technische universität
dortmund

Modeling Count Time Series Following Generalized Linear Models

Dissertation

Tobias Liboschik

In partial fulfillment of the
requirements for the degree of
Doktor der Naturwissenschaften
presented to the

Department of Statistics
TU Dortmund University

Advisors:
Prof. Dr. Roland Fried, TU Dortmund University
Prof. Dr. Konstantinos Fokianos, University of Cyprus

Dortmund, 13th July 2016

Abstract

Count time series are found in many different applications, e.g. from medicine, finance or industry, and have received increasing attention in the last two decades. The class of count time series following generalized linear models is very flexible and can describe serial correlation in a parsimonious way. The conditional mean of the observed process is linked to its past values, to past observations and to potential covariate effects. In this thesis we give a comprehensive formulation of this model class. We consider models with the identity and with the logarithmic link function. The conditional distribution can be Poisson or Negative Binomial. An important special case of this class is the so-called INGARCH model and its log-linear extension.

A key contribution of this thesis is the R package **tscount** which provides likelihood-based estimation methods for analysis and modeling of count time series based on generalized linear models. The package includes methods for model fitting and assessment, prediction and intervention analysis. This thesis summarizes the theoretical background of these methods. It gives details on the implementation of the package and provides simulation results for models which have not been studied theoretically before. The usage of the package is illustrated by two data examples. Additionally, we provide a review of R packages which can be used for count time series analysis. A detailed comparison of **tscount** to those packages demonstrates that **tscount** is an important contribution which extends and complements existing software.

A thematic focus of this thesis is the treatment of all kinds of unusual effects influencing the ordinary pattern of the data. This includes structural changes and different forms of outliers one is faced with in many time series. Our first study on this topic is concerned with retrospective detection of such changes. We analyze different approaches for modeling such intervention effects in count time series based on INGARCH models. Other authors treated a model where an intervention affects the non-observable underlying mean process at the time point of its occurrence and additionally the whole process thereafter via its dynamics. As an alternative, we consider a model where an intervention directly affects the observation at its occurrence, but not the underlying mean, and then also enters the dynamics of the process. While the former definition describes an internal change of the system, the latter can be understood as an external effect on the observations due to e.g. immigration. For our alternative model we develop conditional likelihood estimation and, based on this, develop tests and detection procedures for intervention effects. Both models are compared analytically and using simulated and real data examples. The procedures for our new model work reliably and we find some robustness against misspecification of the intervention model.

The aforementioned methods are applied after the complete time series has been observed. In another study we investigate the prospective detection of structural changes, i.e. in real time. For example in public health, surveillance of infectious diseases aims at recognizing outbreaks of epidemics with only short time delays in order to take adequate action promptly. We point out that serial dependence is present in many infectious disease time series. Nevertheless it is still ignored by many procedures used for infectious disease surveillance. Using historical data, we design a prediction-based monitoring procedure for count time series following generalized linear models. We illustrate benefits but also pitfalls of using dependence models for monitoring.

Moreover, we briefly review the literature on model selection, robust estimation and robust prediction for count time series. We also make a first study on robust model identification using robust estimators of the (partial) autocorrelation.

Keywords: Count time series, generalized linear models, serial correlation, temporal dependence, autoregressive models, regression models, likelihood, mixed Poisson, model selection, prediction, forecasting, statistical software, R, intervention analysis, level shifts, outliers, statistical process control, online monitoring, change point detection, aberration detection, outbreak detection, infectious disease surveillance, epidemiology, public health.

Contents

Chapter 1

Introduction

1.1 Motivation

Count time series appear naturally in various areas whenever a number of events per time period is observed over time. Examples for the wide range of applications in medicine are the weekly number of patients recruited for a clinical trial, the daily number of hospital admissions or the weekly number of epileptic seizures of a patient. An important example from epidemiology is the weekly number of registered infections by certain pathogens, which is routinely collected by public health authorities. Important objectives of such data analysis are the prediction of future values for adequate planning of resources, the detection of unusual values pointing at some epidemics or the proper description of e.g. seasonal patterns for better understanding and interpretation of data generating mechanisms. Examples from other fields are the number of stock market transactions per minute, from finance, or the hourly number of defect items, from industrial quality control.

Models for count time series should take into account that the observations are nonnegative integers and they should capture suitably the dependence among observations. A convenient and flexible approach is to employ the generalized linear model (GLM) methodology (Nelder and Wedderburn, 1972) for modeling the observations conditionally on the past information. This methodology is implemented by choosing a suitable distribution for count data and an appropriate link function. Such an approach is pursued by Fahrmeir and Tutz (2001, Chapter 6) and Kedem and Fokianos (2002, Chapters 1–4), among others. Another important class of models for time series of counts is based on the thinning operator, like the integer autoregressive moving average (INARMA) models, which, in a way, imitate the structure of the common autoregressive moving average (ARMA) models (see the review article by Weiß, 2008). A different type of count time series models are the so-called state space models. We refer to the reviews of Fokianos (2011), Jung and Tremayne (2011), Fokianos (2012), Tjøstheim (2012) and Fokianos (2015) for an in-depth overview of models for count time series. Advantages of GLM-based models compared to the models which are based on the thinning operator are the following:

(a) They can describe covariate effects and negative correlations in a straightforward way.

(b) There is a rich toolkit available for this class of models.

State space models allow to describe even more flexible data generating processes than GLM models but at the cost of a more complicated model specification. On the other hand, GLM-based models yield predictions in a convenient manner due to their explicit formulation.

This thesis is concerned with methods for count time series based on generalized linear models. In the following section we give a comprehensive formulation of this model class.

1.2 Models

Denote a count time series by $\{Y_t : t \in \mathbb{N}\}$. We will denote by $\{\boldsymbol{X}_t : t \in \mathbb{N}\}$ a time-varying r-dimensional covariate vector, say $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,r})^\top$. We model the conditional mean $\mathsf{E}\left(Y_t|\mathcal{F}_{t-1}\right)$ of the count time series by a process, say $\{\lambda_t : t \in \mathbb{N}\}$, such that $\mathsf{E}\left(Y_t|\mathcal{F}_{t-1}\right) = \lambda_t$. Denote by $\mathcal{F}_{t_0}$ the history of the joint process $\{Y_t, \lambda_t, \boldsymbol{X}_{t+1} : t \in \mathbb{N}\}$ up to time t_0 including the covariate information at time $t_0 + 1$. The distributional assumption for Y_t given $\mathcal{F}_{t-1}$ is discussed later. We are interested in models of the general form

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \, \widetilde{g}(Y_{t-i_k}) + \sum_{\ell=1}^{q} \alpha_\ell g(\lambda_{t-j_\ell}) + \boldsymbol{\eta}^\top \boldsymbol{X}_t, \tag{1.1}$$

where $g : \mathbb{R}^+ \to \mathbb{R}$ is a link function and $\widetilde{g} : \mathbb{N}_0 \to \mathbb{R}$ is a transformation function. The parameter vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_r)^\top$ corresponds to the effects of covariates. In the terminology of GLMs we call $\nu_t = g(\lambda_t)$ the linear predictor. To allow for regression on arbitrary past observations of the response, define a set $P = \{i_1, i_2, \ldots, i_p\}$ and integers $0 < i_1 < i_2 \ldots < i_p < \infty$, with $p \in \mathbb{N}_0$. This enables us to regress on the lagged observations $Y_{t-i_1}, Y_{t-i_2}, \ldots, Y_{t-i_p}$. Analogously, define a set $Q = \{j_1, j_2, \ldots, j_q\}$, $q \in \mathbb{N}_0$ and integers $0 < j_1 < j_2 \ldots < j_q < \infty$, for regression on lagged conditional means

$\lambda_{t-j_1}, \lambda_{t-j_2}, \ldots, \lambda_{t-j_q}$. This case is covered by the theory for models with $P = \{1, \ldots, p\}$ and $Q = \{1, \ldots, q\}$ by choosing p and q suitably and setting some model parameters to zero. Our formulation is useful particularly when dealing with modeling stochastic seasonality (see Section 2.6.1, for an example). Specification of the model order, i.e., of the sets P and Q, are guided by considering the empirical autocorrelation functions of the observed data. This approach is described for ARMA models in many time series analysis textbooks and transfers to the above model by employing its ARMA representation (see (A.4) in Appendix A.3). Parameter constraints which ensure stationarity and ergodicity of two important special cases of (1.1) are given in Section 2.2.1.

We give several examples of model (1.1). Consider the situation where g and $\widetilde{g}$ equal the identity, i.e., $g(x) = \widetilde{g}(x) = x$. Furthermore, let $P = \{1, \ldots, p\}$, $Q = \{1, \ldots, q\}$ and $\boldsymbol{\eta} = \mathbf{0}$. Then model (1.1) becomes

$$\lambda_t = \beta_0 + \sum_{k=1}^{p} \beta_k \, Y_{t-k} + \sum_{\ell=1}^{q} \alpha_\ell \lambda_{t-\ell}. \tag{1.2}$$

Assuming further that Y_t given the past is Poisson distributed, then we obtain an *integer-valued GARCH model* of order p and q, abbreviated as INGARCH(p,q). These models are also known as *autoregressive conditional Poisson (ACP) models*. They have been discussed by Heinen (2003), Ferland, Latour, and Oraichi (2006) and Fokianos, Rahbek, and Tjøstheim (2009), among others. An example of an INGARCH model with covariates is given in Section 2.5, where we fit a count time series model which includes intervention effects.

Consider again model (1.1) but now with the logarithmic link function $g(x) = \log(x)$, $\widetilde{g}(x) = \log(x + 1)$ and P, Q as before. Then, we obtain a *log-linear model* of order p and q for the analysis of count time series. Indeed, set $\nu_t = \log(\lambda_t)$ to obtain from (1.1) that

$$\nu_t = \beta_0 + \sum_{k=1}^{p} \beta_k \, \log(Y_{t-k} + 1) + \sum_{\ell=1}^{q} \alpha_\ell \nu_{t-\ell}. \tag{1.3}$$

This log-linear model is studied by Fokianos and Tjøstheim (2011), Woodard, Matteson, and Henderson (2011) and Douc, Doukhan, and Moulines (2013). We follow Fokianos and Tjøstheim (2011) in transforming past observations by employing the function $\widetilde{g}(x) = \log(x + 1)$, such that they are on the same scale as the linear predictor ν_t. These authors show that the addition of a constant c to each observation for avoiding zero values does not affect inference; in addition they argue that a reasonable choice for c is 1. Note that model (1.3) allows modeling of negative serial correlation, whereas model (1.2) accommodates positive serial correlation only. Additionally, (1.3) accommodates covariates easier than (1.2) since the log-linear model implies positivity of the conditional mean process $\{\lambda_t\}$. The linear model (1.2) with covariates should be fitted with some care because it is limited to positive effects on $\{\lambda_t\}$. This is so because we need to ensure that the resulting mean process is positive. The effects of covariates on the response are multiplicative for model (1.3); they are additive though for model (1.2). For a discussion on the inclusion of time-dependent covariates see Fokianos and Tjøstheim (2011, Section 4.3).

In model (1.1) the effect of a covariate fully enters the dynamics of the process and propagates to future observations both by the regression on past observations and by the regression on past conditional means. The effect of such covariates can be seen as an internal influence on the data-generating process, which is why we refer to it as an *internal* covariate effect. We also allow to include covariates in a way that their effect only propagates to future observations by the regression on past observations but not directly by the regression on past conditional means. Following Liboschik, Kerschke, Fokianos, and Fried (2016), who make this distinction for the case of intervention effects described by deterministic covariates, we refer to the effect of such covariates as an *external* covariate effect. Let $\boldsymbol{e} = (e_1, \ldots, e_r)^\top$ be a vector specified by the user with $e_i = 1$ if the i-th component of the covariate vector has an external effect and $e_i = 0$ otherwise, $i = 1, \ldots, r$. Denote by $\mathrm{diag}(\boldsymbol{e})$ a diagonal matrix with diagonal elements given by $\boldsymbol{e}$. The generalization of (1.1) allowing for both internal and external covariate effects is given by

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \, \widetilde{g}(Y_{t-i_k}) + \sum_{\ell=1}^{q} \alpha_\ell \left(g(\lambda_{t-j_\ell}) - \boldsymbol{\eta}^\top \mathrm{diag}(\boldsymbol{e}) \boldsymbol{X}_{t-j_\ell} \right) + \boldsymbol{\eta}^\top \boldsymbol{X}_t. \qquad (1.4)$$

Basically, the effect of all covariates with an external effect is subtracted in the feedback terms such that their effect enters the dynamics of the process only via the observations. We refer to Chapter 3 (based on Liboschik *et al.*, 2016) for an extensive discussion and comparison of internal and external effects. It is our experience with these models that on the one hand an empirical discrimination between internal and external covariate effects is difficult but on the other hand there is some robustness against misspecification of the type of covariate effect.

So far we have only specified the mean of $Y_t|\mathcal{F}_{t-1}$ but not its distribution. Model (1.1) together with the *Poisson* assumption, i.e., $Y_t|\mathcal{F}_{t-1} \sim \mathrm{Poisson}(\lambda_t)$, implies

$$\mathsf{P}\left(Y_t = y|\mathcal{F}_{t-1}\right) = \frac{\lambda_t^y \exp(-\lambda_t)}{y!}, \quad y \in \mathbb{N}_0. \qquad (1.5)$$

It holds $\mathsf{VAR}\left(Y_t|\mathcal{F}_{t-1}\right) = \mathsf{E}\left(Y_t|\mathcal{F}_{t-1}\right) = \lambda_t$. Hence in the case of a conditional Poisson response model the conditional mean is identical to the conditional variance of the observed process.

The *Negative Binomial* distribution allows for a conditional variance to be larger than the mean λ_t, which is often referred to as overdispersion and observed in many time series. Following Christou and Fokianos (2014), it is assumed that $Y_t|\mathcal{F}_{t-1} \sim \text{NegBin}(\lambda_t, \phi)$, where the Negative Binomial distribution is parametrized in terms of its mean with an additional dispersion parameter $\phi \in (0, \infty)$ (Hilbe, 2011), i.e.,

$$\mathsf{P}\left(Y_t = y|\mathcal{F}_{t-1}\right) = \frac{\Gamma(\phi + y)}{\Gamma(y+1)\Gamma(\phi)} \left(\frac{\phi}{\phi + \lambda_t}\right)^{\phi} \left(\frac{\lambda_t}{\phi + \lambda_t}\right)^{y}, \quad y \in \mathbb{N}_0. \tag{1.6}$$

In this case it again holds $\mathsf{E}\left(Y_t|\mathcal{F}_{t-1}\right) = \lambda_t$ but $\mathsf{VAR}\left(Y_t|\mathcal{F}_{t-1}\right) = \lambda_t + \lambda_t^2/\phi$, i.e., the conditional variance increases quadratically with λ_t. The Poisson distribution is a limiting case of the Negative Binomial when $\phi \to \infty$.

Note that the Negative Binomial distribution belongs to the class of mixed Poisson processes. A mixed Poisson process is specified by setting $Y_t = N_t(0, Z_t\lambda_t]$, where $\{N_t\}$ are i.i.d. Poisson processes with unit intensity and $\{Z_t\}$ are i.i.d. random variables with mean 1 and variance σ^2, independent of $\{Y_t\}$. When $\{Z_t\}$ is an i.i.d. process of Gamma random variables, then we obtain the Negative Binomial process with $\sigma^2 = 1/\phi$. We refer to σ^2 as the overdispersion coefficient because it is proportional to the extent of overdispersion of the conditional distribution. The limiting case of $\sigma^2 = 0$ corresponds to the Poisson distribution, i.e., no overdispersion. The estimation procedure we study is not confined to the Negative Binomial case but to any mixed Poisson distribution. However, the Negative Binomial assumption is required for prediction intervals and model assessment; these topics are discussed in Sections 2.3 and 2.4.

1.3 Outline

Chapter 2 introduces the R package **tscount** which implements methods for the class of count time series following GLMs. This package includes methods for model fitting and assessment, prediction and intervention analysis. The chapter summarizes the theoretical background of these methods. It gives details on the implementation of the package and provides simulation results for models which have not been studied theoretically before. The usage of the package is illustrated by two data examples. Additionally, we provide a review of R packages which can be used for count time series analysis. This includes a detailed comparison of **tscount** to those packages. The chapter is based on a manuscript entitled "**tscount**: An R package for Analysis of Count Time Series Following Generalized Linear Models" which is currently under revision for *Journal of Statistical Software*. A previous version of that manuscript has been published as a discussion paper (Liboschik, Fokianos, and Fried, 2015) and the most recent version is available as a vignette of the package.

In many applications, unusual external effects or measurement errors can lead to either sudden or gradual changes in the structure of the data, so-called intervention effects. A goal of an intervention analysis is to examine the effect of known interventions, for example to judge whether a policy change had the intended impact, or to search for unknown intervention effects and to find explanations for them. Chapter 3 studies different approaches for modeling intervention effects by deterministic covariates, focusing on INGARCH models from the class of GLM-based count time series models. Fokianos and Fried (2010) treated a model where an intervention has an internal effect according to the definition in the previous section, which describes an internal change of the system. We consider an alternative model where an intervention has an external effect on the observations due to e.g. immigration. For our alternative model we develop conditional likelihood estimation and, based on this, tests and detection procedures for intervention effects. We compare both models analytically and using simulated and real data examples. Our simulations confirm that the procedures for our new model perform well. It turns out that there is some robustness against misspecification of the intervention model. The chapter is based on the article "Modelling interventions in INGARCH processes" published in the *International Journal of Computer Mathematics* (Liboschik *et al.*, 2016, accepted for publication in July 2014).

The aforementioned methods are applied after the complete time series has been observed. Chapter 4 investigates the prospective detection of structural changes, i.e. in real time. For example in public health, surveillance of infectious diseases aims at recognizing outbreaks of epidemics with only short time delays in order to take adequate action promptly. We point out that serial dependence is present in many infectious disease time series. Nevertheless it is still ignored by many procedures used for infectious disease surveillance. This chapter studies how accommodating temporal dependence can improve monitoring procedures. Using historical data, we design a prediction-based monitoring procedure for count time series following GLMS. Our simulations and a data example demonstrate that such a procedure can substantially improve the immediate detection of outbreaks but that its dependence on previous observations may also yield undesired effects in some situations. We discuss some ideas how to utilize the promising features of dependent models for monitoring and at the same time to overcome their weakness.

Chapter 5 discusses three further topics, all of particular interest in the context of infectious disease surveillance, and points to directions where further research is needed. Section 5.1 reviews some further tools for model selection. The other two sections are concerned with robust methods which work reliably in the presence of outliers or intervention effects. Section 5.2 is a first study on robust identification of the model order for count time series with the robustly estimated (partial) autocorrelation function. This Section is based on Section 4 of the article “On Outliers and Interventions in Count Time Series following GLMs” published in the *Austrian Journal of Statistics* (Fried, Liboschik, Elsaied, Kitromilidou, and Fokianos, 2014, Sections 4 and 5 are written by the author of this thesis). Section 5.3 reviews robust methods for parameter estimation and prediction. Chapter 6 concludes this thesis with a brief summary of its most important results.

Chapter 2

Basic methods and implementation

2.1 Introduction

Recently, there has been an increasing interest in regression models for time series of counts and a considerable number of publications on this subject has appeared in the literature. However, many of the proposed methods are not yet available in a statistical software package and hence they cannot be applied easily. We aim at filling this gap and publish a package named **tscount** for the popular free and open source software environment R (R Core Team, 2016). In fact, our main goal is to develop software for models whose conditional mean depends on previous observations and on its own previous values. These models are quite analogous to the generalized autoregressive conditional heteroscedasticity (GARCH) models (Bollerslev, 1986) which were proposed for describing the conditional variance.

In the first version of the package **tscount** we provide likelihood-based methods for the framework of count time series following GLMs. Some simple autoregressive models can be fitted with standard software by treating the observations as if they were independent (see Section 2.7 and Appendix A.3), for example, using the R function `glm`. However, these procedures are in general not tailored for dependent data and may yield invalid model fits. The implementation in the package **tscount** allows for a more general dependence structure which is specified conveniently by the user. We consider general time series models whose conditional mean may depend on time-varying covariates, previous observations and, similar to the conditional variance of a GARCH model, on its own previous values. The usage and output of our functions is in parts inspired by the R functions `arima` and `glm` in order to provide a familiar user experience. Furthermore **tscount** is object-oriented and provides many standard `S3` methods for well-known generic functions. There are several other R functions available which can be employed for analyzing count time series. Many of those are related to GLMs and have been developed for independent observations but are, with some limitations, also capable to describe simple forms of serial dependence. There are also some functions available for extending such models to time series. Another group of functions handles state space models for count time series. We briefly review these functions and the corresponding model classes in Section 2.7 and compare them to **tscount**. As it turns out, there are special cases for which our model corresponds to existing ones. In these cases we obtain quite similar results with functions from some other packages, thus confirming the reliability of our package. However, many features of **tscount**, like the flexible dependence structure, outreach the capability of other packages. Admittedly, some packages provide features like zero-inflation or more general forms of the linear predictor which cannot be accommodated yet by **tscount** but could possibly be included in future versions. As a conclusion, this package is a valuable addition to the R environment which fills some significant gaps associated with time series fitting.

The functionality of **tscount** partly goes beyond the theory available in the literature since theoretical investigation of these models is still an ongoing research theme. For instance the problem of accommodating covariates in such GLM-type count time series models or fitting a mixed Poisson log-linear model have not been studied theoretically. We have checked their appropriateness by simulations reported in Appendix B. However, some care should be taken when applying the package's programs to situations which are not covered by existing theory.

This chapter is organized as follows. At first the theoretical background of the methods included in the package is briefly summarized with references to the literature for more details. Section 2.2 describes quasi maximum likelihood estimation of the unknown model parameters and gives some details regarding its implementation. Section 2.3 treats prediction with such models. Section 2.4 sums up tools for model assessment. Section 2.5 discusses procedures for the detection of interventions. Section 2.6 demonstrates the usage of the package with two data examples. Section 2.7 reviews other R packages which are capable to model count time series and compares them with our package. Finally, Section 2.8 gives an outlook on possible future extensions of our package. In the Appendix we give further details and we confirm empirically some of the new methods that we discuss but which have not been studied, as of yet.

2.2 Estimation and inference

2.2.1 Estimation

The **tscount** package fits models of the form (1.1) by quasi conditional maximum likelihood (ML) estimation (function `tsglm`). If the Poisson assumption holds true, then we obtain an ordinary ML estimator. However, under the mixed Poisson assumption we obtain a quasi-ML estimator. Denote by $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_q, \eta_1, \ldots, \eta_r)^\top$ the vector of regression parameters. Regardless of the distributional assumption, the parameter space for the INGARCH model (1.2) with covariates is given by

$$\Theta = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{p+q+r+1} : \ \beta_0 > 0, \ \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_q, \eta_1, \ldots, \eta_r \geq 0, \ \sum_{k=1}^{p} \beta_k + \sum_{\ell=1}^{q} \alpha_\ell < 1 \right\}.$$

The intercept β_0 must be positive and all other parameters must be nonnegative to ensure positivity of the conditional mean λ_t. The other condition ensures that the fitted model has a stationary and ergodic solution with moments of any order (Ferland *et al.*, 2006; Fokianos *et al.*, 2009; Doukhan, Fokianos, and Tjøstheim, 2012); see also Tjøstheim (2015) for a recent review. For the log-linear model (1.3) with covariates the parameter space is taken to be

$$\Theta = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{p+q+r+1} : \ |\beta_1|, \ldots, |\beta_p|, |\alpha_1|, \ldots, |\alpha_q| < 1, \ \left| \sum_{k=1}^{p} \beta_k + \sum_{\ell=1}^{q} \alpha_\ell \right| < 1 \right\},$$

see Appendix A.1 for a discussion. Christou and Fokianos (2014) point out that with the parametrization (1.6) of the Negative Binomial distribution the estimation of the regression parameters $\boldsymbol{\theta}$ does not depend on the additional dispersion parameter ϕ. This allows to employ a quasi maximum likelihood approach based on the Poisson likelihood to estimate the regression parameters $\boldsymbol{\theta}$, which is described below. The nuisance parameter ϕ is then estimated separately in a second step. This approach is different from a full maximum likelihood estimation based on the Negative Binomial distribution, which for example has been implemented in the function `glm.nb` in the R package **MASS** (Venables and Ripley, 2002). In that algorithm, maximization of the Negative Binomial likelihood for an estimated dispersion parameter ϕ and estimation of ϕ given the estimated regression parameters $\boldsymbol{\theta}$ are iterated until convergence. The quasi negative binomial approach has been chosen for simplicity and its usefulness on deriving consistent estimators when the model for λ_t has been correctly specified (see also Ahmad and Francq, 2016).

The log-likelihood, score vector and information matrix are derived conditionally on pre-sample values of the time series and the conditional mean process $\{\lambda_t\}$, precisely on $\mathcal{F}_0$. An appropriate initialization is needed for their evaluation, which is discussed in the next subsection. For a vector of observations $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, the conditional quasi log-likelihood function, up to a constant, is given by

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{n} \log p_t(y_t; \boldsymbol{\theta}) = \sum_{t=1}^{n} \Big(y_t \ln(\lambda_t(\boldsymbol{\theta})) - \lambda_t(\boldsymbol{\theta}) \Big), \tag{2.1}$$

where $p_t(y; \boldsymbol{\theta}) = \mathsf{P}(Y_t = y | \mathcal{F}_{t-1})$ is the probability density function of a Poisson distribution as defined in (1.5). The conditional mean is regarded as a function $\lambda_t : \Theta \to \mathbb{R}^+$ and thus it is denoted by $\lambda_t(\boldsymbol{\theta})$ for all t. The conditional score function is the $(p + q + r + 1)$-dimensional vector given by

$$S_n(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} \left(\frac{y_t}{\lambda_t(\boldsymbol{\theta})} - 1 \right) \frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{2.2}$$

The vector of partial derivatives $\partial \lambda_t(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ can be computed recursively by the recursions given in Appendix A.2. Finally, the conditional information matrix is given by

$$\begin{aligned} G_n(\boldsymbol{\theta}; \sigma^2) &= \sum_{t=1}^{n} \mathsf{COV} \left(\frac{\partial \ell(\boldsymbol{\theta}; Y_t)}{\partial \boldsymbol{\theta}} \middle| \mathcal{F}_{t-1} \right) \\ &= \sum_{t=1}^{n} \left(\frac{1}{\lambda_t(\boldsymbol{\theta})} + \sigma^2 \right) \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^\top . \end{aligned} \tag{2.3}$$

In the case of the Poisson assumption it holds $\sigma^2 = 0$ and in the case of the Negative Binomial assumption $\sigma^2 = 1/\phi$. For the ease of notation let $G_n^*(\boldsymbol{\theta}) = G_n(\boldsymbol{\theta}; 0)$, which is the conditional information matrix in case of a Poisson distribution.

The quasi maximum likelihood estimator (QMLE) $\widehat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$, assuming that it exists, is the solution of the non-linear constrained optimization problem

$$\widehat{\boldsymbol{\theta}} := \widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}). \tag{2.4}$$

Denote the fitted values by $\widehat{\lambda}_t = \lambda_t(\widehat{\theta})$. Following Christou and Fokianos (2014), the dispersion parameter ϕ of the Negative Binomial distribution is estimated by solving the equation

$$\sum_{t=1}^{n} \frac{(Y_t - \widehat{\lambda}_t)^2}{\widehat{\lambda}_t + \widehat{\lambda}_t^2 / \widehat{\phi})} = n - (p + q + r + 1), \tag{2.5}$$

which is based on Pearson's χ^2 statistic. The variance parameter σ^2 is estimated by $\widehat{\sigma}^2 = 1/\widehat{\phi}$. For the Poisson distribution we set $\widehat{\sigma}^2 = 0$. Strictly speaking, the log-linear model (1.3) does not fall into the class of models considered by Christou and Fokianos (2014). However, results obtained by Douc *et al.* (2013) (for $p = q = 1$) and Sim (2016) (for $p = q$) allow us to use this estimator also for the log-linear model. This issue is addressed by simulations in Appendix B.2, which support that the estimator obtained by (2.5) provides good results also for models with the logarithmic link function.

2.2.2 Inference

Inference for the regression parameters is based on the asymptotic normality of the QMLE, which has been studied by Fokianos *et al.* (2009) and Christou and Fokianos (2014) for models without covariates. For a well behaved covariate process $\{\boldsymbol{X}_t\}$ we conjecture that

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{d} N_{p+q+r+1}\left(\mathbf{0}, G_n^{-1}(\widehat{\boldsymbol{\theta}}_n; \widehat{\sigma}^2) G_n^{*}(\widehat{\boldsymbol{\theta}}_n) G_n^{-1}(\widehat{\boldsymbol{\theta}}_n; \widehat{\sigma}^2)\right), \tag{2.6}$$

as $n \to \infty$, where $\boldsymbol{\theta}_0$ denotes the true parameter value and $\widehat{\sigma}^2$ is a consistent estimator of σ^2. We suppose that this applies under the same assumptions usually made for the ordinary linear regression model (see for example Demidenko, 2013, p. 140 ff.). For deterministic covariates these assumptions are $||\boldsymbol{X}_t|| < c$, where $||\cdot||$ denotes the usual Euclidean norm, i.e., the covariate process is bounded, and $\lim_{n\to\infty} n^{-1}\sum_{t=1}^{n} \boldsymbol{X}_t\boldsymbol{X}_t^{\top} = A$, where c is a constant and A is a nonsingular matrix. For stochastic covariates it is assumed that the expectations $\mathsf{E}\left(\boldsymbol{X}_t\right)$ and $\mathsf{E}\left(\boldsymbol{X}_t\boldsymbol{X}_t^{\top}\right)$ exist and that $\mathsf{E}\left(\boldsymbol{X}_t\boldsymbol{X}_t^{\top}\right)$ is nonsingular. The assumptions imply that the information on each covariate grows linearly with the sample size and that the covariates are not linearly dependent. Fuller (1996, Theorem 9.1.1) shows asymptotic normality of the least squares estimator for a regression model with time series errors under even more general conditions which allow the presence of certain types of trends in the covariates. For the special case of a Poisson model with the identity link, Agosto, Cavaliere, Kristensen, and Rahbek (2015) show asymptotic normality of the MLE for a model with covariates that are functions of Markov processes with finite second moments and that are not collinearly related to the response. The asymptotic normality of the QMLE in our context is supported by the simulations presented in Appendix B.1. A formal proof requires further research. To avoid numerical instabilities when inverting $G_n(\widehat{\boldsymbol{\theta}}_n; \widehat{\sigma}^2)$ we apply an algorithm which makes use of the fact that it is a real symmetric and positive definite matrix; see Appendix A.4.

As an alternative method to the normal approximation (2.6) for obtaining standard errors and confidence intervals (function `se`) we include a parametric bootstrap procedure (argument `B`), for which computation time is many times higher. Accordingly, B time series are simulated from the model fitted to the original data. The empirical standard errors of the parameter estimates for these B time series are the bootstrap standard errors. Confidence intervals are based on quantiles of the bootstrap sample, see Efron and Tibshirani (1993, Chapter 13). This procedure can compute standard errors and confidence intervals both for $\widehat{\boldsymbol{\theta}}$ and $\widehat{\sigma}^2$. In our experience $B = 500$ yields stable results.

2.2.3 Implementation

This section and Appendix A provide some details on the implementation of the function `tsglm` and explain its technical arguments. The default settings of this arguments are chosen wisely based on plenty of experiments and should be sufficient for most situations though.

The parameter restrictions which are imposed by the condition $\boldsymbol{\theta} \in \Theta$ can be formulated as d linear inequalities. This means that there exists a matrix $\boldsymbol{U}$ of dimension $d \times (p+q+r+1)$ and a vector $\boldsymbol{c}$ of length d, such that $\Theta = \{\boldsymbol{\theta} \mid \boldsymbol{U}\boldsymbol{\theta} \geq \boldsymbol{c}\}$. For the linear model (1.2) one needs $d = p + q + r + 2$ constraints to ensure nonnegativity of the conditional mean λ_t and stationarity of the resulting process. For the log-linear model (1.3) there are not any constraints on the intercept term and on the covariate coefficients; hence $d = 2(p + q + 1)$. In order to enforce strict inequalities the respective constraints are tightened by an arbitrarily small constant $\xi > 0$; this constant is set to $\xi = 10^{-6}$ by default (argument `slackvar`).

For solving numerically the maximization problem (2.4) we employ by default the function `constrOptim`. This function applies an algorithm described by Lange (1999, Chapter 14), which essentially enforces the constraints by adding a barrier value to the objective function and then employs an algorithm for unconstrained optimization of this new objective function, iterating these two steps if necessary. By default the quasi-Newton Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm is employed for the latter task of unconstrained optimization, which additionally makes use of the score vector (2.2). It is possible to tune the optimization algorithm and even to employ an unconstrained optimization (argument `final.control`).

Note that the log-likelihood (2.1) and the score (2.2) are given conditional on unobserved pre-sample values. They depend on the linear predictor and its partial derivatives, which can be computed recursively using any initialization. We give the recursions and present several strategies for their initialization in Appendix A.2 (arguments `init.method` and `init.drop`). Christou and Fokianos (2014, Remark 3.1) show that the effect of the initialization vanishes asymptotically. Nevertheless, from a practical point of view the initialization of the recursions is crucial. Especially in the presence of strong serial dependence, the resulting estimates can differ substantially even for long time series with 1000 observations; see the simulated example in Table A.1 in Appendix A.2.

Solving the non-linear optimization problem (2.4) requires a starting value for the parameter vector $\boldsymbol{\theta}$. This starting value can be obtained from fitting a simpler model for which an estimation procedure is readily available. We consider either to fit a GLM or to fit an ARMA model. A third possibility is to fit a naive i.i.d. model without covariates. Furthermore, the user can assign fixed values. All these possibilities are available by the argument `start.control`. It turns out that the optimization algorithm converges very reliably even if the starting values are not close to the global optimum of the likelihood. A starting value which is closer to the global optimum usually requires fewer iterations until convergence. However, we have encountered some examples where starting values close to a local optimum, obtained by one of the first two aforementioned methods, do not yield the global optimum. Consequently, we recommend fitting the naive i.i.d. model without covariates to obtain starting values. More details on these approaches are given in Appendix A.3.

2.3 Prediction

In terms of the mean square error, the optimal 1-step-ahead predictor $\widehat{Y}_{n+1}$ for Y_{n+1}, given $\mathcal{F}_n$, i.e., the past of the process up to time n and potential covariates at time $n+1$, is the conditional expectation λ_{n+1} given in (1.1) (`S3` method of function `predict`). By construction of the model the conditional distribution of $\widehat{Y}_{n+1}$ is a Poisson (1.5) respectively Negative Binomial (1.6) distribution with mean λ_{n+1}. An h-step-ahead prediction $\widehat{Y}_{n+h}$ for Y_{n+h} is obtained by recursive 1-step-ahead predictions, where unobserved values $Y_{n+1}, \ldots, Y_{n+h-1}$ are replaced by their respective 1-step-ahead prediction, $h \in \mathbb{N}$. The distribution of this h-step-ahead prediction $\widehat{Y}_{n+h}$ is not known analytically but can be approximated numerically by a parametric bootstrap procedure, which is described below.

In applications, λ_{n+1} is substituted by its estimator $\widehat{\lambda}_{n+1} = \lambda_{n+1}(\widehat{\boldsymbol{\theta}})$, which depends on the estimated regression parameters $\widehat{\boldsymbol{\theta}}$. The dispersion parameter ϕ of the Negative Binomial distribution is replaced by its estimator $\widehat{\phi}$. Note that plugging in the estimated parameters induces additional uncertainty to the predictive distribution. This estimation uncertainty is not taken into account for the construction of prediction intervals described in the following paragraphs.

Prediction intervals for Y_{n+h} with a given coverage rate $1-\alpha$ (argument `level`) are designed to cover the true observation Y_{n+h} with a probability of $1-\alpha$. Simultaneous prediction intervals achieving a global coverage rate for $Y_{n+1},\ldots,Y_{n+h}$ can be obtained by a Bonferroni adjustment of the individual coverage rates to $1-\alpha/h$ each (argument `global = TRUE`).

There are two different principles for constructing predictions intervals available which in practice often yield identical intervals. Firstly, the limits can be the $(\alpha/2)$- and $(1-\alpha/2)$-quantile of the (approximated) predictive distribution (argument `type = "quantiles"`). Secondly, the limits can be chosen such that the interval has minimal length given that, according to the (approximated) predictive distribution, the probability that a value falls into this interval is at least as large as the desired coverage rate $1-\alpha$ (argument `type = "shortest"`).

One-step-ahead prediction intervals can be straightforwardly obtained from the conditional distribution (argument `method = "conddistr"`). Prediction intervals obtained by a parametric bootstrap procedure (argument `method = "bootstrap"`) are based on B simulations of realizations $y_{n+1}^{(b)},\ldots,y_{n+h}^{(b)}$ from the fitted model, $b=1,\ldots,B$ (argument `B`). To obtain an approximative prediction interval for Y_{n+h} one can either use the empirical $(\alpha/2)$- and $(1-\alpha/2)$-quantile of $y_{n+h}^{(1)},\ldots,y_{n+h}^{(B)}$ (if `type = "quantiles"`) or find the shortest interval which contains at least $\lceil(1-\alpha)\cdot B\rceil$ of these observations (if `type = "shortest"`). This bootstrap procedure can be accelerated by distributing it to multiple cores simultaneously (argument `parallel = TRUE`), which requires a computing cluster registered by the R package **parallel** (see the help page of the function `setDefaultCluster`).

2.4 Model assessment

Tools originally developed for generalized linear models as well as for time series can be utilized to asses the model fit and its predictive performance. Within the class of count time series following generalized linear models it is desirable to asses the specification of the linear predictor as well as the choice of the link function and of the conditional distribution. The tools presented in this section facilitate the selection of an adequate model for a given data set. Note that all tools are introduced as in-sample versions, meaning that the observations $y_1 \dots, y_n$ are used for fitting the model as well as for assessing the obtained fit. However, it is straightforward to apply such tools as out-of-sample criteria.

Recall that the fitted values are denoted by $\widehat{\lambda}_t = \lambda_t(\widehat{\theta})$. Note that these do not depend on the chosen distribution, because the mean is the same regardless of the response distribution. There are various types of *residuals* available (`S3` method of function `residuals`). Response (or raw) residuals (argument `type = "response"`) are given by

$$r_t = y_t - \widehat{\lambda}_t, \tag{2.7}$$

whereas a standardized alternative are Pearson residuals (argument `type = "pearson"`)

$$r_t^P = (y_t - \widehat{\lambda}_t)/\sqrt{\widehat{\lambda}_t + \widehat{\lambda}_t^2 \widehat{\sigma}^2}, \tag{2.8}$$

or the more symmetrically distributed standardized Anscombe residuals (argument `type = "anscombe"`)

$$r_t^A = \frac{3/\widehat{\sigma}^2 \left(\left(1 + y_t \widehat{\sigma}^2\right)^{2/3} - \left(1 + \widehat{\lambda}_t \widehat{\sigma}^2\right)^{2/3} \right) + 3\left(y_t^{2/3} - \widehat{\lambda}_t^{2/3}\right)}{2\left(\widehat{\lambda}_t + \widehat{\lambda}_t^2 \widehat{\sigma}^2\right)^{1/6}}, \tag{2.9}$$

for $t = 1, \dots, n$ (see for example Hilbe, 2011, Section 5.1). The empirical autocorrelation function of these residuals is useful for diagnosing serial dependence which has not been explained by the fitted model. A plot of the residuals against time can reveal changes of the data generating process over time. Furthermore, a plot of squared residuals r_t^2 against the corresponding fitted values $\widehat{\lambda}_t$ exhibits the relation of mean and variance and might point to the Poisson distribution if the points scatter around the identity function or to the Negative Binomial distribution if there exists a quadratic relation (see Ver Hoef and Boveng, 2007).

Christou and Fokianos (2015b) and Jung and Tremayne (2011) extend tools for assessing the predictive performance to count time series, which were originally proposed by Gneiting, Balabdaoui, and Raftery (2007) and others for continuous data and transferred to independent but not identically distributed count data by Czado, Gneiting, and Held (2009). These tools follow the *prequential principle* formulated by Dawid (1984), depending only on the realized observations and their respective forecast distributions. Denote by $P_t(y) = P\big(Y_t \leq y | \mathcal{F}_{t-1}\big)$ the cumulative distribution function (c.d.f.), by $p_t(y) = P\big(Y_t = y | \mathcal{F}_{t-1}\big)$ the probability density function, $y \in \mathbb{N}_0$, and by $\upsilon_t = \sqrt{\mathsf{VAR}\left(Y_t | \mathcal{F}_{t-1}\right)}$ the standard deviation of the predictive distribution, which is either a Poisson distribution with mean $\widehat{\lambda}_t$ or a Negative Binomial distribution with mean $\widehat{\lambda}_t$ and overdispersion coefficient $\widehat{\sigma}^2$ (recall Section 2.3 on 1-step-ahead prediction).

A tool for assessing the probabilistic calibration of the predictive distribution (see Gneiting *et al.*, 2007) is the *probability integral transform* (PIT), which will follow a uniform distribution if the predictive distribution is correct. For count data Czado *et al.* (2009) define a non-randomized PIT value for the observed value y_t and the predictive distribution $P_t(y)$ by

$$F_t(u|y) = \begin{cases} 0, & u \leq P_t(y-1) \\ \dfrac{u - P_t(y-1)}{P_t(y) - P_t(y-1)}, & P_t(y-1) < u < P_t(y) \\ 1, & u \geq P_t(y) \end{cases} .$$

The mean PIT is then given by

$$\overline{F}(u) = \frac{1}{n} \sum_{t=1}^{n} F_t(u|y_t), \quad 0 \leq u \leq 1.$$

To check whether $\overline{F}(u)$ is the c.d.f. of a uniform distribution Czado *et al.* (2009) propose plotting a histogram with H bins, where bin h has the height $f_j = \overline{F}(h/H) - \overline{F}((h-1)/H)$, $h = 1, \ldots, H$ (function `pit`). By default H is chosen to be 10. A U-shape indicates underdispersion of the predictive distribution, whereas an upside down U-shape indicates overdispersion. Gneiting *et al.* (2007) point out that the empirical coverage of central, e.g., 90% prediction intervals can be read off the PIT histogram as the area under the 90% central bins.

Marginal calibration is defined as the difference of the average predictive c.d.f. and the empirical c.d.f. of the observations, i.e.,

$$\frac{1}{n} \sum_{t=1}^{n} P_t(y) - \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}(y_t \leq y) \tag{2.10}$$

for all $y \in \mathbb{R}$. In practice we plot the marginal calibration for values y in the range of the original observations (Christou and Fokianos, 2015b) (function `marcal`). If the predictions from a model are appropriate the marginal distribution of the predictions resembles the marginal distribution of the observations and (2.10) should be close to zero. Major deviations from zero point to model deficiencies.

Gneiting *et al.* (2007) show that the calibration assessed by a PIT histogram or a marginal calibration plot is a necessary but not sufficient condition for a forecaster to be ideal.

Scoring rule	Abbreviation	Definition
squared error score	`sqerror`	$(y_t - \lambda_t)^2$
normalized squared error score	`normsq`	$(y_t - \lambda_t)^2/\upsilon_t^2$
Dawid-Sebastiani score	`dawseb`	$(y_t - \lambda_t)^2/\upsilon_t^2 + 2\log(\upsilon_t)$
logarithmic score	`logarithmic`	$-\log(p_t(y_t))$
quadratic (or Brier) score	`quadratic`	$-2p_t(y_t) + \lVert p_t \rVert^2$
spherical score	`spherical`	$-p_t(y_t)/\lVert p_t \rVert$
ranked probability score	`rankprob`	$\sum_{y=0}^{\infty} \left(P_t(y) - \mathbb{1}(y_t \leq y)\right)^2$

Table 2.1: Definitions of proper scoring rules $s(P_t, y_t)$ (cf. Czado *et al.*, 2009) and their abbreviations in the package; $\lVert p_t \rVert^2 = \sum_{y=0}^{\infty} p_t^2(y)$.

They advocate to favor the model with the maximal sharpness among all sufficiently calibrated models. Sharpness is the concentration of the predictive distribution and can be measured by the width of prediction intervals. A simultaneous assessment of calibration and sharpness summarized in a single numerical score can be accomplished by *proper scoring rules* (Gneiting *et al.*, 2007). Denote a score for the predictive distribution P_t and the observation y_t by $s(P_t, y_t)$. A number of possible proper scoring rules is given in Table 2.1. The mean score for each corresponding model is given by $\sum_{t=1}^{n} s(P_t, y_t)/n$. Each of the different proper scoring rules captures different characteristics of the predictive distribution and its distance to the observed data (function `scoring`). Except for the normalized error score, the model with the lowest score is preferable. The mean squared error score is the only one which does not depend on the distribution and is also known as mean squared prediction error. The mean normalized squared error score measures the variance of the Pearson residuals and is close to one if the model is adequate. The Dawid-Sebastini score is a variant of this with an extra term to penalize overerstimation of the standard deviation.

Other popular tools are model selection criteria like Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) (functions `AIC` and `BIC`). The model with the lowest value of the respective information criterion is preferable. Denote the log-likelihood by $\widetilde{\ell}(\widehat{\boldsymbol{\theta}}, \widehat{\sigma}^2) = \sum_{t=1}^{n} \log\left(p_t(y_t)\right)$. Note that this is the true and not the quasi log-likelihood given in (2.1). Furthermore, $\widetilde{\ell}(\widehat{\boldsymbol{\theta}}, \widehat{\sigma}^2)$ includes all constant terms which have been omitted on the right hand side of (2.1). The AIC and BIC are given by $\text{AIC} = -2\widetilde{\ell}(\widehat{\boldsymbol{\theta}}, \widehat{\sigma}^2) + 2df$ and $\text{BIC} = -2\widetilde{\ell}(\widehat{\boldsymbol{\theta}}, \widehat{\sigma}^2) + \log(n_{\text{eff}})df$, respectively. Here df is the total number of parameters (including the dispersion coefficient) and n_{eff} the number of effective observations (excluding those only used for initialization when argument `init.drop = TRUE`). The BIC generally yields more parsimonious models than the AIC. Note that for other distributions than the Poisson, $\widehat{\boldsymbol{\theta}}$ maximizes the quasi log-likelihood

(2.1) but not $\widetilde{\ell}(\boldsymbol{\theta}, \sigma^2)$. In such cases the quasi information criterion (QIC), proposed by Pan (2001) for regression analysis based on the generalized estimating equations, is a properly adjusted alternative to the AIC (function `QIC`). We have verified by a simulation reported in Appendix B.3 that in case of a Poisson distribution the QIC approximates the AIC quite satisfactory.

2.5 Intervention analysis

In many applications sudden changes or extraordinary events occur. Box and Tiao (1975) refer to such special events as interventions. This could be for example the outbreak of an epidemic in a time series which counts the weekly number of patients infected with a particular disease. It is of interest to examine the effect of known interventions, for example to judge whether a policy change had the intended impact, or to search for unknown intervention effects and find explanations for them *a posteriori*.

Fokianos and Fried (2010, 2012) model interventions affecting the location by including a deterministic covariate of the form $\delta^{t-\tau}\mathbb{1}(t \geq \tau)$, where τ is the time of occurrence and the decay rate δ is a known constant (function `interv_covariate`). This covers various types of interventions for different choices of the constant δ: a singular effect for $\delta = 0$ (spiky outlier), an exponentially decaying change in location for $\delta \in (0, 1)$ (transient shift) and a permanent change of location for $\delta = 1$ (level shift). Similar to the case of covariates, the effect of an intervention is essentially additive for the linear model and multiplicative for the log-linear model. However, the intervention enters the dynamics of the process and therefore its effect on the linear predictor is not purely additive. Our package includes methods to test for such intervention effects developed by Fokianos and Fried (2010, 2012), suitably adapted to the more general model class described in Section 1.2. The linear predictor of a model with s types of interventions according to parameters $\delta_1, \ldots, \delta_s$ occurring at time points $\tau_1, \ldots, \tau_s$ reads

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \, \widetilde{g}(Y_{t-i_k}) + \sum_{\ell=1}^{q} \alpha_\ell g(\lambda_{t-j_\ell}) + \boldsymbol{\eta}^\top \boldsymbol{X}_t + \sum_{m=1}^{s} \omega_m \delta_m^{t-\tau_m} \mathbb{1}(t \geq \tau_m), \quad (2.11)$$

where ω_m, $m = 1, \ldots, s$ are the intervention sizes. At the time of its occurrence an intervention changes the level of the time series by adding the magnitude ω_m, for a linear model like (1.2), or by multiplying the factor $\exp(\omega_m)$, for a log-linear model like (1.3). In the following paragraphs we briefly outline the proposed intervention detection procedures and refer to Chapter 3 and to the original articles for details.

Our package allows to test whether s interventions of certain types occurring at given time points, according to model (2.11), have an effect on the observed time series, i.e., to test the hypothesis $H_0 : \omega_1 = \ldots = \omega_s = 0$ against the alternative $H_1 : \omega_\ell \neq 0$ for some $\ell \in \{1, \ldots, s\}$. This is accomplished by employing an approximate score test (function `interv_test`). Under the null hypothesis the score test statistic $T_n(\tau_1, \ldots, \tau_s)$ has asymptotically a χ^2-distribution with s degrees of freedom, assuming some regularity conditions (Fokianos and Fried, 2010, Lemma 1).

For testing whether a single intervention of a certain type occurring at an unknown time point τ has an effect, the package employs the maximum of the score test statistics $T_n(\tau)$ and determines a p value by a parametric bootstrap procedure (function `interv_detect`). If we consider a set D of time points at which the intervention might occur, e.g., $D = \{2, \ldots, n\}$, this test statistic is given by $\widetilde{T}_n = \max_{\tau \in D} T_n(\tau)$. The bootstrap procedure can be computed on multiple cores simultaneously (argument `parallel = TRUE`). The time point of the intervention is estimated to be the value τ which maximizes this test statistic. Our empirical observation is that such an estimator usually has a large variability. It is possible to speed up the computation of the bootstrap test statistics by using the model parameters used for generation of the bootstrap samples instead of estimating them for each bootstrap sample (argument `final.control_bootstrap = NULL`). This results in a conservative procedure, as noted by Fokianos and Fried (2012).

If more than one intervention is suspected in the data, but neither their types nor the time points of its occurrences are known, an iterative detection procedure is used (function `interv_multiple`). Consider the set of possible intervention times D as before and a set of possible intervention types Δ, e.g., $\Delta = \{0, 0.8, 1\}$. In a first step the time series is tested for an intervention of each type $\delta \in \Delta$ as described in the previous paragraph and the p values are corrected to account for the multiple testing by the Bonferroni method. If none of the p values is below a previously specified significance level, the procedure stops and does not identify an intervention effect. Otherwise the procedure detects an intervention of the type corresponding to the lowest p value. In case of equal p values preference is given to interventions with $\delta = 1$, that is level shifts, and then to those with the largest test statistic. In a second step, the effect of the detected intervention is eliminated from the time series and the procedures starts anew and continues until no further intervention effects are detected. Finally, model (2.11) with all detected intervention effects can be fitted to the data to estimate the intervention sizes and the other parameters jointly (which are in general different than when estimated in separate steps). Note that statistical inference for this final model fit has to be done with care.

In practical applications, the decay rate δ of a particular intervention effect is often unknown and needs to be estimated. Since the parameter δ is not identifiable when the corresponding intervention size ω is zero, its estimation is nonstandard. As suggested by a reviewer of the *Journal of Statistical Software*, estimation could be carried out by profiling the likelihood over this parameter. For a single intervention effect this could be done by computing the (quasi) ML estimator of all other parameters for a given decay rate δ. This is repeated for all $\delta \in \Delta$, where Δ is a set of possible decay rates, and the value which results in the maximum value of the log-likelihood is chosen (apply the function `tsglm` repeatedly). Note that this approach affects the validity of the usual statistical inference for the other parameters.

Chapter 3 (based on Liboschik *et al.*, 2016) studies a model for external intervention effects (modeled by external covariate effects, recall (1.4) and the related discussion) and compare it to internal intervention effects studied in the two aforementioned publications (argument `external`).

2.6 Usage of the package

The most recent stable version of the **tscount** package is distributed via the Comprehensive R Archive Network (CRAN). A current development version is available from the project's website `http://tscount.r-forge.r-project.org` on the development platform R-Forge. After installation of the package it can be loaded in R by typing `library("tscount")`.

The central function for fitting a GLM for count time series is `tsglm`, whose help page (accessible by `?tsglm`) is a good starting point to become familiar with the usage of the package. The most relevant functions of the package are summarized in Table 2.2. There are many standard `S3` methods available for well-known generic functions. A detailed description of the functions' usage including examples can be found on the accompanying help pages. There is also a number of auxiliary functions which are not intended to be called by the average user. In total the package currently consists of about 1600 lines of code and a manual of more than forty pages. The package provides some data sets which are also listed in Table 2.2.

In the following sections we demonstrate typical applications of the package by two data examples.

	Name	Description
Functions	`tsglm`	Fitting a model to given data (class `"tsglm"`)
	`tsglm.sim`	Simulating from the model
	Generic functions with methods for class `"tsglm"`:	
	`plot`	Diagnostic plots
	`se`	Standard errors and confidence intervals
	`summary`	Summary of the fitted model
	`fitted`	Fitted values
	`residuals`	Residuals
	`AIC`	Akaike's information criterion
	`BIC`	Bayesian information criterion
	`QIC`	Quasi information criterion
	`pit`	Probability integral transform histogram
	`marcal`	Marginal calibration plot
	`scoring`	Proper scoring rules
	`predict`	Prediction
	`interv_test`	Test for intervention effects
	`interv_detect`	Detection of single intervention effects
	`interv_multiple`	Iterative detection of multiple intervention effects
Data sets	`campy`	Campylobacter infections in Québec
	`ecoli`	E. coli infections in North Rhine-Westphalia (NRW)
	`ehec`	EHEC/HUS infections in NRW
	`influenza`	Influenza infections in NRW
	`measles`	Measles infections in NRW

Table 2.2: Most important functions of the R package **tscount** and the included data sets.

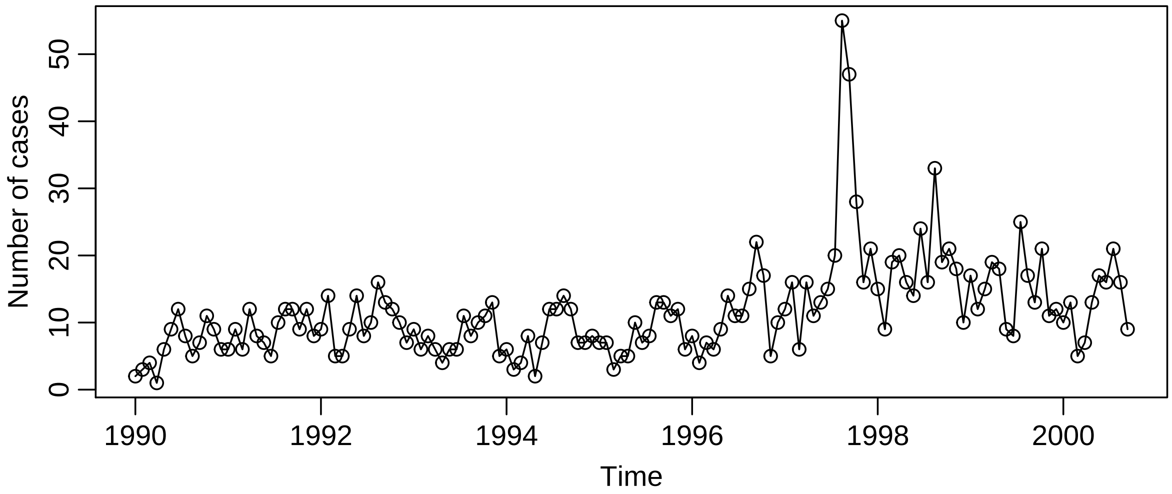


Figure 2.1: Number of campylobacterosis cases (reported every 28 days) in the North of Québec in Canada.

2.6.1 Campylobacter infections in Canada

We first analyze the number of campylobacterosis cases (reported every 28 days) in the North of Québec in Canada. The data are shown in Figure 2.1 and were first reported by Ferland *et al.* (2006). These data are made available in the package (object `campy`). We fit a model to this time series using the function `tsglm`. Following the analysis of Ferland *et al.* (2006) we fit model (1.2) with the identity link function, defined by the argument `link`. For taking into account serial dependence we include a regression on the previous observation. Seasonality is captured by regressing on λ_{t-13}, the unobserved conditional mean 13 time units (which is about one year) back in time. The aforementioned specification of the model for the linear predictor is assigned by the argument `model`, which has to be a list. We also include the two intervention effects detected by Fokianos and Fried (2010) in the model by suitably chosen covariates provided by the argument `xreg`, see also Section 3.5. We compare a fit of a Poisson with that of a Negative Binomial conditional distribution, specified by the argument `distr`. The call for both model fits is then given by:

```
R> interventions <- interv_covariate(n = length(campy), tau = c(84, 100),
+                       delta = c(1, 0))
R> campyfit_pois <- tsglm(campy, model = list(past_obs = 1, past_mean = 13),
+                       xreg = interventions, distr = "poisson")
R> campyfit_nbin <- tsglm(campy, model = list(past_obs = 1, past_mean = 13),
+                       xreg = interventions, distr = "nbinom")
```

The resulting fitted models `campyfit_pois` and `campyfit_nbin` have class `"tsglm"`, for which a number of methods is provided (see help page), including `summary` for a detailed model summary and `plot` for diagnostic plots. The diagnostic plots like in Figure 2.2 can be produced by:

```
R> acf(residuals(campyfit_pois), main = "ACF of response residuals")
R> marcal(campyfit_pois, ylim = c(-0.03, 0.03), main = "Marginal calibration")
R>   lines(marcal(campyfit_nbin, plot = FALSE), lty = "dashed")
R>   legend("bottomright", legend = c("Pois", "NegBin"), lwd = 1,
+          lty = c("solid", "dashed"))
R> pit(campyfit_pois, ylim = c(0, 1.5), main = "PIT Poisson")
R> pit(campyfit_nbin, ylim = c(0, 1.5), main = "PIT Negative Binomial")
```

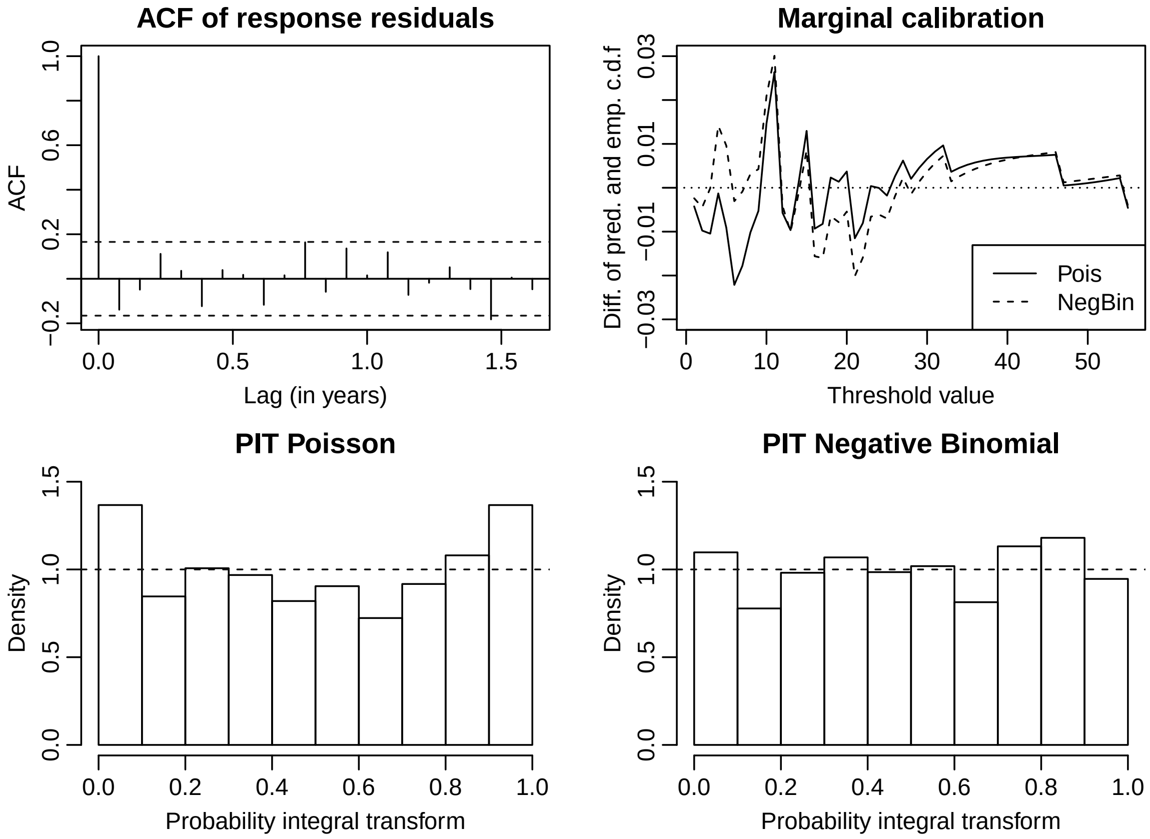


Figure 2.2: Diagnostic plots after model fitting to the campylobacterosis data.

The response residuals are identical for the two conditional distributions. Their empirical autocorrelation function, shown in Figure 2.2 (top left), does not exhibit any serial correlation or seasonality which has not been taken into account by the models. Figure 2.2 (bottom left) points to an approximately U-shaped PIT histogram indicating that the Poisson distribution is not adequate for model fitting. As opposed to this, the PIT histogram which corresponds to the Negative Binomial distribution appears to approach uniformity better. Hence the probabilistic calibration of the Negative Binomial model is satisfactory. The marginal calibration plot, shown in Figure 2.2 (top right), is inconclusive. As a last tool we consider the scoring rules for the two distributions:

```
R> rbind(Poisson = scoring(campyfit_pois), NegBin = scoring(campyfit_nbin))
        logarithmic quadratic spherical rankprob dawseb normsq sqerror
Poisson       2.750  -0.07669   -0.2751    2.200  3.662 1.3081   16.51
NegBin        2.722  -0.07800   -0.2766    2.185  3.606 0.9643   16.51
```

All considered scoring rules are in favor of the Negative Binomial distribution. Based on the PIT histograms and the results obtained by the scoring rules, we decide for the Negative Binomial model. The degree of overdispersion seems to be small, as the estimated overdispersion coefficient `sigmasq` of 0.0297 given in the output below is close to zero.

```
R> summary(campyfit_nbin)

Call:
tsglm(ts = campy, model = list(past_obs = 1, past_mean = 13),
    xreg = interventions, distr = "nbinom")

Coefficients:
             Estimate  Std.Error  CI(lower)  CI(upper)
(Intercept)    3.3184     0.7851     1.7797      4.857
beta_1         0.3690     0.0696     0.2326      0.505
alpha_13       0.2198     0.0942     0.0352      0.404
interv_1       3.0810     0.8560     1.4032      4.759
interv_2      41.9541    12.0914    18.2554     65.653
sigmasq        0.0297         NA         NA         NA
Standard errors and confidence intervals (level =  95 %) obtained
by normal approximation.

Link function: identity
Distribution family: nbinom (with overdispersion coefficient 'sigmasq')
Number of coefficients: 6
Log-likelihood: -381.1
AIC: 774.2
BIC: 791.8
QIC: 787.6
```

The coefficient `beta_1` corresponds to regression on the previous observation, `alpha_13` corresponds to regression on values of the conditional mean thirteen units back in time. The output reports the estimation of the overdispersion coefficient σ^2, which is related to the dispersion parameter ϕ of the Negative Binomial distribution by $\phi = 1/\sigma^2$. Accordingly, the fitted model for the number of new infections Y_t in time period t is given by $Y_t|\mathcal{F}_{t-1} \sim \text{NegBin}(\lambda_t, 33.61)$ with

$$\lambda_t = 3.32 + 0.37Y_{t-1} + 0.22\lambda_{t-13} + 3.08\mathbb{1}(t = 84) + 41.95\mathbb{1}(t \geq 100), \quad t = 1, \ldots, 140.$$

The standard errors of the estimated regression parameters and the corresponding confidence intervals in the summary above are based on the normal approximation given in (2.6). For the additional overdispersion coefficient `sigmasq` of the Negative Binomial distribution there is no analytical approximation available for its standard error. Alternatively, standard errors (and confidence intervals, not shown here) of the regression parameters and the overdispersion coefficient can be obtained by a parametric bootstrap (which takes about 15 minutes computation time on a single 3.2 GHz processor for 500 replications):

```
R> se(campyfit_nbin, B = 500)$se

(Intercept)      beta_1    alpha_13    interv_1    interv_2     sigmasq
    0.89850     0.06941     0.10136     0.93836    11.16856     0.01460
Warning message:
In se.tsglm(campyfit_nbin, B = 500) :
  The overdispersion coefficient 'sigmasq' could not be estimated
in 5 of the 500 replications. It is set to zero for these
replications. This might to some extent result in a biased estimation
of its true variability.
```

Estimation problems for the dispersion parameter (see warning message) occur occasionally for models where the true overdispersion coefficient σ^2 is small, i.e., which are close to a Poisson model; see Appendix B.2. The bootstrap standard errors of the regression parameters are slightly larger than those based on the normal approximation. Note that neither of the approaches reflects the additional uncertainty induced by the model selection.

2.6.2 Road casualties in Great Britain

Next we study the monthly number of killed drivers of light goods vehicles in Great Britain between January 1969 and December 1984 shown in Figure 2.3. This time series is part of a dataset which was first considered by Harvey and Durbin (1986) for studying the effect of compulsory wearing of seatbelts introduced on 31 January 1983. The dataset, including additional covariates, is available in R in the object `Seatbelts`. In their paper Harvey and Durbin (1986) analyze the numbers of casualties for drivers and passengers of cars, which are so large that they can be treated with methods for continuous-valued data. The monthly number of killed drivers of vans analyzed here is much smaller (its minimum is 2 and its maximum 17) and therefore methods for count data are to be preferred.

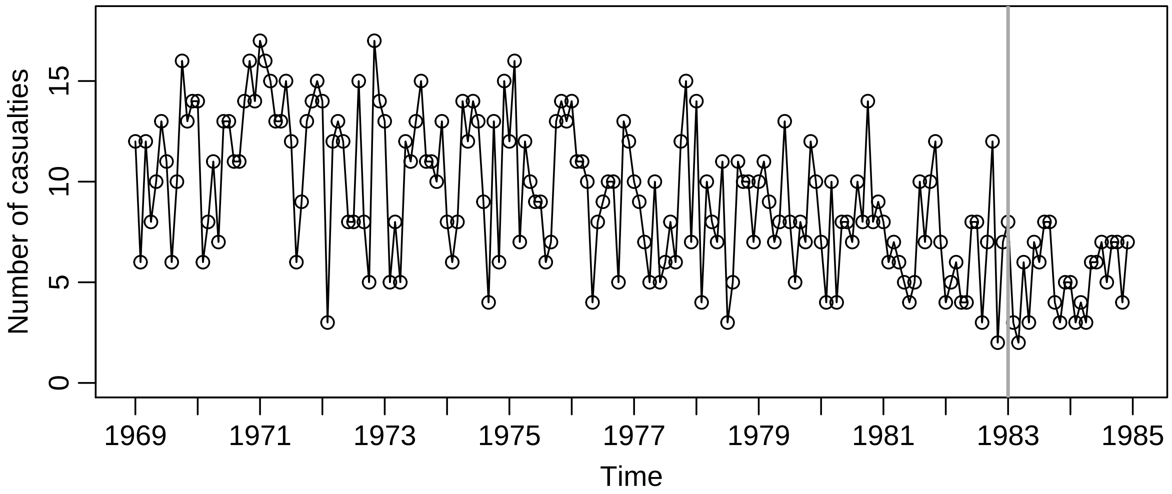


Figure 2.3: Monthly number of killed van drivers in Great Britain. The introduction of compulsory wearing of seatbelts on 31 January 1983 is marked by a vertical line.

For model selection we only use the data until December 1981. We choose the log-linear model with the logarithmic link because it allows for negative covariate effects. We aim at capturing the short range serial dependence by a first order autoregressive term and the yearly seasonality by a 12th order autoregressive term. Both of these terms are declared by the list element named `past_obs` of the argument `model`. Following Harvey and Durbin (1986) we use the real price of petrol as an explanatory variable. We also include a deterministic covariate describing a linear trend. Both covariates are provided by the argument `xreg`. Based on PIT histograms, a marginal calibration plot and the scoring rules (not shown here) we find that the Poisson distribution is sufficient for modeling. The model is fitted by the call:

```
R> timeseries <- Seatbelts[, "VanKilled"]
R> regressors <- cbind(PetrolPrice = Seatbelts[, c("PetrolPrice")],
+                       linearTrend = seq(along = timeseries)/12)
R> timeseries_until1981 <- window(timeseries, end = 1981 + 11/12)
R> regressors_until1981 <- window(regressors, end = 1981 + 11/12)
R> seatbeltsfit <- tsglm(timeseries_until1981,
+    model = list(past_obs = c(1, 12)), link = "log", distr = "poisson",
+    xreg = regressors_until1981)

R> summary(seatbeltsfit, B = 500)


Call:
tsglm(ts = timeseries_until1981, model = list(past_obs = c(1,
    12)), xreg = regressors_until1981, link = "log", distr = "pois")
```

```
Coefficients:
             Estimate  Std.Error  CI(lower)  CI(upper)
(Intercept)    1.7872    0.39925     1.1927      2.727
beta_1         0.0854    0.08515    -0.1055      0.209
beta_12        0.1581    0.10082    -0.0334      0.314
PetrolPrice    1.1893    2.76888    -4.0278      6.427
linearTrend   -0.0306    0.00885    -0.0489     -0.016
Standard errors and confidence intervals (level =  95 %) obtained
by parametric bootstrap with 500 replications.

Link function: log
Distribution family: poisson
Number of coefficients: 5
Log-likelihood: -396.1
AIC: 802.2
BIC: 817.5
QIC: 802.2
```

Accordingly, the fitted model for the number of van drivers Y_t killed in month t is given by $Y_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t)$ with

$$\log(\lambda_t) = 1.9 + 0.09Y_{t-1} + 0.15Y_{t-12} + 0.08X_t - 0.03t/12, \quad t = 1, \ldots, 156,$$

where X_t denotes the real price of petrol at time t. The estimated coefficient `beta_1` corresponding to the first order autocorrelation is very small and even slightly below the size of its approximative standard error, indicating that there is no notable dependence on the number of killed van drivers of the preceding week. We find a seasonal effect captured by the twelfth order autocorrelation coefficient `beta_12`. Unlike in the model for the car drivers by Harvey and Durbin (1986), the petrol price does not seem to influence the number of killed van drivers. An explanation might be that vans are much more often used for commercial purposes than cars and that commercial traffic is less influenced by the price of fuel. The linear trend can be interpreted as a yearly reduction of the number of casualties by a factor of 0.97 (obtained by exponentiating the corresponding estimated coefficient), i.e., on average we expect 2.9% fewer killed van drivers per year (which is below one in absolute numbers).

Based on the model fitted to the training data until December 1981, we can predict the number of road casualties in 1982 given the respective petrol price. Coherent, i.e. integer-valued forecasts could be obtained by rounding the predictions. A graphical representation of the following predictions is given in Figure 2.4.

```
R> timeseries_1982 <- window(timeseries, start = 1982, end = 1982 + 11/12)
R> regressors_1982 <- window(regressors, start = 1982, end = 1982 + 11/12)
```

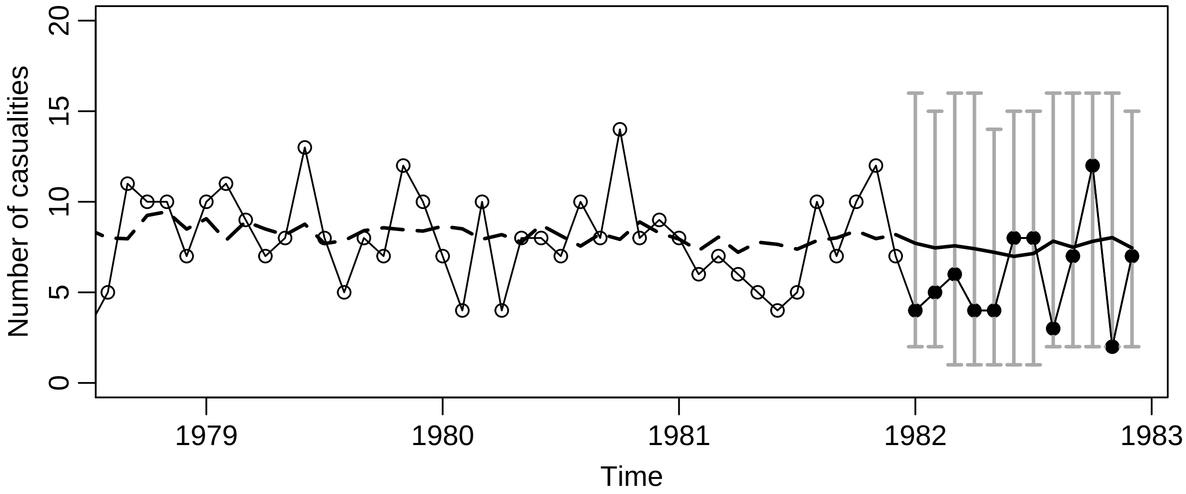


Figure 2.4: Fitted values (dashed line) and predicted values (solid line) according to the model with the Poisson distribution. Prediction intervals (grey bars) are designed to ensure a global coverage rate of 90%. They are chosen to have minimal length and are based on a simulation with 2000 replications.

```
R> predict(seatbeltsfit, n.ahead = 12, level = 0.9, global = TRUE,
+          B = 2000, newxreg = regressors_1982)$pred
      Jan  Feb  Mar  Apr  May  Jun  Jul  Aug  Sep  Oct  Nov  Dec
1982 7.71 7.45 7.57 7.41 7.21 6.99 7.15 7.83 7.49 7.82 8.02 7.45
```

Finally, we test whether there was an abrupt shift in the number of casualties occurring when the compulsory wearing of seatbelts is introduced on 31 January 1983. The approximative score test described in Section 2.5 is applied:

```
R> seatbeltsfit_alldata <- tsglm(timeseries, link = "log",
+                                model = list(past_obs = c(1, 12)),
+                                xreg = regressors, distr = "poisson")

R> interv_test(seatbeltsfit_alldata, tau = 170, delta = 1, est_interv = TRUE)


Score test on intervention(s) of given type at given time

Chisq-Statistic: 1.153 on 1 degree(s) of freedom, p-value: 0.2829

Fitted model with the specified intervention:

Call:
tsglm(ts = fit$ts, model = model_extended, xreg = xreg_extended,
    link = fit$link, distr = fit$distr)
```

```
Coefficients:
(Intercept)       beta_1      beta_12  PetrolPrice  linearTrend
    0.19508      0.08819      0.80446      3.17408     -0.04788
   interv_1
    0.24570
```

With a p value of 0.28 the null hypothesis of no intervention cannot be rejected at a 5% significance level. Note that this result does not rule out that there is an effect of the seatbelts law which is either too small for being significant or of a different type than it is tested for. For illustration we fit the model under the alternative of a level shift after the introduction of the seatbelts law (see the output above). The multiplicative effect size of the intervention is found to be 1.279. This indicates that according to this model fit 27.9% less van drivers are killed after the law enforcement. For comparison, Harvey and Durbin (1986) estimate a reduction of 18% for the number of killed car drivers.

2.7 Comparison with other software packages

In this section we review functions (and the corresponding models) from other R packages which can be employed for count time series analysis. Many of them have been published only very recently, a fact that demonstrates the raising interest in count time series analysis. We discuss how these packages differ from our package **tscount**. For illustration we use the time series of Campylobacter infections analyzed in Section 2.6.1 ignoring the intervention effects. For the presentation of other models we use a notation parallel to the one used in the previous sections to highlight similarities. Interpretation of the final model should be done carefully, though.

We consider a large number of somehow related packages which makes this comparison quite extensive yet interesting for those readers who want guidance on choosing the most appropriate package for their data. In the first subsection we present packages for independent data and in the second subsection we discuss packages for dependent data.

2.7.1 Packages for independent data

We start reviewing functions which have been introduced for independent observations but can, with certain limitations, be employed for time series whose temporal dependence is rather simple. This is exemplarily discussed in the following paragraph.

The function glm in the package **stats** and, for the Negative Binomial distribution, the function glm.nb in the package **MASS** (Venables and Ripley, 2002) can fit standard GLMs to count time series with the iteratively reweighted least squares (IRLS) algorithm. Just like with our tsglm function, one can choose the identity or logarithmic link in combination with a Poisson or Negative Binomial conditional distribution. Standard GLMs have been introduced to model independent but not identically distributed observations. In principle, one could also fit simple models for time series by including lagged values of the time series, i.e., $Y_{t-i_1}, \ldots, Y_{t-i_p}$, as covariates. However, the glm function has several limitations; the most important being that it does not allow for regression on past values of the conditional mean. For example, the glm function cannot be used to fit the model which included stochastic seasonality; recall Section 2.6.1. Furthermore, the glm function does not induce the constraints on the vector of parameters given in Section 2.2.1, which are necessary to ensure stationarity of the fitted process. Models which are violating these parameter constraints are generally not suitable for prediction. We have also experienced that glm occasionally does not find good starting values for its optimization procedure such that it returns an error and requests the user to provide starting values. At least for the very simple case of a Poisson INGARCH(1,0) model fitted to the Campylobacterosis data the glm function performs well and we obtain very similar parameter estimates like with the tsglm function:

```
R> campydata <- data.frame(ts = campy[-1], lag1 = campy[-length(campy)])
R> coef(glm(ts ~ lag1, family = poisson(link = "identity"), data = campydata))
(Intercept)        lag1
     4.0322      0.6556
R> coef(tsglm(campy, model = list(past_obs = 1), link = "identity"))
(Intercept)      beta_1
     4.0083      0.6501
```

As described in more detail in Section A.3, a fit by the glm function can be used as a starting value to the function tsglm.

The class of generalized additive models for location, scale and shape (GAMLSS) has been introduced by Rigby and Stasinopoulos (2005) as an extension of a GLM and a generalized additive model (GAM). In addition to the location parameter further parameters of the conditional distribution can be modeled as functions of the explanatory variables. In the following example we use the package **gamlss** (authored by Rigby and Stasinopoulos, 2005) to fit an INGARCH(1,0) model to the Campylobacterosis data. The overdispersion coefficient σ_t^2 of the Negative Binomial distribution is not constant but changes with time according to the equation

$$\sigma_t^2 = \exp\left(\beta_0^* + \beta_1^* \log(Y_{t-1} + 1)\right).$$

```
R> library("gamlss")
R> gamlss(ts ~ lag1, sigma.formula = ~ log(lag1+1), data = campydata,
+         family = NBI(mu.link = "identity", sigma.link = "log"))[c(25, 43)]
GAMLSS-RS iteration 1: Global Deviance = 803.7
GAMLSS-RS iteration 2: Global Deviance = 803.7
GAMLSS-RS iteration 3: Global Deviance = 803.7
$mu.coefficients
(Intercept)        lag1
     3.8409      0.6768

$sigma.coefficients
  (Intercept) log(lag1 + 1)
      -4.2986        0.7167

GAMLSS-RS iteration 1: Global Deviance = 803.7
GAMLSS-RS iteration 2: Global Deviance = 803.7
GAMLSS-RS iteration 3: Global Deviance = 803.7
GAMLSS-RS iteration 1: Global Deviance = 805.6
GAMLSS-RS iteration 2: Global Deviance = 805.6
GAMLSS-RS iteration 3: Global Deviance = 805.6
```

The possibility of a time dependent dispersion coefficient does not improve the fit for this data example (according to the AIC, which is 811.72 compared to 811.64 for a model with constant overdispersion coefficient) but might be quite useful for other data examples. However, it is clear that such a complex model yields more uncertainty of the parameter estimations (i.e., larger standard errors, which are not shown here).

The package **ZIM** (Yang, Zamba, and Cavanaugh, 2014) fits zero-inflated models (ZIM) for count time series with excess zeros. These models are suitable for data where the value zero occurs more frequently than it would be expected when assuming other count time series models. The main idea of these models is to replace the ordinary Poisson or Negative Binomial distribution by its respective zero-inflated version, which is a mixture of a singular distribution in zero (with probability ω_t) and a Poisson or Negative Binomial distribution (with probability $1-\omega_t$), respectively. The model proposed by Yang, Zamba, and Cavanaugh (2013) allows both, the probability ω_t and the conditional mean λ_t of the ordinary count data distribution, to vary over time. The conditional mean λ_t is modeled by using a logistic regression model. The probability ω_t is modeled by a GLM with the logistic link. Other methods for count data with excess zeros, which also have these limitations, are provided by the well-established functions `zeroinfl` and `hurdle` from the package **pscl** (Zeileis, Kleiber, and Jackman, 2008). However, the package **ZIM** includes an extension of ZIM to state space models, which is treated in the next section. The parameters of a ZIM are fitted by the function `zim` employing an EM algorithm. Zero-inflation models are definitely appealing for count time series which occasionally exhibit excess zeros. For our data example of Campylobacter infections, which does not include any zero observations, ZIM are not applicable.

The current version of the **tscount** package considered in this chapter is limited to modeling univariate data. A possible extension to models for vectors of counts is provided by the package **VGAM** (Yee, 2016) introduced by Yee (2015). The function `vglm` in this package fits a vector GLM (VGLM) (see Yee and Wild, 1996) where the conditional density function of a d-dimensional response vector $\boldsymbol{Y}_t$ given an r-dimensional covariate vector $\boldsymbol{X}_t$ is assumed to be of the form

$$f(\boldsymbol{Y}_t|\boldsymbol{X}_t;\boldsymbol{H}) = h(\boldsymbol{Y}_t, \nu_1, \ldots, \nu_s),$$

where $\nu_j = \boldsymbol{\eta}_j^\top \boldsymbol{X}_t$, $j = 1, \ldots, s$ and $h(\cdot)$ is a suitably defined function. The model parameters are given by the $(r \times s)$-dimensional parameter matrix $\boldsymbol{H} = (\boldsymbol{\eta}_1^\top, \ldots, \boldsymbol{\eta}_s^\top)^\top$. Choosing $d = s = 1$ results in the special case of an ordinary, univariate GLM. We demonstrate a fit of an INGARCH(1,0) model to the Campylobacterosis data by the following code:

```
R> library("VGAM")
R> coef(vglm(ts ~ lag1, family = poissonff(link = "identitylink"),
+            data = campydata))
(Intercept)        lag1
     4.0322      0.6556
```

We note that the function `vglm` produces exactly the same output as the function `glm` for this special case. The function `vgam` from the same package would allow to fit an even more general vector generalized additive model (VGAM), which is a multivariate generalization of a generalized additive model (GAM), see Yee (2015) for more details.

Due to the aforementioned limitations of the procedures developed for independent data we would generally suggest the use of the function `tsglm` for modeling count time series. However, in certain situations, where features of the data are currently not supported by `tsglm`, the aforementioned packages can be employed with care; recall the second paragraph of this section. For count time series with many zeros one might want to consider using, for example, the package **ZIM**. If there are reasons to assume a time-varying overdispersion coefficient, the package **gamlss** is a good choice. Multivariate count time series could be analyzed with the package **VGAM**.

2.7.2 Packages for time series data

In this section we present R packages developed for count time series data.

The package **acp** (Siakoulis, 2015) has been published recently and provides maximum likelihood fitting of autoregressive conditional Poisson (ACP) regression models. These are the INGARCH models given by (1.2); see Section 1.2. The **acp** package also allows to include covariate effects. In its latest version 2.1, which has been published in December 2015, the package has been extended to fit models of general order p and q. The `tsglm` function of our package includes these models as special cases and is more general in the following aspects:

- The **acp** package is different in many technical details. Notably, it does not allow to incorporate the parameter constraints given in Section 2.2.1.
- Quasi maximum likelihood fitting allows to choose a more flexible Negative Binomial model instead of a Poisson model (argument `distr = "nbinom"`).
- The `tsglm` function additionally comprises a log-linear model (argument `link = "log"`), which is more adequate for many count time series.
- The `tsglm` function allows for more flexible dependence modeling by allowing arbitrary specification of dynamics. This flexibility is missing by the `acp` function for model fitting because it requires all variables up to a given order to be included (e.g. $\lambda_{t-1}, \ldots, \lambda_{t-12}$ and not just λ_{t-12}). For instance, `tsglm` allows for stochastic seasonality (see Section 2.6.1).
- The `tsglm` function differentiates between covariates with so-called external and internal effect (see Equation (1.4) and the accompanying discussion).

In the following example, an INGARCH(1,1) model (ignoring the seasonal effect) is fitted to the Campylobacterosis data analyzed in Section 2.6.1:

```
R> library("acp")
R> coef(acp(campy ~ -1, p = 1, q = 1))
[1] 2.5320 0.5562 0.2295
R> coef(tsglm(campy, model = list(past_obs = 1, past_mean = 1)))
(Intercept)      beta_1     alpha_1
     2.3890      0.5183      0.2693
```

The parameter estimations obtained by the `acp` function are very similar to those obtained by the `tsglm` function when fitting the same model.

The class of generalized linear autoregressive moving average (GLARMA) models combines GLM with ARMA processes. A software implementation is available in the package **glarma** (Dunsmuir and Scott, 2015). The GLARMA model assumes the conditional distribution of Y_t given the past $\mathcal{F}_{t-1}$ to be Poisson or Negative Binomial with mean λ_t and density $f(Y_t|\lambda_t)$, with λ_t given by

$$g(\lambda_t) = \boldsymbol{\eta}^\top \boldsymbol{X}_t + O_t + Z_t,$$

where O_t is an offset term. An intercept is included by choosing the first column of the time-varying covariate matrix $\boldsymbol{X}_t$ to be the vector $(1, \dots, 1)^\top$. Serial correlation is induced by an autoregressive moving average (ARMA) structure of Z_t, which is given by

$$Z_t = \sum_{k=1}^{p} \phi_k (Z_{t-i_k} + e_{t-i_k}) + \sum_{\ell=1}^{q} \psi_\ell e_{t-j_\ell}.$$

Hereby the process $\{Z_t : t \in \mathbb{N}\}$ is defined by means of residuals e_t which can be possibly rescaled, see (2.7) and (2.8). In the example below we choose Pearson residuals. For the link function $g(\cdot)$, the **glarma** package currently supports only the logarithm but not the identity, which is available in our function `tsglm`. Like in our package, the user can specify the model order by considering the sets $P = \{i_1, \dots, i_p\}$ and $Q = \{j_1, \dots, j_q\}$. The formulation of the GLARMA model we consider describes the modeling possibilities provided by the **glarma** package. In fact, this formulation is more general than the accompanying article by Dunsmuir and Scott (2015), where the authors consider the case $Q = \{1, \dots, q\}$ and $P = \{1, \dots, p\}$. Choosing P and Q, in the context of GLARMA modeling, should be done cautiously (see Dunsmuir and Scott, 2015, Seection 3.4). Our limited experience shows that the minimum element of the set Q should be chosen in such a way that it is larger than the maximum element of P for avoiding errors. Unlike ordinary ARMA models, GLARMA models are not driven by random innovations but by residuals e_t. Note that the model fitted by the function `tsglm` is also related to ARMA processes, see (A.4) in the Appendix. The function `glarma` implements maximum likelihood fitting of a GLARMA model. We compare the following model fitted to the Campylobacterosis data by the function `glarma` with a fit by `tsglm` (see Section 2.6.1, but without the intervention effects):

```
R> library("glarma")
R> glarmaModelEstimates(glarma(campy, phiLags = 1:3, thetaLags = 13,
+      residuals = "Pearson", X = cbind(intercept=rep(1, length(campy))),
+      type = "NegBin"))[c("Estimate", "Std.Error")]
           Estimate Std.Error
intercept  2.34110   0.10757
phi_1      0.22101   0.03940
phi_2      0.05978   0.04555
phi_3      0.09784   0.04298
theta_13   0.08602   0.03736
alpha     10.50823   1.91232
```

With the notation introduced above the fitted model for the number of new infections Y_t in time period t is given by $Y_t|\mathcal{F}_{t-1} \sim \text{NegBin}(\lambda_t, 10.51)$ with $\log(\lambda_t) = 2.34 + Z_t$ and

$$Z_t = 0.22(Z_{t-1} + e_{t-1}) + 0.06(Z_{t-2} + e_{t-2}) + 0.1(Z_{t-3} + e_{t-3}) + 0.09e_{t-13}.$$

We focus on the models' capability to explain the serial correlation which is present in the data. Considering the GLARMA model, a choice of $P = \{1, 2, 3\}$ and $Q = \{13\}$ leaves approximately uncorrelated residuals (see the autocorrelation function in Figure 2.5 (top right)). For the model class fitted by our function `tsglm` we have chosen the more parsimonious model with $P = \{1\}$ and $Q = \{13\}$ to obtain a fit with approximately uncorrelated residuals. Figure 2.6 shows that both models seem to provide an adequate fit to given data. The package **glarma** provides a collection of functions which can be applied to a fitted GLARMA model. For example it provides a function for testing whether there exists serial dependence and it offers tools for model diagnostics. To conclude, both models are able to explain quite general forms of serial correlation but the role of the dependence parameters is quite different and any results should be interpreted carefully. A more detailed comparison would be interesting but is beyond the scope of this thesis.

Another class of models, which is closely related to the GLARMA models, are the so-called generalized autoregressive moving average (GARMA) models developed by Benjamin, Rigby, and Stasinopoulos (2003). Dunsmuir and Scott (2015, Section 3) remark that both model classes are similar in their structure but they have some important differences. The GARMA model is formulated by

$$g(\lambda_t) = \boldsymbol{\eta}^\top \boldsymbol{X}_t + \sum_{k=1}^{p} \phi_k \left(g(Y_{t-k}) - \boldsymbol{\eta}^\top \boldsymbol{X}_{t-k}\right) + \sum_{\ell=1}^{q} \psi_\ell \left(g(Y_{t-k}) - g(\lambda_{t-\ell})\right),$$

where the notation follows the GLARMA notation. Compared to the GLARMA model, the GARMA model does not include an offset and the ARMA structure applies to values which are transformed by the link function g, i.e., on the scale of the linear predictor. In case of a logarithmic link, the observations Y_t are replaced by $\max(Y_t, c)$ for a threshold $c \in (0, 1)$, such that $g(Y_t)$ is well-defined. In our package this problem is handled replacing Y_t by $Y_t + 1$. The package **gamlss.util** (Stasinopoulos, Rigby, and Eilers, 2015) contains the function `garmaFit` for fitting such GLARMA models. Like ordinary GLMs, these models are fitted by maximum likelihood employing the IRLS algorithm. As pointed out on the accompanying help page, the function `garmaFit` does not guarantee stationarity of the fitted model. Additionally, the function `garmaFit` does not allow to specify serial dependence of higher order without including all lower orders, which would be necessary for parsimoniously describing stochastic seasonality. The following example shows a fit of a Negative Binomial GARMA model of order $p = 1$ and $q = 1$ with link $g(\cdot) = \log(\cdot)$ to the Campybacterosis data:

```
R> library("gamlss.util")
R> coef(garmaFit(campy ~ 1, order = c(1, 1), family = NBI(mu.link = "log")))
deviance of linear model=  891.1
deviance of  garma model=  803.3
beta.(Intercept)              phi            theta
          2.6216           0.7763          -0.2917
```

In the above output the AR coefficient ϕ_1 is named `phi` and the MA coefficient ψ_1 `theta`. The function `garma` from the package **VGAM** (Yee, 2016) is an alternative implementation for fitting GARMA models. However, the accompanying help page warns that this function is still in premature stage and points to potential problems with the initialization (in version 1.0-1 of the package). In addition, `garma` allows only for autoregressive modeling (i.e. $q = 0$) and the Negative Binomial distribution is not supported. Hence our example can only show a fit of a Poisson GARMA model of order $p = 1$ and $q = 0$ to the Campylobacterosis data:

```
R> coef(vglm(campy ~ 1, family = garma(link="loge", p.ar.lag = 1, q.ma.lag = 0,
+                                      coefstart = c(0.1, 0.1))))
(Intercept)      (lag1)
      1.948       1.123
```

In this example the estimated coefficient for the autoregressive term is larger than one, which suggests that the fitted process is not stationary. We could not find settings for which the functions `garmaFit` and `garma` fit the same model and give identical or at least similar results. Due to the close relationship of GARMA and GLARMA models we refrain from presenting a comparison to a fit with our package and refer to the comparison in the previous paragraph made for GLARMA models.

The models presented so far are determined by a single source of randomness, i.e., given all past observations, uncertainty is only induced by the Poisson or Negative Binomial distribution from which the observations are assumed to be drawn. These models belong to the class of observation-driven models according to the classification of Cox (1981). In the following paragraphs we present parameter-driven models. These models are determined by multiple sources of randomness introduced by one or more innovation processes. Helske (2016a) comment on the merits of both approaches. He argues that parameter-driven models are appealing because they allow to introduce even multiple latent structures in a flexible way. On the other hand, he observes that observation-driven models like the ones we consider are of advantage for prediction because of their explicit dependence on past observations and covariates.

The package **surveillance** (Salmon, Schumacher, and Höhle, 2016b) includes methods for online change point detection in count time series. One of the models used here is a hierarchical time series (HTS) model proposed by Manitz and Höhle (2013) based on the work by Heisterkamp, Dekkers, and Heijne (2006). This particular state space model accounts for serial dependence by a time-varying intercept. More precisely, it is assumed that $Y_t \sim \text{NegBin}(\lambda_t, \phi)$, where the conditional mean λ_t is given by

$$\lambda_t = \exp\left(\beta_{0,t} + \delta t + \gamma_t + \boldsymbol{\eta}^\top \boldsymbol{X}_t\right).$$

The time-varying intercept $\beta_{0,t}$ is assumed to depend on its previous values according to

$$\Delta_d \beta_{0,t} | \beta_{0,t-1}, \ldots, \beta_{0,t-d} \sim \text{N}(0, \kappa_{\beta_0}^{-1}),$$

where Δ_d is the difference operator of order $d \in \{0,1,2\}$ and κ_{β_0} a precision parameter. Explicitly, it holds $\beta_{0,t} \sim \text{N}(0, \kappa_{\beta_0}^{-1})$, $\beta_{0,t}|\beta_{0,t-1} \sim \text{N}(\beta_{0,t-1}, \kappa_{\beta_0}^{-1})$ and

$$\beta_{0,t}|\beta_{0,t-1}, \beta_{0,t-2} \sim \text{N}(2\beta_{0,t-1} - \beta_{0,t-2}, \kappa_{\beta_0}^{-1})$$

for $d = 1,2,3$, respectively. For $d > 0$ this induces dependence between successive observations. The other parameters, δ for the linear trend, γ_t for a seasonal effect, and the vector $\boldsymbol{\eta}$ for the effect of a covariate vector $\boldsymbol{X}_t$, are also assumed to be normally distributed with certain priors. Inference is done in a Bayesian framework and utilizes an efficient integrated nested Laplace approximation (INLA) provided by the package **INLA** (Lindgren and Rue, 2015) (available from `http://www.r-inla.org`). In the following example we fit a Negative Binomial model without trend, seasonality or covariate effects but with a time-varying intercept of order $d = 1$ to the Campylobacterosis data:

```
R> library("INLA")
R> campyfit_INLA <- inla(ts ~ f(time, model = "rw1", cyclic = FALSE),
+                   data = data.frame(time = seq(along = campy), ts = campy),
+                   family = "nbinomial", E = mean(campy),
+                   control.predictor = list(compute = TRUE, link = 1),
+                   control.compute = list(cpo = FALSE, config = TRUE),
+                   control.inla = list(int.strategy = "grid", dz = 1,
+                                       diff.logdens = 10))
R> posterior <- inla.posterior.sample(1000, campyfit_INLA)
R> rowMeans(sapply(posterior, function(x) (unname(x$hyperpar))))
```

```
[1]   9.314 169.208
```

The estimates for the parameters ϕ (the former) and κ_{β_0} (the latter) in the output above are based on means of a sample of size 1000 from the posterior distribution. Fitted values $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{140}$ are obtained in the same way (not shown in the above code). With the Bayesian approach it is very natural to obtain prediction intervals for future observations from the posterior distribution which account for the estimation and observation uncertainties. This is a clear advantage over the classical likelihood-based approach pursued in our package (cf. Section 2.3). A disadvantage of the Bayesian approach is its much higher computational effort which could be an obstacle for real-time applications and simultaneous analysis of several time series. The above example runs more than eight seconds on a standard office computer (Intel Xeon CPU with 2.83 GHz); this is seven times longer than `tsglm` takes to fit the model. An additional difference between our approach and that taken by INLA is the specification of temporal dependence. The comparison of the final fitted values, shown in Figure 2.6, illustrates that the model with a time-varying intercept fitted by `inla` possess a much smoother line through the observed values when compared to the model fitted by `tsglm`. For this example, the empirical autocorrelation function of the response residuals in Figure 2.5 (bottom left) is significantly different from zero at lag one; hence short term temporal correlation is not explained sufficiently by the hierarchical model. It also becomes clear by this plot that we should have included seasonality by employing the term γ_t. The residuals of the GLM-based fit by the function `tsglm` do not exhibit any serial correlation which has not been explained by the model (see Figure 2.5 (top left)). In general, the GLM-based model is expected to provide more accurate 1-step-ahead predictions whilst the hierarchical model prediction obtained by `inla` is more stable. Either of these two features could be preferable depending upon the specific application. It would be interesting to study these two ways of modeling temporal dependence in a future work.

The package **KFAS** (Helske, 2016b) treats state space models for multivariate time series where the distribution of the observations belongs to the exponential family (and also includes the Negative Binomial distribution). Its name refers to Kalman filtering and smoothing, which are the two key algorithms employed by the package. This package is able to cope with very general state space models at the cost of a rather big effort for its correct specification. However, some auxiliary functions and the use of symbolic model description reduces this effort. In contrast to the package **INLA**, which is also capable of fitting state space models, **KFAS** implements maximum likelihood estimation (for a comparison of these two packages see Helske, 2016a). One possible univariate model for the Campylobacterosis data could be the state space model $Y_t|\mathcal{F}_{t-1} \sim \text{NegBin}(\lambda_t, \phi)$ where $\lambda_t = \exp(\nu_t)$ and the state equation is

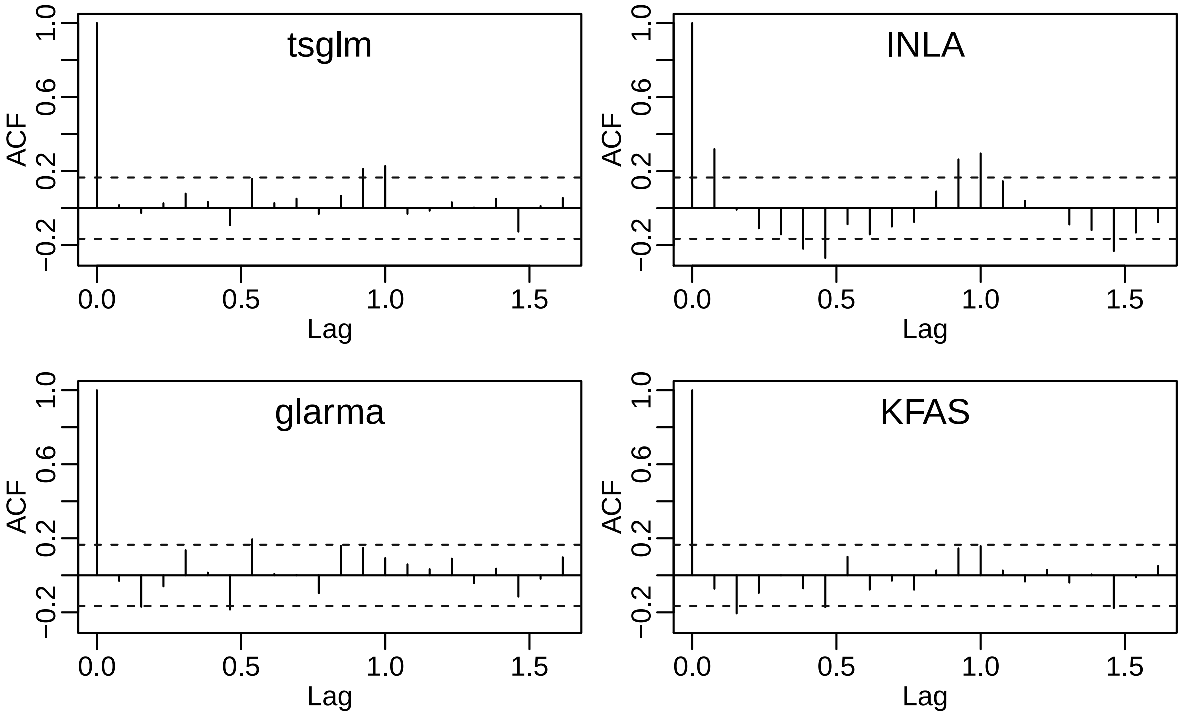


Figure 2.5: Empirical autocorrelation function of the response residuals for a model fit of the campylobacterosis data by our function `tsglm` (see Section 2.6.1, but without the intervention effects) and the packages **glarma**, **INLA** and **KFAS** (see Section 2.7).

$$\nu_t = \nu_{t-1} + \varepsilon_t.$$

The initialization ν_1 is specified by assuming $\nu_1 \sim \mathrm{N}(\lambda, \sigma_\nu^2)$. The degree of serial dependence is induced by independently distributed innovations ε_t for which it is assumed that $\varepsilon_t \sim \mathrm{N}(0, \sigma_\varepsilon^2)$. This model has unknown parameters $\lambda \in \mathbb{R}$, and $\phi, \sigma_{\nu_1}^2, \sigma_\varepsilon^2 \in [0, \infty)$ and can be fitted as follows:

```
R> library("KFAS")
R> model <- SSModel(campy ~ SSMcustom(Z = 1, T = 1, R = 1, Q = 0,
+                                      a1 = NA, P1 = NA) - 1,
+                   distribution = "negative binomial", u = NA)
R> updatefn <- function(pars, model, ...){
+   model$a1[1, 1] <- pars[1]
+   model$u[, 1] <- exp(pars[2])
+   model$P1[1, 1] <- exp(pars[3])
+   model$Q[1,1,1] <- exp(pars[4])
+   return(model)
+ }
R> campyfit_KFAS <- fitSSM(model = model, inits = c(mean(campy), 0, 0, 0),
+                  updatefn = updatefn)
R> exp(campyfit_KFAS$optim.out$par)
[1] 3.427e+00 9.148e+01 2.775e-16 4.334e-02
```

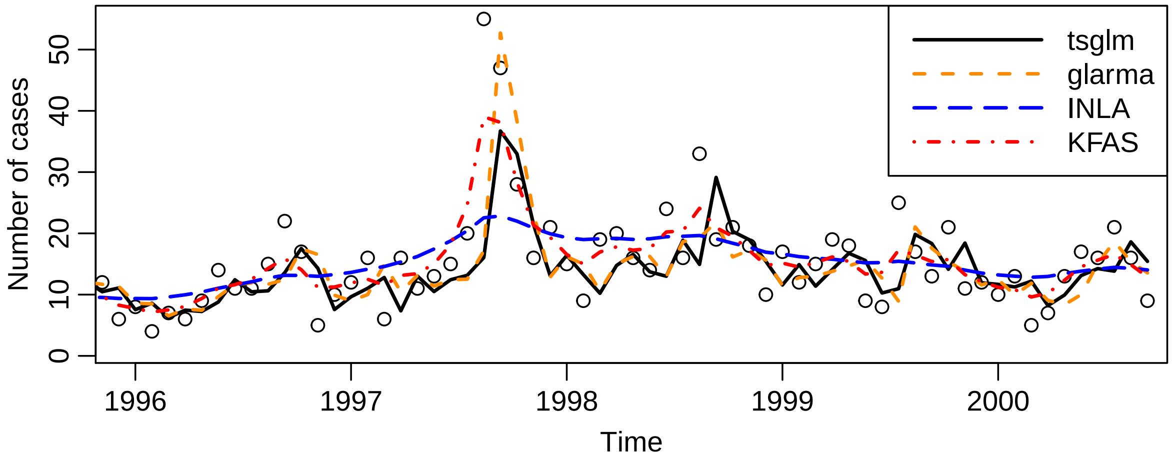


Figure 2.6: Comparison of a model fit of the campylobacterosis data by our function `tsglm` (see Section 2.6.1, but without the intervention effects) and the packages **glarma**, **INLA** and **KFAS** (see Section 2.7).

The output above corresponds to the estimated parameters λ, ϕ, $\sigma^2_{\nu_1}$ and σ^2_{ε}, respectively. We observe that the estimation procedure is quite sensitive to given starting values; this fact has been pointed out by Helske (2016a). As shown by the empirical autocorrelation function of the residuals in Figure 2.5 (bottom right), the fitted model explains the temporal dependence of the data quite adequately. The values fitted by this model (see Figure 2.6) do not show any delay when compared to the fit obtained by `tsglm`. The algorithm used by `tsglm` yields fitted values by 1-step-ahead forecasts based on previous observations; note that only the model parameters are fitted using all available observations. The algorithm of the **KFAS** package for obtaining fitted values includes future observations which naturally lead to a more accurate fit. However, this methodology does not guarantee better out-of-sample forecasting performance since future observations will not be available in general. Further empirical comparison between **KFAS** and **tscount** is required to compare the accuracy of predictions obtained by both models.

Another state space model which could be used to describe count time series is a partially observed Markov process (POMP). The package **pomp** (King, Nguyen, and Ionides, 2016) provides a general and abstract representation of such models. One example by King *et al.* (2016, Sections 4.5 and 4.6) is the so-called Ricker model for describing the size N_t of a population which is assumed to fulfill

$$N_{t+1} = rN_t \exp(-N_t + \varepsilon_t)$$

with innovations $\varepsilon_t \sim \mathrm{N}(0, \sigma^2_\varepsilon)$. The actual observations Y_t are noisy measurements of the population size N_t and it is assumed to hold $Y_t \sim \text{Poisson}(\phi N_t)$, where ϕ is an unknown dispersion parameter. Specification and fitting of this rather simple model with the package **pomp** requires more than thirty lines of code. This complexity is an obstacle for using the package in standard situations but might prove beneficial in very special scenarios.

The package **gcmr** provides methodology for Gaussian copula marginal regression (Masarotto and Varin, 2012), a framework which is also capable to model count time series. The marginal distribution of a time series Y_t given a covariate vector $\boldsymbol{X}_t$ can be modeled by a Poisson or Negative Binomial distribution with mean λ_t using that $g(\lambda_t) = \beta_0 + \boldsymbol{\eta}^\top \boldsymbol{X}_t$. This is similar to model (1.1) but it does not include the terms for regression on past values of Y_t and λ_t. Furthermore, randomness is introduced through an unobserved error process $\{\varepsilon_t : t \in \mathbb{N}\}$ by assuming that $Y_t = F_t^{-1}(\Phi(\varepsilon_t))$, where F_t is the cumulative distribution function of the Poisson or Negative Binomial distribution with mean λ_t and Φ is the cumulative distribution function of the standard normal distribution. Hence the actual value of Y_t is the $\Phi(\varepsilon_t)$-quantile of the Poisson or Negative Binomial distribution with mean λ_t. As pointed out by Masarotto and Varin (2012), copulas with discrete marginals might not be unique (see also Genest and Nešlehová, 2007). Temporal dependence of $\{Y_t\}$ is modeled through the error process $\{\varepsilon_t\}$ by assuming an autoregressive moving average (ARMA) model of order p and q. Note that although this model accounts for serial dependence it does not model the conditional distribution of Y_t given the past but only its time-varying marginal distribution. The mean λ_t of this marginal distribution is not influenced by the actual (unobserved) value of the error ε_t. Hence this model is not suitable for accurate 1-step-ahead predictions but rather for quantifying the additional uncertainty induced by the serial dependence. The following example presents a fit of a Negative Binomial model with an ARMA error process of order $p = 1$ and $q = 1$:

```
R> library("gcmr")
R> gcmr(ts ~ 1, marginal = negbin.marg(link = "identity"),
+       cormat = arma.cormat(p=1, q=1), data = data.frame(ts = campy))


Call:
gcmr(formula = ts ~ 1, data = data.frame(ts = campy),
    marginal = negbin.marg(link = "identity"),
    cormat = arma.cormat(p = 1, q = 1))

Marginal model parameters:
(Intercept)   dispersion
     11.320        0.255

Gaussian copula parameters:
   ar1     ma1
 0.823  -0.269
```

The estimated ARMA parameters of the error process indicate that there is a considerable amount of serial correlation in the data. Some of the observed variability is explained by this serial dependence. As discussed above, the fitted values are based solely on the estimated marginal distribution and are therefore constant over time (equal to the estimated intercept in the above output).

An extension of the zero-inflated models of the package **ZIM**, which was presented in the previous section, are the so-called dynamic zero-inflated models (DZIM) as proposed by Yang, Cavanaugh, and Zamba (2015). These fall within the framework of state space models and introduce serial dependence by an unobserved autoregressive process of order p. Fitting is based on the EM algorithm and MCMC simulation.

The package **tsintermittent** (Kourentzes and Petropoulos, 2016) provides methods for so-called intermittent demand time series (see for example Kourentzes, 2014). These are time series giving the number of requested items of a particular product (e.g., rarely needed spare parts) which is demanded in a sporadic fashion with periods of zero demand. Reliable forecasts of such time series are important for companies to efficiently plan stocking the respective items. In principle, these are count time series which could be analyzed with the methods in our package. However, it is known that classical forecasting methods perform unsatisfactorily for this kind of data (Kourentzes, 2014). More successful approaches, which are included in the **tsintermittent** package, are based on the idea of separately modeling the non-zero demand size and the inter-demand intervals. These methods do not take into account that the observations are integers and do not include covariate effects, as the methods in our package do. Temporal dependence is considered implicitly, by assuming that there are periods where subsequent observations are zero. Our functions consider explicit time dependence. The methods in the **tsintermittent** package are possibly more appropriate for the specific context they are tailored for (which would need further examination), but not suitable for count time series in general. They compete with other types of models for zero excess time series data, for example with DZIM.

2.8 Discussion

We are the first to provide such a general formulation and comprehensive treatment of count time series following generalized linear models. In its current version, the R package **tscount** allows for the analysis of count time series with this quite broad class of models. Our comparison with other packages shows that **tscount** is an important contribution which extends and complements existing software. It will hopefully prove to be useful for a wide range of applications.

There are a number of desirable extensions of the package which could be included in future releases. The most important ideas for extending **tscount** are described in the following paragraphs. We invite other researchers and developers to contribute to this package.

As an alternative to the Negative Binomial distribution, one could consider the so-called Quasi-Poisson model. It allows for a conditional variance of $\phi\lambda_t$ (instead of $\lambda_t + \phi\lambda_t^2$, as for the Negative Binomial distribution), which is linearly and not quadratically increasing in the conditional mean λ_t (for the case of independent data see Ver Hoef and Boveng, 2007). A scatterplot of the squared residuals against the fitted values could reveal whether a linear relation between conditional mean and variance is more adequate for a given time series. A generalization of the test for overdispersion in INGARCH(1,0) processes proposed by Weiß and Schweer (2015) could provide guidance for choosing an appropriate conditional distribution.

The common regression models for count data are often not capable to describe an exceptionally large number of observations with the value zero. In the literature so-called zero-inflated and hurdle regression models have become popular for zero excess count data (for an introduction and comparison see Loeys, Moerkerke, De Smet, and Buysse, 2012). A first attempt to utilize zero-inflation for INGARCH time series models is made by Zhu (2012).

In some applications the variable of interest is not the number of events but the rate, which expresses the number of events per unit. For example the number of infected people per 10 000 inhabitants, where the population size is a so-called exposure variable which varies over time. For models with a logarithmic link function such a rate could be described by a model where the number of events is the response variable and the logarithm of the exposure variable is a so-called offset. An offset is supported by many standard functions for GLMs and could be part of a future release of our package.

Alternative nonlinear models are for example the threshold model suggested by Woodard *et al.* (2011) or the models studied by Fokianos and Tjøstheim (2012). Fokianos and Neumann (2013) propose a class of goodness-of-fit tests for the specification of the linear predictor, which are based on the smoothed empirical process of Pearson residuals. Christou and Fokianos (2015a) develop suitably adjusted score tests for parameters which are identifiable as well as non-identifiable under the null hypothesis. These tests can be employed to test for linearity of an assumed model.

In practical applications one is often faced with outliers. Elsaied and Fried (2014) and Kitromilidou and Fokianos (2016) develop M-estimators for the linear and the log-linear model, respectively. Fried *et al.* (2014) compare robust estimators of the (partial) autocorrelation (see also Dürre, Fried, and Liboschik, 2015a) for time series of counts, which can be useful for identifying the correct model order.

In the long term, related models for binary or categorical time series (Moysiadis and Fokianos, 2014) or potential multivariate extensions of count time series following GLMs could be included as well.

The models which are so far included in the package or mentioned above fall into the class of time series following GLMs. There is also quite a lot of literature on thinning-based time series models but we are not aware of any publicly available software implementations. To name just a few of many publications, Weiß (2008) reviews univariate time series models based on the thinning operation, Pedeli and Karlis (2013) study a multivariate extension and Scotto, Weiß, Silva, and Pereira (2014) consider models for time series with a finite range of counts. For the wide class of state space models there are the R packages **INLA**, **KFAS** and **pomp** available, although it is quite complex to apply these to count time series. A future version of our package could provide simple interfaces to such packages specifically for fitting certain count time series models.

Chapter 3

Retrospective intervention detection

3.1 Introduction

In many applications, unusual external effects or measurement errors can lead to either sudden or gradual changes in the structure of the data. Furthermore such effects can result in several singular observations of a distinct nature than the rest. Following Box and Tiao (1975) we use the term intervention for all kinds of unusual effects influencing the ordinary pattern of the data, including structural changes and different forms of outliers. Considering for example the monthly number infections with a certain disease, where a disease outbreak would be a possible intervention effects. A goal of an intervention analysis is to examine the effect of known interventions, for example to judge whether a policy change had the intended impact (for practical applications see for example Box and Tiao, 1975). Another possible goal is to search for unknown intervention effects and to find explanations for them *a posteriori*. Such sudden events are often included in the model by deterministic covariates (e.g., Box and Tiao, 1975; Abraham and Box, 1979). These goals are fundamentally different from the goal of a robust estimation approach, where methods are expected to ignore the effect of aberrant observations.

The main focus of this chapter is to model various types of interventions in so-called INGARCH (or ACP) processes which are defined by (1.2). An extension to the count time series based on generalized linear models as defined in Section 1.2 is briefly discussed in Section 2.5. This includes log-linear models, models with a Negative Binomial conditional distribution, models with additional covariates as well as those with more than one intervention effect.

We introduce a new model to describe intervention effects within the class of INGARCH processes and compare it to an existing intervention model proposed by Fokianos and Fried (2010). Our proposal is able to describe interventions which enter the dynamics of the process in a different way when compared to the existing model. It allows to describe an external effect on the observed value rather than an internal change of the underlying state of the process. Such a model can be more realistic for some applications, as discussed in the last paragraph of Section 3.2.

Section 3.2 defines both intervention models and compares them analytically. Section 3.3 presents joint maximum likelihood estimation of the ordinary model parameters and the intervention effects and studies its properties by simulations. Section 3.4 presents asymptotic procedures to test for single interventions at a given time, or to detect one or multiple interventions at unknown positions. We verify these procedures by simulations. Furthermore, we discuss the problem of misspecification of the intervention model and shed some light on the question of discriminating between the two intervention models. Section 3.5 applies some of the methods to the weekly number of campylobacterosis infections. Section 3.6 concludes this chapter with a short discussion on related issues.

3.2 Intervention models

As a first intervention model in the framework of INGARCH(p,q) processes, Fokianos and Fried (2010) define a contaminated observed process $\{Z_t\}_{t\in\mathbb{N}}$ with an intervention at time τ by

$$Z_t|\mathcal{F}_{t-1}^{Z,\kappa} \sim \text{Poisson}(\kappa_t), \quad \kappa_t = \beta_0 + \sum_{k=1}^{p} \beta_k Z_{t-k} + \sum_{\ell=1}^{q} \alpha_\ell \kappa_{t-\ell} + \nu X_t, \tag{i}$$

where $X_t = \delta^{t-\tau} I_{[\tau,\infty)}(t)$ is a deterministic process describing the intervention effect, $\nu \geq 0$ denotes the intervention size, the predefined constant $\delta \in [0,1]$ specifies the type of intervention and all other model parameters are as before. The information about the past of the process is denoted by $\mathcal{F}_{t_0-1}^{Z,\kappa} = \sigma(Z_{1-p}, \ldots, Z_{t_0-1}, \kappa_{1-q}, \ldots, \kappa_0)$. Model (i) covers three types of interventions with different choices of δ: a spiky outlier (SO) for $\delta = 0$, a transient shift (TS) for $\delta \in (0,1)$ and a level shift (LS) for $\delta = 1$. Singular effects are modeled by a SO, exponentially decaying effects by a TS and a permanent change of location by a LS.

The intervention effect is added to the underlying conditional mean process $\{\kappa_t\}$ and not to the observations itself. Through the regression on past observations and on past conditional means the intervention effect enters the dynamics of the process. The parameter δ has both a direct and an indirect impact on the observations. Its value directly influences the impact of the intervention on the conditional mean and hence its impact on the future values of the process. In addition, the value of δ affects future values of the process because of the dependence in (i) on past values of $\{\kappa_t\}$ and $\{Z_t\}$. Small values of δ yield less impact than larger ones.

We study an alternative intervention model for INGARCH(p,q) processes $\{Z_t\}$ following the definition

$$Z_t|\mathcal{F}_{t-1}^{Z,\kappa} \sim \text{Poisson}(\kappa_t), \quad \kappa_t = \lambda_t + \nu X_t, \quad \lambda_t = \beta_0 + \sum_{k=1}^{p} \beta_k Z_{t-k} + \sum_{\ell=1}^{q} \alpha_\ell \lambda_{t-\ell}, \tag{ii}$$

where the intervention process $\{X_t\}$, the regular INGARCH parameters and the additional intervention parameters are as before. Models (i) and (ii) look somewhat similar but their difference becomes more obvious, if we rewrite the equation for the conditional mean κ_t of intervention model (ii) as

$$\kappa_t = \beta_0 + \sum_{k=1}^{p} \beta_k Z_{t-k} + \sum_{\ell=1}^{q} \alpha_\ell (\kappa_{t-\ell} \underbrace{-\nu X_{t-\ell}}_{\text{not for model (i)}}) + \nu X_t. \tag{3.1}$$

Apart from the labelled term, it is identical to the conditional mean equation of intervention model (i). For the alternative specification (ii) the intervention effect is not propagated via the feedback mechanism of the conditional mean but only via the contaminated observations. From (3.1) we see that both intervention models can be written in a unified way and are fully specified by their observable process $\{Z_t\}$ and their unobservable conditional mean process $\{\kappa_t\}$, in addition to the deterministic intervention process $\{X_t\}$.

One could gain more insight about the difference between the intervention models by considering another representation, which turns out to be useful for cleaning the processes from any intervention effects (see Section 3.4). For intervention model (i) Fokianos and Fried (2010) show that it is possible to decompose $\{Z_t\}$ into an intervention-free process $\{Y_t\}$ and a contamination process $\{C_t\}$, which are mutually independent conditionally on the past, such that $Z_t = Y_t + C_t$. The intervention-free process $\{Y_t\}$ is an INGARCH(p,q) process with conditional mean process $\{\lambda_t\}$ and the same parameters as $\{Z_t\}$. The contamination process $\{C_t\}$ underlies the same structure as $\{Z_t\}$ and can be described by

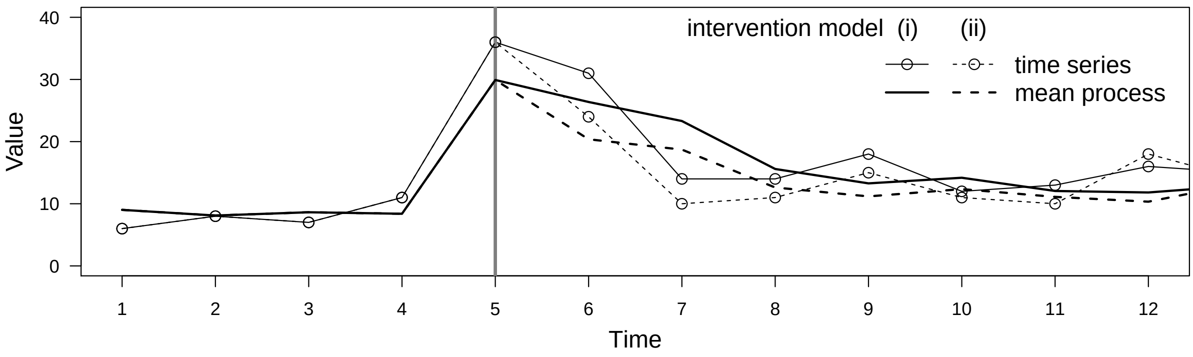


Figure 3.1: Comparison of the intervention models for a simulated INGARCH(1,1) processes with parameters $\beta_0 = 3$, $\beta_1 = 0.4$ and $\alpha_1 = 0.3$ with a transient shift ($\delta = 0.8$) of size $\nu = 20$ at time $\tau = 5$ (vertical line).

$$C_t|\mathcal{F}_{t-1}^{C,\upsilon} \sim \text{Poisson}(\upsilon_t), \qquad \upsilon_t = \sum_{k=1}^{p} \beta_k C_{t-k} + \underbrace{\sum_{\ell=1}^{q} \alpha_\ell \upsilon_{t-j\ell}}_{\text{not for model (ii)}} + \nu X_t, \tag{3.2}$$

with $\{\upsilon_t\}$ the conditional mean process of $\{C_t\}$ and $\kappa_t = \lambda_t + \upsilon_t$. It is straightforward to show that a similar decomposition holds for intervention model (ii) without the labelled summand. Naturally, $\{\upsilon_t\}$ and thus also $\{C_t\}$ are zero for $t < \tau$.

In both models the intervention affects the conditional mean from time τ inwards. However, for model (ii) its effect does not enter the dynamics of the process directly but only via the observations. This can be seen more clearly in (3.2): only for intervention model (i) the intervention effect propagates directly via the conditional mean υ_t. The difference can also be seen from (3.1), where for model (ii) the feedback mechanism is adjusted by the intervention effect. The larger the feedback parameters α_i are, the greater is the difference between both models. For $\alpha_1 = \cdots = \alpha_q = 0$ both models are equivalent.

The simulated example of a transient shift, see Figure 3.1, illustrates the difference between the intervention models. After the intervention occurred, the conditional mean returns faster to its previous level before the intervention for model (ii) rather than for model (i).

An intervention following model (ii) can be interpreted as a sudden external effect, whilst an intervention following model (i) can be rather seen as an internal change of the data generating process. We will therefore refer to model (ii) as the *external* and to model (i) as the *internal* intervention model. This nomenclature corresponds to *external* and *internal* covariate effects as they are defined at the end of Section 1.2.

For illustration consider the weekly number of registered cases of a disease spreading human to human and by contaminated nutrition. A level shift could for instance be caused by a change of the reporting requirements, an enhanced detection method or just a new and more aggressive type of pathogen. A spiky outlier in the external intervention model could be caused by the return of a large number of infected persons from another area, which will affect the future number of cases only by human to human spread. In contrast, a spiky outlier from the internal intervention model could be caused by a large quantity of contaminated food resulting in many more infections than usual. This will affect the future number of cases not only by human to human spread of the additionally infected patients, but also by a larger number of pathogens circulating for example in food processing establishments.

3.3 Estimation and inference

Estimation of model (ii) proceeds along the lines of Fokianos and Fried (2010) after some modifications, treating the intervention as a time-dependent covariate process $\{X_t\}$ and estimating the vector of unknown model parameters $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p,\, \alpha_1, \ldots, \alpha_q, \nu)^\top$ jointly by maximum likelihood. Note that this falls within the more general framework of model (1.1); recall the maximum likelihood estimation procedure in Section 2.2. Note that for the contaminated process we denote observations by z_t and the conditional mean by κ_t unlike for the clean process with y_t and λ_t, respectively. We present estimation formulas valid for both intervention models and point out where these differ from each other. Note that the time τ when the intervention occurs, as well as the parameter δ specifying the type of intervention, are treated as known for estimation. We assume that $\boldsymbol{\theta} \in \Theta$, with

$$\Theta = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{p+q+2} \,\middle|\, \beta_0 > 0,\ \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_q \geq 0,\ \sum_{k=1}^{p} \beta_k + \sum_{\ell=1}^{q} \alpha_\ell < 1,\ \nu \geq 0 \right\}.$$

The following equations are conditional on the unobserved past $\mathcal{F}_0^{Z,\kappa}$ of the process; recall the notation of (i). For an observed time series $\boldsymbol{z} = (z_1, \ldots, z_n)^\top$ the conditional log-likelihood function is up to a constant given by

$$\ell(\boldsymbol{\theta}; \boldsymbol{z}) = \sum_{t=1}^{n} \Big(z_t \ln(\kappa_t(\boldsymbol{\theta})) - \kappa_t(\boldsymbol{\theta}) \Big),$$

where the conditional mean is regarded as a function $\kappa_t : \Theta \to \mathbb{R}^+$ and thus denoted by $\kappa_t(\boldsymbol{\theta})$ for all t. The conditional score function is the $(p+q+2)$-dimensional vector

$$S_{n\tau}(\boldsymbol{\theta}; \boldsymbol{z}) = \frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{z})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} \left(\frac{z_t}{\kappa_t(\boldsymbol{\theta})} - 1 \right) \frac{\partial \kappa_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

Finally, the conditional information matrix is given by

$$G_{n\tau}(\boldsymbol{\theta}) = \sum_{t=1}^{n} \frac{1}{\kappa_t(\boldsymbol{\theta})} \left(\frac{\partial \kappa_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \kappa_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^{\top}.$$

The additional index τ indicates that $S_{n\tau}$ and $G_{n\tau}$ depend on the time of the intervention effect. The conditional maximum likelihood (CML) estimator $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ is the solution of the non-linear constrained optimization problem

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}; \boldsymbol{z}), \tag{3.3}$$

assuming that it exists. As discussed in Section 2.2.3 and Appendix A.2, the evaluation of log-likelihood function, score vector and information matrix is carried out recursively with an appropriate initialization.

3.3.1 Starting value for optimization

To solve the non-linear optimization problem (3.3) subject to the induced constraints, we need to give a starting value for the parameter vector $\boldsymbol{\theta}$. We obtain this starting value from a fit of an ARMA model with the same second order properties as the considered INGARCH model (cf. Appendix A.3). We compare a method of moments (MM) and a conditional least squares (CLS) estimator with respect to their efficiency, computation time and robustness against interventions in a simulation study.

For the simulations in this study we consider an INGARCH(1,1) process with true parameters $\beta_0 = 3$, $\beta_1 = 0.4$ and $\alpha_1 = 0.3$, unless stated otherwise. Without an intervention, such process has a marginal mean of 10 and a marginal variance of 13.14 and thus exhibits moderate overdispersion. The autocorrelation function is 0.47 for lag 1 and decays to a value below 0.01 after 12 lags. Following Fokianos and Fried (2010), the parameter δ of a transient shift is set to 0.8 for simulation as well as for estimation. In the simulations we choose different sizes ν for different types of interventions, because a spiky outlier needs to be of much larger size to be noticeable than a transient shift or even more a level shift (see discussion in Section 3.4.1). Moreover, in applications outlier

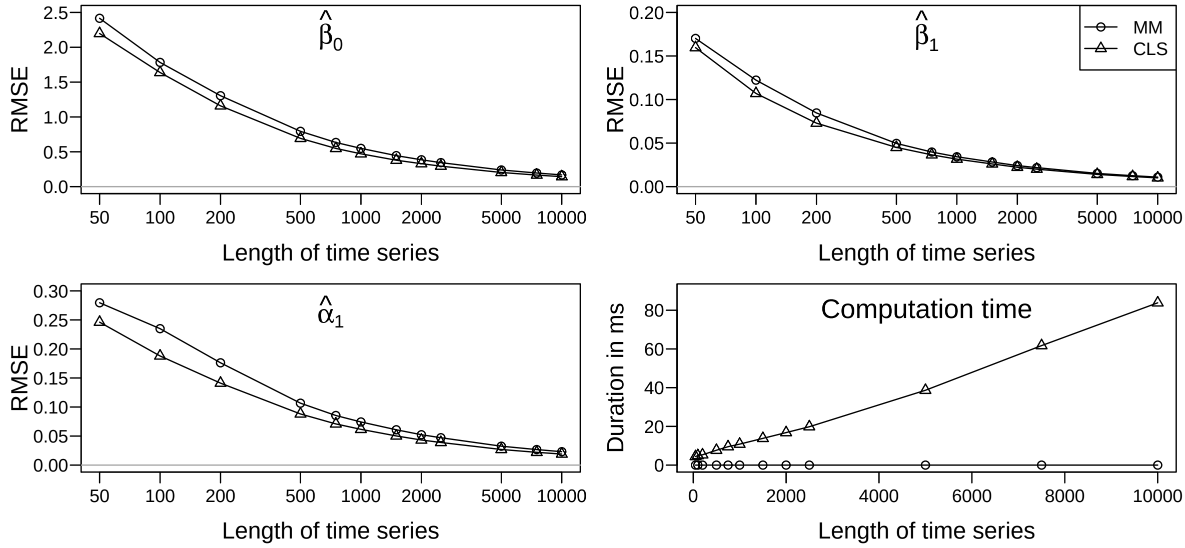


Figure 3.2: Root mean square error (RMSE) and computation time of different start estimators for a simulated intervention-free INGARCH(1,1) process averaged over 5000 repetitions.

effects are particularly relevant if they are large whereas level shifts are of interest even if they are only of moderate size. In the case that the simulation setup is modified it will be stated explicitly.

The results of the simulations in Figure 3.2 suggest that both start estimators are mean square consistent, although several thousand observations are needed to reduce the root mean square error (RMSE) for each parameter to less than 10% of its respective parameter value. Particularly for smaller sample sizes of up to about thousand observations the CLS estimator has a considerably lower RMSE than the MM estimator. The computation time grows linearly in the sample size for both estimators. It is much faster and increases slower with the sample size for the MM estimator than for the CLS estimator (see Figure 3.2 bottom right). However, even though the MM estimator is more than ten times faster for time series with 200 or more observations, the computation times of both start estimators are negligible compared to the computation time required to obtain the final CML estimation.

In this work we are particularly interested in the behavior of the estimators in the presence of interventions. The simulation presented in Figure 3.3 shows that with a TS or LS type of intervention both estimators become heavily biased, whilst the singular event of a SO does not really affect them. If the time of possible interventions is known, the start estimation could be done using only an intervention-free part of the time series. This approach is certainly not always possible, because in many applications the times of possible interventions are not known or the intervention-free part of the time series is too

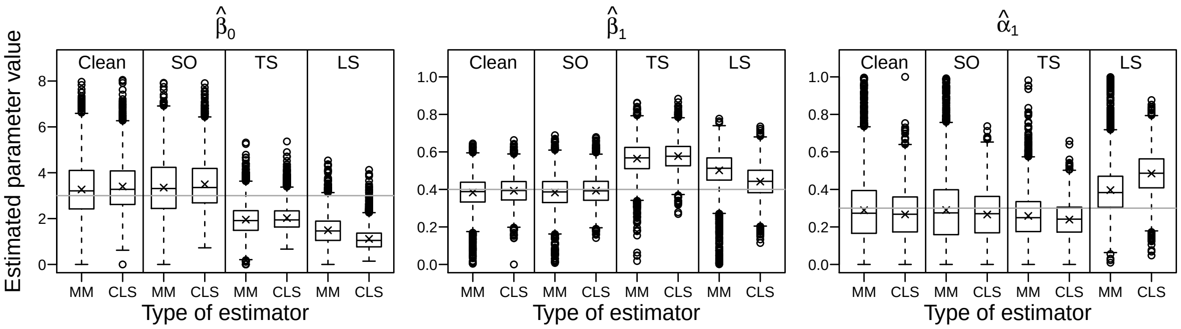


Figure 3.3: Start estimates of a simulated INGARCH(1,1) process with $n = 200$ observations and an intervention in the middle (Clean without intervention, SO with intervention of size $\nu = 24$, TS with $\nu = 18$, LS with $\nu = 3$). The true value is marked by a grey line, the sample mean by a cross. Simulation results are based on 5000 repetitions.

short for a sensible start estimation. However, our simulations presented in Figures 3.4 and 3.5 indicate that the CML estimation of β_0, β_1 and α_1 works well, although the start estimation is biased by an intervention. With respect to its behavior in the presence of an intervention, none of the two initial estimation methods seems to be superior.

Overall we suggest conditional least squares estimation for obtaining starting values because it possesses better properties in small and moderately large samples. This estimator will therefore be used to obtain starting values in all subsequent simulations. Note that we come to a different conclusion when considering start estimation for models with arbitrary covariates; see Appendix A.3.

3.3.2 Properties of the maximum likelihood estimator

Fokianos and Fried (2010) conjecture asymptotic normality of the CML estimator for their intervention model (i) and verify this by simulations. We assume that the same also holds for the alternative intervention model (ii). In other words,

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{d} N_{p+q+2}\left(\mathbf{0}, G_{n\tau}(\widehat{\boldsymbol{\theta}}_n)^{-1}\right), \tag{3.4}$$

as $n \to \infty$, where $\boldsymbol{\theta}_0$ denotes the true parameter value, which has to be in the interior of the parameter space Θ. Note that (3.4) is a special case of (2.6). Simulations reported below suggest that (3.4) holds for $\tau = \lfloor \rho n \rfloor$ with $0 < \rho < 1$. However, a formal proof requires further research. Assuming that the bivariate process $\{(Z_t, \kappa_t)\}$ is ergodic and has moments of at least fourth order, we could in fact show that the score vector forms a square integrable martingale. The main obstacle on proving the asymptotic normality is the convergence of the information matrix because of the non-stationarity of the

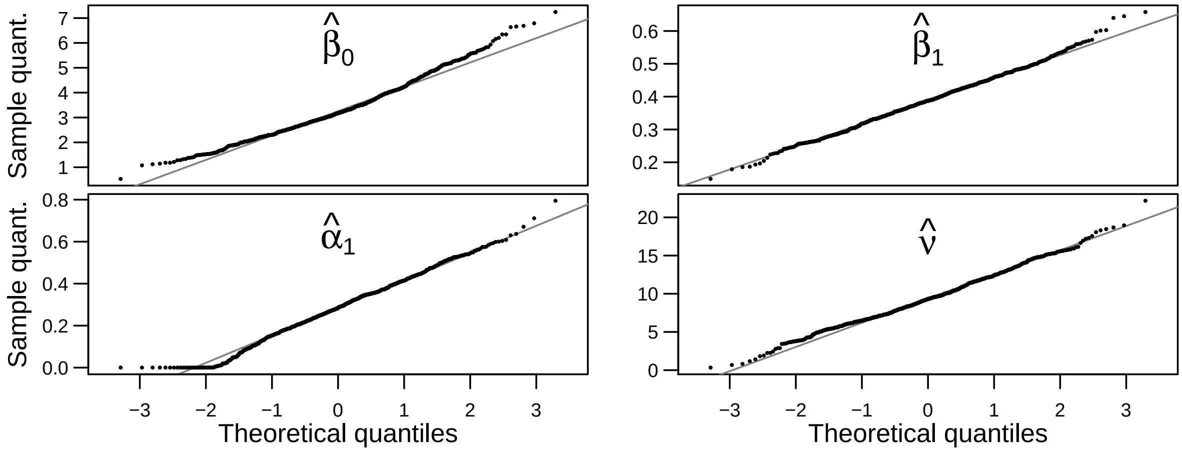


Figure 3.4: Normal QQ-plots of the CML estimation of an INGARCH(1,1) process according to model (ii) with a transient shift of size $\nu = 9$ in the center of $n = 200$ observations. Simulation results are based on 1000 repetitions.

intervention process $\{X_t\}$. However, from our extensive experience with these models the information matrix converges to a non-random limit. In Section 2.2.2 asymptotic normality of the maximum likelihood estimator is discussed for a model with arbitrarily covariates; see also the simulation study in Appendix B.1.

Figure 3.4 and our simulations for several other parameter settings and sample sizes support our conjecture of asymptotic normality, but also show that the convergence is rather slow if the true parameter is close to the boundary of the parameter space. In this case the estimator is skewed for moderate sample sizes. In the example shown in Figure 3.4 there are many estimations of α_1 very close to its lower parameter constraint zero, which is an artifact of the constrained optimization. This problem disappears if the true value of α_1 is further away from the boundary of the parameter space Θ or the sample size n becomes larger and thus the variance of the estimations decreases. For this example with a true value of $\alpha_1 = 0.3$ estimation based on $n = 500$ observations turns out to be sufficient for adequate normal approximation. In contrast, the case of a true value of $\alpha_1 = 0.1$ requires several thousands of observations for achieving approximate normality. The same problem occurs when estimating the constant term β_0, see Fokianos *et al.* (2009), or the intervention size ν. This problem is not inherent in the external intervention model (ii), but arises in the same way for intervention model (i) and also for the intervention-free model (1.2), as we have confirmed by simulations which are not shown here. The practical implication is that if the estimated parameter is close to the boundary of the parameter space, then a larger sample size is needed for reliable inference.

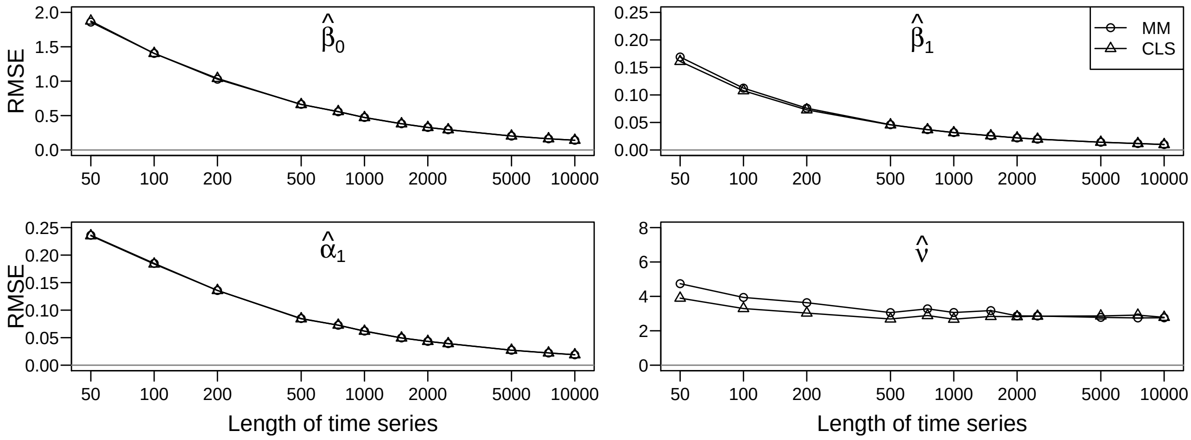


Figure 3.5: Root mean square error (RMSE) for the CML estimator with different initial estimations (see legend) for a simulated INGARCH(1,1) process with TS of size $\nu = 9$. Here the time of intervention is $\tau = n/2$ for a varying length n of the time series. Simulation results are averaged over 1000 repetitions.

Another issue showed up during the simulation experiments for situations with $\delta < 1$ only. In the case of a level shift (i.e. $\delta = 1$) all observations after time τ carry roughly the same amount of information about the unknown intervention size ν, whereas for $\delta < 1$ this amount decreases for observations which are far apart from the time of intervention τ. Hence, estimation of the parameter ν becomes quite challenging in these situations because the log-likelihood function is very flat in the direction of the parameter ν. In our empirical works we observed, for a considerable number of cases, that the algorithm used for optimization had not moved away from its starting value, although the log-likelihood function was not even close to its maximum there. This happens especially when the sample size is large. In most cases, we can overcome this burden by choosing a very strict stopping rule for the optimization algorithm, which leads to a considerably longer computation time but on the other hand also to reasonable results.

The CML estimation of the INGARCH parameters β_0, β_1 and α_1 turns out to be mean square consistent in our simulations under intervention model (ii), as illustrated in Figure 3.5 for the case of a TS. Regarding the parameter ν, with growing sample size the RMSE decays very slowly except if $\delta = 1$, and it does not get close to zero even for a sample size of 10 000 observations. This is in accordance with the above remarks and is also the case in simulations for intervention model (i) which are not shown here.

Based on the asymptotic distribution in (3.4) we can obtain approximative standard errors for the estimated parameters. Another approach is based on a parametric bootstrap, i.e simulate B time series from the fitted model, fit the model to each of this B artificial time series and compute the empirical standard deviation of the parameter estimators

		β_0	β_1	α_1	ν
$n = 200$	Norm. approx.	0.94 (0.227)	0.07 (0.004)	0.13 (0.018)	3.00 (0.359)
	Bootstrap	1.06 (0.220)	0.07 (0.003)	0.14 (0.023)	2.95 (0.273)
500	Norm. approx.	0.62 (0.100)	0.04 (0.002)	0.08 (0.008)	2.80 (0.315)
	Bootstrap	0.67 (0.098)	0.04 (0.001)	0.09 (0.009)	2.77 (0.257)

Table 3.1: Standard errors of the parameters of an INGARCH(1,1) process with a transient shift of size $\nu = 9$ in the center according to model (ii). Bootstrap standard errors are based on $B = 1000$ random samples. Simulation results are averaged over 150 replications with standard deviations given in parentheses.

(see Section 2.2.2). We compare both approaches in a simulation study, see Table 3.3.2. Standard errors for β_0 and α_1 based on the approximate normality are on average slightly lower than those based on the parametric bootstrap method. For the other parameters the average standard errors are pretty close to each other. The normal approximation standard errors suffer from the issue discussed before. To be on the safe side we recommend to rely on bootstrap standard errors whenever the substantially longer computation time is acceptable.

3.4 Testing for intervention effects

In this section we modify the procedures proposed by Fokianos and Fried (2010) for the internal intervention model (i) for the case of the external intervention model (ii), using the conditional score function and information matrix presented in Section 3.3. We extend the simulations presented by Fokianos and Fried (2010) in the new context, verify that these procedures are also valid for the external intervention model (ii) and include further comparisons.

3.4.1 Intervention of known type at known time

The presence of an intervention effect of a given type (i.e. δ known and fixed) occurring at a known time τ can be checked by a score test, which only requires fitting a model under the null hypothesis of no intervention unlike likelihood ratio or Wald tests which are based on the alternative hypotheses. The null hypothesis $H_0 : \nu = 0$ is tested against the alternative $H_1 : \nu \neq 0$. Let $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\eta}}^\top, 0)^\top$, with $\widetilde{\boldsymbol{\eta}} = (\widetilde{\beta}_0, \widetilde{\beta}_1, \ldots, \widetilde{\beta}_p, \widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_q)^\top$ be the CML estimator under the null model, which is the intervention-free model (1.2). The score test statistic is given by

	$\tau =$	$0.25n$			$0.5n$			$0.75n$			
					Significance level						
		1.0	5.0	10.0	1.0	5.0	10.0	1.0	5.0	10.0	*Type*
$n =$	200	1.3	4.9	10.3	0.8	4.0	8.8	1.3	4.8	9.7	spiky
	500	0.3	5.4	10.9	0.7	4.4	9.0	1.2	4.1	9.2	outlier
	1000	0.7	3.3	8.1	1.5	6.4	11.1	0.7	4.4	8.2	
	5000	1.9	5.3	9.6	0.9	5.7	10.2	0.8	5.0	10.3	
	200	1.4	5.2	10.4	0.9	5.0	10.1	1.5	5.7	11.5	transient
	500	0.6	4.2	9.2	1.3	4.9	10.3	1.5	6.8	12.1	shift
	1000	0.5	4.5	10.1	0.7	6.0	10.0	0.9	4.9	10.8	
	5000	1.2	5.4	10.2	0.9	5.3	10.9	1.2	4.6	9.8	
	200	1.6	5.6	11.3	0.8	6.4	12.9	1.3	6.5	12.2	level
	500	1.5	6.0	10.6	1.2	5.0	10.6	1.8	5.7	11.0	shift
	1000	0.7	4.5	9.1	0.8	5.1	8.7	0.9	5.6	10.1	
	5000	0.8	3.3	7.4	0.5	4.3	10.2	1.0	5.5	10.8	

Table 3.2: Size (in percent) of the test for an intervention of a given type occurring at known time from model (ii) under the null hypothesis of no intervention averaged over 1000 replications. We varied the type of intervention we tested for, the position of the outlier τ and the sample size n.

$$T(\tau) = [S_{n\tau}(\widetilde{\boldsymbol{\theta}}; \boldsymbol{Z})]^{\top}\ G_{n\tau}^{-1}(\widetilde{\boldsymbol{\theta}})\ [S_{n\tau}(\widetilde{\boldsymbol{\theta}}; \boldsymbol{Z})].$$

Note that the score vector $S_{n\tau}(\boldsymbol{\theta}; Z)$ and the information matrix $G_{n\tau}(\boldsymbol{\theta})$ depend on δ and τ. Under the null hypothesis $H_0 : \nu = 0$ the test statistic $T(\tau)$ converges in distribution to a chi-square distribution with one degree of freedom, as $n \to \infty$, provided that certain regularity conditions hold (Fokianos and Fried, 2010, Lemma 1). This yields an asymptotic test for this hypothesis, rejecting for large values of the test statistic.

We examine the finite sample behavior of this test procedure for intervention model (ii) by simulation. Table 3.4.1 gives the averaged observed significance levels under the null hypothesis. These results do not reveal any large deviations between the achieved and the nominal significance levels 1%, 5% and 10% for various sample sizes, types and positions of interventions. This supports that the chi-square approximation of the test statistic is adequate also for intervention model (ii).

The bold lines in Figure 3.6 give the average simulated power of the test for an intervention following model (ii). From the left column we see that naturally interventions of larger size are detected better than interventions of smaller size. We also see that a level shift (even of a much lower size) is easier to detect than a transient shift, which is in turn easier to detect than a spiky outlier. The reason for this is that larger values of δ yield more information about the intervention to the subsequent observations. In a situation with a certain type of intervention the corresponding test for this type of intervention has the highest power, among all other test statistics. This will be important for classifying

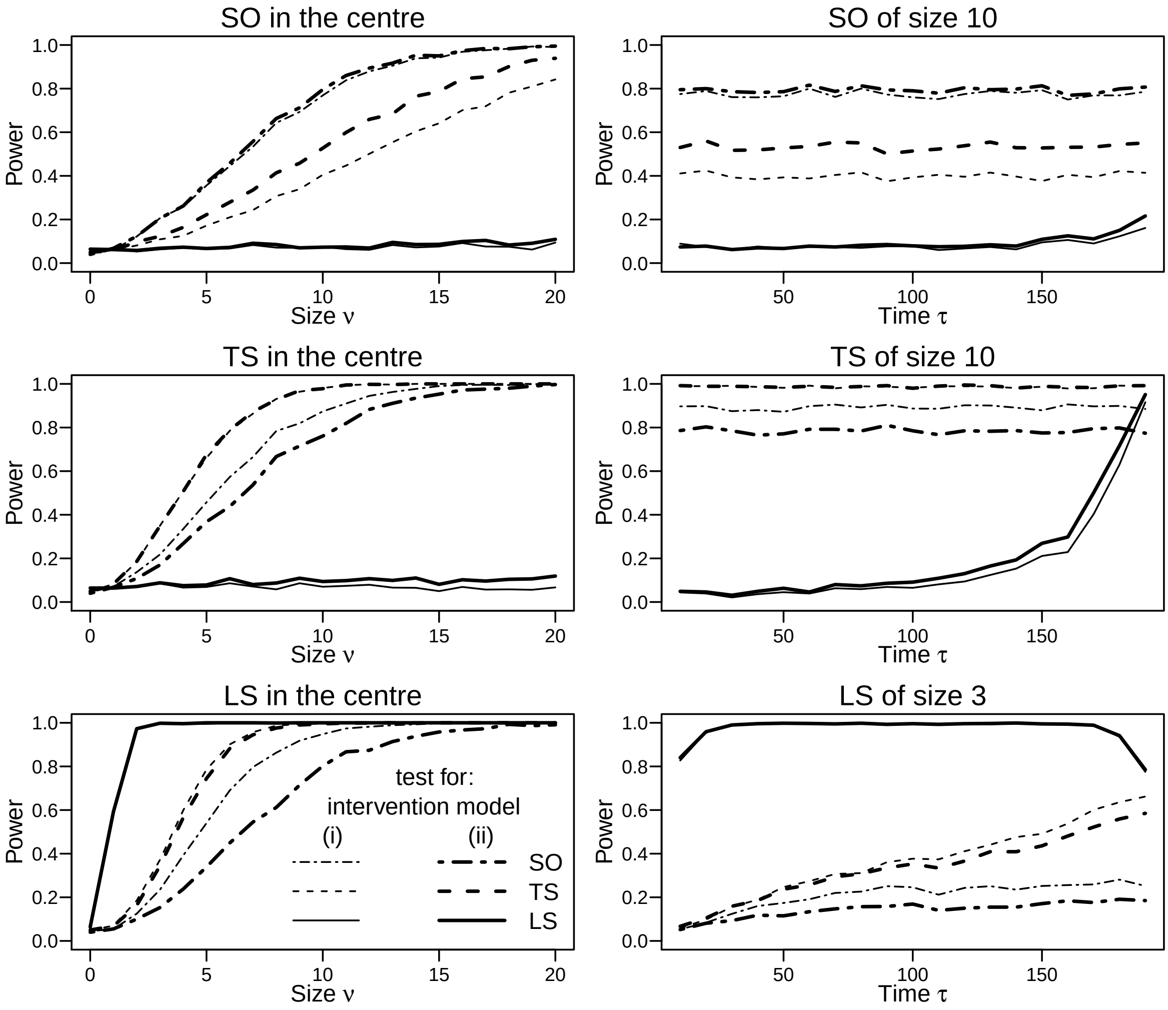


Figure 3.6: Power of the test for an intervention of a given type at known time from a given intervention model (see legend at the bottom) averaged over 1000 replications. The data are simulated under the alternative of an intervention of type spiky outlier (top), transient shift (middle) and level shift (bottom) from model (ii). In the left column the position of the intervention is fixed in the center of the time series and the size ν of the intervention varies. In the right column the time τ of the intervention varies and the size of the intervention is fixed. The time series are of length $n = 200$.

the type of an intervention when it is not known, as it is done in Section 3.4.3. The right column shows that for interventions at the end of a time series, it is more challenging to distinguish a level shift from a transient shift or spiky outlier, since we have little information on the evolution of the time series after the time of intervention.

For further insight, we run simulations based on intervention model (i) with our standard parameter setting as before. The results for both models are quite similar, but as a general observation, it is easier to detect interventions from model (i), since interventions from this model have stronger effects. Model (i) and (ii) represent different forms of intervention effects. We will examine the consequences of misspecifying an intervention model in Section 3.4.4.

3.4.2 Intervention of known type at unknown time

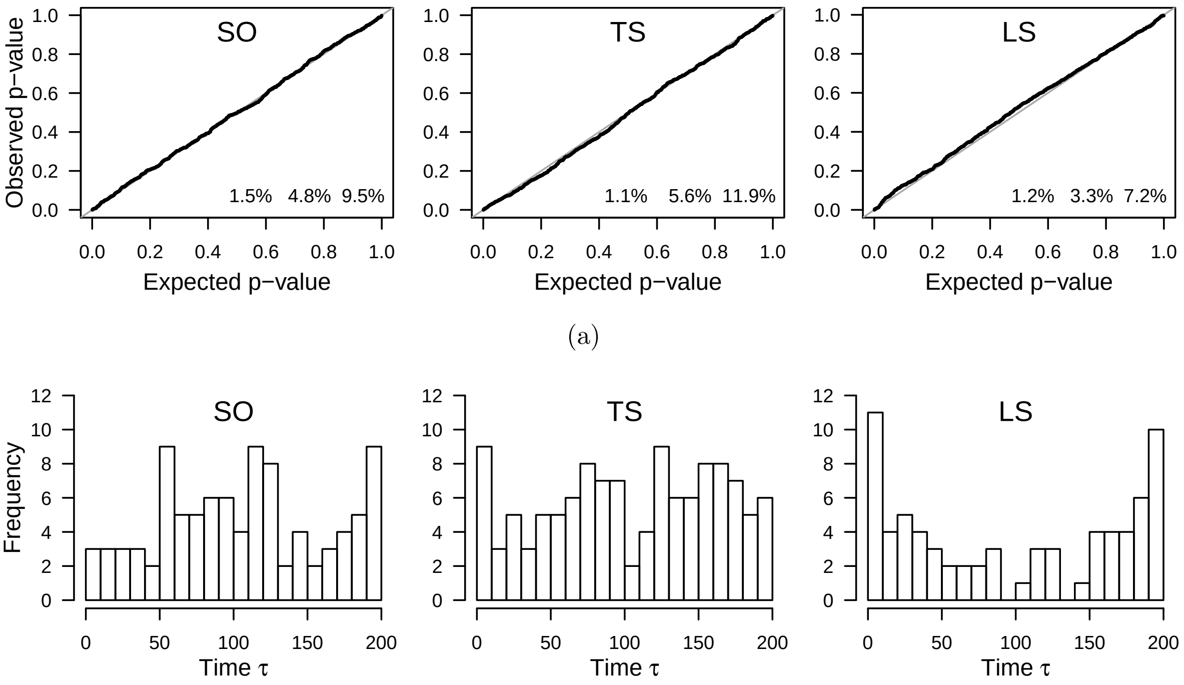
An intervention of a given type occurring at unknown time can be detected using the test statistic

$$T^* = \max_{\tau \in D} T(\tau),$$

i.e. the maximum of the test statistics for an intervention at each possible time. Denote the time of a possible intervention by $\tau^* = \arg\max_{\tau \in D} T(\tau)$. *A priori* knowledge about the time where an intervention might occur could be included by restricting the maximization to a predefined smaller set D of values for τ, which is chosen to be $D = \{2, \ldots, n\}$ by default. The p value of a test for the hypothesis $H_0 : \nu = 0 \; \forall \tau \in D$ is approximated by a parametric bootstrap procedure as proposed by Fokianos and Fried (2010):

1. Generate B bootstrap replicates from an intervention-free INGARCH model with parameter vector $\widetilde{\boldsymbol{\theta}}$, the estimate under H_0 defined in the previous section.
2. Compute the test statistic T_b^* for each bootstrap replicate $b = 1, \ldots, B$.
3. The p value is given by the number of bootstrapped test statistics at least as large as the original test statistic, divided by $B + 1$.

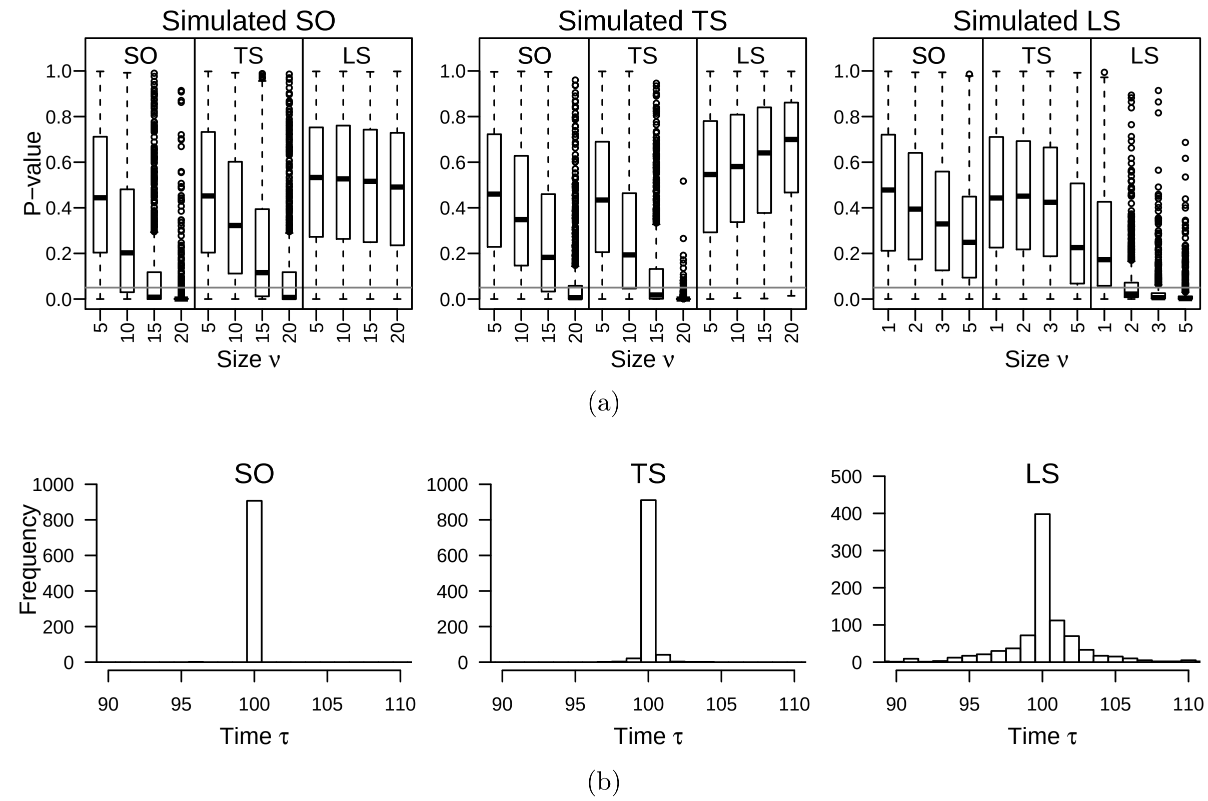
We choose $B = 500$ for our study. Our simulations presented in Figure 3.7 (a) support that this testing procedure behaves well also for intervention model (ii). Under the null hypothesis the p values show a reasonable approximation to the uniform distribution and the achieved significance levels do not deviate from the respective nominal significance levels much. The positions of the erroneously detected interventions in Figure 3.7 (b) show no peculiarities for the test for SO and TS. The test on a LS detects more interventions



(a)

(b)

Figure 3.7: (a) Size of the test for an intervention of given type (see title) at unknown time from intervention model (ii) using $B = 500$ bootstrap samples. The data are simulated under the null hypothesis of no intervention. Simulation results are averaged over 1000 repetitions. A point above the bisecting line indicates that the test is conservative for this particular size. The percentages given at the bottom of each panel are the achieved significance levels for a nominal level of 1%, 5% and 10%, respectively. The time series are of length $n = 200$.
(b) Detected times of intervention τ^* for tests with a p value below 10% in the simulations above.



(a)

(b)

Figure 3.8: (a) Power of the test for an intervention of given type (see segments within each plot) at unknown time from intervention model (ii) using $B = 500$ bootstrap samples. The data are simulated under the alternative of an intervention of given type (see title) from model (ii) in the center of a time series of length $n = 200$. We vary the size of the intervention (see horizontal axis). Simulation results are averaged over 1000 repetitions. (b) Detected times of intervention τ^* for tests with a p value below 10% in the simulations above. The simulated intervention is of the same type which is also tested for (see title) and is of size 20 for SO and TS and of size 5 for LS.

at the beginning and at the end of the time series. For level shifts at these positions there are only very few observations available for estimating the level before and after the shift, respectively. Therefore randomly occurring extreme observations can hardly be distinguished from a LS.

To investigate the power of the test we consider the simulations shown in Figure 3.8 (a). A test for a SO requires an intervention size of about $\nu = 15$ to detect it right in more than 50% of the cases (see left segment in the left plot). For a test regarding a LS, a size of $\nu = 2$ of a LS is needed to attain a detection above 50% (see right segment in the right plot). Note that a LS of size ν does imply a shift in the marginal mean by $\nu/(1 - \beta_1 - \alpha_1)$, i.e. by about 6.6 in this simulation with $\nu = 2$. The timing of the intervention shown in Figure 3.8 (b) is very accurate for SO and TS (over 85% of the

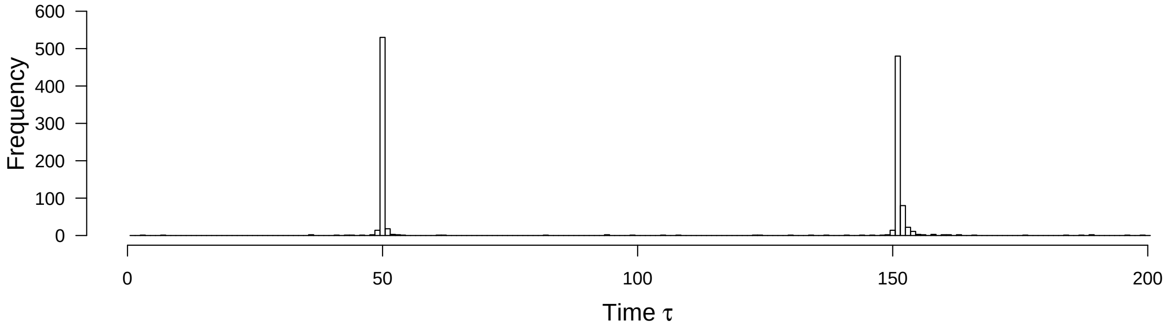


Figure 3.9: Times of interventions found by the iterative detection procedure for model (ii) using $B = 500$ bootstrap samples. The data are simulated with two transient shifts from model (ii) at time 50 ($\delta = 0.7$, size $\nu = 12$) and at time 151 ($\delta = 0.9$, size $\nu = 9$). The time series are of length $n = 200$. Simulation results are based on 1000 repetitions.

cases are detected correctly) and somewhat less accurate for the LS (about 40% detected correctly, the other cases are up to ten observations away from the true value).

3.4.3 Multiple interventions of unknown type at unknown time

A common situation in several applications is that both the position and the type of intervention are unknown. From Figure 3.8 (a) we get an idea that it is possible to classify the type of an intervention by the minimum p value of all (three) tests. On average the p values of the tests for the matching type are lower than the p values of the tests on the other types. To overcome the problem of multiple testing we use a Bonferroni correction, i.e. multiply the p values by the number of intervention types considered. The intervention type is classified according to the lowest p value, if below the chosen significance level. In case of equal p values we prefer a LS with $\delta = 1$ (since its effect is usually more dominant) and then opt for the intervention type with the highest test statistic, as in Fokianos and Fried (2010).

Furthermore, there might be more than one intervention in a time series. Fokianos and Fried (2010) propose a stepwise procedure to detect, classify and eliminate multiple intervention effects in a time series. The elimination of intervention effects is based on (3.2). We adopt their iterative procedure for the alternative intervention model (ii) by using the modified equations from Section 3.2. We refer to their paper for details and show some simulation results for the alternative intervention model (ii).

We apply the iterative detection procedure to simulated time series with two transient shifts at time 50 with size $\nu = 12$ and $\delta = 0.7$, and at time 151 with $\nu = 9$ and $\delta = 0.9$. In 27% of the 1000 repetitions we find both interventions exactly at the time of their occurrence, and no others. Each of the interventions is found in around 50% of the cases, at least one of them in about 72%. Notably, only in about 20% of the repetitions we detect an intervention at a wrong time and this happens quite often close to the actual times of interventions, see Figure 3.9. Although we test for interventions with $\delta = 0.8$ and thus slightly misspecify their type, our results are satisfactory. There are no systematic simulations of Fokianos and Fried (2010) for intervention model (i), but their examples point into the same direction.

3.4.4 Misspecification of the intervention model

An important question is the effect of model misspecification on the detection procedure. Consider the situation with an intervention of a given type occurring at a given time. We look at the thin lines in Figure 3.6 for the test on an intervention of model (i) and compare them with the bold ones of the same kind, which stand for the test based on model (ii) from which the data are in fact simulated. We note that both tests are comparable and misspecification of the data generating process does not affect the power a lot. This also applies when model (ii) is the true data generating process, as we have confirmed by simulations not reported here. Hence, there is some robustness against misspecification of the intervention model.

Another interesting finding from Figures 3.6 is that for example a transient shift from model (ii) is somewhat similar to a spiky outlier from model (i). In general, an intervention effect from model (ii) resembles somewhat one from model (i) with a slightly lower value of δ and vice versa. This is in line with the explanations given in Section 3.2.

We also study the effect of misspecification of the intervention model on the iterative procedure for detection of multiple interventions. Applying the detection procedure for internal interventions to 1000 time series from the external intervention model (ii), identical to those used in Section 3.4.3, yields almost the same rates of truly and falsely detected interventions as for the correctly specified situation. For the converse situation, i.e. data generation according to the internal model (i) and detection assuming the external model (ii), we detect both interventions correctly in 40% of the cases, which is more than a third higher than for data generation from the external model. On the other hand we find interventions at wrong times in another 40% of the cases, compared to 20% when simulating from the external model. In a nutshell, the above empirical results suggest that a successful identification procedure for interventions is based on the amount of information that the data carry. The choice of the model is of secondary importance.

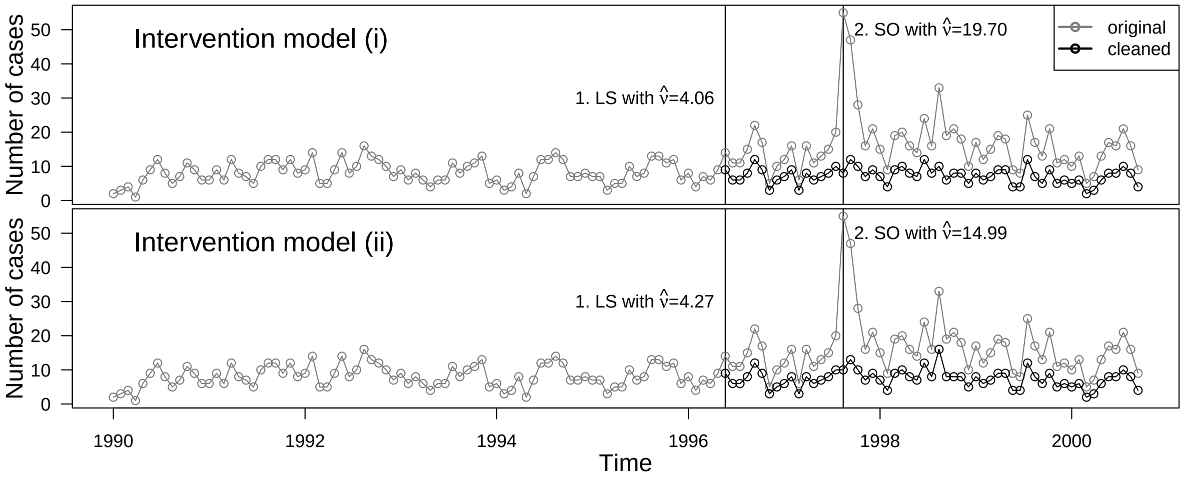


Figure 3.10: Number of campylobacterosis infections (reported every 28 days) in the north of Québec in Canada. The original time series is displayed in grey, the modified time series cleaned by the procedure from Section 3.4.3 in black.

A related question is whether we can deduce from the data which intervention model is more suitable for describing a certain data pattern. Some experiments with classification based on the out-of-sample prediction error, as suggested by a reviewer, did not provide good results.

3.5 Real data application

We again study the time series of campylobacterosis cases considered in Section 2.6.1. The original time series with $n = 140$ observations is shown in grey in Figure 3.10. An infection with the bacterium Campylobacter may be caused mainly by contaminated food but also by contact with infected animals. Human to human spread is possible particularly for infants. The incubation period is typically 2-5 days. Infected patients often do not show any symptoms and are potentially infectious via their excrements for 2-4 weeks on average. The number of cases is higher in warm seasons. More details on modeling Campylobacter infections for epidemiological surveillance (applied to data from Germany) are given by Manitz and Höhle (2013).

Following Ferland *et al.* (2006) we fit an INGARCH(1,13) model to the data with $\alpha_1 = \cdots = \alpha_{12} = 0$ and the regression on the conditional mean for lag 13 accounting for seasonal variation.

Since we do not have any prior information on the occurrence of possible interventions, we apply the iterative detection procedure from Section 3.4.3 to the data. We first search for interventions following our new intervention model (ii). In the first step, the bootstrapped p values for all considered types (SO, TS with $\delta = 0.8$ and LS) are zero, using $B = 500$ bootstrap replicates. The procedure decides in favor of the LS with an estimated size of 4.27 found at time 84, following the classification rule given in Section 3.4.3. The detected intervention effect is removed from the time series according to the decomposition presented in Section 3.2, applying the procedure proposed by Fokianos and Fried (2010). In a second step, both the test on a SO and on a TS give a p value of zero at time 100, the same position already found in the first step for these types of interventions. We decide in favor of the SO with an estimated size of 14.99, because it has a higher value of the test statistic than the TS. Again, the time series is cleaned from the detected intervention effect. In a third step, no further interventions with a p value lower than 5% are detected and the procedure stops with two intervention effects found. The estimated parameters for the cleaned time series are $\widehat{\beta}_0 = 2.28$, $\widehat{\beta}_1 = 0.36$ and $\widehat{\alpha}_{13} = 0.35$. The cleaned time series and both interventions are shown in Figure 3.10 (bottom).

Finally we fit the full model with both detected interventions to the original time series and obtain

$$\begin{aligned}
&Z_t|\mathcal{F}_{t-1}^{Z,\kappa} \sim \text{Poisson}(\kappa_t),\\
&\kappa_t = \lambda_t + 3.52(0.80)\, I_{[84,\infty)}(t) + 36.67(7.31)\, I_{\{100\}}(t),\\
&\lambda_t = 2.25(0.58) + 0.36(0.06)\, Z_{t-1} + 0.35(0.08)\, \lambda_{t-13}.
\end{aligned} \tag{3.5}$$

The jointly estimated parameters in (3.5) clearly differ from those by the iterative detection procedure reported before. The standard errors of the regression coefficients based on the normal approximation (3.4) are given in parentheses. Standard errors obtained by a parametric bootstrap based on 1000 replications are 1.27 ($\widehat{\beta}_0$), 0.07 ($\widehat{\beta}_1$), 0.17 ($\widehat{\alpha}_{13}$), 0.98 (LS at time 84) and 7.65 (SO at time 100). Note that these standard errors do not reflect the additional model uncertainty induced by the intervention detection procedure for either of the two approaches. The bootstrapped standard errors largely agree with the normal approximation ones, but are much higher for $\widehat{\beta}_0$ and $\widehat{\alpha}_{13}$. Because of the long delay of 13 time points the estimation of α_{13} can make less use of the data and is, as a regression coefficient of an unobserved variable, more difficult to estimate anyway. Estimation of the intercept β_0 and α_{13} strongly interfere. In this challenging situation with a small sample size the normal approximation standard errors do not express this uncertainty adequately. However, for a simpler INGARCH(1,1) process their performance was quite good in case of 200 or more observations (see Section 3.3.2).

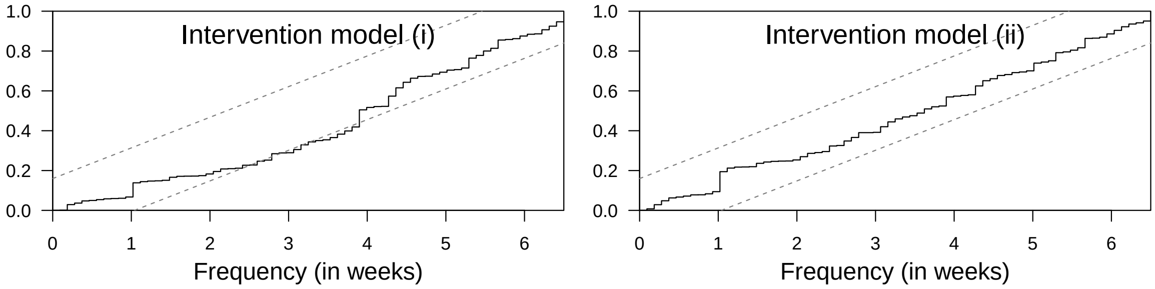


Figure 3.11: Cumulative periodogram of the Pearson residuals of model (3.5) for the campylobacterosis infections time series (right). For intervention model (i) the same detection procedure has found the same two intervention effects and a full model including both interventions is fitted to the data (left). The dashed lines give approximative 95% confidence limits of a Kolmororov-Smirnov test on a constant spectral density (cf. Venables and Ripley, 2002, p. 396), which are used as a visual check for uncorrelated residuals.

We also apply this procedure for intervention model (i) and find the same types of interventions at the same times with only slightly different estimated intervention sizes. These results are presented in Figure 3.10 (top). The average of the squared Pearson residuals for the fitted models with both interventions is 18.8 for model (i), compared to 19.3 for our new intervention model (ii). An INGARCH model without intervention effects has a much higher value of 30.9. Hence the internal intervention model (i) fits in this sense a little better to the data than the external one and both intervention models clearly outperform a model without intervention effects. Note that the cumulative periodograms in Figure 3.11 show that there is less structure left in the residuals of our new intervention model (ii) than for intervention model (i). However, both models might suffer from lacking covariates. Manitz and Höhle (2013) list some studies which suggest an association of campylobacterosis cases with certain weather conditions like absolute humidity, on which we do not have information for our data example.

3.6 Discussion

In this chapter we study a new variant of an intervention model for INGARCH processes and adopt a maximum likelihood approach for detection and estimation of intervention effects proposed by Fokianos and Fried (2010) to this new model. With the obtained procedures one can test for intervention effects at given times but also search for intervention effects at unknown times. Simulations support that the presented procedures work quite reliably. In comparison to the intervention effects studied by Fokianos and Fried (2010), those in the new model have less influence on subsequent observations after the occurrence of an intervention. The new model describes an external effect which propagates only via the observations, whereas an intervention effect from the old model can be seen as an internal change of the underlying mean. Our application to the weekly number of campylobacterosis cases illustrates the usefulness of our new model.

We find some robustness against misspecification of the intervention model. Thus one can expect a reasonable performance for this kind of intervention effects applying either of the two models. On the other hand this finding suggests that a test to discriminate the intervention models might have low power.

One undesirable feature of the procedures for detection of interventions are the high computational costs. For the example in Section 3.5 with $B = 500$ bootstrap samples and $n = 140$ observations each step of the iterative detection procedure runs about 72 minutes for intervention model (i) and about 75 minutes for model (ii) on a single processor unit (Intel Xeon CPU with 2.83 GHz). The reason for the long computation time is the implementation of the parametric bootstrap. Fokianos and Fried (2012) reduced the computation time of the bootstrap considerably by not estimating the parameter vector for each bootstrap replicate but using the true value used for its generation instead. This modification results in quite conservative procedures. With parallelized computation of the bootstrap we have greatly shortened its duration using multiple processors.

This chapter has shown that there are capable tools available for modeling and retrospective detection of intervention effects in count time series based on INGARCH models. These methods have been generalized to the framework of count time series following generalized linear models and are included in the R package **tscount**; see Section 2.5. In many applications it is not sufficient to detect intervention effects retrospectively, i.e. once the complete time series has been observed. Instead, the goal is to detect intervention effects in real-time. This problem of prospective intervention detection is the topic of the next chapter.

Chapter 4

Online monitoring in the context of infectious disease surveillance

4.1 Introduction

Prospective detection of aberrations in count time series is of interest in various applications. For example in public health, surveillance of infectious diseases aims at timely recognizing outbreaks of epidemics for taking proper actions promptly. Other possible applications are industrial quality control, where an unusually large number of faulty pieces may indicate a flawed production process (Weiß, 2015), or finance, where exceptionally large numbers of transactions may point to anomalies at the stock market. Monitoring procedures need to be applicable in an online manner such that the decision that an aberration has occurred is made in real-time with only short time delays. The statistical methods employed in this context have their roots in the literature on statistical process control (SPC), time series analysis, sequential testing and change point detection.

The focus of this chapter is on the application of these statistical methods to the problem of infectious disease surveillance. However, most of the considerations in this study are meant to be of general relevance and do also apply to many data problems beyond public health. Recent introduction to the topic of infectious disease surveillance are given by Unkel, Farrington, Garthwaite, Robertson, and Andrews (2012) and Held and Paul (2013). Salmon, Schumacher, Burmann, Frank, Claus, and Höhle (2016a) describe a system for automated outbreak detection which employs the R package **surveillance** (Salmon *et al.*, 2016b). We summarize the crucial aspects of infectious disease surveillance. Its principal aim is to timely recognize disease outbreaks which manifest by an unusually large number of infections. As a data basis one constantly observes the number of patients per day, week or month which have been infected by a certain pathogen. These time series are typically collected by public health authorities and they are available for a number of different pathogens. For each pathogen, the data is subdivided into different geographical units and sometimes further into age groups. Consequently, there may be thousands of time series under surveillance, requiring the use of automatic procedures for outbreak detection which not require manual (and thus subjective) decisions. Monitoring procedures are also referred to as control charts because they were originally employed as graphical tools.

The general goal of a monitoring procedure is to timely detect potential aberrations in a time series whilst not giving too many false alarms. An aberration is a sudden or gradual change of the data generating process (DGP) occurring at an unknown time point, when the process is said to be *out-of-control*. This change can basically be of any form. The outbreak of a disease for instance is expected to result in an increase of the location and possibly also of the strength of temporal dependence (Höhle and Paul, 2008, Section 4). Notably, control charts based on the likelihood ratio are designed to optimally detect specific out-of-control scenarios (Höhle and Paul, 2008; Weiß and Testik, 2012). In this study we focus on methods which particularly aim at detecting location increases but are not designed for a specific out-of control scenario.

In order to assess whether the DGP has changed one needs to learn about its characteristics before the aberration, when the process is assumed to be *in-control*. This can be done by assuming a parametric model for the time series before a potential change. In the applications we are aiming at, the model parameters are in fact unknown and need to be estimated using historical data. The setting for our study is to have observed a time series up to time n and to start monitoring future observations at time $n + 1$. The time period of the first n observations used for learning about the DGP is referred to as the *set-up phase* (phase I). The time period thereafter, in which the actual monitoring takes place, is referred to as the *operational phase* (phase II).

The characteristics of infectious disease time series may vary with the respective pathogen, geographical unit or age group. Typical features of this kind of data are seasonal variation (caused by environmental effects, varying severity of pathogens and other factors like public holidays), trends, temporal dependence (possibly resulting from the disease transmission mechanism) and the occurrence of irregularities (such as disease outbreaks). The number of infected persons is also driven by the size and structure of the population at risk. Note that the number of reported cases is generally lower than the actual incidence. The extent of this under-reporting is not constant over time but follows a seasonal pattern itself and also reacts to singular events (e.g., increased media coverage on an infectious disease). This may cause seasonal effects as well as irregular artifacts in the observed number of cases. The various aspects of modeling seasonality are studied by Held and Paul (2012).

Classical control charts for count data are based on the assumption of independent and identically distributed observations from a distribution which is fully specified (Gan, 1990). Such procedures do not account for the typical characteristics of infectious disease time series which are described in the previous paragraph. In the context of infectious disease surveillance, monitoring procedures for non-identically distributed observations based on GLMs have been proposed; so-called regression charts which are able to consider trends and seasonal effects (Farrington, Andrews, Beale, and Catchpole, 1996; Rossi, Lampugnani, and Marchi, 1999; Höhle and Paul, 2008; Noufaily, Enki, Farrington, Garthwaite, Andrews, and Charlett, 2013). However, these procedures still assume independent observations. In the recent literature on control charts for count data the case of dependent observations has received more attention (Weiß and Testik, 2012; Manitz and Höhle, 2013). These methods do neither require the observations to be identically distributed nor to be independent. Most research has been on methods based on the marginal distribution of an observation.

In this study we investigate how accommodating temporal dependence can enhance monitoring procedures. We introduce a monitoring procedure which is based on one-step ahead predictions, i.e. the conditional distribution of an observation given the past. Our main finding is that such a procedure can substantially improve the immediate detection of outbreaks (compared to procedures based on the marginal distribution) but that its dependence on previous observations may also yield undesired effects in some situations. Our monitoring procedure assumes that the DGP is from the class of count time series following GLMs as it is presented in Section 1.2. Models from this appealing and flexible class are able to capture typical characteristics of infectious disease data like serial dependence and seasonality. These models are formulated recursively, describing the underlying mean of the time series conditionally on the past. Serial dependence is included by regression on past observations and on past values of the conditional mean. The latter is called feedback mechanism and helps to model serial dependence parsimoniously.

Notwithstanding the model which is assumed for the DGP, there are different types of test statistics which are used to built control charts for count time series: the observations themselves (Manitz and Höhle, 2013), transformed residuals (Rossi *et al.*, 1999; Noufaily *et al.*, 2013), (generalized) likelihood ratios (Höhle and Paul, 2008; Weiß and Testik, 2012) or an estimator of the model parameters. Moreover it is important to distinguish between different construction principles for control charts. A *Shewhart-type* control chart bases its decision on the information of the current time point only, whilst a *CUSUM-type* control chart accumulates information of successive time points. The former type has more power than the latter type to detect large aberrations without delay. Conversely, CUSUM-type charts have more power than Shewhart-type charts to detect moderate aberrations, albeit with some delay. Our simulations show that monitoring based on models with temporal dependence is particularly good for immediate detection, but, at least in our current implementation, not so good for delayed detection of outbreaks. We therefore decide to build a Shewhart-type monitoring procedure, which is also the state of the art at the German infectious disease surveillance system at the Robert Koch institute (see for example Salmon *et al.*, 2016a).

This chapter is organised as follows. Section 4.2 very briefly describes the GLM-based count time series models for the in-control process and compares them to other proposals from the literature. We discuss in some more detail how to model seasonality. Section 4.3 introduces a monitoring procedure based on one-step-ahead predictions. Section 4.4 studies the performance of the monitoring procedure by simulations. We compare the results of the procedure for models with and without temporal dependence in different data situations. We discuss benefits but also pitfalls of using models with temporal dependence for monitoring. In Section 4.5 we present a comprehensive case study using a data example from infectious disease surveillance, again considering models with and without temporal dependence. We demonstrate that these data in fact exhibit temporal dependence. We illustrate the process of model selection and fitting in the set-up phase and apply the monitoring procedure in the operational phase. Section 4.6 summarizes our results and gives an outlook on further directions of research.

4.2 Models for the in-control process

For low counts, standard models based on the assumption of normality, or at least continuously and symmetrically distributed data, are often inappropriate even after suitable transformation (cf. Held and Paul, 2013). Models which have been particularly tailored for count data are usually more appropriate. In the context of infectious disease data, several count data models have been proposed. Many of those assume that the observations are independently but not identically distributed (e.g. Noufaily *et al.*, 2013). However, such models are not adequate if there is a considerable amount of serial dependence among data, as it is usually observed in practice. A comprehensive review of models for count time series which are able to describe serial dependence is given in

Section 2.7. In this section we review models which have been proposed particularly for modeling infectious disease time series. All these models originate from the idea of a generalized linear model (GLM).

We first introduce some notation. Let $\{y_t : t \in \mathbb{N}\}$ be a univariate count time series to be monitored, where y_t is e.g. the number of registered infections in week t. We assume this observed count time series to be a realization of the stochastic process $\{Y_t : t \in \mathbb{N}\}$. In some cases there is additional covariate information available, for instance on weather conditions. Let $\{\boldsymbol{X}_t : t \in \mathbb{N}\}$ be a corresponding time-varying r-dimensional covariate vector, i.e., $\boldsymbol{X}_t = (X_{t,1}, \dots, X_{t,r})^\top$. Denote by $\mathcal{F}_{t_0}$ the information on past observations up to time t_0 as well as on covariates up to time $t_0 + 1$. All presented models assume that $Y_t|\mathcal{F}_{t-1}$ has a Negative Binomial or Poisson distribution but differ in the way they specify the conditional mean $\mathsf{E}\left(Y_t|\mathcal{F}_{t-1}\right) = \lambda_t$.

In the model proposed by Held, Höhle, and Hofmann (2005) the conditional mean is a sum of a deterministic (endemic) component and a first order autoregressive (epidemic) component and is given by

$$\lambda_t = \exp\left(\beta_0 + \delta t + \sum_{s=1}^{S} (\gamma_{1,s}\sin(\omega_s t) + \gamma_{2,s}\cos(\omega_s t))\right) + \beta_1 Y_{t-1},$$

where $\beta_0, \delta, \gamma_{1,1}, \dots, \gamma_{1,S}, \gamma_{2,1}, \dots, \gamma_{2,S}, \beta_1$ are unknown parameters but S is fixed. The endemic component includes a linear trend and a deterministic seasonal figure with frequency ω_s. The exponential function is applied to the covariate effects but not to the autoregressive component such that this model is not a GLM (since the linear predictor is not linear in the parameters) but could instead be regarded as a generalized additive model (GAM). Wei, Schüpbach, and Held (2015) additionally include the effect of time-varying covariates either to the endemic or to the epidemic component. Inference is based on the likelihood approach.

Another proposal by Manitz and Höhle (2013) based on Heisterkamp *et al.* (2006) is a hierarchical time series (HTS) model (see Section 2.7.2). This model does not explicitly have an autoregressive component but is able to account for serial dependence by a time-varying intercept. The conditional mean λ_t is given by

$$\lambda_t = \exp\left(\beta_{0,t} + \delta t + \gamma_t + \boldsymbol{\eta}^\top \boldsymbol{X}_t\right).$$

The time-varying intercept $\beta_{0,t}$ is assumed to depend on its previous values according to

$$\Delta_d \beta_{0,t}|\beta_{0,t-1}, \dots, \beta_{0,t-d} \sim \mathrm{N}(0, \kappa_{\beta_0}^{-1}),$$

where Δ_d is the difference operator of order $d \in \{0, 1, 2\}$ and κ_{β_0} a precision parameter. For $d = 0$ this yields an independence model but for higher orders d this induces dependence between successive observations. The other parameters, δ for the linear trend, γ_t for a seasonal effect and the vector $\boldsymbol{\eta}$ for the effect of a covariate vector $\boldsymbol{X}_t$, are also assumed to be normally distributed with certain priors. Inference is done in a Bayesian framework and utilizes an efficient integrated nested Laplace approximation (INLA) (Rue, Martino, and Chopin, 2009).

In our study the process $\{Y_t : t \in \mathbb{N}\}$ is assumed to belong to the class of count time series following generalized linear models as it has been introduced in Section 1.2. Recall that the regression equation specifying the conditional mean λ_t is of the general form given by (1.1), i.e.,

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \, \widetilde{g}(Y_{t-i_k}) + \sum_{\ell=1}^{q} \alpha_\ell g(\lambda_{t-j_\ell}) + \boldsymbol{\eta}^\top \boldsymbol{X}_t.$$

Here $g : \mathbb{R}^+ \to \mathbb{R}$ is a link function and $\widetilde{g} : \mathbb{R}^+ \to \mathbb{R}$ is a transformation. A choice of $g(x) = \widetilde{g}(x) = x$ yields the INGARCH model whereas choosing $g(x) = \log(x)$ and $\widetilde{g}(x) = \log(x + 1)$ yields the log-linear model studied by Fokianos and Tjøstheim (2011), Woodard *et al.* (2011) and Douc *et al.* (2013). Note that the INGARCH model with the identity link function accommodates only positive temporal correlation and requires the covariates to be nonnegative. In contrast, the log-linear model, which implies a multiplicative effect of the regressors on the response, is able to describe negative and positive temporal correlation and allows for covariates with negative values. The dependence on previous values of the mean, the so-called feedback mechanism, allows for parsimoniously modeling serial correlation. It can, to some extent, also model time-varying means and generally yields smoother models than only with dependence on previous observations. Denote by $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_q, \eta_1, \ldots, \eta_r)^\top$ the vector of all regression parameters. The conditional distribution of Y_t given $\mathcal{F}_{t-1}$ is assumed to belong to the class of mixed Poisson processes (see Christou and Fokianos, 2015a). An important special case is to assume a Poisson distribution, i.e. $Y_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t)$. Another special case studied by Christou and Fokianos (2014) is to assume a Negative Binomial distribution parametrized in terms of its mean with an additional dispersion parameter $\phi \in (0, \infty)$, i.e. $Y_t|\mathcal{F}_{t-1} \sim \text{NegBin}(\lambda_t, \phi)$. This allows for more variability of the process, which is otherwise fully specified by the mean and the degree of serial dependence. The Poisson distribution is a limiting case for $\phi \to \infty$.

Compared to the model by Held *et al.* (2005), the class of GLM-based count time series used in this study allows for more general forms of serial correlation of higher order and with a feedback mechanism. By way of comparison to the model proposed by Manitz and Höhle (2013) recall our findings in Section 2.7.2: An advantage of their Bayesian approach is that prediction intervals which reflect both observation and estimation uncertainty can be obtained in a natural way. A disadvantage is the much higher computational effort, which is particularly a problem for a real-time application to thousands of time series. A fit of their model turns out to be very smooth (cf. Figure 2.5) and is thus expected to provide less accurate one-step-ahead predictions than models including explicit dependence on past observations.

Infectious disease time series often exhibit a seasonal pattern. There are several possibilities to account for seasonal effects within the class of count time series based on GLMs, see Kedem and Fokianos (2002, Section 4) and many textbooks on regression models. These approaches are briefly reviewed in the following paragraphs. The length of one season T is assumed to be known. An extension of the considered approaches to several superimposed periodicities of different lengths, say $T_1, \ldots, T_r$, would be straightforward.

One popular approach is to include deterministic periodic covariates, which implies that the seasonal pattern does not change over time. The most common choice for deterministic seasonality is a superposition of sinusoidal waves of the Fourier frequencies $\omega_s = 2\pi s/T$, $s = 1, \ldots, S$, where S is chosen based on a model selection criterion (see for example Held and Paul, 2012). For each value of s, the two covariates $X_{t,s(1)} = \sin(\omega_s t)$ and $X_{t,s(2)} = \cos(\omega_s t)$ with the corresponding coefficients $\eta_{s(1)}$ and $\eta_{s(2)}$, respectively, are added to the model. The resulting $2S$ covariates are linearly independent from each other and from a constant vector, which guarantees a unique and identifiable solution. Another, more flexible choice of deterministic periodic covariates would be a periodic B-splines basis for polynomial splines using the R package **pbs** (Wang, 2013), where a polynomial degree of three yields good smoothness and the number of knots K can be chosen based on a model selection criterion. The number of knots is equal to the number of covariates and hence the number of parameters which are introduced. The resulting K covariates $X_{t,1}, \ldots, X_{t_K}$ are nonnegative and not linearly independent from a constant vector such that the corresponding parameters interfere with the intercept parameter and there is not necessarily a unique solution. An example of periodic B splines with $K = 5$ knots is shown in Figure 4.1. A variant of the approach with deterministic covariates is employed by the improved Farrington method (Noufaily *et al.*, 2013), where dummy variables account for seasonality. However, the division of the whole period into windows is not fixed but depends on the week for which a prediction is needed (see also Salmon *et al.*, 2016b, p. 8). We do not consider this variant in our study because the time-varying windows conflict with the idea of a global seasonal pattern and because dummy variables usually introduce more parameters than the other types of periodic covariates. Note that

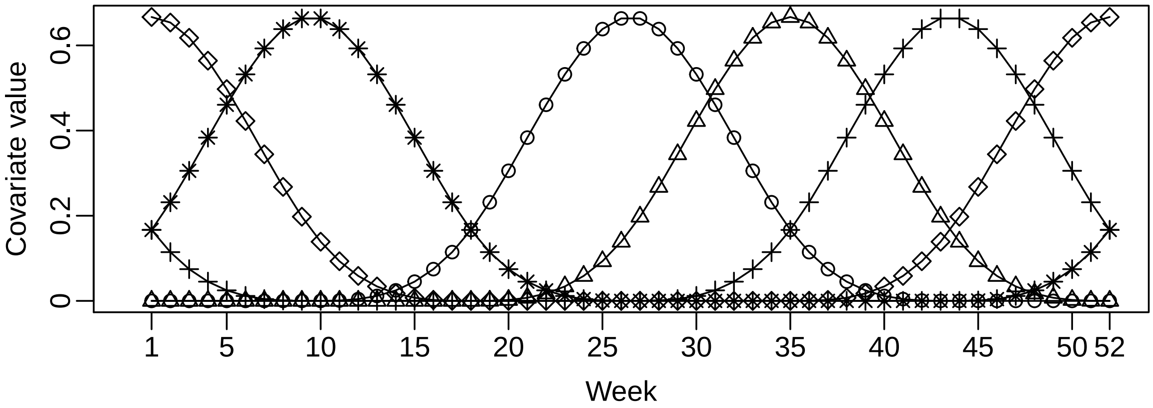


Figure 4.1: Covariates $X_{t,1}$ (circle), $X_{t,2}$ (triangle), $X_{t,1}$ (cross), $X_{t,1}$ (diamond) and $X_{t,1}$ (asterisk) of periodic B splines with $K = 5$ knots for a period length of $T = 52$. The covariates are periodic with $X_{t,k} = X_{t+52,k}$, $k = 1, \ldots, K$.

infectious disease time series do not only show smooth seasonal patterns but sometimes also sharp periodic effects that take place in only one or two weeks of the year (e.g. less cases around Christmas or New Year). Such effects can be described by a single dummy variables (see for example Salmon *et al.*, 2016b).

A second and completely different approach for modeling seasonality accommodates temporal dependence on observations (Y_{t-T}) and/or conditional means (λ_{t-T}) which are T time units (i.e. one period) back in time. This stochastic seasonality is employed by for instance Ferland *et al.* (2006). This approach allows to describe seasonal patterns of any shape which might possibly even change slightly over time. Stochastic seasonality is more flexible than deterministic one but tends to be less suitable for very pronounced seasonal patterns. A third approach pursued by Held and Paul (2012) explains seasonality by time-varying dependence parameters. We do not consider this approach in our study because it is quite complicated and difficult to interpret.

Fitting models of the form (1.1) can be done by quasi conditional maximum likelihood estimation of the unknown regression parameters $\boldsymbol{\theta}$. For the dispersion parameter ϕ of the Negative Binomial distribution one can use an estimator based on Pearson's χ^2 statistics as proposed by Christou and Fokianos (2014). For a sample of n observations, these estimators are denoted by $\widehat{\boldsymbol{\theta}}_n$ and $\widehat{\phi}_n$, respectively. Details on the estimation procedure including formulas, technical considerations regarding the implementation and properties of the estimators are given in Section 2.2.2.

One usually assumes that the process is in-control in the set-up phase. However, it might happen that this assumption is violated. A likely scenario is the presence of outbreaks in the set-up phase which have not been detected previously. Such outbreaks could be included in the above framework model by the intervention model studied by Fokianos and Fried (2010, 2012) and Liboschik *et al.* (2016), which describes intervention effects by deterministic covariates (see Section 2.5 and Chapter 3). Their approach also allows to search for retrospective detection of intervention effects of unknown type occurring at unknown time points. One could also consider robust approaches for model selection and fitting in the set-up phase. In Section 5.2 we investigate robust estimators of the (partial) autocorrelation for identifying the model order. Section 5.3 discusses some existing approaches for robust parameter estimation.

4.3 Prediction-based monitoring

Let us assume that we have observed the time series $y_t, \ldots, y_n$ and possible covariates $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of the complete set-up phase. Based on that data we have selected and fitted an adequate model from the class of count time series based on GLMs. This includes that we have fully specified a model for the conditional mean $\lambda_t(\boldsymbol{\theta})$ and estimated the regression parameters by $\widehat{\boldsymbol{\theta}}_n$. To simplify notation we consider a model with a Negative Binomial conditional distribution, i.e. $Y_t|\mathcal{F}_{t-1} \sim \mathrm{NegBin}(\lambda_t(\boldsymbol{\theta}), \phi)$, of which we have estimated the dispersion parameter by $\widehat{\phi}_n$. The case of $Y_t|\mathcal{F}_{t-1} \sim \mathrm{Poisson}(\lambda_t(\boldsymbol{\theta}))$ is covered as a limiting case, where ϕ does not need to be estimated because it approaches infinity. Note that in our setting the model parameters are not re-estimated in each step of the following monitoring procedure since this bears the risk of overfitting by adapting to gradually increasing outbreaks. For this reason Noufaily *et al.* (2013) do not consider the last six months for designing their monitoring procedure.

4.3.1 Monitoring procedure

Based on the model obtained from the set-up phase, we want to successively monitor incoming observations in the operational phase. We study a monitoring procedure which gives an alarm if the actually observed realization y_{t_0+1} of Y_{t_0+1} is implausible with respect to the one-step-ahead predictive distribution for that time $t_0 + 1$. Because of the conditional construction of the considered model, one-step-ahead prediction is straightforward. Recall that $\mathcal{F}_{t_0}$ is the available information before y_{t_0+1} is observed. If the true parameters $\boldsymbol{\theta}$ and ϕ were known, the distribution of Y_{t_0+1} given $\mathcal{F}_{t_0}$ is $\mathrm{NegBin}(\lambda_{t_0+1}(\boldsymbol{\theta}), \phi)$. In practice we plug in the estimated parameters $\widehat{\boldsymbol{\theta}}_n$ and $\widehat{\phi}_n$ to approximate the distribution of the one-step-ahead prediction $\widehat{Y}_{t_0+1}^{(t_0)}$ of Y_{t_0+1} by

$$\widehat{Y}_{t_0+1}^{(t_0)} \sim \text{NegBin}(\widehat{\lambda}_{t_0+1}, \widehat{\psi}_n) \quad \text{with} \quad \widehat{\lambda}_{t_0+1} = \lambda(\widehat{\boldsymbol{\theta}}_n). \tag{4.1}$$

Note that this distribution does only reflect the sampling uncertainty but neither the estimation uncertainty nor the uncertainty induced by model selection. We will discuss in Section 4.6 how our procedure could be extended to account for estimation uncertainty. In the context of infectious disease surveillance one is only interested in detecting positive deviations from the in-control model. Denote a one-sided one-step-ahead prediction interval with a coverage rate of $1-\alpha$ by

$$\text{PI}_{t_0+1}^{(t_0)}(1-\alpha) = \left[0, U_{t_0+1}^{(t_0)}\right].$$

Its upper bound $U_{t_0+1}^{(t_0)}$ is chosen to be the $(1-\alpha)$-quantile of (4.1) for attaining the given coverage rate; recall Section 2.3. An alarm is given if y_{t_0+1} lies outside this prediction interval, which is referred to as the *stopping rule.* The time since the procedures has started until the first alarm is called *run length*, which is here given by

$$R = \inf\left\{s \geq 1 : Y_{n+s} \notin \text{PI}_{n+s}^{(n+s-1)}(1-\alpha)\right\}. \tag{4.2}$$

4.3.2 Calibration and performance measures

There is a trade off between the chance to timely detect aberrations (out-of-control performance) and the chance of a false alarm (in-control performance). On the one hand the procedure should not produce too many false alarms since this will fatigue the users of the system. On the other hand the procedure should not detect disease outbreaks too late (or even not at all) since this can affect the health of many people. This conflict of objectives is solved by maximizing the out-of-control performance subject to a given in-control performance (chance of a false alarm). Different monitoring procedures are then compared with respect to their out-of-control performance.

We briefly review some performance measures. All of them can be formulated in terms of the distribution of the run length R. If the process is assumed to be in-control, then R is the time until a false alarm. If the process is assumed to be in-control before time $t_0 = n + \tau$ (i.e. at the τ-th observation of the operational phase) and out-of-control thereafter, then $R - \tau$ is the detection delay. One should consider $R - \tau + 1$ instead of R to measure the out-of-sample performance if $\tau > 1$. Note that the support of the run length distribution is the set of positive integers and this distribution is generally positively skewed. In case of i.i.d. observations the run length of certain control charts has a geometric distribution and is quite easy to handle. However, Tartakovsky (2008) demonstrates that this does not hold for non-i.i.d. models and argues that many principles for designing control charts are not suitable in that case. One should bear in mind that in case of time series models some of the performance measures may have a different interpretation than in the classical case. In our study we approximate the distribution of the run length R by simulation. Accordingly, we pay some attention to this approximation in our discussion of the performance measures.

The average run length (ARL) is defined by the expectation of the run length distribution $\mathsf{E}(R)$. It is a standard practice in SPC to fix the in-control ARL (often denoted by ARL_0) to a given value, say a_0. This has the interpretation that we can on average expect a false alarm once every a_0 observations. Approximation of the ARL is problematic since the expectation depends on the upper tail of the run length distribution. In practice one needs to stop the procedure when there has not been an alarm within a certain (large) number of time points, say M. Therefore the simulated approximation of the run length distribution is truncated at M such that the mean of this distribution is a negatively biased estimator of the ARL.

One could also consider quantiles of the run length distribution. Gan (1994) propose to use the median run length (MRL) instead of the ARL to measure the performance of a control chart. He argues that the MRL has a more meaningful interpretation (the probability of run lengths lower (respectively greater) than the MRL is 50%). Whilst the ARL is more convenient for analytical calculations, the MRL is easier to obtain by simulation since one can approximate it from a truncated run length distribution (as long as the truncation is above the MRL).

Another measure is the probability of an alarm within r_0 time units, i.e., $\text{POA}_{r_0} = P(R \leq r_0)$. In the context of infectious disease surveillance one might want to use this as an out-of-control performance measure to attain a large probability of detecting an outbreak within only r_0 weeks. This measure is also meaningful in an in-control situation, where one would choose r_0 sufficiently large and adjust the control chart to given (low) value of POA_{r_0}. For weekly data one could choose $t_0 = 52$ to control the probability of a false alarm within one year. Lai and Chan (2008) propose the probability of a false alarm per unit of time, which coincides with POA_{r_0}/r_0 in the case of independent in-control data. This is identical to the false-positive rate (FPR) used by e.g. Noufaily *et al.* (2013) and Manitz and Höhle (2013). When fixing the in-control FPR one might want account for multiple testing, e.g. by a Bonferroni correction of the desired level. For out-of-control situations Noufaily *et al.* (2013) and Manitz and Höhle (2013) measure the performance of a control chart by the probability of detection (POD). It is given by POA_Δ, where Δ

Scenario	β_0	β_1	α_1	η_1	ϕ
Independent	1.570	0.0	0.0	0.4	8.5
Weak autocorrelation	1.220	0.2	0.0	0.4	14.5
Medium autocorrelation	0.880	0.4	0.0	0.3	17.0
Strong autocorrelation	0.187	0.8	0.0	0.1	1000.0
With feedback	0.266	0.4	0.4	0.1	17.0

Table 4.1: Parameter settings for model (4.3) considered in the simulation study.

is the duration of the aberration which has to be detected. It measures the probability that an aberration is detected during its occurrence.

4.4 Simulation study

We study the performance of the monitoring procedure presented in Section 4.3.1 by a simulation study. The scenarios used in this simulation study are chosen to imitate a realistic application like it is presented in Section 4.5. Our study is a proof of concept for prediction-based monitoring procedures for models with dependence and points out some notable features. We consider a set-up phase with a length of nine years and 52 observations per year, i.e. $n = 468$ observations. The performance of the procedure is assessed in an operational phase of one year, i.e. $m = 52$ observations.

The in-control data are randomly generated trajectories from the GLM-based count time series model (1.1) with a Negative Binomial conditional distribution, the logarithmic link function and dependence on the previous observation as well as on the previous conditional mean. A seasonal effect is added by a cosine function with a period length of one year, i.e. 52 weeks. This model is given by $Y_t|\mathcal{F}_{t-1} \sim \text{NegBin}(\lambda_t, \phi)$, $t = 1, \ldots, n+m$, with

$$\log(\lambda_t) = \beta_0 + \beta_1 \log(Y_{t-1} + 1) + \alpha_1 \log(\lambda_{t-1}) + \eta_1 \cos(2\pi t/52). \tag{4.3}$$

The considered parameter settings for that model are given in Table 4.1 and cover a variety of scenarios. The parameters are chosen such that the resulting process approximately has a marginal mean of 5 and a marginal variance of 10. Model (4.3) is a realistic choice for the DGP since it is very similar to models which have been selected for a real data example; see Section 4.5.1.

For the out-of-control data we need to decide how to model the effect of a disease outbreak. Most proposals from the literature are modifications of the respective in-control model.

This is based on the assumption that at least the general principles of the DGP still hold in the presence of the outbreak. One suggestion is to consider a multiplicative shift of the mean in case of an outbreak (Höhle and Paul, 2008, Section 2). Another possibility is that, due to an outbreak, serial correlation is introduced to the process (Höhle and Paul, 2008, Section 4) or, more generally, the extent of serial correlation is increased. One could also consider changes in the mean which enter the dynamics of the process, like spiky outliers or transient shifts (Fokianos and Fried, 2010, 2012), see also Chapter 3. Based on knowledge about the distribution of the incubation period, the shape of a log-normal distribution might be adequate to describe an outbreak (Lotze, Shmueli, and Yahav, 2007). A similar approach is to first draw the total number of additional cases of an outbreak starting at time t_0 from a Poisson distribution whose mean depends on the number of baseline cases at time t_0, and then to distribute these cases in time according to a discretized log-normal distribution (Noufaily *et al.*, 2013; Salmon, Schumacher, Stark, and Höhle, 2015). Another strategy to synthetically generate realistic outbreak effects is to use data from real outbreaks (Hutwagner, Browne, Seeman, and Fleischauer, 2005) and superimpose those on data from the in-control DGP.

In this study we consider an abrupt additive shift of the mean which does not enter the dynamics of the process. This outbreak scenario might not be realistic. However, this simple outbreak scenario makes it easier to interpret the simulation results and thus to understand the properties of our monitoring procedure. We place an outbreak with a duration of $\ell = 6$ weeks starting at a random position τ, where all positions $\tau \in \{n+1, \dots, n+m-\ell\}$ are equally likely. This random placement of the outbreak will average out an undesired effect of seasonality on our results. The out-of-control data denoted by Z_t are given by $Z_t = Y_t + O_t$, $t = 1, \dots, n+m$. Realizations of $Y_1, \dots, Y_{n+m}$ are generated in the same way as the in-control data. The outbreak effects O_t are independent realizations of Poisson distributions with mean Δ_t, i.e. $O_t \sim \text{Poisson}(\Delta_t)$, and also independent from Y_t, $t = 1, \dots, n+m$. The outbreak size at time t is given by $\Delta_t = \Delta \cdot \upsilon_t \cdot \mathbb{1}(t \in \{\tau, \dots, \tau+\ell\})$, where $\Delta \in \mathbb{R}$ is a global factor determining the outbreak size. The size of the outbreak is proportional to the conditional standard deviation at the respective time, i.e. $\upsilon_t = \sqrt{\lambda_t + \lambda_t^2/\phi}$, such that an outbreak causes more additional infections if the baseline number of infections is large. This seems to be a realistic assumption for an outbreak. From a rather technical point of view this ensures that outliers are more or less equally difficult to detect irrespective of the current mean of the process.

We study the prediction-based monitoring procedure introduced in Section 4.3.1 considering two different models: the *time series model*

Scenario	**Independence model** (4.5)	**Time series model** (4.4)
Independence	2.06 (64.5)	2.12 (64.7)
Weak autocorrelation	2.11 (65.4)	2.08 (65.0)
Medium autocorrelation	2.21 (61.2)	2.08 (64.2)
Strong autocorrelation	2.42 (52.3)	1.63 (56.3)
With feedback	2.60 (59.6)	2.03 (61.5)

Table 4.2: Average false positive rate in percent for different scenarios (procedures calibrated to a FPR of 2.5%). The values in parantheses are the proportion of samples with a false alarm in any of the 52 weeks of the observational period.

$$\log(\lambda_t) = \beta_0 + \beta_1 \log(Y_{t-1} + 1) + \alpha_1 \log(\lambda_{t-1}) + \sum_{s=1}^{2} \big(\eta_{s(1)} \sin(\omega_s t) + \eta_{s(2)} \cos(\omega_s t)\big) \tag{4.4}$$

and the *independence model*

$$\log(\lambda_t) = \beta_0 + \sum_{s=1}^{2} \big(\eta_{s(1)} \sin(\omega_s t) + \eta_{s(2)} \cos(\omega_s t)\big), \tag{4.5}$$

both with $Y_t|\mathcal{F}_{t-1} \sim \text{NegBin}(\lambda_t, \phi)$ and $\omega_s = 2\pi s/52$, $t = 1, \ldots, n + m$. The time series model (4.4) has eight and the independence model (4.5) has six unknown parameters. Note that the time series model covers all models (4.3) of the DGP (although it is overparametrized) whilst the independence model lacks serial dependence for all parameter settings in Table 4.1 except the first. Following Noufaily *et al.* (2013), the monitoring procedures are adjusted to a FPR of 2.5%. Hence the probability of a false alarm for any of the 52 observations of the operational phase is at most $1 - (1 - 0.025)^{52} = 73.2\%$.

The simulations are carried out on a computation cluster using the R packages **batchJobs** and **batchExperiments** (Bischl, Lang, Mersmann, Rahnenführer, and Weihs, 2015). The reported results are averages of 1000 replications.

We first examine the behavior of the monitoring procedure in an in-control situation. The results in Table 4.2 confirm that the procedure complies with the desired FPR of 2.5% in all scenarios and produces usually even less false alarms. This particularly holds for time series with strong temporal correlation if the procedure is based on the time series model (4.4). Considering the complete operational phase of 52 weeks, one can expect at least one false alarm in at most 66% of the cases. This is far below the defined upper bound of 73.2%. Overall, the procedure has a satisfying in-control behavior in all considered scenarios and seems to be a bit conservative.

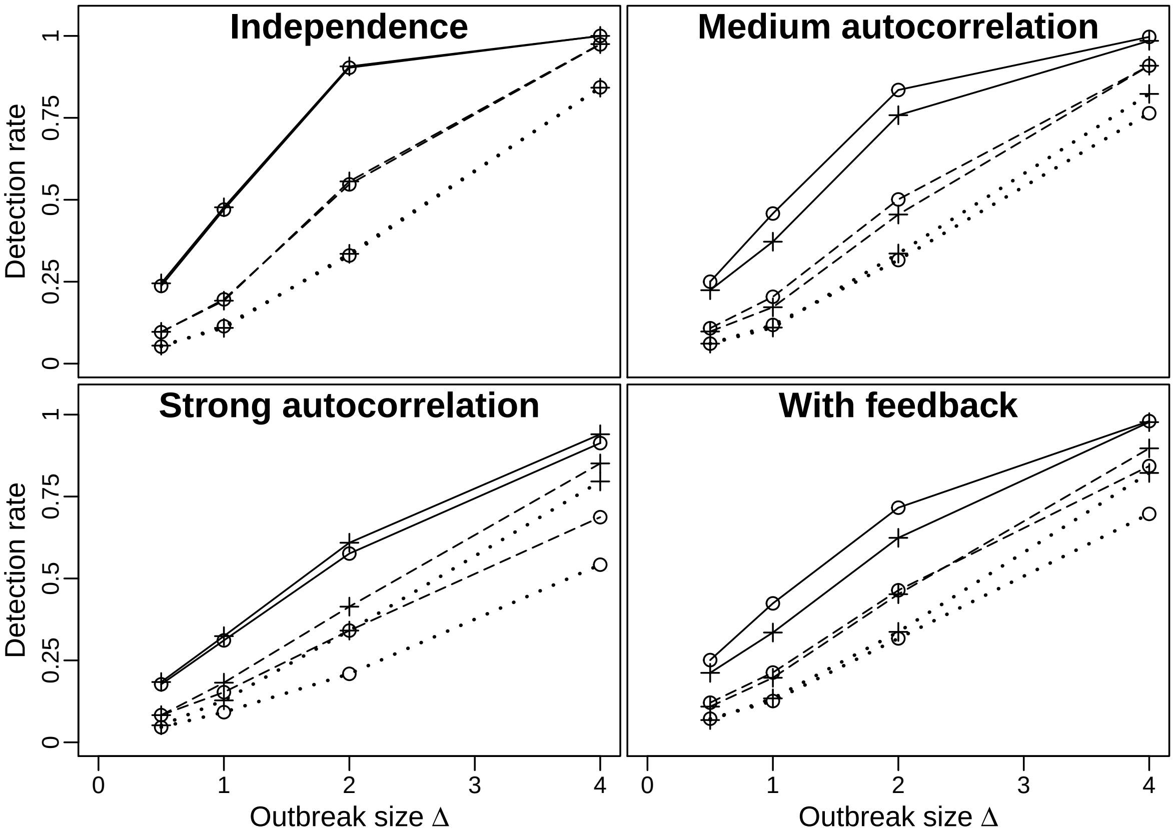


Figure 4.2: Outbreak detection rates of prediction-based monitoring procedures for four different data scenarios (each in a different panel). The plot symbol indicates whether the monitoring procedure is based on the time series model (cross) or on the independence model (circle). The line type indicates detection at the first time point (dotted), up to the second time point (dashed) or up to the last time point (solid) of the outbreak.

Next, we compare the out-of-control performance of monitoring procedures based on the time series model (4.4) and on the independence model (4.5). The dotted lines in Figure 4.2 show the probability of detecting the outbreak in the first week of its occurrence, i.e. POA_1, for different outbreak sizes ω. Naturally, the detection rate becomes larger for increasing outbreak size. If the data are independent, then the detection rate is almost the same for the procedure based on the time series model and the one based on the independence model. However, in case of dependent data the time series model outperforms the independence model. This advantage becomes larger with the degree of temporal dependence in the data. It is not very surprising though, that a model accounting for temporal dependence yields a superior monitoring procedure in case of dependent observations.

The estimated probability of detecting the outbreak with a delay of one time period, i.e. $\mathrm{POA}_2 - \mathrm{POA}_1$, shown by dashed lines in Figure 4.3 reveals that if an outbreak has not

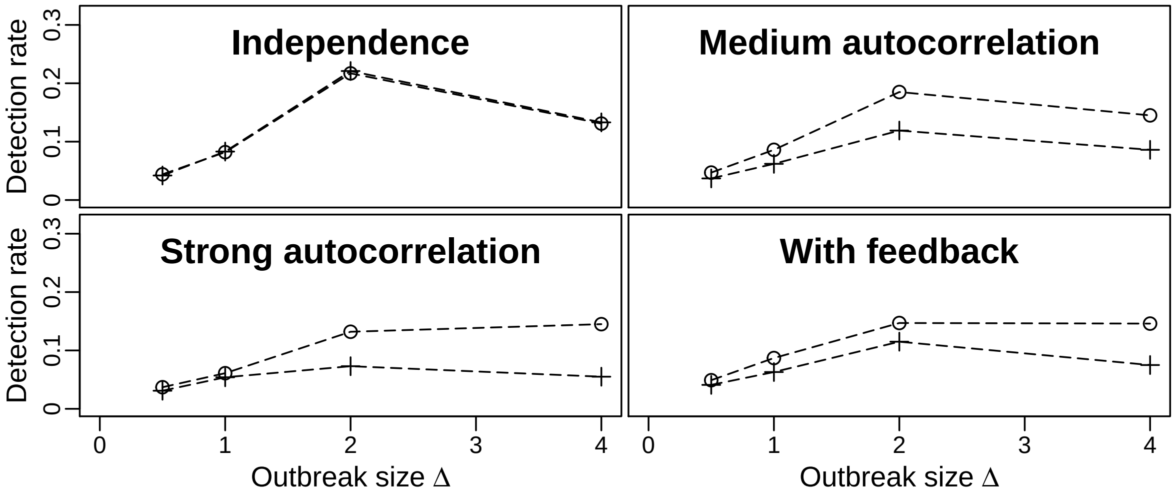


Figure 4.3: Estimated probabilities that the prediction-based monitoring procedure detects an outbreak with a delay of one time period for four different data scenarios (each in a different panel). The plot symbol indicates whether the monitoring procedure is based on the time series model (4.4) (cross) or on the independence model (4.5) (circle). Note that the vertical axis ranges to 0.3 only.

been detected at the first time point of its occurrence, then a procedure based on the time series model has a lower chance to detect it at the second time point than one based on the independence model – in case of actually dependent data. Hence a procedure which includes temporal correlation has a higher chance of immediately detecting an outbreak but a lower chance of detecting it at a later time. The explanation for this phenomenon is as follows: The procedure compares a new observation with the distribution of its one-step-ahead prediction, which depends on previous observations in case of the time series model. If the first observation of an outbreak is large (but not large enough such that the outbreak is detected), then the one-step-ahead prediction for the second observation of the outbreak is shifted upwards in case of positive autocorrelation. With respect to this shifted one-step-ahead predictive distribution, the second observation of the outbreak does not appear to be implausibly large. Hence this observation lies within the one-step-ahead prediction interval and no alarm is given. The described effect becomes stronger when increasing serial correlation of the fitted model. The dashed lines in Figure 4.2 show the estimated probability of detection with a delay of at most one time period, i.e. POA_2, which is equal to the sum of the estimated probabilities of immediate (dotted line in Figure 4.2) and of one time period delayed (Figure 4.3) detection. Overall, the monitoring procedure based on the time series model is superior in case of strong temporal correlation but slightly inferior in case of medium temporal correlation.

Finally, we evaluate the estimated probability of detection (POD) of an outbreak within its total duration of $\Delta = 6$ time points, i.e. POA_6, shown by solid lines in Figure 4.2.

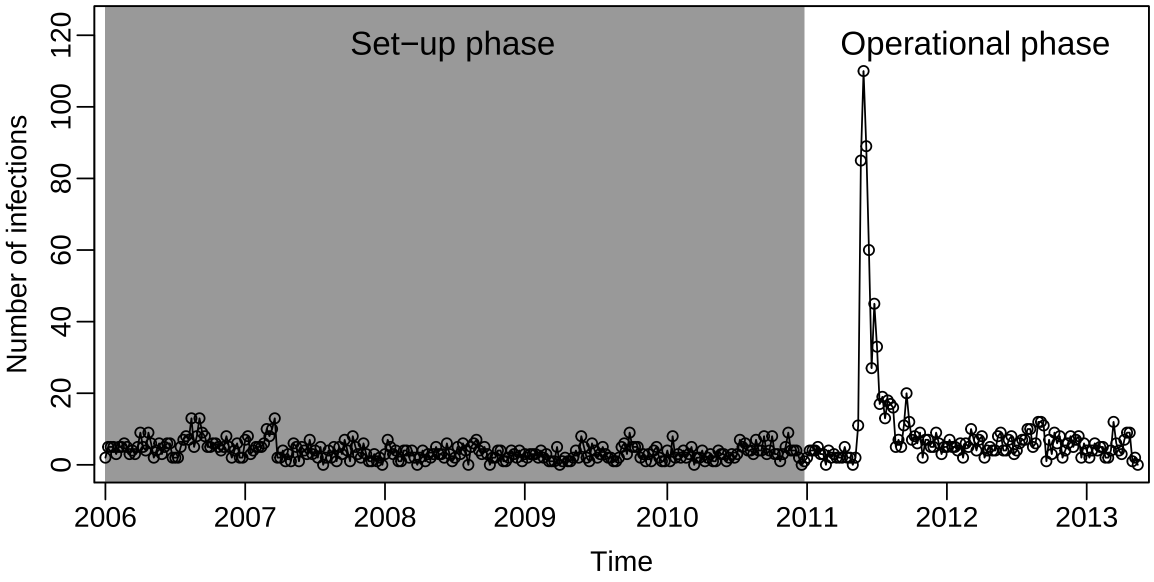


Figure 4.4: Number of registered EHEC cases per week in North Rhine-Westphalia.

The monitoring procedure based on the time series model is only superior if there is strong temporal correlation in the data. In situations with lower correlation its lower rate of delayed outbreak detection is not compensated by its higher rate of immediate detection. We revisit this issue in Section 4.6.

4.5 Case study

We illustrate phases I and II of a monitoring procedure by a real-data problem from infectious disease surveillance. Infection of humans with the bacterium enterohemorrhagic E. coli (EHEC) is notifiable under the German Protection against Infection Act. The weekly number of reported cases broken down to a regional level is recorded by the Robert Koch Institute (RKI) and publicly accessible by their web application SurvStat@RKI 2.0 (Robert Koch Institut, 2015). We consider the number of reported EHEC infections in the German state North Rhine-Westphalia from January 2006 to May 2013 (646 observations), see Figure 4.4. The first five years ($n = 261$ observations) until December 2010 are used as the set-up phase for selecting and fitting an appropriate in-control model. There would be more data available before 2006 but it is quite common in the field of infectious disease surveillance to only consider the previous few years to avoid problems with a potentially unstable DGP (for instance because laboratories change their methods or the number of probes sent to laboratories is varying over time). We will discuss the stability of the DGP in Section 4.5.2.

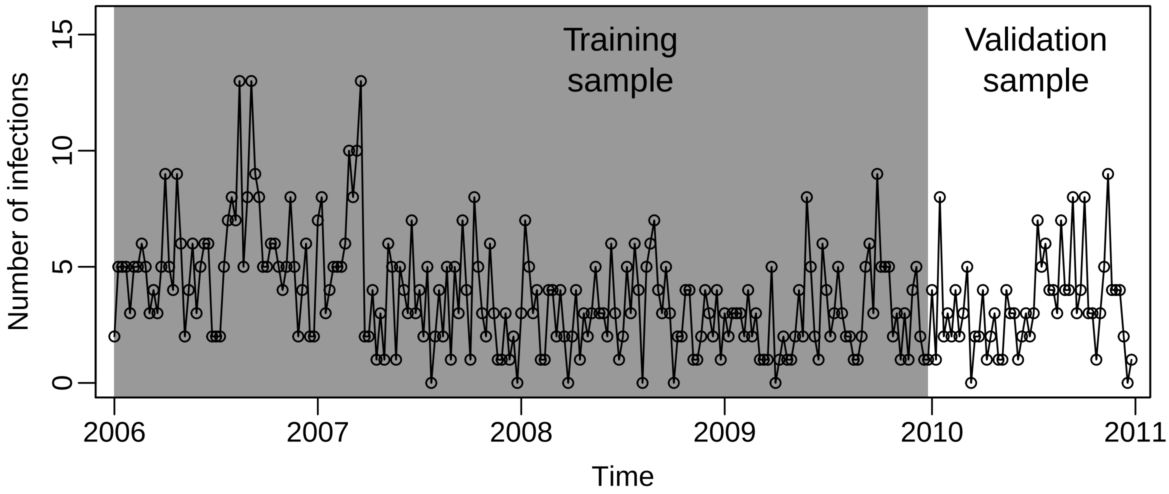


Figure 4.5: Number of registered EHEC cases per week in North Rhine-Westphalia (only data of the set-up phase).

4.5.1 Model selection and fitting in the set-up phase

Building a reliable control chart requires knowledge about the in-control process. Within a parametric approach this means selection of an adequate model and fitting this model to the available data. We split up the set-up phase into a training sample (the first four years until December 2009, 209 observations) and a validation sample (the remaining year, 52 observations).

In a first step we study the characteristics of the DGP with tools not based on a specific model. We hold back the validation sample and use only the training sample highlighted by a dark grey background in Figure 4.5. These considerations shall give an idea which features an adequate model for the data needs to have and allow to identify a range of candidate models.

The empirical marginal mean of 3.83 is quite low. For low means an approximation by a continuous-valued distribution is usually not appropriate and a distribution accounting for the discreteness of the observations is preferable. The data is overdispersed, i.e. the empirical marginal variance of 5.94 is substantially larger than the empirical marginal mean. Although a Poisson model with serial dependence explains overdispersion up to a certain degree, we also consider a more flexible Negative Binomial distribution.

A plot of the training sample against time in Figure 4.5 indicates that the number of registered cases is larger at the beginning and drops to a lower level in 2007. A global trend model does not seem to be adequate for this kind of change in the mean. It seems more promising to use models which allow for a more flexible change in the mean like

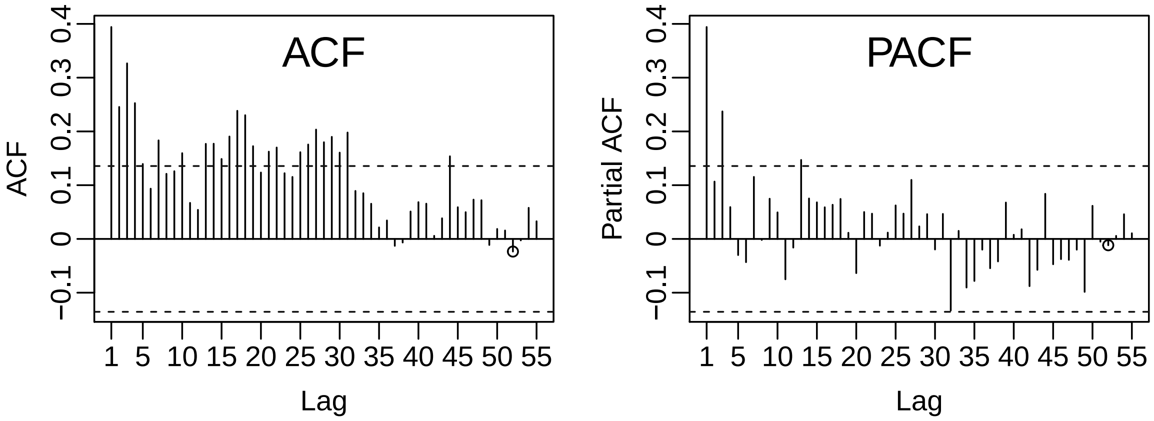


Figure 4.6: Estimated autocorrelation function (left) and partial autocorrelation function (right) of the EHEC training sample. The (partial) autocorrelation at lag 52, corresponding to a seasonality of about one year, is marked by a circle.

those with a feedback mechanism, i.e. a dependence on previous conditional means. We will nevertheless also consider models with a linear trend to make sure that we do not overlook this issue. In case of neglecting an upward trend the monitoring procedure produces too many false alarms, whilst ignoring a downward trend bears the risk of missing a relevant aberration. We also consider models with a level shift at some unknown time point.

The empirical autocorrelation function in Figure 4.6 (left) is about 0.39 at lag one and clearly different from zero for the first few lags. This shows that there is a considerable amount of temporal dependence present in the data. The empirical partial autocorrelation in Figure 4.6 (right) is clearly different from zero for the first and the third lag. This indicates that a regression only on one previous observation is not sufficient for describing the existing serial dependence. In such a case, regression on previous conditional means often yields a more parsimonious model which fits the data equally well or even better than regression on more than one previous observation. We consider models regressing on one or more previous observations $Y_{t-1}, \dots, Y_{t-p}$ and those additionally regressing on the conditional mean one week back in time λ_{t-1}.

The time series plot in Figure 4.5 slightly indicates that there might be a yearly seasonal fluctuation with larger values in the summer. However, the empirical autocorrelation and empirical partial autocorrelation in Figure 4.6 are very close to zero at lags around $T = 52$ and hence do not confirm this periodicity. We consider the techniques for modeling seasonality described in the previous section and also models without seasonal effects. It is neglected that the length of a year is not exactly 52 weeks and varies slightly.

We decide for the log-linear model which has more flexibility in including covariates since it does not induce restrictions on covariates and on the corresponding parameters. We indeed experience that the limitation of the identity link model to positive covariates impede properly fitting the seasonal pattern. Furthermore, the logarithm is a commonly used link function for that kind of data (see for example Höhle and Paul, 2008). This choice implies that the effect of covariates is multiplicative (and not additive as for a model with identity link function). This includes that a seasonal effect modeled by covariates is multiplicative. Trends also have a multiplicative effect, such that a linear trend in the linear predictor is actually an exponential trend on the original scale of the response. For slight trends and short time series such a trend is approximately linear on the original scale. However, this results in instabilities when extrapolating to long time horizons. We stress that multiplicative trends should be used with care.

In a second step we use the training data to fit the candidate models which have been identified initially. Since the conditional mean is fitted independently of the chosen conditional distribution, we assume a Poisson distribution and reassess this at a later stage. These fitted models are compared by the (in-sample) Akaike information criterion (AIC) and the out-of-sample measures ranked probability score (RPS), coverage rate (CR) of a one-sided 95% 1-step-ahead prediction interval and prediction bias (see Section 2.4 on model assessment). The out-of-sample measures are evaluated on the validation data. The results of the model comparison are presented in Tables 4.3 and 4.4 for models with and without serial dependence, respectively.

We have to decide on a) how to model seasonality, b) whether to include a trend and c) the model orders for serial correlation (excluding stochastic seasonality). We expect (and have also verified empirically) that decisions b) and c) are strongly interdependent but that they do not interfere much with decision a). Hence we first decide which approach for modeling seasonality is preferable while fixing the two other decisions. Regarding the Fourier approach, the AIC would opt for the first $S = 3$ Fourier frequencies whilst the out-of-sample RPS would opt for $S = 2$. Regarding the periodic B splines approach, both considered criteria consistently choose $K = 5$ knots (corresponding to 5 additional parameters). Stochastic seasonality seems to be described better by dependence on the conditional mean λ_{t-T} than on the observation Y_{t-T} observed one year ago. The former two approaches employing deterministic covariates clearly outperform the latter stochastic approach, which is not better than not considering seasonality at all. Overall, the periodic B splines approach performs best. In addition to this numerical measures, we examine the seasonal patterns resulting from the Fourier approach with $S = 2$ and from B splines with $K = 5$ which are shown in Figure 4.7 (top). Both patterns have their maximum in August and their minimum in November, though not in exactly the same

Seasonality	**Trend**	**Dependence**	**DoF**	**AIC**	**RPS**	**CR**	**Bias**
no	no	p=1, q=1	3	899.0	1.104	0.942	-0.213
Fourier (S=1)	no	p=1, q=1	5	901.6	1.091	0.962	-0.220
Fourier (S=2)	no	p=1, q=1	7	900.0	1.076	0.962	-0.224
Fourier (S=3)	no	p=1, q=1	9	898.4	1.097	0.942	-0.210
B splines (K=3)	no	p=1, q=1	6	900.9	1.070	0.923	0.204
B splines (K=4)	no	p=1, q=1	7	904.3	1.089	0.962	-0.252
B splines (K=5)	no	p=1, q=1	8	892.5	1.058	0.942	0.218
B splines (K=6)	no	p=1, q=1	9	893.7	1.075	0.942	0.211
stochastic (Y_{t-T})	no	p=1, q=1	4	911.2	1.178	0.923	-0.346
stochastic (λ_{t-T})	no	p=1, q=1	4	900.3	1.102	0.942	-0.191
B splines (K=5)	no	p=0, q=0	6	940.8	1.109	0.962	-0.433
B splines (K=5)	no	p=1, q=0	7	910.7	1.116	0.962	-0.307
B splines (K=5)	no	p=2, q=0	8	910.5	1.101	0.962	-0.285
B splines (K=5)	no	p=2, q=1	9	888.1	1.106	0.904	0.637
B splines (K=5)	no	p=3, q=0	9	903.1	1.115	0.962	-0.254
B splines (K=5)	no	p=4, q=0	10	903.8	1.108	0.962	-0.248
B splines (K=5)	linear	p=0, q=0	7	894.3	1.300	0.846	1.259
B splines (K=5)	linear	p=1, q=0	8	888.1	1.222	0.885	1.062
B splines (K=5)	linear	p=1, q=1	9	887.2	1.283	0.865	1.165
B splines (K=5)	linear	p=2, q=0	9	890.0	1.225	0.885	1.067
B splines (K=5)	linear	p=2, q=1	10	888.0	1.243	0.885	1.070
B splines (K=5)	linear	p=3, q=0	10	889.5	1.196	0.885	0.964
B splines (K=5)	linear	p=4, q=0	11	891.4	1.192	0.885	0.957

Table 4.3: Comparison of the fitted candidate models (assuming a Poisson conditional distribution). Degrees of freedom (DoF) indicates the number of regression parameters. The considered measures are Akaike's information criterion (AIC), ranked probability score (RPS), coverage rate (CR) of a one-sided 95% 1-step-ahead prediction interval and prediction bias. The AIC is an in-sample measure evaluated on the training data, wheras all others are out-of-sample measures evaluated on the validation data.

Seasonality	**Trend**	**DoF**	**AIC**	**RPS**	**CR**	**Bias**
no	no	1	949.4	1.164	0.923	-0.429
Fourier (S=1)	no	3	949.4	1.119	0.962	-0.430
Fourier (S=2)	no	5	940.9	1.102	0.962	-0.432
Fourier (S=3)	no	7	939.0	1.136	0.962	-0.432
B splines (K=3)	no	4	942.5	1.113	0.962	-0.432
B splines (K=4)	no	5	940.5	1.104	0.962	-0.433
B splines (K=5)	no	6	940.8	1.109	0.962	-0.433
B splines (K=6)	no	7	938.7	1.129	0.962	-0.433
Fourier (S=3)	linear	6	893.9	1.302	0.846	1.267

Table 4.4: Comparison of the fitted candidate models (assuming a Poisson conditional distribution) considering only models assuming independent observations. See Figure 4.3 for further explanation.

weeks of the year. This corresponds to more infections in the summer and less towards the end of the year. The seasonal pattern based on the Fourier approach exhibits another peak with a lower amplitude in February which is not present in the pattern based on periodic B splines. This additional peak might be an artifact of the superposition of sinusoidal waves of lengths 52 and 26. Consequently, we decide for periodic B splines with $K = 5$ knots.

Given the chosen type of seasonality we consider a range of different model orders for the serial dependence (including an independence model). We combine each of these with a model without trend as well as with one with a global linear trend. The AIC is in favor of models with a (negative) trend whilst the out-of-sample properties do not support the inclusion of a linear trend. The models with a linear trend have larger values of the RPS, do not hold a coverage rate of 95% of 1-step-ahead prediction intervals and have a large (positive) bias. This indicates that a global trend is not adequate here.

In a last step we decide on the model order for the serial dependence. With respect to the AIC a regression on the two previous observations and on one previous value of the conditional mean ($p = 2$ and $q = 1$) performs best. With respect to the out-of-sample criteria the model with $p = 1$ and $q = 1$ performs best. Since the former model has a very poor out-of-sample performance with a large bias we decide for the latter model with model order $p = 1$ and $q = 1$. The conditional mean λ_t of this model is given by

$$\begin{aligned}\log(\lambda_t) = \beta_0 + \beta_1 \log(Y_{t-1} + 1) + \alpha_1 \log(\lambda_{t-1}) \\ + \eta_1 X_{t,1} + \eta_2 X_{t,2} + \eta_3 X_{t,3} + \eta_4 X_{t,4} + \eta_5 X_{t,5},\end{aligned}$$

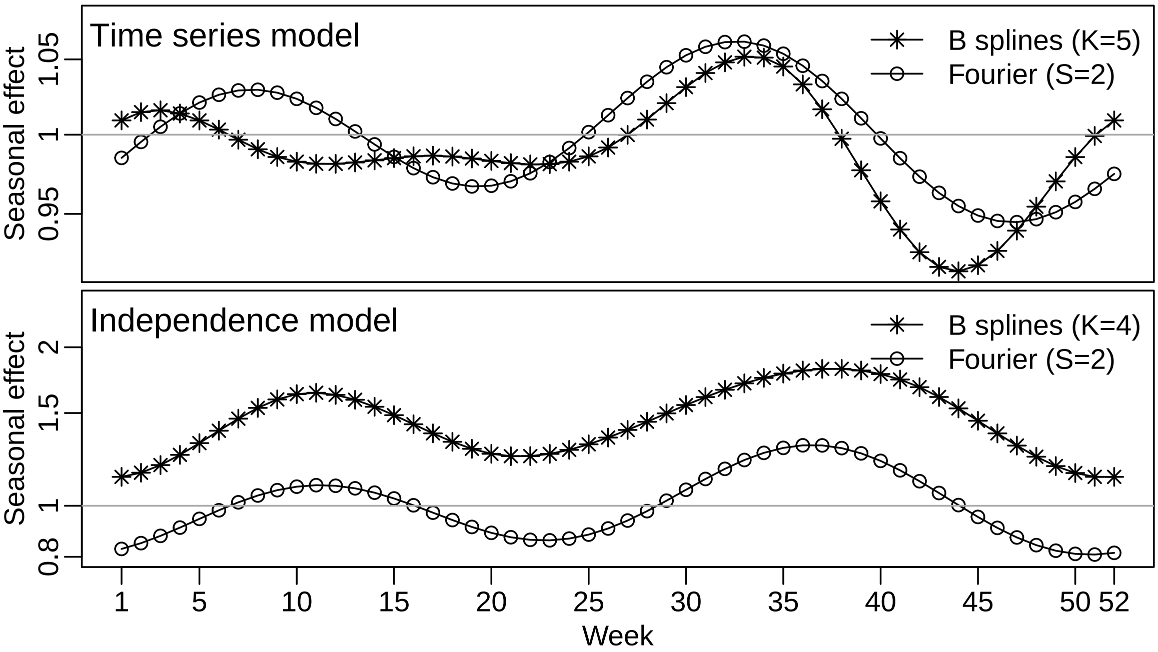


Figure 4.7: Seasonal patterns of B splines with K knots and of a superposition of sine and cosine functions for the first S Fourier frequencies. The seasonal effect expresses the estimated factor by which the conditional mean is multiplied in the respective week and is shown on a logarithmic scale. The time series model (4.4) with serial correlation of order $p = 1$, $q = 1$ (top) and the independence model (4.5) (bottom) are fitted to the training data.

where the covariates $X_{t,1}, \ldots, X_{t,5}$ specify the value of the periodic B splines at time t (see Figure 4.1). We call this best model the (best) *time series model.*

For comparison, we also select the best model assuming independent observations. According to a comparison of the out-of-sample in Table 4.4 there is little difference in the performance of models with seasonality based on the first $S = 2$ Fourier frequencies and based on periodic B splines with $K = 4$ knots with respect to the (out-of-sample) ranked probability score. The seasonal patterns of both approaches shown in Figure 4.7 (bottom) are of quite similar shape but displaced. Unlike for the Fourier approach, the parameters of periodic B splines may interfere with the intercept and actually do so in this data example. Again, a model with a linear trend has a low AIC but a very poor out-of-sample behavior. Consequently we choose a model without a trend and with the first $S = 2$ Fourier frequencies to be the best model of those assuming independent observations. We call this model the (best) *independence model.* The conditional mean λ_t of this model is given by

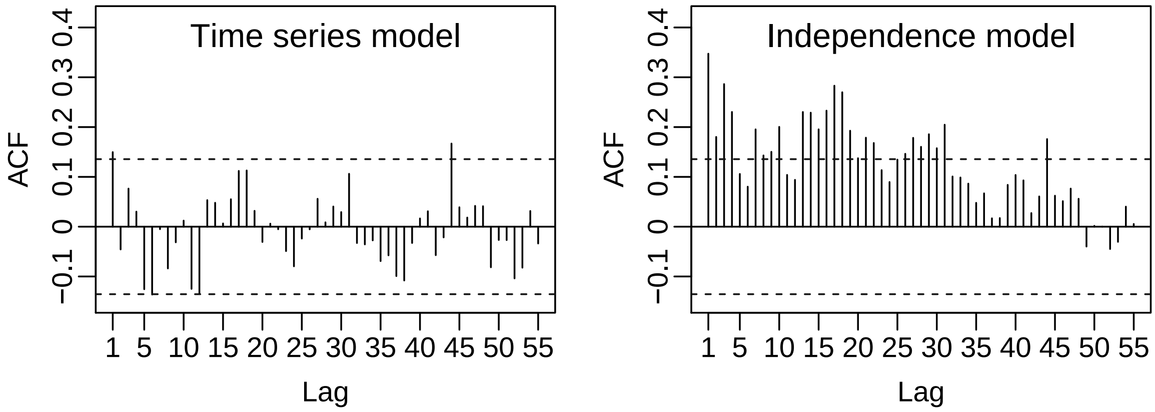


Figure 4.8: Estimated autocorrelation function of the residuals. The time series model with serial correlation of order $p = 1$, $q = 1$ (left) and the independence model without serial correlation (right) are fitted to the training data.

$$\begin{aligned}\log(\lambda_t) = \beta_0 &+ \eta_{1(1)} \sin(2\pi t/52) + \eta_{1(2)} \cos(2\pi t/52) \\ &+ \eta_{2(1)} \sin(2\pi t/26) + \eta_{2(2)} \cos(2\pi t/26).\end{aligned}$$

Given the linear predictor of the model we can now decide which conditional distribution to use. We compare a Poisson and a Negative Binomial model by (in-sample) probability integral transform (PIT) histograms (see Section 2.4) shown in Figure 4.9. For the time series model both PIT histograms appear to be uniform. Hence we decide for the simpler Poisson model. For the independence model the PIT histogram is uniform for the Negative Binomial but U-shaped for the Poisson distribution. It seems that the Negative Binomial distribution needs to account for the variability which is otherwise explained by temporal dependence.

We compare the fit to the training data of the selected best time series model and the best independence model as shown in Figure 4.10. The fit of the independence model repeats the same seasonal pattern each year and is otherwise constant over time. Consequently this also holds for the predictions of the validation sample, which do not depend on previous observations. The fit of the time series model also exhibits a seasonal pattern with a peak in the summer. However, this pattern is superimposed by the effect of serial correlation. The model reacts to the large observations in the summer of the year 2006 where its fit has an especially pronounced peak in the summer as compared to the years 2007 and 2009. As mentioned before, the mean of the data appears to vary slowly over time. The time series model reacts to this change in location and has a larger mean until 2007 than in the years thereafter. By contrast, the independence model tends to underestimate the observations until 2007 and to overestimate them thereafter. The

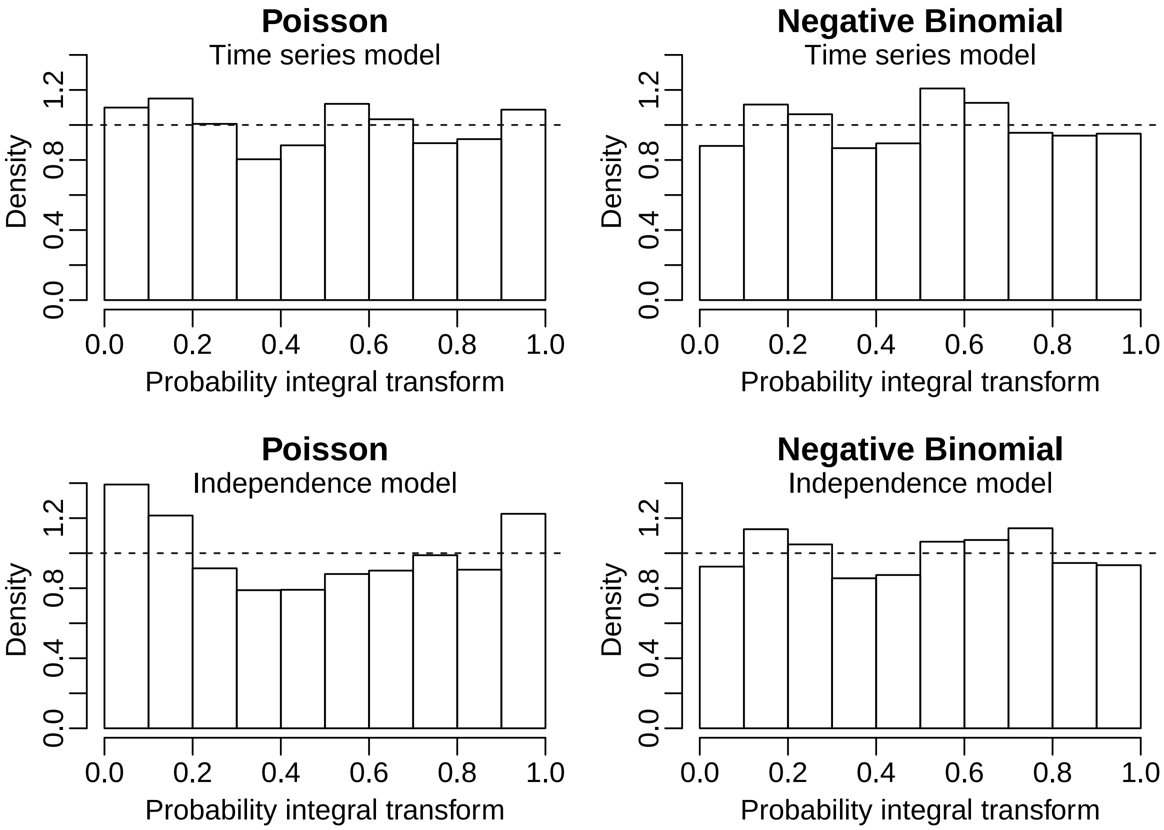


Figure 4.9: Probability integral transform (PIT) histograms for model with Poisson and Negative Binomial conditional distribution. The time series model with serial correlation of order $p = 1$, $q = 1$ (top) and the independence model without serial correlation (bottom) are fitted to the training data.

residuals of the independence model exhibit a considerable amount of serial correlation whilst the residuals of the time series model appear to be uncorrelated, see Figure 4.8. This shows very clearly that the selected model for independent observations does not capture the serial correlation which is present in the data. The purpose of the fitted models is their use for prediction-based monitoring of future observations. This requires that not only the predicted value but also the corresponding prediction interval is reliable. The latter is for example assessed by the ranked probability score (RPS), which is in favor of the time series model: the out-of-sample RPS evaluated on the validation data is 1.148 for the time series model and 1.281 (based on the selected Negative Binomial distribution) for the independence model. Figure 4.10 illustrates the predictions and corresponding prediction intervals of the validation sample. In case of the independence model, the upper limits of the one-sided one-step-ahead prediction intervals reflect the seasonal pattern of the fitted model and are uniformly larger than those of the time series model. Applying the monitoring procedure (calibrated to a FPR of 2.5%) to the validation data would sound no alarm (since all observations lie in the 97.5%PI) if it is

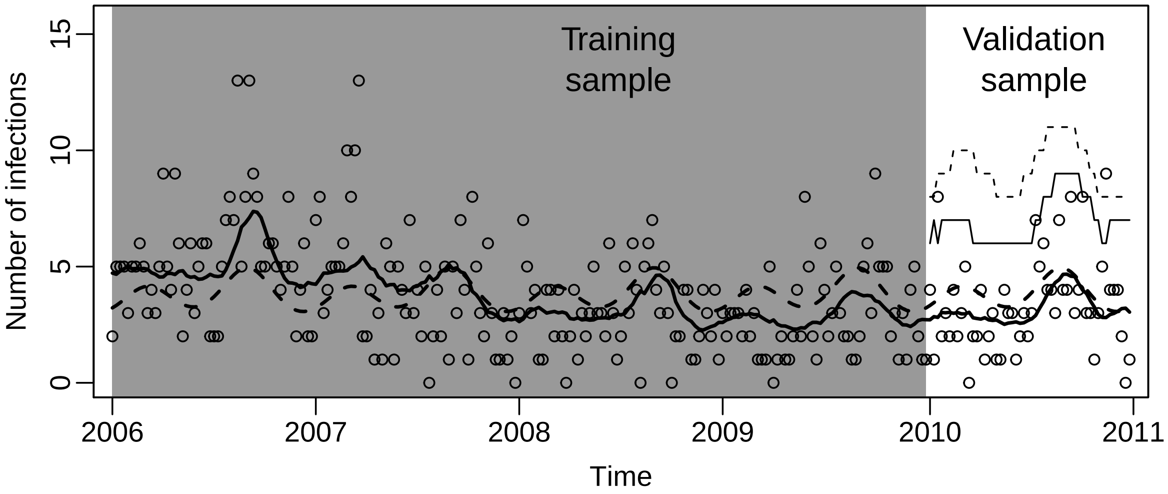


Figure 4.10: Fitted values and 1-step-ahead predictions (thick lines) with upper bounds of one-sided 97.5% PI (thin lines). The time series model with serial correlation of order $p = 1$, $q = 1$ (solid line) and the independence model without serial correlation (dashed line) are fitted to the training data (grey background).

based on the independence model and two alarms if it is based on the time series model. If we assume the data of the validation phase to follow the in-control DGP, these alarms will actually be false alarms. Nevertheless, we see more evidence in favor of the time series model.

Finally, we fit the two models on the complete data of the set-up phase. Assume the number or registered EHEC cases in week t to be a realization of Y_t, $t = 1, \ldots, 261$. The fit of the best time series model is given by $Y_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t)$ with

$$\begin{aligned}\log(\lambda_t) = {} & 1.423 \cdot 10^{-5}(0.461) + 0.062(0.043)\log(Y_{t-1}+1) + 0.938(0.356)\log(\lambda_{t-1}) \\ & - 0.031(0.125)X_{t,1} + 0.114(0.174)X_{t,2} - 0.166(0.207)X_{t,3} \\ & + 0.059(0.107)X_{t,4} - 0.046(0.187)X_{t,5}.\end{aligned}$$

Standard errors obtained by a parametric bootstrap procedure are given in parentheses. Remarkably, the coefficient of $\log(\lambda_{t-1})$ is very close to one. This term allows the model to describe the change in the mean over time. However, this comes along with quite large standard errors, pointing to a possibly unstable model fit. The fit of the best independence model is given by $Y_t|\mathcal{F}_{t-1} \sim \text{NegBin}(\lambda_t, 9.280)$ with

$$\begin{aligned}\log(\lambda_t) = {} & 1.310(0.040) - 0.127(0.052)\sin(2\pi t/52) - 0.031(0.053)\cos(2\pi t/52) \\ & + 0.094(0.053)\sin(2\pi t/26) - 0.140(0.052)\cos(2\pi t/26).\end{aligned}$$

The estimated overdispersion coefficient is $\widehat{\sigma}^2 = 1/9.280 = 0.108$ (0.035). The conditional variance is on average roughly $\overline{y}\widehat{\sigma}^2 = 40\%$ larger under a Poisson model. These two fitted models form the basis for the prediction-based monitoring procedure which is demonstrated in Section 4.5.3.

4.5.2 Nonstationarity vs. serial dependence

Before using the fitted models for monitoring we discuss their reliability. Both models do not assume the DGP to be stationary but allow the expected number of cases to change periodically during the year. Apart from that seasonal effect, the expected number of cases is assumed to be stable over time. For the independence model stable does actually mean constant whilst the time series model only assumes that the process emerges in the same way given its previous expected value and its previous observation. The decrease in the number of infections in the year 2007 questions the assumption that the DGP is stable over time, even if we account for the seasonal effect. Models including a global linear trend have not proven to be appropriate, see the previous section. However, we have not considered other forms of nonstationarity of the DGP like more general trends or abrupt changes in location. To some extent serial correlation is able to describe such nonstationarities. It is a challenging problem to decide whether an explicit model for the trend or change is more appropriate than a model explaining such phenomenons by (usually strong) serial correlation. We investigate this question in some more detail in this section.

As mentioned before, Figure 4.5 suggests that there might be an abrupt shift in location in the year 2007 which could possibly be modeled explicitly. We test whether there is a level shift (LS) at an unknown time point in the training data by applying the procedure described in Section 2.5 (and in more detail for models with the identity link in Section 3.4.2). The overall test statistic of this test is the maximum of local test statistics at each time which are illustrated in Figure 4.11. For the time series model this test is not significant on a 5% level (p value 0.17, test statistic 15.2). For the independence model the test is significant (p value 0, test statistic 23.15) and detects a level shift in week 15 of the year 2007. The estimated size of the LS is -0.63, which corresponds to a decrease of the expected number of cases by the factor $\exp(-0.63) = 0.53$. Considering this level shift for the independence model does indeed yield a model with more or less uncorrelated residuals whose ACF looks very similar to that of the time series model in Figure 4.8 (left). This indicates that the observed serial dependence, recall Figure 4.6, is largely induced by this change in location.

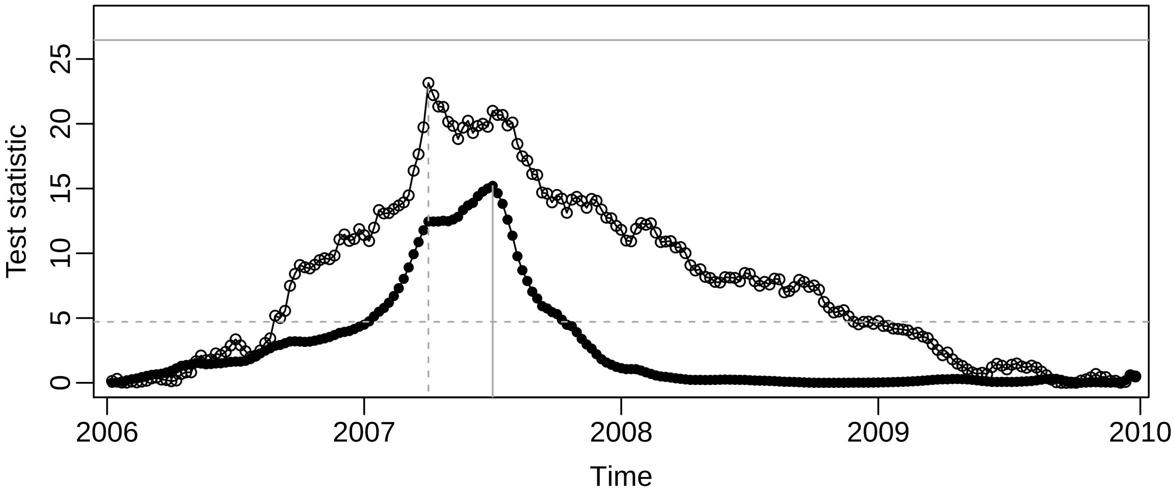


Figure 4.11: Test statistic for detecting a level shift in the training data. The time series model with serial correlation of order $p = 1$, $q = 1$ (filled circle) and the independence model without serial correlation (empty circle) are fitted to the training data. The vertical line indicates at which time the test statistic has its maximum and the horizontal line is the critical value of the test statistic at a 95% level for the time series model (solid lines) and for the independence (dashed lines) model, respectively.

One might come to the conclusion that is not necessary to fit a model with temporal dependence once we haven taken into account the level shift. However, the data of including the years before the set-up phase in Figure 4.12 tell that there is more change in the average location than just one abrupt downward shift in 2007. A moving average with a window width of one year suggests that the location changes in a way which cannot be described adequately by a global trend and/or a few abrupt changes. The model with serial dependence is capable to model such a slowly varying trend.

4.5.3 Monitoring in the operational phase

The prediction-based monitoring procedure is illustrated in Figure 4.13. The procedure is based on the time series respectively independence model fitted to the data of the set-up phase. It starts in the first week of the year 2011. The available information at that time are all previous observations, the fitted values of the set-up phase and the covariates of that time. The alarm threshold is the upper bound of a one-sided 97.5% one-step-ahead prediction interval. This threshold is 8 for the time series model (solid line) and 9 for the independence model, both larger than the value of 2 EHEC cases which has been registered for that week. For both models the procedures continues without an alarm. For the second week of 2011 we can use the observation of the first week and the covariates of the second week. However, monitoring based on the independence model does (by

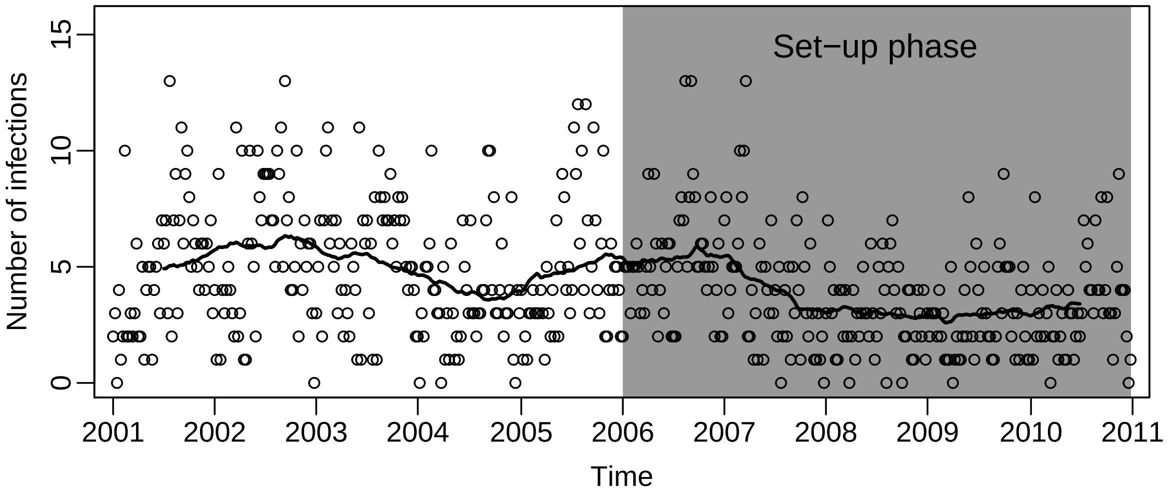


Figure 4.12: Number of registered EHEC cases per week in North Rhine-Westphalia including data before the chosen set-up phase. The thick line is the moving average with a window width of one year.

construction) not make use of the previous observation but only of the current covariates. There have been 4 infections registered in the second week, which is again well below the thresholds of both models. Both procedures successively continue like that. In week 20 the time series model yields a threshold of 7 and the independent one a threshold of 9. In that week the actual observation is 11 and exceeds both thresholds. Both procedures sound an alarm. In fact there is the beginning of an outbreak in that week.

As noted before, the procedure for the independence model is not affected by the previous observations at all. Its threshold values do only depend on the seasonal pattern. As opposed to this, the prediction of the time series model directly depends very little on the previous observation (with a factor of 0.062) and very much on the previous prediction (0.938). Due to the recursive formulation this means that the very large observation in week 20 has a strong positive effect on the prediction and hence also on the threshold for week 22. The following seven observations in weeks 21 to 27 also belong to the outbreak and are even larger (they are not shown in the plot because the vertical axis is truncated at a value of 25). The previously described effect results in a steep increase of the one-step-ahead predictions and hence also of the corresponding thresholds. The effect of serial correlation is desired as long as the DGP is in-control but not when it is out-of-control. In this example the first observation of the outbreak is large enough to detect it immediately. Even if we would not have detected the outbreak in week 20, the gradient of the outbreak is so steep that the following observations of the outbreak are still above the corresponding thresholds, even though these thresholds are increased because of the outbreak. However, we know from the simulations in Section 4.4 that this

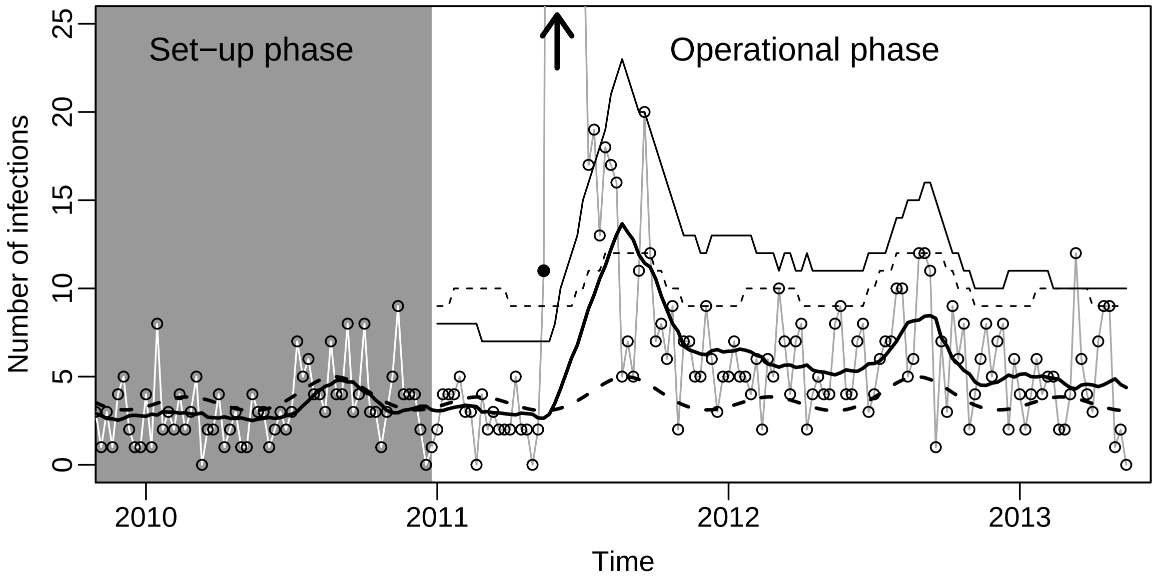


Figure 4.13: Illustration of the prediction-based monitoring procedure in the operational phase for the time series model (solid line) and the independence model (dashed line). If an observation lies above the one-sided one-step-ahead 95% PI (thin lines) an alarm is given. The first alarm (concurrently detected by both models) is marked by a filled circle. The thick lines are the one-step-ahead predictions. Note that the vertical axis is truncated at 25 in order to show more details and hence seven larger observations (see Figure 4.4) are not shown but only represented by the arrow.

behavior of the time series model may also circumvent the detection of only moderate outbreaks.

4.6 Discussion

We demonstrate by the data example of EHEC infections in this chapter and also by the Campylobacter infections analyzed in other chapters (see Sections 2.6.1 and 3.5) that there are infectious disease time series which exhibit serial correlation. This suggests to employ time series models with temporal dependence for monitoring procedures. We introduce a simple monitoring procedure for the class of count time series following generalized linear models which is based on successive one-step-ahead predictions. We investigate by a simulation study how using a model with temporal dependence affects the performance of such a procedure compared to an independence model. Our most important findings are:

- If the data are uncorrelated there is no perceivable loss in performance.
- If the data are correlated, more outbreaks can be immediately detected.

- If an outbreak is not detected immediately, a model with dependence is influenced by this outbreak. This impedes its detection at a later time with the prediction-based monitoring procedure considered in this chapter.

The last result is particularly a problem when outbreaks emerge gradually, which is often the case in practice. It can be explained as follows: Predictions by independence models do only depend on the covariates of the respective time. If we disregard the covariates effect for a moment, these predictions are only based on the marginal distribution. In contrast, time series models make predictions based on the conditional distribution given previous observations. They are using more information which makes them more efficient in many situations. On the other hand this also makes them more vulnerable to the effect of unusual observations.

We have several ideas to utilize the promising features of time series models for monitoring and at the same time to overcome their weakness:

1. Simultaneously apply a control chart based on the conditional distribution (which is good for immediate detection of large outbreaks) and another one based on the marginal distribution (which is good for delayed detection of only moderate outbreaks). We recommend CUSUM-type charts for the latter. Both charts need to be combined into a single procedure by a decision rule (e.g. sound an alarm if any of the charts exceeds its control limits). A challenge will be to calibrate this procedure to attain a desired in-control performance.
2. Limit the influence of previous observations on the prediction. One strategy for that is to employ ideas from robust statistics (e.g. truncate or replace implausibly large previous observations) to obtain robust predictions. Another strategy is to replace one-step-ahead predictions by h-step-ahead predictions (see Section 2.3) such that the previous $h-1$ observations have no influence at all on the prediction, see also Section 5.3. The distribution of h-step-ahead predictions would need to be approximated by simulation, which is of course computationally expensive.
3. Compare a vector of the most recent h subsequent observations to its h-dimensional marginal distribution. Give an alarm if the observed vector is implausible with respect to that distribution. One challenge of this approach is to find a region which covers observed vectors from the in-control model with a probability of $1-\alpha$. This is much more difficult for discrete count data distributions than for continuous distributions, for which this approach has been originally proposed by Gather, Bauer, and Fried (2002). Another challenge is how to account for covariate effects, which are essential for modeling seasonality.

In its current form our monitoring procedure does not account for estimation uncertainty. A recent review on the effect of estimation uncertainty on control charts is provided by Psarakis, Vyniou, and Castagliola (2014). Schmid (1995, Theorem 4.3) give an analytical result for the average run length of a control chart in case of autoregressive processes with normal errors. In the case of count time series, the Bayesian approach by Manitz and Höhle (2013) accommodates estimation uncertainty. Our monitoring procedure could easily be extended to accommodate estimation uncertainty by approximating the one-step-ahead prediction intervals with the parametric bootstrap procedure described in Section 2.3. Some first experiments with this approach yields promising results at the cost of a several times longer computation time. One could also employ the delta method to analytically approximate the prediction distribution by combining the conditional distribution of the observations given the true parameter, and the approximate normal distribution of the maximum likelihood estimator. Freeland and McCabe (2004, Theorem 2) do this for so-called INAR models of order one, which are very different from the models considered in this work.

It has been pointed out in Section 4.5.1 that a linear trend has a multiplicative effect on the conditional mean in case of the logarithm link function. This explosive behavior is usually undesirable. With a generalized additive model (GAM) one could fit models with a conditional mean of $\lambda_t = \eta_0 t + \nu_t$, where ν_t is the linear predictor given on the right hand side of (1.1) with $\widetilde{g}(x) = \log(x+1)$ and $g(x) = \log(x)$. Note that the GAM proposed by Held *et al.* (2005) (see Section 4.1) is different from the above suggestion and has a multiplicative trend. Another issue which requires further investigation is the discrimination between serial dependence and nonstationarity which has been discussed in Section 4.5.2.

Summing up, this chapter is an important step towards efficient monitoring procedures based on models with temporal dependence. However, this topic is still ongoing research and we outlined some directions of further research. The discussed ideas need to be elaborated, implemented and compared with established procedures. The following chapter takes up the issue of model selection and deals with some further topics which are relevant to the problem of online monitoring.

Chapter 5

Further topics

This chapter discusses three further topics which are particularly (but of course not exclusively) of interest for monitoring count time series in the context of infectious disease surveillance. We do not present a conclusive treatment but rather point to directions where further research is needed.

Section 5.1 reviews tools that have been proposed for model selection for count time series and those which have been proposed in another context but could possibly be extended to count time series. These tools complement and enhance the ones presented in Section 2.4. Such methods can from the basis of comprehensive model selection strategy which does not require manual intervention. The other two sections are concerned with robust methods which work reliably in the presence of outliers or intervention effects. Section 5.2 is a first study on robust identification of the model order for count time series by means of autocorrelations. Section 5.3 reviews robust methods for parameter estimation and prediction.

5.1 Towards a comprehensive model selection strategy

An essential step of model-assisted statistical analysis is the selection of a suitable model among the many candidates arising from different choices of the link function, of the model orders, of the type of seasonality (if any), or of the conditional distribution, even when restricting to GLM-based models. Besides well-known model selection criteria like the Akaike Information Criterion (AIC), there are several tools for model selection available in the literature, which are briefly reviewed in the following paragraphs. Some of them have been developed only for a special count time series model or even only for independent or continuous-valued data and need to be generalized.

In order to assess the predictive performance of a model in a general context, a nonparametric diagnostic approach has been proposed by Gneiting *et al.* (2007). This comprises probability integral transform histograms, marginal calibration plots, sharpness diagrams and proper scoring rules. This approach is transferred to count data by Czado *et al.* (2009). Christou and Fokianos (2014, 2015b) employ these tools for comparison of the Poisson and the negative binomial conditional distribution in the context of count time series models. These tools have been presented in 2.4 and are implemented in the package **tscount**. Moreover, there are proposals for hypothesis tests based on scoring rules. For comparison of two models, Paul and Held (2011) suggest a Monte Carlo permutation test for paired individual scores. For independent count data, Wei and Held (2014) propose calibration tests based on conditional exceedance probabilities and on proper scoring rules, which could possibly be generalized to the case of dependent data.

For choosing the conditional mean function, a specification test has been proposed by Neumann (2011) for a first order Poisson model. Fokianos and Neumann (2013) improve this procedure, proposing a goodness-of-fit test with non-trivial power for local alternatives. Christou and Fokianos (2015a) develop a score test which is also applicable when parameters are non-identifiable under the null hypothesis and apply it to test for linearity considering two non-linear specifications of the conditional mean under the alternative.

For choosing the conditional distribution (e.g. Poisson or Negative Binomial), specification tests based on the characteristic function have been proposed by Klar, Lindner, and Meintanis (2012) in the context of GARCH models and by Klar and Meintanis (2012) in the context of GLMs for independent data. These procedures could possibly be adapted to test for the conditional distribution of count time series following GLMs. Hudecová, Hušková, and Meintanis (2015) propose tests based on the probability generating function and demonstrate that they have power against the alternative of a different conditional distribution and even against the alternative of a different model class.

A test proposed by Weiß and Schweer (2015) is based on the empirical index of dispersion for the degree of overdispersion of a Poisson INGARCH(1,0) model. They employ it for testing against the alternative of additional conditional overdispersion and against the alternative of a so-called integer-valued autoregressive (INAR) model, which belongs to the class of thinning-based count time series models. This is a first approach to discriminate between these model classes, though only for a special case.

We have seen in Section 4.5.1 that model selection is a cumbersome task which requires a practitioner to take many decisions. It is often not very clear how to decide such that the whole model selection process is currently not very satisfying. All tools mentioned so far aim at single aspects of the model selection problem only. So far there is no sound advice for a comprehensive model selection strategy. For some applications there is a need for an automated model selection strategy. Consider, for example, the surveillance of infectious diseases in Germany, where several hundred thousand time series are under surveillance. This large number of time series results from about fifty (subtypes of) pathogens which each are observed in more than 400 regional units and further divided by sex and age group. For now all of them are monitored using the same quite general algorithm which works reasonably well for most time series. However, using models more tailored to e.g. each disease could potentially improve monitoring. Such models could be found by automated model selection strategies.

5.2 Robust model identification with the (partial) autocorrelation function

The proper identification of the model orders gets more complicated in the presence of outliers or intervention effects. In the following we provide a first robustness study for the identification of the model orders in case of the INGARCH model (1.2).

Two common tools for the choice of the model orders of linear time series models are the sample autocorrelation function (SACF) and the sample partial autocorrelation function (SPACF). However, these are strongly affected by outlying observations so that there is a need for robust and efficient alternatives. Dürre, Fried, and Liboschik (2015a) review robust alternatives to the SACF and the SPACF in the context of continuous-valued data. However, our investigation shows that some estimators which perform well in their study are not suitable for discrete-valued count data. Implementations of robust estimators are provided in the R package **robts** (Dürre, Fried, Liboschik, and Rathjens, 2015b).

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ be an observed time series. We consider estimation of the autocorrelation at lag h by a robust bivariate correlation estimator applied to the vector $\boldsymbol{y}_t^h = (y_{1+h}, \ldots, y_n)^\top$ and the vector of lagged observations $\boldsymbol{y}_{t-h}^h = (y_1, \ldots, y_{n-h})^\top$. We consider the rank-based correlation estimators Spearman's ρ, Kendall's τ and Gaussian rank (for a comparison in the bivariate context see Boudt, Cornelissen, and Croux, 2012). Another class of autocorrelation estimators, which is based on an idea of Gnanadesikan and Kettenring (1972), employs any robust univariate scale estimator $\widehat{\mathsf{VAR}}(\cdot)$. We use a variant bounded between -1 to $+1$ inclusive, which at lag h is given by

$$\widehat{\mathrm{ACF}}_{GK}(\boldsymbol{y}; h) = \frac{\widehat{\mathsf{VAR}}(\boldsymbol{y}_t^h + \boldsymbol{y}_{t-h}^h) - \widehat{\mathsf{VAR}}(\boldsymbol{y}_t^h - \boldsymbol{y}_{t-h}^h)}{\widehat{\mathsf{VAR}}(\boldsymbol{y}_t^h + \boldsymbol{y}_{t-h}^h) + \widehat{\mathsf{VAR}}(\boldsymbol{y}_t^h - \boldsymbol{y}_{t-h}^h)}.$$

If $\widehat{\mathsf{VAR}}(\cdot)$ is the sample variance, this is equivalent to the ordinary SACF. Ma and Genton (2000) study this Gnanadesikan-Kettenring (GK) approach in the Gaussian framework, using the highly robust Q_n estimator of scale proposed by Croux and Rousseeuw (1992). We additionally consider the median absolute deviation from the median (MAD), the 10% and 20% winsorized variance, the interquartile range (IQR), as well as the highly robust S_n (Croux and Rousseeuw, 1992) and τ (Maronna and Zamar, 2002) estimators of scale. Apart from the winsorized variance, these estimators are on the scale of the original data and need to be squared.

We compare estimators which are corrected such that they achieve consistency at the normal distribution. Note that the normal distribution is a limiting case of a Poisson distribution with mean tending to infinity. However, we cannot expect this Fisher-consistency correction to hold true, especially in the case of a clearly skewed Poisson distribution with a small mean. Moreover, the marginal distribution of a time series from an INGARCH model is strictly speaking only Poisson under the null hypothesis of independence.

In our simulation study we generate time series with 100 observations from an INGARCH(1,0) model. We consider scenarios with a true autocorrelation at lag $h = 1$ of zero ($\beta_1 = 0$) and of 0.5 ($\beta_1 = 0.5$). The results are averaged over 10 000 repetitions for each scenario and reported as a function of the marginal mean $\mu = \beta_0/(1 - \beta_1)$. The shown relative efficiencies are the ratio of the mean square errors of the SACF and the respective estimator.

The GK autocorrelation estimators based on Q_n (see Figure 5.1), S_n, MAD and IQR are unsuitable for small counts, as these estimators are unstable due to the high proportion of ties in such data. It frequently happens that the scale estimations $\widehat{\mathsf{VAR}}(\boldsymbol{y}_t^h + \boldsymbol{y}_{t-h}^h)$ and $\widehat{\mathsf{VAR}}(\boldsymbol{y}_t^h - \boldsymbol{y}_{t-h}^h)$ coincide, resulting in an autocorrelation estimate of zero, or that one or both of them collapse to zero, resulting in an estimate of ±1 or a non-computable autocorrelation estimation, respectively. Particularly for small marginal means, we get zero estimates with high probability, causing a super-efficient performance if the true autocorrelation is zero. Implosion, that is breakdown to zero, is a known problem of many robust scale estimators. But not even the Q_n estimator, which showed the best performance with respect to implosion among many other alternatives in a study of

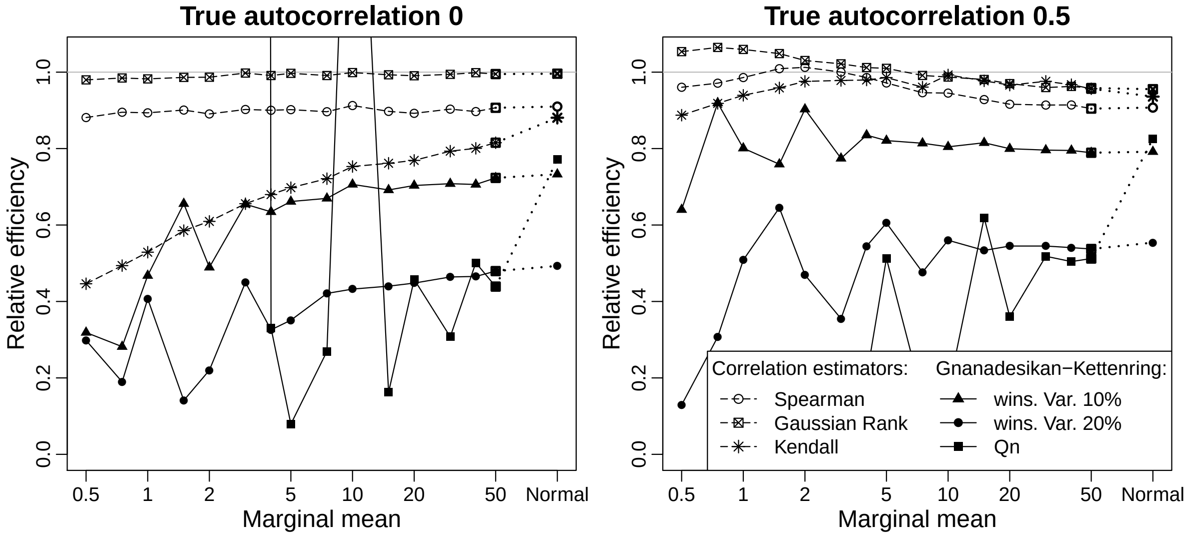


Figure 5.1: Efficiency of autocorrelation estimators at lag $h = 1$ relatively to the SACF. Time series of length 100 are simulated from an INGARCH(1,0) model with the marginal mean given on the horizontal axis and from a $N(\lambda_t, \lambda_t)$ model with a marginal mean of 50 (points on the very right of each plot).

Gather and Fried (2003), does perform acceptably in the case of small counts. We also tried variants of the Q_n using the 50%- and 75%-quantile of the pairwise distances, instead of the 25%-quantile as it is usually employed. Yet, for counts with low means none of these alternatives perform well. The τ estimator of scale as implemented by Maronna and Zamar (2002) is based on the variance estimation of the MAD and hence also performs poorly. We conclude that none of these popular highly robust scale estimators seems to be appropriate for small counts. Particularly for a low winsorizing proportion, the winsorized variance estimator results in smaller problems with stability than the estimators mentioned before and will be considered further.

Figure 5.1 reconfirms the result that the efficiency of the estimators relatively to the SACF tends to its value achieved under a normal distribution. The Gaussian rank estimator has a very high relative efficiency both for uncorrelated and autocorrelated data, which does not depend a lot on the marginal mean. Spearman's ρ correlation estimator behaves in a similar fashion, but has a lower relative efficiency of about 90% on uncorrelated data. In contrast, the relative efficiency of Kendall's τ depends very much on the marginal mean. In case of uncorrelated data its relative efficiency is below 50% for small means and even for large means slightly below Spearman's ρ.

To study the robustness properties of the estimators, we contaminate the time series of independent data, that is $\beta_1 = 0$, with a patch of 5% additive outliers in the center and the autocorrelated ones with 5% of isolated additive outliers. The first outlier scenario is

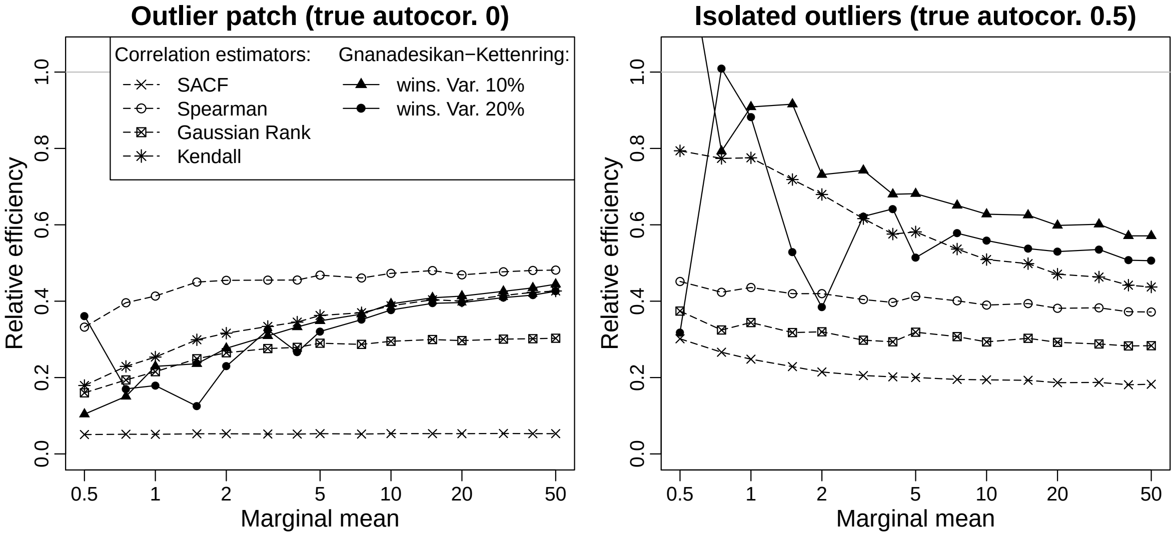


Figure 5.2: Efficiency of autocorrelation estimators at lag $h = 1$ for contaminated Poisson data relatively to the SACF for uncontaminated Poisson data. We contaminated 5% of the 100 observations with additive outliers of size five times the marginal standard deviation. Left: Patchy outliers in the center. Right: Isolated outliers at arbitrarily chosen positions 17, 40, 55, 72 and 92.

known to bias the estimation towards one and the latter one biases towards zero, which is away from the true values of zero and 0.5, respectively. For autocorrelation estimation when $\beta_1 = 0$, outlier patches are the worst case, whereas for time series with $\beta_1 > 0$ they can even compensate for an existing downward bias in finite samples. The simulation results in Figure 5.2 can be interpreted as the loss of efficiency compared to the SACF for uncontaminated data from the same model.

The outlier patch has a strong effect on the efficiency of the autocorrelation estimators for uncorrelated data (see Figure 5.2 left). The ordinary SACF is not robust and drops down to a relative efficiency of around 5%. The rank-based autocorrelation estimators show qualitatively the same pattern of increasing relative efficiency for increasing marginal mean. The Gaussian rank correlation, which has been the most efficient rank-based estimator for clean uncorrelated data, is the least robust one, because it gives more influence to the largest and the smallest observations. The 10%-winsorized variance has an efficiency of around 10% relatively to the SACF for clean data, which also increases with the marginal mean to about 40%. The 20%-winsorized variance is in principle slightly less efficient and shows a similar behavior but is, as for uncontaminated data, quite unstable for low means.

The same number of isolated outliers for moderately correlated data has a weaker effect on the efficiency of the autocorrelation estimators than the outlier patch for uncorrelated

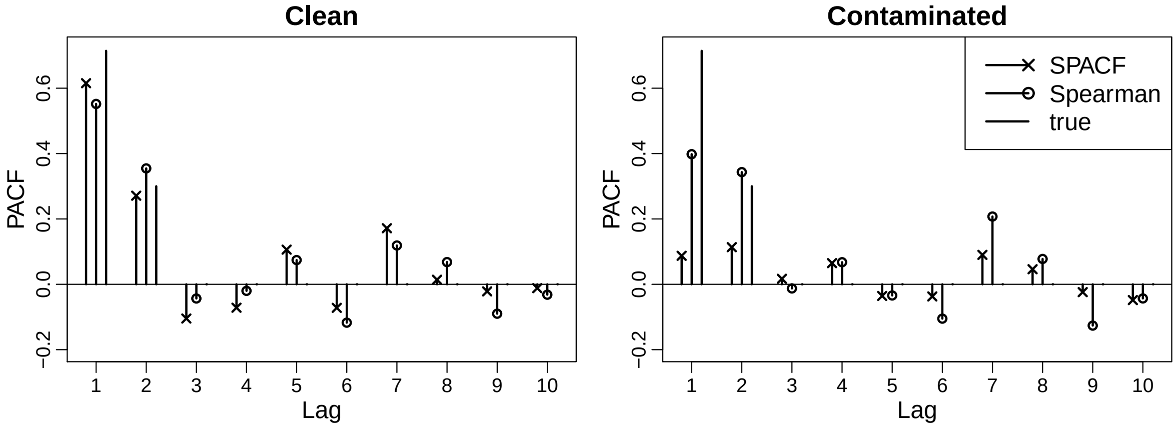


Figure 5.3: Estimated PACF of a simulated INGARCH(2,0) time series of length 100 with parameters $\beta_0 = 0.4$, $\beta_1 = 0.5$ and $\beta_2 = 0.3$. Left: Clean data. Right: Contaminated with five additive outliers of size five times the marginal standard deviation at arbitrarily chosen positions 17, 40, 55, 72 and 92.

data (see Figure 5.2 right). Unlike in the latter situation, we observe a decreasing relative efficiency for an increasing marginal mean for all estimators, except for the instability of the GK estimation based on the 20%-winsorized variance, which has been discussed before. Again, the Gaussian rank based estimator is the least efficient among the rank-based estimators, but this time Kendall's τ is much more efficient than Spearman's ρ, particularly for low marginal means.

Because of the instability of most of the other estimators we recommend to use one of the rank-based autocorrelation estimators for count time series with small counts. When choosing an autocorrelation estimator one should take into account both the desired efficiency at clean data and the desired robustness properties.

We illustrate the usefulness of robust autocorrelation estimation for identification of the model order with a simulated example. Consider a time series $(Y_t : t \in \mathbb{N}_0)$ from an INGARCH(p,0) model of unknown order $p \in \mathbb{N}_0$, with $Y_t|\mathcal{F}_{t-1}^Y \sim Pois(\lambda_t)$ and conditional mean equation $\lambda_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p}$ for $t \geq 1$. We want to determine the model order p. The time series $(Y_t : t \in \mathbb{N}_0)$ has the same second-order properties as an AR(p) model (cf. Ferland *et al.*, 2006). Hence, it is known that the partial autocorrelation function (PACF) is non-zero for lags up to p and zero for larger lags. We obtain the estimated partial autocorrelation function from the estimated autocorrelation function by applying the Durbin-Levinson algorithm (see for example Morettin, 1984).

Looking at Figure 5.3, we see that one can correctly identify the model order of an INGARCH(2,0) model by looking at the SPACF or at the estimated PACF derived from the ACF estimation based on Spearman's ρ: both estimations are clearly larger than zero for the first two lags and close to zero for all other lags. In case of a contamination with isolated outliers the non-robust estimation with the SPACF is pushed towards zero, such that one might falsely identify a model of order $p = 0$. As opposed to this, the robust estimation of the PACF with Spearman's ρ is not so strongly affected by the outliers and would still allow a correct model specification.

5.3 Robust estimation and prediction

Atypical observations occurring in the period used for model fitting complicate any model-based statistical analysis. For example in infectious disease time series there are (sometimes unrecognized) disease outbreaks or singular artifacts resulting from increased reporting due to media coverage. Estimators and predictions should thus preferably be robust against outliers or other unusual events so that prediction is not overly influenced by outliers in past data.

In order to reduce the effects of such extraordinary events, observations with large residuals can be downweighted (Farrington *et al.*, 1996, Section 3.6). This has been implemented in currently applied procedures for disease surveillance, starting from a nonrobust initial fit (Noufaily *et al.*, 2013). However, it is well known in robust statistics that this can give misleading results since such residuals are not reliable and since the degree of uncertainty is misjudged. Moreover, in models for dependent data unusual observations enter the estimation equations several times which complicates the design of appropriate weighting schemes and subsequent statistical inference.

A first approach to robust estimation of INGARCH(p,0) models has been developed by Elsaied (2012), see also Elsaied and Fried (2014, 2016). She studies robust M-estimation in this context, employing Tukey's bisquare function with a bias correction. The developed methods offer good efficiency for clean data and a superior performance as compared to maximum likelihood estimation and uncorrected M-estimation in the presence of outliers. However, note that first attempts to generalize them to INGARCH(p,q) models have failed because the observations lack a Markovian structure (Elsaied, 2012). Kitromilidou (2015); Kitromilidou and Fokianos (2015, 2016) study robust estimation for a log-linear count time series model with the conditionally unbiased bounded-influence estimator and the Mallows quasi-likelihood estimator. Robust fitting based on divergences has been suggested by Kang and Lee (2014). An approach by Agostinelli (2004) for robustly fitting of ARMA models by weighted likelihood could be transferred to count time series. Elsaied and Fried (2016) obtain promising results by applying this weighted likelihood approach (implemented in the R package **wle** by Agostinelli and SLATEC Common Mathematical Library (2013)) to independent Poisson data. For the case of independent observations, Aeberhard, Cantoni, and Heritier (2014) study robust estimation of the overdispersion coefficient from a negative binomial regression model. Such robust parameter estimations would need to be generalized to the class of count time series following GLMs and can form the basis of a robust prediction procedure.

Multi-step ahead prediction in GLM-type models for time series of counts and the derivation of corresponding prediction intervals allows generalization of currently applied procedures for infectious disease surveillance based on independence assumptions. However, the distribution of such multi-step-ahead predictions is not easily tractable so that Monte Carlo algorithms are needed. Another option is usage of better accessible successive 1-step-ahead predictions for the monitoring, but these can be strongly affected after the time of occurrence of a strongly deviating observation, so that subsequent deviating observations are masked. Using robust parameter estimates obtained from an initial sequence, 1-step ahead predictions can be calculated successively, truncating observations outside an $(1-\alpha)$ prediction interval by the respective upper or lower quantile when using it in predictions of later time points.

Alternatively, one could choose an approach which is not strictly model-based. Robust versions of exponential smoothing techniques for prediction of continuous-valued time series in the presence of outliers have been suggested by Cipra (2006), Gelper, Fried, and Croux (2010) and Cipra and Hanzák (2011). In a similar vein, Ruckdeschel, Spangl, and Pupashenko (2014) robustify Kalman filtering approaches for estimation of the state of a time series in the presence of outliers and shifts, see also Ruckdeschel (2001, 2010). These need to be adapted to discrete-valued time series.

Chapter 6

Summary

This thesis provides a unified formulation and comprehensive treatment of the class of count time series following generalized linear models. Such models link the conditional mean of the observed process to its past values, to past observations and to covariate effects. An integral part of these models is the dependence on past values of the conditional mean, the so-called feedback mechanism, which allows for parsimonious modelling of temporal correlation. We present the first systematic study on incorporating covariate effects within this framework. Models with the identity link function accommodate only positive temporal correlation and imply additive effects of necessarily nonnegative covariates. Models with the logarithmic link allow to accommodate negative temporal correlation and imply multiplicative effects of covariates. Assuming a Poisson or Negative Binomial conditional distribution facilitates model-based prediction and model assessment.

We develop likelihood-based methods for model fitting and assessment, prediction and intervention analysis for this unified model framework. This generalizes existing theory for popular special cases from the class of GLM-based count time series, for instance the so-called INGARCH model. A key contribution of this thesis is the open source implementation of these methods for the statistical software R provided by the package **tscount** (available from CRAN). This package is comprehensively documented by a vignette (the basis for Chapter 2) and detailed help pages with examples for all available functions. It complements and extends the functionality of existing software for the analysis of count time series.

A major part of this thesis is concerned with the treatment of unusual effects influencing the ordinary pattern of a count time series. This has important applications in the field of infectious diseases surveillance, which plays a prominent role in this work. In that context one aims at detecting disease outbreaks by analyzing time series of disease counts.

We show that such time series may exhibit temporal correlation and are not described adequately by the commonly used models which assume independent observations. We design procedures for the detection of unusual effects both for retrospective application and for prospective application in real-time which are based on count time series following GLMs. For the former case we study a new model which describes such intervention effects by deterministic covariates which enter the dynamics in a different way than it is the case for an existing model. We develop likelihood-based tests and detection procedures and find some robustness against misspecification of the intervention model. For the latter case we introduce a sequential procedure for online monitoring of count time series based on one-step-ahead predictions. Our procedure relies on the conditional distribution (and thus on previous observations) to decide whether a new observation is unusual. Simulations show that in case of actually dependent data our procedure is superior to a procedure based on an independence model. However, this applies only for the detection at the first unusual observation. If there are many unusually large observations in a row, e.g. due to a disease outbreak, then our procedure is influenced by the first unusual observation and has a lower chance to detect the following ones than under an independence model. We outline some ideas how to utilize the promising features of models with temporal dependence for designing a monitoring procedure whilst avoiding their undesirable shortcomings.

In the last part of this thesis we discuss some important steps on the way to an automated monitoring procedure for count time series. We review tools which could form the basis of a comprehensive model selection strategy for count time series that does not involve manual intervention. There is a need to obtain reliable results even in the presence of atypical observations. We thus present a first study on robust estimation of autocorrelation functions, which is a useful tool for identifying the model orders and potential seasonal effects. It turns out that many robust estimators which are highly efficient for continuous-valued data are not suitable for discrete-valued counts. Instead, it is recommended to use rank-based estimators of the (partial) autocorrelation function for robust model identification. Finally, we review approaches for robust estimation and prediction which could be employed to design a robust monitoring procedure.

Acknowledgements

This thesis is a result of my work in the group of Professor Roland Fried at the Department of Statistics of TU Dortmund University. I am much obliged for his guidance and support. It was also a pleasure to have Professor Konstantinos Fokianos from the University of Cyprus as my second advisor and I am deeply grateful for his commitment.

I thank the whole department and particularly my colleagues and friends Max Wornowizki and Alexander Dürre for many stimulation discussions. I further thank Pascal Kerschke for assisting me to run the simulations in Chapter 3, Philipp Probst for his contributions to an early version of the package and Jonathan Rathjens for carefully checking it. The fruitful discussions with Maëlle Salmon (formerly at Robert Koch Institute, Berlin, Germany; currently at Center for Research in Environmental Epidemiology, Barcelona, Spain) have improved both the package and this thesis. The comments by R. Fried, K. Fokianos, M. Wornowizki and M. Salmon on previous versions of this manuscript have enhanced the presentation of this research. I also thank the editors and referees of the *Journal of Statistical Software* and the *International Journal of Computer Mathematics* for their helpful and constructive comments on the manuscripts which form the basis of Chapters 2 and 3, respectively. I very much appreciated the hospitality of Stella Kitromilidou, Xanthi Pedeli, Vasiliki Christou and Maria Pitsillou during my two visits at the Department of Mathematics and Statistics of the University of Cyprus.

My research and several visits of K. Fokianos have been supported financially by the German Research Foundation (DFG, SFB 823 “Statistical modelling of nonlinear dynamic processes”). Part of this work was done while K. Fokianos was a Gambrinus Fellow at TU Dortmund University in 2014.

References

Abraham B, Box GEP (1979). "Bayesian analysis of some outlier problems in time series." *Biometrika*, **66**(2), 229–236. 10.1093/biomet/66.2.229.

Aeberhard WH, Cantoni E, Heritier S (2014). "Robust inference in the negative binomial regression model with an application to falls data." *Biometrics*, **70**(4), 920–931. 10.1111/biom.12212.

Agostinelli C (2004). "Robust time series estimation via weighted likelihood." In R Dutter, P Filzmoser, U Gather, PJ Rousseeuw (eds.), *Developments in robust statistics*, pp. 1–16. Physisca-Verlag, Heidelberg.

Agostinelli C, SLATEC Common Mathematical Library (2013). "**wle**: Weighted likelihood estimation." R package version 0.9-9, http://cran.r-project.org/package=wle.

Agosto A, Cavaliere G, Kristensen D, Rahbek A (2015). "Modeling corporate defaults: Poisson autoregressions with exogenous covariates (PARX)." *Creates research paper*, School of Economics and Management, University of Aarhus. http://econpapers.repec.org/RePEc:aah:create:2015-11.

Ahmad A, Francq C (2016). "Poisson QMLE of count time series models." *Journal of Time Series Analysis*, **37**(3), 291–314. 10.1111/jtsa.12167.

Benjamin MA, Rigby RA, Stasinopoulos DM (2003). "Generalized autoregressive moving average models." *Journal of the American Statistical Association*, **98**(461), 214–223. 10.1198/016214503388619238.

Bischl B, Lang M, Mersmann O, Rahnenführer J, Weihs C (2015). "**BatchJobs** and **BatchExperiments**: Abstraction mechanisms for using R in batch environments." *Journal of Statistical Software*, **64**(11). 10.18637/jss.v064.i11.

Bollerslev T (1986). "Generalized autoregressive conditional heteroskedasticity." *Journal of Econometrics*, **31**(3), 307–327. 10.1016/0304-4076(86)90063-1.

Boudt K, Cornelissen J, Croux C (2012). "The Gaussian rank correlation estimator: Robustness properties." *Statistics and Computing*, **22**(2), 471–483. 10.1007/s11222-011-9237-0.

Box GEP, Tiao GC (1975). "Intervention analysis with applications to economic and environmental problems." *Journal of the American Statistical Association*, **70**(349), 70–79. 10.2307/2285379.

Christou V, Fokianos K (2014). "Quasi-likelihood inference for negative binomial time series models." *Journal of Time Series Analysis*, **35**(1), 55–78. 10.1111/jtsa.12050.

Christou V, Fokianos K (2015a). "Estimation and testing linearity for non-linear mixed poisson autoregressions." *Electronic Journal of Statistics*, **9**, 1357–1377. 10.1214/15-EJS1044.

Christou V, Fokianos K (2015b). "On count time series prediction." *Journal of Statistical Computation and Simulation*, **85**(2), 357–373. 10.1080/00949655.2013.823612.

Cipra T (2006). "Robust exponential smoothing." *Journal of Forecasting*, **11**(1), 57–69. 10.1002/for.3980110106.

Cipra T, Hanzák T (2011). "Exponential smoothing for time series with outliers." *Kybernetika*, **47**(2), 165–178. http://dml.cz/dmlcz/141565.

Cox DR (1981). "Statistical analysis of time series: Some recent developments." *Scandinavian Journal of Statistics*, **8**(2), 93–115. http://www.jstor.org/stable/4615819.

Croux C, Rousseeuw PJ (1992). "Time-efficient algorithms for two highly robust estimators of scale." In Y Dodge, J Whittaker (eds.), *Computational Statistics*, volume 1, pp. 411–428. Physica-Verlag, Heidelberg.

Czado C, Gneiting T, Held L (2009). "Predictive model assessment for count data." *Biometrics*, **65**(4), 1254–1261. 10.1111/j.1541-0420.2009.01191.x.

Dawid AP (1984). "Statistical theory: The prequential approach." *Journal of the Royal Statistical Society A*, **147**(2), 278–292. 10.2307/2981683.

Demidenko E (2013). *Mixed Models: Theory and Applications with R*. Wiley series in probability and statistics, 2nd edition. John Wiley & Sons, Hoboken.

Douc R, Doukhan P, Moulines E (2013). "Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator." *Stochastic Processes and their Applications*, **123**(7), 2620–2647. 10.1016/j.spa.2013.04.010.

Doukhan P, Fokianos K, Tjøstheim D (2012). "On weak dependence conditions for Poisson autoregressions." *Statistics & Probability Letters*, **82**(5), 942–948. 10.1016/j.spl.2012.01.015.

Dunsmuir WTM, Scott DJ (2015). "The **glarma** package for observation-driven time series regression of counts." *Journal of Statistical Software*, **67**(7). 10.18637/jss.v067.i07.

Dürre A, Fried R, Liboschik T (2015a). "Robust estimation of (partial) autocorrelation." *WIREs Computational Statistics*, **7**(3), 205–222. 10.1002/wics.1351.

Dürre A, Fried R, Liboschik T, Rathjens J (2015b). "**robts**: Robust time series analysis." R package version 0.1.0, http://robts.r-forge.r-project.org.

Efron B, Tibshirani R (1993). *An Introduction to the Bootstrap*. Number 57 in Monographs on statistics and applied probability. Chapman & Hall, New York.

Elsaied H (2012). *Robust Modelling of Count Data*. Ph.D. thesis, Technische Universität Dortmund, Dortmund. `http://hdl.handle.net/2003/29404`.

Elsaied H, Fried R (2014). "Robust fitting of INARCH models." *Journal of Time Series Analysis*, **35**(6), 517–535. `10.1111/jtsa.12079`.

Elsaied H, Fried R (2016). "Tukey's M-estimator of the Poisson parameter with a special focus on small means." *Statistical Methods and Applications*, **25**(2), 191–209. `10.1007/s10260-015-0295-x`.

Fahrmeir L, Tutz G (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York.

Farrington CP, Andrews NJ, Beale AD, Catchpole MA (1996). "A statistical algorithm for the early detection of outbreaks of infectious disease." *Journal of the Royal Statistical Society A*, **159**(3), 547–563. `10.2307/2983331`.

Ferland R, Latour A, Oraichi D (2006). "Integer-valued GARCH process." *Journal of Time Series Analysis*, **27**(6), 923–942. `10.1111/j.1467-9892.2006.00496.x`.

Fokianos K (2011). "Some recent progress in count time series." *Statistics*, **45**(1), 49–58. `10.1080/02331888.2010.541250`.

Fokianos K (2012). "Count time series models." In T Subba Rao, S Subba Rao, C Rao (eds.), *Time series – methods and applications*, Handbook of Statistics, pp. 315–347. Elsevier, Amsterdam.

Fokianos K (2015). "Statistical analysis of count time series models: A glm perspective." In R Davis, S Holan, R Lund, N Ravishanker (eds.), *Handbook of discrete-valued time series*, Handbooks of Modern Statistical Methods, pp. 3–28. Chapman & Hall, London.

Fokianos K, Fried R (2010). "Interventions in INGARCH processes." *Journal of Time Series Analysis*, **31**(3), 210–225. `10.1111/j.1467-9892.2010.00657.x`.

Fokianos K, Fried R (2012). "Interventions in log-linear Poisson autoregression." *Statistical Modelling*, **12**(4), 299–322. `10.1177/1471082X1201200401`.

Fokianos K, Neumann MH (2013). "A goodness-of-fit test for Poisson count processes." *Electronic Journal of Statistics*, **7**, 793–819. `10.1214/13-EJS790`.

Fokianos K, Rahbek A, Tjøstheim D (2009). "Poisson autoregression." *Journal of the American Statistical Association*, **104**(488), 1430–1439. `10.1198/jasa.2009.tm08270`.

Fokianos K, Tjøstheim D (2011). "Log-linear Poisson autoregression." *Journal of Multivariate Analysis*, **102**(3), 563–578. `10.1016/j.jmva.2010.11.002`.

Fokianos K, Tjøstheim D (2012). "Nonlinear Poisson autoregression." *Annals of the Institute of Statistical Mathematics*, **64**(6), 1205–1225. `10.1007/s10463-012-0351-3`.

Freeland R, McCabe B (2004). "Forecasting discrete valued low count time series." *International Journal of Forecasting*, **20**(3), 427–434. `10.1016/S0169-2070(03)00014-1`.

Fried R, Liboschik T, Elsaied H, Kitromilidou S, Fokianos K (2014). “On outliers and interventions in count time series following GLMs.” *Austrian Journal of Statistics*, **43**(3), 181–193. 10.17713/ajs.v43i3.30.

Fuller WA (1996). *Introduction to Statistical Time Series*. 2nd edition. John Wiley & Sons, New York.

Gan FF (1990). “Monitoring Poisson observations using modified exponentially weighted moving average control charts.” *Communications in Statistics - Simulation and Computation*, **19**(1), 103–124. 10.1080/03610919008812847.

Gan FF (1994). “An optimal design of cumulative sum control chart based on median run length.” *Communications in Statistics - Simulation and Computation*, **23**(2), 485–503. 10.1080/03610919408813183.

Gather U, Bauer M, Fried R (2002). “The identification of multiple outliers in online monitoring data.” *Estadística*, **54**, 289–338.

Gather U, Fried R (2003). “Robust estimation of scale for local linear temporal models.” *Tatra Mountains Mathematical Publications*, **26**, 87–101.

Gelper S, Fried R, Croux C (2010). “Robust forecasting with exponential and Holt–Winters smoothing.” *Journal of Forecasting*, **29**(3), 285–300. 10.1002/for.1125.

Genest C, Nešlehová J (2007). “A primer on copulas for count data.” *ASTIN Bulletin*, **37**(02), 475–515. 10.1017/S0515036100014963.

Gnanadesikan R, Kettenring JR (1972). “Robust estimates, residuals, and outlier detection with multiresponse data.” *Biometrics*, **28**(1), 81–124. 10.2307/2528963.

Gneiting T, Balabdaoui F, Raftery AE (2007). “Probabilistic forecasts, calibration and sharpness.” *Journal of the Royal Statistical Society B*, **69**(2), 243–268. 10.1111/j.1467-9868.2007.00587.x.

Harvey AC, Durbin J (1986). “The effects of seat belt legislation on british road casualties: A case study in structural time series modelling.” *Journal of the Royal Statistical Society A*, **149**(3), 187–227. 10.2307/2981553.

Heinen A (2003). “Modelling time series count data: An autoregressive conditional poisson model.” *CORE discussion paper*, **62**. 10.2139/ssrn.1117187.

Heisterkamp SH, Dekkers ALM, Heijne JCM (2006). “Automated detection of infectious disease outbreaks: Hierarchical time series models.” *Statistics in Medicine*, **25**(24), 4179–4196. 10.1002/sim.2674.

Held L, Höhle M, Hofmann M (2005). “A statistical framework for the analysis of multivariate infectious disease surveillance counts.” *Statistical Modelling*, **5**(3), 187–199. 10.1191/1471082X05st098oa.

Held L, Paul M (2012). “Modeling seasonality in space-time infectious disease surveillance data.” *Biometrical Journal*, **54**(6), 824–843. 10.1002/bimj.201200037.

Held L, Paul M (2013). “Statistical modeling of infectious disease surveillance data.” In NM M’ikanatha, R Lynfield, CA Van Beneden, H de Valk (eds.), *Infectious disease surveillance*. John Wiley & Sons, Oxford.

Helske J (2016a). “**KFAS**: Exponential family state space models in R.” Vignette of the R package KFAS submitted to Journal of Statistical Software, `https://cran.r-project.org/web/packages/KFAS/vignettes/KFAS.pdf`.

Helske J (2016b). “**KFAS**: Kalman filter and smoother for exponential family state space models.” R package version 1.1.2, `http://cran.r-project.org/package=KFAS`.

Hilbe JM (2011). *Negative Binomial Regression*. 2nd edition. Cambridge University Press, Cambridge.

Höhle M, Paul M (2008). “Count data regression charts for the monitoring of surveillance time series.” *Computational Statistics & Data Analysis*, **52**(9), 4357–4368. `10.1016/j.csda.2008.02.015`.

Hudecová Š, Hušková M, Meintanis SG (2015). “Tests for time series of counts based on the probability-generating function.” *Statistics*, **49**(2), 316–337. `10.1080/02331888.2014.979826`.

Hutwagner L, Browne T, Seeman G, Fleischauer A (2005). “Comparing aberration detection methods with simulated data.” *Emerging infectious diseases*, **11**(2), 314–316. `10.3201/eid1102.040587`.

Jung R, Tremayne A (2011). “Useful models for time series of counts or simply wrong ones?” *AStA Advances in Statistical Analysis*, **95**(1), 59–91. `10.1007/s10182-010-0139-9`.

Kang J, Lee S (2014). “Minimum density power divergence estimator for Poisson autoregressive models.” *Computational Statistics & Data Analysis*, **80**, 44–56. `10.1016/j.csda.2014.06.009`.

Kedem B, Fokianos K (2002). *Regression Models for Time Series Analysis*. Wiley series in probability and statistics. John Wiley & Sons, Hoboken.

King AA, Nguyen D, Ionides EL (2016). “Statistical inference for partially observed markov processes via the R package **pomp**.” *Journal of Statistical Software*, **69**(12), 1–43. `10.18637/jss.v069.i12`. To appear.

Kitromilidou S (2015). *Robust Inference for Log-Linear Count Time Series Models*. Ph.D. thesis, University of Cyprus, Nicosia.

Kitromilidou S, Fokianos K (2015). “Mallows’ quasi-likelihood estimation for log-linear Poisson autoregressions.” *Statistical Inference for Stochastic Processes*. `10.1007/s11203-015-9131-z`. Published online.

Kitromilidou S, Fokianos K (2016). “Robust estimation methods for a class of log-linear count time series models.” *Journal of Statistical Computation and Simulation*, **86**(4), 740–755. `10.1080/00949655.2015.1035271`.

Klar B, Lindner F, Meintanis SG (2012). "Specification tests for the error distribution in GARCH models." *Computational Statistics & Data Analysis*, **56**(11), 3587–3598. 10.1016/j.csda.2010.05.029.

Klar B, Meintanis SG (2012). "Specification tests for the response distribution in generalized linear models." *Computational Statistics*, **27**(2), 251–267. 10.1007/s00180-011-0253-5.

Kourentzes N (2014). "On intermittent demand model optimisation and selection." *International Journal of Production Economics*, **156**, 180–190. 10.1016/j.ijpe.2014.06.007.

Kourentzes N, Petropoulos F (2016). "**tsintermittent**: Intermittent time series forecasting." R package version 1.9, http://cran.r-project.org/package=tsintermittent.

Lai TL, Chan HP (2008). "Discussion on "Is average run length to false alarm always an informative criterion?" by Yajun Mei." *Sequential Analysis*, **27**(4), 385–388. 10.1080/07474940802445964.

Lange K (1999). *Numerical Analysis for Statisticians.* Statistics and computing. Springer-Verlag, New York.

Liboschik T, Fokianos K, Fried R (2015). "**tscount**: An R package for analysis of count time series following generalized linear models." *TU Dortmund, SFB 823 Discussion Paper*, **06/15**. 10.17877/DE290R-7239.

Liboschik T, Kerschke P, Fokianos K, Fried R (2016). "Modelling interventions in INGARCH processes." *International Journal of Computer Mathematics*, **93**(4), 640–657. 10.1080/00207160.2014.949250.

Lindgren F, Rue H (2015). "Bayesian spatial modelling with R-INLA." *Journal of Statistical Software*, **63**(19). 10.18637/jss.v063.i19.

Loeys T, Moerkerke B, De Smet O, Buysse A (2012). "The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression." *British Journal of Mathematical and Statistical Psychology*, **65**(1), 163–180. 10.1111/j.2044-8317.2011.02031.x.

Lotze T, Shmueli G, Yahav I (2007). "Simulating multivariate syndromic time series and outbreak signatures." *Robert H. Smith School Research Paper*, **RHS-06-054**. 10.2139/ssrn.990020.

Ma Y, Genton MG (2000). "Highly robust estimation of the autocovariance function." *Journal of Time Series Analysis*, **21**(6), 663–684. 10.1111/1467-9892.00203.

Manitz J, Höhle M (2013). "Bayesian outbreak detection algorithm for monitoring reported cases of campylobacteriosis in Germany." *Biometrical Journal*, **55**(4), 509–526. 10.1002/bimj.201200141.

Maronna RA, Zamar RH (2002). "Robust estimates of location and dispersion for high-dimensional datasets." *Technometrics*, **44**(4), 307–317. 10.1198/004017002188618509.

Masarotto G, Varin C (2012). "Gaussian copula marginal regression." *Electronic Journal of Statistics*, **6**, 1517–1549. 10.1214/12-EJS721.

Morettin PA (1984). "The Levinson algorithm and its applications in time series analysis." *International Statistical Review / Revue Internationale de Statistique*, **52**(1), 83–92. 10.2307/1403247.

Moysiadis T, Fokianos K (2014). "On binary and categorical time series models with feedback." *Journal of Multivariate Analysis*, **131**, 209–228. 10.1016/j.jmva.2014.07.004.

Nelder JA, Wedderburn RWM (1972). "Generalized linear models." *Journal of the Royal Statistical Society A*, **135**(3), 370–384. 10.2307/2344614.

Neumann MH (2011). "Absolute regularity and ergodicity of Poisson count processes." *Bernoulli*, **17**(4), 1268–1284. 10.3150/10-BEJ313.

Noufaily A, Enki DG, Farrington P, Garthwaite P, Andrews N, Charlett A (2013). "An improved algorithm for outbreak detection in multiple surveillance systems." *Statistics in Medicine*, **32**(7), 1206–1222. 10.1002/sim.5595.

Pan W (2001). "Akaike's information criterion in generalized estimating equations." *Biometrics*, **57**(1), 120–125. 10.1111/j.0006-341X.2001.00120.x.

Paul M, Held L (2011). "Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts." *Statistics in Medicine*, **30**(10), 1118–1136. 10.1002/sim.4177.

Pedeli X, Karlis D (2013). "On composite likelihood estimation of a multivariate INAR(1) model." *Journal of Time Series Analysis*, **34**(2), 206–220. 10.1111/jtsa.12003.

Psarakis S, Vyniou AK, Castagliola P (2014). "Some recent developments on the effects of parameter estimation on control charts." *Quality and Reliability Engineering International*, **30**(8), 1113–1129. 10.1002/qre.1556.

R Core Team (2016). "R – A language and environment for statistical computing." http://www.r-project.org.

Rigby RA, Stasinopoulos DM (2005). "Generalized additive models for location, scale and shape (with discussion)." *Journal of the Royal Statistical Society C*, **54**(3), 507–554. 10.1111/j.1467-9876.2005.00510.x.

Robert Koch Institut (2015). "SurvStat@RKI 2.0. Web-based query on data reported under the German 'Protection against infection act'."

Rossi G, Lampugnani L, Marchi M (1999). "An approximate CUSUM procedure for surveillance of health events." *Statistics in Medicine*, **18**(16), 2111–2122. 10.1002/(SICI)1097-0258(19990830)18:16<2111::AID-SIM171>3.0.CO;2-Q.

Ruckdeschel P (2001). *Ansätze Zur Robustifizierung Des Kalman-Filters*. Ph.D. thesis, Universität Bayreuth, Bayreuth. http://www.mathematik.uni-kl.de/~ruckdesc/pubs/Diss_hpt.pdf.

Ruckdeschel P (2010). “Optimally (distributional-)robust Kalman filtering.” Unpublished manuscript, `http://arxiv.org/abs/1004.3393`.

Ruckdeschel P, Spangl B, Pupashenko D (2014). “Robust Kalman tracking and smoothing with propagating and non-propagating outliers.” *Statistical Papers*, **55**(1), 93–123. `10.1007/s00362-012-0496-4`.

Rue H, Martino S, Chopin N (2009). “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(2), 319–392. `10.1111/j.1467-9868.2008.00700.x`.

Salmon M, Schumacher D, Burmann H, Frank C, Claus H, Höhle M (2016a). “A system for automated outbreak detection of communicable diseases in Germany.” *Eurosurveillance*, **21**(13). `10.2807/1560-7917.ES.2016.21.13.30180`.

Salmon M, Schumacher D, Höhle M (2016b). “Monitoring count time series in R: Aberration detection in public health surveillance.” *Journal of Statistical Software*, **70**(10), 1–35. `10.18637/jss.v070.i10`.

Salmon M, Schumacher D, Stark K, Höhle M (2015). “Bayesian outbreak detection in the presence of reporting delays.” *Biometrical Journal*, **57**(6), 1051–1067. `10.1002/bimj.201400159`.

Schmid W (1995). “On the run length of a Shewhart chart for correlated data.” *Statistical Papers*, **36**(1), 111–130. `10.1007/BF02926025`.

Scotto MG, Weiß CH, Silva ME, Pereira I (2014). “Bivariate binomial autoregressive models.” *Journal of Multivariate Analysis*, **125**, 233–251. `10.1016/j.jmva.2013.12.014`.

Siakoulis V (2015). “**acp**: Autoregressive conditional Poisson.” R package version 2.1, `http://cran.r-project.org/package=acp`.

Sim T (2016). *Maximum Likelihood Estimation in Partially Observed Markov Models with Applications to Time Series of Counts.* Ph.D. thesis, Télécom ParisTech, Paris.

Stasinopoulos DM, Rigby RA, Eilers P (2015). “**gamlss.util**: GAMLSS utilities.” R package version 4.3-2, `http://cran.r-project.org/package=gamlss.util`.

Tartakovsky AG (2008). “Discussion on “Is average run length to false alarm always an informative criterion?” by Yajun Mei.” *Sequential Analysis*, **27**(4), 396–405. `10.1080/07474940802446046`.

Tjøstheim D (2012). “Some recent theory for autoregressive count time series.” *TEST*, **21**(3), 413–438. `10.1007/s11749-012-0296-0`.

Tjøstheim D (2015). “Count time series with observation-driven autoregressive parameter dynamics.” In R Davis, S Holan, R Lund, N Ravishanker (eds.), *Handbook of discrete-valued time series*, Handbooks of Modern Statistical Methods, pp. 77–100. Chapman & Hall, London.

Unkel S, Farrington CP, Garthwaite PH, Robertson C, Andrews N (2012). "Statistical methods for the prospective detection of infectious disease outbreaks: A review." *Journal of the Royal Statistical Society A*, **175**(1), 49–82. 10.1111/j.1467-985X.2011.00714.x.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. Statistics and computing, 4th edition. Springer-Verlag, New York.

Ver Hoef JM, Boveng PL (2007). "Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data?" *Ecology*, **88**(11), 2766–2772. 10.1890/07-0043.1.

Wang S (2013). "**pbs**: Periodic b splines." R package version 1.1, http://cran.r-project.org/package=pbs.

Wei W, Held L (2014). "Calibration tests for count data." *TEST*, **23**(4), 787–805. 10.1007/s11749-014-0380-8.

Wei W, Schüpbach G, Held L (2015). "Time-series analysis of Campylobacter incidence in Switzerland." *Epidemiology & Infection*, **143**(09), 1982–1989. 10.1017/S0950268814002738.

Weiß CH (2008). "Thinning operations for modeling time series of counts—a survey." *Advances in Statistical Analysis*, **92**(3), 319–341. 10.1007/s10182-008-0072-3.

Weiß CH (2015). "SPC methods for time-dependent processes of counts—A literature review." *Cogent Mathematics*, **2**(1), 1111116. 10.1080/23311835.2015.1111116.

Weiß CH, Schweer S (2015). "Detecting overdispersion in INARCH(1) processes." *Statistica Neerlandica*, **69**(3), 281–297. 10.1111/stan.12059.

Weiß CH, Testik MC (2012). "Detection of abrupt changes in count data time series: Cumulative sum derivations for INARCH (1) models." *Journal of quality technology*, **44**(3), 249–264.

Woodard DB, Matteson DS, Henderson SG (2011). "Stationarity of generalized autoregressive moving average models." *Electronic Journal of Statistics*, **5**, 800–828. 10.1214/11-EJS627.

Yang M, Cavanaugh JE, Zamba GK (2015). "State-space models for count time series with excess zeros." *Statistical Modelling*, **15**(1), 70–90. 10.1177/1471082X14535530.

Yang M, Zamba GK, Cavanaugh JE (2013). "Markov regression models for count time series with excess zeros: A partial likelihood approach." *Statistical Methodology*, **14**, 26–38. 10.1016/j.stamet.2013.02.001.

Yang M, Zamba GK, Cavanaugh JE (2014). "**ZIM**: Zero-inflated models for count time series with excess zeros." R package version 1.0.2, http://cran.r-project.org/package=ZIM.

Yee TW (2015). *Vector generalized linear and additive models: with an implementation in R*. Springer series in statistics. Springer-Verlag, New York.

Yee TW (2016). "**VGAM**: Vector generalized linear and additive models." R package version 1.0-1, `http://cran.r-project.org/package=VGAM`.

Yee TW, Wild CJ (1996). "Vector generalized additive models." *Journal of the Royal Statistical Society B*, **58**(3), 481–493. `http://www.jstor.org/stable/2345888`.

Zeileis A, Kleiber C, Jackman S (2008). "Regression models for count data in R." *Journal of Statistical Software*, **27**(8), 1–25. `10.18637/jss.v027.i08`.

Zhu F (2012). "Zero-inflated Poisson and negative binomial integer-valued GARCH models." *Journal of Statistical Planning and Inference*, **142**(4), 826–839. `10.1016/j.jspi.2011.10.002`.

Appendix A

Implementation details

A.1 Parameter space for the log-linear model

The parameter space Θ for the log-linear model (1.3) which guarantees a stationary and ergodic solution of the process is subject of current research. In the current implementation of `tsglm` the parameters need to fulfill the condition

$$\max\left\{|\beta_1|, \ldots, |\beta_p|, |\alpha_1|, \ldots, |\alpha_q|, \left|\sum_{k=1}^{p} \beta_k + \sum_{\ell=1}^{q} \alpha_\ell\right|\right\} < 1. \tag{A.1}$$

At the time we started developing **tscount**, (A.1) appeared as a reasonable extension of the condition

$$\max\left\{|\beta_1|, |\alpha_1|, |\beta_1 + \alpha_1|\right\} < 1,$$

which Douc *et al.* (2013, Lemma 14) derive for $p = q = 1$. However, in a recent work, Sim (2016, Proposition 5.4.7) derives sufficient conditions for a model of order $p = q$. For the first order model he obtains the weaker condition

$$\max\left\{|\alpha_1|, |\beta_1 + \alpha_1|\right\} < 1.$$

For $p = q = 2$ the required condition is

$$\begin{aligned}\max\Big\{ & |\alpha_1| + |\alpha_2| + |\beta_2|, |\alpha_1\alpha_2| + \left|\alpha_2\alpha_1^2\right| + |\alpha_1 + \beta_2|, |\alpha_2| + |\alpha_1 + \beta_1| + |\beta_2|, \\ & |\alpha_2(\alpha_1 + \beta_1)| + |\alpha_2 + \alpha_1(\alpha_1 + \beta_1)| + |(\alpha_1 + \beta_1)\beta_2|, |\alpha_2| + |\alpha_1| + |\beta_2|,\end{aligned}$$

$$|\alpha_2(\alpha_1+\beta_1)| + |\alpha_2+\alpha_1(\alpha_1+\beta_1)| + |(\alpha_1+\beta_1)\beta_2|\,, |\alpha_2| + |\alpha_1+\beta_1| + |\beta_2|\,,$$
$$|(\alpha_1+\beta_1)\alpha_2| + \left|\alpha_2+(\alpha_1+\beta_1)^2+\beta_2\right| + |(\alpha_1+\beta_1)\beta_2|\,\Big\} \; < 1. \qquad (A.2)$$

There are parameters which fulfill (A.1) but not (A.2) (e.g. $\beta_1 = -0.9$, $\beta_2 = 0.9$, $\alpha_1 = 0$, $\alpha_2 = 0$) and vice versa (e.g. $\beta_1 = -1.8$, $\beta_2 = 0$, $\alpha_1 = 0.9$, $\alpha_2 = 0$). However, there exists a large intersection between values which fulfill (A.1) and (A.2). For the general case $p = q$ the condition can be obtained by considering the maximum among p elements of the norms of matrix products with p factors, where each factor corresponds to a $(2p-1)\times(2p-1)$ matrix. The implementation of this condition is a challenging problem and therefore we have decided in favor of (A.1). Alternatively, we can obtain unconstrained estimates (argument `final.control = list(constrained = NULL)`), which should be examined carefully.

A.2 Recursions for inference and their initialization

Let h be the inverse of the link function g and let $h'(x) = \partial h(x)/\partial x$ be its derivative. In the case of the identity link $g(x) = x$ it holds $h(x) = x$ and $h'(x) = 1$ and in the case of the logarithmic link $g(x) = \log(x)$ it holds $h(x) = h'(x) = \exp(x)$. The partial derivative of the conditional mean $\lambda_t(\boldsymbol{\theta})$ is given by

$$\frac{\partial\lambda_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = h'\left(\nu_t(\boldsymbol{\theta})\right)\frac{\partial\nu_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}},$$

where the vector of partial derivatives of the linear predictor $\nu_t(\boldsymbol{\theta})$,

$$\frac{\partial\nu_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \left(\frac{\partial\nu_t(\boldsymbol{\theta})}{\partial\beta_0}, \frac{\partial\nu_t(\boldsymbol{\theta})}{\partial\beta_1}, \ldots, \frac{\partial\nu_t(\boldsymbol{\theta})}{\partial\beta_p}, \frac{\partial\nu_t(\boldsymbol{\theta})}{\partial\alpha_1}, \ldots, \frac{\partial\nu_t(\boldsymbol{\theta})}{\partial\alpha_q}, \frac{\partial\nu_t(\boldsymbol{\theta})}{\partial\eta_1}, \ldots, \frac{\partial\nu_t(\boldsymbol{\theta})}{\partial\eta_r}\right)^\top,$$

can be computed recursively. The recursions are given by

$$\begin{aligned}
\frac{\partial\nu_t(\boldsymbol{\theta})}{\partial\beta_0} &= 1 + \sum_{\ell=1}^{q}\alpha_\ell\frac{\partial\nu_{t-j_\ell}(\boldsymbol{\theta})}{\partial\beta_0},\\
\frac{\partial\nu_t(\boldsymbol{\theta})}{\partial\beta_s} &= \widetilde{g}(Y_{t-i_s}) + \sum_{\ell=1}^{q}\alpha_\ell\frac{\partial\nu_{t-j_\ell}(\boldsymbol{\theta})}{\partial\beta_s}, \quad s = 1,\ldots,p,\\
\frac{\partial\nu_t(\boldsymbol{\theta})}{\partial\alpha_s} &= \sum_{\ell=1}^{q}\alpha_\ell\frac{\partial\nu_{t-j_\ell}(\boldsymbol{\theta})}{\partial\alpha_s} + \nu_{t-j_s}(\boldsymbol{\theta}), \quad s = 1,\ldots,q,\\
\frac{\partial\nu_t(\boldsymbol{\theta})}{\partial\eta_s} &= \sum_{\ell=1}^{q}\alpha_\ell\frac{\partial\nu_{t-j_\ell}(\boldsymbol{\theta})}{\partial\eta_s} + X_{t,s}, \quad s = 1,\ldots,r.
\end{aligned}$$

The recursions for the linear predictor $\nu_t = g(\lambda_t)$ and its partial derivatives depend on past values of the linear predictor and its derivatives, which are generally not observable. We implemented three possibilities for initialization of these values. The default and preferable choice is to initialize by the respective marginal expectations, assuming a model without covariate effects, such that the process is stationary (argument `init.method = "marginal"`). For the linear model (1.2) it holds (Ferland *et al.*, 2006)

$$\mathsf{E}(Y_t) = \mathsf{E}(\nu_t) = \frac{\beta_0}{1 - \sum_{k=1}^{p} \beta_k - \sum_{\ell=1}^{q} \alpha_\ell} =: \mu(\boldsymbol{\theta}). \tag{A.3}$$

For the log-linear model (1.3) we instead consider the transformed time series $Z_t := \log(Y_t + 1)$, which has approximately the same second order properties as a time series from the linear model (1.2). It approximately holds $\mathsf{E}(Z_t) \approx \mathsf{E}(\nu_t) \approx \mu(\boldsymbol{\theta})$. Specifically, we initialize past values of ν_t by $\mu(\boldsymbol{\theta})$ and past values of $\partial\nu_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ by

$$\frac{\partial\mu(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \left(\frac{\partial\mu(\boldsymbol{\theta})}{\partial\beta_0}, \frac{\partial\mu(\boldsymbol{\theta})}{\partial\beta_1}, \ldots, \frac{\partial\mu(\boldsymbol{\theta})}{\partial\beta_p}, \frac{\partial\mu(\boldsymbol{\theta})}{\partial\alpha_1}, \ldots, \frac{\partial\mu(\boldsymbol{\theta})}{\partial\alpha_q}, \frac{\partial\mu(\boldsymbol{\theta})}{\partial\eta_1}, \ldots, \frac{\partial\mu(\boldsymbol{\theta})}{\partial\eta_r}\right)^\top,$$

which is explicitly given by

$$\begin{aligned}
\frac{\partial\mu(\boldsymbol{\theta})}{\partial\beta_0} &= \frac{1}{1 - \sum_{k=1}^{p} \beta_k - \sum_{\ell=1}^{q} \alpha_\ell}, \\
\frac{\partial\mu(\boldsymbol{\theta})}{\partial\beta_k} &= \frac{\partial\mu(\boldsymbol{\theta})}{\partial\alpha_\ell} = \frac{\beta_0}{\left(1 - \sum_{k=1}^{p} \beta_k - \sum_{\ell=1}^{q} \alpha_\ell\right)^2}, \quad k = 1, \ldots, p, \ \ell = 1, \ldots, q, \quad \text{and} \\
\frac{\partial\mu(\boldsymbol{\theta})}{\partial\eta_m} &= 0, \quad m = 1, \ldots, r.
\end{aligned}$$

Another possibility is to initialize ν_t by β_0 and $\partial\nu_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ by zero. In this case the model corresponds to standard i.i.d. Poisson random variables (argument `init.method = "iid"`). A third possibility would be a data-dependent initialization of ν_t, for example by $\widetilde{g}(y_1)$. In this case, the partial derivatives of ν_t are initialized by zero (argument `init.method = "firstobs"`).

The recursions also depend on unavailable past observations of the time series, prior to the sample which is used for the likelihood computation. The package allows to choose between two strategies to cope with that. The default choice is to replace these pre-sample observations by the same initializations as used for the linear predictor ν_t (see above), transformed by the inverse link function h (argument `init.drop = FALSE`). An alternative is to use the first i_p observations for initialization and to compute the log-likelihood on the remaining observations $y_{i_p+1}, \ldots, y_n$ (argument `init.drop = TRUE`). Recall that i_p is the highest order for regression on past observations.

	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\alpha}_1$	$\ell(\widehat{\boldsymbol{\theta}})$
`init.method = "marginal", init.drop = FALSE`	0.500	0.733	0.249	-3024.7
`init.method = "marginal", init.drop = TRUE`	0.567	0.746	0.236	-2568.0
`init.method = "iid", init.drop = FALSE`	0.867	0.757	0.218	-3037.2
`init.method = "iid", init.drop = TRUE`	0.563	0.738	0.246	-2587.8
`init.method = "firstobs", init.drop = FALSE`	0.559	0.739	0.246	-3018.7
`init.method = "firstobs", init.drop = TRUE`	0.559	0.739	0.246	-2578.1

Table A.1: Estimated parameters and log-likelihood of a time series of length 1000 simulated from model (1.2) for different initialization strategies. The true parameters are $\beta_0 = 0.5$, $\beta_1 = 0.77$ and $\alpha_1 = 0.22$. Likelihood values are included for completeness of the presentation. There are not comparable as they are based on a different number of observations.

Particularly in the presence of strong serial dependence, the different methods for initialization can affect the estimation substantially even for quite long time series with 1000 observations. We illustrate this by the simulated example presented in Table A.1.

A.3 Starting value for optimization

The numerical optimization of the log-likelihood function requires a starting value for the parameter vector $\boldsymbol{\theta}$, which can be obtained by initial estimation based on a simpler model. Different strategies for this (controlled by the argument `start.control`) are discussed in this section. We call this start estimation (and not initial estimation) to avoid confusion with the initialization of the recursions described in the previous section.

The start estimation by the R function `glm` utilizes the fact that a time series following a GLM without feedback (as in Kedem and Fokianos, 2002) can be fitted by employing standard software. Neglecting the feedback mechanism, the parameters of the GLM

$$Y_t|\mathcal{F}^*_{t-1} \sim \text{Poi}(\lambda^*_t), \text{ with } \nu^*_t = g(\lambda^*_t) \text{ and}$$
$$\nu^*_t = \beta^*_0 + \beta^*_1 \widetilde{g}(Y_{t-i_1}) + \ldots + \beta^*_p \widetilde{g}(Y_{t-i_p}) + \eta^*_1 X_{t,1} + \ldots + \eta^*_r X_{t,r},\ t = i_p + 1, \ldots, n,$$

with $\mathcal{F}^*_t$ the history of the joint process $\{Y_t, \boldsymbol{X}_t\}$, are estimated using the R function `glm`. Denote the estimated parameters by $\widehat{\beta}^*_0$, $\widehat{\beta}^*_1, \ldots, \widehat{\beta}^*_p$, $\widehat{\eta}^*_1, \ldots, \widehat{\eta}^*_r$ and set $\widehat{\alpha}^*_1, \ldots, \widehat{\alpha}^*_q$ to zero (argument `start.control$method = "GLM"`).

Fokianos *et al.* (2009) suggest start estimation of $\boldsymbol{\theta}$, for the first order linear model (1.2) without covariates, by employing its representation as an ARMA(1,1) process with identical second-order properties, see Ferland *et al.* (2006). For arbitrary orders P and

Q with $s := \max(P, Q)$ and the general model from Section 1.2 this representation, after straightforward calculations, is given by

$$(\widetilde{g}(Y_t) - \underbrace{\mu(\boldsymbol{\theta})}_{=:\zeta}) - \sum_{i=1}^{s} \underbrace{(\beta_i + \alpha_i)}_{=:\varphi_i} (\widetilde{g}(Y_{t-i}) - \mu(\boldsymbol{\theta})) = \varepsilon_t + \sum_{i=1}^{q} \underbrace{(-\alpha_i)}_{=:\psi_i} \varepsilon_{t-i}, \qquad \text{(A.4)}$$

where $\beta_i := 0$ for $i \notin P$, $\alpha_i := 0$ for $i \notin Q$ and $\{\varepsilon_t\}$ is a white noise process. Recall that $\widetilde{g}$ is defined by $\widetilde{g}(x) = x$ for the linear model and $\widetilde{g}(x) = \log(x + 1)$ for the log-linear model. Given the autoregressive parameters φ_i and the moving average parameters ψ_i of the ARMA representation of $\{Y_t\}$, the parameters of the original process are obtained by $\alpha_i = -\psi_i$ and $\beta_i = \varphi_i + \psi_i$. We get β_0 from $\beta_0 = \zeta\big(1 - \sum_{k=1}^{p} \beta_k - \sum_{\ell=1}^{q} \alpha_\ell\big)$ using the formula for the marginal mean of $\{Y_t\}$. With these formulas estimates $\widehat{\beta}_0^*$, $\widehat{\beta}_i^*$ and $\widehat{\alpha}_i^*$ are obtained from ARMA estimates $\widehat{\zeta}$, $\widehat{\varphi}_i$ and $\widehat{\psi}_i$. Estimation of the ARMA parameters is implemented by conditional least squares (argument `start.control$method = "CSS"`), maximum likelihood assuming normally distributed errors (argument `start.control$method = "ML"`), or, for models up to first order, the method of moments (argument `start.control$method = "MM"`). If covariates are included, a linear regression is fitted to $\widetilde{g}(Y_t)$, whose errors follow an ARMA model like (A.4). Consequently, the covariate effects do not enter the dynamics of the process, as it is the case in the actual model (1.1). It would be preferable to fit an ARMAX model, in which covariate effects are included on the right hand side of (A.4), but this is currently not readily available in R.

We compare both approaches to obtain start estimates. The GLM approach apparently disregards the feedback mechanism, i.e., the dependence on past values of the conditional mean. As opposed to this, the ARMA approach does not treat covariate effects in an appropriate way. From extensive simulations we note that the final estimation results are almost equally good for both approaches.

However, we also found out that in some situations (especially in the presence of certain types of covariates) both approaches occasionally provoke the likelihood optimization algorithms to run into a local optimum. This happens more often for increasing sample size. To overcome this problem we recommend a naive start estimation assuming an i.i.d. model without covariates, which only estimates the intercept and sets all other parameters to zero (argument `start.control$method = "iid"`). This starting value is usually not close to any local optimum of the likelihood function. Hence we expect, possibly, a larger number of steps for the optimization algorithm to converge. This is the default method of start estimation as we do not guarantee a global optimum with the other two methods, in some special cases.

Particularly for the linear model, some of the aforementioned approaches do not yield a starting value $\widehat{\boldsymbol{\theta}}^* = (\widehat{\beta}_0^*, \widehat{\beta}_1^*, \ldots, \widehat{\beta}_p^*, \widehat{\alpha}_1^*, \ldots, \widehat{\alpha}_q^*, \widehat{\eta}_1^*, \ldots, \widehat{\eta}_r^*)^\top$ for $\boldsymbol{\theta}$ which lays in the interior of the parameter space Θ. To overcome this problem, $\widehat{\boldsymbol{\theta}}^*$ is suitably transformed to be used as a starting value. For the linear model (1.2) this transformation is done according to the following procedure (Liboschik *et al.*, 2016):

1a. Set $\widehat{\beta}_k^* := \min\left\{\widehat{\beta}_k^*, \varepsilon\right\}$ and $\widehat{\alpha}_\ell^* := \min\left\{\widehat{\alpha}_\ell^*, \varepsilon\right\}$.

1b. If $c := \sum_{k=1}^p \widehat{\beta}_k^* + \sum_{\ell=1}^q \widehat{\alpha}_\ell^* > 1 - \xi - \varepsilon$, then shrink each $\widehat{\beta}_k^*$ and $\widehat{\alpha}_\ell^*$ by multiplication with the factor $(1 - \xi - \varepsilon)/c$.

2a. Set $\widehat{\beta}_0^* := \widehat{\beta}_0^* \cdot \left(1 - \sum_{k=1}^p \widehat{\beta}_k^* + \sum_{\ell=1}^q \widehat{\alpha}_\ell^*\right)/c$.

2b. Set $\widehat{\beta}_0^* := \max\left\{\widehat{\beta}_0^*, \xi + \varepsilon\right\}$.

3. Set $\widehat{\eta}_m^* := \max\left\{\widehat{\eta}_m^*, \varepsilon\right\}$.

A small constant $\varepsilon > 0$ ensures that the initial value lies inside the parameter space Θ and not on its boundaries. It is chosen to be $\varepsilon = 10^{-6}$ by default (argument `epsilon`). Another small constant $\xi > 0$ enforces the inequalities to be strict (i.e. $<$ instead of $\leq$). This constant is set to $\xi = 10^{-6}$ by default (argument `slackvar`); recall Section 2.2.3. The shrinkage factor in step 1b is chosen such that the sum of the parameters equals $1 - \xi - \varepsilon$ after possible shrinkage in this step. The choice of $\widehat{\beta}_0^*$ in step 2a ensures that the marginal mean remains unchanged after possible shrinkage in step 1b. For the log-linear model (1.3) it is not necessary to ensure positivity of the parameters. A valid starting value $\widehat{\boldsymbol{\theta}}^*$ is transformed with the following procedure:

1a. Set $\widehat{\beta}_k^* := \operatorname{sign}\big(\widehat{\beta}_k^*\big) \cdot \min\left\{\left|\widehat{\beta}_k^*\right|, \varepsilon\right\}$ and $\widehat{\alpha}_\ell^* := \operatorname{sign}\big(\widehat{\alpha}_\ell^*\big) \cdot \min\left\{\left|\widehat{\alpha}_\ell^*\right|, \varepsilon\right\}$.

1b. If $c := \left|\sum_{k=1}^p \widehat{\beta}_k^* + \sum_{\ell=1}^q \widehat{\alpha}_\ell^*\right| > 1 - \xi - \varepsilon$, then shrink each $\widehat{\beta}_k^*$ and $\widehat{\alpha}_\ell^*$ by multiplication with the factor $(1 - \xi - \varepsilon)/c$.

A.4 Stable inversion of the information matrix

In order to obtain standard errors from the normal approximation (2.6) one needs to invert the information matrix $G_n(\widehat{\boldsymbol{\theta}}; \widehat{\sigma}^2)$. To avoid numerical instabilities we make use of the fact that an information matrix is a real symmetric and positive definite matrix. We first compute a Choleski factorization of the information matrix. Then we apply an efficient algorithm to invert the matrix employing the upper triangular factor of the Choleski decomposition (see R functions `chol` and `chol2inv`). This procedure is implemented in the function `invertinfo` in our package.

Appendix B

Additional simulations

In this section we present simulations supporting that the methods that have not yet been treated thoroughly in the literature work reliably.

B.1 Covariates

We present some limited simulation results for the problem of including covariates, in both linear and log-linear models. For simplicity we employ first order models with one covariate and a conditional Poisson distribution, that is, we consider the linear model with the identity link function

$$Y_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t), \quad \lambda_t = \beta_0 + \beta_1 Y_{t-1} + \alpha_1 \lambda_{t-1} + \eta_1 X_t, \quad t = 1, \ldots, n,$$

and the log-linear model with the logarithmic link function

$$Y_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t), \quad \log(\lambda_t) = \beta_0 + \beta_1 \log(Y_{t-1} + 1) + \alpha_1 \log(\lambda_{t-1}) + \eta_1 X_t, \quad t = 1, \ldots, n.$$

The parameters are chosen to be $\beta_1 = 0.3$ and $\alpha_1 = 0.2$. The intercept parameter is $\beta_0 = 4 \cdot 0.5$ for the linear and $\beta_0 = \log(4) \cdot 0.5$ for the log-linear model in order to obtain a marginal mean (without the covariate effect) of about 4 in both cases. We consider the covariates listed in Table B.1, covering a simple linear trend, seasonality, intervention effects, i.i.d. observations from different distributions and a stochastic process. The covariates are chosen to be nonnegative, which is necessary for the linear model but not for the log-linear model. All covariates have a mean of about 0.5, such that their effect

Abbreviation	**Definition**
Linear	t/n
Sine	$(\sin(2\pi \cdot 5 \cdot t/n) + 1)/2$
Spiky outlier	$\mathbb{1}(t = \tau)$
Transient shift	$0.8^{t-\tau}\mathbb{1}(t \geq \tau)$
Level shift	$\mathbb{1}(t \geq \tau)$
GARCH(1,1)	$\sqrt{h_t}\varepsilon_t$ with $\varepsilon_t \sim \mathrm{N}(0.5, 1)$ and $h_t = 0.002 + 0.1X_{t-1}^2 + 0.8h_{t-1}$
Exponential	i.i.d. Exponential with mean 0.5
Normal	i.i.d. Normal with mean 0.5 and variance 0.04

Table B.1: Covariates $\{X_t : t = 1, \ldots, n\}$ considered in the simulation study. The interventions occur at time $\tau = n/2$. The GARCH model is defined recursively (see Bollerslev, 1986).

sizes are somewhat comparable. The regression coefficient is chosen to be $\eta_1 = 2 \cdot \beta_0$ for the linear and $\eta_1 = 1.5 \cdot \beta_0$ for the log-linear model.

Apparently, certain types of covariates can to some extent be confused with serial dependence. This is the case for the linear trend and the level shift, but also for the sinusoidal term, since these lead to data patterns which resemble positive serial correlation; see Figure B.1.

A second finding is that the effect of covariates, like a transient shift or a spiky outlier, is hard to be estimated precisely. Note that both covariates have most of their values values different from zero only at very few time points (especially the spiky outlier) which explains this behavior of the estimation procedure. The estimators for the coefficients of such covariates have a large variance which decreases only very slowly with growing sample size; see the bottom right plots in Figures B.2 and B.3 for the linear and the log-linear model, respectively. This does not affect the estimation of the other parameters, see the other three plots in the same figures. For all other types of covariates the variance of the estimator for the regression parameter decreases with growing sample size, which indicates consistency of the estimator.

The conjectured approximative normality of the model parameters stated in (2.6) seems to hold for most of the covariates considered here even in case of a rather moderate sample size of 100, as indicated by the QQ plots shown in Figure B.4. The only serious deviation from normality happens for the spiky outlier in the linear model, a case where many estimates of the covariate coefficient η_1 lie close to zero. This value is the lower boundary of the parameter space for this model. Due to the consistency problem for this covariate (discussed in the previous paragraph) the observed deviation from normality is still present even for a much larger sample size of 2000 (not shown here). Note that

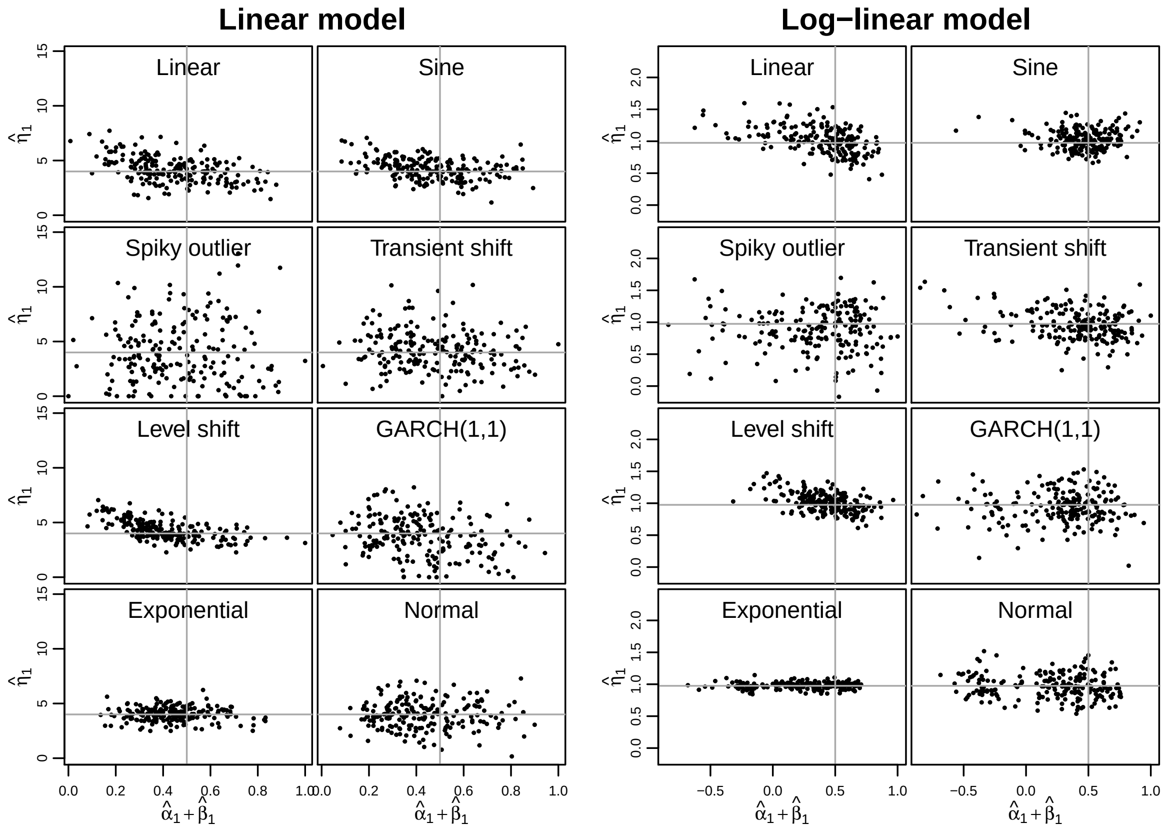


Figure B.1: Scatterplots of the estimated covariate parameter $\widehat{\eta}_1$ against the sum $\widehat{\beta}_1 + \widehat{\alpha}_1$ of the estimated dependence parameters in a linear (left) respectively log-linear (right) model with an additional covariate of the given type. The time series of length $n = 100$ are simulated from the respective model with the true values marked by grey lines. Each dot represents one of 200 replications.

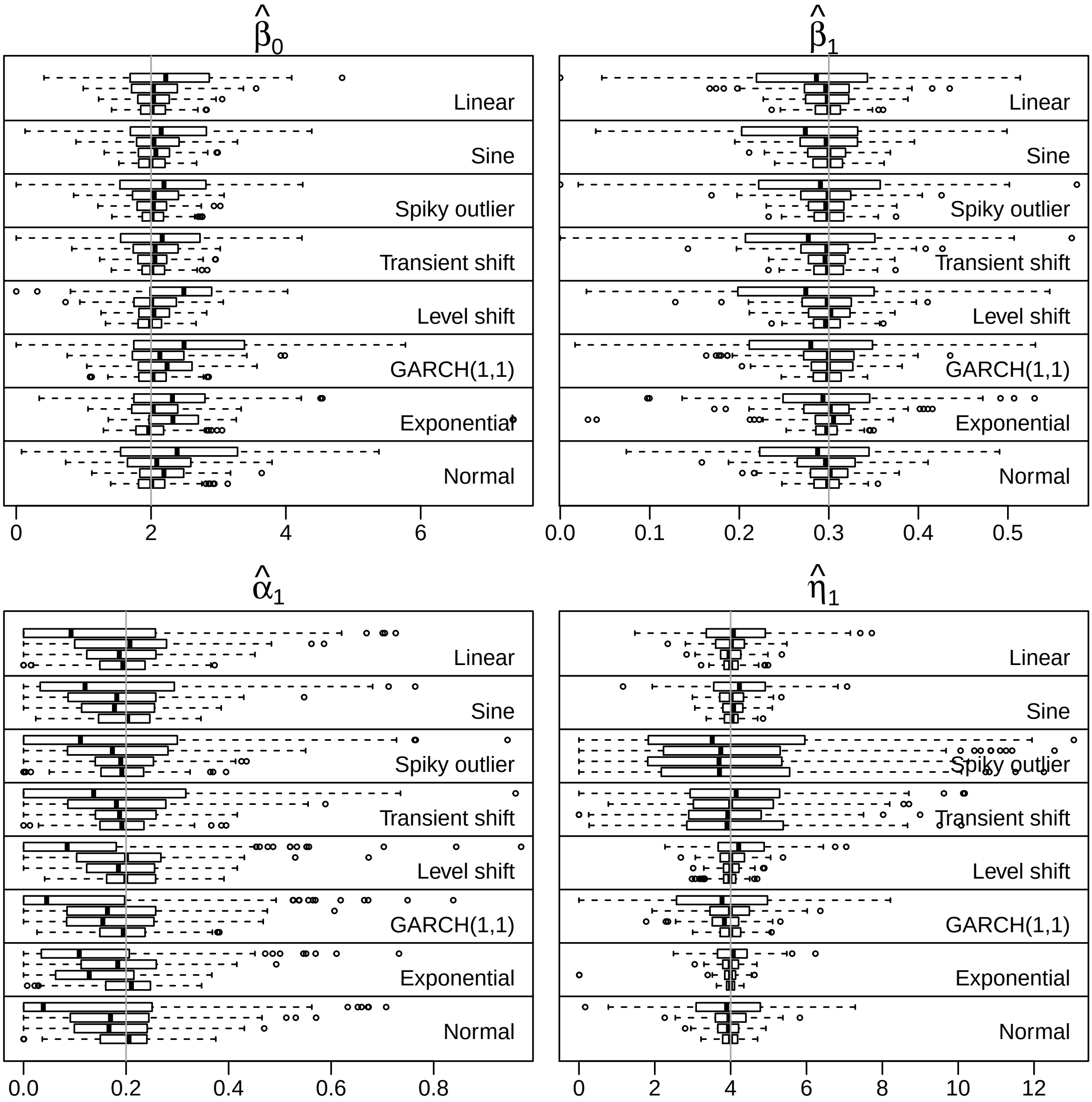


Figure B.2: Estimated coefficients for a linear model of order $p = q = 1$ with an additional covariate of the given type. The time series of length $n = 100, 500, 1000, 2000$ (from top to bottom in each panel) are simulated from the respective model with the true coefficients marked by a grey vertical line. Each boxplot is based on 200 replications.

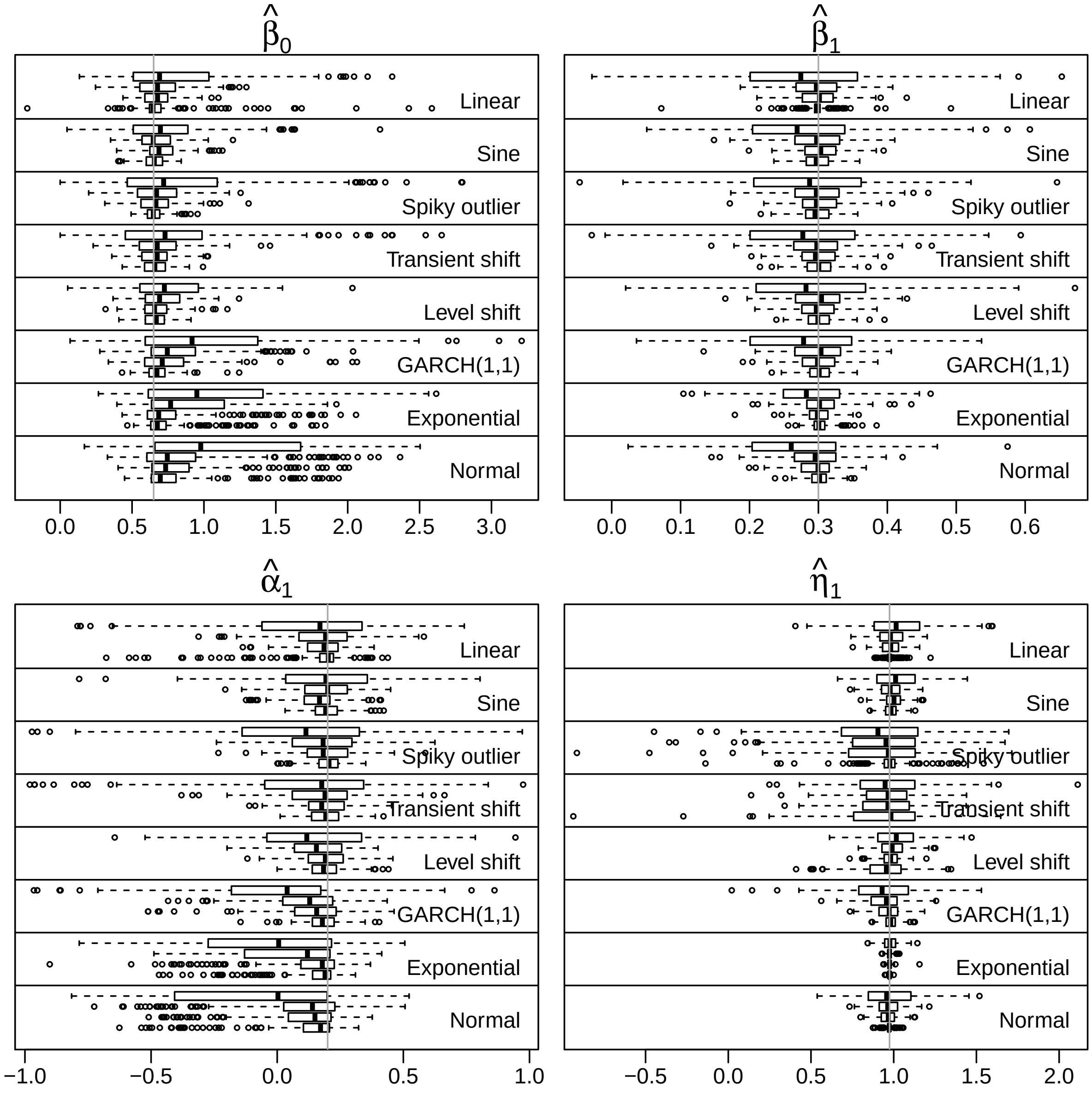


Figure B.3: Identical simulation results as those shown in Figure B.2 but for the log-linear model.

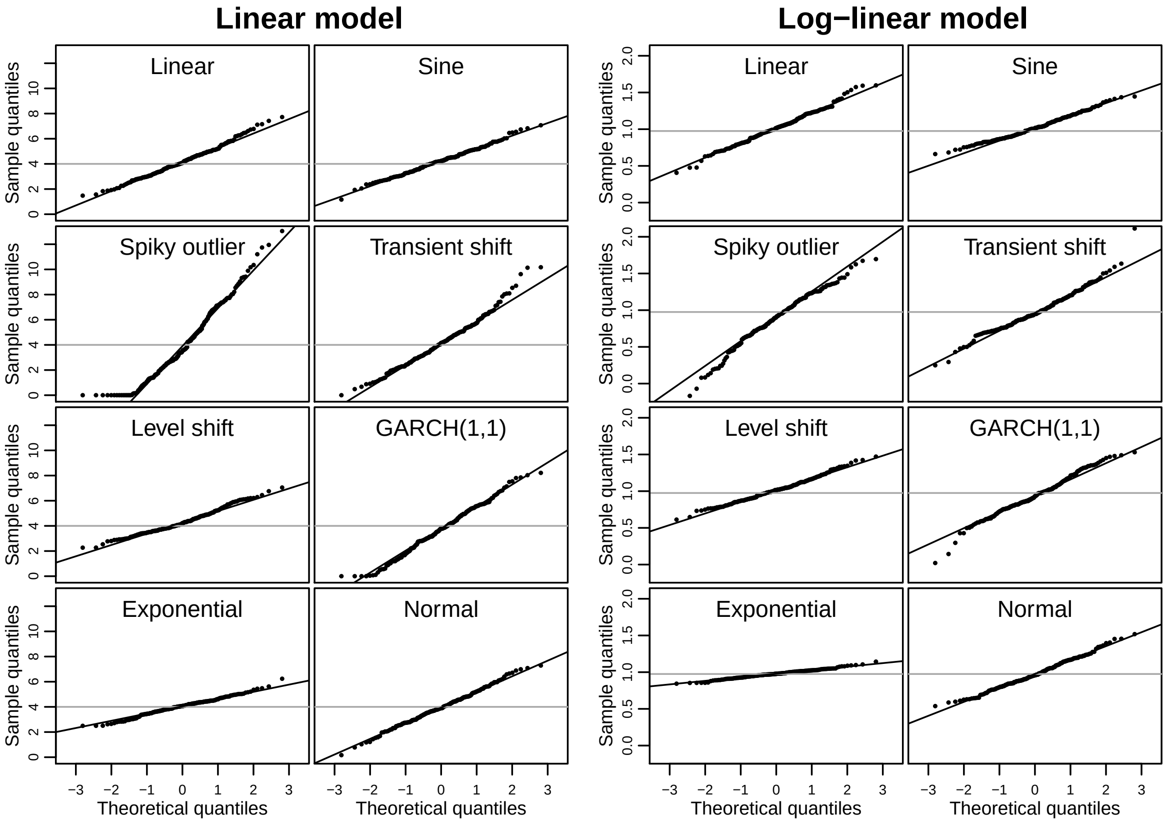


Figure B.4: Normal QQ-plots for the estimated covariate coefficient $\widehat{\eta}_1$ in a linear (left) respectively log-linear (right) model of order $p = q = 1$ with an additional covariate of the given type. The time series of length $n = 100$ are simulated from the respective model with the true coefficient marked by a grey horizontal line. Each plot is based on 200 replications.

	Mean	**Median**	**Std.dev.**	**MAD**	**Failures (in %)**
$\sigma^2 = 1.00$	0.98	0.97	0.18	0.17	0.00
0.20	0.20	0.20	0.05	0.05	0.00
0.10	0.10	0.10	0.03	0.03	0.10
0.05	0.05	0.05	0.03	0.03	3.30
0.00	0.02	0.02	0.01	0.01	52.10

Table B.2: Summary statistics for the estimated overdispersion coefficient $\widehat{\sigma}^2$ of the Negative Binomial distribution. The time series are simulated from a log-linear model with the true overdispersion coefficient given in the rows. Each statistic is based on 200 replications.

for the spiky outlier the conditions for asymptotic normality in linear regression models stated in Section 2.2.2 are not fulfilled. QQ plots for the other model parameters β_0, β_1 and α_1 look satisfactory for all types of covariates and are not shown here.

B.2 Negative Binomial distribution

As mentioned before, the model with the logarithmic link function is not covered by the theory derived by Christou and Fokianos (2014). Consequently, we confirm by simulations that estimating the additional dispersion parameter ϕ of the Negative Binomial distribution by equation (2.5) yields good results. We consider both the linear model with the identity link

$$Y_t|\mathcal{F}_{t-1} \sim \text{NegBin}(\lambda_t, \phi), \quad \lambda_t = \beta_0 + \beta_1 Y_{t-1} + \alpha_1 \lambda_{t-1}, \quad t = 1, \ldots, n,$$

and the log-linear model with the logarithmic link

$$Y_t|\mathcal{F}_{t-1} \sim \text{NegBin}(\lambda_t, \phi), \quad \log(\lambda_t) = \beta_0 + \beta_1 \log(Y_{t-1} + 1) + \alpha_1 \log(\lambda_{t-1}), \quad t = 1, \ldots, n.$$

The parameters β_0, β_1 and α_1 are chosen like in Section B.1. For the dispersion parameter ϕ we employ the values 1, 5, 10, 20 and ∞, which are corresponding to overdispersion coefficients σ^2 of 1, 0.2, 0.1, 0.05 and 0, respectively.

The estimator of the dispersion parameter ϕ has a positively skewed distribution. It is thus preferable to consider the distribution of its inverse $\widehat{\sigma}^2 = 1/\widehat{\phi}$, which is only slightly negatively skewed; see Table B.2. In certain cases it is numerically not possible to solve (2.5) and the estimation fails. This happens when the true value of ϕ is large and we are close to the limiting case of a Poisson distribution (see the proportion of failures

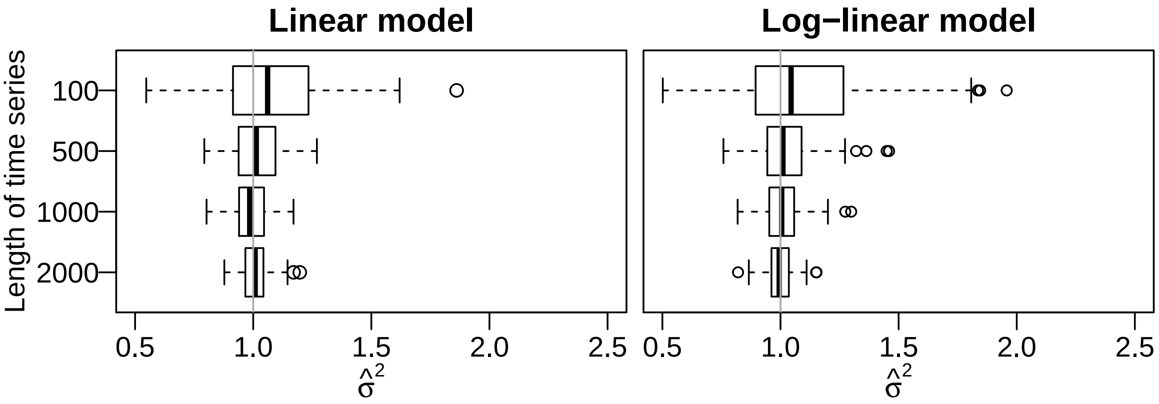


Figure B.5: Estimated overdispersion coefficient $\widehat{\sigma}^2$ of the Negative Binomial distribution for a linear (left) respectively log-linear (right) model of order $p = q = 1$. The time series are simulated from the respective model with the true overdispersion coefficient marked by a grey vertical line. Each boxplot is based on 200 replications.

in the last column of the table). In such a case our fitting function gives an error and recommends fitting a model with a Poisson distribution instead. These results are very similar for the linear model and thus not shown here.

We check the consistency of the estimator by a simulation for a true value of $\sigma^2 = 1/\phi = 1$. Our results shown in Figure B.5 indicate that on average the deviation of the estimation from the true value decreases with increasing sample size for both, the linear and the log-linear model. The boxplots also confirm our above finding that the estimator has a clearly asymmetric distribution for sample sizes up to several hundred.

B.3 Quasi information criterion

We confirm by simulation that the quasi information criterion (QIC) approximates Akaike's information criterion (AIC) in case of a Poisson distribution. Like in Section B.2, we consider both the linear model with the identity link

$$Y_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t), \quad \lambda_t = \beta_0 + \beta_1 Y_{t-1} + \alpha_1 \lambda_{t-1}, \quad t = 1, \ldots, n,$$

and the log-linear model with the logarithmic link

$$Y_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t), \quad \log(\lambda_t) = \beta_0 + \beta_1 \log(Y_{t-1} + 1) + \alpha_1 \log(\lambda_{t-1}), \quad t = 1, \ldots, n,$$

but now with a Poisson distribution. Again, the parameters β_0, β_1 and α_1 are chosen like in Section B.1.

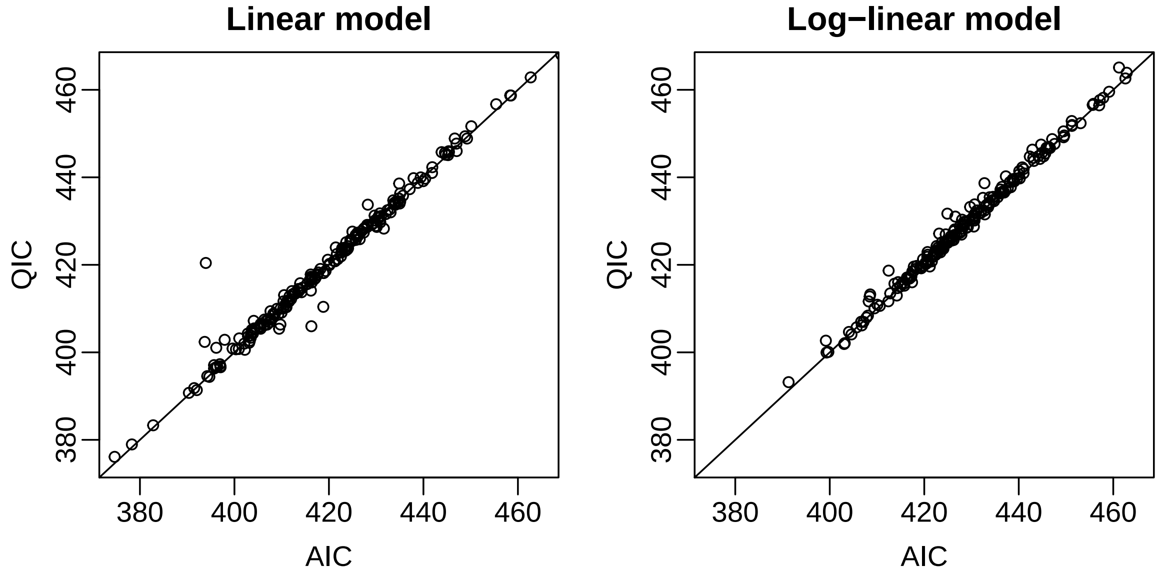


Figure B.6: Relationship of QIC and AIC for a linear (left) respectively log-linear (right) model of order $p = q = 1$. Each of the 200 points represents the QIC and AIC of a fit to a time series of length $n = 100$ simulated from the respective model. The diagonal line is the identity, i.e. it represents values for which the QIC equals the AIC.

From each of the two models we simulate 200 time series of length $n = 100$ and compute the QIC and AIC of the fitted model. Figure B.6 shows that the relationship between QIC and AIC is very close to the identity, i.e. the QIC is approximately equal to the AIC. There is only one out of 200 cases (for the linear model) where the QIC deviates largely from the AIC.