technische universität
dortmund

fakultät
statistik

Double Generalized Linear Models for Spatio-Temporal Data: A Unified Modeling Framework

Dissertation

by

Steffen Maletz

in partial fulfilment of the requirements for the degree of
Doktor der Naturwissenschaften
(Dr. rer. nat.)
presented to the

Department of Statistics
TU Dortmund University

Dortmund, 2026

Referees:
Prof. Dr. Roland Fried
Prof. Dr. Andreas Groll

Date of oral examination:
April 28, 2026

Acting Dean:
Prof. Dr. Philipp Doebler

Supervisors:
Prof. Dr. Roland Fried
Prof. Dr. Konstantinos Fokianos

Disclaimer

Use of AI tools: Parts of this dissertation were assisted by AI-based language models (e.g., ChatGPT, Claude). These tools were used exclusively for language revision, structure, and debugging. All scientific content, analyses, and conclusions originate from the author.

Previously published material: Parts of this dissertation have already been published in the following publication and preprint:

- Maletz, S., Fokianos, K., & Fried, R. (2024). Spatio-Temporal Count Autoregression. *Data Science in Science*, 3(1). DOI: 10.1080/26941899.2024.2425171
- Maletz, S., Fokianos, K., & Fried, R. (2026). *glmSTARMA – An R-Package for fitting autoregressive spatio-temporal models following generalized linear models*. arXiv:2607.08276.

The first applies to the PSTARMA models presented in Chapter 3, as well as the corresponding simulation results in Section 7.1 and results in the appendix. The second applies to Chapter 5, which introduces the Spatio-temporal Double Generalized Linear Models, along with the R package `glmSTARMA` described in Chapter 6. It also covers corresponding simulation results (Section 7.3), the data examples (Chapter 8), and additional results presented in the appendix. Note that the preprint was published after the dissertation was reviewed; due to some changes in parts of the models, some of the results reported there differ from those presented here.

Abstract

> "*I would have written a shorter letter, but I did not have the time.*"
>
> — Blaise Pascal

Spatio-temporal data arises when a variable is measured at multiple locations over time. Modeling such data poses several challenges: In addition to spatial factors, such as the position of sensors or regions, and temporal trends, the data may also be subject to autoregressive dynamics, as is the case with infectious diseases. Furthermore, additional data characteristics, such as integer values or non-negativity, must be taken into account in the modeling process.

This dissertation contributes to the state of research on the modeling of spatio-temporal data on a fixed spatial grid, with a particular focus on non-normal data. We adapt multivariate count data time series models for the spatio-temporal domain. Assuming marginal Poisson distributions for observations conditional on the past, the models describe the conditional expected value by past observations of the same and neighboring locations, a latent feedback process, and optional covariates.

Building on this foundation, we investigate an approach to detect persistent level shifts within this framework. Such level shifts may be caused by political measures, campaigns, or other extraordinary events. Methods from univariate intervention analysis in count time series models are further developed to test for level shifts at unknown times or unknown locations. We propose an iterative procedure to detect multiple intervention effects simultaneously.

Subsequently, the modeling framework is extended by using the idea of double generalized linear models (DGLMs), relaxing the assumption of a marginal Poisson distribution for the conditional observations in favor of arbitrary distributions from the exponential dispersion family. In addition to the conditional expectation, this allows the dispersion parameter to be described by a spatio-temporal model, thereby accommodating overdispersion and other data characteristics. The full framework is implemented in the R package `glmSTARMA`.

The theoretical properties and usefulness of the models are studied in simulations. We illustrate the application using data examples.

Acknowledgements

I would like to thank everyone who has supported and accompanied me throughout the past years.

My special thanks go to Roland and Kostas for their continuous supervision since my Master's thesis. Your insightful feedback, guidance, and encouragement have been invaluable and made this work possible. I am also very grateful to Andreas for agreeing to serve as a reviewer of this dissertation and for his advice and support throughout my studies.

I would like to thank my family and friends for their constant encouragement, patience, and understanding during this journey. In particular, I want to thank Sheila, Jacob, and Maxime. Since the very beginning of our studies, we have shared challenges, successes, and countless memories. I am truly grateful for our friendship and for being able to complete this chapter together.

Finally, I would like to acknowledge the university sports program for providing an important balance to academic work. The diverse activities and the community that formed around them, especially the "Sportis", offered a welcome opportunity to recharge and stay connected beyond research. I'm already excited about the next "class trip".

Contents

1 Introduction

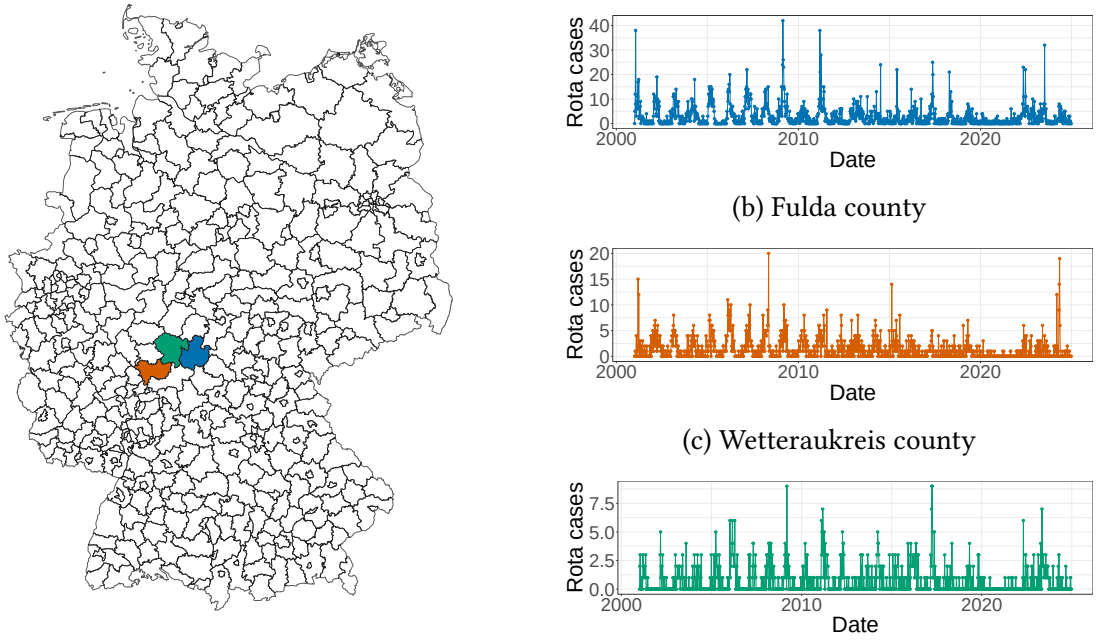
> "*The beginning is the most important part of the work.*"
>
> — Plato

Spatio-temporal data record how a variable evolves over time at a set of fixed locations. Such data arise in many scientific fields. Economic examples include modeling real-estate prices (Otto and Schmid, 2018; Pace et al., 2000) and macroeconomic relationships (Elhorst and Emili, 2022). In the environmental sciences, spatio-temporal models are used to predict water quality in river networks (Clement et al., 2006) or to model urban heat dynamics (Choi et al., 2021). In epidemiology, weekly reported infection counts from different regions are collected and modeled to study disease spread (Dickson et al., 2020; Held et al., 2005). Other applications include analyses of daily hospital admissions in Portugal (Martins et al., 2023), spatio-temporal crime modeling (Clark and Dixon, 2021), and traffic flow modeling (Kamarianakis and Prastacos, 2005).

In practice, users typically prefer models that (i) describe the data with relatively few unknown parameters and (ii) allow potential explanatory variables to be identified and assessed statistically. Excluding spatio-temporal point processes, two main model classes are commonly used; see Habereder et al. (2025) for a detailed review. One class models the observations through covariates and an additive error term, where the error itself follows a spatio-temporal process. The other class augments covariates with autoregressive components implemented via lag structures. These lag-based models incorporate past values and/or values from neighboring locations by using, for example, spatial weight matrices. These are also common in spatial statistics; see Anselin (1988).

Figure 1.1 shows weekly rotavirus infection counts from three German districts, an example well suited for lag-based models. While the time series are characterized by common seasonal patterns, the absolute number of cases varies considerably from region to region, which can mainly be explained by different population sizes. The biological parameters of the virus, i.e., in particular the infective phase lasting several days and an incubation period of one to three days[1] indicate temporal lags: current new infections relate directly to recent infection rates. Spatial dependence complements this dynamic

[1] Robert Koch-Institut (2019) RKI-Ratgeber. Rotaviren-Gastroenteritis., accessed on 03.02.2026



(a) Urban and rural districts in Germany

(b) Fulda county

(c) Wetteraukreis county

(d) Vogelbergkreis county

Figure 1.1: Urban and rural districts in Germany and weekly number of Rota cases in three example regions

as mobility and commuting can carry infection chains into neighboring regions. In this dissertation we concentrate on methodological developments for lag-based approaches that capture these spatio-temporal dependencies.

Lag-based spatio-temporal models are usually extensions of classical time-series models. The first models of this type are the Space–Time Autoregressive Moving Average (STARMA) processes introduced by Cliff and Ord (1975) and Martin and Oeppen (1975) and elaborated in a sequence of papers by Deutsch and Pfeifer (1981) and Pfeifer and Deutsch (1980b,c,d, 1981a,b,c). They are a multivariate extension of the univariate autoregressive moving average (ARMA) processes that were popularized by Box and Jenkins (1970) and whose roots go back to early work by Yule (1926) and Slutzky (1937). Borovkova and Lopuhaa (2012), Hølleland and Karlsen (2020), and Otto et al. (2018) use the generalized autoregressive conditional heteroskedasticity (GARCH) processes (Bollerslev, 1986), which are used for modeling volatility, as the basis for their spatio-temporal models.

Both STARMA and spatio-temporal GARCH models typically assume (conditional) normality; consequently, they are not well suited to epidemiological count data. To address integer-valued counts, Martins et al. (2023) introduced space-time integer ARMA (STINARMA) processes as an extension of thinning-based integer ARMA (INARMA) processes, which can be traced back to McKenzie (1985) and Al-Osh and Alzaid (1987). This model class has an interpretable structure, but is limited to applications with few regions/lo-

cations due to computational demands and additionally does not allow for covariates. The widely used *hhh4* framework, introduced by Held et al. (2005) and implemented in the `surveillance` R package (Meyer et al., 2017), decomposes the conditional mean into endemic, autoregressive, and spatial components and assumes marginal Poisson or negative-binomial distributions. A limitation of standard *hhh4* implementations is that they are typically first order in time and space; higher-order extensions were proposed by Bracher and Held (2022) and implemented in the `hhh4addon` package, but modeling choices for seasonality and spatial anisotropy remain restricted. Univariate integer GARCH (INGARCH) models (Ferland et al., 2006) and log-linear count time-series models (Fokianos and Tjøstheim, 2011) describe the conditional mean via past observations and a latent feedback process; multivariate extensions have been developed by Fokianos et al. (2020).

This dissertation develops methods for modeling spatio-temporal data. First, we consider the case of count data, such as infection counts, e.g., the weekly rotavirus infections shown in Figure 1.1. For modeling, we adapt the ideas of Fokianos et al. (2020), extending their multivariate INGARCH and log-linear time series models to spatio-temporal data in a manner analogous to STARMA and STINARMA models, and additionally allowing for the inclusion of covariates. We use this framework to develop methods for detecting persistent level shifts in spatio-temporal count data. Our approach originates from intervention analysis for univariate time series (Box and Tiao, 1975), which addresses the effects of external shocks or targeted interventions such as policy measures or major events. Specifically, we adapt the methods of Fokianos and Fried (2010, 2012) to the spatio-temporal log-linear model developed earlier in this thesis. Finally, we generalize the modeling approach for count data to a framework based on Double Generalized Linear Models (DGLM) (Smyth, 1989), implemented in the R package `glmSTARMA`. By relaxing the assumption of marginal Poisson distributions and allowing various link functions, the framework can additionally estimate alternative models proposed in the literature, e.g., those of Jahn et al. (2023), for which no prior software implementation existed. We further permit spatially and temporally varying dispersion parameters, each modeled within the same spatio-temporal structure. The result is an analog of ARMA–GARCH ideas: a spatio-temporal model for the mean coupled with a spatio-temporal model for the variance. We demonstrate the utility of these methods through two data examples: the rotavirus data introduced above and Sea Surface Temperature Anomalies in the Pacific. This is supported by extensive simulation studies.

The remainder of this dissertation is organized as follows. The next chapter reviews the key statistical background used throughout this work, including the multivariate models of Fokianos et al. (2020), spatial weight matrices, DGLMs (Smyth, 1989), and intervention analysis for univariate count time series. Chapter 3 develops the theoretical framework for adapting these multivariate models to spatio-temporal count data, and Chapter 4 builds on this by describing the methodology for detecting persistent level shifts. Chapters 5 and 6 then introduce the spatio-temporal DGLMs and the R package `glmSTARMA` in which these methods are implemented, respectively. Chapter 7 presents simulation studies for all previously described methods, and Chapter 8 demonstrates their

application to the rotavirus data and Sea Surface Temperature Anomalies in the Pacific. Finally, Chapter 9 concludes with a discussion and an outlook on possible extensions. An appendix contains further results and supplementary material.

2 Basic Statistical Methods

"*Statistics is the grammar of science.*"
— Karl Pearson

In this chapter, we describe the statistical background underlying the methods developed in this dissertation. The models in Chapter 3 are based on the multivariate count time series models of Fokianos et al. (2020), which we present in Section 2.1. Throughout all spatio-temporal models considered in this thesis, we employ the concept of spatial weight matrices, which we explain in detail and illustrate with examples in Section 2.2. Section 2.3 presents intervention analysis for univariate count time series, which forms the basis for the detection of persistent level shifts in spatio-temporal count data in Chapter 4. Finally, Section 2.4 describes Double Generalized Linear Models, on which we build the spatio-temporal modeling framework introduced in Chapter 5.

2.1 Multivariate Count Time Series Models

In their work, Fokianos et al. (2020) develop multivariate extensions of univariate count time series models, specifically the INGARCH (Integer-valued Generalized Autoregressive Conditional Heteroskedasticity) models (Ferland et al., 2006) and the log-linear models (Fokianos and Tjøstheim, 2011).

The univariate INGARCH(q, r) model for a count time series $\{Y_t\}$ is defined by

$$\begin{aligned} Y_t \mid \mathcal{F}_{t-1} &\sim \text{Poisson}(\mu_t), \\ \mu_t &= \delta + \sum_{i=1}^{q} \alpha_i \mu_{t-i} + \sum_{j=1}^{r} \beta_j Y_{t-j}, \end{aligned} \tag{2.1}$$

where $\mathcal{F}_{t-1}$ denotes the information available up to time $t-1$, and μ_t represents the conditional mean. The parameters satisfy $\delta > 0$, $\alpha_i \geq 0$ for $i = 1, \ldots, q$, and $\beta_j \geq 0$ for $j = 1, \ldots, r$ to ensure positivity of the conditional mean. This model explicitly specifies the conditional expectation $\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = \mu_t$ as a function of past observations and past conditional means.

In the univariate case, a Poisson distribution is typically assumed for the observations conditionally on the past, as in the original INGARCH model. Alternatively, a negative binomial distribution can be used to accommodate overdispersion (Zhu, 2011). In the multivariate context, however, this specification alone is insufficient, as contemporaneous dependencies between the individual components of the time series must also be considered. To model these dependencies, Fokianos et al. (2020) propose a data-generating process based on copulas, which yields marginal Poisson distributions for the conditional observations while simultaneously allowing for flexible contemporaneous dependencies between the components.

We now describe the data-generating process in the multivariate case and present the formal model specifications in detail.

2.1.1 Data Generating Process

Let $\{\mathbf{Y}_t = (Y_{1,t}, \dots, Y_{p,t})', t = 0, 1, \dots, T\}$ denote a p-dimensional count time series with associated latent intensity process $\{\boldsymbol{\mu}_t = (\mu_{1,t}, \dots, \mu_{p,t})', t = 0, 1, \dots, T\}$, where

$$\boldsymbol{\mu}_t := \mathbb{E}(\mathbf{Y}_t \mid \mathcal{F}_{t-1})$$

and $\mathcal{F}_t$ denotes the σ-algebra generated by the observations $\mathbf{Y}_0, \dots, \mathbf{Y}_t$ and a given initial value $\boldsymbol{\mu}_0$.

Conditionally on $\mathcal{F}_{t-1}$, the observation vector $\mathbf{Y}_t$ is constructed using a Poisson process representation with intensity vector $\boldsymbol{\mu}_t$. Specifically, let $\{\mathbf{U}_{t,l} = (U_{t,l,1}, \dots, U_{t,l,p})', l \in \mathbb{N}\}$ be a sequence of independent random vectors from a p-dimensional copula. The components of the observation vector are then defined element-wise by

$$Y_{i,t} = \max\left\{k : \sum_{l=1}^{k} \frac{-\log(U_{t,l,i})}{\mu_{i,t}} \leq 1\right\}, \tag{2.2}$$

where $\max \emptyset := 0$. This construction ensures that the marginal conditional distributions are Poisson, i.e.,

$$Y_{i,t} \mid \mathcal{F}_{t-1} \sim \text{Poisson}(\mu_{i,t}), \qquad i = 1, \dots, p.$$

Crucially, contemporaneous dependencies between the components are introduced exclusively through the underlying copula and are thus decoupled from the temporal dynamics of the intensity process.

2.1.2 Model Equations

Having specified how observations are generated conditionally on the intensity process, we now describe the formula of the conditional mean $\boldsymbol{\mu}_t$ itself.

First-Order Models

The first-order multivariate INGARCH model is given by

$$\boldsymbol{\mu}_t = \boldsymbol{\delta} + \mathbf{A}_1 \boldsymbol{\mu}_{t-1} + \mathbf{B}_1 \mathbf{Y}_{t-1}, \tag{2.3}$$

where $\mathbf{A}_1 \in \mathbb{R}_+^{p \times p}$ and $\mathbf{B}_1 \in \mathbb{R}_+^{p \times p}$ are unknown coefficient matrices to be estimated, and $\boldsymbol{\delta} \in \mathbb{R}_+^p$ is an intercept vector. To ensure that the conditional means remain strictly positive, i.e., $\mu_{i,t} > 0$ for all i and t, the entries of $\mathbf{A}_1$ and $\mathbf{B}_1$ must be non-negative and the entries of $\boldsymbol{\delta}$ must be positive.

Alternatively, the first-order multivariate log-linear model is specified as

$$\boldsymbol{\psi}_t = \boldsymbol{\delta} + \mathbf{A}_1 \boldsymbol{\psi}_{t-1} + \mathbf{B}_1 \log(\mathbf{Y}_{t-1} + \mathbf{1}_p), \tag{2.4}$$

where $\boldsymbol{\psi}_t := \log(\boldsymbol{\mu}_t)$ and the logarithm is applied element-wise. In contrast to (2.3), no sign restrictions are required for the parameters in this specification, as positivity of the conditional means is ensured through the link function.

Higher-Order Models

Both model classes can directly be extended to higher temporal orders. The multivariate INGARCH(q, r) model is defined by

$$\boldsymbol{\mu}_t = \boldsymbol{\delta} + \sum_{i=1}^{q} \mathbf{A}_i \boldsymbol{\mu}_{t-i} + \sum_{j=1}^{r} \mathbf{B}_j \mathbf{Y}_{t-j}, \tag{2.5}$$

where $\mathbf{A}_i \in \mathbb{R}_+^{p \times p}$ for $i = 1, \ldots, q$ and $\mathbf{B}_j \in \mathbb{R}_+^{p \times p}$ for $j = 1, \ldots, r$ are unknown coefficient matrices; the same sign restrictions as in the first-order model apply. The corresponding multivariate log-linear model with temporal orders q and r is given by

$$\boldsymbol{\psi}_t = \boldsymbol{\delta} + \sum_{i=1}^{q} \mathbf{A}_i \boldsymbol{\psi}_{t-i} + \sum_{j=1}^{r} \mathbf{B}_j \log(\mathbf{Y}_{t-j} + \mathbf{1}_p). \tag{2.6}$$

These higher-order specifications allow the conditional mean at time t to depend on multiple past observations and past conditional means, thereby capturing more complex temporal dynamics.

2.1.3 Statistical Properties

For the processes to be identifiable and for their unconditional moments to exist, the true parameters of the models must satisfy certain restrictions. These conditions are also necessary to ensure consistent estimation of the unknown parameters. Fokianos et al.

(2020) establish these properties for first-order models, i.e., for (2.3) and (2.4). For higher-order models, a sufficient condition is conjectured under which the same properties are expected to hold.

Let $\mathbf{X} \in \mathbb{R}^{p\times p}$ be a matrix and denote by $\|\mathbf{X}\|_2$ its spectral norm. We have $\|\mathbf{X}\|_2 = \lambda_{\max}^{1/2}(\mathbf{X}'\mathbf{X})$, where $\lambda_{\max}(\mathbf{X}'\mathbf{X})$ denotes the largest eigenvalue of the matrix product.

For model (2.3), if $\|\mathbf{A}_1 + \mathbf{B}_1\|_2 < 1$ holds, then the joint process $\{(\mathbf{Y}_t', \boldsymbol{\mu}_t')'\}$ is stationary, its m-th moments exist for any $m > 0$, and the process is geometrically ergodic. For the higher-order model (2.5) with temporal orders q and r, it is conjectured that the condition

$$\sum_{i=1}^{\max\{q,r\}} \|\mathbf{A}_i + \mathbf{B}_i\|_2 < 1 \tag{2.7}$$

is sufficient to imply the same properties. Under stationarity, the unconditional expectation of model (2.3) is given by

$$\mathbb{E}[\mathbf{Y}_t] = \mathbb{E}[\boldsymbol{\mu}_t] = \left[\mathbf{I}_p - \mathbf{A}_1 - \mathbf{B}_1\right]^{-1} \boldsymbol{\delta}. \tag{2.8}$$

For the first-order log-linear model (2.4), the condition $\|\mathbf{A}_1\|_2 + \|\mathbf{B}_1\|_2 < 1$ is sufficient to establish the corresponding properties for the joint process $\{(\mathbf{Y}_t', \boldsymbol{\psi}_t')'\}$.

2.1.4 Parameter Estimation

For estimating the unknown parameters, Fokianos et al. (2020) recommend quasi-maximum likelihood estimation (QMLE). This approach maximizes the quasi-log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{t=\max\{q,r\}}^{T} \sum_{i=1}^{p} (Y_{i,t} \log(\mu_{i,t}(\boldsymbol{\theta})) - \mu_{i,t}(\boldsymbol{\theta})), \tag{2.9}$$

where $\boldsymbol{\theta} = (\boldsymbol{\delta}, \mathrm{vec}(\mathbf{A}_1), \ldots, \mathrm{vec}(\mathbf{A}_q), \mathrm{vec}(\mathbf{B}_1), \ldots, \mathrm{vec}(\mathbf{B}_r))'$ denotes the vector of unknown parameters and T is the length of the time series. The $\mathrm{vec}(\cdot)$ operator transforms the $p \times p$ matrices into a vector of length p^2 by stacking the columns. The number of parameters to be estimated is $p + (q + r)p^2$ for both model classes (2.5) and (2.6), thus increasing quadratically with the number of time series components.

The quasi-log-likelihood (2.9) is based on the assumption of contemporaneous independence of the conditional observations and corresponds to a Poisson log-likelihood, which follows directly from the marginal distributions specified in Subsection 2.1.1. Contemporaneous dependencies are initially disregarded. The maximization can be performed using numerical optimization algorithms. A key advantage of this quasi-log-likelihood lies in its simple differentiability, which enables the use of gradient-based optimization methods. However, neglecting contemporaneous dependencies increases the variance of

the estimator compared to using the correct log-likelihood, which fully accounts for the dependencies induced by the data-generating process.

If the true parameter vector $\boldsymbol{\theta}_{\text{true}}$ lies in the interior of the stationarity region implied by the conditions described in Subsection 2.1.3, then the QMLE estimator $\hat{\boldsymbol{\theta}}$ is strongly consistent for $\boldsymbol{\theta}_{\text{true}}$ and asymptotically normally distributed:

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{true}}) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{-1}\right) \quad \text{as} \quad T \to \infty. \tag{2.10}$$

The matrices $\mathbf{G}$ and $\mathbf{H}$ are defined via expectations with respect to the stationary distribution of $\mathbf{Y}_t$ and are replaced by their empirical counterparts in practice. For the multivariate INGARCH model, we have

$$\mathbf{G}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \boldsymbol{\mu}_t'(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \boldsymbol{\Sigma}_t(\boldsymbol{\theta}) \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\mu}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right], \tag{2.11}$$

$$\mathbf{H}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \boldsymbol{\mu}_t'(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\mu}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right], \tag{2.12}$$

where $\mathbf{D}_t = \text{diag}(\mu_{i,t}, i = 1, \dots, p)$ and $\boldsymbol{\Sigma}_t = \text{Var}(\mathbf{Y}_t \mid \mathcal{F}_{t-1})$ is the true covariance matrix of the conditional observations. For the multivariate log-linear model, we have

$$\mathbf{G}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \boldsymbol{\psi}_t'(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \boldsymbol{\Sigma}_t(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\psi}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right], \tag{2.13}$$

$$\mathbf{H}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \boldsymbol{\psi}_t'(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{D}_t(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\psi}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]. \tag{2.14}$$

The sandwich variance form $\mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{-1}$ accounts for the additional variance induced by contemporaneous dependencies. In the special case where the data-generating process employs an independence copula, i.e., when all $U_{t,l,i} \sim \mathcal{U}(0,1)$ are independent and identically distributed for all t, l, i, we have $\mathbf{D}_t(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_t(\boldsymbol{\theta})$ for all t and consequently $\mathbf{H}(\boldsymbol{\theta}) = \mathbf{G}(\boldsymbol{\theta})$ in the sandwich variance. More details about the derivatives $\frac{\partial \boldsymbol{\psi}_t'(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ will be given in the subsequent chapters.

When applying the asymptotic distribution, it should be noted that its validity requires the true parameter vector $\boldsymbol{\theta}_{\text{true}}$ to lie in the interior of the stationary region. This condition is violated in the multivariate INGARCH process when individual true parameters equal zero.

2.2 Spatial Weights

Spatial weight matrices are a fundamental tool in spatial and spatio-temporal statistics and are discussed in the standard literature (among others, see Anselin, 1988; Cressie and Wikle, 2011; Moraga, 2024; Pebesma and Bivand, 2023; Wikle et al., 2019). They encode

spatial dependencies between fixed locations or regions and enter statistical models as prior structural information, thereby influencing model behavior and flexibility.

Many models, such as network autoregressive models (Armillotta and Fokianos, 2024; Zhu et al., 2017) or the *hhh4* model (Held et al., 2005), rely on a single spatial weight matrix that is assumed to capture all relevant spatial dependencies. More flexible approaches, including STARMA processes (Pfeifer and Deutsch, 1980c), allow for higher-order spatial weight matrices. In this setting, multiple matrices can be specified to estimate effects of different strengths for distinct neighborhood orders or to account for directional dependencies. In R, the spdep package (Bivand, 2025) provides a wide range of tools for constructing such matrices. In the following, we present several simple examples to build intuition for spatial weight matrices. We distinguish between regular and irregular grids, as well as settings in which locations are embedded in a continuous coordinate system. The explanations in this section are inspired by the discussion in Anselin (2024, Chaps. 10 and 11), and extended by own ideas.

2.2.1 General Properties

A spatial weight matrix of order $\ell \in \mathbb{N}_0$ is denoted by

$$\mathbf{W}^{(\ell)} = \left(w_{ij}^{(\ell)}\right)_{i=1,\dots,p;\, j=1,\dots,p} \in \mathbb{R}^{p\times p},$$

where p denotes the number of spatial locations. There is a close relationship between a weight matrix $\mathbf{W}^{(\ell)}$ and the associated neighborhoods $\mathcal{N}_i^{(\ell)}$, $i = 1, \dots, p$, of order ℓ. On the one hand, the i-th row of $\mathbf{W}^{(\ell)}$ defines the neighborhood $\mathcal{N}_i^{(\ell)} \subset \{1, \dots, p\}$ of location i as the set of all locations with non-zero weights. More precisely,

$$\mathcal{N}_i^{(\ell)} = \left\{j \in \{1, \dots, p\} \,:\, w_{ij}^{(\ell)} \neq 0\right\}.$$

On the other hand, in practice the neighborhood structure is usually defined first, based on spatial proximity, shared boundaries, or distances between locations. The corresponding weights $w_{ij}^{(\ell)}$ are then specified according to prior knowledge about the strength of spatial dependence or other relevant characteristics.

Depending on the statistical model, additional assumptions on the weight matrices may be imposed. Most commonly, weights are required to be non-negative, that is, $w_{ij}^{(\ell)} \geq 0$ for all i, j, and ℓ. Some models further assume that weight matrices are row-normalized, meaning that $\mathbf{W}^{(\ell)}\mathbf{1}_p = \mathbf{1}_p$. This property often simplifies the derivation of stationarity conditions and other theoretical properties, such as in Pfeifer and Deutsch (1980a) and Pfeifer and Deutsch (1980d). Moreover, row-normalization has important implications for interpretation. For instance, when computing $\mathbf{W}^{(\ell)}\mathbf{Y}$, the i-th component of the resulting vector represents a weighted mean of the entries Y_j with $j \in \mathcal{N}_i^{(\ell)}$.

Analogous to temporal lags, the spatial order ℓ is intended to reflect a notion of distance. For $\ell < k$, locations in $\mathcal{N}_i^{(\ell)}$ are typically closer to location i than those in $\mathcal{N}_i^{(k)}$. This leads to further natural conditions on the neighborhood structure. First, a location should not belong to neighborhoods of different orders, that is, $\mathcal{N}_i^{(\ell)} \cap \mathcal{N}_i^{(k)} = \emptyset$ for $\ell \neq k$. Second, each location is closest to itself, which motivates defining $\mathcal{N}_i^{(0)} = \{i\}$ and, consequently, using $\mathbf{W}^{(0)} = \mathbf{I}_p$.

Many approaches to the construction of spatial weight matrices implicitly follow the *First Law of Geography* formulated by Tobler (1970):

> Everything is related to everything else, but near things are more related than distant things.

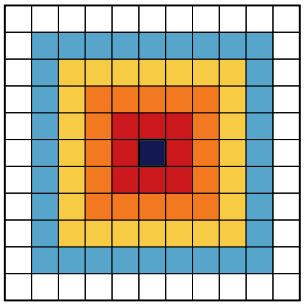
Nevertheless, deliberate deviations from this principle are possible and sometimes desirable. For example, when the objective is to model directional dependencies, spatial order may no longer represent distance in the usual sense. We explored such a setting in the Sea Surface Temperature Anomalies data example in Chapter 8 of this dissertation. There, the spatial orders $\ell \in \{1, 2, 3, 4\}$ correspond to the four cardinal directions (north, east, south, west), implying that all neighboring locations are considered to be at the same distance.

2.2.2 Neighborhood Structures on Spatial Grids and Networks

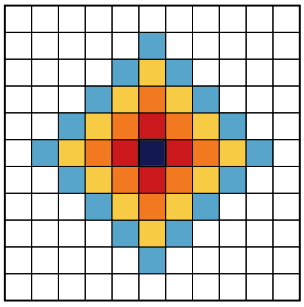
Spatial lattice structures are prevalent in many applications. Regular grids commonly appear in satellite data for meteorological applications, as in the data example of Hølleland and Karlsen (2020). Irregular grids are standard in applications examining socio-economic variables (Otto and Schmid, 2018) or infection counts (Held et al., 2005) across administrative units. Typically, directly adjacent locations are designated as first-order neighbors. On regular grids, a distinction is often made between *rook contiguity* and *queen contiguity* (Moraga, 2024, Chap. 7). These terms originate from chess, where the rook moves only horizontally and vertically, while the queen can also move diagonally. The key difference is that under queen contiguity, locations are already considered adjacent if their boundaries meet at a single point, while the rook contiguity requires multiple points.

These grids can be represented as undirected graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \dots, p\}$ and $\mathcal{E}$ is a set of location pairs $\{i, j\}$ sharing a common boundary, i.e., $\{i, j\} \in \mathcal{E} \Leftrightarrow i \leftrightarrow j$. First-order neighborhoods are naturally defined as $\mathcal{N}_i^{(1)} = \{j \in \{1, \dots, p\} \mid \{i, j\} \in \mathcal{E}\}$. Higher-order neighborhoods of order $\ell > 1$ can be defined as regions reachable by traversing $\ell - 1$ regions that are not contained in any lower-order neighborhood of the same location. This can be written as

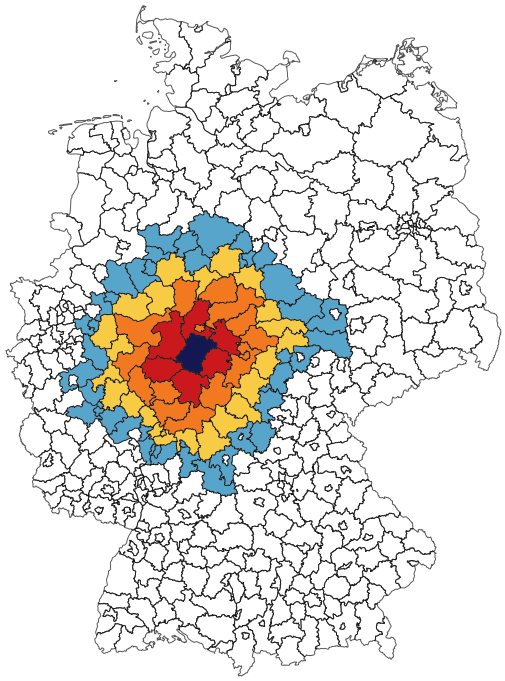
$$\mathcal{N}_i^{(\ell)} = \bigcup_{j \in \mathcal{N}_i^{(\ell-1)}} \mathcal{N}_j^{(1)} \setminus \left(\bigcup_{k=1}^{\ell-1} \mathcal{N}_i^{(k)} \cup \{i\} \right) \tag{2.15}$$



(a) Queen Contiguity



(b) Rook Contiguity



(c) Neighbors of Schwalm-Eder-Kreis County

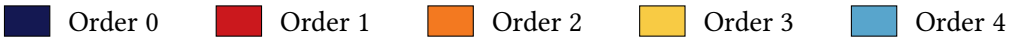


Figure 2.1: Visualization of ℓ-th order spatial neighborhoods ($\ell = 0, \dots, 4$) for one selected location. (a) Regular lattice under Queen contiguity (orthogonal + diagonal neighbors), (b) regular lattice under Rook contiguity (orthogonal neighbors), and (c) an irregular adjacency network of German counties (Landkreise). Colors encode the minimum number of adjacency steps ℓ required to reach other units from the focal location.

The spatial weights are usually defined by the number of elements in each neighborhood, i.e., $w_{ij}^{(\ell)} = 1/\#\mathcal{N}_i^{(\ell)}$. This applies to both regular and irregular grids. Figure 2.1 visualizes neighborhoods up to spatial order 4 constructed by using the principle in (2.15). In Panel 2.1a the adjacency is determined by the queen contiguity, and in Panel 2.1b the rook contiguity is used. Since queen contiguity considers regions adjacent when boundaries meet at a single point, the resulting neighborhoods are larger than those based on rook contiguity. Panel 2.1c shows neighbors of the Schwalm-Eder-Kreis county in Germany, an irregular grid, determined using the same principle (2.15). All examples demonstrate that the number of neighbors increases with higher spatial order, naturally assigning lower weights to more distant locations compared to closer neighbors.

Not all applications can adequately capture spatial dependence using undirected graphs, as neighborhood relationships between two locations imply bidirectional dependence. In such cases, directed graphs provide a better description. River networks (Clement

et al., 2006) are an example for this, as the water typically flows only in one direction. When using (2.15), note that higher-order neighborhoods can only propagate in the direction of the dependence structure. This may cause interpretation issues if directed graphs are used inadvertently. For instance, constructing a directed graph on a regular grid where only the nearest location directly to the east is considered adjacent results in all higher-order neighbors also lying directly east of the origin. Locations in, say, a north-easterly direction can never become neighbors of any order under this construction. Users should choose such constructions carefully and consider which dependencies can be captured.

2.2.3 Continuous Spatial Domains

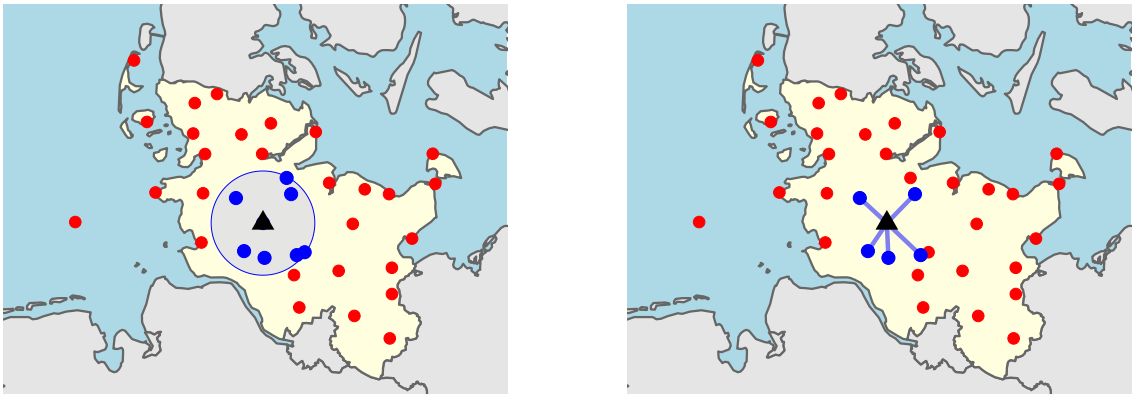
Spatio-temporal data are not always recorded on spatial grids but may instead be observed at irregularly spaced measurement stations, such as weather stations. Although we can assign unique coordinates to these stations, it is no longer intuitive which locations are adjacent to one another, and this must be explicitly defined by the user. In such cases, neighborhoods and spatial weights are typically determined based on the spatial distances between measurement stations, which can be directly computed from the coordinates.

Let $\mathbf{s}_i \in \mathbb{R}^d, i = 1, \dots, p$ denote the coordinates of the fixed measurement locations in the dataset, and let $d_{ij} := d(\mathbf{s}_i, \mathbf{s}_j) \geq 0, i, j \in \{1, \dots, p\}$ denote the distances between coordinates computed using an appropriate metric. Depending on the application, this could be the Euclidean distance, the Manhattan distance, or the Haversine distance when coordinates are placed on a sphere.

Various types of neighborhoods can be defined based on these distances. One simple neighborhood structure is the k-nearest neighbor structure. Let $d_{i(1)} \leq d_{i(2)} \leq \dots \leq d_{i(p)}$ denote the ordered distances to location i. The k-nearest neighbor neighborhood is then defined as $\mathcal{N}_i^{(1)}(k) := \{j \in \{1, \dots, p\} \mid d_{ij} \leq d_{i(k)}\}$.

Another widely used approach involves *distance band* based neighborhood definitions. In this approach, the user specifies a maximum distance $d_{\max}^{(1)}$ below which a location is considered a neighbor. For higher-order neighborhoods, multiple thresholds $d_{\max}^{(\ell)}$ can be specified. The ℓ-th order neighborhood for location i is then defined as $\mathcal{N}_i^{(\ell)} := \{j \in \{1, \dots, p\} \mid d_{\max}^{(\ell-1)} < d_{ij} \leq d_{\max}^{(\ell)}\}$ with $d_{\max}^{(0)} := 0$.

Figure 2.2 illustrates the construction of first-order neighborhoods for a given location. The data example is inspired by Jahn et al. (2023) and shows ($p = 36$) weather stations of the German Weather Service (DWD) measuring precipitation in the federal state of Schleswig-Holstein, Germany. Panel 2.2a defines all measurement stations within a 30km radius as neighbors, while Panel 2.2b defines the 5 nearest measurement stations as neighbors. Both methods require the user to specify an additional parameter that determines which locations qualify as neighbors. In regions with high station density, this yields relatively similar neighborhoods, whereas for isolated stations, substantial



(a) Neighbors within a 30 km radius.

(b) 5 nearest neighbors (5-NN).

Figure 2.2: Weather stations in the federal state of Schleswig-Holstein, Germany (highlighted in light yellow). Blue dots represent neighboring stations of the reference station marked with a black triangle. Red dots indicate non-neighboring stations.

differences may arise. The figures reveal that one measurement station is located on Helgoland, Germany's only offshore island. Under the 30km radius definition, no other station falls within this radius, whereas the 5-nearest neighbor approach selects 5 stations that are far away and may experience weather conditions substantially different from those on Helgoland.

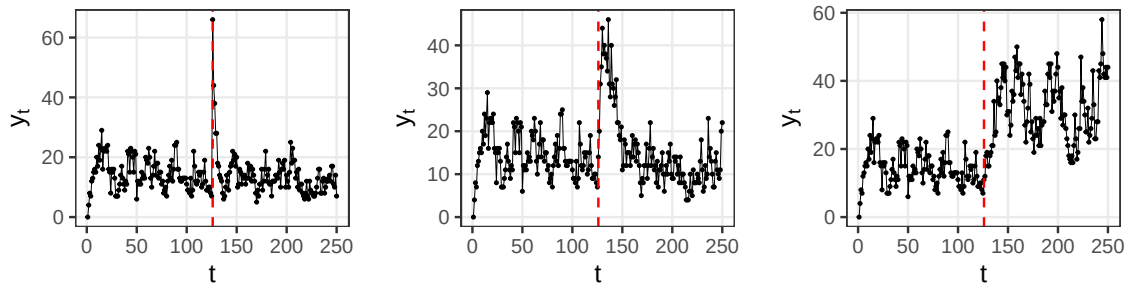
Once the neighborhood structure has been defined, the spatial weights can be chosen to be uniform, similar to grid structures, i.e., $w_{ij}^{(\ell)} = 1/\#\mathcal{N}_i^{(\ell)}$. More commonly, however, they are defined based on inverse or exponentially weighted distances d_{ij}, for instance as

$$w_{ij}^{(\ell)} = \frac{\exp(-d_{ij}\alpha)}{\sum_{k \in \mathcal{N}_i^{(\ell)}} \exp(-d_{ik}\alpha)}, \quad \text{or} \quad w_{ij}^{(\ell)} = \frac{d_{ij}^{-\alpha}}{\sum_{k \in \mathcal{N}_i^{(\ell)}} d_{ik}^{-\alpha}},$$

where $\alpha > 0$ is a decay parameter. These weighting schemes can also be applied to grid structures if distance measures are defined on the graph, for example via shortest path length.

2.3 Intervention Analysis

Intervention analysis, as introduced by Box and Tiao (1975), aims to quantify the effect of an external event on the level of a time series. For count-valued time series, for example, weekly reported infection counts, standard methods may be inappropriate because they do not explicitly account for the discreteness and non-negativity of the observations. Fokianos and Fried (2010, 2012) developed an approach for intervention analysis for univariate INGARCH processes, see Equation (2.1), and for log-linear models. They



(a) Single Outlier (SO) (b) Transient Shift (TS) (c) Level Shift (LS)

Figure 2.3: Simulated trajectories for different intervention types in a first-order log-linear model ($\delta = 0.5$, $\alpha_1 = 0.2$, $\beta_1 = 0.6$) with $\tau = 126$. Intervention magnitudes are $\gamma = 2$ (SO), $\gamma = 0.5$ with $\eta = 0.9$ (TS), and $\gamma = 0.2$ (LS).

represent the effects of m interventions by m covariates $x_{1,t}, \dots, x_{m,t}$ that enter the linear process additively. In particular, for the log-linear model with

$$Y_t \mid \mathcal{F}_{t-1} \sim \text{Poisson}\big(\exp(\psi_t)\big),$$

they consider the linear predictor

$$\psi_t = \delta + \sum_{i=1}^{q} \alpha_i \psi_{t-i} + \sum_{j=1}^{r} \beta_j \log(Y_{t-j} + 1) + \boldsymbol{\gamma}^\top \mathbf{x}_t, \tag{2.16}$$

where $\mathbf{x}_t = (x_{1,t}, \dots, x_{m,t})^\top$ is the vector of intervention covariates and $\boldsymbol{\gamma} \in \mathbb{R}^m$ is the vector of corresponding coefficients which measure the magnitudes of the intervention effects.

The authors propose using covariates of the form

$$x_t = (1 - \eta \mathcal{B})^{-1} 1(t = \tau),$$

where $\mathcal{B}$ denotes the backshift operator and $1(\cdot)$ is the indicator function. The parameter $\eta \in [0, 1]$ controls the temporal pattern of the intervention. When $\eta = 0$, the covariate corresponds to a single impulse at the intervention time τ (a *single outlier*), while $\eta = 1$ produces a persistent level shift from time τ onward. Values $\eta \in (0, 1)$ give a transient shift that decays geometrically. Note that, depending on the autoregressive parameters in (2.16), even a short-lived shock can produce effects that persist for several periods. Figure 2.3 illustrates these three intervention types in a simulated trajectory from a first-order log-linear model. For the single outlier and the transient shift the maximum effect is attained immediately at τ, whereas for the level shift the effect typically builds up over time due to the model dynamics.

2.3.1 Testing for Intervention Effects

A common empirical question is whether an intervention effect is present. For the single-intervention case ($m = 1$) this corresponds to test

$$H_0 : \gamma = 0 \quad \text{versus} \quad H_1 : \gamma \neq 0.$$

If the intervention time τ and the intervention type (i.e., the value of η) are known, the corresponding covariate may simply be included in the model and γ estimated and tested in the usual way. When one or more aspects of the intervention are unknown, Fokianos and Fried (2010, 2012) advocate a score-test approach.

In the score-test procedure the model is first fitted under the null hypothesis (no intervention). Let

$$\hat{\boldsymbol{\theta}}_0 = (\hat{\delta}, \hat{\alpha}_1, \dots, \hat{\alpha}_q, \hat{\beta}_1, \dots, \hat{\beta}_r)^\top$$

be the maximum-likelihood estimator of the model parameters under H_0, and let $(\hat{\boldsymbol{\theta}}_0, 0)^\top$ denote the augmented parameter vector with $\gamma = 0$. For the Poisson log-linear model the score for γ evaluated under H_0 simplifies to

$$S_\gamma(\hat{\boldsymbol{\theta}}_0, 0) = \left.\frac{\partial \ell(\boldsymbol{\theta}, \gamma)}{\partial \gamma}\right|_{(\hat{\boldsymbol{\theta}}_0, 0)} = \sum_{t=1}^{T} \left(y_t - \lambda_t(\hat{\boldsymbol{\theta}}_0, 0)\right) \left.\frac{\partial \psi_t((\boldsymbol{\theta}, \gamma))}{\partial(\boldsymbol{\theta}, \gamma)}\right|_{(\hat{\boldsymbol{\theta}}_0, 0)}$$

The scalar test statistic is then

$$T(\eta, \tau) \;=\; \frac{S_\gamma(\hat{\boldsymbol{\theta}}_0, 0)^2}{\widehat{\mathrm{Var}}\big(S_\gamma(\hat{\boldsymbol{\theta}}_0, 0)\big)},$$

where the variance in the denominator can be obtained from the Schur complement of the Fisher information matrix. Writing the information matrix as

$$\mathbf{H}(\boldsymbol{\theta}, 0) = \begin{pmatrix} \mathbf{H}_{\theta\theta} & \mathbf{H}_{\theta\gamma} \\ \mathbf{H}_{\gamma\theta} & \mathbf{H}_{\gamma\gamma} \end{pmatrix},$$

an estimate of the variance of the score is

$$\widehat{\mathrm{Var}}(S_\gamma) = \widehat{\mathbf{H}}_{\gamma\gamma} - \widehat{\mathbf{H}}_{\gamma\boldsymbol{\theta}} \widehat{\mathbf{H}}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} \widehat{\mathbf{H}}_{\boldsymbol{\theta}\gamma},$$

with the components of the estimated information calculated under H_0 as

$$\widehat{\mathbf{H}}_{\boldsymbol{\nu}\boldsymbol{\nu}} = \sum_{t=1}^{T} \frac{1}{\lambda_t(\hat{\boldsymbol{\theta}}_0, 0)} \left(\frac{\partial \lambda_t}{\partial \boldsymbol{\nu}}\right) \left(\frac{\partial \lambda_t}{\partial \boldsymbol{\nu}}\right)^\top, \qquad \boldsymbol{\nu} \in \{\boldsymbol{\theta}, \gamma\}.$$

Under regularity conditions and under H_0, $T(\eta, \tau)$ is asymptotically χ^2_1-distributed.

When the intervention time τ is unknown, a grid search is used. Let

$$\mathcal{T} = \{\tau_{\min}, \dots, \tau_{\max}\}$$

be a set of candidate time points. The score statistic $T(\eta, \tau)$ is computed for each $\tau \in \mathcal{T}$ and the maximum-score test statistic is

$$\max_{\tau \in \mathcal{T}} T(\eta, \tau),$$

while an estimator of the intervention time is given by

$$\hat{\tau} \;=\; \arg\max_{\tau \in \mathcal{T}} T(\eta, \tau).$$

Fokianos and Fried (2010, 2012) recommend using a parametric bootstrap to obtain critical values for the maximum-score statistic. We omit this procedure here because it is computationally intensive and not used elsewhere in this dissertation. In Chapter 4 we will use a wild bootstrap to determine the distribution of the test statistic. Fokianos and Fried (2010, 2012) also discuss an approach to detect multiple interventions. This will not be treated here, as we cannot transfer this procedure exactly to spatio-temporal data. Instead we will develop a different extension of the maximum-score test approach, which will be described in Chapter 4.

2.4 Double Generalized Linear Models

Smyth (1989) introduced the class of Double Generalized Linear Models (DGLMs) as an extension of the Generalized Linear Models (GLMs) presented by Nelder and Wedderburn (1972). While the main interest in GLMs lies in modeling the expected value of the observations, DGLMs assume an additional second model for the so-called dispersion parameter of the underlying distribution. In GLMs, this is assumed to be a fixed, possibly unknown value. The observations are assumed to follow a distribution from the Exponential Dispersion Model Family (Jørgensen, 1987).

2.4.1 Exponential Dispersion Models

Let Y be a univariate random variable. We say that Y belongs to the family of Exponential Dispersion Models (EDMs) (Jørgensen, 1987) if the probability density function with respect to an appropriate dominating measure can be written as

$$f(y; \theta, \phi) = \exp\left(\frac{y\theta - A(\theta)}{\phi} + c(\phi, y)\right), \tag{2.17}$$

where $\theta \in \Theta \subseteq \mathbb{R}$ is the canonical parameter, $\phi > 0$ is the dispersion parameter, and $c(\cdot, \cdot)$, $A(\cdot)$ are known functions. The function $A : \Theta \to \mathbb{R}$ is the *log-partition function* and is twice continuously differentiable.

We write $Y \sim \text{ED}(\mu, \phi)$, where $\mu := \mathbb{E}(Y) = A'(\theta)$ is the expected value of the distribution. From the properties of the the function A, it follows that

$$\text{Var}(Y) = \phi A''(\theta) = \phi V(\mu), \tag{2.18}$$

where $V(\mu) := A''(\theta(\mu))$ is the variance function. Classical examples of EDMs include the normal distribution with $V(\mu) = 1$, the Poisson distribution with $V(\mu) = \mu$, and the gamma distribution with $V(\mu) = \mu^2$.

Since $A'(\cdot)$ is strictly monotone, there exists a bijective mapping between the canonical parameter θ and the expected value μ. The probability density function can therefore equivalently be parametrized in terms of μ:

$$f(y; \mu, \phi) = b(y, \phi) \exp\left(-\frac{1}{2\phi} d(y, \mu)\right), \tag{2.19}$$

where $d(y, \mu)$ is the distribution-specific *unit deviance*. For the normal distribution, $d(y, \mu) = (y - \mu)^2$, for the Poisson distribution $d(y, \mu) = 2[y \log(y/\mu) - (y - \mu)]$, and for the gamma distribution $d(y, \mu) = 2[-\log(y/\mu) + (y - \mu)/\mu]$.

2.4.2 Model Structure

Let $Y_1, \ldots, Y_n$ be independent random variables with $Y_i \sim \text{ED}(\mu_i, \phi_i)$ for $i = 1, \ldots, n$. A DGLM consists of two coupled regression models:

Mean model: The expected value μ_i is linked to a linear predictor via a strictly monotone, twice continuously differentiable link function $g : \mathbb{R} \to \mathbb{R}$:

$$g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}, \tag{2.20}$$

where $\mathbf{x}_i \in \mathbb{R}^k$ is a vector of covariates and $\boldsymbol{\beta} \in \mathbb{R}^k$ is the unknown parameter vector.

Dispersion model: Analogously, the dispersion parameter ϕ_i is described by a second model:

$$\tilde{g}(\phi_i) = \mathbf{z}_i^\top \boldsymbol{\gamma}, \tag{2.21}$$

with a link function $\tilde{g} : \mathbb{R}_{>0} \to \mathbb{R}$, a covariate vector $\mathbf{z}_i \in \mathbb{R}^q$, and the parameter vector $\boldsymbol{\gamma} \in \mathbb{R}^q$. Frequently, $\tilde{g}(x) = \log(x)$ is chosen to guarantee the positivity of ϕ_i.

This dual structure enables the modeling of heterogeneous variances and is particularly relevant for distributions such as the Poisson distribution. While theoretically $\phi = 1$ holds there, the DGLM approach allows explicit modeling of over- or underdispersion through $\text{Var}(Y_i) = \phi_i \mu_i$ with $\phi_i \neq 1$. This corresponds to a quasi-likelihood approach in which the variance structure is designed more flexibly than in the canonical distribution. The resulting distribution can be interpreted as an EDM, even if it does not necessarily correspond to a known distributional family.

2.4.3 Parameter Estimation

The joint estimation of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is typically performed through an iterative algorithm that alternates between the estimation of both models (Smyth, 1989). The central idea is that with fixed values of the other model, each submodel can be treated as an ordinary GLM.

The estimation of the dispersion model is based on *pseudo-observations* $d_i := d(Y_i, \hat{\mu}_i)$, the empirical unit deviances. Under the assumption of a correctly specified mean model, $\mathbb{E}(d_i \mid \mu_i) \approx \phi_i$ holds, and asymptotically these follow approximately a gamma distribution with constant dispersion parameter 2, that is, $d_i \sim \text{Gamma}(\phi_i/2, 2)$.

The algorithm proceeds as follows:

1. **Initialization:** Set initial values $\phi_i^{(0)} = \phi^{(0)}$ for all i, for example through a constant estimate.
2. **Mean estimation:** Given $\phi_i^{(t)}$, estimate $\boldsymbol{\beta}^{(t+1)}$ by maximum likelihood estimation of the GLM with weights $w_i = 1/(\phi_i^{(t)} V(\mu_i))$. Compute fitted values $\hat{\mu}_i^{(t+1)} = g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta}^{(t+1)})$.
3. **Dispersion estimation:** Compute pseudo-observations $d_i^{(t+1)} = d(Y_i, \hat{\mu}_i^{(t+1)})$ and estimate $\boldsymbol{\gamma}^{(t+1)}$ through a gamma GLM with link function $\tilde{g}$ and dispersion parameter 2. Compute $\hat{\phi}_i^{(t+1)} = \tilde{g}^{-1}(\mathbf{z}_i^\top \boldsymbol{\gamma}^{(t+1)})$.
4. **Convergence check:** Iterate until convergence according to a predetermined convergence criterion.

3 Spatio-Temporal Count Autoregression

> "*The past is never dead. It's not even past.*"
> — William Faulkner

As we treat spatio-temporal data as multivariate time series, standard multivariate time series models are a natural starting point. For count data, these include the multivariate models discussed in Section 2.1. In their generic formulation, however, such models involve $\mathcal{O}(p^2)$ unknown parameters, i.e. the parameter count grows quadratically with the number of locations, rendering them impractical for realistic spatio-temporal applications.

The STARMA approach of Pfeifer and Deutsch (1980c) addresses this by replacing the full parameter matrices with linear combinations of known, pre-specified matrices whose coefficients are scalar unknowns. In Maletz et al. (2024) we applied this principle to the multivariate count time series models of Fokianos et al. (2020) and further extended them to accommodate external covariates. This chapter describes the resulting methodology.

Section 3.1 introduces the model assumptions and derives the spatio-temporal modifications for both the linear and the log-linear time series models. Section 3.2 establishes the corresponding statistical properties, and Section 3.3 presents the adapted parameter estimation procedure.

3.1 Model Equations

Let $\{\mathbf{Y}_t = (Y_{1,t}, \dots, Y_{p,t})' \mid t = 0, 1, \dots\}$ denote a p-dimensional count time series, and let $\{\mathbf{X}_{k,t} = (X_{1,k,t}, \dots, X_{p,k,t})' \mid t = 0, 1, \dots\}$, $k = 1, \dots, m$, be covariate processes observed simultaneously with $\{\mathbf{Y}_t\}$. Denote by $\mathcal{F}_{t-1}$ the σ-algebra generated by all observations and covariates up to time $t-1$, including an initial value $\boldsymbol{\mu}_0$, and let $\{\boldsymbol{\mu}_t\}$ be the associated latent intensity process.

Following Fokianos et al. (2020), we assume that, conditional on $\mathscr{F}_{t-1}$, each marginal component follows a Poisson distribution,

$$Y_{i,t} \mid \mathscr{F}_{t-1} \sim \text{Poisson}(\mu_{i,t}).$$

Each univariate component $\{Y_{i,t}\}$, $\{\mu_{i,t}\}$, and $\{X_{i,k,t}\}$ represents the trajectory of a variable at one of p fixed locations or regions. Contemporaneous dependence between the components of $\mathrm{Y}_t \mid \mathscr{F}_{t-1}$ is captured through a copula, as described in Subsection 2.1.1. As an example, recall the weekly rota virus infections from the introduction: The counts shown in the Panels 1.1b, 1.1c, and 1.1d are three of the in total 411 components $\{Y_{i,t}\}$ of the multivariate time series containing these infections count for each region in Germany.

Spatial structure is encoded via a collection of weight matrices $\mathbf{W}^{(\ell)}$, $\ell = 0, \dots, \ell_{\max}$, constructed as described in Section 2.2. We assume these matrices are row-normalised, i.e. $\mathbf{W}^{(\ell)}\mathbf{1}_p = \mathbf{1}_p$, which simplifies the stationarity conditions in Section 3.2. The zeroth-order matrix is the identity, $\mathbf{W}^{(0)} = \mathbf{I}_p$.

3.1.1 Linear Poisson-STARMA Models

Replacing the freely estimated matrices of the linear model (2.5) with linear combinations of the weight matrices yields the linear Poisson spatio-temporal model, in which the latent intensity evolves as

$$\boldsymbol{\mu}_t = \boldsymbol{\delta} + \sum_{i=1}^{q} \sum_{\ell=0}^{a_i} \alpha_{i,\ell}\, \mathbf{W}^{(\ell)} \boldsymbol{\mu}_{t-i} + \sum_{j=1}^{r} \sum_{\ell=0}^{b_j} \beta_{j,\ell}\, \mathbf{W}^{(\ell)} \mathrm{Y}_{t-j}. \tag{3.1}$$

Here, $\boldsymbol{\delta} \succeq \mathbf{0}$ is the intercept vector; r and q denote the maximum temporal lags for regression on past observations Y_{t-j} and past intensities $\boldsymbol{\mu}_{t-i}$, respectively; and a_i, b_j are the corresponding maximum spatial orders. Non-negativity of all scalar parameters, $\alpha_{i,\ell} \geq 0$ and $\beta_{j,\ell} \geq 0$, guarantees $\mu_{i,t} > 0$.

The total number of free parameters in (3.1) is $p + \sum_{i=1}^{q} a_i + \sum_{j=1}^{r} b_j$, a substantial reduction compared with the $\mathcal{O}(p^2)$ parameters of (2.5). A further reduction is achieved by imposing a common intercept $\boldsymbol{\delta} = \delta_0 \mathbf{1}_p$; we call the model *homogeneous* in this case and *inhomogeneous* otherwise. The linear Poisson network autoregression (PNAR) model of order 1 due to Armillotta and Fokianos (2024) corresponds to the homogeneous special case with $b_1 = 1$ and $\alpha_{1,\ell} = 0$ for all ℓ.

Following STARMA terminology, we refer to model (3.1) as a linear PSTARMA($q_{a_1,\dots,a_q}, r_{b_1,\dots,b_r}$) model. The parameter $\alpha_{i,0}$ (resp. $\beta_{j,0}$), associated with $\mathbf{W}^{(0)} = \mathbf{I}_p$, captures the self-excitation at each location, while parameters at higher spatial orders $\ell > 0$ quantify the influence of neighbouring locations as defined by $\mathbf{W}^{(\ell)}$.

Linear PSTARMA models are attractive for several reasons: they are analytically tractable, admit a transparent interpretation since all variables enter the conditional expectation linearly, and are well-suited for positively correlated count data. Their main limitation is precisely this restriction to positive dependence, which motivates the log-linear extension introduced next.

3.1.2 Log-Linear Poisson-STARMA Models

The log-linear variant connects the conditional mean to a linear process via a logarithmic link function. Setting $\boldsymbol{\psi}_t = \log(\boldsymbol{\mu}_t)$ element-wise, the latent process evolves on the log scale as

$$\boldsymbol{\psi}_t = \boldsymbol{\delta} + \sum_{i=1}^{q} \sum_{\ell=0}^{a_i} \alpha_{i,\ell}\, \mathbf{W}^{(\ell)} \boldsymbol{\psi}_{t-i} + \sum_{j=1}^{r} \sum_{\ell=0}^{b_j} \beta_{j,\ell}\, \mathbf{W}^{(\ell)} \log(\mathbf{Y}_{t-j} + \mathbf{1}_p). \tag{3.2}$$

This formulation relaxes the sign restrictions on the parameters: because $\boldsymbol{\psi}_t$ is unrestricted, the model can capture both positive and negative dependence structures. In addition, covariates may be included without imposing positivity on their regression coefficients.

3.1.3 Covariates

In practice, count processes are often driven by external factors beyond their own history, for instance, demographic variables such as population size or unemployment rates, or environmental variables such as temperature or air pollution. Covariates $\{\mathbf{X}_{k,t}\}$ can be incorporated in both the linear and log-linear specifications, giving

$$\boldsymbol{\mu}_t = \boldsymbol{\delta} + \sum_{i=1}^{q} \sum_{\ell=0}^{a_i} \alpha_{i,\ell}\, \mathbf{W}^{(\ell)} \boldsymbol{\mu}_{t-i} + \sum_{j=1}^{r} \sum_{\ell=0}^{b_j} \beta_{j,\ell}\, \mathbf{W}^{(\ell)} \mathbf{Y}_{t-j} + \sum_{k=1}^{m} \sum_{\ell=0}^{c_k} \gamma_{k,\ell}\, \mathbf{W}^{(\ell)} \mathbf{X}_{k,t}, \tag{3.3}$$

and analogously

$$\begin{aligned} \boldsymbol{\psi}_t = \boldsymbol{\delta} &+ \sum_{i=1}^{q} \sum_{\ell=0}^{a_i} \alpha_{i,\ell}\, \mathbf{W}^{(\ell)} \boldsymbol{\psi}_{t-i} + \sum_{j=1}^{r} \sum_{\ell=0}^{b_j} \beta_{j,\ell}\, \mathbf{W}^{(\ell)} \log(\mathbf{Y}_{t-j} + \mathbf{1}_p) \\ &+ \sum_{k=1}^{m} \sum_{\ell=0}^{c_k} \gamma_{k,\ell}\, \mathbf{W}^{(\ell)} \mathbf{X}_{k,t}, \end{aligned} \tag{3.4}$$

where $c_k \in \mathbb{N}_0$ is the maximum spatial order for the k-th covariate. In (3.3), non-negativity of the covariates and of $\gamma_{k,\ell} \geq 0$ is required to ensure $\mu_{i,t} > 0$; no such restriction applies in (3.4). By analogy with ARMAX processes, we refer to models of these forms as linear or log-linear PSTARMAX($q_{a_1,\ldots,a_q}, r_{b_1,\ldots,b_r}, m_{c_1,\ldots,c_m}$) models.

3.2 Statistical Properties

We discuss three aspects of the model's statistical properties: identifiability of the parameters (Subsection 3.2.1), the connection to classical STARMA processes (Subsection 3.2.2), and conditions for stationarity and ergodicity (Subsection 3.2.3). Second order properties, such as the unconditional mean are included in the Appendix in Section A.1.2. They are similar to those of the models described in Section 2.1.

3.2.1 Identifiability

Identifiability of the model parameters is not automatically guaranteed and must be verified by the user.

Regarding the weight matrices, we adopt the standard conditions for STARMA processes (Pfeifer and Deutsch, 1980c): the main diagonals of $\mathbf{W}^{(\ell)}$ are zero for all $\ell > 0$, and the matrices are linearly independent of each other, i.e. $\mathbf{W}^{(\ell)} \neq \sum_{\tilde{\ell}\neq\ell} c_{\tilde{\ell}}\mathbf{W}^{(\tilde{\ell})}$ for all ℓ and all $c_{\tilde{\ell}} \in \mathbb{R}$. Spatial lagging, which assigns $w_{i,j}^{(\ell)} = 0$ whenever $w_{i,j}^{(\tilde{\ell})} \neq 0$ for some $\tilde{\ell} \neq \ell$, is one way to satisfy this condition.

Beyond the weight matrices, the same identifiability requirements as in Fokianos et al. (2020) apply. In particular, models (3.1) and (3.2) are identifiable only if $\beta_{j,\ell} \neq 0$ for at least one pair (j, ℓ); models that regress solely on the feedback process $\{\boldsymbol{\mu}_t\}$ or $\{\boldsymbol{\psi}_t\}$ are not identifiable.

For PSTARMAX models (3.3) and (3.4), identifiability additionally requires that the design matrix of observed regressors has full rank. This matrix is formed, column by column, from the spatially lagged past observations $\mathbf{W}^{(\ell)}\mathbf{Y}_{t-j}$ and the spatially lagged covariates $\mathbf{W}^{(\ell)}\mathbf{X}_{k,t}$, augmented by $\mathbf{1}_p$ (homogeneous model) or $\mathbf{I}_p$ (inhomogeneous model). Two practically important special cases deserve attention.

Spatially constant covariates. If a covariate is purely temporal, i.e. $\mathbf{X}_{k,t} = \tilde{X}_{k,t} \cdot \mathbf{1}_p$ for a univariate process $\{\tilde{X}_{k,t}\}$, then spatial orders $c_k > 0$ lead to non-identifiability, since $\sum_{\ell=0}^{c_k} \gamma_{k,\ell}\mathbf{W}^{(\ell)}\mathbf{X}_{k,t} = \tilde{\gamma}_{k,0}\,\mathbf{X}_{k,t}$ with $\tilde{\gamma}_{k,0} = \sum_\ell \gamma_{k,\ell}$. Such covariates must be included with $c_k = 0$, so that $\gamma_{k,0}$ represents a global, spatially uniform effect.

Time-constant covariates. Covariates that do not vary over time, $\mathbf{X}_{k,t} \equiv \tilde{\mathbf{X}}_k$, are collinear with the intercept $\boldsymbol{\delta}$: in an inhomogeneous model, their effects cannot be separated from the location-specific intercepts. Any effect that is constant in both space and time is already absorbed by the intercept.

3.2.2 Relationship to STARMA Processes

The models introduced above are built on the same structural idea as the STARMA processes of Pfeifer and Deutsch (1980c). More precisely, a linear PSTARMA($q_{a_1,\dots,a_q}, r_{b_1,\dots,b_r}$) process satisfies the model equation of a STARMA($v_{u_1,\dots,u_v}, q_{a_1,\dots,a_q}$) process with $v = \max\{q, r\}$ and $u_i = \max\{a_i, b_i\}$ (setting $a_i = -1$ for $i > q$ and $b_i = -1$ for $i > r$); see Section A.1.1 in the Appendix for the formal definition.

For log-linear PSTARMA processes, this exact correspondence no longer holds. Nevertheless, since $\log(\mathbf{x}) \approx \log(\mathbf{x}+\mathbf{1}_p)$ when the components of $\mathbf{x}$ are large, a log-linear PSTARMA process approximately satisfies a STARMA model for the process $\{\log(\mathbf{Y}_t + \mathbf{1}_p)\}$.

3.2.3 Stationarity and Ergodicity

Since models (3.1) and (3.2) are parsimonious special cases of the multivariate models in Section 2.1, the general conditions for stationarity and ergodicity from Fokianos et al. (2020) apply. Those conditions cover first-order models ($q = r = 1$); higher-order models were treated by Debaly and Truquet (2021). Adapting their results to the spatio-temporal setting yields the following.

Lemma 1. *Define* $\mathbf{A}_i = \sum_{\ell=0}^{a_i} \alpha_{i,\ell}\mathbf{W}^{(\ell)}$ *for* $i = 1, \dots, q$ *and* $\mathbf{B}_j = \sum_{\ell=0}^{b_j} \beta_{j,\ell}\mathbf{W}^{(\ell)}$ *for* $j = 1, \dots, r$. *A sufficient condition for stationarity, ergodicity, and the existence of all moments of the (log-)linear PSTARMA process is*

$$\left\| \sum_{i=1}^{\max\{q,r\}} \left(|\mathbf{A}_i|_{\text{elem}} + |\mathbf{B}_i|_{\text{elem}} \right) \right\|_2 < 1, \tag{3.5}$$

where $|\cdot|_{\text{elem}}$ *denotes element-wise absolute value and* $\|\cdot\|_2$ *denotes the spectral norm.*

For the linear model, non-negativity of all parameters makes the absolute values in (3.5) redundant. A more explicit sufficient condition follows from a result first derived in Maletz (2021) for first-order models.

Proposition 1. *Let a linear or log-linear PSTARMA*($q_{a_1,\dots,a_q}, r_{b_1,\dots,b_r}$) *process satisfy*

$$\sum_{i=1}^{q} \sum_{\ell=0}^{a_i} |\alpha_{i,\ell}| + \sum_{j=1}^{r} \sum_{\ell=0}^{b_j} |\beta_{j,\ell}| < \frac{1}{\sqrt{\tau}}, \tag{3.6}$$

where $\tau = \max_\ell \left\| \mathbf{W}^{(\ell)} \right\|_1$ *is the maximum column-sum norm over all weight matrices. Then the joint process* $\{(\mathbf{Y}_t, \boldsymbol{\mu}_t)\}$ *(respectively* $\{(\mathbf{Y}_t, \boldsymbol{\psi}_t)\}$*) is stationary, ergodic, and possesses finite moments of all orders. The bound satisfies* $\tau = 1$ *when all* $\mathbf{W}^{(\ell)}$ *are symmetric, and* $\tau \in [1, p-1]$ *in general.*

The proof directly follows by using the matrix norm inequality $\|\mathbf{X}\|_2 \leq \sqrt{\|\mathbf{X}\|_1 \cdot \|\mathbf{X}\|_\infty}$ together with $\left\|\mathbf{W}^{(\ell)}\right\|_\infty = 1$ (row-normalisation).

Condition (3.6) is conservative. For PNAR processes, for instance, the spectral condition (3.5) reduces to requiring the sum in (3.6) to be less than 1 (Armillotta and Fokianos, 2024), which is less restrictive than the bound $1/\sqrt{\tau}$. This weaker condition moreover guarantees second-order stationarity of linear PSTARMA processes through their connection to STARMA processes (Pfeifer and Deutsch, 1980a), but does not in general imply (3.5).

Finally, the stationarity conditions above apply to models without covariates or with time-constant covariates only. PSTARMAX processes with time-varying deterministic covariates are not stationary in general. If the covariates are stochastic, stationarity of the composite covariate term $\tilde{\mathbf{X}}_t = \sum_{k=1}^{m}\sum_{\ell=0}^{c_k}\gamma_{k,\ell}\mathbf{W}^{(\ell)}\mathbf{X}_{k,t}$, which holds whenever all individual covariate processes are stationary, is sufficient. Formal conditions for this case can be derived from Doukhan et al. (2023) and Agosto et al. (2016).

3.3 Parameter Estimation

The data-generating process introduced in Subsection 2.1.1 allows contemporaneous dependence between components through a copula construction based on a Poisson process, while retaining Poisson marginals. Because the resulting likelihood is simulation-based and intractable in closed form, exact maximum likelihood estimation is not feasible. As Fokianos et al. (2020), we employ quasi-maximum likelihood estimation (QMLE), whose asymptotic properties are derived below. Temporal and spatial dependencies are fully accounted for through the model equations (3.3) and (3.4); the only simplification is that the working likelihood treats the contemporaneous components as conditionally independent given the past.

3.3.1 Quasi-Maximum Likelihood Estimation

Let $\{\mathbf{Y}_t,\ t = 0, \dots, T\}$ be a realisation of a PSTARMA or PSTARMAX process. The parameters to be estimated are $\boldsymbol{\delta}$, $\alpha_{i,\ell}$, $\beta_{j,\ell}$, and $\gamma_{k,\ell}$, collected in the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\delta}, \alpha_{i\ell}, (i = 1, \dots, q, \ell = 0, \dots, a_i), \beta_{j\ell}, (j = 1, \dots, r, \ell = 0, \dots, b_j)\gamma_{k,\ell}, (k = 1, \dots, m, \ell = 0, \dots, s_k))'$

The quasi-log-likelihood is

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{T}\sum_{i=1}^{p}\big(Y_{i,t}\log\mu_{i,t}(\boldsymbol{\theta}) - \mu_{i,t}(\boldsymbol{\theta})\big). \tag{3.7}$$

Linear PSTARMAX process. The quasi-score is

$$S_T(\boldsymbol{\theta}) = \sum_{t=1}^{T} \frac{\partial \boldsymbol{\mu}_t'}{\partial \boldsymbol{\theta}} \, \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \big(\mathbf{Y}_t - \boldsymbol{\mu}_t(\boldsymbol{\theta}) \big), \tag{3.8}$$

where $\mathbf{D}_t(\boldsymbol{\theta}) = \operatorname{diag}(\mu_{i,t}(\boldsymbol{\theta}),\, i = 1, \dots, p)$. The Jacobian $\partial \boldsymbol{\mu}_t' / \partial \boldsymbol{\theta}$ is computed component-wise via the recursions

$$\begin{aligned}
\frac{\partial \boldsymbol{\mu}_t}{\partial \boldsymbol{\delta}} &= \mathbf{I}_p + \sum_{j=1}^{q} \mathbf{A}_j \frac{\partial \boldsymbol{\mu}_{t-j}}{\partial \boldsymbol{\delta}}, & \frac{\partial \boldsymbol{\mu}_t}{\partial \alpha_{i,\ell}} &= \mathbf{W}^{(\ell)} \boldsymbol{\mu}_{t-i} + \sum_{j=1}^{q} \mathbf{A}_j \frac{\partial \boldsymbol{\mu}_{t-j}}{\partial \alpha_{i,\ell}}, \\
\frac{\partial \boldsymbol{\mu}_t}{\partial \beta_{i,\ell}} &= \mathbf{W}^{(\ell)} \mathbf{Y}_{t-i} + \sum_{j=1}^{q} \mathbf{A}_j \frac{\partial \boldsymbol{\mu}_{t-j}}{\partial \beta_{i,\ell}}, & \frac{\partial \boldsymbol{\mu}_t}{\partial \gamma_{k,\ell}} &= \mathbf{W}^{(\ell)} \mathbf{X}_{k,t} + \sum_{j=1}^{q} \mathbf{A}_j \frac{\partial \boldsymbol{\mu}_{t-j}}{\partial \gamma_{k,\ell}}.
\end{aligned} \tag{3.9}$$

For the homogeneous model, $\partial \boldsymbol{\mu}_t / \partial \boldsymbol{\delta}$ is replaced by the scalar derivative $\partial \boldsymbol{\mu}_t / \partial \delta_0 = \mathbf{1}_p + \sum_{j=1}^{q} \mathbf{A}_j \, \partial \boldsymbol{\mu}_{t-j} / \partial \delta_0$.

Log-linear PSTARMAX process. The quasi-score takes the form

$$S_T(\boldsymbol{\theta}) = \sum_{t=1}^{T} \frac{\partial \boldsymbol{\psi}_t'}{\partial \boldsymbol{\theta}} \big(\mathbf{Y}_t - \exp(\boldsymbol{\psi}_t(\boldsymbol{\theta})) \big). \tag{3.10}$$

The component-wise derivatives in $\partial \boldsymbol{\psi}_t' / \partial \boldsymbol{\theta}$ follow the same recursions as (3.9), with $\boldsymbol{\mu}$ replaced by $\boldsymbol{\psi}$ and $\mathbf{Y}_{t-i}$ replaced by $\log(\mathbf{Y}_{t-i} + \mathbf{1}_p)$.

Quasi-information matrices. The quasi-information matrix for the linear process is

$$\mathbf{G}_T(\boldsymbol{\theta}) = \sum_{t=1}^{T} \frac{\partial \boldsymbol{\mu}_t'}{\partial \boldsymbol{\theta}} \, \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \, \Sigma_t(\boldsymbol{\theta}) \, \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \, \frac{\partial \boldsymbol{\mu}_t}{\partial \boldsymbol{\theta}}, \tag{3.11}$$

where $\Sigma_t = \operatorname{Var}(\mathbf{Y}_t \mid \mathscr{F}_{t-1})$ is determined by the copula. For the log-linear process it simplifies to

$$\mathbf{G}_T(\boldsymbol{\theta}) = \sum_{t=1}^{T} \sum_{i=1}^{p} \exp(\psi_{i,t}(\boldsymbol{\theta})) \, \frac{\partial \psi_{i,t}}{\partial \boldsymbol{\theta}} \, \frac{\partial \psi_{i,t}'}{\partial \boldsymbol{\theta}}. \tag{3.12}$$

Asymptotic properties. Under mild regularity conditions, discussed in Section A.1.3 of the Appendix, which include the stationarity and ergodicity of the process without covariates, the QMLE $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$ is strongly consistent and asymptotically normal,

$$\sqrt{T} \big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{true}} \big) \xrightarrow{D} \mathcal{N}\big(\mathbf{0},\, \mathbf{H}^{-1} \mathbf{G} \mathbf{H}^{-1} \big) \quad \text{as} \quad T \to \infty. \tag{3.13}$$

For the linear process (3.3), the sandwich matrices are

$$\mathbf{G}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \boldsymbol{\mu}_t'}{\partial \boldsymbol{\theta}} \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \, \Sigma_t(\boldsymbol{\theta}) \, \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\mu}_t}{\partial \boldsymbol{\theta}}\right], \quad \mathbf{H}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \boldsymbol{\mu}_t'}{\partial \boldsymbol{\theta}} \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\mu}_t}{\partial \boldsymbol{\theta}}\right], \tag{3.14}$$

and for the log-linear process (3.4),

$$\mathbf{G}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \boldsymbol{\psi}_t'}{\partial \boldsymbol{\theta}} \, \Sigma_t(\boldsymbol{\theta}) \, \frac{\partial \boldsymbol{\psi}_t}{\partial \boldsymbol{\theta}}\right], \quad \mathbf{H}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \boldsymbol{\psi}_t'}{\partial \boldsymbol{\theta}} \, \mathbf{D}_t(\boldsymbol{\theta}) \, \frac{\partial \boldsymbol{\psi}_t}{\partial \boldsymbol{\theta}}\right]. \tag{3.15}$$

In practice, $\mathbf{H}$ and $\mathbf{G}$ are estimated by their empirical counterparts evaluated at $\hat{\boldsymbol{\theta}}$. Note that the conditions underlying (3.13) require $\boldsymbol{\theta}_{\text{true}}$ to lie in the interior of the parameter space; for the linear model this implies $\boldsymbol{\theta}_{\text{true}} \succ \mathbf{0}$. A detailed discussion of all regularity conditions is given in the Appendix in Section A.1.3.

3.3.2 Model Selection

Model selection in regression is commonly based on the Akaike information criterion (AIC, Akaike, 1974) or the Bayesian information criterion (BIC, Schwarz, 1978). Both criteria rely on a correctly specified likelihood, however, so their direct application under QMLE can introduce bias. To account for the additional uncertainty from a potentially misspecified likelihood, Pan (2001) proposed the quasi-information criterion (QIC),

$$\text{QIC} = -2\,\ell(\hat{\boldsymbol{\theta}}) + 2\,\text{tr}(\hat{\mathbf{G}}\hat{\mathbf{H}}^{-1}), \tag{3.16}$$

where ℓ, $\mathbf{G}$, and $\mathbf{H}$ are defined in (3.7), (3.14), and (3.15). The penalty term $\text{tr}(\hat{\mathbf{G}}\hat{\mathbf{H}}^{-1})$ reduces to the number of free parameters when the likelihood is correctly specified, in which case QIC coincides with AIC.

3.3.3 Testing Linear Hypotheses

The asymptotic distribution (3.13) enables Wald tests for hypotheses of the form

$$H_0 : \mathbf{C}\boldsymbol{\theta} = \mathbf{c}_0 \qquad \text{vs.} \qquad H_1 : \mathbf{C}\boldsymbol{\theta} \neq \mathbf{c}_0,$$

where $\mathbf{C} \in \mathbb{R}^{u \times v}$ with $\text{rank}(\mathbf{C}) = f \leq u \leq v$ and $\boldsymbol{\theta} \in \mathbb{R}^v$. Under H_0, as $T \to \infty$,

$$T \cdot (\mathbf{C}\hat{\boldsymbol{\theta}} - \mathbf{c}_0)' \big(\mathbf{C}\,\hat{\Sigma}_T\,\mathbf{C}'\big)^{-1} (\mathbf{C}\hat{\boldsymbol{\theta}} - \mathbf{c}_0) \xrightarrow{D} \chi^2_f, \tag{3.17}$$

with $\hat{\Sigma}_T = \hat{\mathbf{H}}^{-1}\hat{\mathbf{G}}\hat{\mathbf{H}}^{-1}$ estimated via the empirical counterparts of (3.14) and (3.15).

Boundary issues in the linear model. In the linear PSTARMAX model, all parameters are constrained to be non-negative. If $\boldsymbol{\theta}_{\text{true}}$ lies on the boundary of this parameter space, i.e. one or more components equal zero, the QMLE is no longer asymptotically normal, and the chi-squared limit in (3.17) no longer applies. This is analogous to the situation in GARCH models (Francq and Zakoïan, 2019, Chap. 8), where the parameters are also constrained to be non-negative. The correct limiting distribution is then a chi-bar-squared distribution (Self and Liang, 1987), i.e. a weighted mixture of chi-squared distributions whose exact form depends on the hypothesis being tested (Shapiro, 1988). Procedures for determining the corresponding critical values are discussed in Molenberghs and Verbeke (2007) and Francq and Zakoïan (2019, Sec. 8.2).

A practically important special case is the one-sided test for a single parameter,

$$H_0 : \theta_i = 0 \qquad \text{vs.} \qquad H_1 : \theta_i > 0.$$

Here the asymptotic null distribution of the Wald statistic is $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$, where χ_0^2 denotes a point mass at zero. For significance level $\alpha < 0.5$, the $(1-\alpha)$-quantile of this mixture equals the $(1-2\alpha)$-quantile of the χ_1^2 distribution, so it suffices to double the nominal significance level, i.e. use $\tilde{\alpha} = 2\alpha$.

Log-linear model. No such boundary correction is needed for the log-linear model, where parameters are unrestricted. The relevant test is two-sided,

$$H_0 : \theta_i = 0 \qquad \text{vs.} \qquad H_1 : \theta_i \neq 0,$$

and the Wald statistic converges to a standard χ_1^2 distribution under H_0.

4 Detection of Persistent Level Shifts

> "*A difference, to be a difference, must make a difference.*"
>
> — Gertrude Stein

Similar to univariate time series, spatio-temporal data are also subject to effects that are caused by external events or political measures. In epidemiological applications such events include for example, the introduction of a vaccine against the observed disease or a lockdown to stem its spread. An other event might be an unusually severe outbreak with a long-term impact. While univariate intervention analysis primarily focuses on the time point of an event and type of the intervention effect, the spatio-temporal context introduces more complexity: the spatial extent, i.e., which locations are affected, might be unknown and must be estimated alongside the time point or the intervention type.

Following Fokianos and Fried (2010, 2012), we model intervention effects via a covariate process $\{\mathbf{x}_t\}$, that can be included in a suitable spatio-temporal model. We focus specifically on the log-linear spatio-temporal count data model from Chapter 3, as this framework naturally accommodates both positive and negative correlations. Let $\mathbf{v} \in \{0,1\}^p$ denote a spatial indicator vector, that defines in which locations the intervention occurs. We define the spatio-temporal intervention covariate process by

$$\mathbf{x}_t = (1 - \eta\mathscr{B})^{-1} 1(t = \tau)\mathbf{v}, \tag{4.1}$$

where $\eta \in [0,1]$ is the type and τ the time-point of the intervention, similar to the univariate case; recall Section 2.3. Similar to the univariate case, for $\eta = 0$ there is a single impulse at time τ at the affected regions j for which $v_j = 1$. For $\eta = 1$ there is a persistent impulse, and for $\eta \in (0,1)$ there is a geometrically decaying impulse.

Figure 4.1 shows the effect of such interventions using a simple example. Only one central region is affected by the intervention. One observation here is the diffusion of intervention effects. Due to the inherent spatial dependencies of the underlying model, an intervention at a specific site j will propagate to neighboring sites over time, even if those neighbors were not directly hit by the intervention. This also affects the cumulative

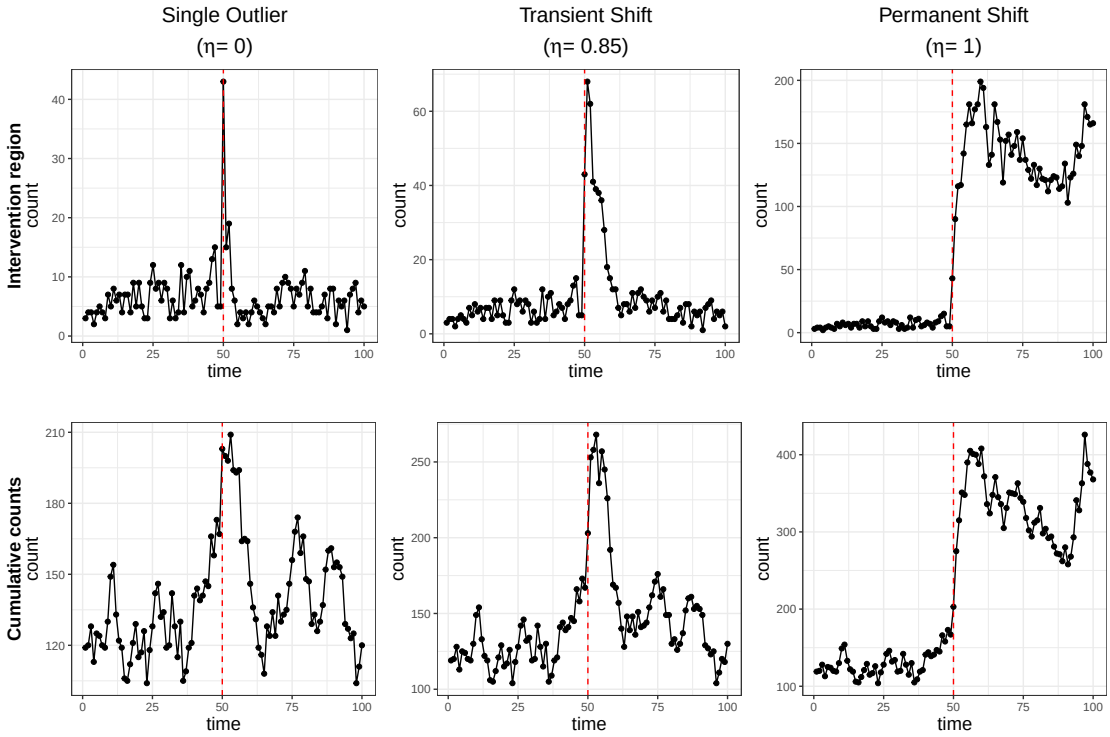


Figure 4.1: Simulated effects of three types of interventions ($\eta = 0$, $\eta = 0.85$, and $\eta = 1$) on the directly affected location (top row) and on the spatially aggregated series across all locations (bottom row). Data were generated from model (3.4) using parameters $\delta_0 = 0.15$, $\beta_{1,0} = 0.3$, $\beta_{1,1} = 0.45$, $\beta_{1,2} = 0.1$, and $\gamma = 2$. The spatial structure is based on a regular 5×5 grid with first order neighborhood with rook contiguity.

counts across all regions and makes it difficult to identify the type of intervention if it is unknown. Hence, in the remainder of this chapter, we will focus on interventions with known intervention effects; more precisely, we restrict ourselves to the case $\eta = 1$, i.e., to persistent level shifts. Nevertheless, this parameterization also allows for temporary changes when multiple intervention covariates are combined in the model.

Note on the Identifiability of Intervention effects In the spatio-temporal log-linear model (3.4), we allow spatial weight matrices of orders greater than 0 for the covariates. This allows the model directly capture indirect interventions effects in addition to direct ones. The parameter $\gamma_{k,0}$ captures the direct intervention effect of the k-th intervention on the locations directly affected. The parameter $\gamma_{k,1}$, on the other hand, describes the indirect (spatial spillover) effect of this intervention on locations that are not themselves affected but are adjacent to affected locations.

This leads to a fundamental challenge regarding parameter identifiability. As the spatial extent of an intervention becomes more *global*, i.e., as the number of affected sites in $\mathbf{v}$ increases, it becomes increasingly difficult to distinguish between the direct intervention

effect and the indirect effects propagated through higher-order spatial lags. Specifically, for the case of a global intervention, $\mathbf{v} = \mathbf{1}_p$, the intervention parameters $\gamma_{k,\ell}$ for spatial orders $\ell > 0$ are not identifiable. To get around this problem, we confine ourself the detection of direct intervention effects, i.e. we reduce the log-linear model to

$$\boldsymbol{\psi}_t = \boldsymbol{\delta} + \sum_{i=1}^{q} \sum_{\ell=0}^{a_i} \alpha_{i,\ell} \mathbf{W}^{(\ell)} \boldsymbol{\psi}_{t-i} + \sum_{j=1}^{r} \sum_{\ell=0}^{b_j} \beta_{j,\ell} \mathbf{W}^{(\ell)} \log(\mathbf{Y}_{t-j} + \mathbf{1}_p) + \sum_{k=1}^{m} \gamma_{k,0} \mathbf{x}_{k,t}. \tag{4.2}$$

4.1 Score Test

When the time point of a level shift and the locations affected are known, the corresponding covariate process can be directly included in the model estimation, and significance can be determined based on the estimated parameter $\hat{\gamma}_{k,0}$. Alternatively, a score test can be constructed, analogous to the univariate case; see Section 2.3. This approach is advantageous when the times or locations affected are not known, as it does not require the covariate to be considered in the estimation itself and allows for the selection of the most suitable candidate from several options.

Let $\{\mathbf{x}_{k,t}\}, k = 1, \dots, m$ denote a set of m intervention covariate processes, with corresponding time-points $\boldsymbol{\tau}_{(m)} = (\tau_1, \dots, \tau_m)'$ and locations $\mathbf{V}_{(m)} = (\mathbf{v}_1, \dots, \mathbf{v}_m)$. We test the following hypothesis pair:

$$H_0 : \boldsymbol{\gamma} = \mathbf{0}_m \quad \text{versus} \quad H_1 : \boldsymbol{\gamma} \neq \mathbf{0}_m,$$

where $\boldsymbol{\gamma} = (\gamma_{1,0}, \dots, \gamma_{m,0})'$ denotes the vector of unknown intervention parameters. The test statistic is derived using the score function of the log-linear model (4.2) evaluated under the null hypothesis of no intervention effects. For given candidates $\boldsymbol{\tau}_{(m)}$, $\mathbf{V}_{(m)}$, the score test statistic is given by the quadratic form:

$$T(\mathbf{V}_{(m)}, \boldsymbol{\tau}_{(m)}) = S_\gamma(\hat{\boldsymbol{\theta}}_0, \mathbf{0}_m)^\top \widehat{\Sigma}^{-1} (\hat{\boldsymbol{\theta}}_0, \mathbf{0}_m) S_\gamma(\hat{\boldsymbol{\theta}}_0, \mathbf{0}_m), \tag{4.3}$$

where $\hat{\boldsymbol{\theta}}_0$ is the QMLE estimate of the log-linear model (4.2) under H_0. The variance $\widehat{\Sigma}$ of the Score vector is estimated using the following formula, see for example Armillotta and Fokianos (2023):

$$\Sigma = \mathbf{G}_{\gamma\gamma} - \mathbf{H}_{\gamma\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{G}_{\theta\gamma} - \mathbf{G}_{\gamma\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{H}_{\theta\gamma} + \mathbf{H}_{\gamma\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{G}_{\theta\theta}\mathbf{H}_{\theta\theta}^{-1}\mathbf{H}_{\theta\gamma}. \tag{4.4}$$

The matrices $\mathbf{H}$ and $\mathbf{G}$ correspond to matrices in the sandwich-estimator, see (2.10), partitioned analogously to the univariate case, i.e.

$$\mathbf{H}(\boldsymbol{\theta}, \mathbf{0}_m) = \begin{pmatrix} \mathbf{H}_{\theta\theta} & \mathbf{H}_{\theta\gamma} \\ \mathbf{H}_{\gamma\theta} & \mathbf{H}_{\gamma\gamma} \end{pmatrix}, \quad \text{and} \quad \mathbf{G}(\boldsymbol{\theta}, \mathbf{0}_m) = \begin{pmatrix} \mathbf{G}_{\theta\theta} & \mathbf{G}_{\theta\gamma} \\ \mathbf{G}_{\gamma\theta} & \mathbf{G}_{\gamma\gamma} \end{pmatrix}$$

Under the null hypothesis, the test statistic approximately follows a chi-square distribution with m degrees of freedom, i.e. it holds

$$T(\mathbf{V}_{(m)}, \boldsymbol{\tau}_{(m)}) \sim \chi^2_m.$$

4.2 Detecting Level Shifts at unknown Space and Time

In practical applications, both the exact time τ and the set of affected locations $\mathbf{v}$ might be unknown. To address this, we propose an iterative forward selection procedure building on maximum score tests that detects one or more level shifts in the data. This procedure has the advantage over the method proposed by Fokianos and Fried (2010, 2012) that the data does not need to be adjusted for intervention effects before the next one can be detected.

The user specifies a set of candidate time points $\mathscr{T} = \{\tau_{\min}, \dots, \tau_{\max}\}$ and a collection of candidate spatial indicator vectors $\mathscr{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$. Spatial candidates may overlap (i.e., $\mathbf{v}_i^\top \mathbf{v}_j \neq 0$ for $i \neq j$), allowing the procedure to choose the most suitable configuration.

The iterative procedure constructs a sequential hypothesis testing problem with a *marginal False Discovery Rate (mFDR)* of $\alpha \in (0, 1)$. The mFDR is defined as

$$\text{mFDR} := \frac{\mathbb{E}(V)}{\mathbb{E}(\max(R, 1))},$$

where V is the number of false rejections (Type I errors) and R is the total number of rejections (Foster and Stine, 2008). The iterative procedure is defined as follows:

1. **Initialization:** Estimate the baseline spatio-temporal log-linear model under H_0 to obtain $\hat{\boldsymbol{\theta}}_0$. Define the initial sets of selected parameters as empty: $\boldsymbol{\tau}_{(0)} = \emptyset$ and $\mathbf{V}_{(0)} = \emptyset$. Let $\mathscr{K}^{(0)} = \mathscr{T} \times \mathscr{V}$ be the set of all possible candidate pairs $(\tau, \mathbf{v})$.
2. **Iterative Search:** For $k = 1, 2, \dots$:
   a) Search for the best additional candidate $(\tau^*, \mathbf{v}^*) \in \mathscr{K}^{(k-1)}$ that maximizes the joint score test statistic:
   $$(\tau^{(k)}, \mathbf{v}^{(k)}) = \arg \max_{(\tau, \mathbf{v}) \in \mathscr{K}^{(k-1)}} T\left((\mathbf{V}_{(k-1)}, \mathbf{v}), (\boldsymbol{\tau}_{(k-1)}, \tau)\right), \tag{4.5}$$
   where $(\mathbf{V}_{(k-1)}, \mathbf{v})$ denotes the matrix of locations augmented by a new column, and $(\boldsymbol{\tau}_{(k-1)}, \tau)$ the vector of time-points augmented by a new entry.
   b) Update the sets of selected interventions:
   $$\boldsymbol{\tau}_{(k)} = (\boldsymbol{\tau}_{(k-1)}, \tau^{(k)}), \quad \mathbf{V}_{(k)} = \left(\mathbf{V}_{(k-1)}, \mathbf{v}^{(k)}\right). \tag{4.6}$$
   c) Remove the selected candidate from the search space:
   $$\mathscr{K}^{(k)} = \mathscr{K}^{(k-1)} \setminus \{(\tau^{(k)}, \mathbf{v}^{(k)})\}.$$
3. **Stopping Criterion:** The algorithm terminates when the incremental gain in the test statistic falls below a threshold $\kappa_\alpha^{(k)}$, i.e., when
   $$T(\mathbf{V}_{(k)}, \boldsymbol{\tau}_{(k)}) - T(\mathbf{V}_{(k-1)}, \boldsymbol{\tau}_{(k-1)}) \leq \kappa_\alpha^{(k)}$$

After stopping at iteration k, the procedure has selected the first $k-1$ level-shift candidates. If the algorithm already stops at the first iteration, the null hypothesis of no level shifts cannot be rejected. The thresholds $\kappa_\alpha^{(k)}$ corresponds to the $(1-\alpha_k)$-Quantile of the underlying distribution $F^{(k)}$ of the increments $D_k := T(\mathbf{V}_{(k)}, \boldsymbol{\tau}_{(k)}) - T(\mathbf{V}_{(k-1)}, \boldsymbol{\tau}_{(k-1)})$, i.e. $D_k \sim F^{(k)}$. One option for choosing α_k is a Bonferroni-like adjustment which ensures $\sum_{k=1}^{\infty} \alpha_k \leq \alpha$. A condition that fulfills this is

$$\alpha_k := \alpha\gamma(1-\gamma)^{k-1}, 0 < \gamma < 1.$$

Another option is based on an α-investing strategy (Foster and Stine, 2008). Since the procedure is designed to stop at the first failure to reject $H_k : \gamma_{k,1} = 0$, one might employ an All-In-α-investing rule. In each iteration, the entire remaining α-budget is invested in the current test. If the null hypothesis is not rejected, the budget is exhausted and the procedure terminates; if it is rejected, the invested amount is fully recovered and the procedure continues. Consequently, each test is conducted at local significance level α. The first option is more restrictive, while the other one is more liberal.

4.2.1 Remark on the distribution of the increments.

At iteration k, the test statistic is maximized over a potentially large number of correlated candidates. For each remaining pair $(\mathbf{v}, \tau) \in \mathscr{K}^{(k-1)}$, the distribution of the test statistic under $H_k : \gamma_{k,1} = 0$, conditional on the $k-1$ previously detected level shifts, is approximately

$$T\big((\mathbf{V}_{(k-1)}, \mathbf{v}), (\boldsymbol{\tau}_{(k-1)}, \tau)\big) \mid \big(\mathbf{V}_{(k-1)}, \boldsymbol{\tau}_{(k-1)}\big) \;\sim\; T\big(\mathbf{V}_{(k-1)}, \boldsymbol{\tau}_{(k-1)}\big) + \chi^2_1.$$

Because the maximum is subsequently taken over a set of highly correlated statistics, determining the distribution of the maximum increment D_k is difficult. Provided that the candidate set is large relative to the number of level shifts selected per iteration, a natural heuristic is that the distribution of the maximum changes only slowly across iterations, so that $F^{(k)} \approx F^{(1)}$ for k small. Under this heuristic, it is sufficient to approximate $F^{(1)}$ for the entire procedure.

For univariate count time series, Fokianos and Fried (2010, 2012) employ a parametric bootstrap. However, this is impractical for spatio-temporal count data for two reasons: (i) it requires generating data using the estimated model under the null, for which the copula structure must be identified – a complex problem in itself; and (ii) it requires fitting the model in each bootstrap iteration, which is computationally prohibitive in high-dimensional settings. Armillotta and Fokianos (2023) propose a wild bootstrap procedure in which, at each bootstrap iteration, the addends of the quasi-score (3.10) are multiplied by an i.i.d. $\mathcal{N}(0,1)$ sequence. This approach was infeasible for the log-linear model considered here: the resulting critical values were excessively conservative, preventing rejection of the null hypothesis in all cases. Another option is to use the

Davies bound (Davies, 1987) to obtain an upper bound for the p-values, which are again too conservative, as noted by Armillotta and Fokianos (2023).

We now explain a heuristic that can be used to decide whether to stop the forward selection or to continue. One key observation is that under the null hypothesis of no level shifts at all, D_1 is the maximum of $n \cdot (\tau_{\max} - \tau_{\min})$ positively correlated χ^2_1 random variables, where n is the number of candidates in space. Let $X \sim \chi^2_1$ be independent of D_1. Since the probability of the maximum is bounded below by that of a single variable (all correlations equal to one) and above by the complement of independent variables, we obtain

$$P(X \geq k_\alpha) \leq P(D_1 \geq k_\alpha) \leq P(X \geq k_\alpha)^{\frac{1}{n \cdot (\tau_{\max} - \tau_{\min})}}.$$

The upper bound is attained when the maximum is taken over independent random variables; the lower bound corresponds to the degenerate case in which all correlations equal one, so that no additional information is gained from additional candidates.

When the calculated maximum increment is smaller than the $1 - \alpha_k$ quantile of the χ^2_1 distribution, the procedure should stop. When it exceeds the $(1 - \alpha_k)^{1/(n \cdot (\tau_{\max} - \tau_{\min}))}$ quantile of the χ^2_1 distribution, the procedure should continue. If the value falls between these two thresholds, one has to apply a decsision rule. In this thesis we use a rather conservative rule. In the data example we stop when the increment does not exceed the $(1 - \alpha_k)^{1/(n \cdot (\tau_{\max} - \tau_{\min}))}$ quantile of the χ^2_1 distribution, with $\alpha_k = \alpha 0.5^k$. This implements a very conservative rule.

5 Spatio-Temporal Double Generalized Linear Models

> "*The purpose of models is not to fit the data but to sharpen the questions.*"
> — Samuel Karlin

In the two models presented in Chapter 3, the conditional expectation of the observations is described by a spatio-temporal model. The main difference between the models lies in the link function, which connects the conditional expectation with a linear process under the assumption that the conditional observations marginally follow a Poisson distribution. Related models have been studied by Jahn et al. (2023), including a model using the so-called soft-plus link for the Poisson distribution and a model for bounded counts assuming marginal binomial distributions.

The Poisson assumption is not essential to the models from Chapter 3; it can be replaced by any distribution from the exponential family, analogously to generalized linear models, with the parameter estimation procedure derived accordingly. For univariate count time series, for example, the Poisson distribution is commonly replaced by a negative binomial distribution (Zhu, 2011) to accommodate overdispersion, with the dispersion parameter typically assumed to be constant. In some spatio-temporal models, spatial variation in the dispersion parameter has also been considered, see Meyer et al. (2017).

For independent data, Double Generalized Linear Models (DGLMs), described in Section 2.4, provide a framework for modeling varying dispersion parameters within a GLM by specifying a separate sub-model for the dispersion. Joint modeling of mean and dispersion parameters has a long history in the context of independent data; early references include Nelder (1985), Efron (1986), and Nelder and Pregibon (1987). In univariate time series analysis, analogous models for non-normal data remain rare; a Bayesian state space approach to time-varying generalized linear models was proposed by West et al. (1985). Jørgensen (1997) provides a comprehensive treatment of exponential dispersion models; Smyth and Jørgensen (2002) illustrates their connection with DGLMs through an application to insurance claims data; and Bonat et al. (2018) introduces a class of discrete generalized linear models based on Poisson–Tweedie factorial dispersion models. Recent

contributions for time varying dispersion models include Barreto-Souza et al. (2025) and Zheng et al. (2022).

This chapter introduces a framework based on DGLMs in which the conditional mean of the observations is modeled spatio-temporally, following a structure similar to the models from Chapter 3, while a second model describes the marginal dispersion parameter spatio-temporally. This framework can be seen as an extension of the models from Chapter 3 to describe spatio-temporal varying overdispersion, and additional as a generalization of the STARMA-GARCH procedure to non-normal data.

Section 5.1 describes the basic model framework. Section 5.2 presents the spatio-temporal model for the mean, and Section 5.3 introduces the dispersion model. Section 5.5 discusses heuristics for model stability and conjectured asymptotic properties of the parameter estimators. Details on the software implementation are given in Chapter 6.

5.1 Model Assumptions

Let $\{\mathbf{Y}_t = (Y_{1,t}, \dots, Y_{p,t})' \mid t = 0, 1, \dots\}$ denote a p-dimensional time series representing the evolution of a target variable, say Y, observed at p fixed spatial locations over time. Each component $\{Y_{i,t},\ t = 0, 1, \dots\}$, $i = 1, \dots, p$, corresponds to a univariate time series of the response variable at location i and time t. We simultaneously observe two – not necessarily distinct – sets of covariate processes: $\{\mathbf{X}_{k,t} = (X_{1,k,t}, \dots, X_{p,k,t})'\}$ for $k = 1, \dots, m$, and $\{\tilde{\mathbf{X}}_{k,t} = (\tilde{X}_{1,k,t}, \dots, \tilde{X}_{p,k,t})'\}$ for $k = 1, \dots, \tilde{m}$. The covariates $\mathbf{X}_{k,t}$ will enter the mean model, while $\tilde{\mathbf{X}}_{k,t}$ are included in the dispersion model.

The proposed framework is based on the DGLM concept of Smyth (1989). Two distinct model components are specified: one for the conditional mean of Y and one for the dispersion parameter of the assumed distribution. Let $\mathcal{F}_t$ denote the history of the processes up to time t, as defined in Chapter 3. The framework is specified through the following components.

1. **Distribution.** The univariate conditional distributions $Y_{i,t} \mid \mathcal{F}_{t-1}$ are assumed to belong to the exponential dispersion family, i.e. $Y_{i,t} \mid \mathcal{F}_{t-1} \sim \mathrm{ED}(\mu_{i,t}, \phi_{i,t})$, with density

$$f(y_{i,t} \mid \theta_{i,t}, \phi_{i,t}, \mathcal{F}_{t-1}) = \exp\left(\frac{\theta_{i,t} y_{i,t} - A(\theta_{i,t})}{\phi_{i,t}} + c(\phi_{i,t}, y_{i,t})\right), \tag{5.1}$$

   where $\theta_{i,t}$ and $\phi_{i,t}$ denote the natural and dispersion parameters, respectively. The conditional mean and variance are given by

$$\mu_{i,t} := \mathbb{E}(Y_{i,t} \mid \mathcal{F}_{t-1}) = \frac{\partial A(\theta_{i,t})}{\partial \theta_{i,t}}, \quad \sigma^2_{i,t} := \mathrm{Var}(Y_{i,t} \mid \mathcal{F}_{t-1}) = \phi_{i,t} \cdot V(\mu_{i,t}),$$

   where V is the variance function of the distribution.

2. **Linear Predictors.** Two linear processes $\{\boldsymbol{\psi}_t = (\psi_{1,t}, \dots, \psi_{p,t})'\}$ and $\{\boldsymbol{\zeta}_t = (\zeta_{1,t}, \dots, \zeta_{p,t})'\}$ serve as predictors for the mean and the dispersion, respectively. Their precise definitions are given in equations (5.2) and (5.4) below.

3. **Link Functions.** The relationships between the conditional mean, the dispersion parameter, and the linear predictors are established through link functions g and $\tilde{g}$:

$$g(\mu_{i,t}) = \psi_{i,t}, \quad \tilde{g}(\phi_{i,t}) = \zeta_{i,t}.$$

The collection of conditional means is referred to as the mean process $\{\boldsymbol{\mu}_t := (\mu_{1,t}, \dots, \mu_{p,t})'\}$, and the corresponding dispersion parameters form the dispersion process $\{\boldsymbol{\phi}_t := (\phi_{1,t}, \dots, \phi_{p,t})'\}$.

The assumptions above do not specify the joint conditional distribution of $\mathrm{Y}_t \mid \mathscr{F}_{t-1}$. We assume copula-based data-generating processes with marginal distributions satisfying the assumptions above, following the approach of Fokianos et al. (2020). These copulas introduce dependence between the components $Y_{i,t} \mid \mathscr{F}_{t-1}$. Concrete examples are discussed in Chapter 6.

5.2 Conditional Mean Model

The linear process $\{\boldsymbol{\psi}_t\}$ from Assumption 2 is modeled as

$$\begin{aligned}\boldsymbol{\psi}_t = \boldsymbol{\delta} &+ \sum_{i \in \mathscr{Q}^{(\mu)}} \sum_{\ell \in \mathscr{A}_i^{(\mu)}} \alpha_{i\ell} \mathbf{W}_\alpha^{(\ell)} h(\boldsymbol{\psi}_{t-i}) + \sum_{j \in \mathscr{R}^{(\mu)}} \sum_{\ell \in \mathscr{B}_j^{(\mu)}} \beta_{j\ell} \mathbf{W}_\beta^{(\ell)} \tilde{h}\big(\mathrm{Y}_{t-j}\big) \\ &+ \sum_{k=1}^{m} \sum_{\ell \in \mathscr{C}_k^{(\mu)}} \gamma_{k\ell} \mathbf{W}_\gamma^{(\ell)} \mathbf{X}_{k,t}.\end{aligned} \tag{5.2}$$

This model comprises an intercept, past values of the linear process, transformed past observations, and covariate effects. The weight matrices may differ across model components. In contrast to the fixed maximum lags q and r used in Chapter 3, the temporal lags are here indexed by sets $\mathscr{Q}^{(\mu)}, \mathscr{R}^{(\mu)} \subset \mathbb{N}$, enabling the modeling of stochastic seasonality. The spatial orders are similarly replaced by index sets $\mathscr{A}_i, \mathscr{B}_j, \mathscr{C}_k \subset \mathbb{N}_0$ for each temporal lag.

Let $\boldsymbol{\theta}_\mu$ denote the parameter vector containing the intercept $\boldsymbol{\delta}$ and the coefficients $\alpha_{i\ell}$, $\beta_{j\ell}$, $\gamma_{k\ell}$. Depending on the link function, non-negativity constraints may be necessary to ensure that the conditional mean process $\boldsymbol{\mu}_t$ is compatible with the assumed distribution in (5.1); for count responses, for instance, $\boldsymbol{\mu}_t > 0$ is required.

As in Chapter 3, we distinguish between a *homogeneous* intercept $\boldsymbol{\delta} = \delta_0 \mathbf{1}_p$ and an *inhomogeneous* intercept $\boldsymbol{\delta} = (\delta_1, \dots, \delta_p)'$.

Transformation of the linear process and past observations. The transformations h and $\tilde{h}$ in (5.2) are applied to past values of the linear process and the observations, respectively. In the `glmSTARMA` package (see Chapter 6), these functions are determined by the link function connecting $\boldsymbol{\psi}_t$ and $\boldsymbol{\mu}_t$. For example, in the log-linear model (3.4) one has $\tilde{h}(x) = \log(x+1)$ and $h(x) = x$, while in the linear model (3.3) both reduce to the identity, $h = \tilde{h} = \mathrm{id}$. For the models of Jahn et al. (2023), the function h corresponds to the inverse link, so that $h(\boldsymbol{\psi}) = \boldsymbol{\mu}$.

5.3 Conditional Dispersion Model

Unlike the conditional mean, the dispersion parameters cannot be inferred directly from the observations. Following Smyth (1989), we introduce pseudo-observations $d_{i,t}$ with $\mathbb{E}(d_{i,t} \mid \mathcal{F}_{t-1}) = \phi_{i,t}$, based on the conditional deviance residuals. Specifically,

$$d_{i,t} = 2 \cdot \log\left[\frac{f(y_{i,t} \mid \mu_{i,t} = y_{i,t},\, \mathcal{F}_{t-1})}{f(y_{i,t} \mid \mu_{i,t} = \mu_{i,t},\, \mathcal{F}_{t-1})}\right], \tag{5.3}$$

where the numerator corresponds to a saturated model with $\mu = y_{i,t}$. We collect the pseudo-observations into the process $\{\mathbf{d}_t = (d_{1,t}, \dots, d_{p,t})'\}$.

Using $\mathbf{d}_t$ as observations, the linear predictor $\boldsymbol{\zeta}_t$ for the dispersion process is modeled analogously to (5.2):

$$\begin{aligned}\boldsymbol{\zeta}_t = \tilde{\boldsymbol{\delta}} &+ \sum_{i \in \mathcal{Q}^{(\phi)}} \sum_{\ell \in \mathcal{A}_i^{(\phi)}} \tilde{\alpha}_{i\ell} \tilde{\mathbf{W}}_\alpha^{(\ell)} \boldsymbol{\zeta}_{t-i} + \sum_{j \in \mathcal{R}^{(\phi)}} \sum_{\ell \in \mathcal{B}_j^{(\phi)}} \tilde{\beta}_{j\ell} \tilde{\mathbf{W}}_\beta^{(\ell)} \tilde{h}_\phi(\mathbf{d}_{t-j}) \\ &+ \sum_{k=1}^{\tilde{m}} \sum_{\ell \in \mathcal{C}_k^{(\phi)}} \tilde{\gamma}_{k\ell} \tilde{\mathbf{W}}_\gamma^{(\ell)} \tilde{\mathbf{X}}_{k,t}.\end{aligned} \tag{5.4}$$

The model orders, transformations, and covariates in (5.4) are allowed to differ from those of the mean model. Most existing models in the literature consider only the mean model (5.2) and thus implicitly assume a constant dispersion, corresponding to $\boldsymbol{\zeta}_t = \tilde{\delta}_0 \mathbf{1}_p$ for some constant $\tilde{\delta}_0$.

Remark 1 (Distribution of $d_{i,t}$). Under certain regularity conditions, the conditional distribution of $d_{i,t} \mid \mathcal{F}_{t-1}$ is approximately Gamma with a fixed dispersion parameter of 2. This approximation is exact for the normal and inverse Gaussian distributions. For the Poisson distribution, it holds provided that $\mu_{i,t} > 3$; for the binomial, both $\mu_{i,t} > 3$ and $n - \mu_{i,t} > 3$ are required. For the Gamma distribution, the approximation holds when $\phi_{i,t} < 1/3$, although in this case $d_{i,t}$ follows a Digamma distribution exactly; see Smyth (1989) for details.

Remark 2 (Count data and overdispersion). For standard Poisson models, the conditional mean and variance coincide, implying $\phi_{i,t} = 1$ for all i, t. In practice, quasi-Poisson models are often used to accommodate overdispersion or underdispersion, with $\text{Var}(Y_{i,t} \mid \mathscr{F}_{t-1}) = \phi_{i,t}\mu_{i,t}$ for $\phi_{i,t} > 0$. A distribution with this variance structure is the generalized Poisson distribution (Consul and Jain, 1973). An analogous issue arises for the binomial distribution. Over- or underdispersion can be generated if the n_i binary trials are not independent (Ahn and Chen, 1995).

Remark 3 (Negative binomial models). Consider the negative binomial distribution parameterized as

$$f(y_{i,t} \mid \mathscr{F}_{t-1}) = \frac{\Gamma(\nu_{i,t} + y_{i,t})}{\Gamma(y_{i,t}+1)\Gamma(\nu_{i,t})} \left(\frac{\nu_{i,t}}{\nu_{i,t}+\mu_{i,t}}\right)^{\nu_{i,t}} \left(\frac{\mu_{i,t}}{\nu_{i,t}+\mu_{i,t}}\right)^{y_{i,t}}, \quad y_{i,t} \in \mathbb{N}_0, \tag{5.5}$$

with shape parameter $\nu_{i,t} > 0$, so that $\mathbb{E}(Y_{i,t} \mid \mathscr{F}_{t-1}) = \mu_{i,t}$ and $\text{Var}(Y_{i,t} \mid \mathscr{F}_{t-1}) = \mu_{i,t} + \mu_{i,t}^2/\nu_{i,t}$. The negative binomial distribution naturally accommodates conditional overdispersion since $\text{Var}(Y_{i,t} \mid \mathscr{F}_{t-1}) \geq \mathbb{E}(Y_{i,t} \mid \mathscr{F}_{t-1})$. To exploit this property, we fix the GLM dispersion parameter at 1 and instead model the shape parameter $\nu_{i,t}$ or its inverse $\tilde{\phi}_{i,t} = 1/\nu_{i,t}$; see, e.g., Venables and Ripley (2002) and Liboschik et al. (2017).

For the negative binomial distribution, we compute the deviance pseudo-observations (5.3) via Pearson residuals of the Poisson distribution,

$$d_{i,t}^{\star} = \frac{(y_{i,t} - \mu_{i,t})^2}{\mu_{i,t}}, \tag{5.6}$$

since the shape parameter is unknown. These are approximately distributed as $(1 + \phi_{i,t}\mu_{i,t}) \cdot \chi_1^2$, with $\phi_{i,t}^{\star} = \mathbb{E}(d_{i,t}^{\star} \mid \mathscr{F}_{t-1})$ modeled by (5.4). The inverse shape parameter is then recovered as

$$\tilde{\phi}_{i,t} = \max\left\{0, \frac{\phi_{i,t}^{\star} - 1}{\mu_{i,t}}\right\}, \tag{5.7}$$

by solving the conditional expectation of $d_{i,t}^{\star}$ for $\tilde{\phi}_{i,t}$ and truncating at zero, since negative values would imply underdispersion, which the negative binomial distribution cannot represent. When $\tilde{\phi}_{i,t} = 0$, the distribution reduces to the Poisson.

Remark 4 (Deviance vs. Pearson residuals). The pseudo-observations $d_{i,t}$ can alternatively be defined using squared Pearson residuals,

$$d_{i,t} = \frac{(y_{i,t} - \mu_{i,t})^2}{V(\mu_{i,t})}, \tag{5.8}$$

as proposed by Wu and Li (2016). However, deviance residuals have a stronger theoretical foundation and are more widely used in practice (Jørgensen, 1997). A key advantage is that, under deviance residuals, the parameter vectors $\boldsymbol{\theta}_{\mu}$ and $\boldsymbol{\theta}_{\phi}$ are orthogonal (Smyth, 1989), which simplifies the estimation procedure described in Section 5.4.

5.4 Parameter Estimation

Parameters of each submodel are estimated using a quasi-maximum likelihood (QMLE) approach. The estimation proceeds iteratively, alternating between fitting the mean model (5.2) and the dispersion model (5.4), following the standard approach for DGLMs (Smyth, 1989). In the `glmSTARMA` package (see Chapter 6) where this framework is implemented, estimation is performed subject to the stability constraints given in (5.13) by default.

5.4.1 Quasi-Maximum-Likelihood Estimation

In each iteration parameters are estimated by maximizing a conditional quasi-log-likelihood of the form

$$\ell(\boldsymbol{\theta}) = \sum_{t=\tau}^{T} \sum_{i=1}^{p} \left(\frac{\theta_{i,t}(\boldsymbol{\theta}) y_{i,t} - A(\theta_{i,t}(\boldsymbol{\theta}))}{\phi_{i,t}} + c(\phi_{i,t}, y_{i,t}) \right), \tag{5.9}$$

where $\tau = \max\left(\mathcal{Q}^{(\mu)} \cup \mathcal{R}^{(\mu)} \cup \mathcal{Q}^{(\phi)} \cup \mathcal{R}^{(\phi)}\right)$ is the largest time lag used in either model. Although the log-likelihood is constructed under the assumption of conditional independence among the components of $\mathbf{Y}_t$, spatial and temporal correlations are still accounted for through the model equations (5.2) and (5.4).

The quasi-score function for the mean model is

$$S_T(\boldsymbol{\theta}_\mu) = \sum_{t=\tau}^{T} \left[\frac{\partial \boldsymbol{\psi}_t}{\partial \boldsymbol{\theta}} \right]' \mathbf{D}_t (\mathbf{Y}_t - g^{-1}(\boldsymbol{\psi}_t)), \tag{5.10}$$

where $\mathbf{D}_t = \operatorname{diag}\left((g^{-1})'(\psi_{i,t})/\sigma^2_{i,t},\ i = 1, \ldots, p\right)$ and $\sigma^2_{i,t} = \phi_{i,t} \cdot A''(\theta_{i,t}(\boldsymbol{\theta}_\mu))$. The quasi-Fisher information matrix is

$$\mathbf{G}_T(\boldsymbol{\theta}_\mu) = \sum_{t=\tau}^{T} \left[\frac{\partial \boldsymbol{\psi}_t}{\partial \boldsymbol{\theta}_\mu} \right]' \tilde{\mathbf{D}}_t \left[\frac{\partial \boldsymbol{\psi}_t}{\partial \boldsymbol{\theta}_\mu} \right], \tag{5.11}$$

with $\tilde{\mathbf{D}}_t = \operatorname{diag}\big(((g^{-1})'(\psi_{i,t}))^2/\sigma_{i,t}^2,\ i = 1,\dots,p\big)$; a derivation is given in Appendix A.2. Denoting by $\mathbf{X} \circ \mathbf{Y}$ the Hadamard (element-wise) product of matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{p\times q}$, the derivatives $\partial\boldsymbol{\psi}_t/\partial\boldsymbol{\theta}_\mu$ are computed recursively as

$$
\begin{aligned}
\frac{\partial\boldsymbol{\psi}_t'}{\partial\delta_0} &= \mathbf{1}_p + \sum_{j=1}^{q}\sum_{\ell=0}^{a_j} \alpha_{j\ell}\mathbf{W}^{(\ell)}\left[h'(\boldsymbol{\psi}_{t-j}) \circ \frac{\partial\boldsymbol{\psi}_{t-j}'}{\partial\delta_0}\right], \\
\frac{\partial\boldsymbol{\psi}_t'}{\partial\alpha_{jo}} &= \mathbf{W}^{(o)}h(\boldsymbol{\psi}_{t-j}) + \sum_{i=1}^{q}\sum_{\ell=0}^{a_i} \alpha_{i\ell}\mathbf{W}^{(\ell)}\left[h'(\boldsymbol{\psi}_{t-j}) \circ \frac{\partial\boldsymbol{\psi}_{t-j}'}{\partial\alpha_{jo}}\right], \\
\frac{\partial\boldsymbol{\psi}_t'}{\partial\beta_{io}} &= \mathbf{W}^{(o)}\tilde{h}(\mathbf{Y}_{t-i}) + \sum_{j=1}^{q}\sum_{\ell=0}^{a_j} \alpha_{j\ell}\mathbf{W}^{(\ell)}\left[h'(\boldsymbol{\psi}_{t-j}) \circ \frac{\partial\boldsymbol{\psi}_{t-j}'}{\partial\beta_{io}}\right], \\
\frac{\partial\boldsymbol{\psi}_t'}{\partial\gamma_{ko}} &= \mathbf{W}^{(o)}\mathbf{X}_{k,t} + \sum_{j=1}^{q}\sum_{\ell=0}^{a_j} \alpha_{j\ell}\mathbf{W}^{(\ell)}\left[h'(\boldsymbol{\psi}_{t-j}) \circ \frac{\partial\boldsymbol{\psi}_{t-j}'}{\partial\gamma_{ko}}\right].
\end{aligned}
\tag{5.12}
$$

5.4.2 Iterative Estimation of Mean and Dispersion Models

In each iteration of the alternating model estimation procedure, the quasi-log-likelihood (5.9) is maximized via numerical optimization. For the mean model, the density of the observations is used directly. For the dispersion model, a Gamma likelihood with dispersion parameter fixed at 2 is evaluated on the pseudo-observations $\mathbf{d}_t$ derived from the mean model residuals. The quasi-log-likelihood for the dispersion model has the same structural form as (5.9), with $\mathbf{Y}_t$ replaced by $\mathbf{d}_t$, $\boldsymbol{\psi}_t$ replaced by $\boldsymbol{\zeta}_t$, and all other components adapted accordingly.

Two enhancements improve numerical stability. First, a step-halving strategy monitors the joint log-likelihood after each sub-model update: if an update decreases the likelihood, the parameter change is scaled by successive factors of 0.5 until a likelihood improvement is achieved or a maximum number of halvings is reached. Second, a dual convergence criterion terminates the iteration when either (i) the change in the combined parameter vector falls below a tolerance ε, or (ii) the relative change in the log-likelihood between successive iterations is negligible. Algorithm 1 provides a detailed description of the procedure.

5.5 Asymptotics and Stability

5.5.1 Some Heuristics on Model Stability

Regarding identifiability, the same arguments apply as for the count data models in Chapter 3, Section 3.2: both the mean and the dispersion sub-models must individually be identifiable for the overall model to be identifiable.

Algorithm 1 Iterative Estimation with Likelihood Monitoring

Initialize iteration counter: $k \leftarrow 0$
Set constant dispersion: $\phi_{i,t}^{(0)} \leftarrow \phi$ for all i, t
Compute initial joint log-likelihood: $\ell^{(0)}$
repeat
    Mean Model Update:
        Estimate mean model via QMLE using current $\phi_{i,t}^{(k)}$: $\tilde{\boldsymbol{\theta}}_{\mu}^{(k)}$
        Compute trial log-likelihood: $\tilde{\ell}_{\mu}^{(k)}$
    if $\tilde{\ell}_{\mu}^{(k)} < \ell^{(k)}$ **then**
            Apply step-halving to $\tilde{\boldsymbol{\theta}}_{\mu}^{(k)}$ until ℓ increases
    end if
        Accept update: $\hat{\boldsymbol{\theta}}_{\mu}^{(k)} \leftarrow \tilde{\boldsymbol{\theta}}_{\mu}^{(k)}$

    Dispersion Model Update:
        Compute pseudo-observations $\{\mathbf{d}_t^{(k)}\}$ from current mean model
        Estimate dispersion model on $\{\mathbf{d}_t^{(k)}\}$ via QMLE: $\tilde{\boldsymbol{\theta}}_{\phi}^{(k)}$
        Update dispersion: $\tilde{\phi}_{i,t}^{(k)}$ based on $\tilde{\boldsymbol{\theta}}_{\phi}^{(k)}$
        Compute trial log-likelihood: $\tilde{\ell}_{\phi}^{(k)}$
    if $\tilde{\ell}_{\phi}^{(k)} < \ell^{(k)}$ **then**
            Apply step-halving to $\tilde{\boldsymbol{\theta}}_{\phi}^{(k)}$ until ℓ increases
    end if
        Accept update: $\hat{\boldsymbol{\theta}}_{\phi}^{(k)} \leftarrow \tilde{\boldsymbol{\theta}}_{\phi}^{(k)}$, $\phi_{i,t}^{(k+1)} \leftarrow \tilde{\phi}_{i,t}^{(k)}$

    Compute joint log-likelihood: $\ell^{(k+1)}$
    $k \leftarrow k + 1$
until $\|(\hat{\boldsymbol{\theta}}_{\mu}^{(k)}, \hat{\boldsymbol{\theta}}_{\phi}^{(k)}) - (\hat{\boldsymbol{\theta}}_{\mu}^{(k-1)}, \hat{\boldsymbol{\theta}}_{\phi}^{(k-1)})\|_2 < \varepsilon$ **or** $|\ell^{(k)} - \ell^{(k-1)}|/|\ell^{(k-1)}| < \delta$
return $(\hat{\boldsymbol{\theta}}_{\mu}, \hat{\boldsymbol{\theta}}_{\phi})$

To the best of our knowledge, no existing work in multivariate time series analysis jointly models a time- and location-dependent mean and dispersion in the spatio-temporal setting considered here, so the theoretical properties of such models remain to be investigated. In the univariate case, it is common practice to pair an ARMA process with a GARCH component; see Francq and Zakoïan (2019, Sec. 7) or the GARMA–GARCH framework of Zheng et al. (2022). For count data, Barreto-Souza et al. (2025) study univariate INGARCH processes with time-varying dispersion under the negative binomial assumption.

A necessary condition for stationarity of $\{\mathrm{Y}_t\}$ is stationarity of both $\{\boldsymbol{\mu}_t\}$ and $\{\boldsymbol{\phi}_t\}$, which is generally guaranteed only if (a) no covariates are present, (b) all covariates are time-

invariant, or (c) all covariate processes are strictly stationary (Aknouche and Francq, 2021). The existence of higher-order moments may additionally be required.

To prevent the components of (5.2) and (5.4) from diverging, the transformation functions $h, \tilde{h}, h_\phi$, and $\tilde{h}_\phi$ should be "well-behaved". We conjecture that it suffices for these functions to be twice continuously differentiable and to grow at most linearly: specifically, $|h(x)| \leq c|x| + d$ for constants $c, d > 0$, and $|\tilde{h}(g^{-1}(x))| \leq \tilde{c}|x| + \tilde{d}$ for $\tilde{c}, \tilde{d} > 0$. Note that these conditions are satisfied by all transformations implemented in the R package presented in the next Chapter; see Table 6.2.

We conjecture that, under these growth conditions, the constraint

$$\sum_{i \in \mathcal{Q}^{(\mu)}} \sum_{\ell \in \mathcal{A}_i^{(\mu)}} |\alpha_{i\ell}| + \sum_{j \in \mathcal{R}^{(\mu)}} \sum_{\ell \in \mathcal{B}_j^{(\mu)}} |\beta_{j\ell}| < 1, \tag{5.13}$$

together with the analogous condition for $\tilde{\alpha}_{i\ell}$ and $\tilde{\beta}_{j\ell}$ in (5.4), is sufficient to guarantee stationarity, ergodicity, and the existence of all moments of $\{\mathbf{Y}_t\}$. Similar conditions are standard for GARCH and INGARCH processes, as well as for spatio-temporal count models (Armillotta and Fokianos, 2024; Maletz et al., 2024) and generalized network models (Knight et al., 2020).

For the Softplus and Softclipping processes studied by Jahn et al. (2023), weaker sufficient conditions are available, at the cost of assuming a constant dispersion parameter:

$$\sum_{i \in \mathcal{Q}} \sum_{\ell \in \mathcal{A}_i} \max\{0, \alpha_{i\ell}\} + \sum_{j \in \mathcal{R}} \sum_{\ell \in \mathcal{B}_j} \max\{0, \beta_{j\ell}\} < 1 \quad \text{and} \quad \sum_{i \in \mathcal{Q}} \sum_{\ell \in \mathcal{A}_i} |\alpha_{i\ell}| < 1. \tag{5.14}$$

Under stationarity, the expected value of $\boldsymbol{\psi}_t$ is approximately

$$\mathbb{E}(\boldsymbol{\psi}_t) \approx \left(\mathbf{I}_p - \sum_{i \in \mathcal{Q}} \sum_{\ell \in \mathcal{A}_i} \alpha_{i\ell} \mathbf{W}_\alpha^{(\ell)} + \sum_{j \in \mathcal{R}} \sum_{\ell \in \mathcal{B}_j} \beta_{j\ell} \mathbf{W}_\beta^{(\ell)} \right)^{-1} \boldsymbol{\delta}^\star, \tag{5.15}$$

and $\mathbb{E}(\mathbf{Y}_t) \approx g^{-1}(\mathbb{E}(\boldsymbol{\psi}_t))$, where $\boldsymbol{\delta}^\star$ collects the intercept and the expected covariate contributions. Analogous expressions hold for $\mathbb{E}(\boldsymbol{\zeta}_t)$ and $\mathbb{E}(\mathbf{d}_t)$.

5.5.2 Theoretical Properties of the Estimator

Smyth (1989) shows that $\boldsymbol{\theta}_\mu$ and $\boldsymbol{\theta}_\phi$ are orthogonal and can therefore be estimated separately. We assume that this orthogonality carries over to the present time-series setting, even though Smyth (1989) considers independent observations.

Consistent estimation further requires that the parameters satisfy (5.13) or (5.14), that covariates are bounded with finite variance, and that the information accrued by the covariates grows sufficiently fast with the sample size; see the discussion in Appendix Section A.1.3 for the PSTARMA processes.

Under these conditions, we conjecture that the estimates $\hat{\boldsymbol{\theta}}_\mu$ and $\hat{\boldsymbol{\theta}}_\phi$ satisfy

$$\sqrt{T}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\big) \xrightarrow{D} \mathcal{N}\big(\mathbf{0}, \mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{-1}\big) \quad \text{as} \quad T \to \infty, \tag{5.16}$$

where $\boldsymbol{\theta}_0$ denotes the true parameter vector. The asymptotic variance has the sandwich form with matrices (for the mean model)

$$\mathbf{G}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \boldsymbol{\psi}_t^\top}{\partial \boldsymbol{\theta}} \tilde{\mathbf{D}}_t(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\psi}_t}{\partial \boldsymbol{\theta}}\right], \qquad \mathbf{H}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \boldsymbol{\psi}_t^\top}{\partial \boldsymbol{\theta}} \mathbf{D}_t(\boldsymbol{\theta}) \boldsymbol{\Sigma}_t(\boldsymbol{\theta}) \mathbf{D}_t(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\psi}_t}{\partial \boldsymbol{\theta}}\right], \tag{5.17}$$

where

$$\begin{aligned} \mathbf{D}_t &= \operatorname{diag}\big(1/[g'(g^{-1}(\psi_{i,t}))\sigma^2_{i,t}],\ i = 1, \dots, p\big), \\ \tilde{\mathbf{D}}_t &= \operatorname{diag}\big(((g^{-1})'(\psi_{i,t}))^2/\sigma^2_{i,t},\ i = 1, \dots, p\big), \end{aligned} \tag{5.18}$$

with $\sigma^2_{i,t} = \phi_{i,t} V(\mu_{i,t})$. For the dispersion model, the analogous matrices use $\boldsymbol{\zeta}_t$ and $\mathbf{d}_t$ in place of $\boldsymbol{\psi}_t$ and $\mathbf{Y}_t$. Both matrices are estimated empirically at $\hat{\boldsymbol{\theta}}$: $\mathbf{G}$ via the expected Hessian and $\mathbf{H}$ via the outer product of the score. A derivation of the expected Hessian is given in the Appendix A.2.

These asymptotic results hold when the true parameters lie in the interior of the parameter space. If some parameters lie on the boundary (e.g., due to non-negativity constraints), test statistics must be adjusted. See Section 3.3.3 for a detailed discussion.

5.6 Information Criteria

To facilitate model selection, including the choice of covariates, temporal orders, and spatial dependencies, one might consider several information criteria, such as in Subsection 3.3.2: the Akaike Information Criterion (AIC) (Akaike, 1974), the Bayesian Information Criterion (BIC) (Schwarz, 1978), and the Quasi Information Criterion (QIC) (Pan, 2001).

General Structure. All criteria follow the form

$$\text{IC} = -2\tilde{\ell}(\boldsymbol{\theta}) + \text{Penalty}, \tag{5.19}$$

where $\tilde{\ell}(\boldsymbol{\theta})$ denotes the scaled (quasi-)log-likelihood and the penalty term varies by criterion. To ensure comparability between models with different temporal orders, one might scale the log-likelihood (5.9) by the effective number of observations:

$$\tilde{\ell}(\boldsymbol{\theta}) = \frac{T}{T - \max(\mathcal{Q} \cup \mathcal{R})} \ell(\boldsymbol{\theta}), \tag{5.20}$$

where T denotes the total number of time points.

Quasi-Likelihood Models. For quasi models, where the mean structure of the Poisson or binomial distribution is paired with a dispersion model, a proper likelihood function may not exist or cannot be uniquely defined. We therefore employ adjusted profile likelihoods (McCullagh and Tibshirani, 1990; Smyth and Verbyla, 1999). For the quasi-Poisson case, the quasi log-likelihood for one observation with mean $\mu_{i,t}$ and dispersion-parameter $\phi_{i,t}$ is

$$\begin{aligned}\ell(y_{i,t}; \mu_{i,t}, \phi_{i,t}) = -\frac{1}{2}\Bigg[&\log(\phi_{i,t}) + 2y_{i,t} + 2\log\Gamma(y_{i,t}+1) - 2y_{i,t}\log(y_{i,t}) \\ &+ \frac{y_{i,t}\log\left(\frac{y_{i,t}}{\mu_{i,t}}\right) - (y_{i,t}-\mu_{i,t})}{\phi_{i,t}}\Bigg],\end{aligned} \tag{5.21}$$

with the convention $0\log(0) := 0$.

For the quasi-binomial case with observations $y_{i,t} \in \{0, \dots, n_i\}$, $\mu_{i,t} = n_i\pi_{i,t}$ and dispersion $\phi_{i,t}$, the quasi-log-likelihood is

$$\begin{aligned}\ell(y_{i,t}; \mu_{i,t}, \phi_{i,t}, n_i) = -\frac{1}{2}\Bigg[&\log(\phi_{i,t}) + 2\log\binom{n_i}{y_{i,t}} \\ &- 2y_{i,t}\log(y_{i,t}) - 2(n_i - y_{i,t})\log(n_i - y_{i,t}) \\ &+ \frac{1}{\phi_{i,t}}\left(2y_{i,t}\log\left(\frac{y_{i,t}}{\mu_{i,t}}\right) + 2(n_i - y_{i,t})\log\left(\frac{n_i - y_{i,t}}{n_i - \mu_{i,t}}\right)\right)\Bigg].\end{aligned} \tag{5.22}$$

Penalty Terms. The penalty terms for the three information criteria differ in how they penalize model complexity. For the AIC, the penalty is given by $2k$, where k denotes the total number of parameters in the mean and dispersion models. The BIC employs a stronger penalty of $k\log(Tp)$, where p is the number of spatial units, thereby favoring more parsimonious models as the sample size increases. The QIC accounts for model complexity through the sandwich variance estimator of the QMLE and is defined as

$$\text{Penalty}_{\text{QIC}} = 2\left(\text{tr}(\mathbf{G}_\mu^{-1}\mathbf{H}_\mu) + \text{tr}(\mathbf{G}_\phi^{-1}\mathbf{H}_\phi)\right), \tag{5.23}$$

where $\mathbf{G}_\mu$ and $\mathbf{H}_\mu$ are the sandwich components for the mean model (see Equation (5.17)), and $\mathbf{G}_\phi$ and $\mathbf{H}_\phi$ are defined analogously for the dispersion model.

6 Software

> "*Computers are useless. They can only give you answers.*"
>
> — Pablo Picasso

The model framework presented in Chapter 5 is implemented in the R package `glmSTARMA` (Maletz et al., 2026b). The most recent stable version is available on the Comprehensive R Archive Network (CRAN); versions with newer and possibly untested features are available on GitHub at `https://github.com/stmaletz/glmSTARMA`. Since the core computational routines are written in C++, installing the development version from GitHub requires suitable compilers – `Rtools` on Windows (`https://cran.r-project.org/bin/windows/Rtools/`), and `XCode` together with the `GNU Fortran compiler` on macOS (see `https://mac.r-project.org/tools/`). Once installed, the package is loaded via `library("glmSTARMA")`. Table 6.1 gives an overview of its most important functions.

The package is designed for ease of use: arguments are self-explanatory and allow different model specifications to be compared quickly. Spatio-temporal models within the framework of Chapter 5 are fitted using the functions `glmstarma` and `dglmstarma`, and data can be simulated from these processes with `glmstarma.sim` and `dglmstarma.sim`, respectively. The functions `glmstarma` and `glmstarma.sim` operate on the mean model (5.2) only, while their counterparts `dglmstarma` and `dglmstarma.sim` additionally handle a dispersion model (5.4). Detailed documentation for each function is available via the corresponding help pages.

Section 6.1 presents the most important arguments of the main functions, their expected input, and available options. Section 6.2 discusses the simulation functions, including additional arguments and the data-generating algorithms required to introduce contemporaneous dependence via copulas, illustrated with two examples. In the Appendix A.3, we additionally provide a short comparison with other existing R packages for spatio-temporal modelling on the datasets analyzed in Chapter 8.

	Name	Description
Main Functions	`glmstarma`	Fit model for the mean to given data
	`dglmstarma`	Fit mean and dispersion model to given data
	`glmstarma.sim`	Simulate observations applying the mean model
	`dglmstarma.sim`	Simulate observations applying a mean and dispersion model
Help Functions	`TimeConstant`	Create a time-constant covariate, i.e. $\forall t : \mathbf{X}_t = \mathbf{X} \in \mathbb{R}^p$
	`SpatialConstant`	Create a spatial constant covariate, i.e. $\mathbf{X}_t = x_t \cdot \mathbf{1}_p$, $x_t \in \mathbb{R}$
	`generateW`	Create spatial weight matrices on regular grids.
Generic Functions	`summary`	Summary of fitted model
	`fitted`	Fitted values
	`residuals`	Different types of residuals
	`AIC`	Akaike Information Criterion
	`BIC`	Bayesian Information Criterion
	`QIC`	Quasi Information Criterion
	`predict`	Predict future values
	`vcov`	Variance-covariance matrix of estimated coefficients
Data	`load_data`	Download and return preprocessed datasets
	`delete_data`	Remove datasets from the cache

Table 6.1: Most important functions in the `R` package `glmSTARMA`.

6.1 Important Arguments

The key arguments of the main functions concern the marginal distribution of the observations, the model orders, the covariates, and the spatial weight matrices. The functions `dglmstarma` and `dglmstarma.sim` have the additional argument `pseudo_observations`, which determines how the pseudo-observations entering the dispersion model are computed; it defaults to the deviance residuals.

6.1.1 Marginal Distribution

Table 6.2 summarises the distributions and link functions currently implemented in `glmSTARMA`, together with the corresponding variance functions. Analogously to the `glm` function in base R, all fitting and simulation functions include a `family` argument that simultaneously specifies the marginal conditional distribution and the link function. Each distribution is implemented via a generator function named after the distribution and prefixed with `v` (for *vector*), and the link function is selected through the `link` argument. Specifying a link function also implicitly determines the two transformation functions h and $\tilde{h}$, which appear in the mean model (5.2) and the dispersion model (5.4). The dispersion model always uses the `vgamma` family internally, so the link functions available for `vgamma` also apply to (5.4); the corresponding link is specified via the `dispersion_link` argument in `dglmstarma` and `dglmstarma.sim`.

Some link functions or transformations involve an additional tuning parameter c, which defaults to one and can be changed via the `const` argument of the family function. The

copula used in the simulation-based data-generating process can be specified through the `copula` and `copula_param` arguments of the family function; see Section 6.2 for more details.

Bounded count data are modelled using the `vbinomial` or `vquasibinomial` family. Both accept an argument `size`, which specifies the upper bound for the counts – either as a scalar applying to all locations, or as an integer vector of length p with a separate bound for each location.

All family functions beside `vpoisson` and `vbinomial` have an argument `dispersion`. If this is `NULL` a global dispersion parameter is estimated in `glmstarma`. Alternatively a scalar, a vector of length p, or a $p \times T$ matrix can be supplied to either set a global, location-specific or a space-time specific known dispersion parameter.

In a more recent version of the R package (see Maletz et al. (2026a) for the associated preprint), we replace the transformation of the observations in the gamma distribution with $\tilde{h}(\mathbf{Y_t}) = \log(\mathbf{Y}_t + c\mathbf{1}_p)$, using $c = 1$ as the default. As a result, past pseudo-observations in the dispersion model are also transformed by default, via $\log(\mathbf{d}_t + \mathbf{1}_p)$, which improves the stability of the dispersion model and allows for more reliable parameter estimation. This change leads to results that differ from those reported in this thesis.

6.1.2 Spatial Weight Matrices

Spatial weight matrices are passed as a list to the arguments `wlist`, `wlist_past_mean`, and `wlist_covariates`, where the $(\ell + 1)$-th list element corresponds to the weight matrix of spatial order ℓ. List elements may be dense matrices (class `matrix` in base R) or sparse matrices (class `dgCMatrix` from the `Matrix` package, Bates et al., 2025). If `wlist_past_mean` or `wlist_covariates` are `NULL`, the matrices supplied in `wlist` are used for all terms. For `dglmstarma` and `dglmstarma.sim`, the additional arguments `wlist_pseudo_obs`, `wlist_past_dispersion`, and `wlist_covariates_dispersion` serve the same purpose for the dispersion model, with identical fallback behaviour. A warning is issued whenever the supplied matrices are not row-normalised.

Example weight matrices based on Euclidean distances on regular grids can be constructed with the helper function `generateW`. For real applications, the `spdep` package (Bivand, 2025) provides more advanced functionality for generating spatial weight matrices from arbitrary geometries.

6.1.3 Covariates

Covariates are supplied as a (named) list via the arguments `covariates` and, if applicable, `dispersion_covariates`, with each list element representing one covariate process. Each element must cover the same observation times as the data and is provided either

as a $p \times T$ matrix or via the helper functions `SpatialConstant` and `TimeConstant`. `SpatialConstant` takes a vector of length T and returns a covariate of the form $\mathbf{X}_t = x_t \mathbf{1}_p$, which varies over time but is spatially uniform. `TimeConstant` takes a vector of length p and returns a covariate of the form $\mathbf{X}_t = \mathbf{X} \in \mathbb{R}^p$, which varies across locations but is constant in time.

6.1.4 Model Orders

The model orders from (5.2) and (5.4) are passed via the `model` argument (or `mean_model` and `dispersion_model` in the `dglmstarma` functions) as a named list.

Two types of intercepts are distinguished: a *homogeneous* intercept $\boldsymbol{\delta} = \delta_0 \mathbf{1}_p$ and an *inhomogeneous* intercept $\boldsymbol{\delta} = (\delta_1, \dots, \delta_p)'$. The default is a homogeneous intercept; setting `intercept = "inhomogeneous"` in the `model` list selects the inhomogeneous variant. Note that an inhomogeneous intercept should not be combined with time-invariant covariates, i.e., covariates satisfying $\mathbf{X}_{k,t} = \mathbf{X}_{k,\tilde{t}}$ for all $t, \tilde{t}$, as this leads to identifiability problems.

The temporal index sets $\mathscr{Q} \subset \mathbb{N}$ and $\mathscr{R} \subset \mathbb{N}$ default to $\{1, \dots, q\}$ and $\{1, \dots, r\}$, where q and r are determined by the length of the vectors, or number of columns in matrices supplied in `past_mean` and `past_obs`, respectively. In applications such as modelling stochastic periodicities it may be preferable to regress on a sparse subset of lags; in that case, custom index sets can be passed as integer vectors via `past_mean_time_lags` and `past_obs_time_lags`.

The spatial index sets $\mathscr{A}_i \subset \mathbb{N}_0$, $\mathscr{B}_j \subset \mathbb{N}_0$, and $\mathscr{C}_k \subset \mathbb{N}_0$ are controlled by the list elements `past_mean`, `past_obs`, and `covariates`. In the simplest case, each is an integer vector whose i-th (resp. j-th, k-th) entry gives the maximum spatial order for the corresponding time lag or covariate; the sets are then $\mathscr{A}_i = \{0, \dots, a_i\}$ and analogously for the others. Covariates default to spatial order 0, i.e., purely local effects. For finer control, binary matrices $\mathbf{A} \in \{0,1\}^{a \times q}$, $\mathbf{B} \in \{0,1\}^{b \times r}$, and $\mathbf{C} \in \{0,1\}^{c \times k}$ may be supplied instead, defining the index sets directly as $\mathscr{A}_i = \{\ell : a_{\ell+1,i} = 1\}$ and analogously.

Data Type	**Distribution**	$V(\mu_{i,t})$	**Link**	$g(\boldsymbol{\mu}_t)$	$h(\boldsymbol{\psi}_t)$	$\tilde{h}(\mathbf{Y}_t)$	**NNP**
Gaussian	Normal (`vnormal`)	1	`"identity"`	$\boldsymbol{\mu}_t$	$\boldsymbol{\psi}_t$	$\mathbf{Y}_t$	
			`"log"`	$\log(\boldsymbol{\mu}_t)$	$\boldsymbol{\psi}_t$	$\log(\lvert\mathbf{Y}_t\rvert)$	
			`"inverse"`	$1/\boldsymbol{\mu}_t$	$\boldsymbol{\psi}_t$	$1/\mathbf{Y}_t$	
Positive continuous	Gamma (`vgamma`)	$\mu_{i,t}^2$	`"identity"`	$\boldsymbol{\mu}_t$	$\boldsymbol{\psi}_t$	$\mathbf{Y}_t$	✓
			`"log"`	$\log(\boldsymbol{\mu}_t)$	$\boldsymbol{\psi}_t$	$\log(\mathbf{Y}_t)$	
			`"inverse"`	$1/\boldsymbol{\mu}_t$	$\boldsymbol{\psi}_t$	$1/\mathbf{Y}_t$	✓
	Inverse-Gaussian (`vinverse.gaussian`)	$\mu_{i,t}^3$	`"1/mu^2"`	$1/\boldsymbol{\mu}_t^2$	$\boldsymbol{\psi}_t$	$1/\mathbf{Y}_t^2$	✓
			`"inverse"`	$1/\boldsymbol{\mu}_t$	$\boldsymbol{\psi}_t$	$1/\mathbf{Y}_t$	✓
			`"identity"`	$\boldsymbol{\mu}_t$	$\boldsymbol{\psi}_t$	$\mathbf{Y}_t$	✓
			`"log"`	$\log(\boldsymbol{\mu}_t)$	$\boldsymbol{\psi}_t$	$\log(\mathbf{Y}_t)$	
Count data	Poisson (`vpoisson`, `vquasipoisson`)	$\mu_{i,t}$	`"log"`	$\log(\boldsymbol{\mu}_t)$	$\boldsymbol{\psi}_t$	$\log(\mathbf{Y}_t + c \cdot \mathbf{1}_p)$	
			`"identity"`	$\boldsymbol{\mu}_t$	$\boldsymbol{\psi}_t$	$\mathbf{Y}_t$	✓
			`"sqrt"`	$\sqrt{\boldsymbol{\mu}_t}$	$\boldsymbol{\psi}_t$	$\sqrt{\mathbf{Y}_t}$	✓
			`"softplus"`	$c\log(\exp(\boldsymbol{\mu}_t/c) - \mathbf{1}_p)$	$c\log(\mathbf{1}_p + \exp(\boldsymbol{\psi}_t/c))$	$\mathbf{Y}_t$	
	Negative-binomial (`vnegative.binomial`)	$\mu_{i,t} + \phi_{i,t} \cdot \mu_{i,t}^2$	`"log"`	$\log(\boldsymbol{\mu}_t)$	$\boldsymbol{\psi}_t$	$\log(\mathbf{Y}_t + c \cdot \mathbf{1}_p)$	
			`"identity"`	$\boldsymbol{\mu}_t$	$\boldsymbol{\psi}_t$	$\mathbf{Y}_t$	✓
			`"sqrt"`	$\sqrt{\boldsymbol{\mu}_t}$	$\boldsymbol{\psi}_t$	$\sqrt{\mathbf{Y}_t}$	✓
			`"softplus"`	$c\log(\exp(\boldsymbol{\mu}_t/c) - \mathbf{1}_p)$	$c\log(\mathbf{1}_p + \exp(\boldsymbol{\psi}_t/c))$	$\mathbf{Y}_t$	
	Binomial (`vbinomial`, `vquasibinomial`)	$\mu_{i,t}(1 - \mu_{i,t}/n_i)$	`"identity"`	$\boldsymbol{\mu}_t/\mathbf{n}$	$\boldsymbol{\psi}_t$	$\mathbf{Y}_t/n$	✓
			`"softclipping"`	$c\log\left[\frac{\exp(\boldsymbol{\mu}_t/c\mathbf{n}) - \mathbf{1}_p}{\mathbf{1}_p - \exp((\boldsymbol{\mu}_t - \mathbf{n})/c\mathbf{n})}\right]$	$c\log\left(\frac{1+\exp(\boldsymbol{\psi}_t/c)}{1+\exp((\boldsymbol{\psi}_t - 1)/c)}\right)$	$\mathbf{Y}_t/n$	
			`"logit"`	$\log\left(\frac{\boldsymbol{\mu}_t}{\mathbf{n} - \boldsymbol{\mu}_t}\right)$	$(1 + \exp(-\boldsymbol{\psi}_t))^{-1}$	$\mathbf{Y}_t/n$	
			`"probit"`	$\Phi^{-1}(\boldsymbol{\mu}_t/\mathbf{n})$	$\Phi(\boldsymbol{\psi}_t)$	$\mathbf{Y}_t/n$	

Table 6.2: Distributions and link functions implemented in `glmSTARMA`, with the corresponding transformations of past feedback-process values and observations. A checkmark in the NNP column indicates that the model formula (5.2) or (5.4) requires all parameters to be non-negative. The constant $c > 0$ is a hyperparameter with default value 1, adjustable via the family argument `const`. The vector $\mathbf{n}$ of upper bounds in the binomial distribution is set via the argument `size`; a scalar value is applied to all locations.

6.2 Simulation Functions

The simulation functions `glmstarma.sim` and `dglmstarma.sim` take an argument `ntime` specifying the length of the simulated time series; the spatial dimension p is inferred automatically from the weight matrices. The argument `parameters` (or `parameters_mean` and `parameters_dispersion` in `dglmstarma.sim`) has the same structure as the corresponding `model` argument, but supplies the true parameter values rather than the model orders. Any parameter entries that do not match the specified model orders are silently set to zero, and list elements for terms not included in the model may be omitted.

An optional argument `n_start` controls the length of the burn-in phase used to initialise the process. Time-varying covariates are excluded during burn-in; the default is 100 time points.

6.2.1 Simulation Procedure

The simulation proceeds in two phases. During the burn-in phase, $\max\left(\mathcal{Q}^{(\mu)} \cup \mathcal{Q}^{(\phi)} \cup \mathcal{R}^{(\mu)} \cup \mathcal{R}^{(\phi)}\right)$ independent observations are drawn from the family specified by the `family` argument, and the initial values of $\mu_{i,t}$ and $\phi_{i,t}$ are set to the approximate stationary expectation from (5.15).

After this initialisation, the model equations (5.2) and (5.4) are applied iteratively to generate updated values of $\boldsymbol{\mu}_t$ and $\boldsymbol{\phi}_t$, from which new observations are drawn. Covariate effects are suppressed during burn-in. Once the burn-in is complete, the simulation runs for a further $T =$ `ntime` time points with covariates incorporated.

6.2.2 Sampling from the Joint Distribution

The model framework of Chapter 5 does not impose conditional independence among the components of Y_t given $\mathcal{F}_{t-1}$. To introduce contemporaneous dependence during simulation, the package supports copula-based data generation following the approach of Fokianos et al. (2020): the marginal distributions are of a known parametric form, and the joint dependence structure is governed by a copula. This preserves the marginal families while allowing flexible dependence modelling; applications to count data are given in Armillotta and Fokianos (2024) and Maletz et al. (2024).

The `copula` argument accepts one of `"normal"`, `"t"`, `"clayton"`, `"frank"`, `"gumbel"`, or `"joe"`, each corresponding to a copula implemented in the `copula` package (Yan, 2007); the scalar `copula_param` supplies the associated copula parameter. When no copula is specified, observations are generated under conditional independence using the standard random number generators.

Two simulation examples are provided in Subection 6.2.3.

Inversion method For the Normal, Poisson (with $\phi_{i,t} = 1$), Negative-binomial, Gamma, and Binomial distributions, contemporaneous dependence is introduced via the inversion method (see e.g. Devroye, 1986, Chapter 2). A random vector $\mathbf{u}_{0,t} = (u_{0,t,1}, \dots, u_{0,t,p})'$ is drawn from the specified copula; by the copula property, the marginal distributions of its components are uniform on $[0, 1]$ but not necessarily independent. Each observation is then obtained element-wise as

$$Y_{i,t} = F^{-1}(u_{0,t,i};\, \mu_{i,t},\, \phi_{i,t}),$$

where F^{-1} denotes the quantile function of the chosen distribution. For discrete distributions with non-unique quantile functions, the smallest value for which the CDF exceeds $u_{0,t,i}$ is used.

Michael–Schucany–Haas method The inverse-Gaussian distribution lacks a closed-form quantile function, so the inversion method is impractical. Instead, the Michael–Schucany–Haas algorithm (Michael et al., 1976) is used, drawing two independent random vectors $\mathbf{U}_{1,t}$ and $\mathbf{U}_{2,t}$ from the copula and computing $Y_{i,t}$ as follows:

1. Set $Z_{i,t} = \Phi^{-1}(U_{1,t,i})$, so that $Z_{i,t} \sim \mathcal{N}(0, 1)$.
2. Let $\tilde{Z}_{i,t} = Z_{i,t}^2$ and compute

$$\tilde{Y}_{i,t} = \mu_{i,t} + \frac{\mu_{i,t}^2 \tilde{Z}_{i,t}}{2\phi_{i,t}} - \frac{\mu_{i,t}}{2\phi_{i,t}} \sqrt{4\mu_{i,t}\phi_{i,t}\tilde{Z}_{i,t} + \mu_{i,t}^2 \tilde{Z}_{i,t}^2}.$$

3. Set

$$Y_{i,t} = \begin{cases} \tilde{Y}_{i,t}, & \text{if } U_{2,t,i} \le \dfrac{\mu_{i,t}}{\mu_{i,t} + \tilde{Y}_{i,t}}, \\ \dfrac{\mu_{i,t}^2}{\tilde{Y}_{i,t}}, & \text{otherwise.} \end{cases}$$

Poisson process method For the Poisson distribution without a dispersion model (5.4), an alternative data-generating algorithm based on Fokianos et al. (2020) is available in `glmstarma.sim`. It can be selected by specifying `sampling_method = "poisson_process"` in the `vpoisson` family; see Fokianos (2024) and Fokianos et al. (2020) for details. It is additionally described in Subsection 2.1.1.

Quasi-Poisson models For quasi-Poisson models with time-varying dispersion, simulated via `dglmstarma.sim`, the sampling algorithm is selected through the `sampling_method` argument of the `vquasipoisson` family. The options `"build_up"`, `"chop_down"`, and `"branching"` all draw from a suitably parameterised generalised Poisson distribution (Consul and Jain, 1973), following the implementation in the `RNGforGPD`

package (Li et al., 2021). The option `"negbin"` instead uses the negative-binomial distribution from Equation (5.5) with $\nu_{i,t} = \mu_{i,t}/(\phi_{i,t} - 1)$. The methods `"branching"` and `"negbin"` cannot reproduce underdispersion ($\phi_{i,t} < 1$); in that case they fall back to the standard Poisson distribution.

Quasi-Binomial models Simulation of quasi-binomial models differs between the over- and underdispersed regimes. Writing $Y_{i,t} \mid \mathcal{F}_{t-1} \sim \text{QuasiBin}(n_i, \pi_{i,t}, \phi_{i,t})$ with

$$\mathbb{E}(Y_{i,t} \mid \mathcal{F}_{t-1}) = n_i \pi_{i,t}, \qquad \text{Var}(Y_{i,t} \mid \mathcal{F}_{t-1}) = \phi_{i,t}\, n_i \pi_{i,t}(1 - \pi_{i,t}),$$

the overdispersed case ($\phi_{i,t} \geq 1$) is handled following Ahn and Chen (1995). One generates $n_i + 1$ independent Bernoulli variables $Z_j \sim \text{Bin}(1, \pi_{i,t})$ for $j = 0, \ldots, n_i$ and n_i independent Bernoulli variables $\tilde{Z}_j \sim \text{Bin}(1, \rho)$ for $j = 1, \ldots, n_i$, where $\rho = \sqrt{(\phi_{i,t} - 1)/(n_i - 1)}$. Setting $\tilde{Y}_j = \tilde{Z}_j Z_0 + (1 - \tilde{Z}_j) Z_j$ then yields $Y_{i,t} = \sum_{j=1}^{n_i} \tilde{Y}_j$, which has the correct mean and variance exactly.

For the underdispersed case ($\phi_{i,t} < 1$), a normal approximation is employed:

$$\tilde{Y}_{i,t} \sim \mathcal{N}\big(n_i \pi_{i,t},\ \phi_{i,t}\, n_i \pi_{i,t}(1 - \pi_{i,t})\big),$$

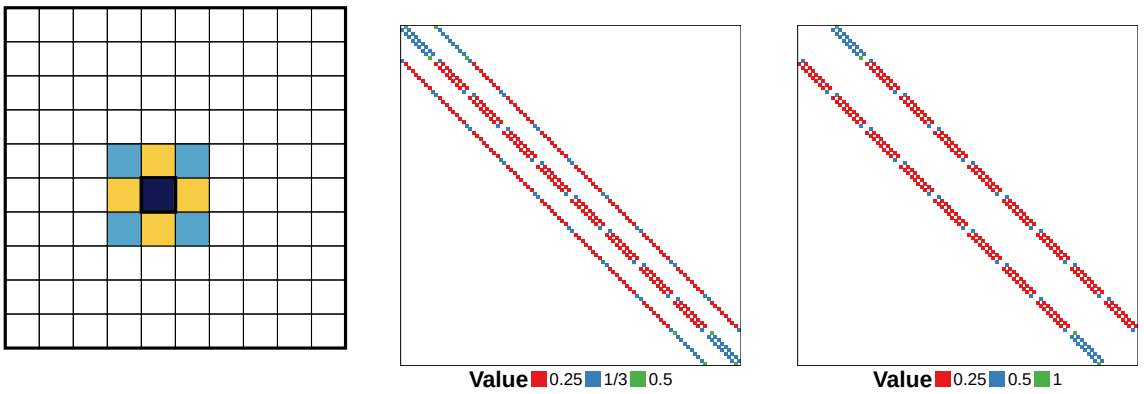
and $Y_{i,t}$ is set to the integer in $\{0, 1, \ldots, n_i\}$ closest to $\tilde{Y}_{i,t}$. This approximation may be inaccurate when $\pi_{i,t}$ is close to 0 or 1.

6.2.3 Examples

We illustrate the simulation functionality with two examples on a regular 10×10 grid ($p = 100$), including neighbours up to spatial order 2. The neighbourhood structure is depicted in Figure 6.1: Panel 6.1a highlights, for a representative central location (highlighted in dark), its first-order neighbours (highlighted in green) and second-order neighbours (highlighted in teal), with equal weights within each order. The helper function `generateW` constructs the corresponding weight matrices via `generateW("rectangle", dim = 100, maxOrder = 2, width = 10)`, returning a list of row-normalised matrices $\mathbf{W}^{(\ell)} \in \mathbb{R}^{p \times p}$ for $\ell = 0, 1, 2$. Panels 6.1b and 6.1c visualise $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$, respectively; $\mathbf{W}^{(0)}$ is the 100×100 identity matrix.

Linear PSTARMA Process

In the first example, data are generated from the linear PSTARMA process introduced in Chapter 3. Since that model specifies only the conditional mean, we use `glmstarma.sim`. Contemporaneous dependence is induced via the Poisson process method of Fokianos



(a) Neighbours (b) Spatial weights $\mathbf{W}^{(1)}$ (c) Spatial weights $\mathbf{W}^{(2)}$

Figure 6.1: Neighbourhood structure generated by `generateW` on a uniform 10×10 rectangular grid up to second-order neighbours.

et al. (2020) with a Frank copula with parameter 2. We simulate a process of length $T = 150$ according to

$$\boldsymbol{\mu}_t = 1 \cdot \mathbf{1}_{100} + 0.2\,\boldsymbol{\mu}_{t-1} + \big(0.3\,\mathbf{I}_{100} + 0.2\,\mathbf{W}^{(1)} + 0.1\,\mathbf{W}^{(2)}\big)\mathbf{Y}_{t-1} + 0.1\,\mathbf{Y}_{t-7}.$$

The corresponding R code is given in Listing 1. The function returns a named list with the simulated observations (`observations`), the linear predictor values (`link_values`), the model orders (`model`), and the true parameters (`parameters`).

Under stationarity, the unconditional global mean, see the discussion in A.1.2 of the Appendix, equals

$$\mu = \frac{1}{1 - 0.2 - 0.3 - 0.2 - 0.1 - 0.1} = 10,$$

which agrees closely with `mean(sim$observations)` $\approx$ 10.0264. Using the true conditional means, the average Pearson residual `mean((sim$observations - sim$link_values)^2 / sim$link_values)` is approximately 0.9963, consistent with the Poisson assumption.

Overdispersion can be introduced by replacing the `vpoisson` family. For instance, `vquasipoisson` with a `dispersion` argument yields draws from the generalised Poisson distribution, and `vnegative.binomial` produces negative-binomial marginals.

Continuous-data Process

The second example simulates data on the same spatial grid with marginal normal distributions,

$$Y_{i,t} \mid \mathcal{F}_{t-1} \sim \mathcal{N}\big(\mu_{i,t},\, \phi_{i,t}\big),$$

```r
set.seed(42)
sim <- glmstarma.sim(
    150,
    parameters = list(
        intercept = 1,
        past_mean = 0.2,
        past_obs  = cbind(
            c(0.3, 0.2, 0.1),
            c(0.1, 0,   0  )
        )
    ),
    model = list(
        intercept = "homogeneous",
        past_mean = 0,
        past_obs = c(2, 0),
        past_obs_time_lags = c(1, 7)
    ),
    family = vpoisson(
        "identity",
        copula = "frank",
        copula_param = 2,
        sampling_method = "poisson_process"
    ),
    wlist = W
)
```

Listing 1: Simulation of a linear PSTARMA process.

using a log link $\boldsymbol{\psi}_t = \log(\boldsymbol{\mu}_t)$ and a Joe copula (parameter 1.5) for contemporaneous dependence. The conditional mean follows

$$\boldsymbol{\psi}_t = 0.1 \cdot \mathbf{1}_{100} + \big(0.3\,\mathbf{I}_{100} + 0.2\,\mathbf{W}^{(1)} + 0.1\,\mathbf{W}^{(2)}\big)\mathrm{Y}_{t-1} + 0.3\sin\!\Big(\tfrac{2\pi}{52}t\Big)\,\mathbf{1}_{100} - 0.2\cos\!\Big(\tfrac{2\pi}{52}t\Big)\,\mathbf{1}_{100},$$

and the dispersion process exhibits GARCH-like dynamics with a seasonal component and a spatial gradient in the column direction. Letting $\mathbf{X} = (1, \dots, 10)' \otimes \mathbf{1}_{10}$ denote the column-index covariate, the dispersion model reads

$$\boldsymbol{\phi}_t = 2 \cdot \mathbf{1}_{100} + \big(0.1\,\mathbf{I}_{100} + 0.05\,\mathbf{W}^{(2)}\big)\boldsymbol{\phi}_{t-1} + 0.3\big(\mathrm{Y}_{t-1} - \boldsymbol{\mu}_{t-1}\big)^2 + 2\exp\!\Big(\sin\!\Big(\tfrac{2\pi}{26}t\Big)\Big)\,\mathbf{1}_{100} + 0.5\,\mathbf{X}.$$

This specification yields a dispersion that varies in both space and time. Simulation is performed with `dglmstarma.sim`; the R code is given in Listing 2.

```r
## Define covariates
covariates_mean <- list(
    sine   = SpatialConstant(sin(2 * pi / 52 * seq(250))),
    cosine = SpatialConstant(cos(2 * pi / 52 * seq(250)))
)
covariates_dispersion <- list(
    sine   = SpatialConstant(exp(sin(2 * pi / 26 * seq(250)))),
    column = TimeConstant(c(1:10 %x% rep(1, 10)))
)

## Mean model
mean_model <- list(
    intercept  = "homogeneous",
    past_obs   = matrix(c(1, 1, 1), ncol = 1),
    covariates = c(0, 0)
)
mean_parameters <- list(
    intercept  = 0.1,
    past_obs   = matrix(c(0.3, 0.2, 0.1), ncol = 1),
    covariates = matrix(c(0.3, -0.2), nrow = 1)
)

## Dispersion model
dispersion_model <- list(
    intercept  = "homogeneous",
    past_mean  = matrix(c(1, 0, 1), ncol = 1),
    past_obs   = 0,
    covariates = c(0, 0)
)
dispersion_parameters <- list(
    intercept  = 2,
    past_mean  = matrix(c(0.1, 0, 0.05), ncol = 1),
    past_obs   = matrix(0.3),
    covariates = matrix(c(2, 0.5), nrow = 1)
)

## Simulation
sim <- dglmstarma.sim(
    250,
    mean_parameters,
    dispersion_parameters,
    mean_model,
    dispersion_model,
    mean_family = vnormal("log", copula = "joe", copula_param = 1.5),
    dispersion_link = "identity",
    wlist = W,
    mean_covariates = covariates_mean,
    dispersion_covariates = covariates_dispersion
)
```

Listing 2: Simulation of a DGLM-STARMA model with covariates and normal marginals.

7 Simulation

> "*Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin.*"
>
> — John von Neumann

In this chapter, we present simulation results for the methods introduced in the preceding chapters. Section 7.1 considers the PSTARMA processes from Chapter 3, Section 7.2 examines the level-shift detection procedure from Chapter 4 within this framework, and Section 7.3 addresses the spatio-temporal Double Generalized Linear Models from Chapter 5. All simulations were conducted using the R package `glmSTARMA` introduced in Chapter 6.

All simulations are carried out on regular rectangular grids with an equal number of locations in both directions, i.e., square grids of size $k \times k$ which results in $p = k^2$. Each setting is replicated 1000 times, with the goal of assessing the finite-sample performance of the derived theoretical results.

We restrict attention to first-order spatial neighbourhoods, i.e., dependencies among directly adjacent locations, as illustrated in Figure 6.1. In addition, some simulations employ anisotropic spatial dependencies: one neighbourhood captures dependencies in the north–south direction, and another captures dependencies in the west–east direction. In total, we use three distinct weight matrices: $\mathbf{W}_{4\text{NN}} = (w_{ij}^{4\text{NN}})$ for dependencies in all four directions, $\mathbf{W}_{\text{NS}} = (w_{ij}^{\text{NS}})$ for column-wise dependencies among northern and southern neighbours, and $\mathbf{W}_{\text{WE}} = (w_{ij}^{\text{WE}})$ for row-wise dependencies among western and eastern neighbours.

The weights w_{ij}^{WE} are defined as

$$w_{ij}^{\text{WE}} = \begin{cases} 0, & \text{if } |i-j| \neq k, \\ 1, & \text{if } i \in \{1, \dots, k, k(k-1)+1, \dots, k^2\} \text{ and } |i-j| = k, \\ 0.5, & \text{otherwise,} \end{cases} \tag{7.1}$$

where a weight of 1 is assigned to locations on the left or right boundary of the grid. Similarly, the weights w_{ij}^{NS} are defined as

$$w_{ij}^{\text{NS}} = \begin{cases} 0, & \text{if } |i-j| \neq 1, \\ 1, & \text{if } i \in \{\ell(k-1)+1,\ \ell k;\ \ell = 1,\dots,k\} \text{ and } |i-j| = 1, \\ 0.5, & \text{otherwise,} \end{cases} \tag{7.2}$$

where a weight of 1 is assigned to locations on the top or bottom boundary of the grid. In the 4-nearest-neighbour case, the weights are given by

$$w_{ij}^{\text{4NN}} = \begin{cases} 0, & \text{if } w_{ij}^{\text{WE}} = w_{ij}^{\text{NS}} = 0, \\ 0.5, & \text{if } w_{ij}^{\text{WE}} = 1 \wedge w_{ij}^{\text{NS}} = 1, \\ 1/3, & \text{if } (w_{ij}^{\text{WE}} = 1 \wedge w_{ij}^{\text{NS}} = 0.5) \vee (w_{ij}^{\text{WE}} = 0.5 \wedge w_{ij}^{\text{NS}} = 1), \\ 0.25, & \text{otherwise,} \end{cases} \tag{7.3}$$

where $\wedge$ denotes the logical "and" and $\vee$ denotes the logical "or".

7.1 Spatio-temporal Count Autoregression

This section presents simulation results for the linear and log-linear PSTARMA(X) processes. In particular, we examine different strategies for initializing the feedback mechanism when it is present in the model, and assess the consequences of misspecifying individual model components, i.e., of choosing an incorrect neighbourhood structure. Additional results for this model class are deferred to the Appendix but briefly discussed here.

The simulations were carried out using a preliminary version of the R package which had higher RAM requirements than the version currently available on CRAN; computation times and estimation results are not materially affected by this. Table 7.1 summarises all simulation settings considered in this section. Contemporaneous dependencies between observations during data generation, see Section 2.1.1, were induced via a Clayton copula with parameter 2.

7.1.1 Computational Resources

The results presented here were obtained on a compute cluster of the Department of Statistics at TU Dortmund University using `R 4.3.2`, running on a node equipped with two AMD EPYC 7453 processors and a maximum allocated RAM of 48 GB per setup. Computation times were measured using the R package `tictoc`.

Table 7.2 reports mean computation times in seconds for linear and log-linear first-order PSTARMAX processes across different grid sizes and numbers of observation times,

<table>
<tr><th>Study</th><th>Description</th><th>Model Orders & Data</th><th colspan="5">True Parameters</th></tr>
<tr><td rowspan="2">Feedback Initialization</td><td rowspan="2">Impact of different initialisations of the feedback term on parameter estimation.
• Linear:
– $\boldsymbol{\mu}_i = \mathbf{Y}_i$
– $\boldsymbol{\mu}_i = \frac{1}{T}\sum_{t=1}^{T}\mathbf{Y}_t$
– $\boldsymbol{\mu}_i = \mathbf{0}_p$
• Log-Linear:
– $\boldsymbol{\psi}_i = \log(\mathbf{Y}_i + \mathbf{1}_p)$
– $\boldsymbol{\psi}_i = \log(\bar{\mathbf{Y}} + \mathbf{1}_p)$
– $\boldsymbol{\psi}_i = \mathbf{0}_p$</td><td rowspan="2">• $T = 250$
• Grid size: 9×9 ($p = 81$)
• Maximum time lags: $(q, r) \in \{(1,1),(2,2)\}$
• Number of covariates: $m \in \{0, 1\}$
• Neighbourhood matrices: $\mathbf{W}^{(0)} = \mathbf{I}_p$, $\mathbf{W}^{(1)} = \mathbf{W}_{\text{4NN}}$</td><td colspan="5">linear</td></tr>
<tr><td>δ_0</td><td>5</td><td>5</td><td>5</td><td>5</td></tr>
<tr><td></td><td>$\alpha_{0,1}$</td><td>0.2</td><td>0.2</td><td>0.2</td><td>0.2</td></tr>
<tr><td></td><td>$\alpha_{1,1}$</td><td>0.1</td><td>0.1</td><td>0.1</td><td>0.1</td></tr>
<tr><td></td><td>$\alpha_{0,2}$</td><td>-</td><td>0.05</td><td>-</td><td>0.05</td></tr>
<tr><td></td><td>$\alpha_{1,2}$</td><td>-</td><td>0.05</td><td>-</td><td>0.05</td></tr>
<tr><td></td><td>$\beta_{0,1}$</td><td>0.2</td><td>0.2</td><td>0.2</td><td>0.2</td></tr>
<tr><td></td><td>$\beta_{1,1}$</td><td>0.1</td><td>0.1</td><td>0.1</td><td>0.1</td></tr>
<tr><td></td><td>$\beta_{0,2}$</td><td>-</td><td>0.05</td><td>-</td><td>0.05</td></tr>
<tr><td></td><td>$\beta_{1,2}$</td><td>-</td><td>0.05</td><td>-</td><td>0.05</td></tr>
<tr><td></td><td>$\gamma_{0,1}$</td><td>-</td><td>-</td><td>2</td><td>2</td></tr>
<tr><td></td><td></td><td></td><td colspan="5">log-linear</td></tr>
<tr><td></td><td>δ_0</td><td>0.6</td><td>0.6</td><td>0.6</td><td>0.6</td></tr>
<tr><td></td><td>$\alpha_{0,1}$</td><td>0.2</td><td>0.2</td><td>0.2</td><td>0.2</td></tr>
<tr><td></td><td>$\alpha_{1,1}$</td><td>0.1</td><td>0.1</td><td>0.1</td><td>0.1</td></tr>
<tr><td></td><td>$\alpha_{0,2}$</td><td>-</td><td>0.05</td><td>-</td><td>0.05</td></tr>
<tr><td></td><td>$\alpha_{1,2}$</td><td>-</td><td>0.05</td><td>-</td><td>0.05</td></tr>
<tr><td></td><td>$\beta_{0,1}$</td><td>0.2</td><td>0.2</td><td>0.2</td><td>0.2</td></tr>
<tr><td></td><td>$\beta_{1,1}$</td><td>0.1</td><td>0.1</td><td>0.1</td><td>0.1</td></tr>
<tr><td></td><td>$\beta_{0,2}$</td><td>-</td><td>0.05</td><td>-</td><td>0.05</td></tr>
<tr><td></td><td>$\beta_{1,2}$</td><td>-</td><td>0.05</td><td>-</td><td>0.05</td></tr>
<tr><td></td><td>$\gamma_{0,1}$</td><td>-</td><td>-</td><td>0.9</td><td>0.9</td></tr>
<tr><td rowspan="10">Misspecification of Spatial Dependencies</td><td rowspan="10">Different types of spatial dependence and detection of misspecification.</td><td rowspan="10">• $T = 250$
• Grid size: 9×9 ($p = 81$)
• Maximum time lags: $(q, r) = (1, 1)$
• Number of covariates: $m = 0$
• Neighbourhood matrices: $\mathbf{W}^{(0)} = \mathbf{I}_p$, $\mathbf{W}^{(1)} = \mathbf{W}_{\text{NS}}$, $\mathbf{W}^{(2)} = \mathbf{W}_{\text{WE}}$</td><td></td><td colspan="2">linear</td><td colspan="2">log-linear</td></tr>
<tr><td>δ_0</td><td colspan="2">5</td><td colspan="2">0.6</td></tr>
<tr><td>$\alpha_{0,1}$</td><td colspan="2">0.2</td><td colspan="2">0.2</td></tr>
<tr><td>$\alpha_{1,1}$</td><td colspan="2">0.1</td><td colspan="2">0.1</td></tr>
<tr><td>$\alpha_{2,1}$</td><td colspan="2">0.05</td><td colspan="2">0.05</td></tr>
<tr><td>$\beta_{0,1}$</td><td colspan="2">0.2</td><td colspan="2">0.2</td></tr>
<tr><td>$\beta_{1,1}$</td><td colspan="2">0.1</td><td colspan="2">0.1</td></tr>
<tr><td>$\beta_{2,1}$</td><td colspan="2">0.05</td><td colspan="2">0.05</td></tr>
<tr><td colspan="5"></td></tr>
<tr><td colspan="5"></td></tr>
</table>

Table 7.1: Summary of the simulation settings. Covariate values for each location are generated independently as a realisation of a stationary ARMA(1, 1) process with $\mathcal{U}([0, 1])$-distributed innovations, fixed in advance so that they remain identical across all repetitions. Contemporaneous correlations between the counts are induced via the data-generating mechanism of Fokianos et al. (2020) using a Clayton copula with parameter 2.

aggregated over multiple simulation settings. The runtime scales linearly with T and quadratically with the number of locations. The log-linear model incurs slightly higher computation times than the linear model, and the inclusion of the feedback term roughly doubles the runtime, depending on T. Additional computation times for larger T and smaller grid sizes are provided in the Appendix.

Feedback order	Model	Grid size	T				
			5	10	20	50	100
$q = 1$	log-linear	10×10	0.006	0.014	0.027	0.065	0.126
		20×20	0.030	0.057	0.110	0.278	0.548
		30×30	0.087	0.153	0.271	0.623	1.173
		40×40	0.234	0.357	0.575	1.226	2.465
		50×50	0.561	0.810	1.134	2.317	4.084
	linear	10×10	0.004	0.007	0.015	0.039	0.083
		20×20	0.018	0.033	0.066	0.177	0.375
		30×30	0.062	0.102	0.180	0.436	0.895
		40×40	0.181	0.259	0.406	0.930	1.800
		50×50	0.502	0.650	0.860	1.802	3.143
$q = 0$	log-linear	10×10	0.003	0.005	0.010	0.025	0.051
		20×20	0.016	0.026	0.046	0.112	0.231
		30×30	0.058	0.082	0.129	0.286	0.563
		40×40	0.164	0.210	0.304	0.596	1.106
		50×50	0.492	0.550	0.680	1.129	1.863
	linear	10×10	0.002	0.004	0.008	0.019	0.040
		20×20	0.013	0.020	0.035	0.086	0.178
		30×30	0.052	0.067	0.105	0.228	0.446
		40×40	0.162	0.189	0.254	0.478	0.876
		50×50	0.406	0.438	0.579	1.009	1.657

Table 7.2: Mean computation time in seconds to estimate a PSTARMAX process for different grid sizes and numbers of observation times T.

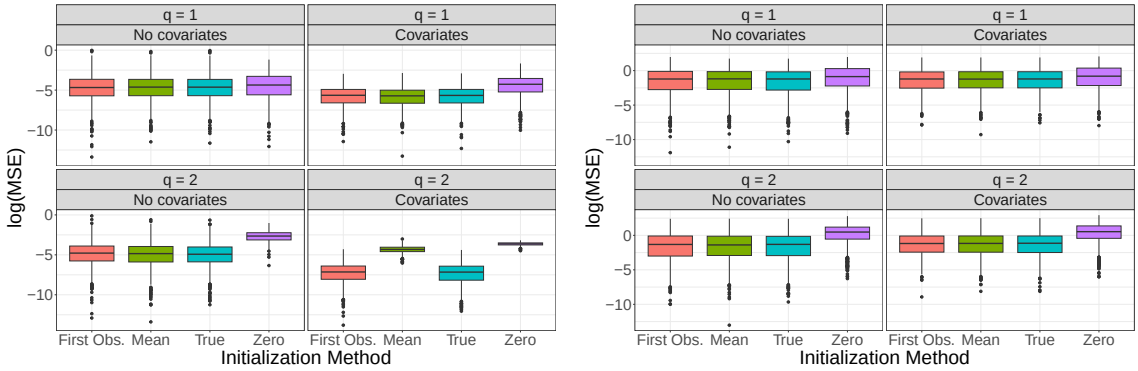
7.1.2 Initialization of the Feedback Process

Processes with a feedback term, i.e., with regression on $\{\boldsymbol{\mu}_t\}$ or $\{\boldsymbol{\psi}_t\}$, require initial values for $\boldsymbol{\mu}_0, \dots, \boldsymbol{\mu}_{q-1}$, and the choice of these values can affect the resulting parameter estimates. Natural candidates are $\boldsymbol{\mu}_i = \bar{\mathbf{y}}$ or $\boldsymbol{\mu}_i = \mathbf{y}_i$, and correspondingly $\boldsymbol{\psi}_i = \log(\bar{\mathbf{y}} + \mathbf{1}_p)$ or $\boldsymbol{\psi}_i = \log(\mathbf{y}_i + \mathbf{1}_p)$ for $i = 0, \dots, q-1$. A further option is to set $\boldsymbol{\mu}_i = \boldsymbol{\psi}_i = \mathbf{0}_p$ for all i. We compare these three strategies and additionally include the true latent values as an oracle benchmark. Performance is assessed via the mean squared error of the QMLE,

$$\text{MSE} = \frac{1}{K}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{true}})'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{true}}),$$

where $K = 1 + \sum_{i=1}^{q}(a_i + 1) + \sum_{i=1}^{r}(b_i + 1)$ is the total number of unknown model parameters.

Figure 7.1a shows the logarithmic MSE for log-linear PSTARMA and PSTARMAX processes with temporal orders 1 and 2 and $T = 250$ observations under the different initialization strategies. For first-order models without covariates, the MSE is essentially



(a) log-linear

(b) linear

Figure 7.1: MSE (on a logarithmic scale) of the QMLE in linear and log-linear model settings for different initialization strategies for the feedback process.

the same across all methods. For second-order models, zero initialization yields noticeably higher MSE than the other strategies. Among the remaining methods, initialization via the global mean performs worse than initialization via the first observations for second-order models with covariates. Notably, initializing with the first observations yields MSEs close to those obtained using the true (oracle) values of the latent process. We therefore recommend this strategy.

The same qualitative pattern holds for linear PSTARMA and PSTARMAX processes (Figure 7.1b), though the differences between methods are less pronounced. Zero initialization again yields the highest MSE. Based on these findings, initialization via the first observations is used throughout the remainder of this work.

7.1.3 Misspecification of Spatial Dependence

The choice of neighbourhood matrices determines which spatial dependencies a model can capture. This study examines the consequences of a mild misspecification: data are generated from a PSTARMA process with an anisotropic neighbourhood structure, where the coefficients of $\mathbf{W}_{\text{NS}}$ and $\mathbf{W}_{\text{WE}}$ differ (0.1 and 0.05 for both observations and feedback, respectively), while estimation is carried out under a misspecified isotropic model with a single neighbourhood matrix assuming equal effects in all directions. We also assess how reliably the anisotropy can be detected.

We first examine whether any spatial dependence is detected when fitting the isotropic model to anisotropically generated data, using Wald tests for whether the estimated spatial parameters $\alpha_{1,1}$ and $\beta_{1,1}$ differ significantly from zero. The empirical rejection rates at the 5% level (Table 7.3) show that spatial dependence in the observation process is detected more readily than in the feedback process. For instance, in the log-linear model

with $T = 250$, spatial dependence in the observations is identified in approximately 88% of replications, but only in around 19% for the feedback process.

Table 7.4 reports the frequency with which QIC, see (5.23), favours the isotropic model over the true anisotropic model. For $T = 50$, the isotropic model is preferred in roughly 44% of cases in the linear setting, decreasing to about 25% for $T = 100$. Across all models, the true anisotropic model is selected more consistently as T grows, consistent with the results of Wald tests for anisotropy fitted under the anisotropic model. These tests assess whether the differences $\alpha_{2,1} - \alpha_{1,1}$ and $\beta_{2,1} - \beta_{1,1}$ between directional spatial parameters deviate significantly from zero (Table 7.5). For $T = 500$, the observation-process parameters are correctly identified as distinct in the vast majority of cases, whereas for the feedback process this occurs in fewer than 20% of replications in both the linear and log-linear models.

Overall, these results indicate that for small sample sizes, isotropy and anisotropy are difficult to distinguish in practice. As a consequence, isotropic models can often capture spatial dependencies reasonably well even when moderate anisotropy is present, and may be preferred by QIC. However, when anisotropic structures are known to exist and sufficient data are available, an anisotropic model should be used.

7.1.4 Further Remarks

Additional simulation results, including an analysis of the asymptotic behaviour of the QMLE and settings with high-dimensional grids and short time series, are provided in the Appendix. A notable finding is that linear PSTARMAX processes exhibit faster convergence of the test statistics than their log-linear counterparts. For models with a feedback term, the asymptotic approximations perform poorly when $T \leq 50$. In contrast, for a linear PSTARMAX process without feedback ($q = 0$), good approximations are already obtained for $T = 5$.

Further simulations based on the Master's thesis of Maletz (2021) are included in the Appendix. These investigate the impact of contemporaneous dependencies on the QMLE as well as additional types of misspecification. As expected, stronger contemporaneous

Model	Coefficient	T			
		50	100	250	500
log-linear	$\beta_{1,1}$	0.264	0.488	0.876	0.990
	$\alpha_{1,1}$	0.173	0.153	0.186	0.304
linear	$\beta_{1,1}$	0.345	0.627	0.944	0.999
	$\alpha_{1,1}$	0.147	0.180	0.311	0.430

Table 7.3: Empirical rejection rates at the 5% level for $\beta_{1,1}$ and $\alpha_{1,1}$ in an isotropic model fitted to data generated from an anisotropic model.

Model	T			
	50	100	250	500
log-linear	0.361	0.257	0.100	0.027
linear	0.444	0.247	0.075	0.017

Table 7.4: Relative frequency with which QIC favours the isotropic model over the true anisotropic model.

Model	Anisotropy test	T			
		50	100	250	500
log-linear	β	0.295	0.465	0.815	0.980
	α	0.122	0.124	0.126	0.173
linear	β	0.216	0.510	0.876	0.992
	α	0.040	0.051	0.092	0.146

Table 7.5: Empirical power of Wald tests at the 5% level for differences between directional spatial coefficients in the anisotropic model; α denotes the test for the feedback process, β for the observation process.

dependencies increase the MSE of the QMLE, though the effect on the bias is small. Log-linear PSTARMA processes are found to partially capture the structure of linear PSTARMA processes, whereas linear models cannot represent negative dependencies; fitting both model types is therefore advisable in practice. Finally, homogeneous models with a global intercept applied to data from an inhomogeneous model perform competitively at small sample sizes, suggesting that a homogeneous model augmented with location-specific covariates may be a practical alternative.

7.2 Detection of Persistent Level Shifts

We examine the behavior of the procedure introduced in Chapter 4 under several scenarios involving zero, one, or two persistent level shifts. For each scenario, we consider two distinct settings: one in which only the time point of the level shift is unknown while the affected locations are assumed to be known to be global, and one in which both the time points and the affected locations are unknown, though the true location vectors are included among the candidates. Since bootstrapping did not yield a satisfactory approximation of the null distribution, all results reported below are size-adjusted based on the empirical quantiles obtained under the null hypothesis.

Data are generated on the grid described above from the following log-linear models. Each of the 1000 simulation iterations produces $T = 500$ time points. Contemporaneous dependence is introduced via the data-generating process described in Section 2.1.1, using a Frank copula with parameter 2.

$$\boldsymbol{\psi}_t = 0.5\mathbf{1}_p + (0.3\mathbf{I}_p + 0.2\mathbf{W}^{(1)}) \log(\mathbf{Y}_{t-1} + \mathbf{1}_p) + 0.75\mathbf{X}_{1,t} + 0.5\mathbf{X}_{2,t} + \sum_{k=1}^{2} \gamma_k \mathbf{1}(t \geq \tau_k)\, \mathbf{v}_k \tag{7.4}$$

$$\boldsymbol{\psi}_t = 0.5\mathbf{1}_p + (0.1\mathbf{I}_p + 0.05\mathbf{W}^{(1)})\boldsymbol{\psi}_{t-1} + (0.3\mathbf{I}_p + 0.2\mathbf{W}^{(1)}) \log(\mathbf{Y}_{t-1} + \mathbf{1}_p) + 0.75\mathbf{X}_{1,t} + 0.5\mathbf{X}_{2,t} + \sum_{k=1}^{2} \gamma_k \mathbf{1}(t \geq \tau_k)\, \mathbf{v}_k \tag{7.5}$$

The covariate processes are constructed as follows. The first covariate, $\mathbf{X}_{1,t} = \sin\left(\frac{2\pi}{12}t\right)\mathbf{1}_p$, induces a seasonal pattern. The second covariate is a fixed spatial vector $\mathbf{X}_{2,t} = \mathbf{X} \in \mathbb{R}^p$ with $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.9)$, drawn once and held constant across all iterations and models.

We consider three shift scenarios: (1) no level shift, i.e. $\gamma_1 = \gamma_2 = 0$; (2) a single level shift with $\gamma_1 = 0.3$ and $\gamma_2 = 0$; and (3) two level shifts with $\gamma_1 = 0.3$ and $\gamma_2 = -0.2$. In all cases, the time points of the shifts are fixed at $\tau_1 = 100$ and $\tau_2 = 200$.

Regarding the spatial structure, we distinguish two cases. In the *global* case, $\mathbf{v}_1 = \mathbf{v}_2 = \mathbf{1}_{100}$, so each shift affects all locations simultaneously. In the *local* case, $\mathbf{v}_1 = (\mathbf{1}'_{50}, \mathbf{0}'_{50})'$ and $\mathbf{v}_2 = (\mathbf{0}'_{50}, \mathbf{1}'_{50})'$, so each shift affects only half of the locations.

Following the procedure described in Chapter 4, each model is estimated on the simulated data under the null hypothesis, i.e. without level-shift regressors. The temporal search grid is set to $\{10, \ldots, 490\}$. In the global level-shift scenario, only the time points of the shifts is searched, treating the affected locations as known. In the local level-shift scenario, the two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are included as spatial candidates. The procedure described in Section 4.2 to estimate the time-points and locations of the shifts is run for exactly four iterations in all cases even when the procedure would terminate earlier.

Global level shifts. Figure 7.2 shows the distribution of candidate time points selected at each iteration for model (7.4), across the three shift scenarios. Under the null hypothesis of no shifts, estimated time points tend to cluster near the boundaries of the search grid, an artifact that diminishes with increasing iteration number. When a single level shift is present, the procedure recovers the true change point $\tau_1 = 100$ accurately and with little spread. When two shifts are present, the second change point $\tau_2 = 200$ is likewise recovered accurately in the second iteration. Once the iteration count exceeds the true number of shifts, subsequent candidates tend to cluster near previously identified time points. Note, however, that the distribution of estimated candidate time points alone does not determine which effects are statistically significant; this is governed by the sequential testing step.

The empirical size-adjusted detection rates are reported in Table 7.6. Using the empirical 95% quantile of the maximum score statistic from the first iteration under the null as

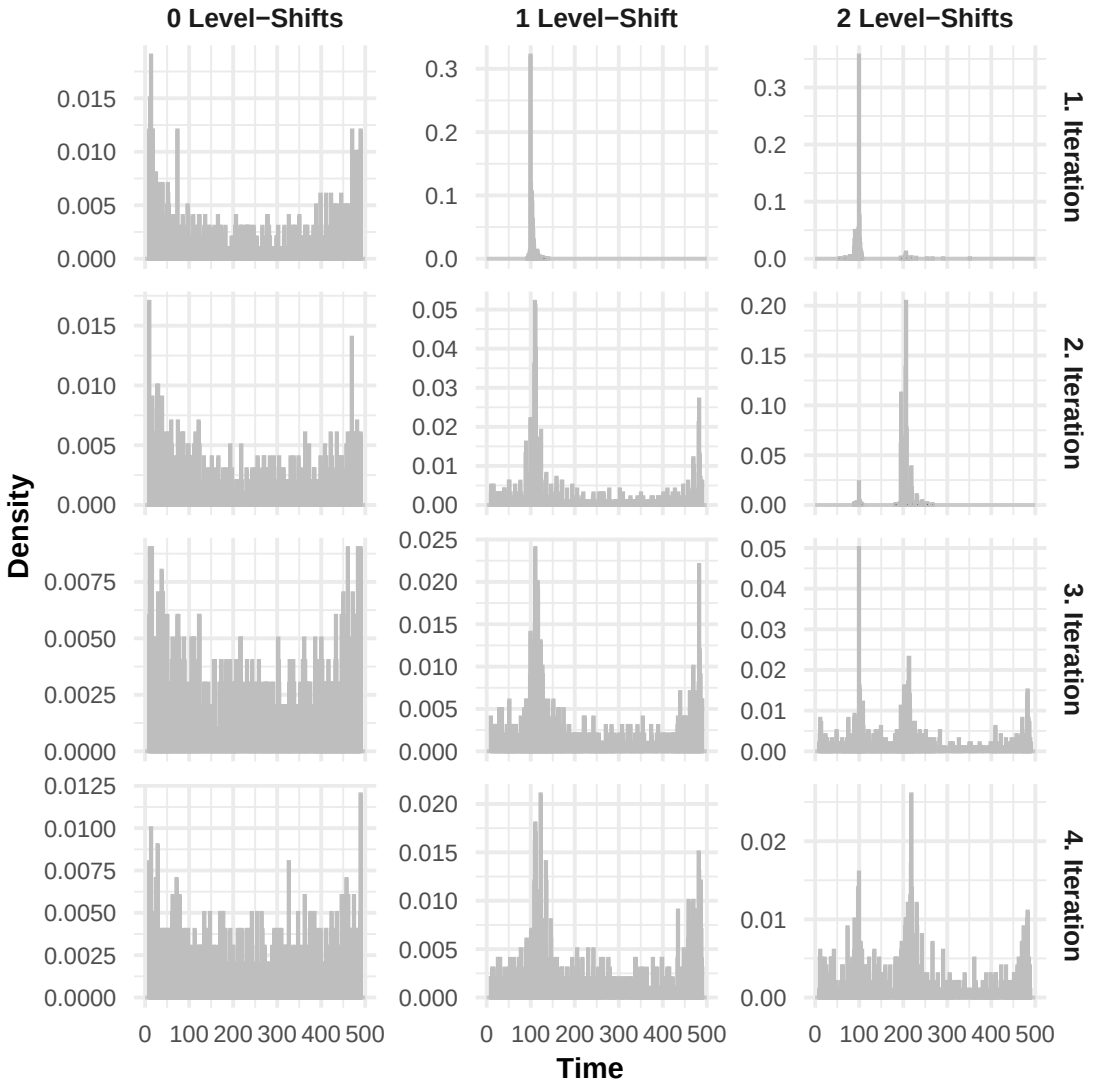


Figure 7.2: Distribution of estimated change points after each of four iterations, for zero, one, or two global level shifts in a setting with *global* level shifts having an effect on all locations. True model (7.4) is without feedback process.

Model	Number of LS	Iteration			
		1	2	3	4
(7.4)	0	0.050	0.000	0.000	0.000
	1	1.000	0.000	0.000	0.000
	2	1.000	1.000	0.006	0.000
(7.5)	0	0.050	0.003	0.000	0.000
	1	1.000	0.056	0.019	0.019
	2	1.000	1.000	0.060	0.007

Table 7.6: Empirical size-adjusted detection rates for *global* level shifts using the sequential procedure described in Section 4.2. Each cell reports the fraction of replications in which a level shift is detected at the given iteration provided that in the previous iteration a level shift was detected. Values in iterations beyond the true number of shifts correspond to false detection rates.

the critical value, the procedure terminates after the first iteration in all replications under the null, yielding an empirical false detection rate of exactly 5% by construction. With one level shift present, the shift is detected in every replication and the procedure consistently terminates after a single step. With two shifts present, both are detected in every replication, while a spurious third shift is flagged in only 0.6% of replications.

Results for model (7.5) are broadly similar, as shown in the lower panel of Table 7.6. Under the null, only 0.3% of replications proceed beyond the first iteration. In the one- and two-shift settings, all true shifts are detected in every replication. After the true shifts have been identified, the empirical false detection rates for a spurious additional shift are 5.6% and 6.0%, respectively. The corresponding distribution of candidate time points is shown in Figure A-12 in Subsection A.3.3 of the Appendix.

Local level shifts. When level shifts affect only a subset of locations, the results deteriorate noticeably. Under the null hypothesis of no level shifts, the procedure behaves similarly to the global case: it terminates after at most one iteration, and the empirical false detection rate remains at the nominal 5% by construction.

With a single local shift present, the correct spatial candidate is selected in 92.5% of replications. However, the estimated change point is severely biased, frequently collapsing to the boundary of the search grid, see Figure 7.3. While the null hypothesis is still rejected in all replications at the first iteration, the false detection rate rises sharply in subsequent iterations, reaching 42.9% at the second iteration for model (7.4).

The same pattern persists when two local shifts are present. Both shifts are detected at the first two iterations, but false detection rates in later iterations reach problematic levels: 87.7% and 36.6% at iterations 3 and 4, respectively, for model (7.4). Results for model (7.5) follow a qualitatively similar pattern, as shown in Table 7.7.

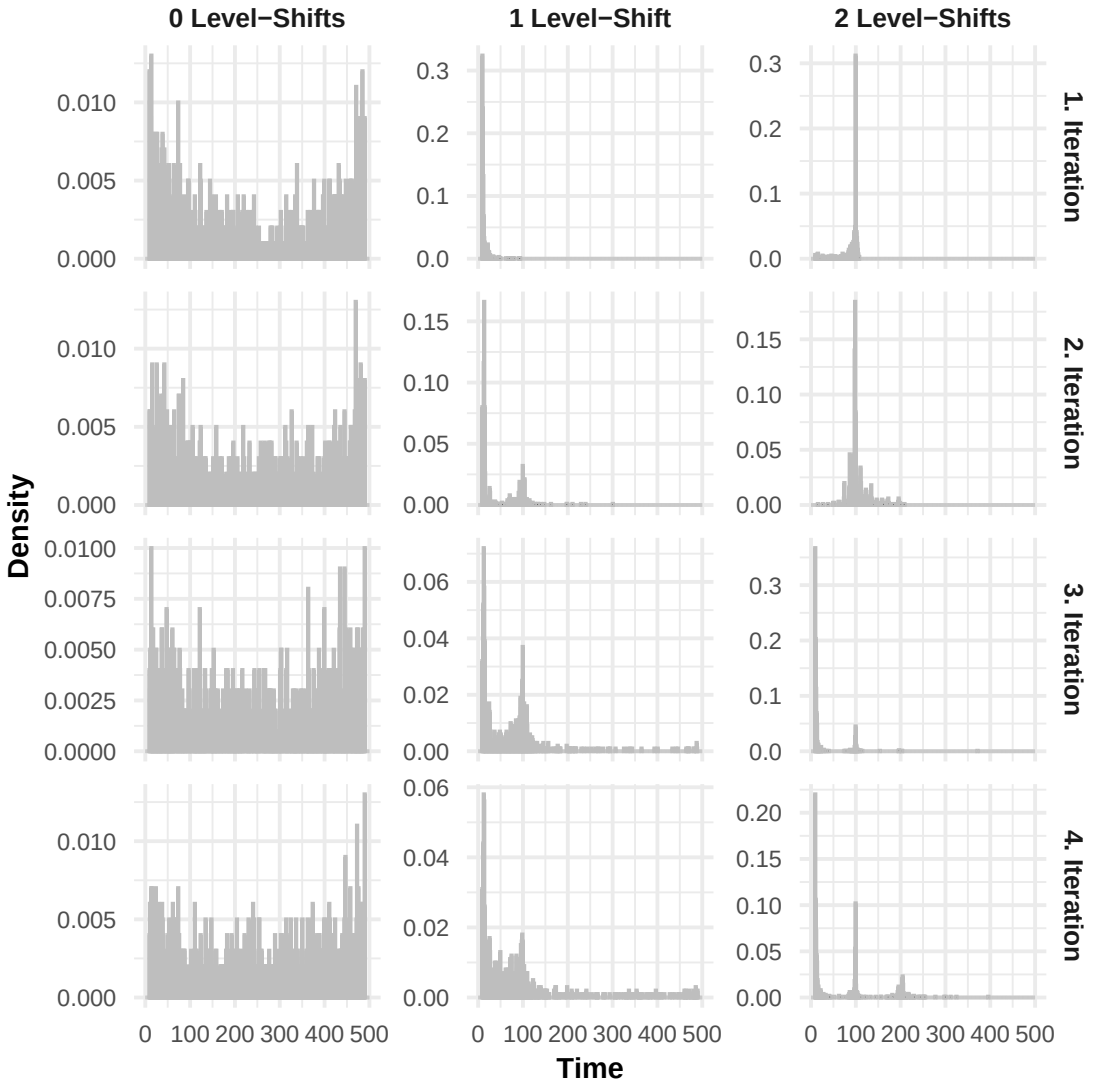


Figure 7.3: Distribution of estimated change points after each of four iterations, for zero, one, or two global level shifts in a setting with *local* level shifts having an effect only on half of the locations. True model (7.4) is without feedback process.

Model	Number of LS	Iteration			
		1	2	3	4
(7.4)	0	0.050	0.010	0.000	0.000
	1	1.000	0.429	0.034	0.000
	2	1.000	0.999	0.877	0.366
(7.5)	0	0.050	0.010	0.000	0.000
	1	1.000	0.964	0.566	0.239
	2	1.000	0.876	0.202	0.075

Table 7.7: Empirical size-adjusted detection rates for *local* level shifts using the sequential procedure described in Section 4.2. Each cell reports the fraction of replications in which a level shift is detected at the given iteration provided that in the previous iteration a level shift was detected. Values in iterations beyond the true number of shifts correspond to false detection rates.

Summary. The simulation study yields two main conclusions. First, when the spatial structure of the level shift is known to be global, the procedure seems to accurately identify both the number and time points of the level shifts, and controls false detection rates close to the nominal level. Second, performance degrades substantially when the spatial structure is unknown. In this setting, the estimated change points tend to be biased toward the boundaries of the search grid, and false detection rates inflate severely in subsequent iterations.

Taken together, the results suggest that the procedure seems to be well suited for settings in which the affected locations can be specified in advance, but should be applied with caution when there might be multiple spatial candidates. In either case, the practical performance of the procedure depends on the quality of the size adjustment: since the null distribution must be approximated rather than computed analytically, the empirical results reported here represent a best-case benchmark.

Further studies are necessary to improve the performance of the algorithm when multiple spatial candidates are available an to evaluate the performance of the procedure in more general and different settings.

7.3 Spatio-temporal Double Generalized Linear Models

In this section, we assess the theoretical properties stated in Section 5.5 through simulations. We adopt the same spatial grid structure as described at the beginning of this chapter, restricting attention to first-order models on a 10×10 grid ($p = 100$).

7.3.1 Construction

We consider two types of marginal distributions: (i) count data and (ii) marginally Gaussian data. For the count-data experiments, observations are generated from the `vquasipoisson` family with a log link; for the normal-data experiments, the identity link is used. Contemporaneous dependence between locations is induced by a Frank copula with parameter 2.

For the conditional mean, we consider two alternative dynamic specifications:

$$\boldsymbol{\psi}_t = \delta_0 \cdot \mathbf{1}_{100} + (0.2\mathbf{I}_{100} + 0.1\mathbf{W}^{(1)})\,\boldsymbol{\psi}_{t-1} + (0.2\mathbf{I}_{100} + 0.1\mathbf{W}^{(1)})\,h(\mathbf{Y}_{t-1}) + \gamma\,\mathbf{X}_t, \tag{7.6}$$

$$\boldsymbol{\psi}_t = \delta_0 \cdot \mathbf{1}_{100} + (0.4\mathbf{I}_{100} + 0.2\mathbf{W}^{(1)})\,h(\mathbf{Y}_{t-1}) + \gamma\,\mathbf{X}_t, \tag{7.7}$$

where for count data we set $\delta_0 = 0.6$, $\gamma = 0.9$, and $h(\mathbf{Y}_{t-1}) = \log(\mathbf{Y}_{t-1} + \mathbf{1}_{100})$, while for normally distributed data we set $\delta_0 = 5$, $\gamma = 2$, and $h(\mathbf{Y}_{t-1}) = \mathbf{Y}_{t-1}$.

The covariate process $\{\mathbf{X}_t\}$ was generated once and reused across all scenarios using the following R code:

```
set.seed(42)
covariates <- t(replicate(100,
    arima.sim(n = 1000,
              list(ar = c(0.89, -0.3), ma = c(-0.1, 0.28)),
              rand.gen = runif)))
covariates <- covariates / max(covariates)
covariates[covariates < 0] <- 0
```

For the dispersion process, we consider three alternative specifications. The dispersion parameter vector $\boldsymbol{\phi}_t$ is linked to a linear process $\boldsymbol{\zeta}_t$ via the log link (i.e. $\boldsymbol{\phi}_t = \exp(\boldsymbol{\zeta}_t)$):

$$\boldsymbol{\phi}_t = 2 \cdot \mathbf{1}_{100}, \tag{7.8}$$

$$\boldsymbol{\zeta}_t = 0.5 \cdot \mathbf{1}_{100} + (0.15\mathbf{I} + 0.05\mathbf{W}^{(1)})\,\boldsymbol{\zeta}_{t-1} + (0.3\mathbf{I} + 0.2\mathbf{W}^{(1)})\,\log(\mathbf{d}_{t-1}), \tag{7.9}$$

$$\boldsymbol{\zeta}_t = 0.5 \cdot \mathbf{1}_{100} + (0.5\mathbf{I} + 0.2\mathbf{W}^{(1)})\,\log(\mathbf{d}_{t-1}). \tag{7.10}$$

Each combination of mean specification (7.6)–(7.7) with dispersion specification (7.8)–(7.10) defines one data-generating scenario. Table 7.8 summarizes all configurations and the corresponding fitted models. When data are generated under a time-varying dispersion process, we estimate both (i) the correctly specified model with time-varying dispersion and (ii) a misspecified model assuming a single, time-constant global dispersion parameter, thereby examining the robustness of estimation and the consequences of misspecification. When the true dispersion is constant (7.8), only the time-varying model is fitted in order to study the asymptotic behaviour of the overparameterized model under the null hypothesis of no temporal dispersion dynamics.

True Model	Fitted Models	Normal		Generalized Poisson	
		Boxplots	QQ-Plots	Boxplots	QQ-Plots
(7.6) + (7.8)	(7.6) + (7.10)	-	-	-	-
(7.7) + (7.8)	(7.7) + (7.10)	-	-	-	-
(7.6) + (7.9)	(7.6) + (7.8) & (7.6) + (7.9)	Fig. A-18	Fig. A-19	Fig. A-26	Fig. A-27
(7.6) + (7.10)	(7.6) + (7.8) & (7.6) + (7.10)	Fig. A-16	Fig. A-17	Fig. A-22	Fig. A-23
(7.7) + (7.9)	(7.7) + (7.8) & (7.7) + (7.9)	Fig. 7.4	Fig. 7.5	Fig. A-24	Fig. A-25
(7.7) + (7.10)	(7.7) + (7.8) & (7.7) + (7.10)	Fig. A-14	Fig. A-15	Fig. A-20	Fig. A-21

Table 7.8: Overview of data-generating and fitted models, with references to the corresponding simulation figures.

Each scenario is simulated for $T \in \{50, 100, 250, 500, 1000\}$ over 1000 independent replications. For models including a feedback mechanism, i.e. (7.6) and (7.9), the feedback is initialized using the first observations or pseudo-observations.

7.3.2 Results

Table 7.8 provides an overview of all simulation figures. For brevity, only two representative figures for the marginally normal case are shown in the main text; the remaining figures are collected in the Appendix.

The figures contain boxplots of parameter estimates for the mean and dispersion models. For the mean model, estimates obtained under correctly specified and misspecified (constant) dispersion are shown side by side. In all settings, the accuracy of the mean-model estimates improves with T regardless of whether the dispersion is correctly specified. Misspecification, however, entails a cost: the parameter estimates exhibit higher variability, though this difference diminishes when a feedback mechanism is incorporated into at least one model component.

The dispersion-model estimates show greater uncertainty, which may be attributable to the model being estimated on a residual process or to non-identifiability of individual parameters. For marginally normal data, the estimates converge to the true parameter values as T grows. For small sample sizes in the count-data setting, a notable bias is observed: in particular, the intercept and the autoregressive parameters for past pseudo-observations are systematically underestimated.

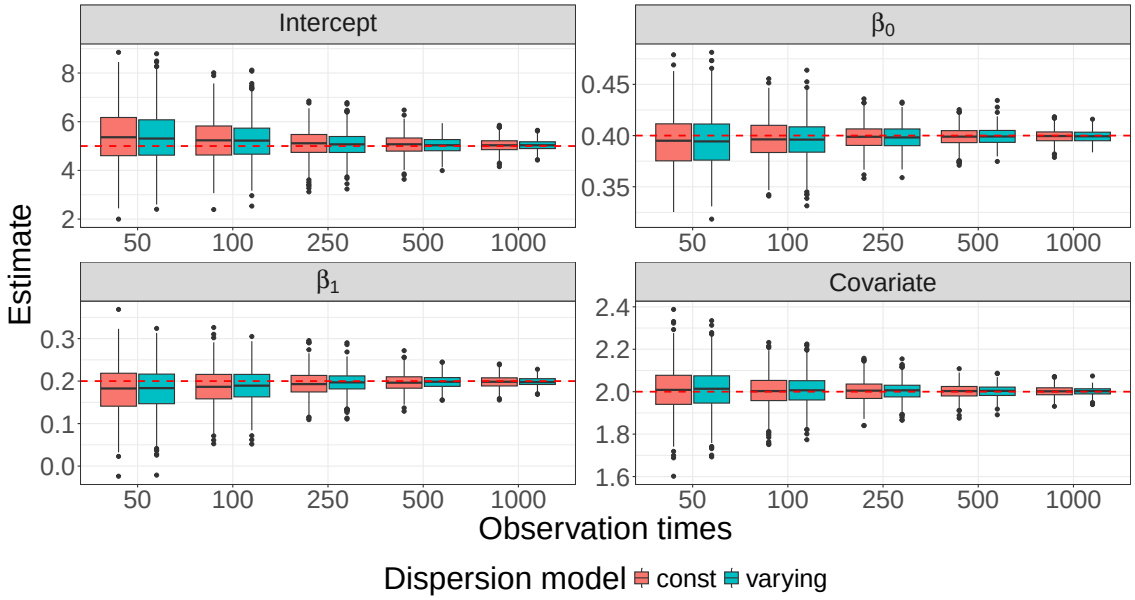
Despite this, normal Q–Q plots of the parameter estimates at $T = 100$ confirm rapid convergence to normality, with the most relevant deviations occurring when a feedback mechanism is present in the respective model component. Table 7.9 reports the empirical sizes of an approximate Wald test at the 5% significance level, derived from the normal approximation (5.16). We test the two null hypotheses $\mathrm{H}_0 : \beta_{0,1} = 0$ and $\mathrm{H}_0 : \beta_{1,1} = 0$ in the first two scenarios of Table 7.8; rejection of either hypothesis would indicate that a constant global dispersion parameter is no longer tenable within the model framework.

Mean model	Distribution	Parameter	T				
			50	100	250	500	1000
(7.7)	Normal	$\beta_{0,1}$	0.064	0.058	0.052	0.051	0.045
		$\beta_{1,1}$	0.062	0.058	0.051	0.046	0.050
	Gen. Poisson	$\beta_{0,1}$	0.066	0.064	0.062	0.051	0.048
		$\beta_{1,1}$	0.062	0.059	0.058	0.054	0.054
(7.6)	Normal	$\beta_{0,1}$	0.066	0.059	0.055	0.051	0.045
		$\beta_{1,1}$	0.063	0.060	0.050	0.047	0.049
	Gen. Poisson	$\beta_{0,1}$	0.068	0.060	0.056	0.054	0.046
		$\beta_{1,1}$	0.064	0.060	0.056	0.054	0.051

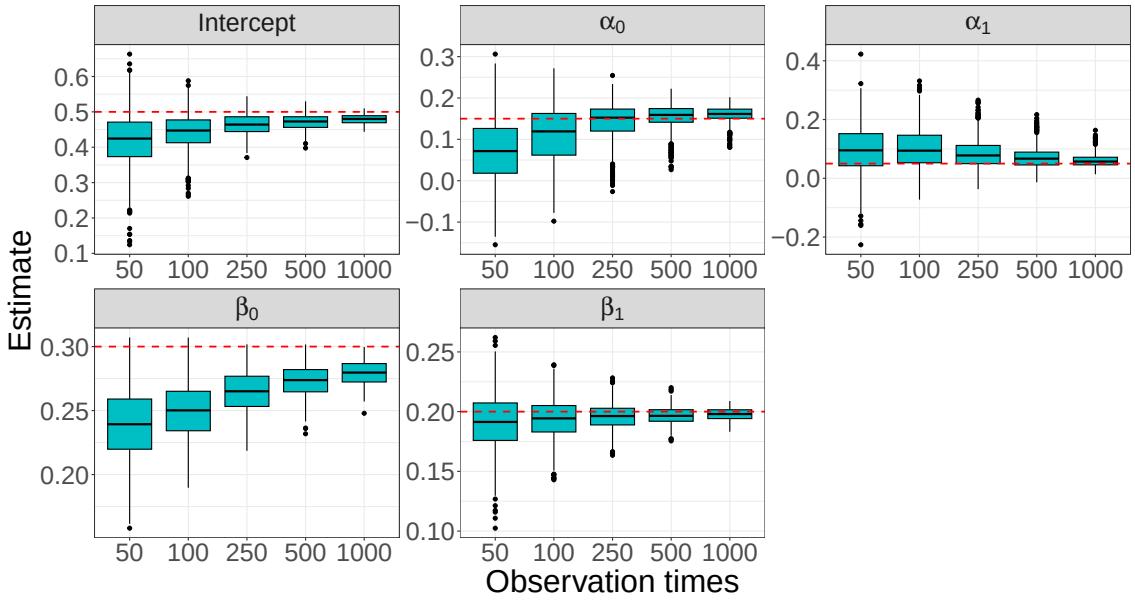
Table 7.9: Empirical size of the Wald test at the 5% significance level for the spatio-temporal dispersion parameters of (7.10) under the null hypothesis of constant dispersion (7.8).

Even for $T = 50$, the empirical size closely approximates the nominal level of 5%, and this accuracy is maintained as T increases.

In summary, jointly modelling the mean and dispersion in this framework allows us to detect and describe spatial and temporal dependence in the mean and the dispersion process. However, the effects in the dispersion process tend to be systematically underestimated when the sample size is small. Further research is therefore needed to determine under which conditions the parameters of the dispersion process can be estimated more reliably. One possible direction is the use of Restricted (Quasi) Maximum Likelihood (REML), originally developed for DGLMs with independent data (Smyth and Verbyla, 1999). In contrast, the parameters of the mean model are already estimated with high accuracy, even when the dispersion model is misspecified, i.e. constant dispersion is assumed.

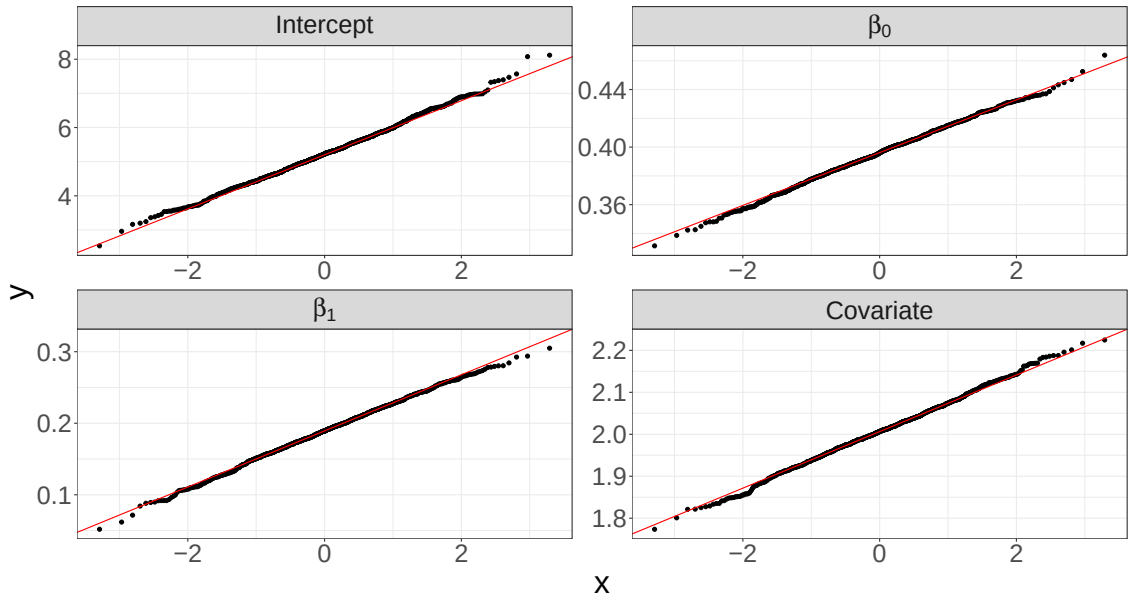


(a) Parameter estimates of the mean model (7.7) under constant dispersion (7.8) and time-varying dispersion (7.9).

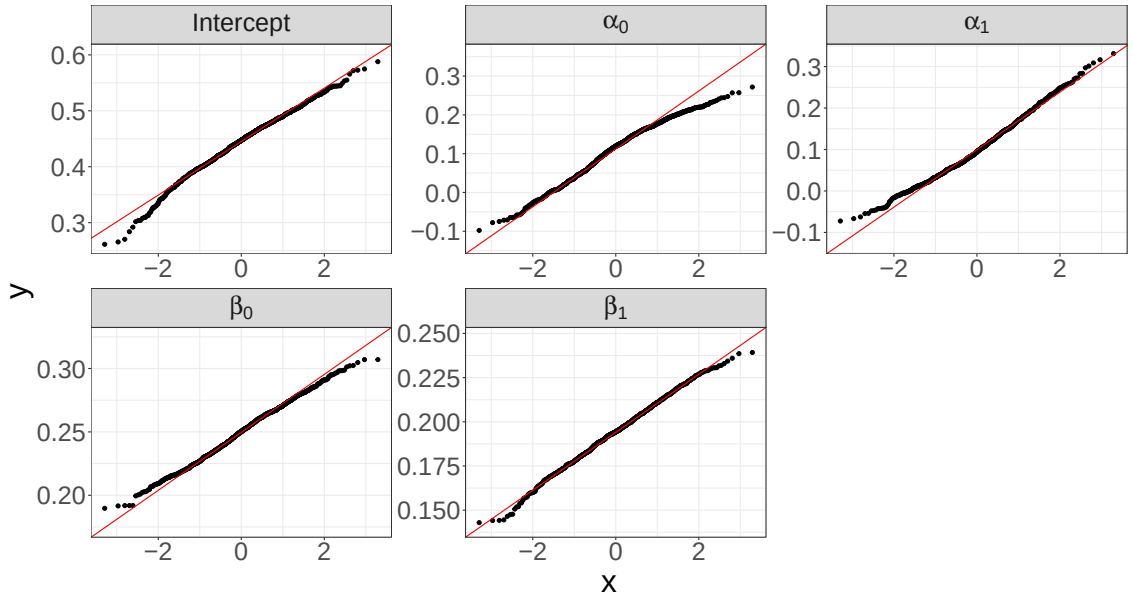


(b) Parameter estimates of the time-varying dispersion model (7.9) combined with the mean model (7.7).

Figure 7.4: Boxplots of parameter estimates under alternative dispersion specifications. Panel (a) shows mean-model estimates under constant and time-varying dispersion; panel (b) shows estimates of the time-varying dispersion model. Data are generated from (7.7) and (7.9) with marginal normal distributions, i.e. the mean model has no feedback mechanism while the dispersion process does.



(a) Mean model (7.7) (no feedback mechanism).



(b) Dispersion model (7.9) (with feedback mechanism).

Figure 7.5: Normal Q–Q plots of parameter estimates under the true data-generating process (7.7) and (7.9) with marginal normal distributions, based on $T = 100$.

8 Data Examples

> "*Torture numbers, and they'll confess to anything.*"
>
> — Gregg Easterbrook

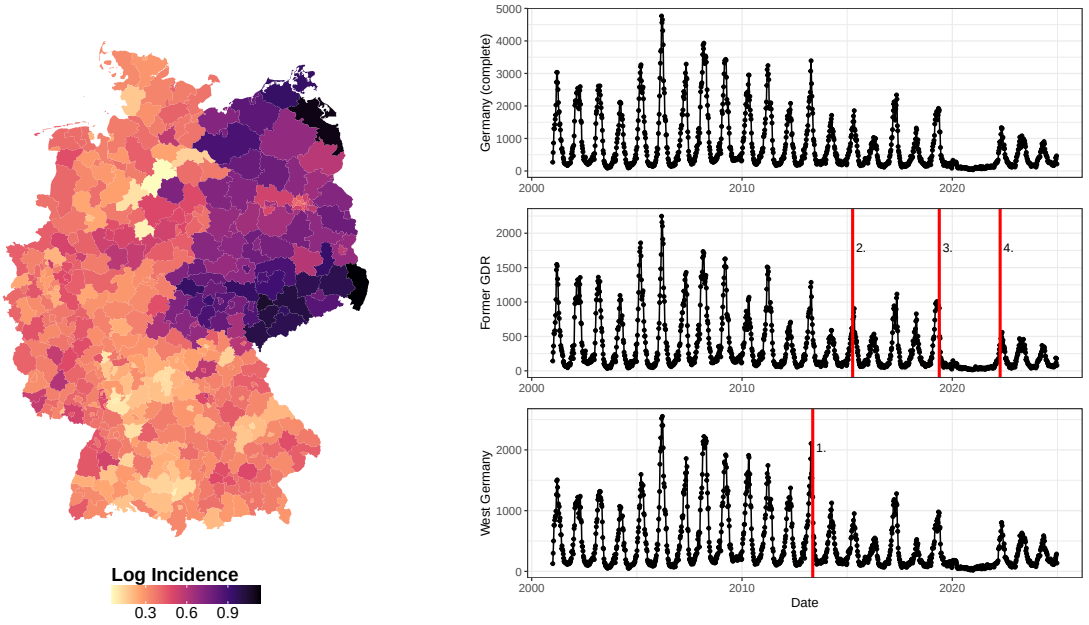
In this chapter, we present two examples in which we apply the developed methodology to the data. In the first example, we look at rotavirus infections in Germany, and in the second example, we look at sea surface temperature anomaly data in the Pacific. While the first example considers count data, in the second example we assume normally distributed data.

8.1 Rota Virus Infections in Germany

As a first example, we analyze weekly reported rotavirus infections in $p = 411$ urban and rural districts in Germany from 2001 to 2024, comprising $T = 1,252$ weekly observations. An earlier version of this dataset (limited to 2018) was examined in our previous work (Maletz et al., 2024). Note that Germany's 2021 territorial reform merged the city of Eisenach into Wartburg district, reducing the total number of districts by one. The data were obtained from the Robert Koch Institute's SurvStat@RKI 2.0 platform (`https://survstat.rki.de/`). It can be loaded in preprocessed form in the `glmSTARMA` package via `load_data("rota")`. Demographic information was derived from the Federal Statistical Office of Germany (Destatis). Such population information is only available on a yearly basis for each district. For each district we interpolated the population size between two years linearly to have a smooth transition and get weekly estimates.

The rotavirus is one of the most frequent causes of diarrhoeal diseases among children, according to the Robert Koch Institut[1]. The viruses are easily transmitted through smear infections, contaminated food or water. After an incubation period of around 1 to 3 days, the symptoms typically last for 2 to 6 days. Infected people are usually infectious for up to 8 days, but individual cases may remain infectious much longer. Vaccination for infants under the age of 6 months is recommended in Germany since July 2013.

[1] Robert Koch-Institut (2019) RKI-Ratgeber. Rotaviren-Gastroenteritis.



(a) Log Mean Incidence of Rota Virus infections in Germany.

(b) Cumulative Rota Virus infections for entire Germany, former GDR regions, and West Germany.

Figure 8.1: Rota Virus in Germany.

The administrative subdivision of Germany into the urban and rural districts gives rise to an irregular grid on a natural way. We consider two regions to be neighbours if they share a common border. The number of neighbours of a region varies from 1 to 12, with an average of 5.250. In total, there are 27 regions that only have one neighbour in the German territory. These are regions completely included in another region or located on the border of Germany. Second order neighbors are defined as the regions, that can be reached by crossing one directly neighbored region, recall Figure 2.1c.

The reported infection numbers within the individual regions varies from 0 to 192. On average, 1.885 cases are registered per region and week. Figure 8.1a shows the spatial distribution of the local incidences on a logarithmic scale, i.e., the number of cases per 100,000 inhabitants, averaged over all time points. It reveals high values for regions belonging to the former German Democratic Republic (GDR) in the northeast of Germany. These regions were administered in a fundamentally different way politically and structurally than the other regions for a long period of time.

Figure 8.1b shows the cumulative weekly infection cases for Germany in total, for the former GDR regions and for the western part of Germany. The figures reveal seasonality with peaks during spring and dips towards the end of summer. In the cumulated times series one observes multiple extraordinary patterns. After Summer 2013 the counts peaks are less high, probably due to the vaccine which was released in Germany in July 2013, and in the time period between the end of 2019 and 2023 there are nearly no counts, most likely due to interventions against Covid-19.

8.1.1 Detecting Level Shifts

To analyse this further we used the model from our previous work (Maletz et al., 2024) as a baseline to test for persistent level-shifts. In that work we fitted a log-linear PSTARMAX process (3.4) with temporal order 4 and maximum spatial order of two for each time lag. We can write this baseline as

$$\boldsymbol{\psi}_t = \delta_0 \mathbf{1}_p + \sum_{i=1}^{4} \sum_{\ell=0}^{2} \beta_{i,\ell} \log(\mathbf{Y}_{t-i} + \mathbf{1}_p) + \sum_{k=1}^{4} \gamma_k \mathbf{X}_{k,t}. \tag{8.1}$$

The model includes these covariates:

- **Population**: log-transformed, linearly interpolated population size (per 100,000) for each region
- **GDR**: time-constant binary indicator for regions in former East Germany
- **Seasonality**: $\mathbf{X}_t = \cos\left(\frac{2\pi}{52}t\right)$ and $\mathbf{X}_t = \sin\left(\frac{2\pi}{52}t\right)$ for annual fluctuations

We now want to extend this to identify possible multiple level shifts, present for whole Germany, in the former GDR regions, and in the western Germany regions. We use the procedure described in Chapter 4. We define $\tau_{\text{min}} = 50$ and $\tau_{\text{max}} = 1200$, i.e. we are not interested in changes within the first or the last year of the observation period. The spatial candidates are $\mathbf{v}_{\text{Germany}} = \mathbf{1}_{411}$, representing whole Germany, then $\mathbf{v}_{\text{GDR}} = 1(\text{Region lies in former GDR})$, and $\mathbf{v}_{\text{West}} = \mathbf{v}_{\text{Germany}} - \mathbf{v}_{\text{GDR}}$ which are the regions in Western Germany. In total this results in 3,453 candidates for level shifts.

We apply the procedure described in Chapter 4, where we use the $(1-\alpha_k)^{\frac{1}{3453}}$-quantile of a χ^2_1-distribution in the k-th iteration as the stopping criterium. We set $\alpha_1 = 0.05$. After successfully finding a level-shift, for the next iteration divide the previous α_k by two for the next iteration, which implements a conservative detection behaviour.

The procedure stops after 4 iterations. Figure 8.1b shows at the red lines the time points of the level shift. The first one occurs at time point 646 in Western Germany, which is mid-2013 might be attributed to the vaccination released in July 2013. The second level shift is found in former GDR-regions and occurs at time point 745, which is approximatly two years after the release of the vaccination. This level shift also might be caused by the vaccination but with some time delay, due to maybe different people behaviour, i.e. more conservative people, lower vaccination rates and so on. The last two level-shifts are detected in the former GDR Regions and fall around a time period where Covid-19 was popular. The third intervention is found at time-point 960 in the former GDR, i.e. mid 2019 after the spring peak. The fourth one is found at time-point 1,111, which is in Mid April 2022, and in a time period where most of the interventions against Covid-19 were cancelled. If the procedure were to be run for further iterations, level shifts for western Germany would initially be detected in the first half of the time series. Only in the eighth iteration a level shift with a global effect would be detected. We re-estimate

the model from Maletz et al. (2024) and include the 4 level shifts detected. The estimated coefficients are in Table 8.1.

The recognized effects are also recognized as significant in the output. In West Germany, the number of Rota infections after the introduction of vaccination is approximately 26% lower than predicted by the model before the introduction of vaccination. In East Germany, there has been a slight increase of about 5% since 745. Due to COVID-19, there was a sharp drop in numbers in eastern Germany and a sharp increase after COVID-19. However, this will not be the case for Germany as a whole, nor specifically for western Germany. This may be because the modeled values for this area are already well described by the model, meaning that there were no unusual changes there during this period. Another possibility is that the assumption of a Poisson distribution is not justified and the detection procedure is biased.

8.1.2 Time varying dispersion

The default Poisson assumption fixes the dispersion parameter at $\phi_{i,t} = 1$ for all i, t, although count data often exhibit overdispersion. Alternatively one can refit the model using a Quasi-Poisson assumption or a negative binomial family. All three marginal distributions yield identical point estimates, differing only in standard error calculations due to different marginal variances. The sandwich estimator of the variance further reduces these differences. For this dataset, the quasipoisson dispersion is estimated at 1.804, while the negative binomial dispersion is 1.068, both indicating overdispersion.

We now extend the model by Maletz et al. (2024) via a time varying dispersion model on a log-linear scale, i.e. $\boldsymbol{\zeta} = \log(\boldsymbol{\phi}_t)$, and

$$\boldsymbol{\zeta}_t = \delta_0 \mathbf{1}_p + \beta_{1,0} \log(\mathbf{d}_{t-1}) + \beta_{1,1} \mathbf{W}^{(1)} \log(\mathbf{d}_{t-1}) + \sum_{k=1}^{8} \gamma_k \mathbf{X}_{k,t},$$

with $\mathbf{d}_t$ the pseudo-observations based on deviance residuals of the marginal poisson assumption, and the covariates are the same as in the previous section, including the detected Level-Shifts.

Table 8.2 shows the parameter estimates for the combined model. The parameters of the mean model change slightly compared to Table 8.1. Part of the seasonality shifts to the dispersion model. We also observe that one of the level shift effects disappears, specifically the second one that occurred approximately two years after the introduction of vaccination in eastern Germany. In the dispersion model, we observe a weak autoregressive dependence, which, however, estimates significant spatial-temporal dependencies. We also see that larger population sizes result in greater overdispersion, i.e., greater uncertainty in the model. We also have greater overdispersion and thus volatility in the regions of the former GDR. The level shifts have similar effects in the dispersion model as in the mean model. The level shifts appear to affect both the expectation and the overdispersion of the data.

Coefficient	Estimate	Std. Error	z value	Pr(> |z|)	
Intercept					
(Intercept)	−0.7574	0.0152	−49.732	<0.001	***
Past observations (Time Lag 1)					
$\beta_{1,0}$	0.4647	0.0062	75.131	<0.001	***
$\beta_{1,1}$	0.1000	0.0105	9.535	<0.001	***
$\beta_{1,2}$	0.0385	0.0203	1.898	0.058	.
Past observations (Time Lag 2)					
$\beta_{2,0}$	0.2128	0.0050	42.901	<0.001	***
$\beta_{2,1}$	0.0018	0.0106	0.171	0.864	
$\beta_{2,2}$	0.0012	0.0194	0.060	0.952	
Past observations (Time Lag 3)					
$\beta_{3,0}$	0.1100	0.0050	22.207	<0.001	***
$\beta_{3,1}$	−0.0003	0.0116	−0.024	0.981	
$\beta_{3,2}$	−0.0005	0.0196	−0.027	0.978	
Past observations (Time Lag 4)					
$\beta_{4,0}$	0.0527	0.0051	10.294	<0.001	***
$\beta_{4,1}$	−0.0025	0.0112	−0.226	0.821	
$\beta_{4,2}$	−0.0016	0.0189	−0.085	0.932	
Covariates					
$\text{population}_{\{s_0\}}$	0.3355	0.0063	53.592	<0.001	***
GDR	0.3014	0.0121	24.856	<0.001	***
$\cos(\frac{2\pi}{52}t)$	0.0626	0.0136	4.611	<0.001	***
$\sin(\frac{2\pi}{52}t)$	0.4459	0.0128	34.836	<0.001	***
Level shift effects					
$\tau = 646, \mathbf{v} = \mathbf{v}_{\text{West}}$	−0.2670	0.0207	−12.908	<0.001	***
$\tau = 745, \mathbf{v} = \mathbf{v}_{\text{GDR}}$	0.0584	0.0198	2.949	0.003	**
$\tau = 960, \mathbf{v} = \mathbf{v}_{\text{GDR}}$	−0.5513	0.0480	−11.488	<0.001	***
$\tau = 1111, \mathbf{v} = \mathbf{v}_{\text{GDR}}$	0.4924	0.0516	9.540	<0.001	***

Table 8.1: Estimation results of the log-linear PSTARMAX model on the rota virus data including the level shift covariates

Coefficient	Estimate	Std. Error	z value	Pr(> |z|)	
Mean Model *(Quasi-Poisson, log link)*					
Intercept	-0.8511	0.0154	-55.40	<0.001	***
Past observations (Time Lag 1)					
$\beta_{1,0}$	0.4563	0.0065	70.67	<0.001	***
$\beta_{1,1}$	0.1131	0.0107	10.53	<0.001	***
$\beta_{1,2}$	0.0106	0.0211	0.50	0.615	
Past observations (Time Lag 2)					
$\beta_{2,0}$	0.2214	0.0051	43.12	<0.001	***
$\beta_{2,1}$	0.0043	0.0110	0.39	0.694	
$\beta_{2,2}$	0.0025	0.0206	0.12	0.903	
Past observations (Time Lag 3)					
$\beta_{3,0}$	0.1175	0.0050	23.59	<0.001	***
$\beta_{3,1}$	0.0018	0.0109	0.17	0.866	
$\beta_{3,2}$	0.0012	0.0199	0.06	0.951	
Past observations (Time Lag 4)					
$\beta_{4,0}$	0.0575	0.0051	11.32	<0.001	***
$\beta_{4,1}$	0.0006	0.0110	0.06	0.954	
$\beta_{4,2}$	0.0006	0.0194	0.03	0.974	
Covariates					
log(population)	0.3886	0.0069	56.52	<0.001	***
GDR	0.3449	0.0133	25.88	<0.001	***
$\cos(\frac{2\pi}{52}t)$	0.0206	0.0137	1.50	0.134	
$\sin(\frac{2\pi}{52}t)$	0.4868	0.0136	35.85	<0.001	***
Level shift effects					
$\tau = 646, \mathbf{v} = \mathbf{v}_{\text{West}}$	-0.2745	0.0197	-13.96	<0.001	***
$\tau = 745, \mathbf{v} = \mathbf{v}_{\text{GDR}}$	0.0297	0.0195	1.52	0.128	
$\tau = 960, \mathbf{v} = \mathbf{v}_{\text{GDR}}$	-0.5087	0.0432	-11.78	<0.001	***
$\tau = 1111, \mathbf{v} = \mathbf{v}_{\text{GDR}}$	0.4831	0.0467	10.34	<0.001	***
Dispersion Model					
Intercept	0.3204	0.0118	27.23	<0.001	***
Past pseudo observations (Time Lag 1)					
$\beta_{1,0}$	0.0360	0.0049	7.36	<0.001	***
$\beta_{1,1}$	0.0119	0.0063	1.87	0.061	.
Covariates					
log(Population)	0.3128	0.0070	44.63	<0.001	***
GDR	0.5859	0.0163	35.92	<0.001	***
$\cos(\frac{2\pi}{52}t)$	0.1314	0.0115	11.43	<0.001	***
$\sin(\frac{2\pi}{52}t)$	0.4079	0.0097	42.03	<0.001	***
Level shift effects					
$\tau = 646, \mathbf{v} = \mathbf{v}_{\text{West}}$	-0.1765	0.0158	-11.20	<0.001	***
$\tau = 745, \mathbf{v} = \mathbf{v}_{\text{GDR}}$	0.0161	0.0304	0.53	0.596	
$\tau = 960, \mathbf{v} = \mathbf{v}_{\text{GDR}}$	-0.5847	0.0545	-10.72	<0.001	***
$\tau = 1111, \mathbf{v} = \mathbf{v}_{\text{GDR}}$	0.4131	0.0568	7.27	<0.001	***

Table 8.2: Parameter estimates of the spatio-temporal Double GLMs on the Rota data.

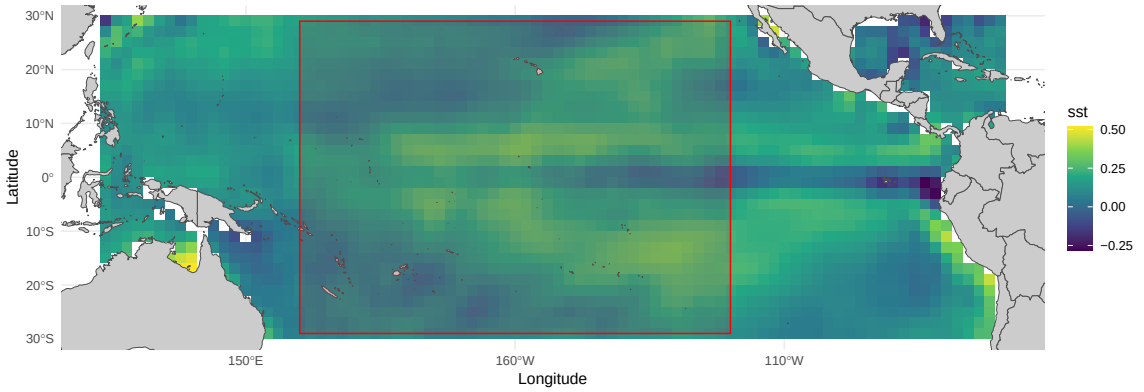


Figure 8.2: Pixel-wise mean of sea surface temperature anomalies (in °C) in the Pacific. The red rectangle, 29°S to 29°N latitude and 160°E to 240°E (120°W), marks the area modeled in this data example.

8.2 Sea Surface Temperature Anomalies

In the second example, we consider Sea Surface Temperature Anomaly (SST) data in the Pacific Ocean. The data describe monthly deviations (in °C) from a climatology over the period from January 1970 to December 2003, measured on a regular grid from 120°E to 290°E (70°W) longitude and 29°S to 29°N latitude in 2° increments. The complete data set is included in the R package `STRbook` (Wikle et al., 2019) under the name `SST_df`. Following Hølleland and Karlsen (2020), we consider the rectangular area from 29°S to 29°N latitude and 160°E to 240°E (120°W), which contains no land masses and has complete spatial coverage.

Figure 8.2 shows the area considered for modeling, along with the coordinate-wise mean values of the SST variable. In total, we use monthly observations from this 30×41 grid ($p = 1230$) over the time period from January 1970 to December 2002 ($T = 396$). We observe minor average temperature anomalies near the equator, which increase until 6°N and 6°S and then decrease again. Moving eastward, i.e., with increasing longitude, higher anomalies are also observed compared to the western part of the study area. Based on these observations, we construct several covariates to account for in the modeling: (1) a linear temporal trend with values ranging from 0 (beginning of the observation period) to 1 (end), to capture a potential increase in temperature; (2) longitude/360 to describe higher temperatures further east; (3) $\cos(2\pi/12 \cdot t)$ and $\sin(2\pi/12 \cdot t)$ to capture potential seasonality; and (4) two covariates $\min\{|\text{latitude}|, 6\}/90$ and $\max\{|\text{latitude}| - 6, 0\}/90$ to

describe the increasing temperature anomalies up to the sixth degree and their decline beyond that in a piecewise linear manner.

8.2.1 Mean Model

We cannot reconstruct exactly how the anomalies were calculated. It is plausible that they were determined using a simple regression model that ignores spatial dependencies. Such dependencies, however, can arise naturally, as these are surface water temperature data that may be transported to nearby locations by ocean currents. Additionally, weather data typically exhibit time-varying variability influenced by factors such as seasonal changes. A combined modeling approach addressing both the mean and variance is therefore appropriate for such data. We consider a model under the marginal normality assumption, i.e.,

$$Y_{i,t} \mid \mathscr{F}_{t-1} \sim \mathcal{N}(\mu_{i,t}, \phi_{i,t})$$

and $\mathrm{Var}(Y_{i,t} \mid \mathscr{F}_{t-1}) = \phi_{i,t}$. We first use the mean model

$$\boldsymbol{\mu}_t = \delta_0 \mathbf{1}_{1230} + \left(\beta_0 \mathbf{I}_{1230} + \beta_{\mathrm{N}} \mathbf{W}^{(\mathrm{N})} + \beta_{\mathrm{E}} \mathbf{W}^{(\mathrm{E})} + \beta_{\mathrm{S}} \mathbf{W}^{(\mathrm{S})} + \beta_{\mathrm{W}} \mathbf{W}^{(\mathrm{W})}\right) \mathbf{Y}_{t-1} + \sum_{k=1}^{6} \gamma_k \mathbf{X}_{k,t},$$

where $\mathbf{W}^{(\ell)}, \ell \in \{\mathrm{N}, \mathrm{E}, \mathrm{S}, \mathrm{W}\}$ indicates, row-wise for each location, exactly the adjacent grid point to the north, east, south, and west, respectively. Each row contains exactly one entry of 1 and zeros elsewhere. For grid points located at the boundary of the region shown in Figure 8.2, such a neighbor does not necessarily exist, meaning some rows contain only zeros. This violates the condition $\mathbf{W}^{(\ell)} \mathbf{1}_p = \mathbf{1}_p$ mentioned earlier in this work. However, given the large number of locations $p = 1230$ relative to the small number of boundary points (138), the impact of this violation is negligible. The cyclic neighborhood matrices used by Hølleland and Karlsen (2020) would, in our opinion, introduce a bias into the analysis, as they treat locations at the northern edge of the grid as direct neighbors of points on the southern edge. We do not use higher-order neighborhoods, as it is not clear how to construct them properly without causing the number of parameters to grow excessively. We also refrain from including a feedback mechanism in the model, as this would introduce long-range dependence spanning several months from neighbors in a single direction. Such dependencies would complicate the interpretation of the model, as they would imply dependence from locations that are not directly adjacent. Since water flows through these adjacent regions, first-order neighbors are sufficient to capture the relevant spatial dependence structure.

8.2.2 Dispersion Model

We describe the variance process $\boldsymbol{\phi}_t$ on a log-linear scale to also capture negative dependencies. We set $\boldsymbol{\zeta}_t = \log(\boldsymbol{\phi}_t)$ and assume

$$\boldsymbol{\zeta}_t = \delta_0 \mathbf{1}_{1230} + \left(\beta_0 \mathbf{I}_{1230} + \beta_{\mathrm{N}} \mathbf{W}^{(\mathrm{N})} + \beta_{\mathrm{E}} \mathbf{W}^{(\mathrm{E})} + \beta_{\mathrm{S}} \mathbf{W}^{(\mathrm{S})} + \beta_{\mathrm{W}} \mathbf{W}^{(\mathrm{W})}\right) \log((\mathbf{Y}_{t-1} - \boldsymbol{\mu}_t)^2) + \sum_{k=1}^{6} \gamma_k \mathbf{X}_{k,t},$$

i.e., we adopt the same model order as in the mean model and use the same set of covariates. The structure of the dispersion model is closely related to that of a spatio-temporal log-GARCH model with additional covariates (Otto et al., 2025).

8.2.3 Interpretation

Table 8.3 contains the parameter estimates along with standard errors and p-values obtained by normal approximation. For the mean model parameters, we detect significant spatial dependence. The coefficient associated with $\mathbf{W}^{(\mathrm{E})}$ is estimated to be substantially larger than the remaining spatial dependence parameters, indicating a stronger dependence from the eastern direction. This can be explained by the northern and southern equatorial currents, which are driven by trade winds and transport water masses westward, thereby inducing spatial dependence from the east. Dependencies from other directions are also significant, which can be attributed to additional ocean currents that are, however, dominated by the equatorial currents on the scale of this grid section. Furthermore, the mean model exhibits a positive temporal trend, likely attributable to rising ocean temperatures due to climate change. The remaining covariates in the mean model are not significant at the 5% level.

In the dispersion model, we also detect weak spatio-temporal autoregressive dependence, with only minor differences across the directional coefficients. A negative trend is evident in the dispersion, suggesting that variance decreases over time. Additionally, an annual seasonality in variance is present. Longitude does not appear to have a notable effect on either the mean or the variance. Latitude leads to a decrease in variance up to the sixth degree north or south, followed by an increase beyond that point. This pattern suggests a higher concentration of extreme anomalies in this region, possibly an artefact of equatorial currents. An asymptotic Wald test of the hypothesis $\beta_{\mathrm{N}} = \beta_{\mathrm{E}} = \beta_{\mathrm{S}} = \beta_{\mathrm{W}}$ in the dispersion model does not reject the null at the 5% level, indicating that the directional coefficients do not differ significantly. One might therefore fit the dispersion model using a row-normalized version of $\mathbf{W} = \mathbf{W}^{(\mathrm{N})} + \mathbf{W}^{(\mathrm{E})} + \mathbf{W}^{(\mathrm{S})} + \mathbf{W}^{(\mathrm{W})}$, which does not appreciably change the remaining parameter estimates.

Coefficient	Estimate	Std. Error	z value	Pr(> |z|)	
Mean Model *(Gaussian, identity link)*					
Intercept	−0.1059	0.0392	−2.700	0.007	**
Past observations (Time Lag 1)					
β_0	0.1526	0.0224	6.819	<0.001	***
β_N	0.1473	0.0125	11.781	<0.001	***
β_E	0.2583	0.0104	24.860	<0.001	***
β_S	0.1529	0.0116	13.138	<0.001	***
β_W	0.1095	0.0119	9.190	<0.001	***
Covariates					
trend	0.1090	0.0170	6.431	<0.001	***
longitude /360	0.1304	0.0696	1.873	0.061	.
$\cos(2\pi/12 \cdot t)$	−0.0081	0.0062	−1.304	0.192	
$\sin(2\pi/12 \cdot t)$	−0.0113	0.0063	−1.788	0.074	.
$\min\{\|\text{latitude}\|, 6\}/90$	0.2778	0.2693	1.032	0.302	
$\max\{\|\text{latitude}\| - 6, 0\}/90$	−0.0898	0.0540	−1.662	0.096	.
Dispersion Model *(log link)*					
Intercept	−0.5505	0.1201	−4.584	<0.001	***
Past pseudo observations (Time Lag 1)					
β_0	0.0060	0.0011	5.380	<0.001	***
β_N	0.0292	0.0031	9.586	<0.001	***
β_E	0.0245	0.0022	10.936	<0.001	***
β_S	0.0270	0.0030	9.089	<0.001	***
β_W	0.0170	0.0021	8.155	<0.001	***
Covariates					
trend	−0.6708	0.0631	−10.625	<0.001	***
longitude /360	0.1183	0.2264	0.523	0.601	
$\cos(2\pi/12 \cdot t)$	0.1489	0.0260	5.727	<0.001	***
$\sin(2\pi/12 \cdot t)$	0.0290	0.0265	1.093	0.274	
$\min\{\|\text{latitude}\|, 6\}/90$	−17.0994	1.2660	−13.507	<0.001	***
$\max\{\|\text{latitude}\| - 6, 0\}/90$	2.0457	0.2452	8.344	<0.001	***

Table 8.3: Estimated spatio-temporal DGLM for the Sea Surface Temperature Anomalies.

8.2.4 Comparison to alternative models

Since we assume conditional normal distributions for this data set, a wide range of alternative models is available. Natural candidates are the STARMA models of Pfeifer and Deutsch (1980c) and their combination with a GARCH process (Hølleland and Karlsen, 2020). Both models describe the data solely through autoregressive dependence structures, meaning that location-specific differences cannot easily be incorporated via covariates. Furthermore, the data must be detrended beforehand. The theoretical properties of these models rely, among other things, on spatial stationarity assumptions, so their application requires several complex preprocessing steps that additionally compromise interpretability. The `starma` R package (Cheysson, 2016), which implements STARMA processes, works exclusively with dense spatial weight matrices. This is particularly problematic in high-dimensional settings such as this application, where the weight matrices are sparse, drastically increasing computation time. On an Apple MacBook Air with M2 processor, estimating the DGLM reported in Table 8.3 takes approximately 34 seconds using the `glmSTARMA` package. By contrast, estimating a first-order STAR process with the same neighborhood matrices, i.e.,

$$\mathbf{Y}_t = \left(\beta_0 \mathbf{I}_{1230} + \beta_{\mathrm{N}} \mathbf{W}^{(\mathrm{N})} + \beta_{\mathrm{E}} \mathbf{W}^{(\mathrm{E})} + \beta_{\mathrm{S}} \mathbf{W}^{(\mathrm{S})} + \beta_{\mathrm{W}} \mathbf{W}^{(\mathrm{W})}\right) \mathbf{Y}_{t-1} + \boldsymbol{\varepsilon}_t,$$

takes approximately 250 seconds with the `starma` package, making it unsuitable for data of this scale.

A more practical alternative is provided by Generalized Network Autoregression (GNAR) models (Knight et al., 2016) and their extension to include exogenous covariates (GNARX) (Nason and Wei, 2022), both implemented in the R package `GNAR` (Knight et al., 2020).

Users of this package need only supply a graph defining the spatial neighborhood structure, as higher-order neighbors are computed automatically when required. While this simplifies the model specification, it also means that users may be unaware of the implied spatial dependency structure when relying on this feature. Furthermore, it is not possible to estimate direction-specific dependence parameters, as in the example above, since only a single network can be provided.

The GNAR framework supports time-varying networks, which is currently not possible within our framework. Additionally, GNAR models can readily estimate individual autoregressive parameters for each location. In the `glmSTARMA` package, this can only be achieved by supplying a large list of neighborhood matrices $\mathbf{W}^{(i)}, i = 1, \ldots, p$, where the i-th diagonal element of $\mathbf{W}^{(i)}$ equals 1 and all other entries are zero. The spatial dependence matrices are then appended to this list, and the spatial orders must be specified accordingly.

For comparison, we estimate a simple GNARX model of the form

$$\mathbf{Y}_t = \beta_0 \mathbf{Y}_{t-1} + \beta_1 \mathbf{W}\mathbf{Y}_{t-1} + \sum_{k=1}^{7} \gamma_k \mathbf{X}_{k,t} + \boldsymbol{\varepsilon}_t,$$

Model		Intercept	β_0	β_1	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6
GNARX	Est.	−0.0899	−0.1603	0.9749	0.1080	0.0874	−0.0069	−0.0084	0.4124	−0.0964
	Std. Err.	0.0053	0.0100	0.0104	0.0020	0.0085	0.0008	0.0008	0.0419	0.0075
`glmstarma`	Est.	−0.0889	−0.0937	0.9062	0.1086	0.0867	−0.0068	−0.0086	0.4025	−0.0974
	Std. Err.	0.0406	0.0356	0.0393	0.0185	0.0711	0.0069	0.0070	0.2674	0.0541

Table 8.4: Parameter estimates and standard errors of the GNARX and `glmstarma` models.

where $\mathbf{W}$ is the row-normalized adjacency matrix and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ are independent and identically distributed. A grid point is considered adjacent if it lies directly to the north, east, south, or west of a given point, as defined above. The covariates are identical to those used above; the additional seventh covariate is the intercept, which is not included automatically.

We also estimate a comparable mean model using `glmstarma` under the assumption of marginal normal distributions, i.e.,

$$\boldsymbol{\mu}_t = \delta_0 \mathbf{1}_p + \beta_0 \mathbf{Y}_{t-1} + \beta_1 \mathbf{W}\mathbf{Y}_{t-1} + \sum_{k=1}^{6} \gamma_k \mathbf{X}_{k,t}.$$

Estimation of the GNARX model takes approximately 15 seconds on a MacBook Air M2, while the `glmstarma` model requires approximately 1.7 seconds. Table 8.4 reports the parameter estimates of both models together with standard errors obtained from supplementary calculations. Both models produce nearly identical estimates for the covariate parameters. Differences arise in β_0 and β_1: we enforce stationarity in `glmstarma`, whereas the `GNAR` package estimates the parameters via the `lm` function. Furthermore, the standard errors we obtain are larger than those from the GNARX model, as we do not assume independent and identically distributed errors but also account for contemporaneous spatial dependencies.

9 Summary and Outlook

> "*What we know is not much. What we do not know is immense.*"
> — Pierre-Simon Laplace

9.1 Summary

This dissertation develops a comprehensive framework for modeling spatio-temporal data based on Double Generalized Linear Models (DGLMs), with a focus on lag-based autoregressive approaches that explicitly capture dependencies across time and space. The starting point is the class of multivariate count time series models introduced by Fokianos et al. (2020), which generalizes univariate INGARCH and log-linear autoregressive models to a multivariate setting. These models are extended to the spatio-temporal domain by replacing unstructured parameter matrices with parsimonious linear combinations of known matrices encoding spatial and temporal dependence. This structural restriction ensures scalability to high-dimensional settings with many regions while retaining interpretability and statistical tractability.

Chapter 3 lays the theoretical foundations for the proposed model class. We adapt the multivariate count time series framework and establish conditions for stationarity and ergodicity, which simplify considerably in the spatio-temporal setting. Asymptotic properties of the quasi-maximum likelihood estimator (QMLE), consistency and asymptotic normality, are discussed, providing a solid theoretical basis for the simulation studies and applications that follow. The use of quasi-likelihood, rather than full likelihood, is a deliberate modeling choice: it avoids specifying contemporaneous dependencies between regional components, thereby reducing misspecification risk and simplifying computation.

Chapter 4 introduces a methodology for detecting multiple persistent level-shifts in spatio-temporal counts, which may arise from policy interventions, singular events, or structural changes. The procedure adapts and extends the sequential testing approaches of Fokianos and Fried (2010, 2012) and naturally accommodates simultaneous shifts across different regions, providing a practical tool for post-hoc analysis of count time series subject to external shocks.

A major contribution is presented in Chapters 5 and 6. We relax the restrictive assumption of marginal Poisson distributions underlying the baseline framework by developing a considerably more general DGLM-based framework (Smyth, 1989). This introduces two linked submodels: one for the conditional mean and one for the dispersion, where dispersion is modeled dynamically through the same spatio-temporal lag structure. The resulting model class constitutes a spatio-temporal analog of the ARMA-GARCH paradigm, now applicable to non-normal data such as counts. It subsumes several existing models as special cases and, through flexible link functions and distributional families, substantially broadens the scope of applications addressable within a unified framework.

All methods are implemented in the `R` package `glmSTARMA`, documented in Chapter 6. The package provides a user-friendly interface for model specification, estimation, and diagnostics, making the developed methods accessible to practitioners without specialist theoretical background. To our knowledge, `glmSTARMA` is the first software implementation accommodating both varying mean and varying dispersion in non-normal spatio-temporal settings, closing an important gap between methodological development and applied practice.

The theoretical developments are complemented by extensive simulation studies in Chapter 7, which systematically assess finite-sample behavior of the proposed estimators and testing procedures across a range of model configurations, sample sizes, and dependence structures. The results confirm good finite-sample performance of the QMLE even at moderate sample sizes. The level-shift detection procedure is able to detect the time point very good, but lacks an easy termination criterion.

Two real-data applications in Chapter 8 illustrate the practical utility of the developed methods. Weekly rotavirus infection counts from German districts demonstrate that the spatio-temporal count autoregressive framework successfully captures seasonal patterns, district-level heterogeneity, and spatial contagion; the level-shift detection procedure finds meaningful time points, i.e. it proves informative in identifying periods of elevated or suppressed incidence. Sea Surface Temperature Anomalies in the Pacific Ocean showcase that the DGLM framework extends naturally beyond epidemiology to continuous-valued environmental data with time-varying variability. Together, these applications underscore the breadth and practical relevance of the methodology.

In summary, this dissertation advances spatio-temporal modeling for non-Gaussian data along three interconnected dimensions: it provides a theoretically rigorous and parsimonious modeling framework, extends it with tools for intervention analysis and dispersion modeling, and makes all estimation methods computationally accessible through open-source software, offering a coherent and flexible toolkit for the analysis of spatio-temporal data across a wide range of scientific disciplines.

9.2 Outlook

While the developed methods represent a substantial advance over existing approaches, several natural directions for future research remain.

A first direction concerns temporally and spatially varying parameters. Throughout this work, dependence parameters governing the autoregressive dynamics are assumed constant over time and space. While this facilitates theoretical analysis and estimation, it may be too restrictive in applications where spatio-temporal dependence evolves over time, for instance due to seasonal variation in behavior or gradual changes in spatial connectivity. Allowing for time-varying autoregressive parameters, analogous to the treatment of seasonality in the `hhh4` framework (Meyer et al., 2017), would substantially increase modeling flexibility. Similarly, incorporating time-varying spatial weight matrices, as in GNAR processes, would allow the dependence structure to adapt to changing spatial conditions and could be integrated into the existing framework in a relatively straightforward manner.

A second direction concerns contemporaneous dependencies. Quasi-likelihood estimation, used throughout this dissertation, remains valid without specifying how regions co-move at the same point in time, but comes at the cost of statistical efficiency relative to full likelihood methods. Explicitly modeling contemporaneous dependence, for instance via a copula approach that jointly specifies the cross-sectional dependence structure while preserving marginal autoregressive dynamics, could yield considerable efficiency gains in applications with strong contemporaneous correlations, such as geographically proximate regions during epidemic outbreaks or extreme weather events, albeit at substantially increased complexity and computational cost.

Third, several extensions of the level-shift detection procedure proposed in Chapter 4 are warranted. The stopping criterion relies on quantiles of a distribution that is difficult to determine analytically; bootstrap-based approximations are feasible but computationally demanding in high-dimensional settings, making closed-form alternatives desirable. Furthermore, the current framework addresses only persistent level shifts; extending it to transient or exponentially attenuating interventions, already studied in univariate settings, would add practical value for applications where external shocks dissipate over time, although the identifiability of such effects in the presence of autoregressive dynamics requires careful investigation. The behavior of the procedure under model misspecification also remains to be characterized.

A fourth open problem is model selection. Standard information criteria such as AIC and BIC were designed for independent data, and their application to time series raises well-known concerns; in particular, increasing model order reduces the number of observations contributing to the quasi-likelihood, potentially biasing the criterion. Whether rescaling approaches can correct for this bias in spatio-temporal count settings deserves further investigation. Beyond order selection, the choice of spatial weight matrices constitutes an additional layer of model selection not addressed by standard criteria. Spatially resolved

goodness-of-fit measures and proper scoring rules for probabilistic forecasts at each location would be a valuable complement, and the development of computationally efficient procedures for simultaneous spatial and temporal order selection, possibly based on penalized estimation, represents an important open problem.

Fifth, scalability to very high-dimensional settings merits further attention. While the parsimonious parameterization via linear combinations of known matrices already mitigates the curse of dimensionality, settings with very many regions or dense, unstructured weight matrices may still strain estimation. LASSO-type regularization of the weight matrices or autoregressive coefficient vectors could enforce additional sparsity and improve numerical stability. More broadly, how to incorporate uncertainty in the spatial weight matrices themselves into inference and prediction remains largely unexplored in the spatio-temporal count time series literature.

Sixth, the spatio-temporal DGLM framework of Chapter 5 rests on several restrictive assumptions, and formal theoretical results for stationarity, ergodicity, and asymptotic properties of the QMLE remain outstanding. The coupling between mean and dispersion submodels introduces additional technical challenges, as dispersion dynamics depend on past realized values whose distribution is itself governed by the mean submodel. Moreover, while the reduction to purely temporal univariate models is straightforward in principle, a systematic development of this special case, including tailored software and comparative simulations, would broaden accessibility.

Finally, the spatial dependence structures considered throughout assume static weight matrices, reflecting time-invariant spatial relationships. Extending the framework to time-varying weight matrices, represents a promising avenue for future work. Taken together, these directions suggest that the framework developed in this dissertation provides not only a self-contained toolkit for practitioners, but also a fertile starting point for a broad agenda of methodological research in spatio-temporal statistics.

Bibliography

Agosto, A., G. Cavaliere, D. Kristensen, and A. Rahbek (2016). "Modeling corporate defaults: Poisson autoregressions with exogenous covariates (PARX)". In: *Journal of Empirical Finance* 38, pp. 640–663.

Ahn, H. and J. J. Chen (1995). "Generation of Over-Dispersed and Under-Dispersed Binomial Variates". In: *Journal of Computational and Graphical Statistics* 4 (1), pp. 55–64.

Akaike, H. (1974). "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* 19 (6), pp. 716–723.

Aknouche, A. and C. Francq (2021). "Count and Duration Time Series with Equal Conditional Stochastic and Mean Order". In: *Econometric Theory* 37 (2), pp. 248–280.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Vol. 4. Studies in Operational Regional Science. Dordrecht: Springer Netherlands.

Anselin, L. (2024). *An Introduction to Spatial Data Science with GeoDa: Volume 1: Exploring Spatial Data*. 1st ed. Boca Raton: Chapman and Hall/CRC.

Armillotta, M. and K. Fokianos (2023). "Nonlinear network autoregression". In: *The Annals of Statistics* 51 (6), pp. 2526–2552.

Armillotta, M. and K. Fokianos (2024). "Count network autoregression". In: *Journal of Time Series Analysis* 45 (4), pp. 584–612.

Barreto-Souza, W., L. S. C. Piancastelli, K. Fokianos, and H. Ombao (2025). "Time-Varying Dispersion Integer-Valued GARCH Models". In: *Journal of Time Series Analysis*, jtsa.12838.

Bates, D., M. Maechler, and M. Jagan (2025). *Matrix: Sparse and Dense Matrix Classes and Methods*.

Bivand, R. (2025). *spdep: Spatial Dependence: Weighting Schemes, Statistics*.

Bollerslev, T. (1986). "Generalized autoregressive conditional heteroskedasticity". In: *Journal of Econometrics* 31 (3), pp. 307–327.

Bonat, W. H., B. Jørgensen, C. C. Kokonendji, J. Hinde, and C. G. B. Demétrio (2018). "Extended Poisson-Tweedie: properties and regression models for count data". In: *Statistical Modelling. An International Journal* 18 (1), pp. 24–49.

Borovkova, S. and R. Lopuhaa (2012). "Spatial GARCH: A Spatial Approach to Multivariate Volatility Modeling". In: *SSRN Electronic Journal*.

Box, G. E. P. and G. C. Tiao (1975). "Intervention Analysis with Applications to Economic and Environmental Problems". In: *Journal of the American Statistical Association* 70 (349), pp. 70–79.

Box, G. E. P. and G. M. Jenkins (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Bracher, J. and L. Held (2022). "Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction". In: *International Journal of Forecasting* 38 (3), pp. 1221–1233.

Chen, K., I. Hu, and Z. Ying (1999). "Strong Consistency of Maximum Quasi-Likelihood Estimators in Generalized Linear Models with Fixed and Adaptive Designs". In: *The Annals of Statistics* 27 (4), pp. 1155–1163.

Cheysson, F. (2016). *starma: Modelling Space Time AutoRegressive Moving Average (STARMA) Processes*.

Choi, B., M. Bergés, E. Bou-Zeid, and M. Pozzi (2021). "Short-term probabilistic forecasting of meso-scale near-surface urban temperature fields". In: *Environmental Modelling & Software* 145, p. 105189.

Clark, N. J. and P. M. Dixon (2021). "A class of spatially correlated self-exciting statistical models". In: *Spatial Statistics* 43, p. 100493.

Clement, L., O. Thas, P. Vanrolleghem, and J. Ottoy (2006). "Spatio-temporal statistical models for river monitoring networks". In: *Water Science and Technology* 53 (1), pp. 9–15.

Cliff, A. D. and J. K. Ord (1975). "Space-Time Modelling with an Application to Regional Forecasting". In: *Transactions of the Institute of British Geographers* (64), p. 119.

Consul, P. C. and G. C. Jain (1973). "A Generalization of the Poisson Distribution". In: *Technometrics* 15 (4), pp. 791–799.

Cressie, N. A. C. and C. K. Wikle (2011). *Statistics for spatio-temporal data*. Wiley series in probability and statistics. Hoboken, N.J: Wiley. 1 p.

Davies, R. B. (1987). "Hypothesis testing when a nuisance parameter is present only under the alternative". In: *Biometrika* 74 (1), pp. 33–43.

Debaly, Z. M. and L. Truquet (2021). "A note on the stability of multivariate non-linear time series with an application to time series of counts". In: *Statistics & Probability Letters* 179, p. 109196.

Deutsch, S. J. and P. E. Pfeifer (1981). "Space-Time ARMA Modeling With Contemporaneously Correlated Innovations". In: *Technometrics* 23 (4), pp. 401–409.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. New York, NY: Springer.

Dickson, M. M., G. Espa, D. Giuliani, F. Santi, and L. Savadori (2020). "Assessing the effect of containment measures on the spatio-temporal dynamic of COVID-19 in Italy". In: *Nonlinear Dynamics* 101 (3), pp. 1833–1846.

Doukhan, P., M. H. Neumann, and L. Truquet (2023). "Stationarity and ergodic properties for some observation-driven models in random environments". In: *The Annals of Applied Probability* 33 (6), pp. 5145–5170.

Efron, B. (1986). "Double exponential families and their use in generalized linear regression". In: *Journal of the American Statistical Association* 81 (395), pp. 709–721.

Elhorst, J. P. and S. Emili (2022). "A spatial econometric multivariate model of Okun's law". In: *Regional Science and Urban Economics* 93, p. 103756.

Fahrmeir, L. and H. Kaufmann (1985). "Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models". In: *The Annals of Statistics* 13 (1), pp. 342–368.

Ferland, R., A. Latour, and D. Oraichi (2006). "Integer-Valued GARCH Process". In: *Journal of Time Series Analysis* 27 (6), pp. 923–942.

Fokianos, K. (2024). "Multivariate count time series modelling". In: *Econometrics and Statistics* 31, pp. 100–116.

Fokianos, K. and R. Fried (2010). "Interventions in INGARCH processes". In: *Journal of Time Series Analysis* 31 (3), pp. 210–225.

Fokianos, K. and R. Fried (2012). "Interventions in log-linear Poisson autoregression". In: *Statistical Modelling* 12 (4), pp. 299–322.

Fokianos, K., B. Støve, D. Tjøstheim, and P. Doukhan (2020). "Multivariate count autoregression". In: *Bernoulli* 26 (1).

Fokianos, K. and D. Tjøstheim (2011). "Log-linear Poisson autoregression". In: *Journal of Multivariate Analysis* 102 (3), pp. 563–578.

Foster, D. P. and R. A. Stine (2008). "alpha-Investing: a Procedure for Sequential Control of Expected False Discoveries". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 70 (2), pp. 429–444.

Francq, C. and J.-M. Zakoïan (2019). *GARCH Models: Structure, Statistical Inference and Financial Applications*. John Wiley & Sons, Ltd.

Gao, Q.-B., J.-G. Lin, C.-H. Zhu, and Y.-H. Wu (2012). "Asymptotic properties of maximum quasi-likelihood estimators in generalized linear models with adaptive designs". In: *Statistics* 46 (6), pp. 833–846.

Habereder, I., T. Kneib, I. Echizen, and T. Spinde (2025). *A Systematic Review of Spatio-Temporal Statistical Models: Theory, Structure, and Applications*. arXiv: 2511.00422 [stat.AP].

Held, L., M. Höhle, and M. Hofmann (2005). "A statistical framework for the analysis of multivariate infectious disease surveillance counts". In: *Statistical Modelling* 5 (3), pp. 187–199.

Hølleland, S. and H. A. Karlsen (2020). "A Stationary Spatio-Temporal GARCH Model". In: *Journal of Time Series Analysis* 41 (2), pp. 177–209.

Jahn, M., C. H. Weiß, and H.-Y. Kim (2023). "Approximately linear INGARCH models for spatio-temporal counts". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 72 (2), pp. 476–497.

Jørgensen, B. (1987). "Exponential Dispersion Models". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 49 (2), pp. 127–145.

Jørgensen, B. (1997). *The theory of dispersion models*. Monographs on statistics and applied probability 76. London: Chapman & Hall. 237 pp.

Kamarianakis, Y. and P. Prastacos (2005). "Space–time modeling of traffic flow". In: *Computers & Geosciences* 31 (2), pp. 119–133.

Knight, M. I., M. A. Nunes, and G. P. Nason (2016). *Modelling, Detrending and Decorrelation of Network Time Series*.

Knight, M., K. Leeming, G. Nason, and M. Nunes (2020). "Generalized Network Autoregressive Processes and the GNAR Package". In: *Journal of Statistical Software* 96, pp. 1–36.

Li, H., H. Demirtas, and R. Chen (2021). "RNGforGPD: An R Package for Generation of Univariate and Multivariate Generalized Poisson Data". In: *The R Journal* 12 (2), pp. 120–133.

Liboschik, T., K. Fokianos, and R. Fried (2017). "tscount: An R package for analysis of count time series following generalized linear models". In.

Maletz, S. (2021). "Spatio-temporal Models for Count Data". Master's Thesis. Deutschland: TU Dortmund.

Maletz, S., K. Fokianos, and R. Fried (2024). "Spatio-Temporal Count Autoregression". In: *Data Science in Science* 3 (1), p. 2425171.

Maletz, S., K. Fokianos, and R. Fried (2026a). *glmSTARMA – An R-Package for fitting autoregressive spatio-temporal models following generalized linear models*. arXiv: 2607.08276 [stat.CO].

Maletz, S., K. Fokianos, and R. Fried (2026b). *glmSTARMA: (Double) Generalized Linear Models for Spatio-Temporal Data*.

Martin, R. L. and J. E. Oeppen (1975). "The Identification of Regional Forecasting Models Using Space: Time Correlation Functions". In: *Transactions of the Institute of British Geographers* (66), p. 95.

Martins, A., M. G. Scotto, C. H. Weiß, and S. Gouveia (2023). "Space-time integer-valued ARMA modelling for time series of counts". In: *Electronic Journal of Statistics* 17 (2).

McCullagh, P. and R. Tibshirani (1990). "A Simple Method for the Adjustment of Profile Likelihoods". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 52 (2), pp. 325–344.

McKenzie, E. (1985). "Some Simple Models for Discrete Variate Time Series". In: *JAWRA Journal of the American Water Resources Association* 21 (4), pp. 645–650.

Meyer, S., L. Held, and M. Höhle (2017). "Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance". In: *Journal of Statistical Software* 77 (11).

Michael, J. R., W. R. Schucany, and R. W. Haas (1976). "Generating Random Variates Using Transformations with Multiple Roots". In: *The American Statistician* 30 (2), pp. 88–90.

Molenberghs, G. and G. Verbeke (2007). "Likelihood Ratio, Score, and Wald Tests in a Constrained Parameter Space". In: *The American Statistician* 61 (1), pp. 22–27.

Moraga, P. (2024). *Spatial statistics for data science: theory and practice with R*. First edition. Chapman and Hall/CRC Data Science Series. Boca Raton London New York: CRC Press. 1 p.

Nason, G. P. and J. L. Wei (2022). "Quantifying the Economic Response to COVID-19 Mitigations and Death Rates Via Forecasting Purchasing Managers' Indices Using Generalised Network Autoregressive Models with Exogenous Variables". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 185 (4), pp. 1778–1792.

Nelder, J. A. (1985). "Quasi-likelihood and GLIM". In: *Generalized linear models*. Vol. 32. Lect. Notes Stat. Springer, Berlin, pp. 120–127.

Nelder, J. A. and D. Pregibon (1987). "An extended quasilikelihood function". In: *Biometrika* 74 (2), pp. 221–232.

Nelder, J. A. and R. W. M. Wedderburn (1972). "Generalized Linear Models". In: *Journal of the Royal Statistical Society. Series A (General)* 135 (3), p. 370.

Al-Osh, M. A. and A. A. Alzaid (1987). "First-order Integer Values Autoregressive (INAR(1)) Process". In: *Journal of Time Series Analysis* 8 (3), pp. 261–275.

Otto, P., O. Doğan, S. Taşpınar, W. Schmid, and A. K. Bera (2025). "Spatial and spatiotemporal volatility models: A review". In: *Journal of Economic Surveys* 39 (3), pp. 1037–1091.

Otto, P. and W. Schmid (2018). "Spatiotemporal analysis of German real-estate prices". In: *The Annals of Regional Science* 60 (1), pp. 41–72.

Otto, P., W. Schmid, and R. Garthoff (2018). "Generalised spatial and spatiotemporal autoregressive conditional heteroscedasticity". In: *Spatial Statistics* 26, pp. 125–145.

Pace, R., R. Barry, O. W. Gilley, and C. Sirmans (2000). "A method for spatial–temporal forecasting with an application to real estate prices". In: *International Journal of Forecasting* 16 (2), pp. 229–246.

Pan, W. (2001). "Akaike's Information Criterion in Generalized Estimating Equations". In: *Biometrics* 57 (1), pp. 120–125.

Pebesma, E. and R. Bivand (2023). *Spatial Data Science: With Applications in R*. 1st ed. New York: Chapman and Hall/CRC.

Pfeifer, P. E. and S. J. Deutsch (1980a). "Stationarity and invertibility regions for low order starma models: Stationarity and invertibility regions". In: *Communications in Statistics - Simulation and Computation* 9 (5), pp. 551–562.

Pfeifer, P. E. and S. J. Deutsch (1980b). "A STARIMA Model-Building Procedure with Application to Description and Regional Forecasting". In: *Transactions of the Institute of British Geographers* 5 (3), p. 330.

Pfeifer, P. E. and S. J. Deutsch (1980c). "A Three-Stage Iterative Procedure for Space-Time Modeling". In: *Technometrics* 22 (1), p. 35.

Pfeifer, P. E. and S. J. Deutsch (1980d). "Identification and Interpretation of First Order Space-Time ARMA Models". In: *Technometrics* 22 (3), pp. 397–408.

Pfeifer, P. E. and S. J. Deutsch (1981a). "Seasonal Space-Time ARIMA Modeling". In: *Geographical Analysis* 13 (2), pp. 117–133.

Pfeifer, P. E. and S. J. Deutsch (1981b). "Variance of the Sample Space-Time Autocorrelation Function". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 43 (1), pp. 28–33.

Pfeifer, P. E. and S. J. Deutsch (1981c). "Variance of the Sample Space-Time Correlation Function of Contemporaneously Correlated Variables". In: *SIAM Journal on Applied Mathematics* 40 (1), pp. 133–136.

Schwarz, G. (1978). "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6 (2), pp. 461–464.

Self, S. G. and K.-Y. Liang (1987). "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions". In: *Journal of the American Statistical Association* 82 (398), pp. 605–610.

Shapiro, A. (1988). "Towards a Unified Theory of Inequality Constrained Testing in Multivariate Analysis". In: *International Statistical Review / Revue Internationale de Statistique* 56 (1), p. 49.

Slutzky, E. (1937). "The Summation of Random Causes as the Source of Cyclic Processes". In: *Econometrica* 5 (2), p. 105.

Smyth, G. K. (1989). "Generalized Linear Models with Varying Dispersion". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 51 (1), pp. 47–60.

Smyth, G. K. and B. Jørgensen (2002). "Fitting Tweedie's Compound Poisson Model to Insurance Claims Data: Dispersion Modelling". In: *ASTIN Bulletin* 32 (1), pp. 143–157.

Smyth, G. K. and A. P. Verbyla (1999). "Adjusted likelihood methods for modelling dispersion in generalized linear models". In: *Environmetrics* 10 (6), pp. 695–709.

Tobler, W. R. (1970). "A Computer Movie Simulating Urban Growth in the Detroit Region". In: *Economic Geography* 46, p. 234.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. New York: Springer.

West, M., P. J. Harrison, and H. S. Migon (1985). "Dynamic generalized linear models and Bayesian forecasting". In: *Journal of the American Statistical Association* 80 (389), pp. 73–97.

Wikle, C. K., A. Zammit Mangion, and N. A. C. Cressie (2019). *Spatio-temporal statistics with R*. Chapman & Hall/CRC the R series. Boca Raton London New York: CRC Press.

Wu, K. Y. K. and W. K. Li (2016). "On a dispersion model with Pearson residual responses". In: *Computational Statistics & Data Analysis* 103, pp. 17–27.

Yan, J. (2007). "Enjoy the Joy of Copulas: With a Package copula". In: *Journal of Statistical Software* 21, pp. 1–21.

Yin, C., L. Zhao, and C. Wei (2006). "Asymptotic normality and strong consistency of maximum quasi-likelihood estimates in generalized linear models". In: *Science in China Series A* 49 (2), pp. 145–157.

Yule, G. U. (1926). "Why do we Sometimes get Nonsense-Correlations between Time-Series?–A Study in Sampling and the Nature of Time-Series". In: *Journal of the Royal Statistical Society* 89 (1), p. 1.

Zhang, S., Y. Liao, and W. Ning (2011). "Asymptotic Properties of Quasi-Maximum Likelihood Estimates in Generalized Linear Models". In: *Communications in Statistics - Theory and Methods* 40 (24), pp. 4417–4430.

Zheng, T., H. Xiao, and R. Chen (2022). "Generalized autoregressive moving average models with GARCH errors". In: *Journal of Time Series Analysis* 43 (1), pp. 125–146.

Zhu, F. (2011). "A negative binomial integer-valued GARCH model". In: *Journal of Time Series Analysis* 32 (1), pp. 54–67.

Zhu, X., R. Pan, G. Li, Y. Liu, and H. Wang (2017). "Network vector autoregression". In: *The Annals of Statistics* 45 (3).

Appendix

This Appendix contains additional content that supplements the main part of this dissertation. Section A.1 contains theoretical additions to Chapter 3, in which the linear and log-linear PSTARMA(X) processes are introduced. Section A.2 contains notes on Chapter 6 and the associated R package `glmSTARMA`. The last section A.3 contains supplementary simulation results for Chapter 7.

A.1 Spatio-temporal Count Autoregession

This section contains additions to the theory of PSTARMA processes. Subsection A.1.1 contains the definition of a STARMA process, which gives PSTARMA processes their name. In the subsequent subsection, we specify the second-order properties of the processes, i.e., the conditional expected value and, for first-order models, the unconditional variance. In subection A.1.3, we take a closer look at the asymptotic properties of the QMLE and discuss the necessary regularity conditions.

A.1.1 Definition STARMA processes

Let $\{\mathbf{Z}_t = (Z_{1,t}, \dots, Z_{p,t})'\}$ denote a p-dimensional stochastic process and let $\{\boldsymbol{\varepsilon}_t = (\varepsilon_{1,t}, \dots, \varepsilon_{p,t})'\}$ denote a p-dimensional white noise process, i.e. $\mathbb{E}(\boldsymbol{\varepsilon}_t) = \mathbf{0}_p$ and

$$\mathbb{E}(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_{t+s}') = \begin{cases} \Sigma \in \mathbb{R}^{p \times p}, & \text{if } s = 0, \\ \mathbf{0}_{p \times p}, & \text{otherwise} \end{cases} \tag{A-1}$$

We refer to $\{\mathbf{Z}_t\}$ as STARMA($v_{u_1,\dots,u_v}, q_{a_1,\dots a_q}$) process with autoregressive order v and moving-average order q with associated spatial orders u_i and a_j, see Pfeifer and Deutsch (1980c), if it holds

$$\mathbf{Z}_t = \sum_{i=1}^{v} \sum_{\ell=0}^{u_i} \phi_{i\ell} \mathbf{W}^{(\ell)} \mathbf{Z}_{t-i} - \sum_{j=1}^{q} \sum_{\ell=0}^{a_j} \theta_{j\ell} \mathbf{W}^{(\ell)} \boldsymbol{\varepsilon}_{t-j} + \boldsymbol{\varepsilon}_t. \tag{A-2}$$

A.1.2 Second Order Properties

As discussed in Section 3.2.2, the linear PSTARMA($q_{a_1,\dots,a_q}, r_{b_1,\dots,b_r}$) process conforms to the model equation of a STARMA($v_{u_1,\dots,u_v}, q_{a_1,\dots,a_q}$) process, where $v = \max\{q, r\}$ and $u_i = \max\{a_i, b_i\}$ for $i = 1, \dots, v$, with $a_i = -1$ for $i > q$ and $b_i = -1$ for $i > r$. When considered stationary, i.e. fulfilling (3.5), various properties such as the unconditional expected values $\mathbb{E}(\mathbf{Y}_t)$ and $\mathbb{E}(\boldsymbol{\mu}_t)$, or the autocorrelation function, can be derived from the STARMA processes. To be more precise, for the linear PSTARMA process (3.1) it holds

$$\mathbb{E}(\mathbf{Y}_t) = \mathbb{E}(\boldsymbol{\mu}_t) = \left(\mathbf{I}_p - \sum_{i=1}^{q} \mathbf{A}_i - \sum_{j=1}^{r} \mathbf{B}_j\right)^{-1} \boldsymbol{\delta}. \tag{A-3}$$

In the case of a homogeneous PSTARMA process, (A-3) reduces to

$$\mathbb{E}(\mathbf{Y}_t) = \mathbb{E}(\boldsymbol{\mu}_t) = \delta_0 \cdot \left(1 - \sum_{i=1}^{q} \sum_{\ell=0}^{a_i} \alpha_{i\ell} - \sum_{j=1}^{r} \sum_{\ell=0}^{b_j} b_{j\ell}\right)^{-1} \mathbf{1}_p. \tag{A-4}$$

If the covariates in the linear PSTARMAX process (3.3) are stationary or constant in time, so that $\boldsymbol{\gamma} := \mathbb{E}\left(\sum_{k=1}^{m} \sum_{\ell=0}^{s_k} \gamma_{k,\ell} \mathbf{W}^{(\ell)} \mathbf{X}_{k,t}\right)$ exists and is independent of t, then it holds

$$\mathbb{E}(\mathbf{Y}_t) = \mathbb{E}(\boldsymbol{\mu}_t) = \left(\mathbf{I}_p - \sum_{i=1}^{q} \mathbf{A}_i - \sum_{j=1}^{r} \mathbf{B}_j\right)^{-1} (\boldsymbol{\delta} + \boldsymbol{\gamma}). \tag{A-5}$$

As STARMA processes are derived from modified VARMA processes, they exhibit identical autocovariance functions and expected values. In general, there's no straightforward expression for the autocovariance. However, for low-temporal-order temporal models, such as the linear PSTARMA($1_{a_1}, 1_{b_1}$) process, we can do exact calculations.

Define $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{p \times p}$ via

$$\operatorname{vec}(\boldsymbol{\Gamma}_0) := \operatorname{vec}(\Sigma) + \left(\mathbf{I}_p \otimes \mathbf{I}_p - (\mathbf{A}_1 + \mathbf{B}_1) \otimes (\mathbf{A}_1 + \mathbf{B}_1)\right)^{-1} \cdot \operatorname{vec}(\mathbf{B}_1 \Sigma \mathbf{B}_1'),$$

with vec as the vec-operator, and $\otimes$ the Kronecker product. Here, $\Sigma := \mathbb{E}[\Sigma_t]$ denotes the expected value of the covariance matrix conditioned on the past $\Sigma_t = \operatorname{Var}(\mathbf{Y}_t \mid \mathcal{F}_{t-1})$. Consequently, the autocovariance function of the linear process in (3.1) can be expressed as follows:

$$\boldsymbol{\Gamma}(h) = \begin{cases} \boldsymbol{\Gamma}_0, & h = 0 \\ (\mathbf{A}_1 + \mathbf{B}_1)\boldsymbol{\Gamma}_0 - \mathbf{A}_1 \Sigma, & h = 1 \\ (\mathbf{A}_1 + \mathbf{B}_1)^{h-1} \boldsymbol{\Gamma}(1), & h \geq 2 \end{cases}. \tag{A-6}$$

When the components of the conditional distribution $\mathbf{Y}_t \mid \mathcal{F}_{t-1}$ are uncorrelated, the covariance matrix Σ_t becomes a diagonal matrix with the diagonal elements determined

by $\boldsymbol{\mu}_t$. As a result, the covariance matrix Σ is also diagonal, with the elements of (A-3) on its main diagonal.

The expected values and autocovariance function can be explicitly specified solely for the linear PSTARMA process. For instance, considering the expected value of $(\mathbf{Y}_t, \boldsymbol{\psi}_t)$ in a stationary log-linear PSTARMA$(1_{a_1}, 1_{b_1})$ process as $(\boldsymbol{\mu}, \boldsymbol{\psi})$, we have

$$\boldsymbol{\psi} := \mathbb{E}[\boldsymbol{\psi}_t] = \left(\mathbf{I}_p - \mathbf{A}\right)^{-1} \left[\mathbf{d} + \mathbf{B} \cdot \mathbb{E}\left[\log\left(\mathbf{Y}_t + \mathbf{1}_p\right)\right]\right] \preceq \log(\boldsymbol{\mu}). \tag{A-7}$$

However, the expected value $\mathbb{E}\left[\log\left(\mathbf{Y}_t + \mathbf{1}_p\right)\right]$ of $\log\left(\mathbf{Y}_t + \mathbf{1}_p\right)$ is unknown. Consequently, we cannot rely solely on the process's parameters to determine its expected value. Nevertheless, if the elements of $\boldsymbol{\mu}$ are large (i.e., $\mu_i \gg 0$), then $\mathbb{E}\left[\log\left(\mathbf{Y}_t + \mathbf{1}_p\right)\right]$ is approximately equal to $\boldsymbol{\psi}$, which can be further approximated by $\log(\boldsymbol{\mu})$. Then $\boldsymbol{\nu}$ ist approximately equal to (A-3). Moreover, the autocovariance function (A-6) is almost equal to that of $\log\left(\mathbf{Y}_t + \mathbf{1}_p\right)$ with using $\Sigma_t = \mathrm{Var}(\log\left(\mathbf{Y}_t + \mathbf{1}_p\right) \mid \mathscr{F}_{t-1})$.

A.1.3 Asymptotic Properties of the QMLE

In the case of PSTARMA processes, the same asymptotic properties apply to parameter estimation as in the work by Fokianos et al. (2020). If the true parameter vector $\boldsymbol{\theta}_{\text{true}}$ satisfies the stability condition (3.5), then, as stated in the main part, the quasi-maximum likelihood estimator defined by $\hat{\boldsymbol{\theta}} := \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$ is strongly consistent and asymptotically normally distributed as $T \to \infty$,

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{true}}\right) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{-1}). \tag{A-8}$$

The condition that $\boldsymbol{\theta}_{\text{true}}$ is in the interior of the parameter space means that the linear PSTARMA-processes are only allowed to have positive parameter values $\boldsymbol{\theta}_{\text{true}} \succ \mathbf{0}$. The matrices $\mathbf{H}$ and $\mathbf{G}$ in (A-8) for the linear PSTARMA-process (3.1) are given by

$$\mathbf{G}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \boldsymbol{\mu}_t'}{\partial \boldsymbol{\theta}} \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \Sigma_t(\boldsymbol{\theta}) \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\mu}_t}{\partial \boldsymbol{\theta}}\right] \text{ and } \mathbf{H}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \boldsymbol{\mu}_t'}{\partial \boldsymbol{\theta}} \mathbf{D}_t^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\mu}_t}{\partial \boldsymbol{\theta}}\right], \tag{A-9}$$

and for the log-linear PSTARMA-process (3.2) by

$$\mathbf{G}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \boldsymbol{\psi}_t'}{\partial \boldsymbol{\theta}} \Sigma_t(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\psi}_t}{\partial \boldsymbol{\theta}}\right] \text{ and } \mathbf{H}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \boldsymbol{\psi}_t'}{\partial \boldsymbol{\theta}} \mathbf{D}_t(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\psi}_t}{\partial \boldsymbol{\theta}}\right]. \tag{A-10}$$

For the calculation of standard errors, the matrices $\mathbf{H}$ and $\mathbf{G}$ can be estimated by their empirical counterparts by substituting $\hat{\boldsymbol{\theta}}$.

If covariates are included in PSTARMAX processes, they must fulfil certain conditions to ensure consistent parameter estimation. First, the model has to be identifiable, see subsection 3.2.1. In addition, the PSTARMAX processes are related to multivariate GLMs,

so that it can be assumed that the conditions that ensure the consistency and asymptotic normality of maximum and quasi-maximum likelihood estimates, as elaborated in the work of Chen et al. (1999), Fahrmeir and Kaufmann (1985), Gao et al. (2012), Yin et al. (2006), and Zhang et al. (2011), are also sufficient here. In total we suppose that the QMLE $\hat{\boldsymbol{\theta}}$ in the PSTARMAX-processes for the parameter vector $\boldsymbol{\theta}_{\text{true}}$ is consistent and asymptotically normal (A-8), if the following conditions hold:

1. The parameters of the true underlying process fulfil condition (3.5).
2. All covariate processes are bounded and have finite variance, i.e.,
   - for deterministic covariate processes $\{\mathbf{X}_{k,t}\}, k = 1, \dots, m$, there is a $c > 0$ such that for all $t, k : \|\mathbf{X}_{k,t}\|_2^2 < c$.
   - for random covariate processes $\{\mathbf{X}_{k,t}\}, k = 1, \dots, m$, there is a $c > 0$ such that $\forall t, \forall k : \mathbb{E}\left[\|\mathbf{X}_{k,t}\|_2^2\right] < c$.
3. The processes
$$\{\mathbf{W}^{(\ell)}\mathbf{X}_{k,t}\},\ k = 1, \dots, m,\ \ell = 0, \dots, c_k,$$
are linearly independent.
4. There is a $\varepsilon > 1$ such that $\frac{\overline{\sigma_T}^{0.5}\log(\overline{\sigma_T})^{0.5\varepsilon}}{\underline{\sigma_T}} \to 0$ as $T \to \infty$, where $\overline{\sigma_T}(\underline{\sigma_T})$ is the largest (smallest) eigenvalue of the matrix $\mathbf{F}_T = \sum_{t=1}^{T} \tilde{\mathbf{X}}_t' \tilde{\mathbf{X}}_t$ with
$$\tilde{\mathbf{X}}_t = \left(\mathbf{1}_p, \mathbf{W}^{(0)}\mathbf{X}_{1,t}, \dots, \mathbf{W}^{(s_1)}\mathbf{X}_{1,t}, \dots, \mathbf{W}^{(s_m)}\mathbf{X}_{m,t}\right).$$

Remark to conditions: Condition 1 ensures the stability of the autoregressive parts of the model, while condition 2 ensures the stability of the covariates. Condition 3 implies that the covariate parameters are identifiable. Condition 4 implies that the *information* from the covariate processes increases fast enough as the number of observation points increases and thus the consistency of the parameter estimates, see Zhang et al. (2011). In the case of an inhomogeneous intercept structure the vector $\mathbf{1}_p$ is replaced by the matrix $\mathbf{I}_p$ to verify condition 4.

In practice, conditions 2-4 are satisfied by numerous examples of covariate processes. For instance, consider a time-constant covariate $\mathbf{X}_{k,t} = \tilde{\mathbf{X}} \quad \forall t$, where $\tilde{\mathbf{X}} \in \mathbb{R}^p$ and $\tilde{\mathbf{X}} \neq c \cdot \mathbf{1}_p, c \in \mathbb{R}$. Such a covariate, for example the population size of a region, with maximum spatial order $s_k = 0$ satisfies the conditions. Furthermore, two spatially constant covariates $\mathbf{X}_{1,t} = \sin(2\pi t/n)\mathbf{1}_p$ and $\mathbf{X}_{2,t} = \cos(2\pi t/n)\mathbf{1}_p$ also fulfill these conditions. In time series analysis, these covariates are employed to model a seasonality of length $n \in \mathbb{N}$, $n > 1$. For $T = kn, k \in \mathbb{N}$, the matrix $\mathbf{F}_T$ from condition 4 becomes a diagonal matrix with $\text{diag}(\mathbf{F}_T) = (Tp, 0.5Tp, 0.5Tp)'$, thereby satisfying condition 4.

A.2 Derivation of the expected Hessian for spatio-temporal Double Generalized Linear Models

This section provides a derivation of the expected Hessian for the spatio-temporal double generalized linear models, which is relevant for determining the variance of the parameter estimators.

The Hessian is obtained with the common derivation rules as $\mathbf{H}_T(\boldsymbol{\theta}) = \frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} = \frac{\partial S_T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}$. For the Hessian we define $\tilde{\mathbf{D}}_t = \operatorname{diag}\left\{\left[g'(g^{-1}(\psi_{i,t}))\right]^2 g^{-1}(\psi_{i,t}), i = 1, \dots, n\right\}$.

$$
\begin{aligned}
\mathbf{H}_T(\boldsymbol{\theta}) &= \sum_{t=1}^{T}\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\left[\frac{\partial \psi_{i,t}}{\partial \boldsymbol{\theta}} \frac{y_{i,t} - g^{-1}(\psi_{i,t})}{g'(g^{-1}(\psi_{i,t}))g^{-1}(\psi_{i,t})}\right] \\
&= \underbrace{\sum_{t=1}^{T}\sum_{i=1}^{n} \frac{\partial^2 \psi_{i,t}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \frac{y_{i,t} - g^{-1}(\psi_{i,t})}{g'(g^{-1}(\psi_{i,t}))g^{-1}(\psi_{i,t})}}_{(\star)} + \sum_{t=1}^{T}\sum_{i=1}^{n} \frac{\partial \psi_{i,t}}{\partial \boldsymbol{\theta}} \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\left[\frac{y_{i,t} - g^{-1}(\psi_{i,t})}{g'(g^{-1}(\psi_{i,t}))g^{-1}(\psi_{i,t})}\right] \\
&= (\star) + \sum_{t=1}^{T}\sum_{i=1}^{n} \frac{\partial \psi_{i,t}}{\partial \boldsymbol{\theta}} \frac{\frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}}(y_{i,t} - g^{-1}(\psi_{i,t})) \cdot g'(g^{-1}(\psi_{i,t}))g^{-1}(\psi_{i,t})}{\left[g'(g^{-1}(\psi_{i,t}))g^{-1}(\psi_{i,t})\right]^2} \\
&\quad - \sum_{t=1}^{T}\sum_{i=1}^{n} \frac{\partial \psi_{i,t}}{\partial \boldsymbol{\theta}} \frac{(y_{i,t} - g^{-1}(\psi_{i,t}))\frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\left(g'(g^{-1}(\psi_{i,t}))g^{-1}(\psi_{i,t})\right)}{\left[g'(g^{-1}(\psi_{i,t}))g^{-1}(\psi_{i,t})\right]^2} \\
&= (\star) - \sum_{t=1}^{T}\sum_{i=1}^{n} \frac{\partial \psi_{i,t}}{\partial \boldsymbol{\theta}} \frac{\frac{g'(g^{-1}(\psi_{i,t}))g^{-1}(\psi_{i,t})}{g'(g^{-1}(\psi_{i,t}))}}{\left[g'(g^{-1}(\psi_{i,t}))g^{-1}(\psi_{i,t})\right]^2} \frac{\partial \psi_{i,t}}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \\
&\quad - \sum_{t=1}^{T}\sum_{i=1}^{n} \frac{\partial \psi_{i,t}}{\partial \boldsymbol{\theta}} \frac{(y_{i,t} - g^{-1}(\psi_{i,t}))\left[\frac{g''(g^{-1}(\psi_{i,t}))}{g'(g^{-1}(\psi_{i,t}))} g^{-1}(\psi_{i,t}) + \frac{g'(g^{-1}(\psi_{i,t}))}{g'(g^{-1}(\psi_{i,t}))}\right]}{\left[g'(g^{-1}(\psi_{i,t}))g^{-1}(\psi_{i,t})\right]^2} \frac{\partial \psi_{i,t}}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \\
&= (\star) - \sum_{t=1}^{T}\sum_{i=1}^{n} \frac{\partial \psi_{i,t}}{\partial \boldsymbol{\theta}} \frac{1}{\left[g'(g^{-1}(\psi_{i,t}))\right]^2 g^{-1}(\psi_{i,t})} \frac{\partial \psi_{i,t}}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \\
&\quad - \sum_{t=1}^{T}\sum_{i=1}^{n} \frac{\partial \psi_{i,t}}{\partial \boldsymbol{\theta}} \frac{(y_{i,t} - g^{-1}(\psi_{i,t}))\left[g''(g^{-1}(\psi_{i,t}))g^{-1}(\psi_{i,t}) + g'(g^{-1}(\psi_{i,t}))\right]}{\left[g'(g^{-1}(\psi_{i,t}))\right]^3 \left[g^{-1}(\psi_{i,t})\right]^2} \frac{\partial \psi_{i,t}}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \\
&= \sum_{t=1}^{T}\sum_{i=1}^{n} \frac{\partial^2 \psi_{i,t}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \frac{y_{i,t} - g^{-1}(\psi_{i,t})}{g'(g^{-1}(\psi_{i,t}))g^{-1}(\psi_{i,t})} \\
&\quad - \sum_{t=1}^{T}\sum_{i=1}^{n} \frac{\partial \psi_{i,t}}{\partial \boldsymbol{\theta}} \frac{(y_{i,t} - g^{-1}(\psi_{i,t}))\left[g''(g^{-1}(\psi_{i,t}))g^{-1}(\psi_{i,t}) + g'(g^{-1}(\psi_{i,t}))\right]}{\left[g'(g^{-1}(\psi_{i,t}))\right]^3 \left[g^{-1}(\psi_{i,t})\right]^2} \frac{\partial \psi_{i,t}}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \\
&\quad - \sum_{t=1}^{T} \frac{\partial \boldsymbol{\psi}_t'}{\partial \boldsymbol{\theta}} \tilde{\mathbf{D}}_t^{-1} \frac{\partial \boldsymbol{\psi}_t'}{\partial \boldsymbol{\theta}^{\mathrm{T}}}.
\end{aligned}
$$

By taking expectation, we receive the conditional information matrix as

$$\mathbf{G}_T(\boldsymbol{\theta}) = \mathbb{E}\left[\mathbf{H}_T(\boldsymbol{\theta})\right] = -\sum_{t=1}^{T} \frac{\partial \boldsymbol{\psi}_t'}{\partial \boldsymbol{\theta}} \tilde{\mathbf{D}}_t^{-1} \frac{\partial \boldsymbol{\psi}_t'}{\partial \boldsymbol{\theta}^{\mathrm{T}}}$$

A.3 Additional Simulation Results

In this section, we present additional simulation results supplementing the investigations from Chapter 7. Subsection A.3.1 contains extended results on PSTARMA processes, building on the master's thesis of the author of this dissertation. Section A.3.2 repeats the simulations from Section 7.1 under a different copula dependency structure to verify the consistency of the results. The remaining figures from Section 7.3 are provided in Section A.3.4.

A.3.1 Further Simulation Results for PSTARMAX Processes

We provide further simulation results extending those of Maletz (2021) by adjusting the underlying parameters and increasing the number of repetitions to 1,000. A summary of these simulations is presented in Table A-1. Except for the *Copula* simulation, all simulations use the same copula in the data-generating process as in Section 7.1.

Normal Approximation

Figure A-1 shows QQ-plots of the parameter estimates for $T = 100$ observations on a 5×5 grid in the log-linear model, indicating an adequate approximation to the normal distribution. Similar results are observed for other process dimensions; see Figures A-2, A-3, and A-4.

For the linear model, we observe stronger deviations from the asymptotic normal distribution, attributable to the parameter constraints. At $T = 250$, the distributions of all parameter estimators except $\alpha_{0,0}$ and $\alpha_{0,1}$ are well approximated by a normal distribution. For $T = 750$ on the 5×5 grid, approximately 12% of the estimates for $\alpha_{0,1}$ still lie in the range $[0, 10^{-5})$, i.e., are effectively zero, despite the true value being 0.2.

Figure A-2 shows QQ-plots of the asymptotic parameter distribution for the log-linear PSTARMAX model on a 9×9 grid. The results are qualitatively similar to those on the 5×5 grid in Figure A-1, with only minor differences. Figures A-3 and A-4 present results for the linear PSTARMAX model on the same 9×9 grid at two different sample sizes. While substantial deviations from normality are evident at $T = 100$ due to parameter constraints, the approximation is adequate at $T = 750$, with the exception of $\alpha_{0,0}$ and $\alpha_{0,1}$.

<table>
<tr><th>Study</th><th>Description</th><th>Model Orders & Data</th><th>True Parameters</th></tr>
<tr><td>Normal Approximation</td><td>Verification of (3.17) and (A-8).</td><td>• $T \in \{5, 10, 20\} \cup \{50, 100, 250, 500, 750\}$
• Grid sizes: $n \times n$ ($p = n^2$) $n \in \{5, 7, 9, 10, 20, 30, 40, 50\}$
• Maximum time lags: $(q, r) \in \{(1, 1), (0, 1)\}$
• Number of covariates: $m \in \{0, 1\}$
• Neighborhood matrices: $\mathbf{W}^{(0)} = \mathbf{I}_p$, $\mathbf{W}^{(1)} = \mathbf{W}_{\text{4NN}}$</td><td>
<table>
<tr><th></th><th>linear</th><th>log-linear</th></tr>
<tr><td>δ_0</td><td>5</td><td>0.6</td></tr>
<tr><td>$\alpha_{0,1}$</td><td>$[0, 0.5]$</td><td>$[-0.25, 0.25]$</td></tr>
<tr><td>$\alpha_{1,1}$</td><td>0.1</td><td>0.1</td></tr>
<tr><td>$\beta_{0,1}$</td><td>$[0, 0.5]$</td><td>$[-0.25, 0.25]$</td></tr>
<tr><td>$\beta_{1,1}$</td><td>0.1 0.25</td><td>0.1 0.25</td></tr>
<tr><td>$\gamma_{0,1}$</td><td>$[0, 0.5]$</td><td>$[-0.25, 0.25]$</td></tr>
</table>
Remark: At any time, only one parameter is varied over the above intervals; the remaining two are held fixed ($\alpha_{0,1} = 0.2$, $\beta_{0,1} = 0.2$, and $\gamma_{0,1} = 2$ for linear and $\gamma_{0,1} = 0.9$ for log-linear models). Parameter sets without the feedback process are only used for $p \geq 100$.
</td></tr>
<tr><td>Copula</td><td>Effect of different copulas on parameter estimation, compared with contemporaneously independent data. Copulas used with parameters:
• Frank: $\{0.5, 1, \dots, 3\}$
• Clayton: $\{0.5, 1, \dots, 3\}$
• Joe: $\{1.5, 2, 2.5, 3\}$
• Independent</td><td>• $T \in \{50, 100, 250, 500\}$
• Grid size: 9×9 ($p = 81$)
• Maximum time lags: $(q, r) = (1, 1)$
• Number of covariates: $m = 0$
• Neighborhood matrices: $\mathbf{W}^{(0)} = \mathbf{I}_p$, $\mathbf{W}^{(1)} = \mathbf{W}_{\text{4NN}}$</td><td>
<table>
<tr><th></th><th>linear</th><th>log-linear</th></tr>
<tr><td>δ_0</td><td>5</td><td>0.6</td></tr>
<tr><td>$\alpha_{0,1}$</td><td>0.2</td><td>0.2</td></tr>
<tr><td>$\alpha_{1,1}$</td><td>0.1</td><td>0.1</td></tr>
<tr><td>$\beta_{0,1}$</td><td>0.2</td><td>0.2</td></tr>
<tr><td>$\beta_{1,1}$</td><td>0.1</td><td>0.1</td></tr>
</table>
</td></tr>
<tr><td>Intercept</td><td>Estimation of models with a misspecified intercept structure.</td><td>• $T \in \{50, 100, 250, 500\}$
• Grid size: 9×9 ($p = 81$)
• Maximum time lags: $(q, r) = (1, 1)$
• Number of covariates: $m = 0$
• Neighborhood matrices: $\mathbf{W}^{(0)} = \mathbf{I}_p$, $\mathbf{W}^{(1)} = \mathbf{W}_{\text{4NN}}$</td><td>
<table>
<tr><th></th><th>linear</th><th>log-linear</th></tr>
<tr><td>δ_0</td><td>5</td><td>0.6</td></tr>
<tr><td>$\alpha_{0,1}$</td><td>0.2</td><td>0.2</td></tr>
<tr><td>$\alpha_{1,1}$</td><td>0.1</td><td>0.1</td></tr>
<tr><td>$\beta_{0,1}$</td><td>0.2</td><td>0.2</td></tr>
<tr><td>$\beta_{1,1}$</td><td>0.1</td><td>0.1</td></tr>
</table>
Remark: The intercept vectors are defined elementwise as $\boldsymbol{\delta}_{\text{linear}} = (\delta_i)$ and $\boldsymbol{\delta}_{\log} = (\tilde{\delta}_i)$ for $i = 1, \dots, 81$, with $\delta_i = 2 + 3 \cdot \frac{i}{81}$ and $\tilde{\delta}_i = 0.6 \cdot \frac{i}{81}$.
</td></tr>
<tr><td>Link</td><td>Estimation under a misspecified model type (linear vs. log-linear).</td><td>• $T \in \{50, 100, 250, 500\}$
• Grid size: 9×9 ($p = 81$)
• Maximum time lags: $(q, r) = (1, 1)$
• Number of covariates: $m = 0$
• Neighborhood matrices: $\mathbf{W}^{(0)} = \mathbf{I}_p$, $\mathbf{W}^{(1)} = \mathbf{W}_{\text{4NN}}$</td><td>
<table>
<tr><th></th><th>linear</th><th colspan="2">log-linear</th></tr>
<tr><td>δ_0</td><td>5</td><td>0.6</td><td>0.6</td></tr>
<tr><td>$\alpha_{0,1}$</td><td>0.2</td><td>0.2</td><td>-0.2</td></tr>
<tr><td>$\alpha_{1,1}$</td><td>0.1</td><td>0.1</td><td>0.1</td></tr>
<tr><td>$\beta_{0,1}$</td><td>0.2</td><td>0.2</td><td>-0.2</td></tr>
<tr><td>$\beta_{1,1}$</td><td>0.1</td><td>0.1</td><td>0.1</td></tr>
</table>
</td></tr>
</table>

Table A-1: Summary of additional simulations. Contemporaneous correlations between the counts are generated following the data-generating process of Fokianos et al. (2020), using the Clayton copula with parameter 2, for the *Normal Approximation*, *Intercept*, and *Link* simulations.

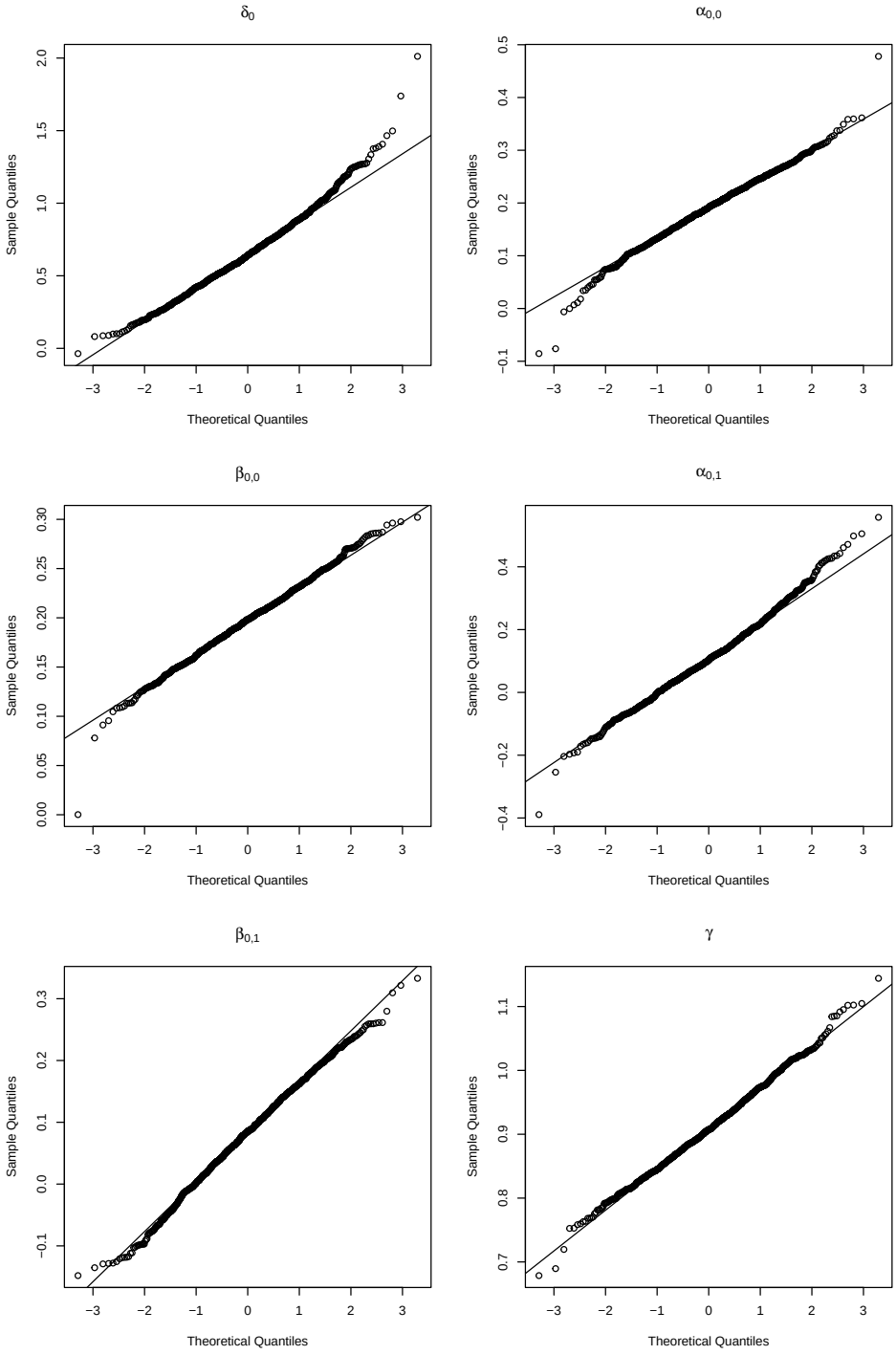


Figure A-1: QQ-plots for the QMLE of the log-linear PSTARMAX process on a 5×5 grid with $T = 100$ observation times in the "Normal Approximation" simulation from Table A-1.

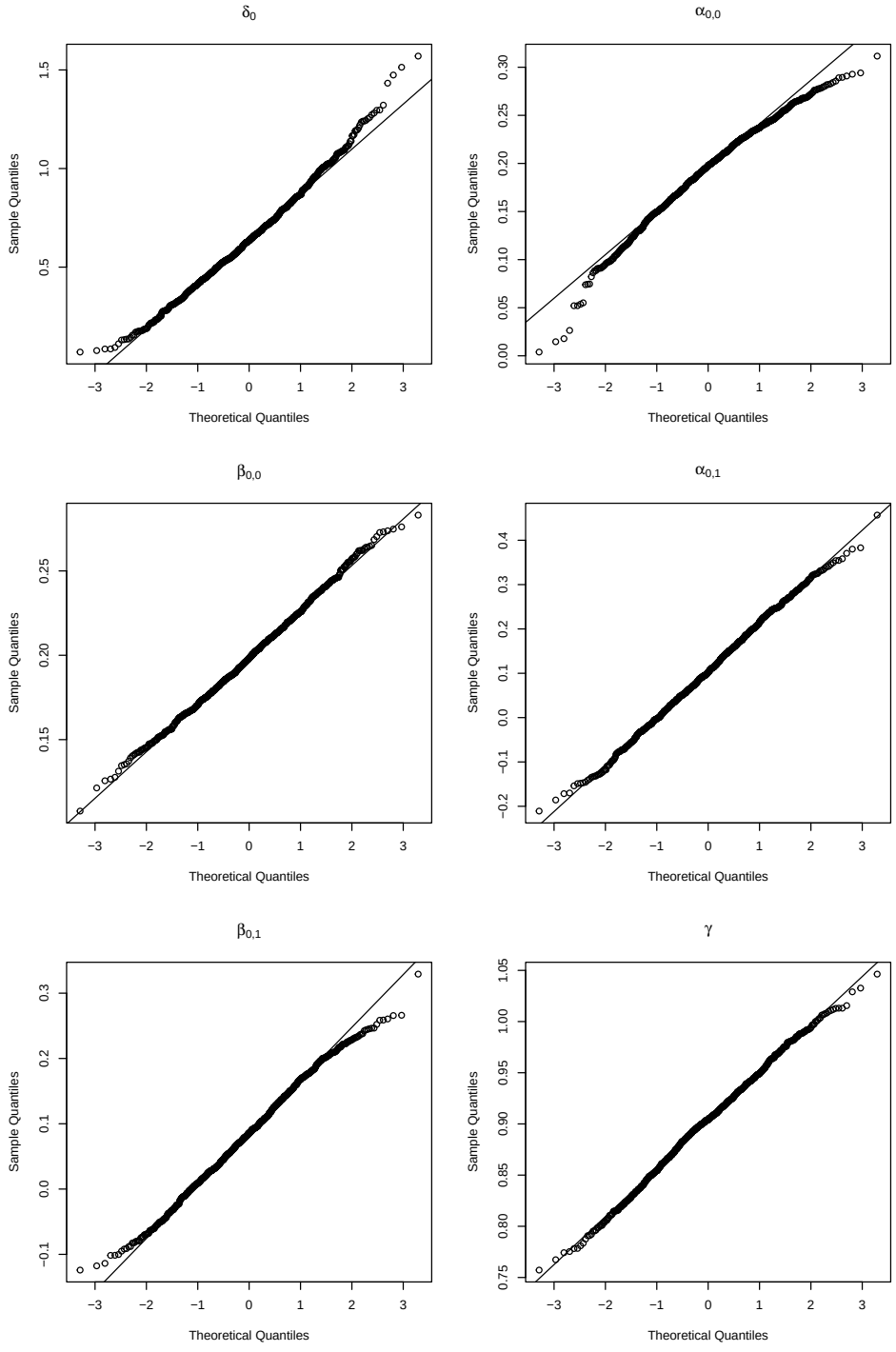


Figure A-2: QQ-plots for the QMLE of the log-linear PSTARMAX process on a 9×9 grid with $T = 100$ observation times in the "Normal Approximation" simulation from Table A-1.

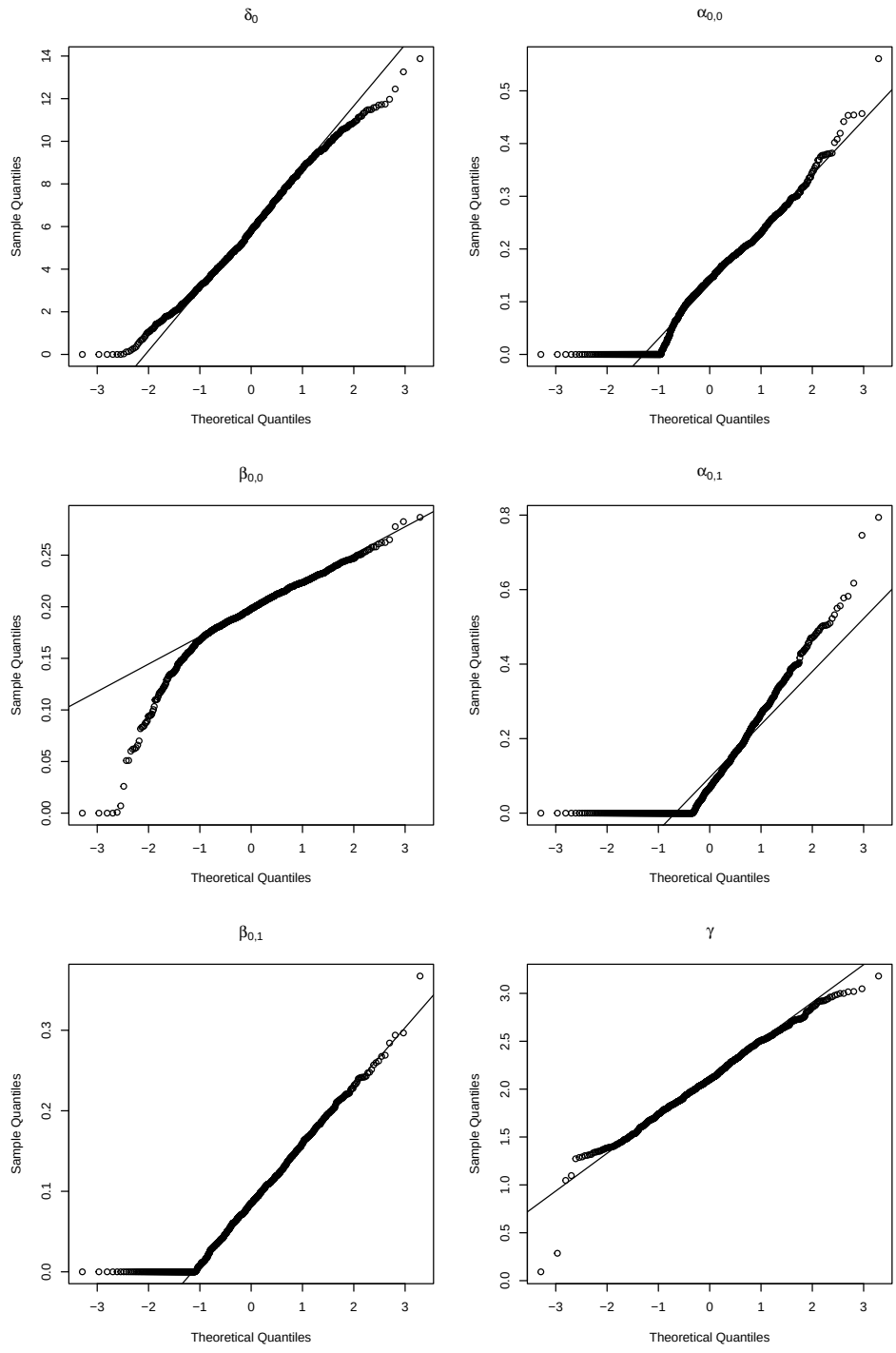


Figure A-3: QQ-plots for the QMLE of the linear PSTARMAX process on a 9×9 grid with $T = 100$ observation times in the "Normal Approximation" simulation from Table A-1.

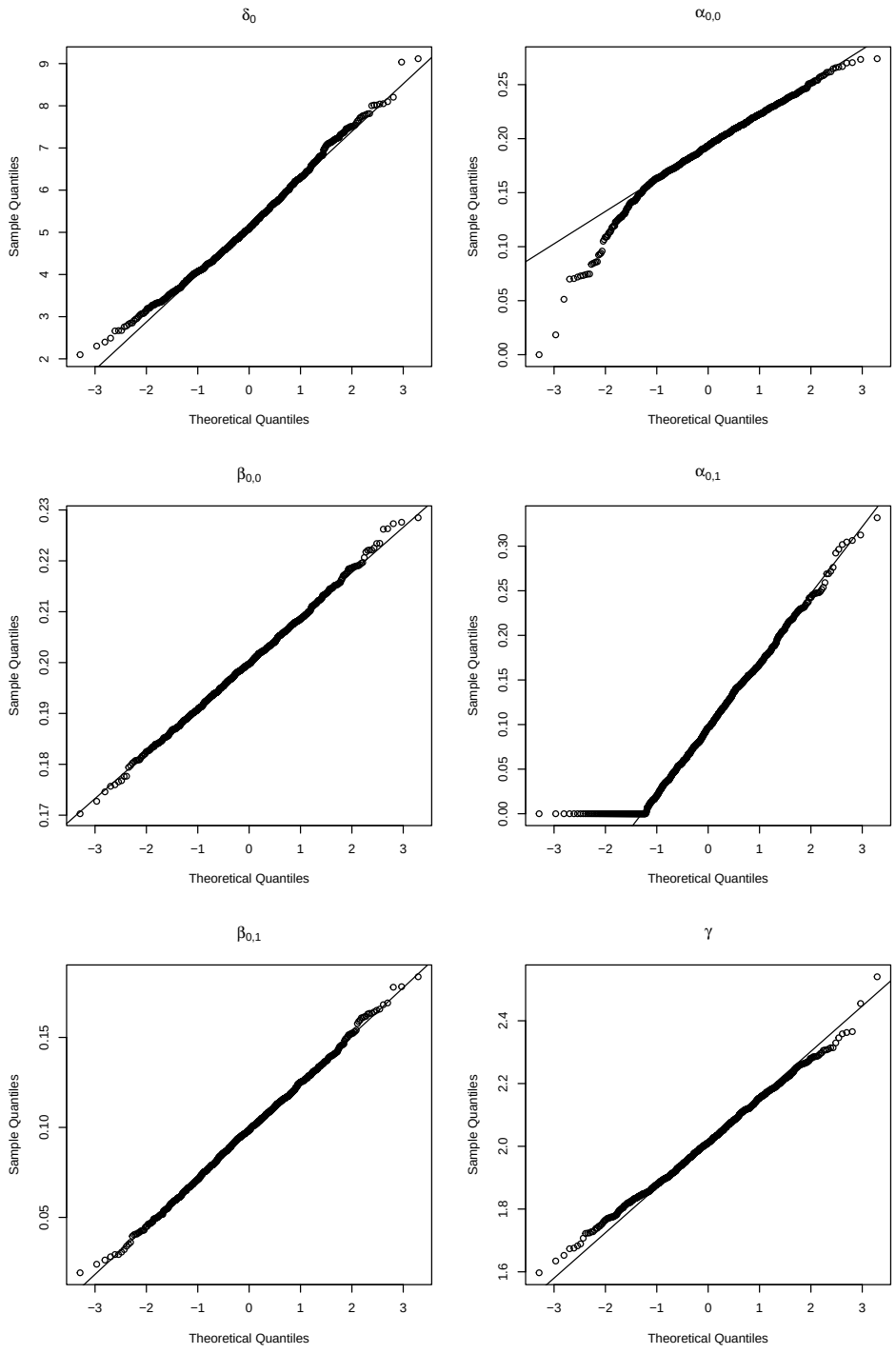


Figure A-4: QQ-plots for the QMLE of the linear PSTARMAX process on a 9×9 grid with $T = 750$ observation times in the "Normal Approximation" simulation from Table 7.1.

Wald Test

We now investigate the properties of the asymptotic Wald test. Table A-2 reports its Type I error at the nominal 5% significance level across different grid sizes and sample sizes. For each data-generating process, one of the parameters $\alpha_{0,1}$, $\beta_{0,1}$, or $\gamma_{0,1}$ is set to zero while the remaining parameters are held constant; see Table A-1. At $T = 50$, most configurations deviate from the target level of 0.05, but convergence to the nominal level is achieved as T increases. Notably, the test for $\alpha_{0,1}$ in the linear model tends to be conservative, with empirical sizes around 0.04, likely reflecting the slower convergence to normality for this parameter discussed above.

Models without a feedback term converge more quickly to the nominal significance level, as shown in Table A-4. At $T = 5$, the empirical size for the autoregressive parameter $\beta_{0,1}$ in the linear model is already close to 0.05, suggesting less conservative behavior than for processes with a feedback term. The test for covariate effects requires larger T, with adequate convergence achieved at $T = 50$ in the linear model.

Comparing the linear and log-linear models, the linear model reaches the nominal 5% level with fewer observations. This also holds in high-dimensional settings with few observation times, as shown in Table A-3. The number of locations appears to affect inference primarily for the feedback term: at $T = 5$, larger grids yield higher Type I errors, e.g., 0.345 on the 50×50 grid compared to 0.279 on the 10×10 grid in the linear model.

In summary, asymptotic approximations are adequate for models including the feedback process when a sufficiently large number of observations is available. Detecting significance for the feedback process is more demanding than for the lagged observation process, which is also reflected in the power analysis.

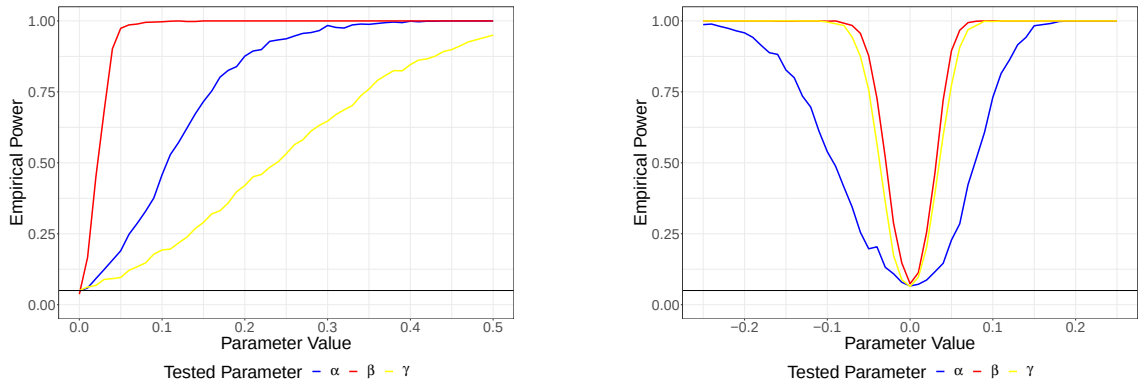
Figures A-5a and A-5b show power curves at $T = 250$ for linear and log-linear models on the 9×9 grid. In the log-linear model, deviations of $\beta_{0,1}$ and $\gamma_{0,1}$ from zero are detected equally well and more readily than deviations of $\alpha_{0,1}$. The linear model shows a similar pattern, with $\alpha_{0,1}$ being detected less efficiently than $\beta_{0,1}$; however, the power for the covariate parameter is lower than in the log-linear model. This is consistent with the underlying model structure: in the linear model, covariates enter the conditional mean additively, whereas in the log-linear model, their effect is multiplicative.

When T is small and observations are collected on a large grid, a strong autoregressive effect is needed for reliable detection. Figures A-6a and A-6b display power curves for $\beta_{0,1}$ and $\gamma_{0,1}$ in linear and log-linear PSTARMAX processes without a feedback term on a 50×50 grid at $T = 5$. As expected, the power for both parameters is lower than in the lower-dimensional case with feedback term shown in Figures A-5a and A-5b.

Table A-5 reports mean computation times for the models in Table A-2. As discussed in Section 7.1.1, computation times are short and increase with both the number of observation times and the grid size.

Parameter	Model	Gridsize	T				
			50	100	250	500	750
$\gamma_{0,1}$	log-linear	5×5	0.076	0.061	0.052	0.052	0.052
		7×7	0.057	0.045	0.051	0.053	0.042
		9×9	0.066	0.058	0.064	0.052	0.063
	linear	5×5	0.067	0.054	0.056	0.042	0.047
		7×7	0.062	0.054	0.053	0.057	0.059
		9×9	0.060	0.066	0.050	0.046	0.046
$\alpha_{0,1}$	log-linear	5×5	0.151	0.108	0.075	0.064	0.062
		7×7	0.149	0.096	0.069	0.052	0.050
		9×9	0.196	0.116	0.066	0.065	0.051
	linear	5×5	0.047	0.035	0.038	0.040	0.035
		7×7	0.054	0.035	0.037	0.034	0.041
		9×9	0.046	0.043	0.048	0.037	0.042
$\beta_{0,1}$	log-linear	5×5	0.128	0.093	0.081	0.056	0.051
		7×7	0.131	0.101	0.085	0.049	0.057
		9×9	0.145	0.115	0.074	0.058	0.046
	linear	5×5	0.043	0.043	0.053	0.057	0.055
		7×7	0.044	0.044	0.0046	0.070	0.057
		9×9	0.046	0.042	0.037	0.058	0.045

Table A-2: Empirical size of the asymptotic Wald test for different parameters and grid sizes.



(a) linear

(b) log-linear

Figure A-5: Power curves of the PSTARMAX process on a 9×9 grid with 250 observation times.

Copula

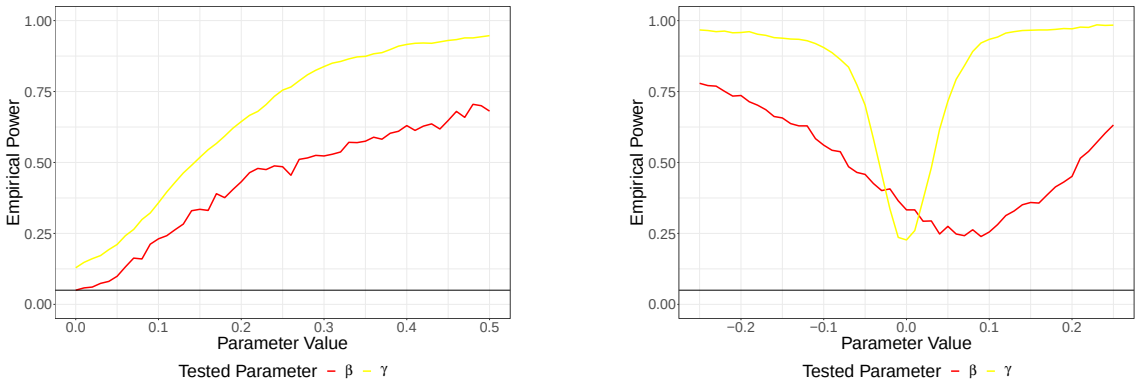
In Section 7.1, we employed the Clayton copula with parameter 2 following the data-generating process of Fokianos et al. (2020). The present study extends this by examining the effect of different copula families and their parameters on estimation accuracy. We

Parameter	Model	Gridsize	T				
			5	10	20	50	100
$\gamma_{0,1}$	log-linear	10×10	0.211	0.111	0.091	0.063	0.057
		20×20	0.238	0.141	0.106	0.061	0.051
		30×30	0.190	0.153	0.094	0.069	0.067
		40×40	0.193	0.175	0.109	0.073	0.051
		50×50	0.196	0.134	0.122	0.067	0.064
	linear	10×10	0.122	0.107	0.079	0.070	0.057
		20×20	0.149	0.109	0.072	0.057	0.054
		30×30	0.108	0.129	0.070	0.069	0.061
		40×40	0.111	0.108	0.082	0.086	0.080
		50×50	0.114	0.129	0.087	0.075	0.059
$\alpha_{0,1}$	log-linear	10×10	0.232	0.319	0.296	0.156	0.110
		20×20	0.289	0.350	0.285	0.177	0.110
		30×30	0.282	0.337	0.301	0.202	0.135
		40×40	0.317	0.354	0.302	0.205	0.119
		50×50	0.313	0.347	0.294	0.204	0.134
	linear	10×10	0.279	0.222	0.134	0.071	0.048
		20×20	0.284	0.240	0.149	0.089	0.061
		30×30	0.344	0.254	0.158	0.087	0.050
		40×40	0.357	0.265	0.167	0.082	0.067
		50×50	0.345	0.273	0.162	0.066	0.051
$\beta_{0,1}$	log-linear	10×10	0.216	0.205	0.179	0.138	0.103
		20×20	0.223	0.237	0.209	0.176	0.104
		30×30	0.193	0.210	0.219	0.165	0.128
		40×40	0.200	0.211	0.194	0.165	0.115
		50×50	0.208	0.238	0.209	0.166	0.123
	linear	10×10	0.006	0.032	0.038	0.043	0.057
		20×20	0.006	0.026	0.037	0.055	0.059
		30×30	0.007	0.028	0.037	0.045	0.067
		40×40	0.010	0.023	0.028	0.053	0.071
		50×50	0.004	0.022	0.027	0.056	0.059

Table A-3: Empirical size of the asymptotic Wald test for different parameters and grid sizes in high-dimensional settings with few observation times.

Parameter	Model	Gridsize	T				
			5	10	20	50	100
$\gamma_{0,1}$	log-linear	10×10	0.222	0.126	0.096	0.072	0.060
		20×20	0.250	0.122	0.093	0.070	0.053
		30×30	0.210	0.143	0.096	0.056	0.051
		40×40	0.222	0.156	0.100	0.073	0.066
		50×50	0.227	0.149	0.087	0.059	0.048
	linear	10×10	0.133	0.094	0.086	0.063	0.055
		20×20	0.145	0.105	0.085	0.060	0.054
		30×30	0.153	0.133	0.078	0.065	0.051
		40×40	0.119	0.118	0.084	0.062	0.063
		50×50	0.129	0.112	0.072	0.051	0.047
$\beta_{0,1}$	log-linear	10×10	0.299	0.176	0.117	0.073	0.064
		20×20	0.329	0.192	0.111	0.074	0.069
		30×30	0.310	0.185	0.123	0.063	0.061
		40×40	0.320	0.165	0.119	0.083	0.068
		50×50	0.333	0.210	0.125	0.086	0.070
	linear	10×10	0.060	0.044	0.061	0.049	0.045
		20×20	0.060	0.042	0.030	0.044	0.049
		30×30	0.050	0.039	0.032	0.054	0.054
		40×40	0.055	0.047	0.047	0.050	0.053
		50×50	0.050	0.039	0.037	0.043	0.061

Table A-4: Empirical size of the asymptotic Wald test for different parameters and grid sizes in high-dimensional settings with few observation times, for processes without a feedback term.



(a) linear (b) log-linear

Figure A-6: Power curves of the PSTARMAX process without regression on the feedback process on a 50×50 grid with 5 observation times.

Model	Gridsize	T				
		50	100	250	500	750
linear	5×5	0.012	0.024	0.063	0.135	0.209
	7×7	0.022	0.045	0.121	0.262	0.409
	9×9	0.035	0.074	0.206	0.466	0.736
log-linear	5×5	0.020	0.039	0.093	0.194	0.306
	7×7	0.036	0.071	0.179	0.388	0.593
	9×9	0.058	0.111	0.307	0.703	1.015

Table A-5: Mean computation time (in seconds) of PSTARMAX processes for different grid sizes and observation times, including regression on the feedback process.

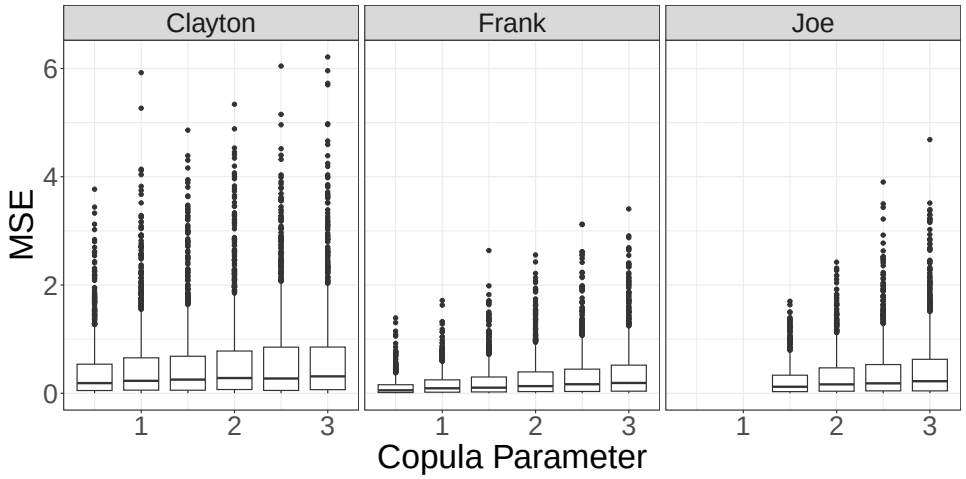
evaluate the MSE and bias (i.e., $\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{true}}$) of the parameter estimates, as well as the Mean Absolute Error (MAE) of the fitted values, defined as

$$\text{MAE} = \frac{1}{pT} \sum_{t=1}^{T} \sum_{i=1}^{p} |y_{i,t} - \mu_{i,t}(\hat{\boldsymbol{\theta}})|, \tag{A-11}$$

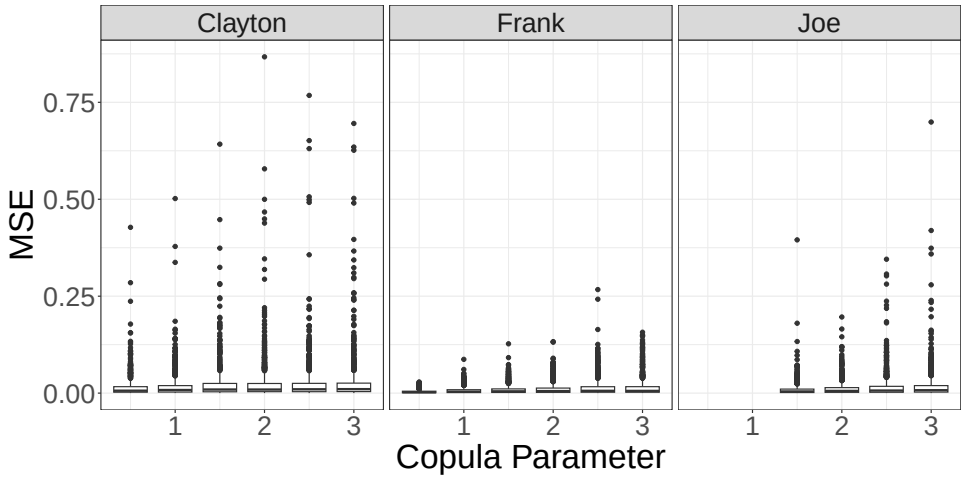
where p denotes the number of locations.

Across all copulas, the MSE increases with the copula parameter and exhibits greater variability at higher dependence levels. Table A-6 reports mean bias values for the linear model at $T = 250$ under the Clayton copula. The bias increases slightly with the copula parameter, but the mean squared bias is small relative to the total MSE, indicating that the increase in MSE is driven primarily by higher estimation variance rather than increased bias. For reference, results under contemporaneous independence are also included; these are substantially smaller.

These findings are corroborated by the MAE in the data. Figure A-8 shows boxplots of the MAEs across simulations. The central tendency remains roughly constant across copula



(a) linear



(b) log-linear

Figure A-7: MSE of parameter estimation under different copula specifications for $T = 250$ observation times.

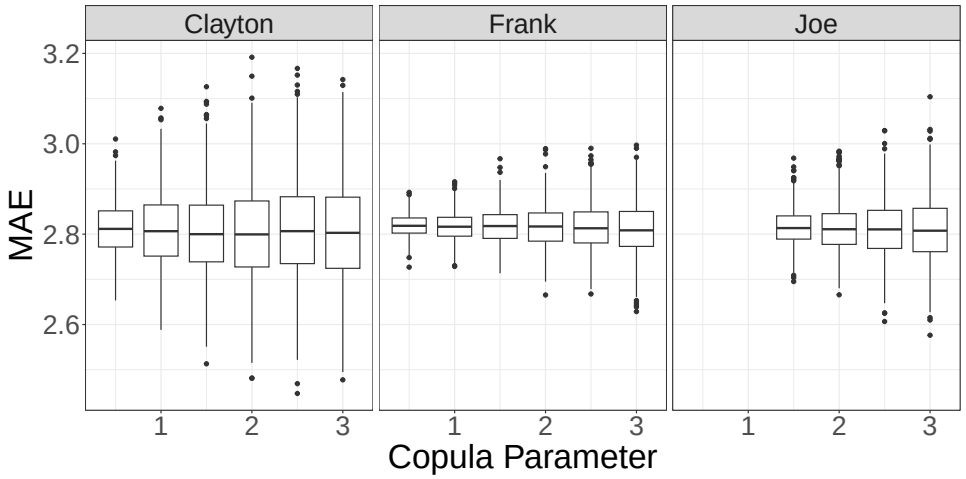
specifications, but variability increases with the copula parameter. Overall, stronger contemporaneous dependencies do not lead to substantially biased estimates, but do result in greater estimation variability.

Intercept

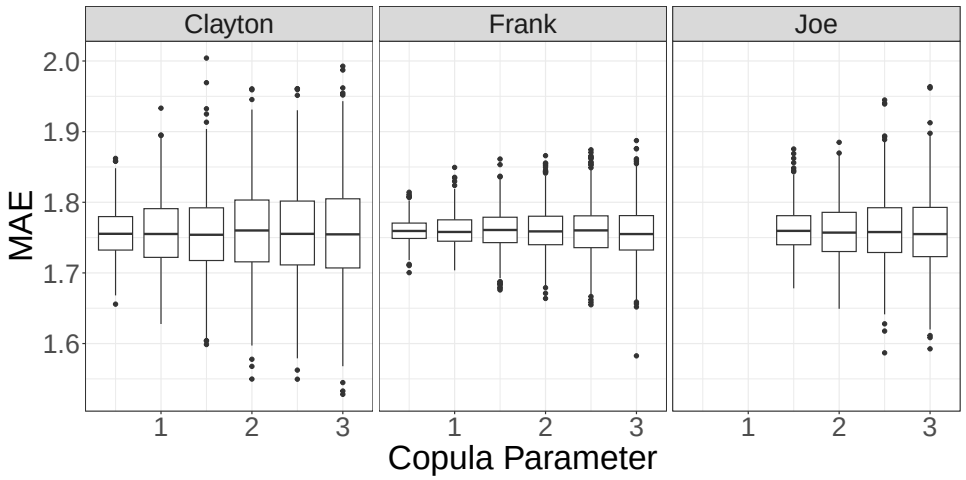
The simulations in Section 7.1 employed a homogeneous intercept structure, which reduces the number of parameters to be estimated and allows for the inclusion of time-constant covariates capturing, for instance, demographic effects such as population size. When such information is unavailable, a homogeneous structure may be adopted even if

Copula	δ_0	$\alpha_{0,1}$	$\alpha_{1,1}$	$\beta_{0,1}$	$\beta_{1,1}$	Bias^2	MSE	Variance
0.5	0.3556	-0.0323	0.0050	0.0008	-0.0025	0.0255	0.3959	0.3704
1.0	0.3855	-0.0378	0.0115	-0.0001	-0.0041	0.0300	0.5210	0.4910
1.5	0.4317	-0.0404	0.0084	-0.0004	-0.0031	0.0376	0.5500	0.5124
2.0	0.4919	-0.0415	0.0092	-0.0011	-0.0055	0.0488	0.6057	0.5570
2.5	0.4848	-0.0436	0.0125	-0.0010	-0.0062	0.0474	0.6326	0.5852
3.0	0.4455	-0.0459	0.0132	-0.0003	-0.0037	0.0402	0.6614	0.6213
Indep.	0.1218	-0.0176	0.0070	0.0007	0.0001	0.0030	0.0652	0.0622

Table A-6: Bias, MSE, and variance of the QMLE in the linear model under the Clayton copula for different copula parameters at $T = 250$.



(a) linear



(b) log-linear

Figure A-8: MAE in the data for different copula specifications at $T = 250$ observation times.

the true intercept is inhomogeneous. In this study, we examine the consequences of this misspecification.

Model	T	inhomogeneous				homogeneous			
		$\alpha_{0,1}$	$\alpha_{1,1}$	$\beta_{0,1}$	$\beta_{1,1}$	$\alpha_{0,1}$	$\alpha_{1,1}$	$\beta_{0,1}$	$\beta_{1,1}$
linear	50	-0.140	0.056	-0.033	0.001	0.240	0.030	-0.006	0.029
	100	-0.108	0.048	-0.013	0.001	0.306	-0.001	-0.012	0.022
	250	-0.058	0.030	-0.004	-0.002	0.368	-0.026	-0.024	0.014
	500	-0.029	0.012	-0.002	-0.001	0.388	-0.042	-0.027	0.017
log-linear	50	-0.174	-0.075	-0.043	-0.005	0.297	0.000	-0.012	0.049
	100	-0.118	-0.018	-0.018	0.001	0.385	-0.024	-0.024	0.032
	250	-0.057	-0.002	-0.006	0.002	0.456	-0.046	-0.040	0.021
	500	-0.028	0.004	-0.003	0.000	0.493	-0.061	-0.051	0.014

Table A-7: Bias of non-intercept parameter estimates under inhomogeneous and homogeneous intercept specifications.

Table A-7 reports the bias of the non-intercept parameter estimates for both model specifications. While the bias of the correctly specified model converges to zero as T increases, the homogeneous model exhibits a persistent bias. For $\beta_{0,1}$ and $\beta_{1,1}$, this bias is relatively small: $\beta_{0,1}$ is slightly underestimated on average (-0.024 at $T = 250$ in the linear model), while $\beta_{1,1}$ tends to be slightly overestimated (0.014 at $T = 250$). In contrast, $\alpha_{0,1}$ exhibits a pronounced and growing bias, reaching 0.306 at $T = 100$ and 0.388 at $T = 500$ in the linear model. The bias for $\alpha_{1,1}$ remains small and negative throughout.

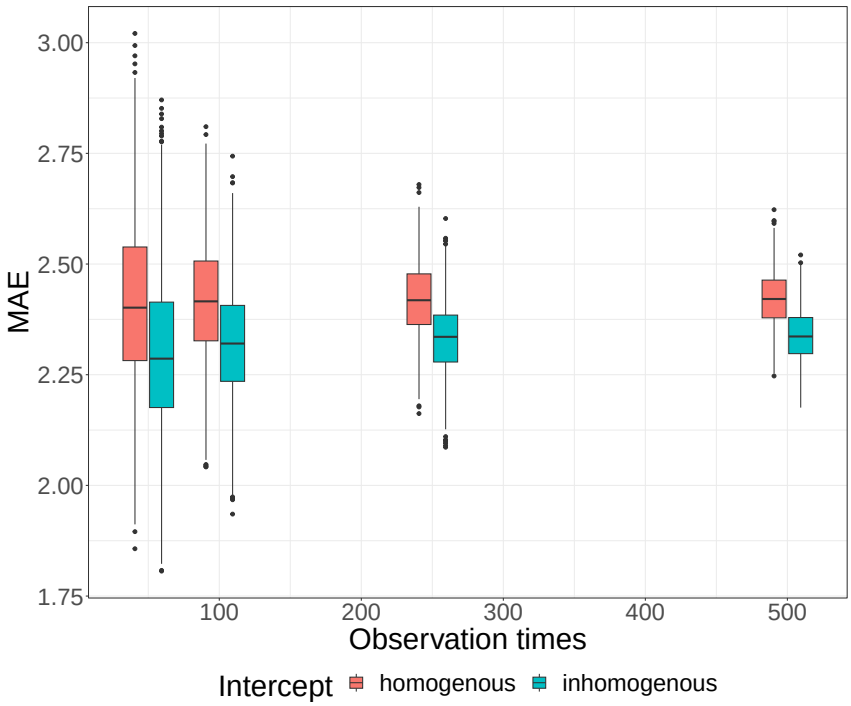
These results suggest that the model uses the feedback process $\{\boldsymbol{\mu}_t\}$ to compensate for the intercept misspecification, while the spatial dependency structure is still captured reasonably well.

The overall bias is also reflected in the in-sample MAE, shown in Figures A-9a and A-9b. For small T, the boxplots of the homogeneous and inhomogeneous models overlap considerably, indicating that the larger parameter count of the inhomogeneous model partially offsets the advantage of correct specification.

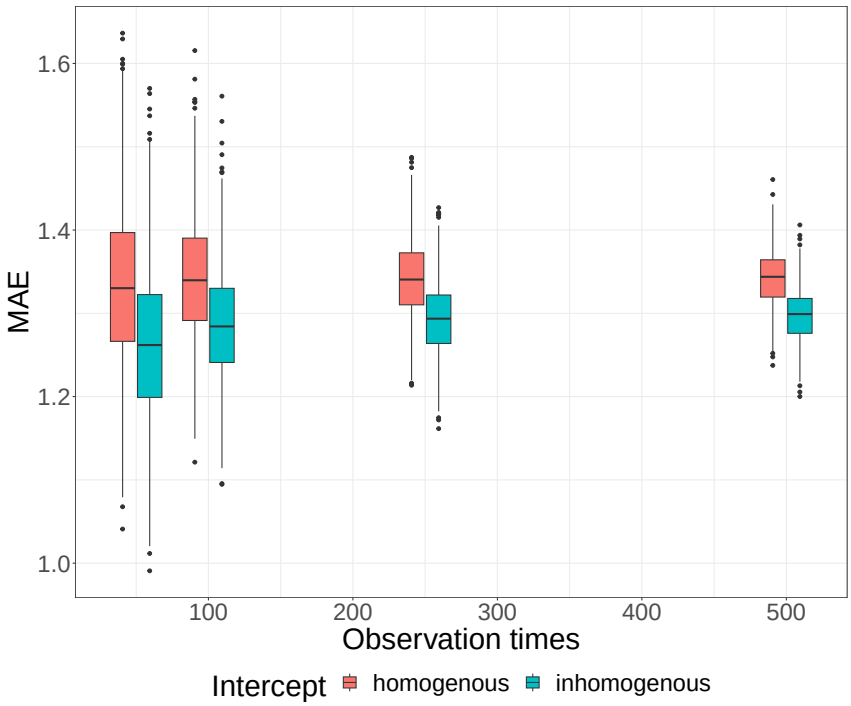
Link

The preceding simulations assumed a correctly specified model type. We now examine the consequences of selecting the wrong model class, i.e., fitting a linear model when the true model is log-linear and vice versa.

Table A-8 reports the proportion of replications in which the QIC correctly selects the true model. The correct model is identified in the majority of cases. When the true log-linear model has a negative parameter, the selection frequency increases markedly, as the parameter constraints of the linear model prevent it from adequately capturing the data. This is confirmed in Table A-9, which shows that the linear model consistently



(a) linear



(b) log-linear

Figure A-9: MAE in the data for inhomogeneous and homogeneous intercept specifications.

True Model	Parameters	T			
		50	100	250	500
log-linear	all positive	0.707	0.660	0.736	0.780
	$\alpha_{0,1} < 0$	0.743	0.733	0.798	0.884
linear	all positive	0.368	0.513	0.590	0.701

Table A-8: Relative frequencies with which the QIC selects the log-linear over the linear model.

estimates the negative $\alpha_{0,1}$ at values close to zero. Overall, the mean QIC values are similar across all settings, indicating that even misspecified models provide an adequate fit to the data. In practice, log-linear models should be considered whenever parameter estimates of the linear model are at or near their lower bound.

True Model	Parameters	T	log-linear					linear				
			$\alpha_{0,1}$	$\alpha_{1,1}$	$\beta_{0,1}$	$\beta_{1,1}$	QIC	$\alpha_{0,1}$	$\alpha_{1,1}$	$\beta_{0,1}$	$\beta_{1,1}$	QIC
linear	all positive	50	0.114	0.006	0.193	0.081	21324.10	0.099	0.127	0.179	0.084	21348.15
		100	0.152	0.038	0.203	0.092	43031.34	0.123	0.116	0.194	0.091	43036.43
		250	0.182	0.083	0.206	0.094	107952.99	0.156	0.115	0.199	0.094	107924.85
		500	0.192	0.095	0.207	0.096	216235.15	0.177	0.106	0.200	0.097	216168.35
log-linear	$\alpha_{0,1} = -0.2$	50	-0.115	-0.045	0.185	0.082	14314.99	0.024	0.116	0.117	0.069	14374.42
		100	-0.143	-0.022	0.191	0.090	28805.66	0.008	0.095	0.130	0.072	28868.96
		250	-0.170	0.031	0.197	0.096	72318.83	0.001	0.062	0.138	0.077	72415.19
		500	-0.182	0.070	0.199	0.098	144791.65	0.000	0.035	0.139	0.080	144945.90
	all positive	50	0.098	-0.036	0.185	0.082	17516.16	0.093	0.116	0.152	0.085	17558.87
		100	0.136	0.000	0.195	0.097	35289.53	0.108	0.120	0.168	0.094	35326.11
		250	0.177	0.072	0.199	0.099	88537.88	0.144	0.113	0.173	0.099	88605.01
		500	0.190	0.090	0.199	0.098	177117.87	0.168	0.108	0.173	0.099	177239.26

Table A-9: Mean parameter estimates and mean QIC of linear and log-linear models in the link misspecification settings.

A.3.2 Repetition of Simulations with a Different Copula

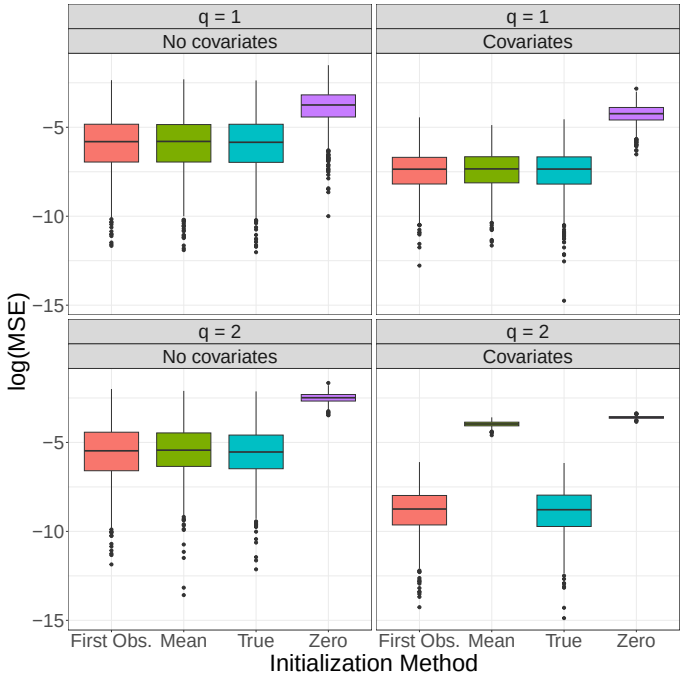
In Sections 7.1 and A.3.1, the Clayton copula with parameter 2 was used throughout to generate contemporaneous dependencies. To validate these results, we repeated the simulations described in Table 7.1 using the Frank copula with parameter 1 in the data-generating process. The two copulas generate different, non-directly-comparable dependency structures; the Frank copula used here tends to induce weaker contemporaneous dependencies between components.

The results from Section 7.1 remain broadly consistent, with two notable differences in the Frank copula setting: faster convergence to the asymptotic distribution and higher test power, both attributable to the weaker contemporaneous dependencies that are better accommodated by the quasi-likelihood (3.7). These findings are in line with the results from Section A.3.1, which showed that estimation variability increases with stronger contemporaneous dependencies.

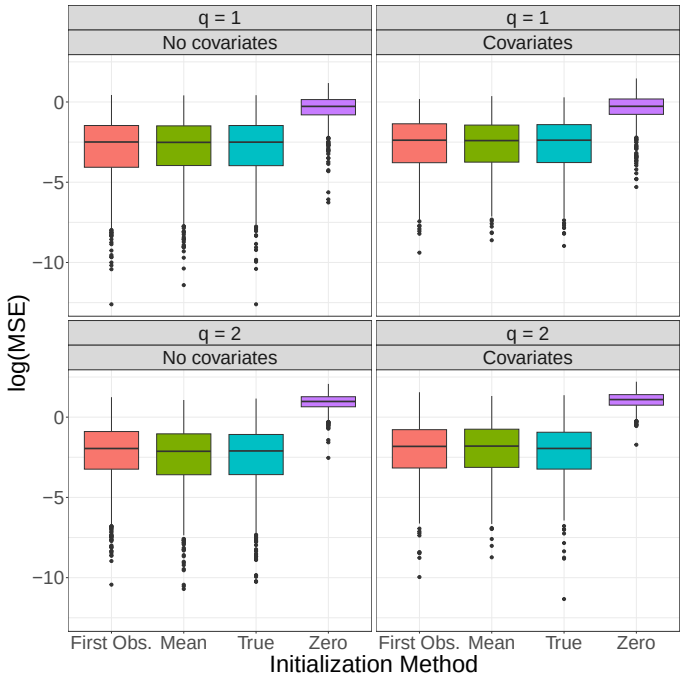
Figures A-10a and A-10b illustrate the behavior of the MSEs under different initializations, exhibiting trends similar to those in Figures 7.1a and 7.1b, though with slightly lower MSEs owing to the weaker dependence induced by the Frank copula.

Table A-10 reports empirical sizes analogous to those in Table A-2. Convergence to the 5% significance level is faster, often achieved at $T = 50$ across most scenarios. Power curves in Figures A-11a and A-11b also show consistently higher power, in line with Figures A-5b and A-5a.

Under spatial dependency misspecification, the correct model is identified via QIC with fewer observations than in the Clayton copula setting, as shown in Table A-11. Accordingly, the power for detecting the anisotropic neighborhood structure is enhanced (Table A-12), and the isotropic model also exhibits improved detection of spatial dependence (Table A-13).



(a) log-linear

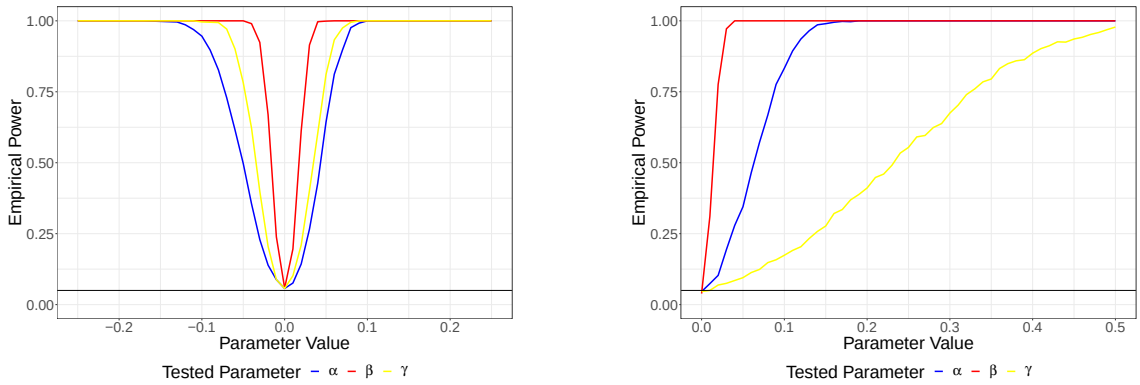


(b) linear

Figure A-10: (Log) MSE of the QMLE under different initialization methods for the feedback process in the Frank copula setting.

Parameter	Model	Gridsize	T				
			50	100	250	500	750
$\gamma_{0,1}$	log-linear	5×5	0.070	0.047	0.053	0.059	0.051
		7×7	0.068	0.061	0.052	0.042	0.036
		9×9	0.061	0.060	0.057	0.039	0.052
	linear	5×5	0.067	0.054	0.056	0.042	0.047
		7×7	0.062	0.054	0.053	0.057	0.059
		9×9	0.060	0.066	0.050	0.046	0.046
$\alpha_{0,1}$	log-linear	5×5	0.095	0.072	0.054	0.043	0.035
		7×7	0.077	0.060	0.050	0.051	0.056
		9×9	0.075	0.042	0.057	0.051	0.048
	linear	5×5	0.049	0.043	0.044	0.046	0.042
		7×7	0.059	0.066	0.052	0.044	0.056
		9×9	0.064	0.047	0.046	0.038	0.047
$\beta_{0,1}$	log-linear	5×5	0.047	0.054	0.044	0.072	0.068
		7×7	0.054	0.060	0.051	0.077	0.072
		9×9	0.058	0.061	0.051	0.062	0.072
	linear	5×5	0.049	0.049	0.054	0.038	0.054
		7×7	0.053	0.049	0.050	0.045	0.042
		9×9	0.039	0.038	0.042	0.043	0.044

Table A-10: Empirical size of the asymptotic Wald test for different parameters and grid sizes in the Frank copula setting.



(a) log-linear

(b) linear

Figure A-11: Power curves of the PSTARMAX process on a 9×9 grid with 250 observation times in the Frank copula setting.

Model	T			
	50	100	250	500
log-linear	0.299	0.093	0.002	0.000
linear	0.271	0.082	0.000	0.000

Table A-11: Relative frequencies with which the QIC favors the isotropic over the true anisotropic model in the Frank copula simulations.

Model	Anisotropy-Test	T			
		50	100	250	500
log-linear	β	0.410	0.672	0.965	1.000
	α	0.109	0.121	0.141	0.222
linear	β	0.454	0.704	0.975	1.000
	α	0.066	0.087	0.150	0.256

Table A-12: Empirical power of the Wald test at the 5% significance level for testing coefficient differences in an anisotropic model with separate neighborhood matrices for each direction, in the Frank copula simulations. α refers to the feedback process coefficients and β to the observation process coefficients.

Model	Coefficient	T			
		50	100	250	500
log-linear	$\beta_{1,1}$	0.858	0.995	1.000	1.000
	$\alpha_{1,1}$	0.164	0.205	0.380	0.593
linear	$\beta_{1,1}$	0.923	0.993	1.000	1.000
	$\alpha_{1,1}$	0.198	0.313	0.515	0.750

Table A-13: Relative frequencies of significant coefficients ($\beta_{1,1}$ and $\alpha_{1,1}$, $p < 0.05$) in an isotropic model fitted to data from an anisotropic model, in the Frank copula simulations.

A.3.3 Figures for the Detection of Persistent Level Shifts

Recall from Section 7.2:

We examine the behavior of the procedure introduced in Chapter 4 under several scenarios involving zero, one, or two persistent level shifts. For each scenario, we consider two distinct settings: one in which only the time point of the level shift is unknown while the affected locations are assumed to be known to be global, and one in which both the time points and the affected locations are unknown, though the true location vectors are included among the candidates. Since bootstrapping did not yield a satisfactory approximation of the null distribution, all results reported below are size-adjusted based on the empirical quantiles obtained under the null hypothesis.

Data are generated on the grid described above from the following log-linear models. Each of the 1000 simulation iterations produces $T = 500$ time points. Contemporaneous dependence is introduced via the data-generating process described in Section 2.1.1, using a Frank copula with parameter 2.

$$\begin{aligned}\boldsymbol{\psi}_t = 0.5\mathbf{1}_p + (0.3\mathbf{I}_p + 0.2\mathbf{W}^{(1)})\log(\mathbf{Y}_{t-1} + \mathbf{1}_p) + 0.75\mathbf{X}_{1,t} + 0.5\mathbf{X}_{2,t} \\ + \sum_{k=1}^{2} \gamma_k \mathbf{1}(t \geq \tau_k)\,\mathbf{v}_k\end{aligned} \tag{A-12}$$

$$\begin{aligned}\boldsymbol{\psi}_t = 0.5\mathbf{1}_p + (0.1\mathbf{I}_p + 0.05\mathbf{W}^{(1)})\boldsymbol{\psi}_{t-1} + (0.3\mathbf{I}_p + 0.2\mathbf{W}^{(1)})\log(\mathbf{Y}_{t-1} + \mathbf{1}_p) \\ + 0.75\mathbf{X}_{1,t} + 0.5\mathbf{X}_{2,t} + \sum_{k=1}^{2} \gamma_k \mathbf{1}(t \geq \tau_k)\,\mathbf{v}_k\end{aligned} \tag{A-13}$$

The covariate processes are constructed as follows. The first covariate, $\mathbf{X}_{1,t} = \sin\left(\frac{2\pi}{12}t\right)\mathbf{1}_p$, induces a seasonal pattern. The second covariate is a fixed spatial vector $\mathbf{X}_{2,t} = \mathbf{X} \in \mathbb{R}^p$ with $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.9)$, drawn once and held constant across all iterations and models.

We consider three shift scenarios: (1) no level shift, i.e. $\gamma_1 = \gamma_2 = 0$; (2) a single level shift with $\gamma_1 = 0.3$ and $\gamma_2 = 0$; and (3) two level shifts with $\gamma_1 = 0.3$ and $\gamma_2 = -0.2$. In all cases, the time points of the shifts are fixed at $\tau_1 = 100$ and $\tau_2 = 200$.

Regarding the spatial structure, we distinguish two cases. In the *global* case, $\mathbf{v}_1 = \mathbf{v}_2 = \mathbf{1}_{100}$, so each shift affects all locations simultaneously. In the *local* case, $\mathbf{v}_1 = (\mathbf{1}'_{50}, \mathbf{0}'_{50})'$ and $\mathbf{v}_2 = (\mathbf{0}'_{50}, \mathbf{1}'_{50})'$, so each shift affects only half of the locations.

Following the procedure described in Chapter 4, each model is estimated on the simulated data under the null hypothesis, i.e. without level-shift regressors. The temporal search grid is set to $\{10, \ldots, 490\}$. In the global level-shift scenario, only the time points of the shifts is searched, treating the affected locations as known. In the local level-shift scenario, the two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are included as spatial candidates. The procedure described in

Section 4.2 to estimate the time-points and locations of the shifts is run for exactly four iterations in all cases even when the procedure would terminate earlier.

Figure A-12 shows the distribution of the estimated time points of the *global* level shifts in case of model (A-13). The results for model (A-12) were already shown in Figure 7.2 in the main part.

Figure A-13 shows the distribution of the estimated time points of the *local* level shifts in case of model (A-13). The results for model (A-12) were already shown in Figure 7.3 in the main part.

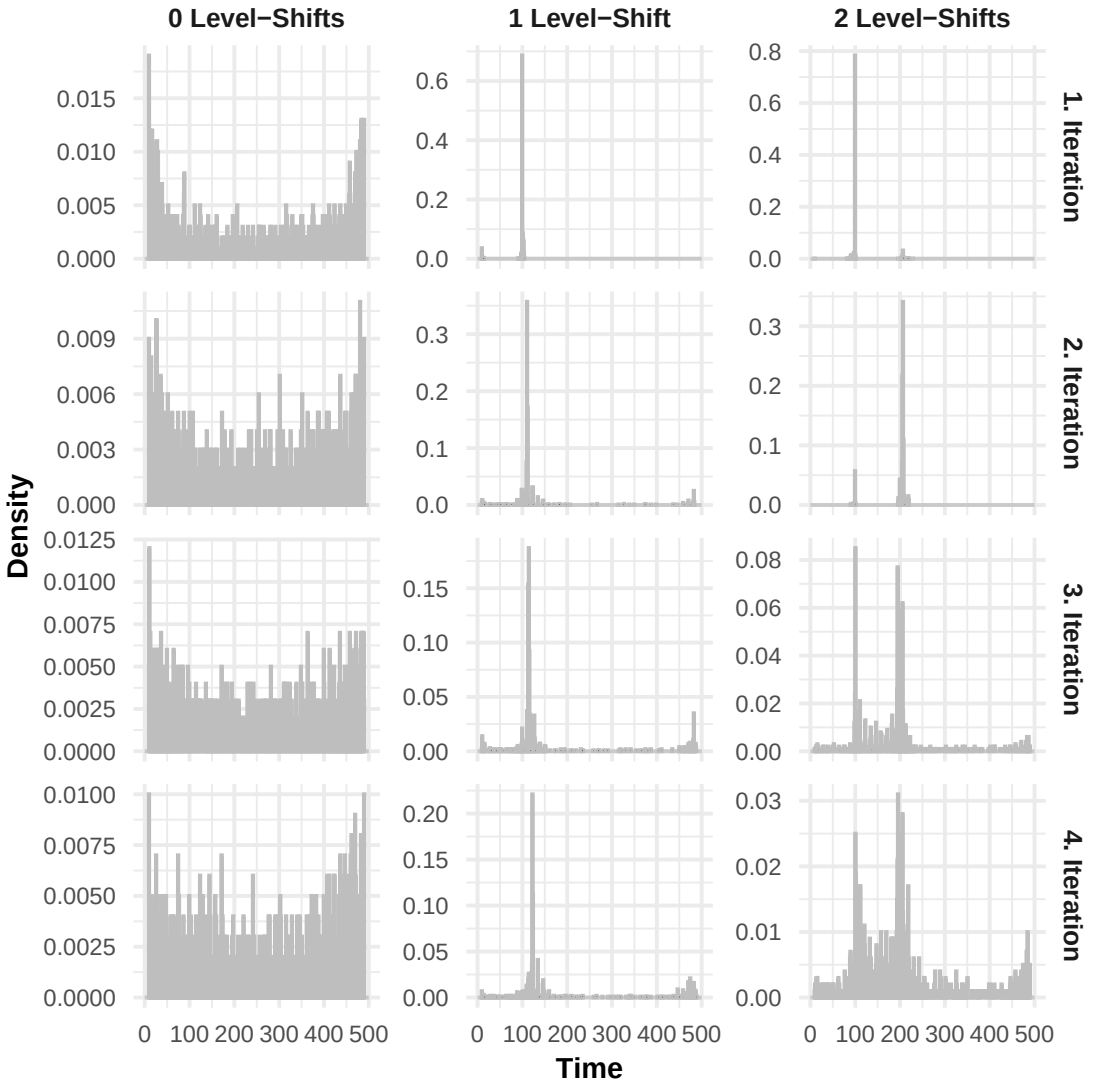


Figure A-12: Distribution of estimated change points after each of four iterations, for zero, one, or two level shifts in a setting with global level shifts having an effect on all locations. True model is with feedback process.

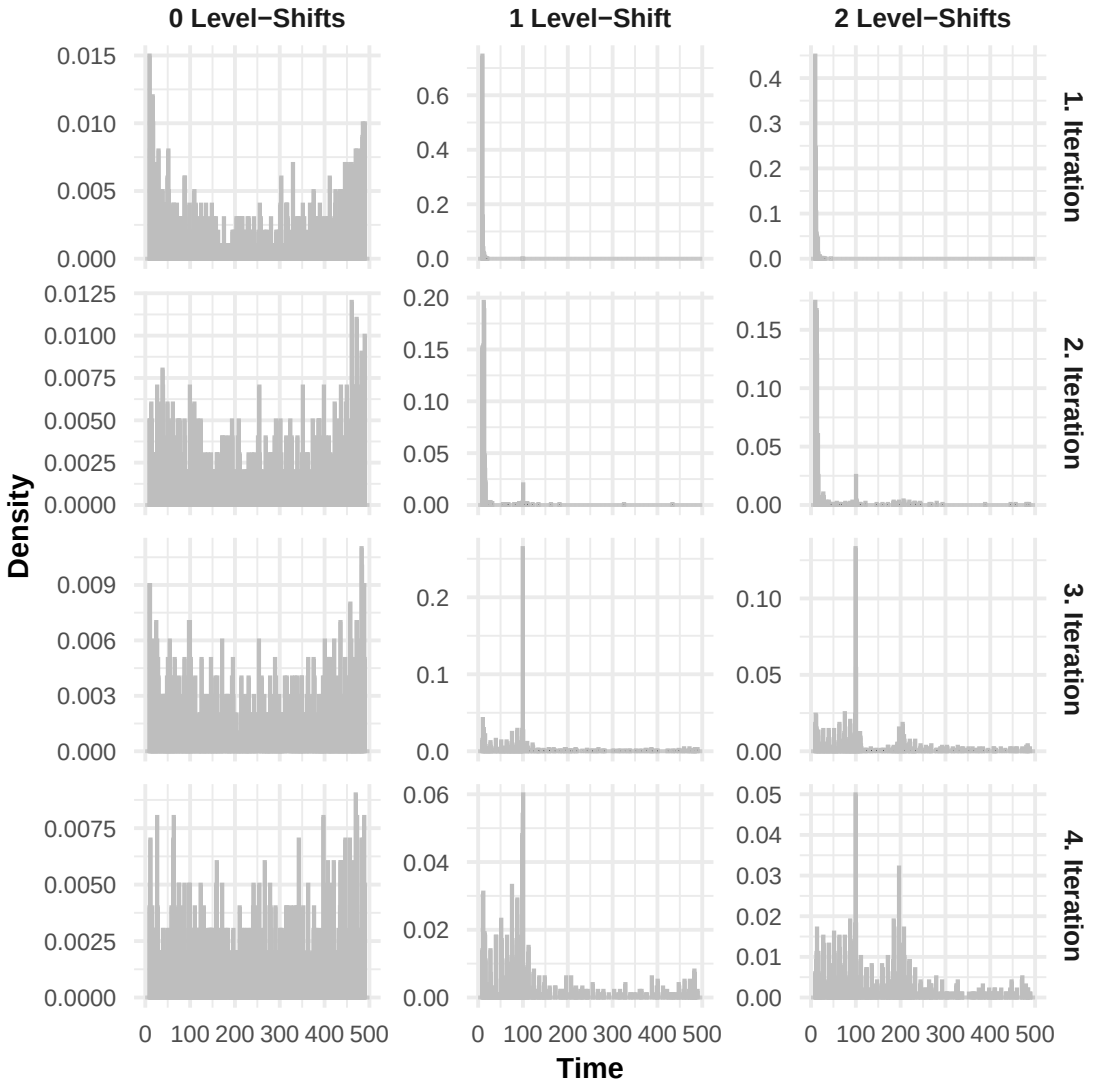


Figure A-13: Distribution of estimated change points after each of four iterations, for zero, one, or two level shifts in a setting with local level shifts having an effect only on half of the locations. True model is with feedback process.

A.3.4 Figures for Spatio-temporal DGLMs

Recall from Section 7.3:

For the conditional mean, we consider two alternative dynamic specifications:

$$\boldsymbol{\psi}_t = \delta_0 \cdot \mathbf{1}_{100} + (0.2\mathbf{I}_{100} + 0.1\mathbf{W}^{(1)})\,\boldsymbol{\psi}_{t-1} + (0.2\mathbf{I}_{100} + 0.1\mathbf{W}^{(1)})\,h(\mathbf{Y}_{t-1}) + \gamma\,\mathbf{X}_t, \quad \text{(A-14)}$$

$$\boldsymbol{\psi}_t = \delta_0 \cdot \mathbf{1}_{100} + (0.4\mathbf{I}_{100} + 0.2\mathbf{W}^{(1)})\,h(\mathbf{Y}_{t-1}) + \gamma\,\mathbf{X}_t, \quad \text{(A-15)}$$

where for count data we set $\delta_0 = 0.6$, $\gamma = 0.9$, and $h(\mathbf{Y}_{t-1}) = \log(\mathbf{Y}_{t-1} + \mathbf{1}_{100})$, while for normally distributed data we set $\delta_0 = 5$, $\gamma = 2$, and $h(\mathbf{Y}_{t-1}) = \mathbf{Y}_{t-1}$.

The covariate process $\{\mathbf{X}_t\}$ was generated once and reused across all scenarios using the following R code:

```
set.seed(42)
covariates <- t(replicate(100,
    arima.sim(n = 1000,
              list(ar = c(0.89, -0.3), ma = c(-0.1, 0.28)),
              rand.gen = runif)))
covariates <- covariates / max(covariates)
covariates[covariates < 0] <- 0
```

For the dispersion process, we consider three alternative specifications. The dispersion parameter vector $\boldsymbol{\phi}_t$ is linked to a linear process $\boldsymbol{\zeta}_t$ via the log link (i.e. $\boldsymbol{\phi}_t = \exp(\boldsymbol{\zeta}_t)$):

$$\boldsymbol{\phi}_t = 2 \cdot \mathbf{1}_{100}, \quad \text{(A-16)}$$

$$\boldsymbol{\zeta}_t = 0.5 \cdot \mathbf{1}_{100} + (0.15\mathbf{I} + 0.05\mathbf{W}^{(1)})\,\boldsymbol{\zeta}_{t-1} + (0.3\mathbf{I} + 0.2\mathbf{W}^{(1)})\,\log(\mathbf{d}_{t-1}), \quad \text{(A-17)}$$

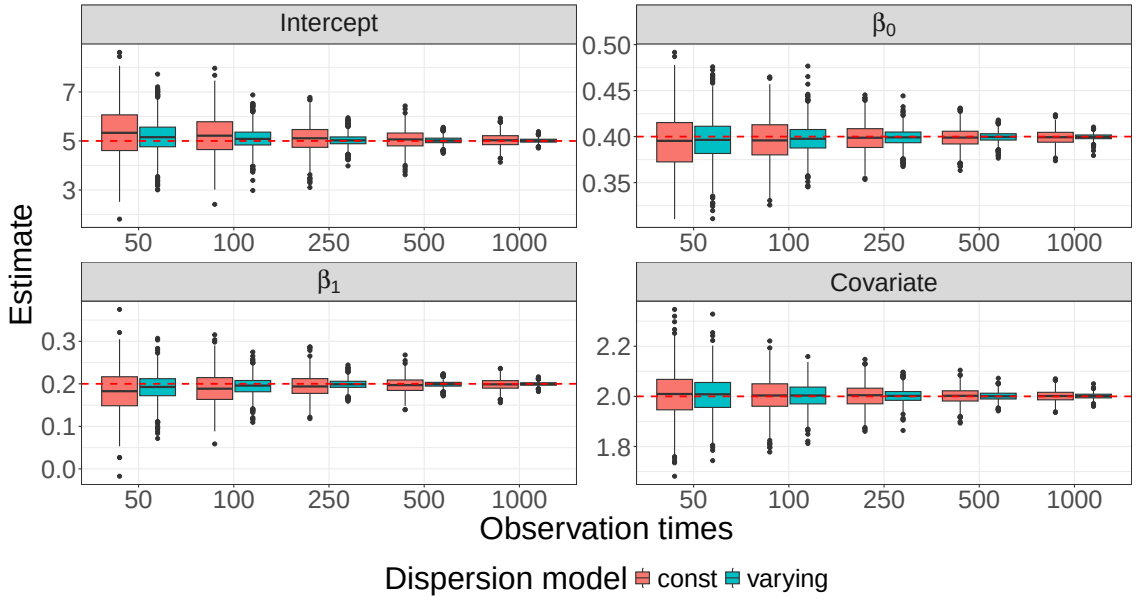
$$\boldsymbol{\zeta}_t = 0.5 \cdot \mathbf{1}_{100} + (0.5\mathbf{I} + 0.2\mathbf{W}^{(1)})\,\log(\mathbf{d}_{t-1}). \quad \text{(A-18)}$$

Each combination of mean specification (7.6)–(7.7) with dispersion specification (7.8)–(7.10) defines one data-generating scenario. Table A-14 summarizes all configurations and the corresponding fitted models. When data are generated under a time-varying dispersion process, we estimate both (i) the correctly specified model with time-varying dispersion and (ii) a misspecified model assuming a single, time-constant global dispersion parameter, thereby examining the robustness of estimation and the consequences of misspecification. When the true dispersion is constant (7.8), only the time-varying model is fitted in order to study the asymptotic behaviour of the overparameterized model under the null hypothesis of no temporal dispersion dynamics.

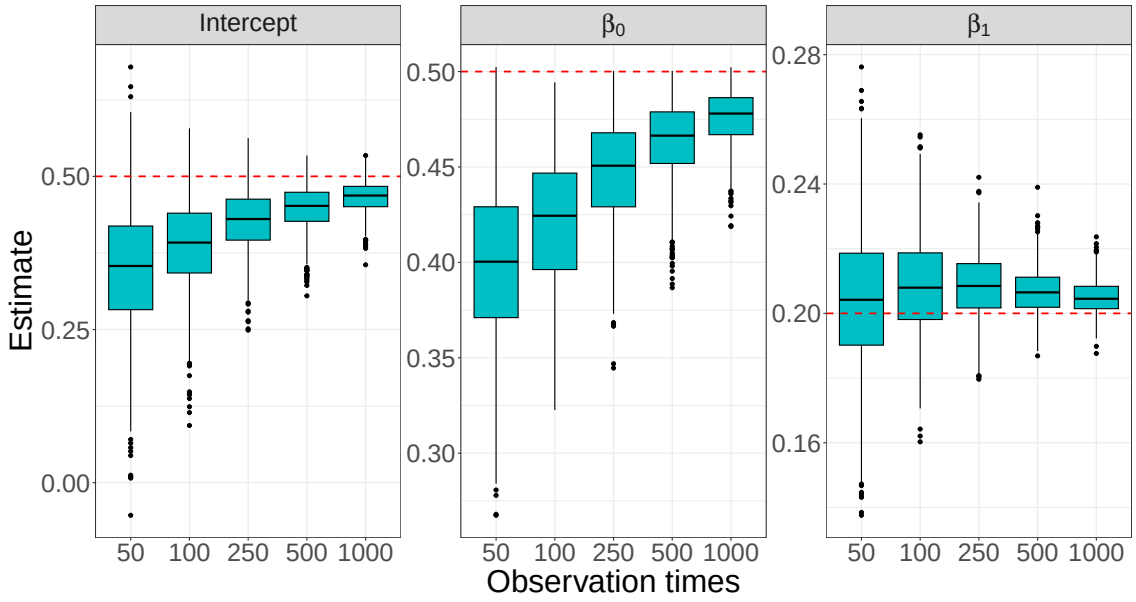
Each scenario is simulated for $T \in \{50, 100, 250, 500, 1000\}$ over 1000 independent replications. For models including a feedback mechanism, i.e. (7.6) and (7.9), the feedback process is initialized using the first observations or pseudo-observations.

True Model	Fitted Models	Normal		Generalized Poisson	
		Boxplots	QQ-Plots	Boxplots	QQ-Plots
(A-14) + (A-16)	(A-14) + (A-18)	-	-	-	-
(A-15) + (A-16)	(A-15) + (A-18)	-	-	-	-
(A-14) + (A-17)	(A-14) + (A-16) & (A-14) + (A-17)	Fig. A-18	Fig. A-19	Fig. A-26	Fig. A-27
(A-14) + (A-18)	(A-14) + (A-16) & (A-14) + (A-18)	Fig. A-16	Fig. A-17	Fig. A-22	Fig. A-23
(A-15) + (A-17)	(A-15) + (A-16) & (A-15) + (A-17)	Fig. 7.4	Fig. 7.5	Fig. A-24	Fig. A-25
(A-15) + (A-18)	(A-15) + (A-16) & (A-15) + (A-18)	Fig. A-14	Fig. A-15	Fig. A-20	Fig. A-21

Table A-14: Overview of data-generating and fitted models, with references to the corresponding simulation figures.

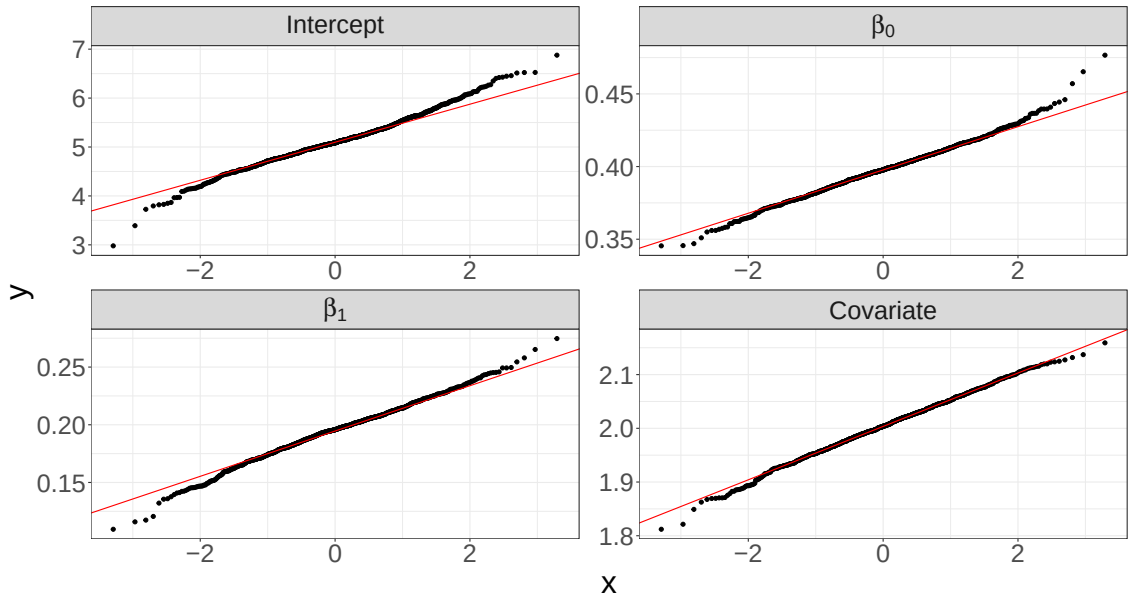


(a) Parameter estimates of the mean model (7.7) under the constant dispersion specification (7.8) and the varying dispersion specification (7.10).

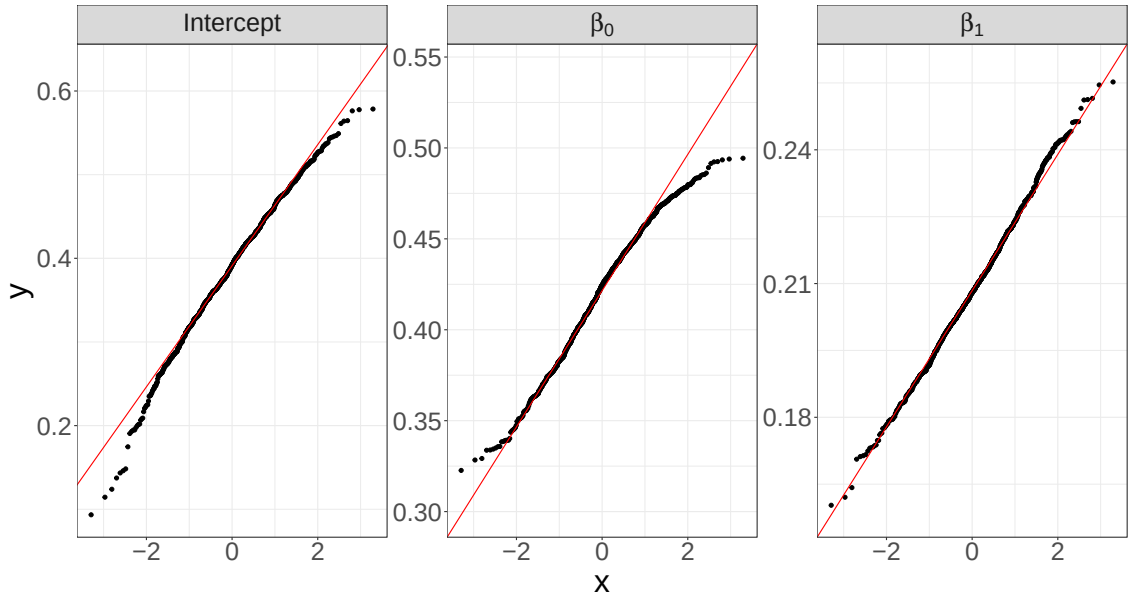


(b) Parameter estimates of the varying dispersion model (7.10) combined with the mean model (7.7).

Figure A-14: Box plots of parameter estimates under alternative dispersion specifications. Panel (a) shows estimates of the mean model under constant and varying dispersion, whereas panel (b) displays estimates of the varying dispersion model. The data-generating process with marginal normal distributions corresponds to (7.7) and (7.10), i.e. both models without feedback mechanism.

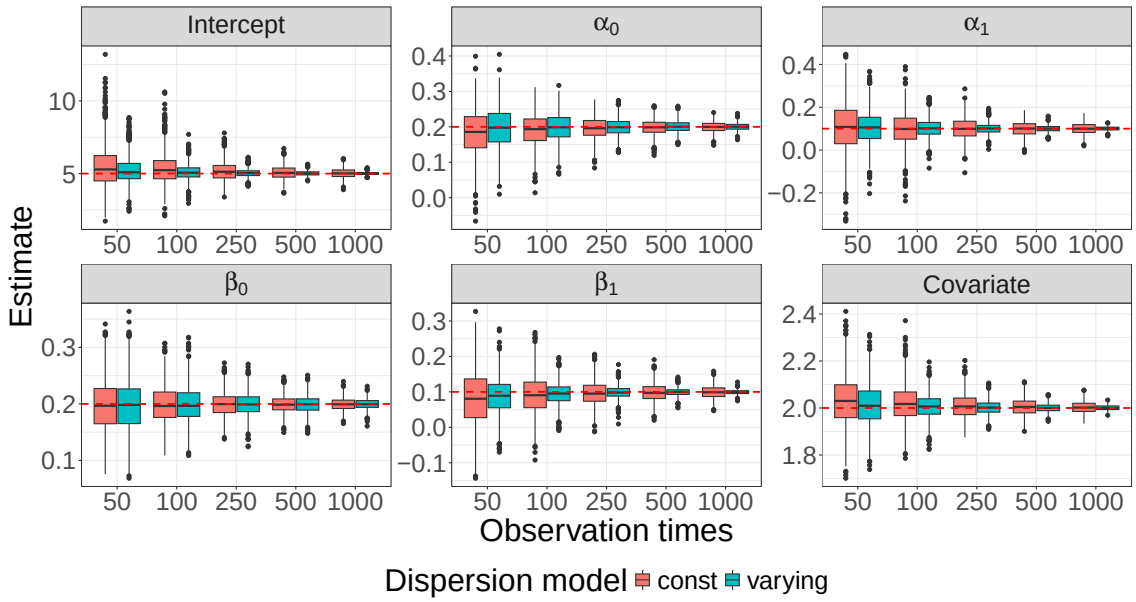


(a) Mean model (7.7), i.e. without feedback mechanism.

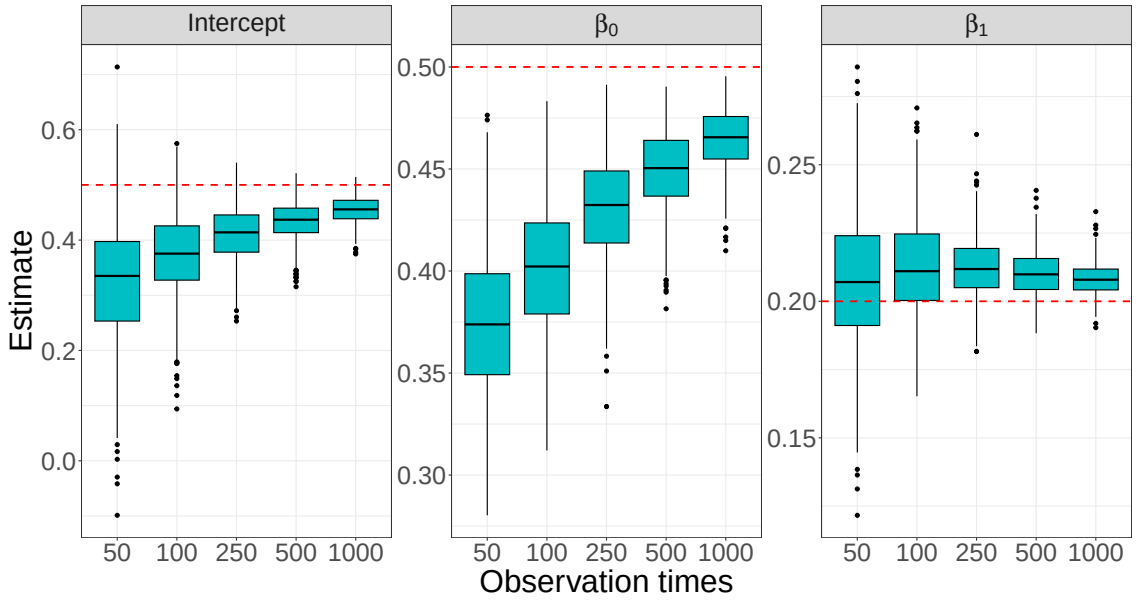


(b) Dispersion model (7.10), i.e. without feedback mechanism.

Figure A-15: Normal Q–Q plots of parameter estimates under the true data-generating process (7.7) and (7.10) with marginal normal distributions, based on $T = 100$.

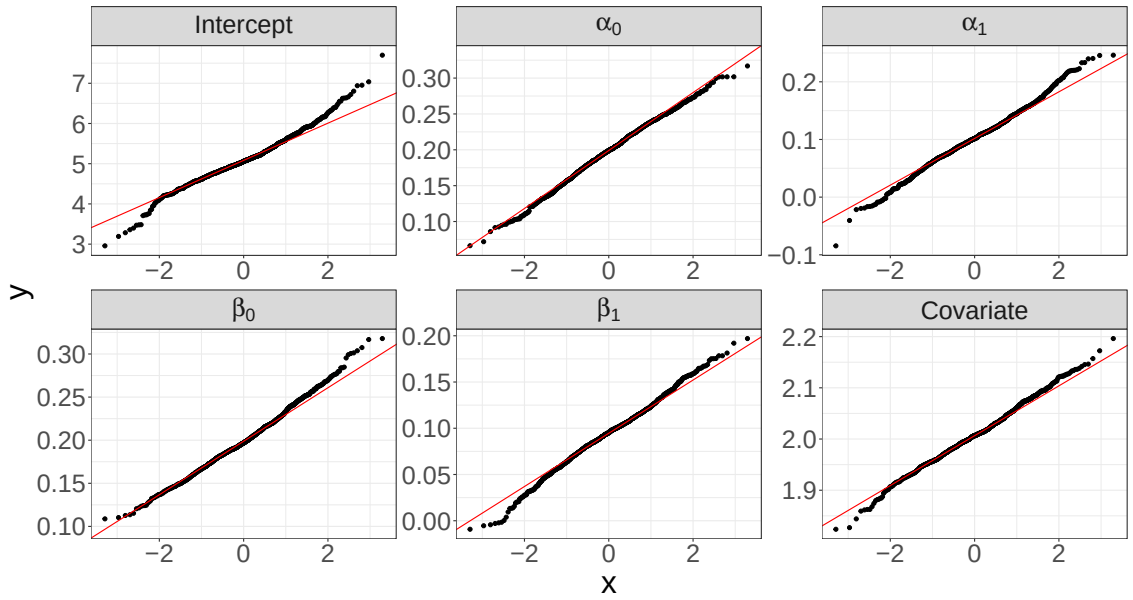


(a) Parameter estimates of the mean model (7.6) under the constant dispersion specification (7.8) and the varying dispersion specification (7.10).

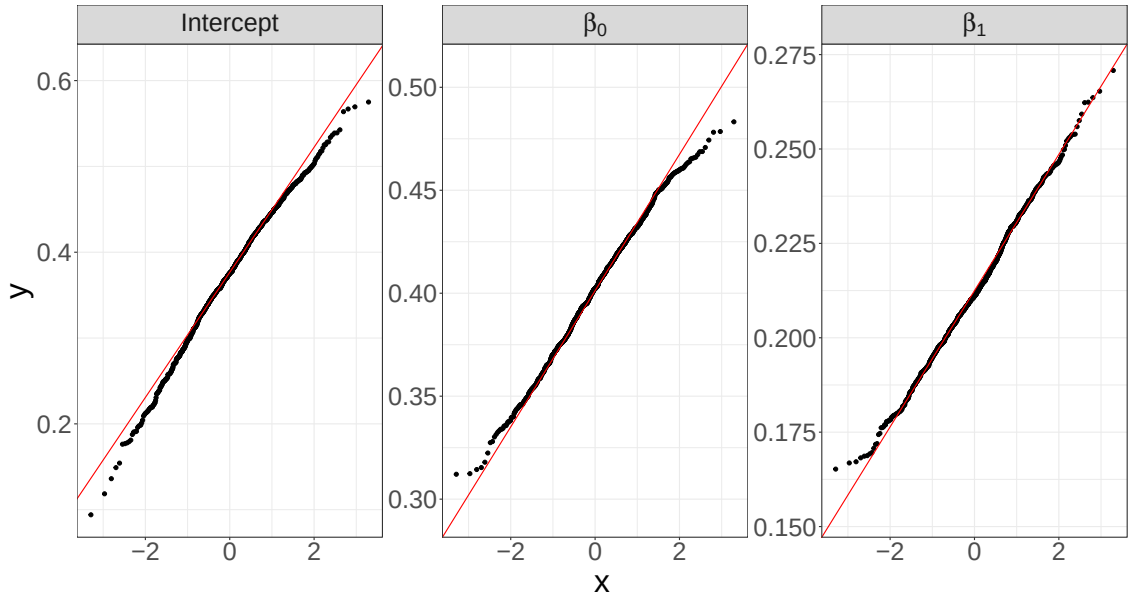


(b) Parameter estimates of the varying dispersion model (7.10) combined with the mean model (7.6).

Figure A-16: Box plots of parameter estimates under alternative dispersion specifications. Panel (a) shows estimates of the mean model under constant and varying dispersion, whereas panel (b) displays estimates of the varying dispersion model. The data-generating process with marginal normal distributions corresponds to (7.6) and (7.10), i.e. the mean model includes a feedback mechanism and the dispersion process not.

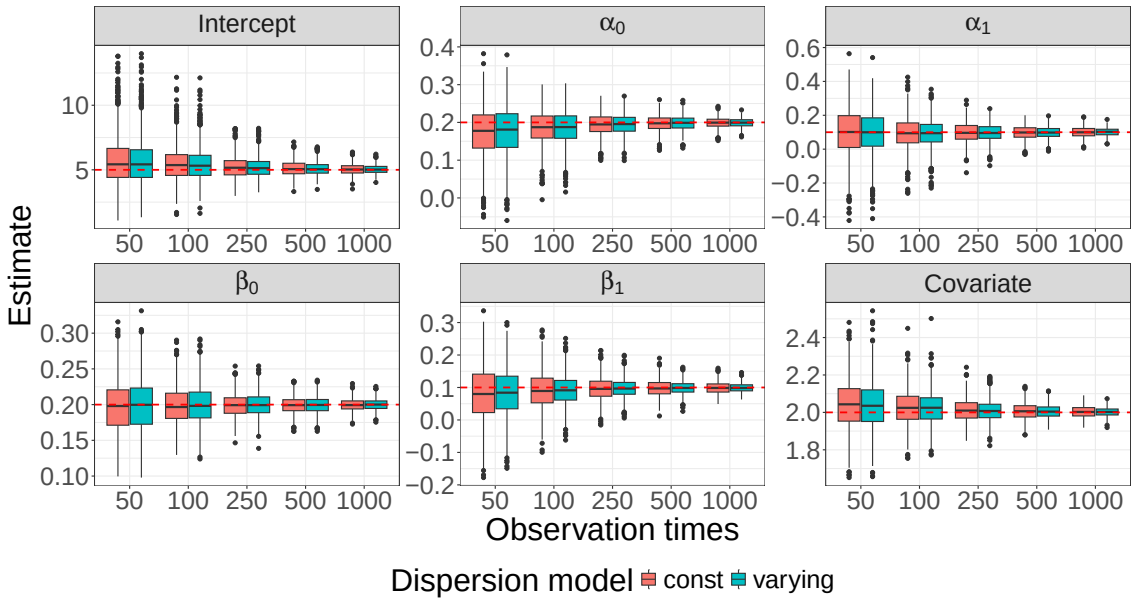


(a) Mean model (7.6), i.e. with feedback mechanism.

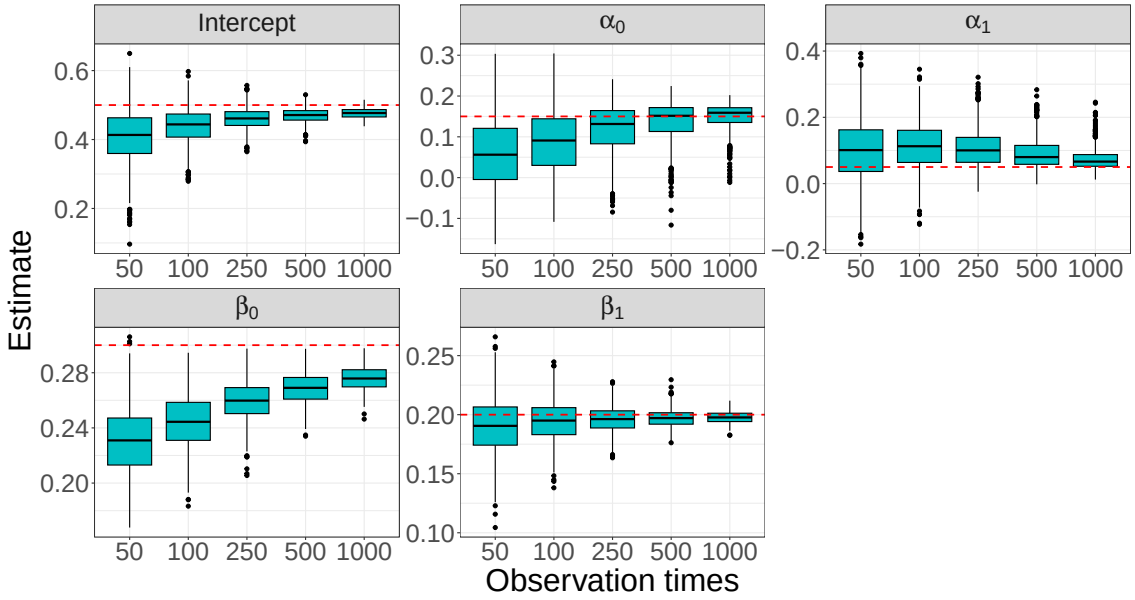


(b) Dispersion model (7.10), i.e. without feedback mechanism.

Figure A-17: Normal Q–Q plots of parameter estimates under the true data-generating process (7.6) and (7.10) with marginal normal distributions, based on $T = 100$.

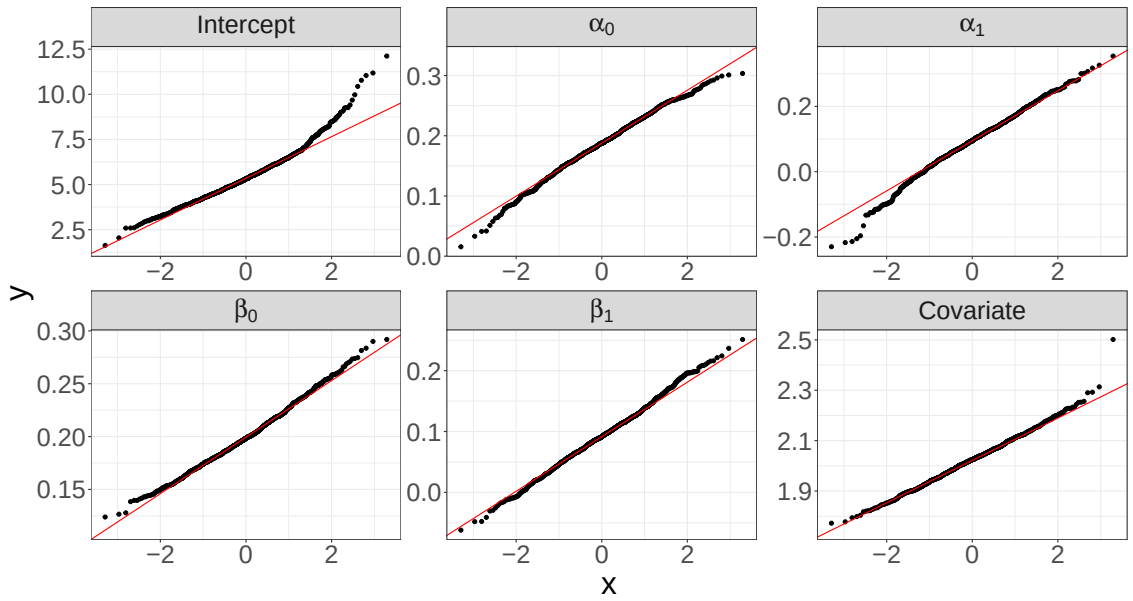


(a) Parameter estimates of the mean model (7.6) under the constant dispersion specification (7.8) and the varying dispersion specification (7.9).

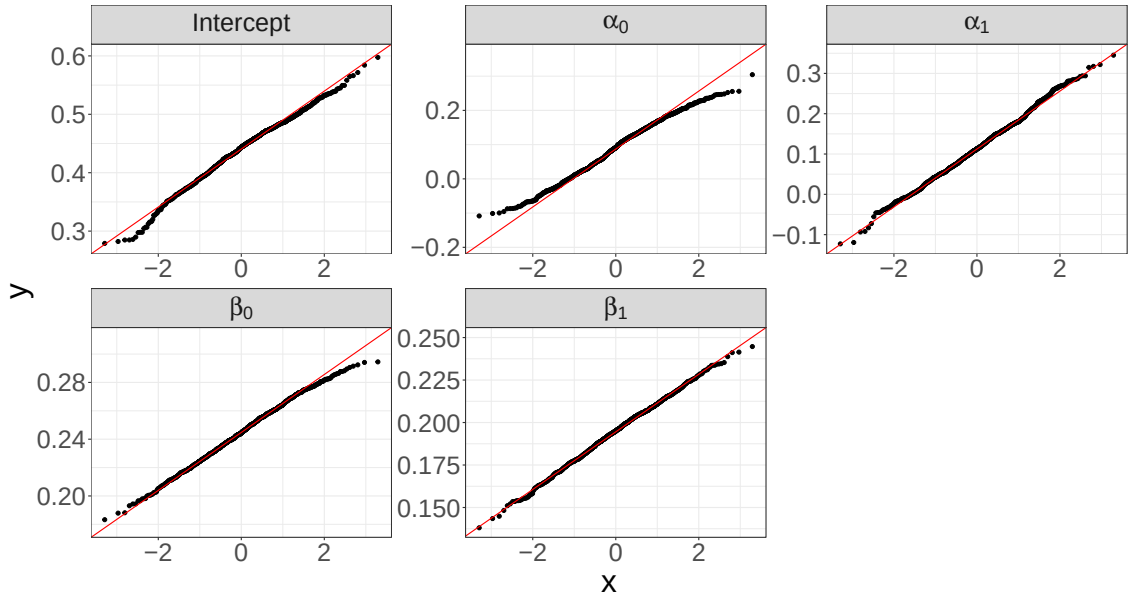


(b) Parameter estimates of the varying dispersion model (7.9) combined with the mean model (7.6).

Figure A-18: Box plots of parameter estimates under alternative dispersion specifications. Panel (a) shows estimates of the mean model under constant and varying dispersion, whereas panel (b) displays estimates of the varying dispersion model. The data-generating process with marginal normal distributions corresponds to (7.7) and (7.9), i.e. both models include a feedback mechanism

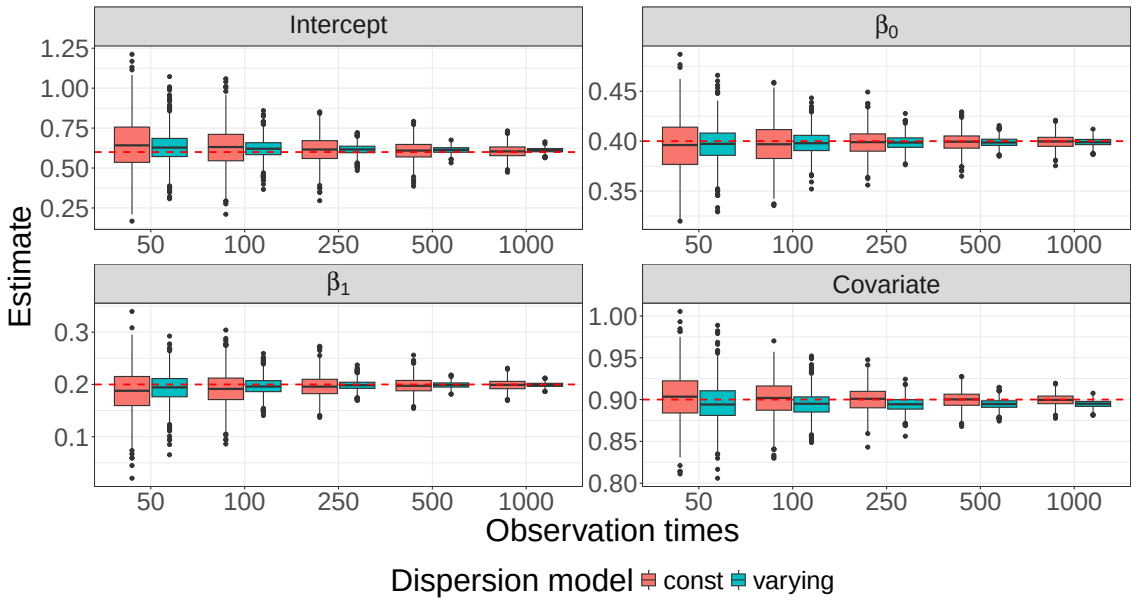


(a) Mean model (7.6), i.e. with feedback mechanism.

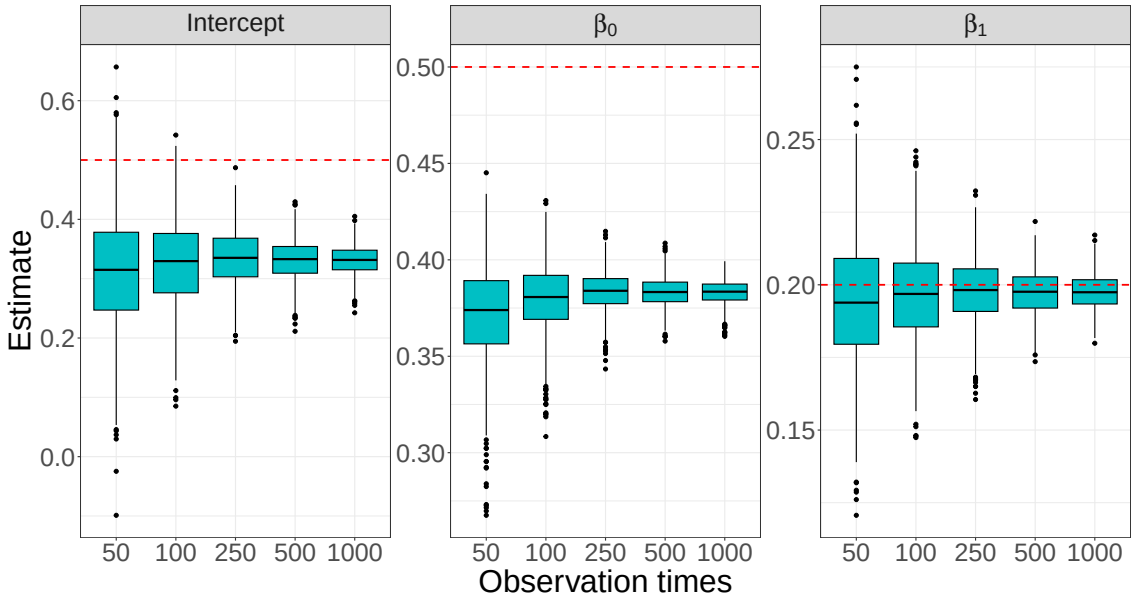


(b) Dispersion model (7.9), i.e. with feedback mechanism.

Figure A-19: Normal Q–Q plots of parameter estimates under the true data-generating process (7.6) and (7.9) with marginal normal distributions, based on $T = 100$.

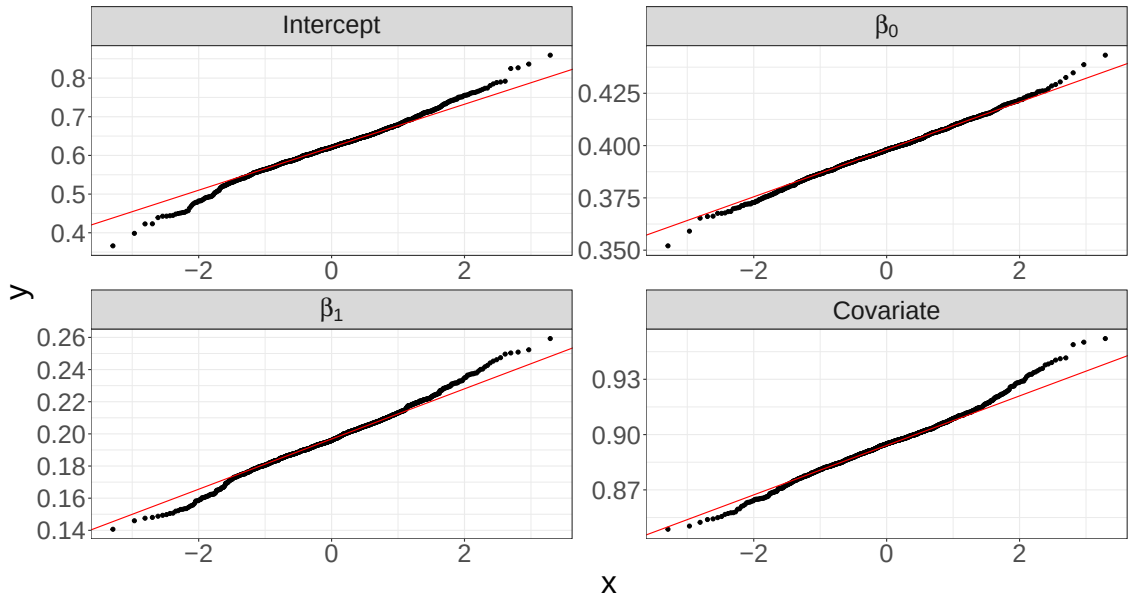


(a) Parameter estimates of the mean model (7.7) under the constant dispersion specification (7.8) and the varying dispersion specification (7.10).

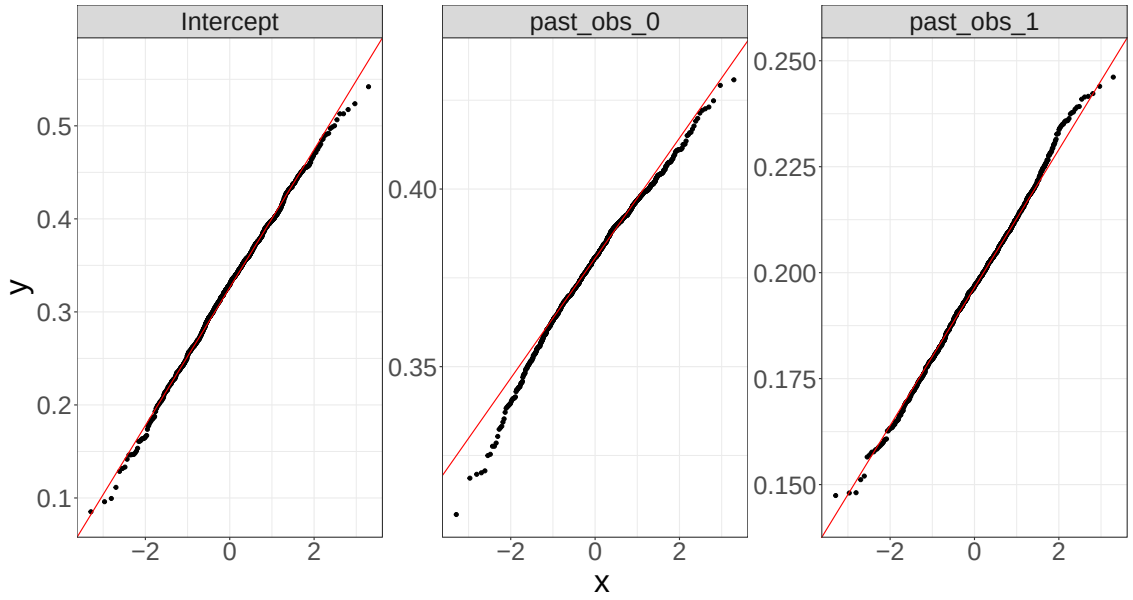


(b) Parameter estimates of the varying dispersion model (7.10) combined with the mean model (7.7).

Figure A-20: Box plots of parameter estimates under alternative dispersion specifications. Panel (a) shows estimates of the mean model under constant and varying dispersion, whereas panel (b) displays estimates of the varying dispersion model. The data-generating process with generalized Poisson marginals corresponds to (7.7) and (7.10), i.e. both models without feedback mechanism.

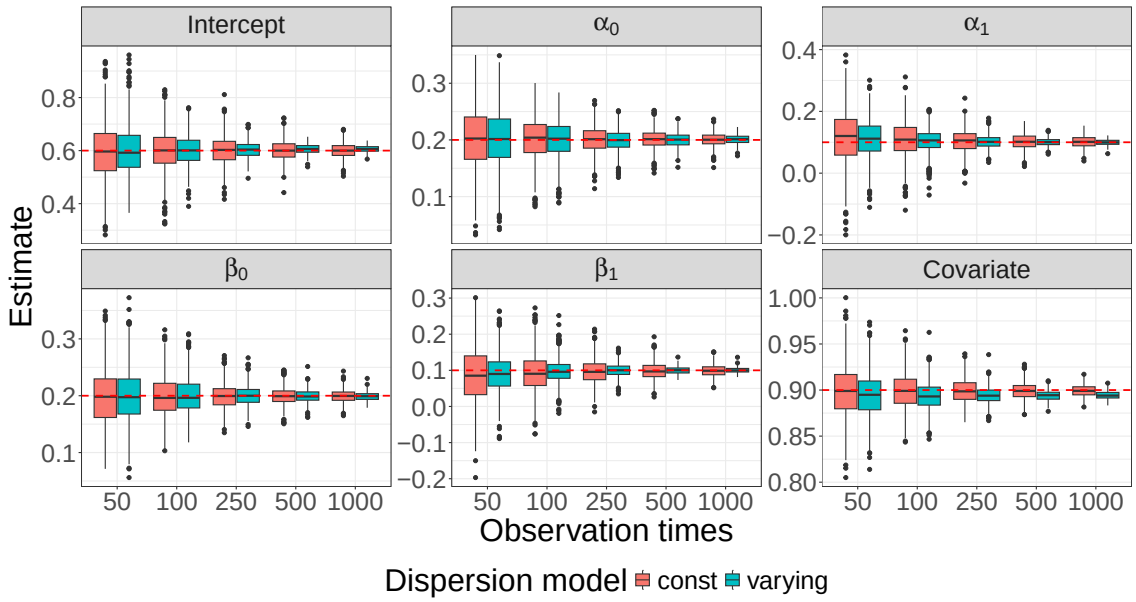


(a) Mean model (7.7), i.e. without feedback mechanism.

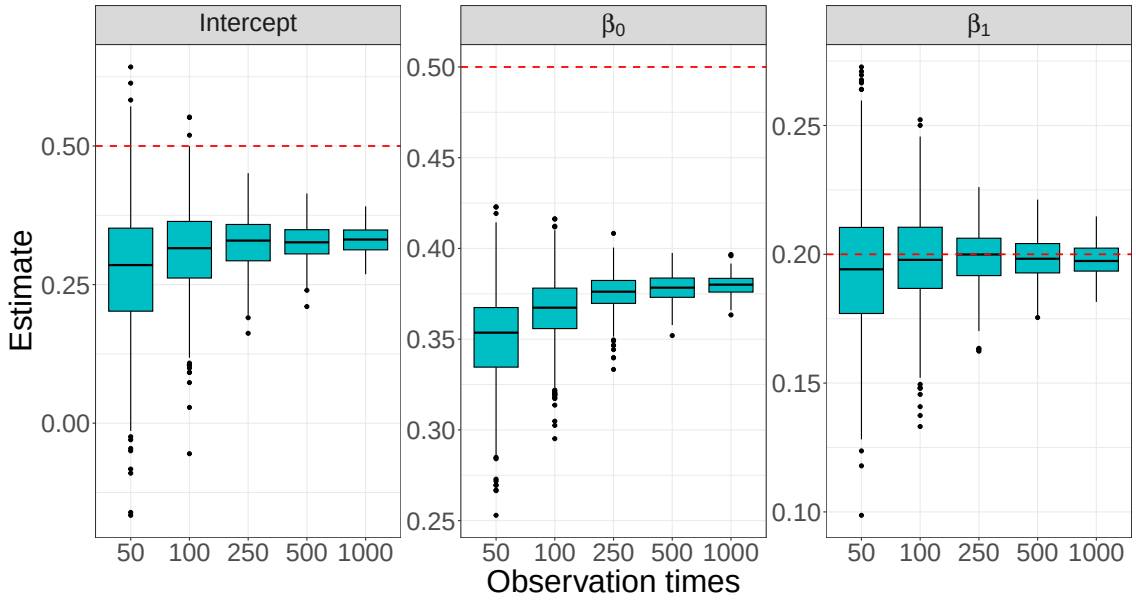


(b) Dispersion model (7.10), i.e. without feedback mechanism.

Figure A-21: Normal Q–Q plots of parameter estimates under the true data-generating process (7.7) and (7.10) with generalized Poisson marginals, based on $T = 100$.

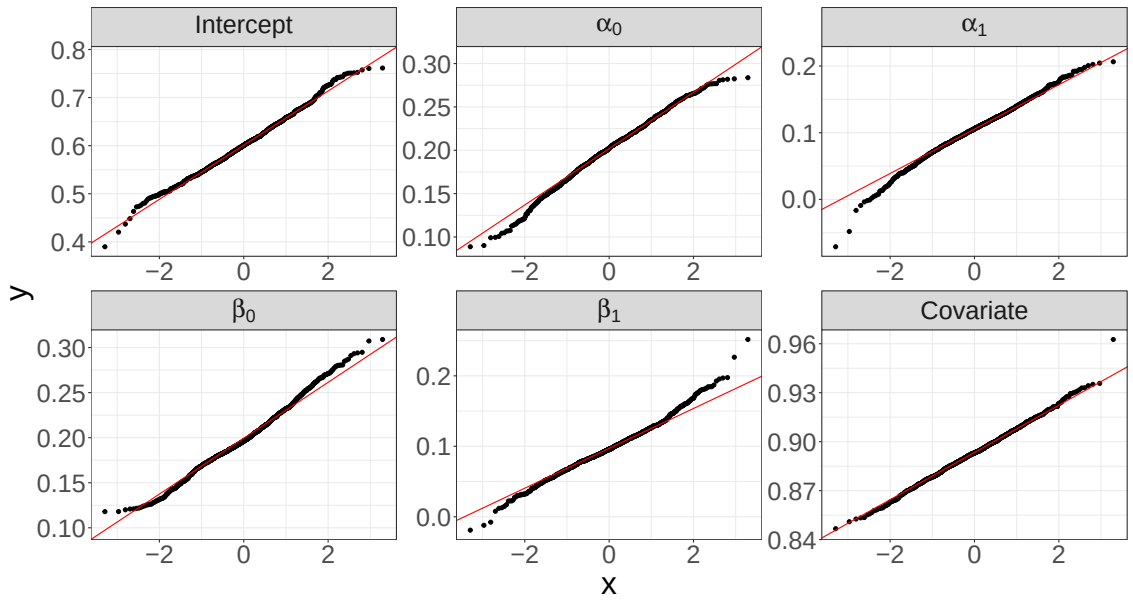


(a) Parameter estimates of the mean model (7.6) under the constant dispersion specification (7.8) and the varying dispersion specification (7.10).

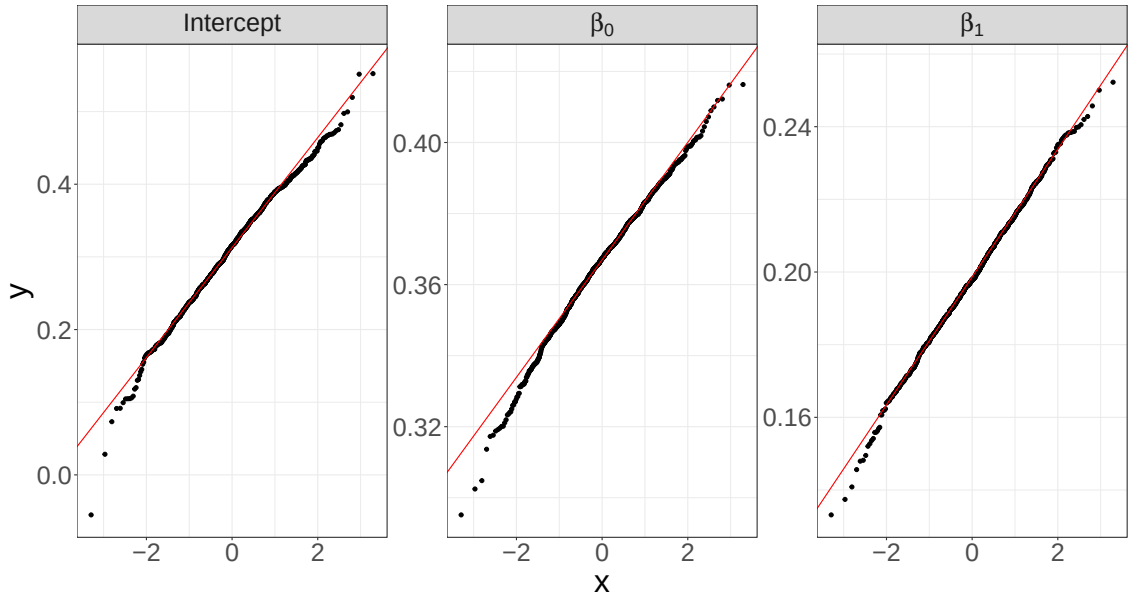


(b) Parameter estimates of the varying dispersion model (7.10) combined with the mean model (7.6).

Figure A-22: Box plots of parameter estimates under alternative dispersion specifications. Panel (a) shows estimates of the mean model under constant and varying dispersion, whereas panel (b) displays estimates of the varying dispersion model. The data-generating process with generalized Poisson marginals corresponds to (7.6) and (7.10), i.e. the mean model includes a feedback mechanism and the dispersion process not.

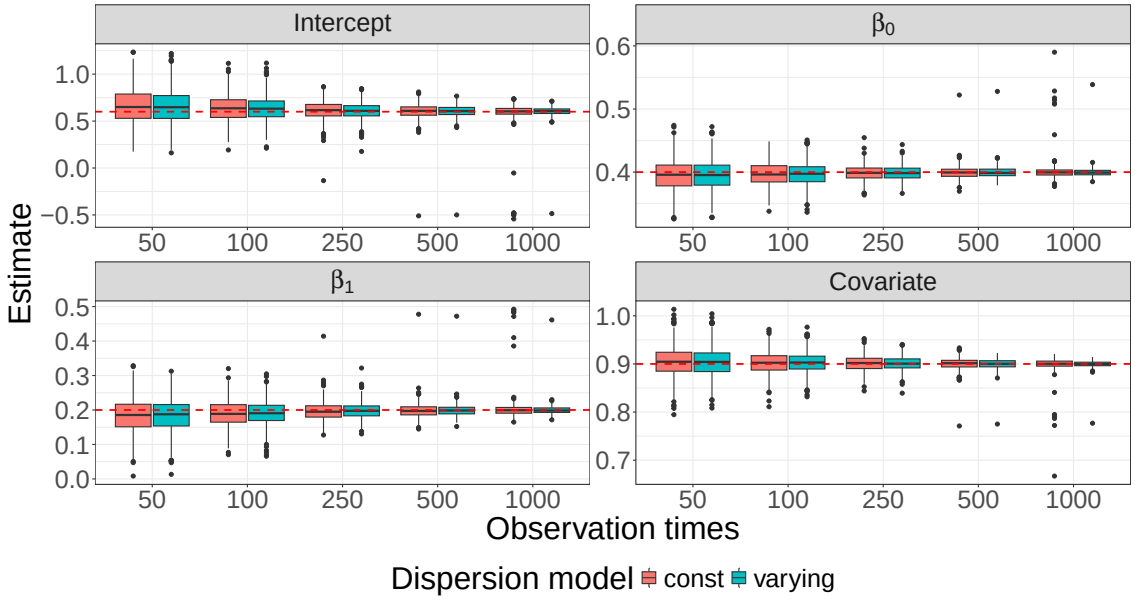


(a) Mean model (7.6), i.e. with feedback mechanism.

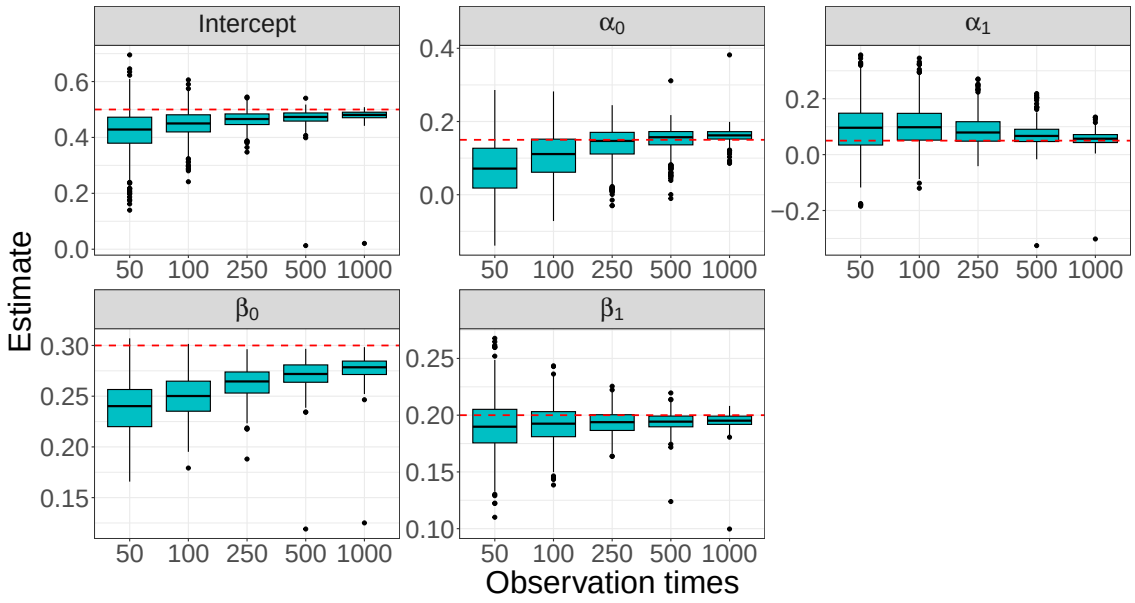


(b) Dispersion model (7.10), i.e. without feedback mechanism.

Figure A-23: Normal Q–Q plots of parameter estimates under the true data-generating process (7.6) and (7.10) with generalized Poisson marginals, based on $T = 100$.

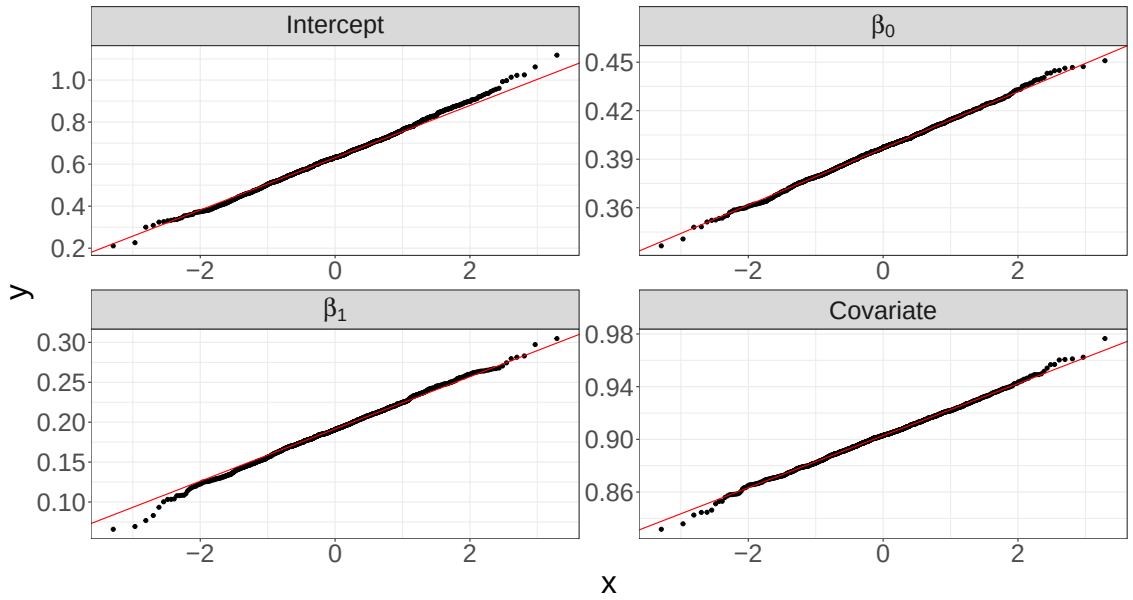


(a) Parameter estimates of the mean model (7.7) under the constant dispersion specification (7.8) and the varying dispersion specification (7.9).

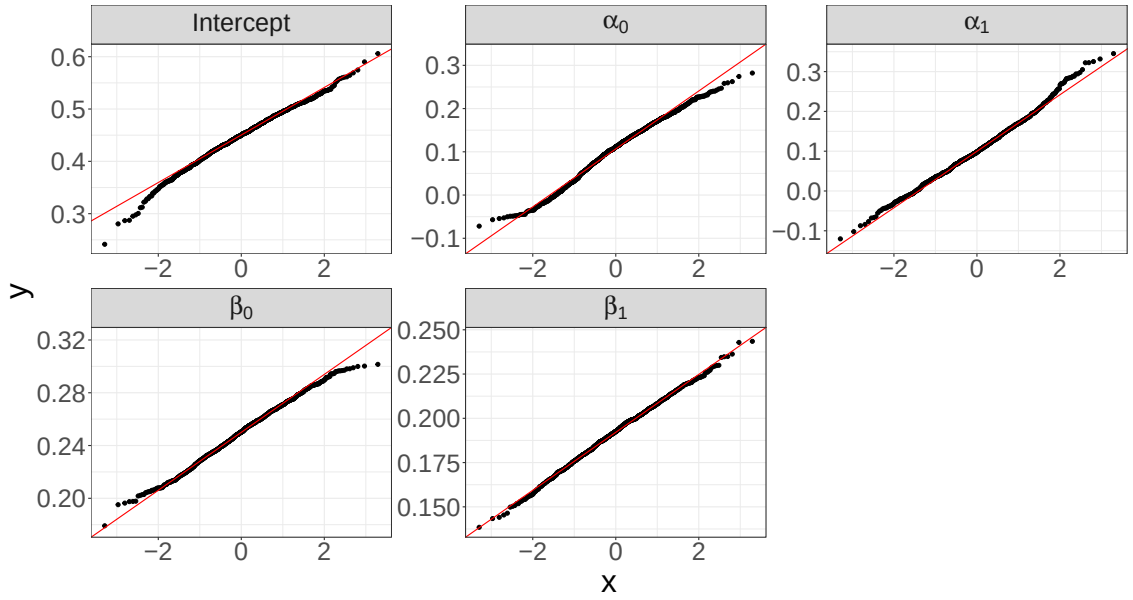


(b) Parameter estimates of the varying dispersion model (7.9) combined with the mean model (7.7).

Figure A-24: Boxplots of parameter estimates under alternative dispersion specifications. Panel (a) shows estimates of the mean model under constant and varying dispersion, whereas panel (b) displays estimates of the varying dispersion model. The data-generating process with generalized Poisson marginals corresponds to (7.7) and (7.9), i.e. the mean model does not include a feedback mechanism while the dispersion process does.

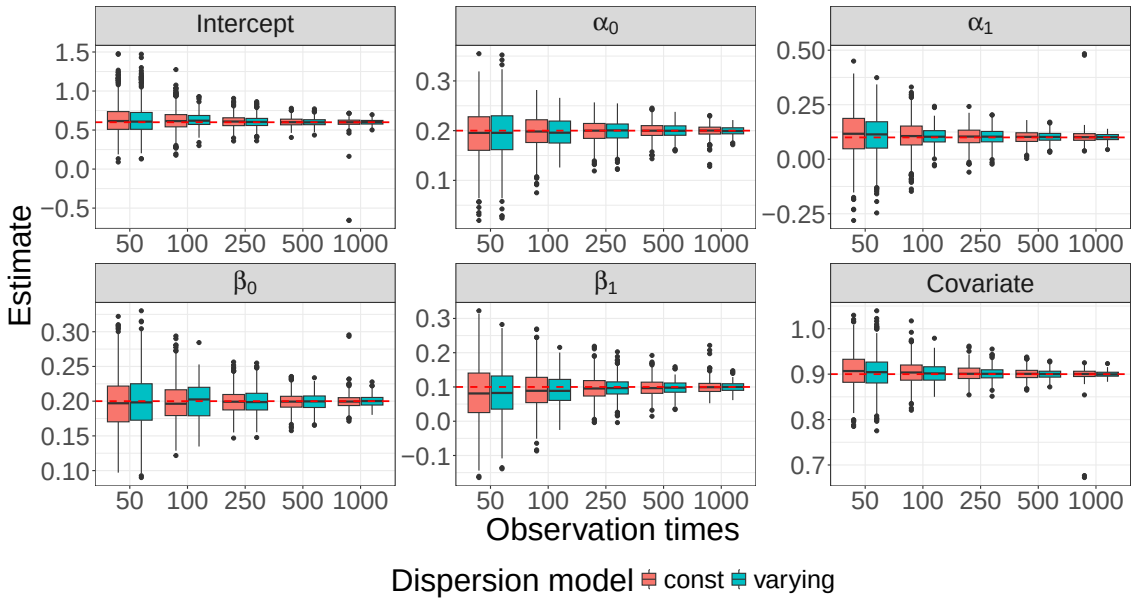


(a) Mean model (7.7), i.e. without feedback mechanism.

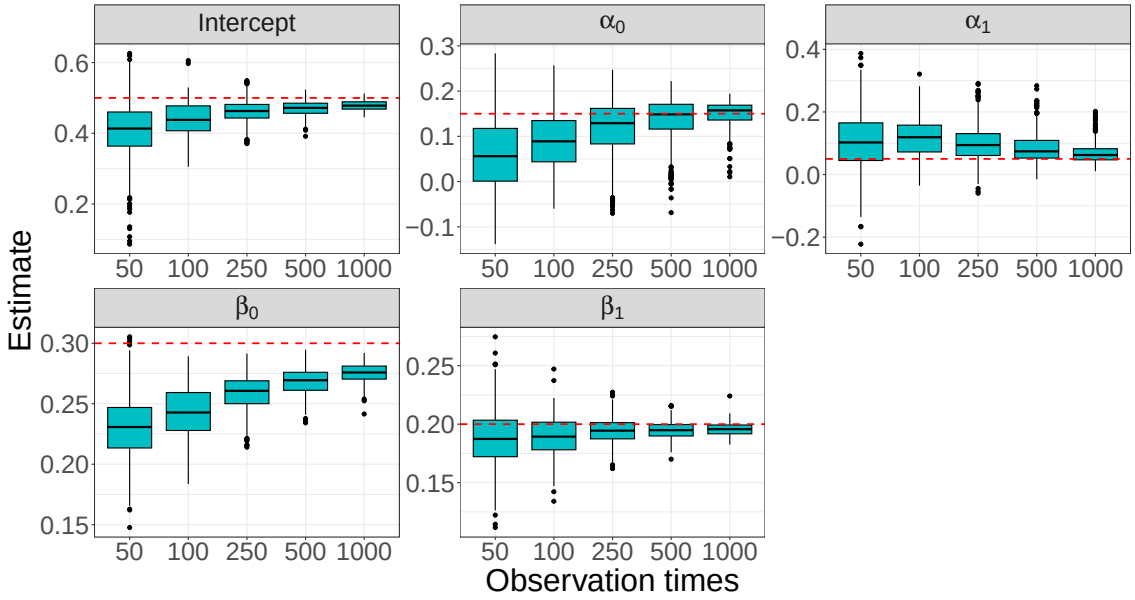


(b) Dispersion model (7.9), i.e. with feedback mechanism.

Figure A-25: Normal Q–Q plots of parameter estimates under the true data-generating process (7.7) and (7.9) with generalized Poisson marginals, based on $T = 100$.

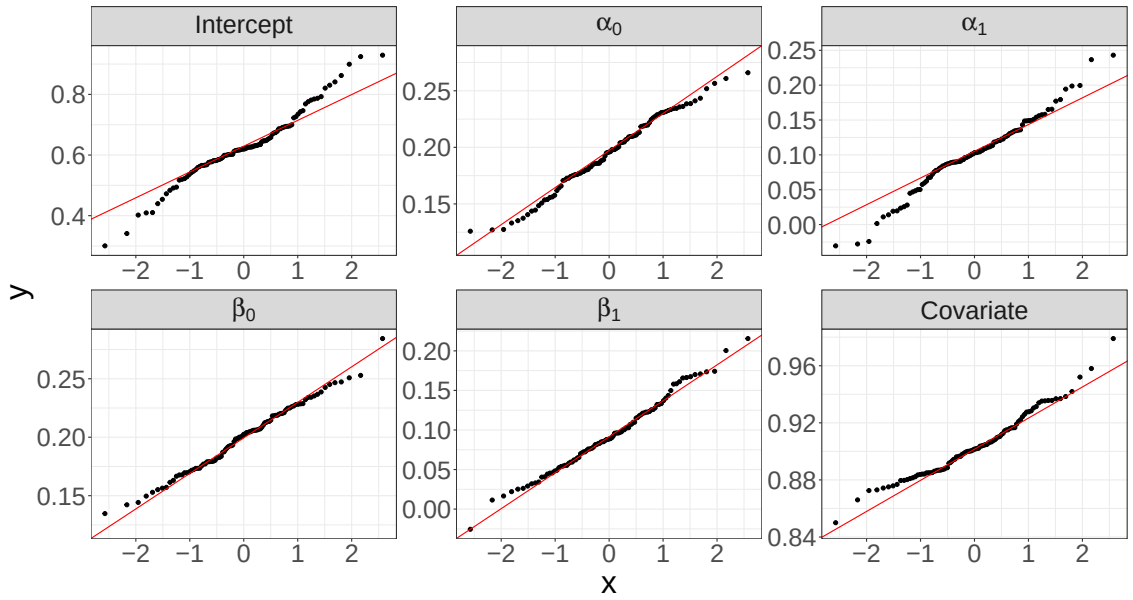


(a) Parameter estimates of the mean model (7.6) under the constant dispersion specification (7.8) and the varying dispersion specification (7.9).

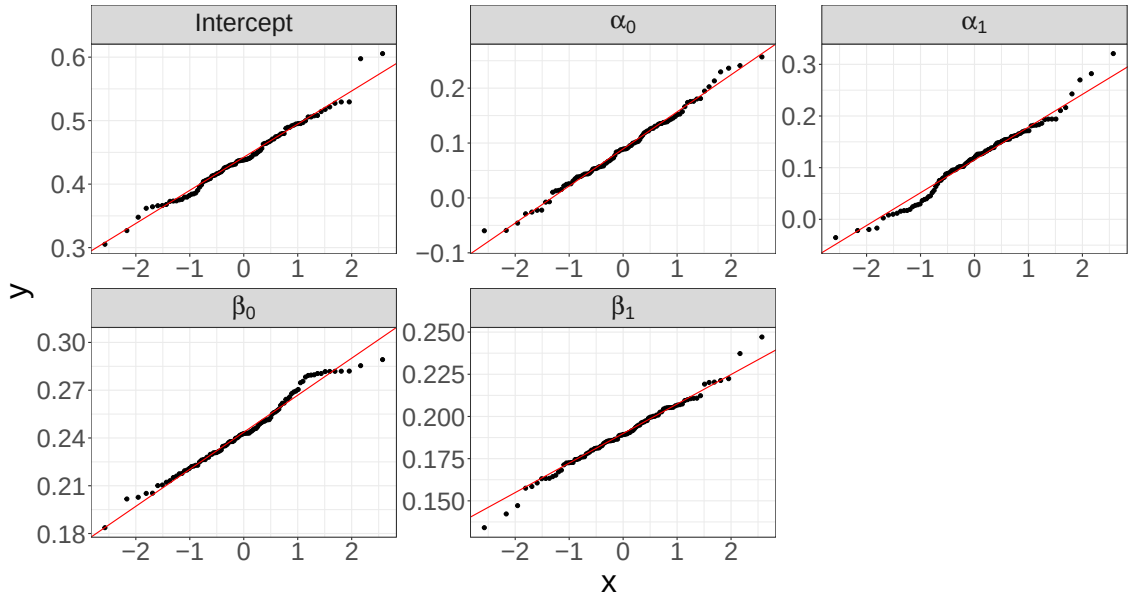


(b) Parameter estimates of the varying dispersion model (7.9) combined with the mean model (7.6).

Figure A-26: Boxplots of parameter estimates under alternative dispersion specifications. Panel (a) shows estimates of the mean model under constant and varying dispersion, whereas panel (b) displays estimates of the varying dispersion model. The data-generating process with generalized Poisson marginals corresponds to (7.7) and (7.9), i.e. both models include a feedback mechanism



(a) Mean model (7.6), i.e. with feedback mechanism.



(b) Dispersion model (7.9), i.e. with feedback mechanism.

Figure A-27: Normal Q–Q plots of parameter estimates under the true data-generating process (7.6) and (7.9) with generalized Poisson marginals, based on $T = 100$.