &= \sum_{\substack{s_{max\{1,p\}},\ldots,s_{max\{p+q,K'\}}, \\ r_q,\ldots,r_{max\{p+q,K'\}}}} \prod_{n=0}^{max\{p+q,K'\}} p_e(r_n) \prod_{k\in\mathcal{F}} \left(\sum_{l=1}^{K}\Big[\sum_{i=1}^{p} \alpha_{kl}^{(i)} \delta_{s_{n,k}s_{n+i,l}}\Big]\right) \\ & \qquad \prod_{k\in\mathcal{F}^c} \left(\sum_{l=1}^{K}\Big[\sum_{i=1}^{p} \alpha_{kl}^{(i)} \delta_{s_{n,k}s_{n+i,l}} + \beta_{kk}^{(0)} \delta_{s_{n,k}r_{n,k}} + \sum_{j=1}^{q} \beta_{kl}^{(j)} \delta_{s_{n,k}r_{n+j,l}}\Big]\right) \end{aligned}$$

The two cases $k \in \mathcal{F}$ and $k \in \mathcal{F}^c$ are now considered separately.
$\underline{k \in \mathcal{F}}$:
The argumentation for this case is similar to the proof in Lemma 4.17 for part i).
$\underline{k \in \mathcal{F}^c}$: As the process $(X_t, t \in \mathbb{Z})$ has an irreducible state space, there always exists for $k_1, k_2 \in \mathcal{F}^c$ and $k_1 \neq k_2$ two tuple of indices $(d_{k_i}, l_{k_i}), i = 1, 2$ such that $r_{n+d_{k_1},l_{k_1}} \neq r_{n+d_{k_2},l_{k_2}}$.
So we can find a set of indices (k, d_k, l_k) to chose the entries of the summands $s_{\max\{q,1\}}, \ldots, s_{\max\{p+q,K'\}}$ and $r_q, \ldots, r_{\max\{p+q,K'\}}$. The following cases have to be considered:

1. for $n = p, \ldots, \max\{p+q, K'\}$:

$$s_{n,k} = r_{n+j_k,l_k}$$

2. $n + j_k \geq q$:

$$r_{n+j_k,l_k} := s_{n,k}$$

If the second case holds $\forall k \in \mathcal{F}^c$ and $\forall r$ then the proof is done as for every $k = 1, \ldots, K$ a positive summand can be found for all $n = 0, \ldots, \max\{p+q, K'\}$ and so the transition probability $p_{s|r}\left(\max\{p+q, K'\}+1\right) > 0$.
If $\exists k \in \mathcal{F}^c$ with $n + j_k < q$ for $n \in \{0, \ldots, \max\{p+q, K'\}\}$ and the first case doesn't occur (so $n < p$) then there exists again a path within the rows, where just a predecessor is taken. For these rows, we know that $\beta_{kl_k}^{(j_k)} = 0, j_k \in \{0, \ldots, q-1\}$ and $\alpha_{kl_k}^{(i_k)} = 0, i_k \in \{1, \ldots, p-q\}$ and so since no vector entry is free to chose the entry within the row is chosen from a previous value. Thus, we get a positive transition probability.

□

4.8.2. Proof of Theorem 4.14

Proof.
The proof follows the same arguments as proof Theorem 3.15 given in 3.8.2 for the state space $\mathcal{V}^K = \{0, 1, \ldots, m\}$. □

5. Conclusion and Outlook

In this thesis, we introduced three time series models for categorical and especially binary data as extensions to the so far known models. The new models have a nice interpretable form due to their ARMA structure and are still parsimonious in the number of parameters. The latter is a main issue for multivariate time series, where the curse of dimensionality has to be considered. The special case of binary data, and consequently only two categories, are discussed in Chapters 2 and 3. The introduced *generalized binary ARMA* and *generalized binary vector ARMA* processes include the extension to also allow negative parameters, indicating a systematic switching to the opposite state and thus they can additional handle also negative dependence structure.

In a simulation study in Chapter 2.3.2, we compared the estimation of univariate gbAR processes with the well-known Markov models by using the goodness of prediction via the transition probability. Interestingly, the gbAR model performs better for small sample sizes even when the underlying process is Markovian. Additionally, the gbAR(2) model can capture nearly the same amount of dependence structure as a Markov process of second order. However, this observation is only based on the processes used in the simulation study. An open question is, whether a theoretical investigation can also confirm the two latter findings. It might be interesting to investigate how much more dependence structure a Markov process can model compared to gbARMA processes in general.

The multivariate extension of the gbARMA and the NDARMA model to the *gbVARMA* and *new discrete vector ARMA* model provides the possibility to include cross-sectional dependence structures, however the multivariate extension leads to identification issues which we have to keep in mind. For this purpose, several problems e.g. the diagonal structure of the parameter matrix for the innovation processes of lag zero, as well as the identification of the innovation distribution are discussed. Additionally, the concept of processes with non-reducible state space is introduced.

Both multivariate models make use of a row-wise independent selection of the multinomial distributed random vectors, such that only one entry of the vector valued processes is chosen, in contrast to the GDARMA model of Möller and Weiß (2020) where the new value is chosen vector-wise. These approaches form two opposing methods, but nevertheless it is conceivable to combine both approaches as a further extension.

Hence, the selection mechanism of the models is one point for possible further extensions, we shortly discussed in Remark 3.11. It is also imaginable that the choice of one dimension does not always depend on the choice of the other dimension, but its dependence can be modeled by a random variable, selecting either a perfect dependence or a independent chosen random vector by the multinomial selection. For instance, let $z \sim Bernoulli(1, \eta)$ and $k_1, k_2 \in \{1, \dots, K\}$, the choice of the new state of X_t in dimension k_1 follows

$$P_{t,k_1\bullet} = zP_{t,k_2\bullet} + (1-z)P_{t,k_1\bullet} \quad t \in \mathbb{Z}.$$

Whenever the new random variable z is equal to one, the multinomial selection mech-

anism in dimension k_1 takes the same selection as in dimension k_2, whereas with probability $(1 - \eta)$ a non depending selection is used in dimension k_1.

Some different options are conceivable here, which change the model to such an extent that the stochastic properties might change, as e.g. the transition probability and the stochastic properties depending on it.

One potential data with binary and time dependent structure are networks that evolve over time. Applying our gbVARMA model to such data, two further extensions are thinkable to describe the additional structure of network data. For instance in friendship networks it is more likely to be in contact with the friends of your friends, such that an edge is more likely to appear to the neighbor nodes of distance two. Whereas the actual gbVARMA model chooses a value from either all edges or no edges, the feature of having more likely an edge to the friends of my friends is not totally included. We might think of calculating first the neighborhood for each node and then include the additional information into the model to emphasize the constitution of the neighbors in a possible forecast for the next step. Furthermore, as the number of nodes K can be large, a sparsity within the parameter matrices is possible and this characteristic is not captured by the classical Yule-Walker estimators. An investigation on the goodness of a sparse estimator, as e.g. the Lasso technique, has to be performed.

Bibliography

Bai, C., S. Kumar, J. Leskovec, M. Metzger, J. Nunamaker, and V. S. Subrahmanian (July 2019). "Predicting the Visual Focus of Attention in Multi-Person Discussion Videos". In: *IJCAI 2019*. International Joint Conferences on Artificial Intelligence, pp. 4504–4510.

Bartlett, M. S. (1951). "The frequency goodness of fit test for probability chains". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 47.1, pp. 86–95.

Bellégo, C. and L. Ferrara (2009). *Forecasting Euro-area recessions using time-varying binary response models for financial*. Working papers. Banque de France. URL: `https://EconPapers.repec.org/RePEc:bfr:banfra:259`.

Bellman, R. E. (2016). *Adaptive Control Processes: A Guided Tour*. Princeton Legacy Library. Princeton University Press. ISBN: 9780691652214.

Berchtold, A. and A. E. Raftery (2002). "The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series". In: *Statistical Science* 17.3, pp. 328–356.

Billingsley, P. (1961). "Statistical Methods in Markov Chains". In: *The Annals of Mathematical Statistics* 32.1, pp. 12–40.

— (1968). *Convergence of probability measures*. 1. Edition. John Wiley and Sons, Inc., New York.

Biswas, A. and P. X. - K. Song (2009). "Discrete - valued ARMA processes". In: *Statistics and Probability Letters* 79.17, pp. 1884–1889.

Bradley, R. C. (2007). *Introduction to strong mixing conditions*. Vol. 1-3. Kendrick Press.

Brockwell, P. J. and R. A. Davis (1991). *Time Series: Theory and Methods*. Springer-Verlag New York: Springer-Verlag.

Bühlmann, P. and A. J. Wyner (1999). "Variable length Markov Chains". In: *The Annals of Statistics* 27.2, pp. 480–513.

Buishand, T. A. (1978). *The Binary DARMA(1,1) Process as a Model for Wet-Dry-Sequences*. Technical Note. Wageningen: Departement of Mathematics, Statistics Division, Agricultural University Wageningen.

Cheng, L., Y. Wang, C. Wu, H. Wu, and Y. Zhang (2013). "Signal processing for a positioning system with binary sensory outputs". In: *Sensors and Actuators A: Physical* 201, pp. 86–92.

Dai, B., S. Ding, and G. Wahba (2013). "Multivariate Bernoulli distribution". In: *Bernoulli* 19(4), pp. 1465–1483.

Davis, R. A., S. H. Holan, R. Lund, and N. Ravishanker (2016). *Handbook of Discrete-Valued Time Series*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Taylor & Francis. ISBN: 9781466577732.

Dedecker, J., P. Doukhan, G. Lang, J.R. Leon, S. Louhichi, and C. Prieur (2007). *Weak Dependence: With Examples and Applications*. Lecture Notes in Statistics, Springer - Verlag New York.

Dehnert, M., R. Plaumann, W. E. Helm, and M. T. Hütt (2005). "Genome phylogeny based on short-range correlations in DNA sequences". In: *Journal of computational Biology* 12(5), pp. 545–553.

Doukhan, P. (1994). *Mixing - Properties and Examples*. Lecture Notes in Statistics, Springer - Verlag New York.

Fitzmaurice, G. M. and S. R. Lipsitz (1995). "A Model for Binary Time Series Data with Serial Odds Ratio Patterns". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 44.1, pp. 51–61.

Gouveia, S., T. Möller, C.H. Weiß, and M. Scotto (2018). "A full ARMA model for counts with bounded support and its application to rainy-days time series". In: *Stochastic Environmental Research and Risk Assessment* 32. DOI: 10.1007/s00477-018-1584-3.

Jackson, M. O. (2008). *Social and Economic Networks*. Princeton University Press, Princeton.

Jacobs, P. A. and P. A. W. Lewis (1978a). "Discrete time series generated by mixtures. I: Correlation and runs properties". In: *Journal of the Royal Statistical Society B* 40.1, pp. 94–105.

— (1978b). "Discrete time series generated by mixtures. II: Asymptotic properties". In: *Journal of the Royal Statistical Society B* 40.2, pp. 222–228.

— (1978c). "Discrete time series generated by mixtures. III: Autoregressive processes (DAR(p))". In: *Naval Postgraduate School Technical Report NPS55-78-022*.

— (1983). "Stationary discrete autoregressive-moving average time series generated by mixtures". In: *Journal of Time Series Analysis* 4.1, pp. 19–36.

Jentsch, C. and L. Reichmann (2019). "Generalized Binary Time Series Models". In: *Econometrics* 7,49.

— (2020). "Generalized binary vector AR processes". In: *Working Paper*.

Jentsch, C. and C. H. Weiß (2019). "Bootstrapping INAR models". In: *Bernoulli 25* 3, pp. 2359–2408.

Jiang, W., S. Song, L. Hou, and H. Zhao (2020). "A set of efficient methods to generate high-dimensional binary data with specified correlation structures". In: *ArXiv:2007.14080v1*.

Johnson, N., S. Kotz, and N. Balakrishnan (1997). *Discrete Multivariate Distributions*. Wiley & Sons.

Kanter, M. (1975). "Autoregression for Discrete Processes Mod 2". In: *Journal of Applied Probability* 12.2, pp. 371–375.

Kedem, B. (1980). *Binary Time Series*. Lecture notes in pure and applied mathematics. M. Dekker.

Kedem, B. and K. Fokianos (2002). *Regression models for time series analysis*. Vol. 488. John Wiley & Sons.

Kolaczyk, E. (2009). *Statistical Analysis of Network Data - Methods and Models*. Springer, New York.

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

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer - Verlag Berlin Heidelberg.

Mazzi, G. L. (2015). "Complementing scoreboards with composite indicators: the new business cycle clock". In: *EURONA - Eurostat Review on National Accounts and Macroeconomic Indicators, Luxembourg* 02.

McKenzie, E. (1981). "Extending the correlation structure of exponential autoregressive–moving-average processes". In: *Journal of Applied Probability* 18.1, pp. 181–189.

— (2003). "Ch. 16. Discrete variate time series". In: *Stochastic Processes: Modelling and Simulation*. Vol. 21. Handbook of Statistics. Elsevier, pp. 573–606.

Möller, T. A. and C. H. Weiß (2020). "Generalized discrete autoregressive moving-average models". In: *Applied Stochastic Models in Business and Industry* 36.4.

Moysiadis, T. and K. Fokianos (2014). "On binary and categorical time series models with feedback". In: *Journal of Multivariate Analysis* 131, pp. 209–228.

Newman, M. E. J. (2010). *Networks - An Introduction*. Oxford University Press, New York.

Paranjape, A., A. R. Benson, and J. Leskovec (2016). "Motifs in Temporal Networks". In: *https://arxiv.org/abs/1612.09259*.

Raftery, A. E. (1985). "A Model for High-Order Markov Chains". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 47.3, pp. 528–539.

Ritzwoller, D. M. and J. P. Romano (2020). "Uncertainty in the hot hand fallacy: Detecting streaky alternatives to random Bernoulli sequences". In: *Technical Report No. 2020-02, February 2020, Department of Statistics, Stanford University*.

Seneta, E. (1981). *Non-negative Matrices and Markov Chains*. Springer - Verlag New York.

Startz, R. (2008). "Binomial Autoregressive Moving Average Models With an Application to U.S. Recessions". In: *Journal of Business & Economic Statistics* 26.1, pp. 1–8.

Stoffer, D. S., M. S. Scher, G. A. Richardson, N. L. Day, and P. A. Coble (1988). "A Walsh—Fourier Analysis of the Effects of Moderate Maternal Alcohol Consumption on Neonatal Sleep-State Cycling". In: *Journal of the American Statistical Association* 83.404, pp. 954–963.

Wei, W. (Jan. 1989). *Time Series Analysis: Univariate and Multivariate Methods*. Vol. 33.

Weiß, C. H. (2009a). *Categorical Time Series Analysis and Applications in Statistical Quality Control*. Dissertation Premium. dissertation.de. ISBN: 9783866244429.

— (2009b). "Properties of a class of binary ARMA models". In: *A Journal of Theoretical and Applied Statistics* 43.2, pp. 131–138.

— (2011a). "Empirical measures of signed serial dependence in categorical time seried". In: *Journal of Statistical Computation and Simulation* 81 Issue 4, pp. 411–429.

— (2011b). "Generalized choice models for categorical time series". In: *Journal of Statistical Planning and Inference* 141.8, pp. 2849–2862.

— (2013). “Serial dependence of NDARMA processes”. In: *Computational Statistics & Data Analysis Vol 68*, pp. 213–238.

— (2018). *An introduction to discrete-valued time series*. John Wiley & Sons.

— (2020). “Regime-Switching Discrete ARMA Models for Categorical Time Series”. In: *Entropy* 22.4.

Weiß, C. H. and R. Göb (Feb. 2008). “Measuring serial dependence in categorical time series”. In: *AStA Advances in Statistical Analysis* 92.1, pp. 71–89.

Zarnowitz, V. (1992). “What is a Business Cycle?” In: *The Business Cycle: Theories and Evidence*. Ed. by Belongia M.T. and Garfinkel M.R. Springer, Dordrecht, pp. 3–83.

A. Appendix

A.1. A second representation of the Yule-Walker equation of lag zero

An additional representation of the Yule-Walker equation for gbVAR and also NDVAR processes of order $p \in \mathbb{N}$ is discussed in the following Section. In contrast to the given representation in Theorem 3.7, we follow Lütkepohl (2005), by first vectorizing the autocovariance formula for a direct calculation of the expectation terms. Furthermore, the proof can be stated independent of the state space $\mathcal{V}^K$, such that the calculations hold for gbVAR and NDVAR processes without adjustments.

The vectorize operator *vec* stacks the entries of a matrix column-wise and has the spin-off that a product of matrices can be rearranged involving a Kronecker products denoted by $\otimes$. Besides, in our case we have to include also Hadamard products denoted by $\circ$ due to the multinomial selection mechanism and the product of dependent random variables. For this purpose, we make use of three auxiliary matrices ψ, ψ_1 and ψ_2, which are suitable to deal with expectation of the matrices having entries of the form $a^{(+,i)}_{k_1 l_1,t} a^{(+,i)}_{k_2 l_2,t}$, $a^{(-,i)}_{k_1 l_1,t} a^{(-,i)}_{k_2 l_2,t}$ and $a^{(+,i)}_{k_1 l_1,t} a^{(-,i)}_{k_2 l_2,t}$ for $k, l = 1, \dots, K$ and $i = 1, \dots, p$, also discussed in Lemma 3.23. Let

$$\psi = \begin{pmatrix} 0_{1\times K^2} \\ \mathbb{1}_{K\times K^2} \\ 0_{1\times K^2} \\ \mathbb{1}_{K\times K^2} \\ \vdots \\ 0_{1\times K^2} \end{pmatrix} \qquad \psi_1 = \begin{pmatrix} u_1' & u_2' & \dots & u_K' \\ 0_{K\times K} & 0_{K\times K} & \dots & 0_{K\times K} \\ u_1' & u_2' & \dots & u_K' \\ 0_{K\times K} & 0_{K\times K} & \dots & 0_{K\times K} \\ & \vdots & & \\ u_1' & u_2' & \dots & u_K' \end{pmatrix} \qquad \psi_2 = \begin{pmatrix} -\mathbb{1}_{1\times K} \\ 0_{K\times K^2} \\ -\mathbb{1}_{1\times K} \\ 0_{K\times K^2} \\ \vdots \\ -\mathbb{1}_{1\times K} \end{pmatrix} \tag{A.1}$$

with u_i is the i-th unit vector.

Theorem A.1 (Yule - Walker equation of lag $h = 0$)
Let ψ, ψ_1, ψ_2 be $K^2 \times K^2$ auxiliary matrices defined in (A.1)*. Let $(X_t, t \in \mathbb{Z})$ be a K-dimensional gbVAR(p) process with autocovariance $\Gamma_X(0)$.*

$$\Big(I_{K^2\times K^2} - \sum_{i=1}^{p} \big[\psi \circ \mathcal{A}^{(i)} \otimes \mathcal{A}^{(i)} + \psi_1 \circ \mathcal{A}^{(i)}_{|\cdot|} \otimes \mathbb{1}_{K\times K}\big]\Big) vec\left(\Gamma_X(0)\right) =$$

$$\sum_{\substack{i_1,i_2=1,\\ i_1\neq i_2}}^{p} \big[\psi \circ \mathcal{A}^{(i_2)} \otimes \mathcal{A}^{(i_1)}\big] vec\left(\Gamma_X(i_2 - i_1)\right) + \big[\psi \circ \mathcal{B} \otimes \mathcal{B} + \psi_1 \circ \mathcal{B} \otimes \mathbb{1}_{K\times K}\big] vec\left(\Sigma_e\right) +$$

$$\Big(\sum_{i_1,i_2=1}^{p} \big[\psi_2 \circ \mathcal{A}^{(i_2)} \otimes \mathcal{A}^{(i_1)}\big] + \sum_{i=1}^{p} \big[\psi_1 \circ \mathcal{A}^{(i)}_{|\cdot|} \otimes \mathbb{1}_{K\times K}\big]\Big) vec\big(\mu_X \mu_X'\big) +$$

$$\Big(\sum_{i_1,i_2=1}^{p} \big[\psi_2 \circ \mathcal{A}^{(-,i_2)} \otimes \mathcal{A}^{(i_1)}\big] + \sum_{i=1}^{p} \big[-\psi_1 \circ \mathcal{A}^{(-,i)} \otimes \mathbb{1}_{K\times K}\big]\Big) vec\big(\mu_X \mathbb{1}'\big) +$$

$$
\begin{aligned}
&\sum_{i=1}^{p}\big[\psi_2 \circ \mathcal{B} \otimes \mathcal{A}^{(i)}\big] vec\big(\mu_X \mu_e'\big) + \\
&\Big(\sum_{i_1,i_2=1}^{p}\big[\psi_2 \circ \mathcal{A}^{(i_2)} \otimes \mathcal{A}^{(-,i_1)}\big] + \sum_{i=1}^{p}\big[-\psi_1 \circ \mathcal{A}^{(-,i)} \otimes \mathbb{1}_{K\times K}\Big) vec\big(\mathbb{1}\mu_X'\big) + \\
&\Big(\sum_{i_1,i_2=1}^{p}\big[\psi_2 \circ \mathcal{A}^{(-,i_2)} \otimes \mathcal{A}^{(-,i_1)}\big] + \sum_{i=1}^{p}\big[\psi_1 \circ \mathcal{A}^{(-,i)} \otimes \mathbb{1}_{K\times K}\big]\Big) vec\big(\mathbb{1}\mathbb{1}'\big) + \\
&\sum_{i=1}^{p}\big[\psi_2 \circ \mathcal{B} \otimes \mathcal{A}^{(-,i)}\big] vec\big(\mathbb{1}\mu_e'\big) + \sum_{i=1}^{p}\big[\psi_2 \circ \mathcal{A}^{(i)} \otimes \mathcal{B}\big] vec\big(\mu_e \mu_X'\big) + \\
&\sum_{i=1}^{p}\big[\psi_2 \circ \mathcal{A}^{(-,i)} \otimes \mathcal{B}\big] vec\big(\mu_e \mathbb{1}'\big) + \big[\psi_2 \circ \mathcal{B} \otimes \mathcal{B} + \psi_1 \circ \mathcal{B} \otimes \mathbb{1}_{K\times K}\big] vec\big(\mu_e \mu_e'\big)
\end{aligned}
$$

with $\mu_X := E[X_t]$ *and* $\mu_e := E[e_t]$ *and* $\Sigma_e = Cov(e_t, e_t)$.

The above formula can be also applied for NDVAR(p) processes by setting $\mathcal{A}_t^{(\cdot,-)} := 0_{K\times K}$ and take $\mathcal{A}_{|\cdot|}^{(\cdot)} = \mathcal{A}^{(\cdot)}$.

Proof.
We fragment the autocovariance equation in the same way as given in proof in Chapter 3.7.5 but by considering first its expectation.

$$\Gamma_X(0) = E\Big[X_t X_t'\Big] - E\Big[X_t\Big] E\Big[X_t'\Big]$$

Initially, we calculate the first part of the right hand side and afterwards combining it with the product of the expectation vectors $\mu_X = E(X_t)$.

$$
\begin{aligned}
E\Big[X_t X_t'\Big] = E\Bigg[&\underbrace{\sum_{i_1,i_2=1}^{p} A_t^{(+,i_1)} X_{t-i_1} X_{t-i_2}' A_t^{(+,i_2)\prime}}_{I_1} + \underbrace{\sum_{i_1,i_2=1}^{p} A_t^{(+,i_1)} X_{t-i_1} \mathbb{1}' A_t^{(-,i_2)\prime}}_{I_2} \\
&+ \underbrace{\sum_{i=1}^{p} A_t^{(+,i)} X_{t-i} e_t' B_t'}_{I_3} + \underbrace{\sum_{i_1,i_2=1}^{p} A_t^{(-,i_1)} \mathbb{1} X_{t-i_2} A_t^{(+,i_2)\prime}}_{I_4} + \underbrace{\sum_{i_1,i_2=1}^{p} A_t^{(-,i_1)} \mathbb{1}\mathbb{1}' A_t^{(-,i_2)\prime}}_{I_5} \\
&+ \underbrace{\sum_{i=1}^{p} A_t^{(-,i)} \mathbb{1} e_t' B_t'}_{I_6} + \underbrace{\sum_{i=1}^{p} B_t e_t X_{t-i}' A_t^{(+,i)\prime}}_{I_7} + \underbrace{\sum_{i=1}^{p} B_t e_t \mathbb{1}' A_t^{(-,i)\prime}}_{I_8} + \underbrace{B_t e_t e_t' B_t'}_{I_9}\Bigg]
\end{aligned}
\tag{A.2}
$$

Applying now the vectorization operator to each summand and make use of

$$vec(ABC) = C' \otimes A \; vec(B).$$

we can examine each term $I_n,\ n = 1, \ldots, 9$ individually.

Additionally, we distinguish between terms with the same order index $i_1 = i_2$ and $i_1 \neq i_2$ for $i_1, i_2 = 1, \ldots, p$.

$$
\begin{aligned}
vec\big(I_1\big) &= \sum_{i=1}^{p} E\Big[vec\big(A_t^{(+,i)} X_{t-i} X_{t-i}' A_t^{(+,i)\prime}\big)\Big] + \sum_{\substack{i_1,i_2=1,\\ i_1\neq i_2}}^{p} E\Big[vec\big(A_t^{(+,i_1)} X_{t-i_1} X_{t-i_2}' A_t^{(+,i_2)\prime}\big)\Big] \\
&= \sum_{i=1}^{p} E\Big[A_t^{(+,i)} \otimes A_t^{(+,i)} vec\big(X_{t-i} X_{t-i}'\big)\Big] + \sum_{\substack{i_1,i_2=1,\\ i_1\neq i_2}}^{p} E\Big[A_t^{(+,i_1)} \otimes A_t^{(+,i_2)} vec\big(X_{t-i_1} X_{t-i_2}'\big)\Big] \\
&= \sum_{i=1}^{p} E\Big[A_t^{(+,i)} \otimes A_t^{(+,i)}\Big] E\Big[vec\big(X_{t-i} X_{t-i}'\big)\Big] \\
&+ \sum_{\substack{i_1,i_2=1,\\ i_1\neq i_2}}^{p} E\Big[A_t^{(+,i_1)} \otimes A_t^{(+,i_2)}\Big] E\Big[vec\big(X_{t-i_1} X_{t-i_2}'\big)\Big]
\end{aligned}
$$

The evaluation of the Kronecker product of the random matrices $A_t^{(+,\cdot)}$, we utilize the considerations of Lemma 3.23. This results in the following matrices ψ and ψ_1, which selects the corresponding matrix entries, either dependent, when both entries are contained in the same dimension $P_{t,k\bullet}$ or independent, corresponding to different dimensions $P_{t,k_1\bullet}$ and $P_{t,k_2\bullet}$. The first matrix ψ selects the independent entries and the dependent entries are set to zero. This yields into the following expectation for $k_1 \neq k_2$

$$
E\Big[\psi \circ A_t^{(+,i_1)} \otimes A_t^{(+,i_2)}\Big] = \psi \circ E\Big[A_t^{(+,i_1)}\Big] \otimes E\Big[A_t^{(+,i_2)}\Big] = \psi \circ \mathcal{A}^{(i_1)} \otimes \mathcal{A}^{(i_2)}.
$$

In rows with dependent entries, we have to keep in mind, that the entries are together chosen by the multinomial selection mechanism. The Hadamard product with the defined matrix ψ_1 collects the dependent entries within the Kronecker product of the modified random matrices and we apply Lemma 3.23, such that

$$
E[a_{t,kl_1}^{(+,i_1)} a_{t,kl_2}^{(+,i_2)}] = \begin{cases} E(a_{t,kl_1}^{(+,i)}) & \text{for } l_1 = l_2,\ i_1 = i_2 \\ 0 & otherwise \end{cases}
$$

Overall it follows with the same argumentation for I_4, I_5 and I_9.

$$
E\big[A_t^{(+,i)} \otimes A_t^{(+,i)}\big] = \psi \circ \mathcal{A}^{(i)} \otimes \mathcal{A}^{(i)} + \psi_1 \circ \mathcal{A}_{|\cdot|}^{(i)} \otimes \mathbb{1}_{K\times K} \tag{A.3}
$$

$$
E\big[A_t^{(-,i)} \otimes A_t^{(+,i)}\big] = \psi \circ \mathcal{A}^{(-,i)} \otimes \mathcal{A}^{(i)} + \psi_1 \circ (-\mathcal{A}^{(-,i)}) \otimes \mathbb{1}_{K\times K} \tag{A.4}
$$

$$
E\big[A_t^{(-,i)} \otimes A_t^{(-,i)}\big] = \psi \circ \mathcal{A}^{(-,i)} \otimes \mathcal{A}^{(-,i)} + \psi_1 \circ \mathcal{A}^{(-,i)} \otimes \mathbb{1}_{K\times K} \tag{A.5}
$$

$$
E\big[B_t \otimes B_t\big] = \psi \circ \mathcal{B} \otimes \mathcal{B} + \psi_1 \circ \mathcal{B} \otimes \mathbb{1}_{K\times K}. \tag{A.6}
$$

Therefore the above consideration of the entries we get for the evaluation of I_1

$$
\begin{aligned}
E\left(I_1\right) &= \sum_{i=1}^{p} \Big[\psi \circ \mathcal{A}^{(i)} \otimes \mathcal{A}^{(i)} + \psi_1 \circ \mathcal{A}_{|\cdot|}^{(i)} \otimes \mathbb{1}_{K\times K}\Big] vec\Big(E\big[X_{t-i} X_{t-i}'\big]\Big) \\
&+ \sum_{\substack{i_1,i_2=1,\\ i_1\neq i_2}}^{p} \Big[\psi \circ \mathcal{A}^{(i_2)} \otimes \mathcal{A}^{(i_1)}\Big] vec\Big(E\big[X_{t-i_1} X_{t-i_2}'\big]\Big)
\end{aligned}
$$

$$vec(I_2) = \sum_{\substack{i_1,i_2=1,\\ i_1\neq i_2}}^{p} \Big[\psi \circ \mathcal{A}^{(-,i_2)} \otimes \mathcal{A}^{(i_1)}\Big] vec\big(E\big[X_{t-i_1}\mathbb{1}'\big]\big) + \sum_{i=1}^{p} \big[\psi \circ \mathcal{A}^{(-,i)} \otimes \mathcal{A}^{(i)} - \psi_1 \circ \mathcal{A}^{(-,i)} \otimes \mathbb{1}_{K\times K}\big] vec\big(E\big[X_{t-i}\mathbb{1}'\big]\big)$$

$$vec\,(I_3) = \sum_{i=1}^{p} \psi \circ \mathcal{B} \otimes \mathcal{A}^{(i)} vec\Big[E\big(X_{t-i}\big)E\big(e_t'\big)\Big] = \sum_{i=1}^{p} \psi \circ \mathcal{B} \otimes \mathcal{A}^{(i)} vec\Big[\mu_X \mu_e'\Big]$$

The next-to-last equation follows from the independence of the innovation term e_t and the $(X_s, s < t)$.

By evaluating the next term we have to distinguish again between the summands with the same index and the others.

$$\begin{aligned}
vec\big(I_4\big) &= \sum_{\substack{i_1,i_2=1,\\ i_1\neq i_2}}^{p} \big[\psi \circ \mathcal{A}^{(i_2)} \otimes \mathcal{A}^{(-,i_1)}\big] vec\big(\mathbb{1}\mu_X'\big) \\
&\quad + \sum_{i=1}^{p} \big[\psi \circ \mathcal{A}^{(i)} \otimes \mathcal{A}^{(-,i)} - \psi_1 \circ \mathcal{A}^{(-,i)} \otimes \mathbb{1}_{K\times K}\big] vec\big(\mathbb{1}\mu_X'\big) \\
vec\big(I_5\big) &= \sum_{\substack{i_1,i_2=1,\\ i_1\neq i_2}}^{p} \psi \circ \mathcal{A}^{(-,i_2)} \otimes \mathcal{A}^{(-,i_1)} vec\big(\mathbb{1}\mathbb{1}'\big) \\
&\quad + \sum_{i=1}^{p} \big[\psi \circ \mathcal{A}^{(-,i)} \otimes \mathcal{A}^{(-,i)} + \psi_1 \circ \mathcal{A}^{(-,i)} \otimes \mathbb{1}_{K\times K}\big] vec\big(\mathbb{1}\mathbb{1}'\big) \\
vec\big(I_6\big) &= \sum_{i=1}^{p} \big[\psi \circ \mathcal{B} \otimes \mathcal{A}^{(-,i)}\big] vec\big(\mathbb{1}\mu_e'\big) \\
vec\big(I_7\big) &= \sum_{i=1}^{p} \big[\psi \circ \mathcal{A}^{(i)} \otimes \mathcal{B}\big] vec\big(\mu_e \mu_X'\big) \\
vec\big(I_8\big) &= \sum_{i=1}^{p} \big[\psi \circ \mathcal{A}^{(-,i)} \otimes \mathcal{B}\big] vec\big(\mu_e \mathbb{1}'\big) \\
vec\big(I_9\big) &= \big[\psi \circ \mathcal{B} \otimes \mathcal{B} + \psi_1 \circ \mathcal{B} \otimes \mathbb{1}_{K\times K}\big] vec\Big(E\Big[e_t e_t'\Big]\Big)
\end{aligned}$$

Finally, the multiplied mean vectors are first vectorized and then substracted from the above evaluated equations.

$$vec\big(\mu_X \mu_X'\big) = \sum_{i_1,i_2=1}^{p} \Big[\mathcal{A}^{(i_2)} \otimes \mathcal{A}^{(i_1)}\Big]\, vec\big(\mu_X \mu_X'\big) + \sum_{i_1,i_2=1}^{p} \Big[\mathcal{A}^{(-,i_2)} \otimes \mathcal{A}^{(i_1)}\Big]\, vec\big(\mu_X \mathbb{1}'\big) \tag{A.7}$$

$$+ \sum_{i=1}^{p} \Big[\mathcal{B} \otimes \mathcal{A}^{(i)}\Big]\, vec\big(\mu_X \mu_e'\big) + \sum_{i_1,i_2=1}^{p} \Big[\mathcal{A}^{(i_2)} \otimes \mathcal{A}^{(-,i_1)}\Big]\, vec\big(\mathbb{1}\mu_X'\big)$$

$$+ \sum_{i_1,i_2=1}^{p} \Big[\mathcal{A}^{(-,i_2)} \otimes \mathcal{A}^{(-,i_1)}\Big]\, vec\big(\mathbb{1}\mathbb{1}'\big) + \sum_{i=1}^{p} \Big[\mathcal{B} \otimes \mathcal{A}^{(-,i)}\Big]\, vec\big(\mathbb{1}\mu_e'\big)$$

$$+\sum_{i=1}^{p}\Big[\mathcal{A}^{(i)}\otimes\mathcal{B}\Big]\,vec\big(\mu_e\mu_X'\big)+\sum_{i=1}^{p}\Big[\mathcal{A}^{(-,i)}\otimes\mathcal{B}\Big]\,vec\big(\mu_e\mathbb{1}'\big)+[\mathcal{B}\otimes\mathcal{B}]\,vec\big(\mu_e\mu_e'\big)$$

We additional use the following decompositions for further calculations.

$$\begin{aligned}E\big[X_{t-i}X_{t-i}'\big]&=\Gamma_X(0)+E\big[X_{t-i}\big]E\big[X_{t-i}'\big]\\E\big[X_{t-i_1}X_{t-i_2}'\big]&=\Gamma_X(i_2-i_1)+E\big[X_{t-i_1}\big]E\big[X_{t-i_2}'\big]\\E\big[e_te_t'\big]&=\Sigma_e+E\big[e_t\big]E\big[e_t'\big].\end{aligned}$$

Combining now these decompositions with (A.2) and substracting (A.7), the difference between the two Kronecker products lead to a sparse matrix. Especially the rows of independent entries within (A.3) to (A.6) eliminates to zero, more precisely, only the sums multiplied with ψ_1 remain. For this, we need another auxiliary matrix, which picks up the remaining entries of the Kronecker product

$$\psi_2:=\begin{pmatrix}-1&-1&\dots&-1\\0&0&\dots&0\\\vdots&\vdots&&\vdots\\0&0&\dots&0\\-1&-1&\dots&-1\\&\vdots&&\\-1&-1&\dots&-1\end{pmatrix}\text{ with dimension }K^2\times K^2.$$

The rows of zero entries have dimension $K\times K^2$ and therefore a -1 appears in a dimension corresponding to a diagonal element of the autocovariance matrix $\Gamma_X(0)$. By using again the Hadamard product it follows

$$\begin{aligned}vec\left(\Gamma_X\left(0\right)\right)&=\sum_{i=1}^{p}\Big[\psi\circ\mathcal{A}^{(i)}\otimes\mathcal{A}^{(l)}+\psi_1\circ\mathcal{A}^{(i)}_{|\cdot|}\otimes\mathbb{1}_{K\times K}\Big]vec\big(\Gamma_X\left(0\right)\big)\\&+\sum_{\substack{i_1,i_2=1,\\i_1\neq i_2}}^{p}\Big[\psi\circ\mathcal{A}^{(i_2)}\otimes\mathcal{A}^{(i_1)}\Big]vec\Big(\Gamma_X\left(i_2-i_1\right)\Big)\\&+\left(\sum_{i=1}^{p}\big[\psi_1\circ\mathcal{A}^{(i)}_{|\cdot|}\otimes\mathbb{1}_{K\times K}+\psi_2\circ\mathcal{A}^{(i)}\otimes\mathcal{A}^{(i)}\big]+\sum_{\substack{i_1,i_2=1,\\i_1\neq i_2}}^{p}\big[\psi_2\circ\mathcal{A}^{(i_2)}\otimes\mathcal{A}^{(i_1)}\big]\right)vec\big(\mu_X\mu_X'\big)\\&+\left(\sum_{\substack{i_1,i_2=1,\\i_1\neq i_2}}^{p}\big[\psi_2\circ\mathcal{A}^{(-,i_2)}\otimes\mathcal{A}^{(i_1)}\big]+\sum_{i=1}^{p}\big[-\psi_1\circ\mathcal{A}^{(-,i)}\otimes\mathbb{1}_{K\times K}+\psi_2\circ\mathcal{A}^{(-,i)}\otimes\mathcal{A}^{(i)}\big]\right)vec\big(\mu_X\mathbb{1}'\big)\\&+\left(\sum_{i=1}^{p}\psi_2\circ\mathcal{B}\otimes\mathcal{A}^{(i)}\right)vec\big(\mu_X\mu_e'\big)\\&+\left(\sum_{\substack{i_1,i_2=1,\\i_1\neq i_2}}^{p}\big[\psi_2\circ\mathcal{A}^{(i_1)}\otimes\mathcal{A}^{(-,i_2)}\big]+\sum_{i=1}^{p}\big[-\psi_1\circ\mathcal{A}^{(-,i)}\otimes\mathbb{1}_{K\times K}+\psi_2\circ\mathcal{A}^{(i)}\otimes\mathcal{A}^{(-,i)}\big]\right)vec\big(\mathbb{1}\mu_X'\big)\\&+\left(\sum_{\substack{i_1,i_2=1,\\i_1\neq i_2}}^{p}\psi_2\circ\mathcal{A}^{(-,i_2)}\otimes\mathcal{A}^{(-,i_1)}+\sum_{i=1}^{p}\big[\psi_1\circ\mathcal{A}^{(-,i)}\mathbb{1}_{K\times K}+\psi_2\circ\mathcal{A}^{(-,i)}\otimes\mathcal{A}^{(-,i)}\big]\right)vec\big(\mathbb{1}\mathbb{1}'\big)\end{aligned}$$

$$
\begin{aligned}
&+\Big(\sum_{i=1}^{p}\psi_2\circ\mathcal{B}\otimes\mathcal{A}^{(-,i)}\Big)vec\big(\mathbb{1}\mu_e'\big)+\Big(\sum_{i=1}^{p}\psi_2\circ\mathcal{A}^{(i)}\otimes\mathcal{B}\Big)vec\big(\mu_e\mu_X'\big)\\
&+\Big(\sum_{i=1}^{p}\psi_2\circ\mathcal{A}^{(-,i)}\otimes\mathcal{B}\Big)vec\big(\mu_e\mathbb{1}'\big)+\Big[\psi_2\circ\mathcal{B}\otimes\mathcal{B}+\psi_1\circ\mathcal{B}\otimes\mathbb{1}_{K\times K}\Big]vec\big(\mu_e\mu_e'\big)\\
&+\big[\psi\circ\mathcal{B}\otimes\mathcal{B}+\psi_1\circ\mathcal{B}\otimes\mathbb{1}_{K\times K}\big]vec\left(\Sigma_e\right)
\end{aligned}
$$

Rearranging the sums of different indices $i_1 \neq i_2$ and expressions we end up with the autocovariance of lag zero.

$$
\begin{aligned}
\Big(I_{K^2\times K^2}-\sum_{i=1}^{p}\big[\psi\circ\mathcal{A}^{(i)}\otimes\mathcal{A}^{(i)}+\psi_1\circ\mathcal{A}^{(i)}_{|\cdot|}\otimes\mathbb{1}_{K\times K}\big]\Big)vec\left(\Gamma_X\left(0\right)\right)=\\
\sum_{\substack{i_1,i_2=1,\\ i_1\neq i_2}}^{p}\big[\psi\circ\mathcal{A}^{(i_2)}\otimes\mathcal{A}^{(i_1)}\big]vec\left(\Gamma_X\left(i_2-i_1\right)\right)+\big[\psi\circ\mathcal{B}\otimes\mathcal{B}+\psi_1\circ\mathcal{B}\otimes\mathbb{1}_{K\times K}\big]vec\left(\Sigma_e\right)\\
+\Big(\sum_{i_1,i_2=1}^{p}\big[\psi_2\circ\mathcal{A}^{(i_2)}\otimes\mathcal{A}^{(i_1)}\big]+\sum_{i=1}^{p}\big[\psi_1\circ\mathcal{A}^{(i)}_{|\cdot|}\otimes\mathbb{1}_{K\times K}\big]\Big)vec\big(\mu_X\mu_X'\big)\\
+\Big(\sum_{i_1,i_2=1}^{p}\big[\psi_2\circ\mathcal{A}^{(-,i_2)}\otimes\mathcal{A}^{(i_1)}\big]+\sum_{i=1}^{p}\big[-\psi_1\circ\mathcal{A}^{(-,i)}\otimes\mathbb{1}_{K\times K}\big]\Big)vec\big(\mu_X\mathbb{1}'\big)\\
+\sum_{i=1}^{p}\big[\psi_2\circ\mathcal{B}\otimes\mathcal{A}^{(i)}\big]vec\big(\mu_X\mu_e'\big)\\
+\Big(\sum_{i_1,i_2=1}^{p}\big[\psi_2\circ\mathcal{A}^{(i_2)}\otimes\mathcal{A}^{(-,i_1)}\big]+\sum_{i=1}^{p}\big[-\psi_1\circ\mathcal{A}^{(-,i)}\otimes\mathbb{1}_{K\times K}\Big)vec\big(\mathbb{1}\mu_X'\big)\\
+\Big(\sum_{i_1,i_2=1}^{p}\big[\psi_2\circ\mathcal{A}^{(-,i_2)}\otimes\mathcal{A}^{(-,i_1)}\big]+\sum_{i=1}^{p}\big[\psi_1\circ\mathcal{A}^{(-,i)}\otimes\mathbb{1}_{K\times K}\big]\Big)vec\big(\mathbb{1}\mathbb{1}'\big)\\
+\sum_{i=1}^{p}\big[\psi_2\circ\mathcal{B}\otimes\mathcal{A}^{(-,i)}\big]vec\big(\mathbb{1}\mu_e'\big)+\sum_{i=1}^{p}\big[\psi_2\circ\mathcal{A}^{(i)}\otimes\mathcal{B}\big]vec\big(\mu_e\mu_X'\big)\\
+\sum_{i=1}^{p}\big[\psi_2\circ\mathcal{A}^{(-,i)}\otimes\mathcal{B}\big]vec\big(\mu_e\mathbb{1}'\big)+\big[\psi_2\circ\mathcal{B}\otimes\mathcal{B}+\psi_1\circ\mathcal{B}\otimes\mathbb{1}_{K\times K}\big]vec\big(\mu_e\mu_e'\big)
\end{aligned}
$$

□

Even the formula has no matrix form, the auxiliary matrices ψ, ψ_1 and ψ_2 indicate, which parts have influence in the diagonal entries of the variance matrix Σ_e or not. All sums with Hadamard product ψ influence the off- diagonal entries and therefore the cross - dependence, whereas ψ_1 and ψ_2 just have influence on the diagonal entries.

Besides, the above representation of Theorem A.1 and Theorem 3.7 for gbVAR(p) processes can be transferred into each other by utilizing the following: When re-

vectorizing the above sums, it is

$$(\psi \circ \mathcal{A}^{(i_2)} \otimes \mathcal{A}^{(i_1)})vec(\mu_X \mu_X') = vec(\mathbb{1}_{K\times K} - I_K) \circ \left[(\mathcal{A}^{(i_2)} \otimes \mathcal{A}^{(i_1)})vec(\mu_X \mu_X')\right]$$

$$(\psi_1 \circ \mathcal{A}^{(i)} \otimes \mathbb{1}_{K\times K})vec(\Gamma_X(0)) = vec(I_K) \circ \left[(\mathcal{A}^{(i)} \otimes \mathbb{1}_{K\times K})vec(\Gamma_X(0))\right]$$

$$(\psi_2 \circ \mathcal{A}^{(i_2)} \otimes \mathcal{A}^{(i_1)})vec(\mu_X \mu_X') = -vec(I_K) \circ \left[(\mathcal{A}^{(i_2)} \otimes \mathcal{A}^{(i_1)})vec(mu_X \mu_X')\right]$$

and accordingly for the other terms. Additional, by calculation rules for the Kronecker and Hadamard product, it is

$$vec\,(D \circ (ABC)) = vec(D) \circ \left[C' \otimes A\right] vec(B)$$

for square matrices A, B, C, D. This in turn gives again the matrix representation.