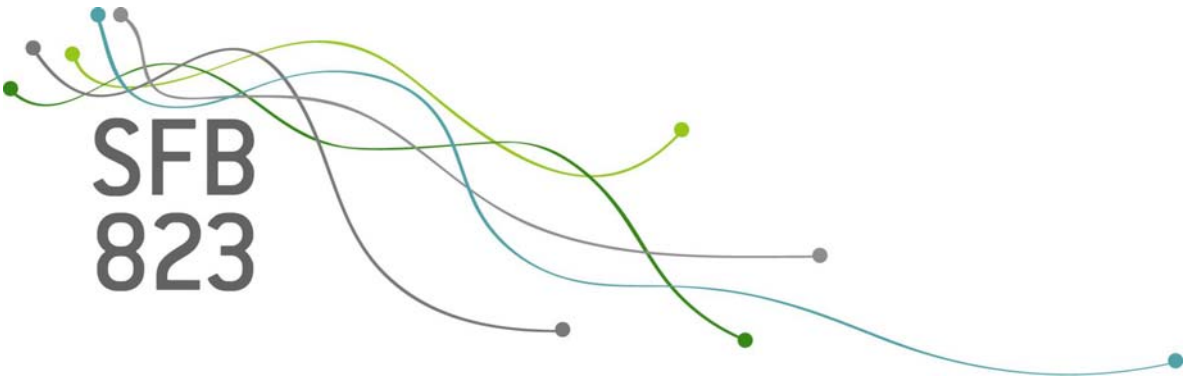
SFB 823

Discussion Paper

tscount: An R package for analysis of count time series following generalized linear models

Tobias Liboschik, Konstantinos Fokianos, Roland Fried

Nr. 6/2015



tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models

Tobias Liboschik
TU Dortmund University

Konstantinos Fokianos
University of Cyprus

Roland Fried
TU Dortmund University

Abstract

The R package **tscount** provides likelihood-based estimation methods for analysis and modelling of count time series following generalized linear models. This is a flexible class of models which can describe serial correlation in a parsimonious way. The conditional mean of the process is linked to its past values, to past observations and to potential covariate effects. The package allows for 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. The package includes methods for model fitting and assessment, prediction and intervention analysis. This paper summarizes the theoretical background of these methods with references to the literature for further details. 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 demonstrated by two data examples.

Keywords: intervention analysis, mixed Poisson, model selection, prediction, R, regression model, serial correlation.

1. Introduction

Recently, there has been an increasing interest in regression models for time series of counts and a quite considerable number of publications on this subject has appeared in the literature. However, most of the proposed methods are not yet available in statistical software packages 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 2014). In fact, our main goal is to develop software for models that include a latent process similar to the case of ordinary generalized autoregressive conditional heteroscedasticity (GARCH) models (Bollerslev 1986).

Count time series appear naturally in various areas whenever a number of events per time period is observed over time. Examples showing the wide range of applications are the daily number of hospital admissions from public health, 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,

choosing a distribution suitable for count data and an appropriate link function. This approach is pursued in detail 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 (for a recent review see Weiß 2008). Another 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.

In the first version of the **tscount** package we provide likelihood-based methods for the framework of count time series following GLMs. For independent data or some simple dependence structures of low order these models can be fitted with standard software for GLMs (see Section A.2); for example the R function `glm` produces accurate results. The implementations in our package **tscount** allows for a more general dependence structure which can be specified conveniently by the user. Accordingly we fit time series models which include a latent process, similarly to the GARCH class of models. The usage and output of our functions is inspired by the R functions `glm` and `arima`. We provide many standard `S3` methods which are known from other functions. The related R package **acp** (Siakoulis 2014) has been published recently and provides maximum likelihood fitting of a simplified first order version of models that we consider. Our package **tscount** covers a much wider class of models and includes this model as a special case.

The functionality of our package **tscount** partly goes beyond the theory available in the literature since theoretical investigation of these models is still an ongoing research theme. For instance consider the problem of accommodating covariates in such GLM-type count time series models or fitting a mixed Poisson log-linear model. These topics 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 paper 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 introduces the models we consider. Section 3 describes quasi maximum likelihood estimation of the unknown model parameters and gives some details regarding its implementation. Section 4 treats prediction with such models. Section 5 sums up tools for model assessment. Section 6 discusses procedures for detection of interventions. Section 7 demonstrates the usage of the package with two data examples. Finally, Section 8 gives an outlook on possible future extensions of the package. In the Appendix we give some additional details and confirm by simulation some of the new methods which have not yet been treated theoretically in the literature.

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 $\mathsf{E}(Y_t|\mathcal{F}_{t-1})$ of the count time series by a latent mean process, say $\{\lambda_t : t \in \mathbb{N}\}$, such that

$\mathsf{E}(Y_t|\mathcal{F}_{t-1}) = \lambda_t$. Denote by $\mathcal{F}_t$ the history of the joint process $\{Y_t, \lambda_t, \boldsymbol{X}_{t+1} : t \in \mathbb{N}\}$ up to time t including the covariate information at time $t+1$. The distributional assumption for Y_t given $\mathcal{F}_{t-1}$ is discussed later. We are interested in models of the general form

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

where $g : \mathbb{R}^+ \to \mathbb{R}$ is a link function and $\widetilde{g} : \mathbb{R}^+ \to \mathbb{R}$ 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 $P = \{i_1, i_2, \ldots, i_p\}$ with $p \in \mathbb{N}_0$ and integers $0 < i_1 < i_2 \ldots < i_p < \infty$. 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 $Q = \{j_1, j_2, \ldots, j_q\}$ with $q \in \mathbb{N}_0$ and integers $0 < j_1 < j_2 \ldots < j_q < \infty$ for regression on lagged latent means $\lambda_{t-j_1}, \lambda_{t-j_2}, \ldots, \lambda_{t-j_q}$. This more general case is covered by the theory for models with $P = \{1, \ldots, p\}$ and $Q = \{1, \ldots, q\}$, which are usually treated in the literature, by choosing p and q sufficiently large and setting unnecessary model parameters to zero.

We give several examples of model (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 we obtain from (1) that

$$\lambda_t = \beta_0 + \sum_{k=1}^{p} \beta_k \, Y_{t-k} + \sum_{\ell=1}^{q} \alpha_\ell \lambda_{t-\ell}. \tag{2}$$

Assuming further that Y_t given the past is Poisson distributed, then we obtain an *integer-valued GARCH model* of order p and q, in short INGARCH(p,q). These models have been discussed by Heinen (2003), Ferland, Latour, and Oraichi (2006) and Fokianos, Rahbek, and Tjøstheim (2009), among others. When $\boldsymbol{\eta} \neq \mathbf{0}$, then our package fits INGARCH models with nonnegative covariates; this is so because we need to ensure that the resulting mean process is positive. An example of an INGARCH model with covariates is given in Section 6, where we fit a count time series model which includes intervention effects.

Consider again model (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) that

$$\nu_t = \beta_0 + \sum_{k=1}^{p} \beta_k \, \log(Y_{t-k} + 1) + \sum_{\ell=1}^{q} \alpha_\ell \nu_{t-\ell}. \tag{3}$$

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 (see Fokianos and Tjøstheim (2011) for a discussion and for showing that the addition of a constant to each observation to avoid zeros does not affect inference). Note that model (3) allows modeling of negative serial correlation, whereas (2) accommodates positive serial correlation only. Additionally, (3) accommodates covariates easier than (2) since the log-linear model implies positivity of the conditional mean process $\{\lambda_t\}$. The effects of covariates on the response is multiplicative for

model (3); it is additive for model (2). For a discussion on the inclusion of time-dependent covariates see Fokianos and Tjøstheim (2011, Section 4.3).

Model (1) together with the *Poisson* assumption, i.e., $Y_t|\mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t)$, implies that

$$\mathsf{P}\left(Y_t = y|\mathcal{F}_{t-1}\right) = \frac{\lambda_t^y \exp(-\lambda_t)}{y!}, \quad y = 0, 1, \ldots. \tag{4}$$

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

The *Negative Binomial* distribution allows for a conditional variance larger than λ_t. 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 $\phi \in (0, \infty)$, i.e.,

$$\mathsf{P}\left(Y_t = y|\mathcal{F}_{t-1}\right) = \frac{\Gamma(\phi + y)}{\Gamma(y+1)\Gamma(\phi)} \left(\frac{\phi}{\phi + \lambda_t}\right)^{\phi} \left(\frac{\lambda_t}{\phi + \lambda_t}\right)^{y}, \quad y = 0, 1, \ldots. \tag{5}$$

In this case, $\mathsf{VAR}\left(Y_t|\mathcal{F}_{t-1}\right) = \lambda_t + \lambda_t^2/\phi$, 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. 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 4 and 5.

In model (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 latent 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 latent means. Following Liboschik, Kerschke, Fokianos, and Fried (2014), 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 $\text{diag}(\boldsymbol{e})$ a diagonal matrix with diagonal elements given by $\boldsymbol{e}$. The generalization of (1) allowing for both internal and external covariate effects then reads

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \, \tilde{g}(Y_{t-i_k}) + \sum_{\ell=1}^{q} \alpha_\ell \left(g(\lambda_{t-j_\ell}) - \boldsymbol{\eta}^\top \text{diag}(\boldsymbol{e}) \boldsymbol{X}_{t-j_\ell}\right) + \boldsymbol{\eta}^\top \boldsymbol{X}_t. \tag{6}$$

Basically, the effect of all covariates with an external effect is subtracted in the feedback terms such that their effect does enter the dynamics of the process only via the observations. We

refer to Liboschik *et al.* (2014) for an extensive discussion of internal versus external effects. It is our experience with these models that an empirical discrimination between internal and external covariate effects is difficult and that it is not crucial which type of covariate effect to fit in practical applications.

3. Estimation and inference

The **tscount** package fits models of the form (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 (2) with covariates is given by

$$\Theta = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{p+q+r+1} : \ \beta_0 > 0, \ \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_q, \eta_1, \ldots, \eta_r \geqslant 0, \ \sum_{i=1}^{p} \beta_i + \sum_{j=1}^{q} \alpha_j < 1 \right\}.$$

The intercept β_0 must be positive and all other parameters must be nonnegative to ensure positivity of the latent mean process. The other condition ensures that the fitted model has a stationary solution (cf. Ferland *et al.* 2006, Proposition 1). For the log-linear model (3) with covariates the parameter space is taken to be

$$\Theta = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{p+q+r+1} : \ |\beta_1|, \ldots, |\beta_p|, |\alpha_1|, \ldots, |\alpha_q| < 1, \ \left| \sum_{i=1}^{p} \beta_i + \sum_{j=1}^{q} \alpha_j \right| < 1 \right\}.$$

This is intended to be the generalization (for model order p,q) of the conditions $|\beta_1| < 1$, $|\alpha_1| < 1$ and $|\beta_1 + \alpha_1| < 1$, which Douc *et al.* (2013, Lemma 14) derive for the first order model. Christou and Fokianos (2014) point out that with the parametrization (5) 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.

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

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{n} \log p_t(y_t; \boldsymbol{\theta}) \propto \sum_{t=1}^{n} \Big(y_t \ln(\lambda_t(\boldsymbol{\theta})) - \lambda_t(\boldsymbol{\theta}) \Big), \tag{7}$$

where $p_t(y; \boldsymbol{\theta}) = \mathsf{P}(Y_t = y | \mathcal{F}_{t-1})$ is the p.d.f. of a Poisson distribution as defined in (4). The latent mean process 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

$$S_n(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} \left(\frac{y_t}{\lambda_t(\boldsymbol{\theta})} - 1 \right) \frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{8}$$

The vector of partial derivatives $\partial\lambda_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ can be computed recursively by the recursions given in Appendix A.1. Finally, the conditional information matrix is given by

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

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 (QML) estimator $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ is, assuming that it exists, the solution of the non-linear constrained optimization problem

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta}\ell(\boldsymbol{\theta}). \tag{9}$$

As proposed by Christou and Fokianos (2014), the dispersion parameter ϕ of the Negative Binomial distribution is estimated by solving the equation

$$\sum_{t=1}^{n}\frac{(Y_t-\widehat{\lambda}_t)^2}{\widehat{\lambda}_t(1+\widehat{\lambda}_t/\widehat{\phi})} = n-m, \tag{10}$$

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 (3) does not fall into the class of models considered by Christou and Fokianos (2014). However, results obtained by Douc *et al.* (2013) allow to use this estimator also for the log-linear model (for $p = q = 1$). This issue is addressed by simulations in Appendix B.2, which support that the estimator obtained by (10) provides good results also for models with the logarithmic link function.

Inference for the regression parameters is based on the asymptotic normality of the QML estimator, which has been shown 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

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right) \xrightarrow{d} N_{p+q+r+1}\left(\mathbf{0}, G_n^{-1}(\widehat{\boldsymbol{\theta}};\widehat{\sigma}^2)G_n^*(\widehat{\boldsymbol{\theta}})G_n^{-1}(\widehat{\boldsymbol{\theta}};\widehat{\sigma}^2)\right), \tag{11}$$

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 $||\boldsymbol{X}_t|| < c$, 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 $\mathsf{E}(\boldsymbol{X}_t)$ 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. The asymptotic normality of the QML estimator 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 $G_n(\widehat{\boldsymbol{\theta}};\widehat{\sigma}^2)$ we apply

an algorithm which makes use of the fact that it is a real symmetric and positive definite matrix; see Appendix A.3.

An alternative to the normal approximation (11) for obtaining standard errors is a parametric bootstrap procedure, which is part of our package (function `se`). 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. This procedure can compute standard errors not only for the estimated regression parameters but also for the dispersion coefficient $\widehat{\sigma}^2$.

Implementation

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 $\Theta = \{\boldsymbol{\theta} | \boldsymbol{U}\boldsymbol{\theta} \geqslant \boldsymbol{c}\}$. For the linear model (2) one needs $d = p+q+r+2$ constraints to ensure nonnegativity of the latent mean λ_t and stationarity of the resulting process. For the log-linear model (3) there are neither constraints on the intercept nor on the covariate coefficients and the total number of constraints is $d = 2(p+q+1)$. 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 (9) we employ 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 (8).

Note that the log-likelihood (7) and the score (8) are 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.1 (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 2 in Appendix A.1.

Solving the non-linear optimization problem (9) 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 autoregressive moving average (ARMA) model. A third possibility is to fit a naive i.i.d. model without covariates. As a last choice, note that we could use fixed values which need to be provided by the statistician. All possibilities are available in our package (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. Of course, a starting value closer to the global optimum usually requires fewer iterations until convergence. However, we have encountered some data examples where starting values close to a local optimum, obtained by one of the first two estimation methods, can even prevent finding 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.2.

4. Prediction

In terms of the mean square error, the optimal predictor $\widehat{Y}_{n+1}$ for Y_{n+1}, given potential covariates at time $n+1$ and the past $\mathcal{F}_n$ of the process up to time n, is the conditional expectation λ_{n+1} given in (1). By construction of the model the conditional distribution of Y_{n+1} is a Poisson (4) respectively Negative Binomial (5) distribution with mean λ_{n+1}.

An h-step-ahead prediction $\widehat{Y}_{n+h}$ for Y_{n+h} is obtained by recursive one-step-ahead predictions, where unobserved values $Y_{n+1}, \ldots, Y_{n+h-1}$ are replaced by their respective one-step-ahead prediction (S3 method of function `predict`), $h \in \mathbb{N}$. The distribution of this h-step-ahead prediction $\widehat{Y}_{n+h}$ is not known analytically but can be approximated numerically by simulation, 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 model parameters $\widehat{\boldsymbol{\theta}}$. The dispersion parameter ϕ of the Negative Binomial distribution is replaced by its estimator $\widehat{\phi}$. The additional uncertainty induced by plugging in the estimated model coefficients is not taken into account for the construction of prediction intervals.

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. The prediction intervals 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 $y_{n+h}^{(1)}, \ldots, y_{n+h}^{(B)}$ or find the shortest interval which contains at least $\lceil (1-\alpha) \cdot B \rceil$ of these observations (the function `predict` returns both types of prediction intervals and additionally the empirical median of the simulated predictive distribution). The computation of prediction intervals can be accelerated by distributing it to multiple cores simultaneously (argument `parallel=TRUE`), which requires a computing cluster registered by the R package **parallel**.

5. 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. Note that all tools are introduced as in-sample versions, meaning that the observations $y_1 \ldots, y_n$ 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.

Denote the fitted values 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

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

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

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

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

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

for $t = 1, \ldots, n$ (see for example Hilbe 2011, Section 5.1). The empirical autocorrelation function of these residuals can demonstrate 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 (2015) 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 $P_t(y) = P\left(Y_t \leqslant y|\mathcal{F}_{t-1}\right)$ the c.d.f., by $p_t(y) = P\left(Y_t = y|\mathcal{F}_{t-1}\right)$ the p.d.f., $y \in \mathbb{N}_0$, and by σ_t the standard deviation of the predictive distribution (recall Section 4 on one-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

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

The mean PIT is then given by

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

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.,

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

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 2015) (function `marcal`). If the predictions from a model are appropriate the marginal distribution of the predictions resembles the marginal distribution of the observations and its plotted difference is close to zero. Major deviations from zero point at 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. 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 1. The mean score for each corresponding model is given by $\sum_{t=1}^{n} s(P_t, y_t)/n$. The model with the lowest score is preferable. Each of the different proper scoring rules captures different characteristics of the predictive distribution and its distance to the observed data (function `scoring`).

Scoring rule	**Abbreviation**	**Definition**
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`	$\sum_{y=0}^{\infty}(P_t(y) - \mathbb{1}(y_t \leqslant y))^2$
Dawid-Sebastiani score	`dawseb`	$(y_t - \lambda_t)^2/\sigma_t^2 + 2\log(\sigma_t)$
normalized squared error score	`normsq`	$(y_t - \lambda_t)^2/\sigma_t^2$
squared error score	`sqerror`	$(y_t - \lambda_t)^2$

Table 1: Definitions of proper scoring rules $s(P_t, y_t)$ (cf. Czado *et al.* 2009; Christou and Fokianos 2015) and their abbreviations in the package; $\lVert p_t \rVert^2 = \sum_{y=0}^{\infty} p_t(y)^2$.

6. 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 $\delta^{t-\tau}\mathbb{1}(t \geqslant \tau)$, where τ is the time of occurrence and 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 hence its effect on the linear predictor is not purely additive. Our package includes methods to test on such intervention effects developed by Fokianos and Fried (2010, 2012), suitably adapted to the more general model class described in Section 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

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \, \widetilde{g}(Y_{t-i_k}) + \sum_{\ell=1}^{q} \alpha_\ell g(\lambda_{t-j_\ell}) + \boldsymbol{\eta}^\top \boldsymbol{X}_t + \sum_{m=1}^{s} \omega_m \delta_m^{t-\tau_m} \mathbb{1}(t \geqslant \tau_m), \tag{12}$$

where $\omega_1, \ldots, \omega_s$ are the respective intervention sizes. At the time of its occurrence an intervention of size ω_ℓ increases the level of the time series by adding the magnitude ω_ℓ for a linear model like (2) or by multiplying the factor $\exp(\omega_\ell)$ for a log-linear model like (3). In the following paragraphs we briefly outline the proposed intervention detection procedures and refer 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 (12) 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 $\ell \in \{1, \ldots, s\}$, 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 and for a sufficiently large sample size.

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

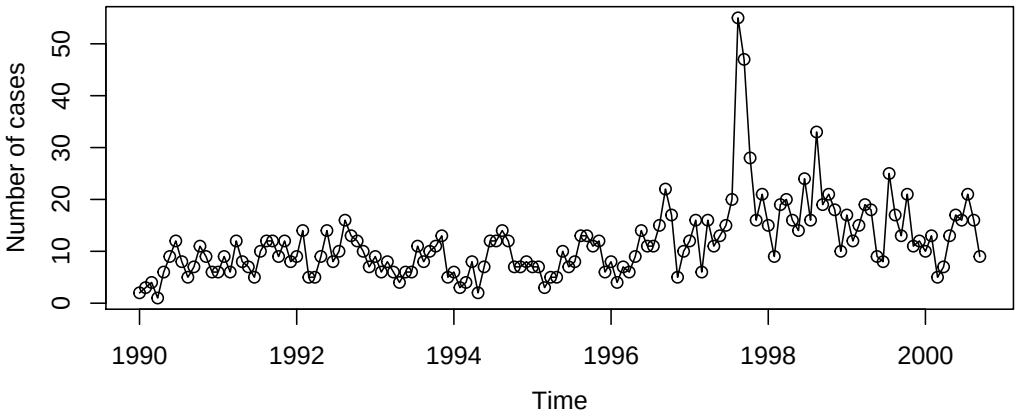


Figure 1: Number of campylobacterosis cases (reported every 28 days) in the north of Québec in Canada.

Bonferroni-corrected to account for the multiple testing. 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 a new step and continues until no further intervention effects are detected. Finally, model (12) with all detected intervention effects can be fitted to the data to estimate the intervention sizes and the other parameters jointly. Note that statistical inference for this final model fit has to be done with care. Further details are given in Fokianos and Fried (2010, 2012).

Liboschik *et al.* (2014) study a model for external intervention effects (modeled by external covariate effects, recall (6) and the related discussion) and compare it to internal intervention effects studied in the two aforementioned publications (argument `external`).

7. 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 `help(tsglm)`) is a good starting point to become familiar with the usage of the package. We demonstrate typical applications of the package by two data examples.

7.1. Campylobacter infections in Canada

We first analyze the number of campylobacterosis cases (reported every 28 days) in the north of Québec in Canada shown in Figure 1, which was first reported by Ferland *et al.* (2006). These data are made available in our package by the 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 (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 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`. 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, dist="poisson")
R> campyfit_nbin <- tsglm(campy, model=list(past_obs=1, past_mean=13),
+                    xreg=interventions, dist="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 in Figure 2 are 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")
```

The response residuals are identical for the two conditional distributions. Their empirical autocorrelation function, shown in Figure 2 top left, does not exhibit remaining serial correlation or seasonality which is not described by the models. The U-shape of the non-randomized PIT histogram in Figure 2 bottom left indicates that the Poisson distribution does not fully capture this dispersion well, although the U-shape is not very pronounced. 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 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
Poisson    2.749861 -0.07669775 -0.2751097 2.199937 3.661858
NegBin     2.721916 -0.07800114 -0.2765980 2.184902 3.605720
           normsq  sqerror
Poisson 1.3075592 16.50839
NegBin  0.9642855 16.50839
```

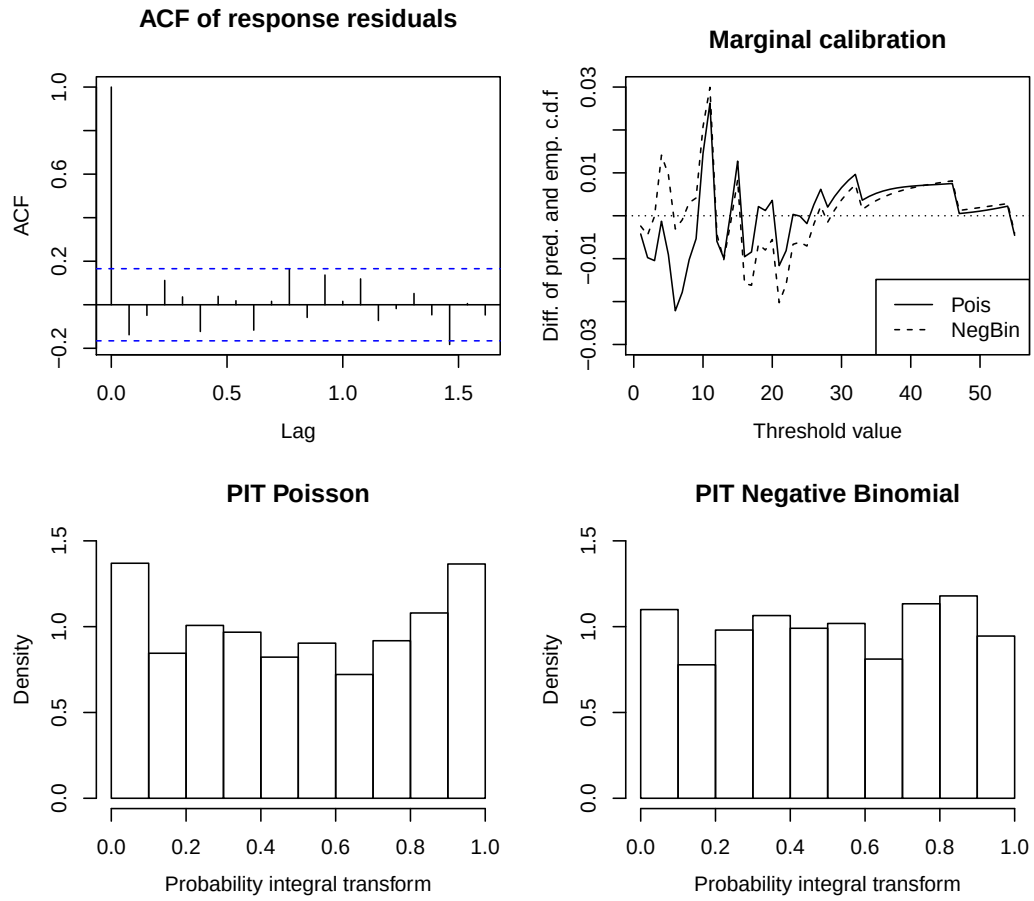


Figure 2: Diagnostic plots for a fit to the campylobacterosis data.

Two of the scoring rules are slightly in favor of the Poisson distribution (quadratic and spherical score), while the other scoring rules support the Negative Binomial distribution. Based on the the PIT histograms and the results obtained by most of the scoring rules, we decide for a Negative Binomial model. The degree of overdispersion seems to be small, as the estimated overdispersion coefficient `'sigmasq'` 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
(Intercept)   3.3169      0.7850
beta_1        0.3689      0.0696
alpha_13      0.2201      0.0942
interv_1      3.0864      0.8561
interv_2     41.8628     12.0693
```

```
sigmasq        0.0297          NA
Standard errors obtained by normal approximation.

Link function: identity
Distribution family: nbinom (with overdispersion coefficient 'sigmasq')
Number of coefficients: 6
Log-likelihood: -381.0682
AIC: 774.1364
BIC: 791.7862
```

The coefficient `beta_1` corresponds to regression on the previous observation, `alpha_13` corresponds to regression on values of the latent mean thirteen units back in time. The standard errors of the estimated regression parameters in the summary above are based on the normal approximation given in (11). For the additional overdispersion coefficient `sigmasq` of the Negative Binomial distribution there is no analytical approximation available for its standard error. Alternatively, standard errors 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
 0.89710834  0.07362390  0.10266237  0.89253577 11.71983693
    sigmasq
 0.01417654

Warning message:
In se.tsglm(campyfit_nbin, B = 500) :
  The overdispersion coefficient 'sigmasq' could not be estimated
in 13 of the 500 replications. It is set to zero for these
replications. This might to some extent result in an overestimation
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.

7.2. 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 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

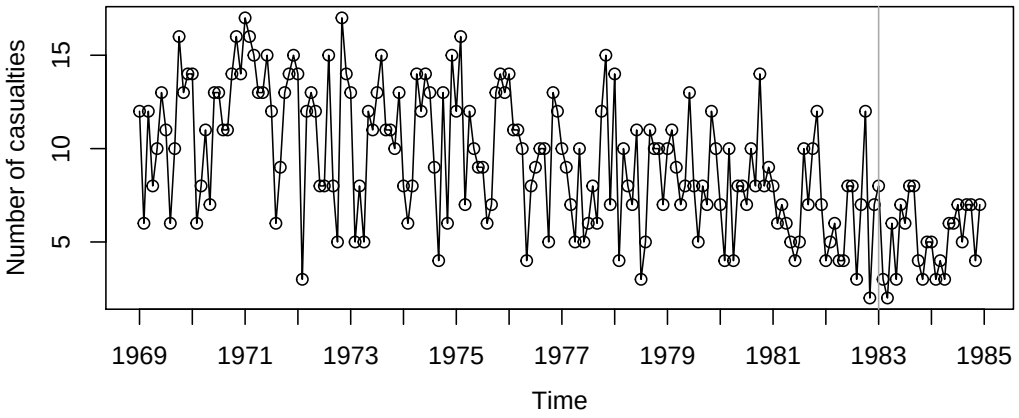


Figure 3: Monthly number of killed van drivers in Great Britain. The introduction of compulsory wearing of seatbelts on 31 January 1983 is marked by a vertical line.

large that they can be treated with methods for continuous-valued data in good approximation. 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.

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 try to capture the short range serial dependence by a first order autoregressive term and the yearly seasonality by a 12th order autoregressive term, both 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(ts=timeseries_until1981,
+    model=list(past_obs=c(1, 12)), link="log", distr="pois",
+    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
```

```
(Intercept)    1.8315     0.36817
beta_1         0.0862     0.08194
beta_12        0.1558     0.08782
PetrolPrice    0.7980     2.37008
linearTrend   -0.0307     0.00853
Standard errors obtained by parametric bootstrap with 500 replications.

Link function: log
Distribution family: poisson
Number of coefficients: 5
Log-likelihood: -396.152
AIC: 802.3039
BIC: 817.5532
```

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 3% 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. A graphical representation of the following predictions is given in Figure 4.

```
R> timeseries_1982 <- window(timeseries, start=1982, end=1982+11/12)
R> regressors_1982 <- window(regressors, start=1982, end=1982+11/12)
R> predict(seatbeltsfit, n.ahead=12, level=1-0.1/12, B=2000,
+          newxreg=regressors_1982)$fit

          Jan      Feb      Mar      Apr      May      Jun      Jul
1982 7.702261 7.413734 7.532874 7.382316 7.177786 6.967197 7.129638
          Aug      Sep      Oct      Nov      Dec
1982 7.844506 7.511275 7.840156 8.049270 7.456153
```

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 6 is applied:

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

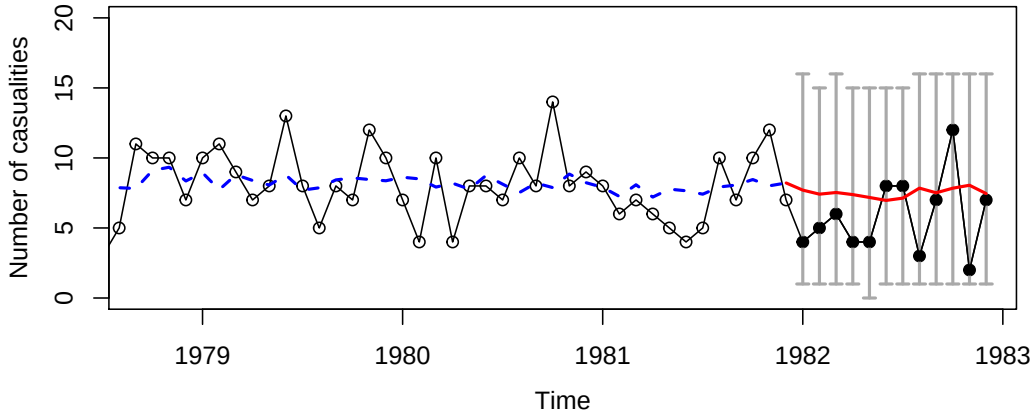


Figure 4: Fitted values (blue dashed line) and predicted values (red solid line) according to the model with the Poisson distribution. Prediction intervals (grey bars) are designed to ensure a global coverage rate of 90%. They are chosen to have minimal length and are based on a simulation with 2000 replications.

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

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

Chisq-Statistic: 46.80599 on 1 degree(s) of freedom, p-value: 7.837397e-12

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
    1.93358      0.08185      0.13918      0.41617     -0.03463
   interv_1
   -0.21696
```

The null hypothesis of no intervention is rejected at a 5% significance level. The multiplicative effect size of the intervention is found to be 0.805. This indicates that according to this model fit 19.5% 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.

8. Outlook

In its current version the R package `tscount` allows the analysis of count time series with a quite broad class of models. It will hopefully proof to be useful for a wide range of applications.

Nevertheless, there is a number of desirable extensions of the package which could be included in future releases. 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 distribution. 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.

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).

Alternative nonlinear models are for example the threshold model suggested by Douc *et al.* (2013) or the examples given 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 (2013) 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 (2013) develop M-estimators for the linear and the log-linear model respectively. Fried, Liboschik, Elsaied, Kitromilidou, and Fokianos (2014) compare robust estimators of the (partial) autocorrelation 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.

Acknowledgements

Part of this work was done while K. Fokianos was a Gambrinus Fellow at TU Dortmund University. The research of R. Fried and T. Liboschik was supported by the German Research Foundation (DFG, SFB 823 “Statistical modelling of nonlinear dynamic processes”). The authors thank Philipp Probst for his considerable contribution to the development of the package and Jonathan Rathjens for carefully checking the package.

References

Bollerslev T (1986). "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, **31**(3), 307–327. ISSN 0304-4076. doi:10.1016/0304-4076(86)90063-1.

Box GEP, Tiao GC (1975). "Intervention Analysis with Applications to Economic and Environmental Problems." *Journal of the American Statistical Association*, **70**(349), 70–79. ISSN 0162-1459. doi:10.2307/2285379.

Christou V, Fokianos K (2013). "Estimation and Testing Linearity for Mixed Poisson Autoregressions." Submitted for publication.

Christou V, Fokianos K (2014). "Quasi-Likelihood Inference for Negative Binomial Time Series Models." *Journal of Time Series Analysis*, **35**(1), 55–78. ISSN 1467-9892. doi:10.1111/jtsa.12050.

Christou V, Fokianos K (2015). "On Count Time Series Prediction." *Journal of Statistical Computation and Simulation*, **85**(2), 357–373. ISSN 0094-9655. doi:10.1080/00949655.2013.823612.

Czado C, Gneiting T, Held L (2009). "Predictive Model Assessment for Count Data." *Biometrics*, **65**(4), 1254–1261. ISSN 1541-0420. doi:10.1111/j.1541-0420.2009.01191.x.

Dawid AP (1984). "Statistical Theory: The Prequential Approach." *Journal of the Royal Statistical Society. Series A (General)*, **147**(2), 278–292. ISSN 0035-9238. doi:10.2307/2981683.

Demidenko E (2013). *Mixed models: theory and applications with R*. Wiley series in probability and statistics, 2nd edition. Wiley, Hoboken. ISBN 978-1-11-809157-9.

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. ISSN 0304-4149. doi:10.1016/j.spa.2013.04.010.

Elsaied H, Fried R (2014). "Robust Fitting of INARCH Models." *Journal of Time Series Analysis*, **35**(6), 517–535. ISSN 1467-9892. doi:10.1111/jtsa.12079.

Fahrmeir L, Tutz G (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York. ISBN 9780387951874.

Ferland R, Latour A, Oraichi D (2006). "Integer-Valued GARCH Process." *Journal of Time Series Analysis*, **27**(6), 923–942. ISSN 1467-9892. doi:10.1111/j.1467-9892.2006.00496.x.

Fokianos K (2011). "Some Recent Progress in Count Time Series." *Statistics: A Journal of Theoretical and Applied Statistics*, **45**(1), 49. ISSN 0233-1888. doi: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*, number 30 in Handbook of Statistics, pp. 315–347. Elsevier, Amsterdam. ISBN 978-0-444-53858-1.

Fokianos K (2015). "Statistical Analysis of Count Time Series Models." In R Davis, S Holan, R Lund, N Ravishanker (eds.), *Handbook of Discrete-Valued Time Series*, Handbooks of Modern Statistical Methods. Chapman & Hall, London. To appear.

Fokianos K, Fried R (2010). "Interventions in INGARCH Processes." *Journal of Time Series Analysis*, **31**(3), 210–225. ISSN 01439782. doi: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. ISSN 1471-082X, 1477-0342. doi:10.1177/1471082X1201200401.

Fokianos K, Neumann MH (2013). "A Goodness-Of-Fit Test for Poisson Count Processes." *Electronic Journal of Statistics*, **7**, 793–819. ISSN 1935-7524. doi:10.1214/13-EJS790.

Fokianos K, Rahbek A, Tjøstheim D (2009). "Poisson Autoregression." *Journal of the American Statistical Association*, **104**(488), 1430–1439. ISSN 0162-1459. doi:10.1198/jasa.2009.tm08270.

Fokianos K, Tjøstheim D (2011). "Log-Linear Poisson Autoregression." *Journal of Multivariate Analysis*, **102**(3), 563–578. ISSN 0047-259X. doi: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. ISSN 0020-3157, 1572-9052. doi:10.1007/s10463-012-0351-3.

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. ISSN 1026-597X. URL http://www.ajs.or.at/index.php/ajs/article/view/vol43-3-2.

Fuller WA (1996). *Introduction to Statistical Time Series*. 2nd edition. Wiley, New York. ISBN 0-471-55239-9.

Gneiting T, Balabdaoui F, Raftery AE (2007). "Probabilistic Forecasts, Calibration and Sharpness." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**(2), 243–268. ISSN 1467-9868. doi: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. Series A (General)*, **149**(3), 187–227. ISSN 0035-9238. doi:10.2307/2981553.

Heinen A (2003). "Modelling Time Series Count Data: An Autoregressive Conditional Poisson Model." *CORE discussion paper*, **62**. doi:10.2139/ssrn.1117187.

Hilbe JM (2011). *Negative Binomial Regression*. 2nd edition. Cambridge University Press, Cambridge. ISBN 9780521198158.

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. ISSN 1863-8171. doi:10.1007/s10182-010-0139-9.

Kedem B, Fokianos K (2002). *Regression Models for Time Series Analysis*. Wiley series in probability and statistics. Wiley-Interscience, Hoboken. ISBN 0-471-36355-3.

Kitromilidou S, Fokianos K (2013). "Robust Estimation Methods for a Class of Count Time Series Log-Linear Models." Under revision.

Lange K (1999). *Numerical Analysis for Statisticians*. Statistics and computing. Springer, New York. ISBN 0-387-94979-8.

Liboschik T, Kerschke P, Fokianos K, Fried R (2014). "Modelling Interventions in INGARCH Processes." *International Journal of Computer Mathematics*. ISSN 0020-7160. doi:10.1080/00207160.2014.949250. Published online.

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. ISSN 2044-8317. doi:10.1111/j.2044-8317.2011.02031.x.

Moysiadis T, Fokianos K (2014). "On Binary and Categorical Time Series Models with Feedback." *Journal of Multivariate Analysis*, **131**, 209–228. ISSN 0047-259X. doi:10.1016/j.jmva.2014.07.004.

Nelder JA, Wedderburn RWM (1972). "Generalized Linear Models." *Journal of the Royal Statistical Society. Series A (General)*, **135**(3), 370–384. ISSN 0035-9238. doi:10.2307/2344614.

Pedeli X, Karlis D (2013). "On Composite Likelihood Estimation of a Multivariate INAR(1) Model." *Journal of Time Series Analysis*, **34**(2), 206–220. ISSN 1467-9892. doi:10.1111/jtsa.12003.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.r-project.org.

Scotto MG, Weiß CH, Silva ME, Pereira I (2014). "Bivariate Binomial Autoregressive Models." *Journal of Multivariate Analysis*, **125**, 233–251. ISSN 0047-259X. doi:10.1016/j.jmva.2013.12.014.

Siakoulis V (2014). ***acp*** *– Autoregressive Conditional Poisson*. R package version 1.0. URL http://cran.r-project.org/package=acp.

Tjøstheim D (2012). "Some Recent Theory for Autoregressive Count Time Series." *TEST*, **21**(3), 413–438. ISSN 1133-0686. doi:10.1007/s11749-012-0296-0.

Ver Hoef JM, Boveng PL (2007). "Quasi-Poisson vs. Negative Binomial Regression: How Should We Model Overdispersed Count Data?" *Ecology*, **88**(11), 2766–2772. ISSN 0012-9658. doi:10.1890/07-0043.1.

Weiß CH (2008). "Thinning Operations for Modeling Time Series of Counts – a Survey." *Advances in Statistical Analysis*, **92**(3), 319–341. ISSN 1863-8171. doi:10.1007/s10182-008-0072-3.

Woodard DB, Matteson DS, Henderson SG (2011). "Stationarity of generalized autoregressive moving average models." *Electronic Journal of Statistics*, **5**, 800–828. ISSN 1935-7524. doi:10.1214/11-EJS627.

Zhu F (2012). "Zero-Inflated Poisson and Negative Binomial Integer-Valued GARCH Models." *Journal of Statistical Planning and Inference*, **142**(4), 826–839. ISSN 0378-3758. doi:10.1016/j.jspi.2011.10.002.

A. Implementation details

A.1. Recursions for inference and their initialization

Let $h : \mathbb{R}^+ \to \mathbb{R}$ 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 latent 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 of 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 (2) it holds (Ferland *et al.* 2006)

$$\mathsf{E}(Y_t) = \mathsf{E}(\nu_t) = \frac{\beta_0}{1 - \sum_{k=1}^{p} \beta_k - \sum_{\ell=1}^{q} \alpha_\ell} =: \mu(\boldsymbol{\theta}). \tag{13}$$

For the log-linear model (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 (2). It approximately holds $\mathsf{E}(Z_t) = \mathsf{E}(\nu_t) = \mu(\boldsymbol{\theta})$. 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

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

		$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\alpha}_1$	$\ell(\widehat{\boldsymbol{\theta}})$
`init.method="marginal",`	`init.drop=FALSE`	0.507	0.738	0.244	-3024.5
`init.method="marginal",`	`init.drop=TRUE`	0.567	0.746	0.237	-2568.5
`init.method="iid",`	`init.drop=FALSE`	0.750	0.750	0.229	-3036.2
`init.method="iid",`	`init.drop=TRUE`	0.562	0.738	0.246	-2587.5
`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 2: Estimated parameters and log-likelihood of a time series of length 1000 simulated from model (2) for different initialization strategies. The true parameters are $\beta_0 = 0.5$, $\beta_1 = 0.77$ and $\alpha_1 = 0.22$.

which is explicitly given by

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

Another possibility is to initialize ν_t by β_0 and $\partial \nu_t(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ by zero, which corresponds to the marginal expectations assuming a model without covariate effects and without serial dependence (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.

The different methods for initialization can affect the estimation substantially even for quite long time series with 1000 observations, particularly in the presence of strong serial dependence. We illustrate this by the simulated example presented in Table 2.

A.2. Starting value for optimization

The numerical optimization of the log-likelihood function requires a starting value for the parameter vector $\boldsymbol{\theta}$. This starting value can be obtained by initial estimation based on a simpler model than the one of interest. 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

$$Y_t|\mathcal{F}_{t-1}^* \sim \text{Poi}(\lambda_t^*), \text{ with } \nu_t^* = g(\lambda_t^*) \text{ and}$$
$$\nu_t^* = \beta_0^* + \beta_1^* \tilde{g}(Y_{t-i_1}) + \ldots + \beta_p^* \tilde{g}(Y_{t-i_p}) + \eta_1^* X_{t,1} + \ldots + \eta_r^* X_{t,r}, \; t = i_p + 1, \ldots, n,$$

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 (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 $k := \max(P, Q)$ and the general model from Section 2 this representation, after straightforward calculations, is given by

$$(\tilde{g}(Y_t) - \underbrace{\mu(\boldsymbol{\theta})}_{=:\zeta}) - \sum_{i=1}^{k} \underbrace{(\beta_i + \alpha_i)}_{=:\varphi_i} (\tilde{g}(Y_{t-i}) - \mu(\boldsymbol{\theta})) = \varepsilon_t + \sum_{i=1}^{q} \underbrace{(-\alpha_i)}_{=:\psi_i} \varepsilon_{t-i}, \tag{14}$$

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 $\tilde{g}$ is defined by $\tilde{g}(x) = x$ for the linear model and $\tilde{g}(x) = \log(x + 1)$ 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 our original process are obtained by $\alpha_i = -\psi_i$ and $\beta_i = \varphi_i + \psi_i$. We get β_0 from $\beta_0 = \zeta\big(1 - \sum_{i=1}^{p} \beta_i - \sum_{j=1}^{q} \alpha_j\big)$ 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 can be done 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 $\tilde{g}(Y_t)$, whose errors follow an ARMA model like (14). Consequently, the covariate effects do not enter the dynamics of the process, as it is the case in the actual model (1). It would be preferable to fit an ARMAX model, in which covariate effects are included on the right hand side of (14), 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 discovered that in some situations (in the presence of certain types of covariates) both approaches occasionally provoke the algorithm for likelihood optimization to run into a local but not the global optimum. This happens even 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 needed for the optimization algorithm to be terminated. Nevertheless, we prefer the longer overall computation time to the risk of an improper final estimation and make this the default method in our package.

Particularly for the linear model, neither of the aforementioned approaches supplies 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 is ensured to lay in the interior of the parameter space Θ, as it is required for the applied optimization algorithm. To overcome this problem $\widehat{\boldsymbol{\theta}}^*$ is suitably transformed to be used as a starting value. For the linear model (2) this transformation is done according to the procedure described by Liboschik *et al.* (2014) and for the log-linear model (3) this procedure is modified appropriately.

A.3. Stable inversion of the information matrix

In order to obtain standard errors from the normal approximation (11) 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.

B. 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. 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

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

The dependence 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 3, 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 values of about 0.5, such that their effect 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.

Abbreviation	Definition
Linear	t/n
Sine	$(\sin(2\pi \cdot 5 \cdot t/n) + 1)/2$
Spiky outlier	$\mathbb{1}(t = \tau)$
Transient shift	$0.8^{t-\tau}\mathbb{1}(t \geqslant \tau)$
Level shift	$\mathbb{1}(t \geqslant \tau)$
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 3: 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).

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 5.

A second finding is that the effect of covariates like a transient shift or a spiky outlier is hard to estimate precisely. Note that both covariates have values considerably 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 plot in Figures 6 and 7 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 (11) 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 8. The only serious deviation from normality happens for the spiky outlier in the linear model, where many estimates of the covariate coefficient η_1 lie close to zero, which 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 for the spiky outlier the conditions for asymptotic normality in linear regression models stated in Section 3 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 (10) yields good results. We consider both, the linear model with the identity link

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

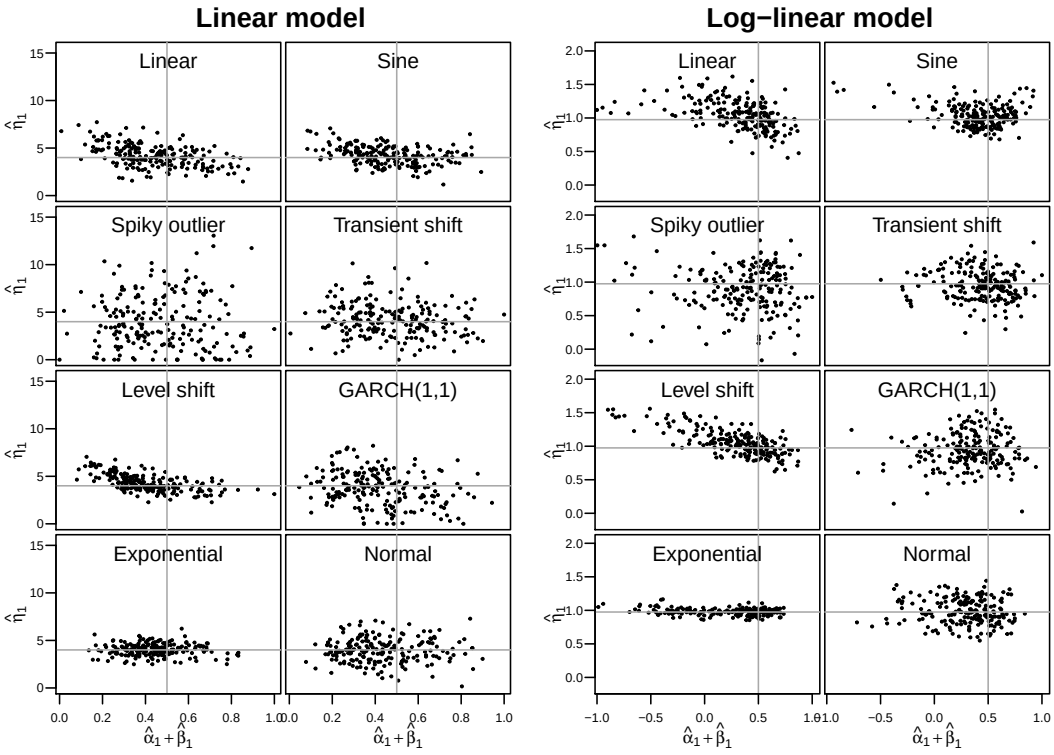


Figure 5: Scatterplots of the estimated covariate parameter against the sum of the estimated dependence parameters in a linear (left) respectively log-linear (right) model of order $p = q = 1$ with an additional covariate of the given type. The time series of length 100 are simulated from the respective model with the true values marked by grey lines. Each dot represents one of 200 replications.

and the log-linear model with the logarithmic link

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

The parameters β_0, β_1 and α_1 are chosen like in the previous section. 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 4. In certain cases it is numerically not possible to solve (10) 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 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 9 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.

	Mean	**Median**	**Std.dev.**	**MAD**	**Failures (in %)**
$\sigma^2 = 1.00$	0.99	0.97	0.18	0.16	0.00
0.20	0.20	0.20	0.05	0.05	0.00
0.10	0.10	0.10	0.04	0.03	0.00
0.05	0.05	0.05	0.03	0.03	3.10
0.00	0.02	0.02	0.02	0.02	51.40

Table 4: Summary statistics for the estimated overdispersion coefficient $\hat{\sigma}^2$ 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.

The boxplots also confirm our above finding that the estimator has a clearly asymmetric distribution for sample sizes up to several hundred.

Affiliation:

Tobias Liboschik
Department of Statistics
TU Dortmund University
44221 Dortmund, Germany
E-mail: liboschik@statistik.tu-dortmund.de
URL: http://www.statistik.tu-dortmund.de/liboschik-en.html

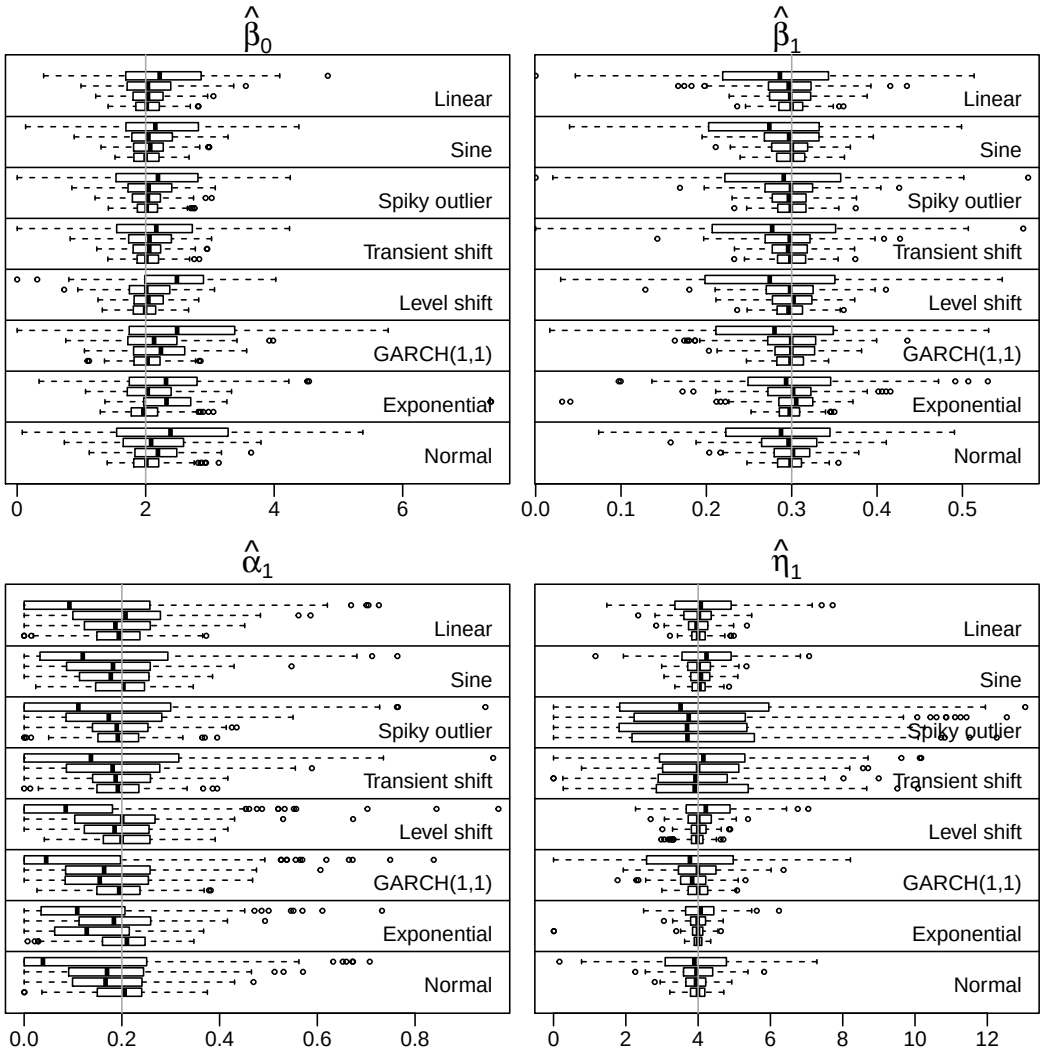


Figure 6: Estimated coefficients for a linear model of order $p = q = 1$ with an additional covariate of the given type. The time series of length $100, 500, 1000, 2000$ (from top top to bottom in each panel) are simulated from the respective model with the true coefficients marked by a grey vertical line. Each boxplot is based on 200 replications.

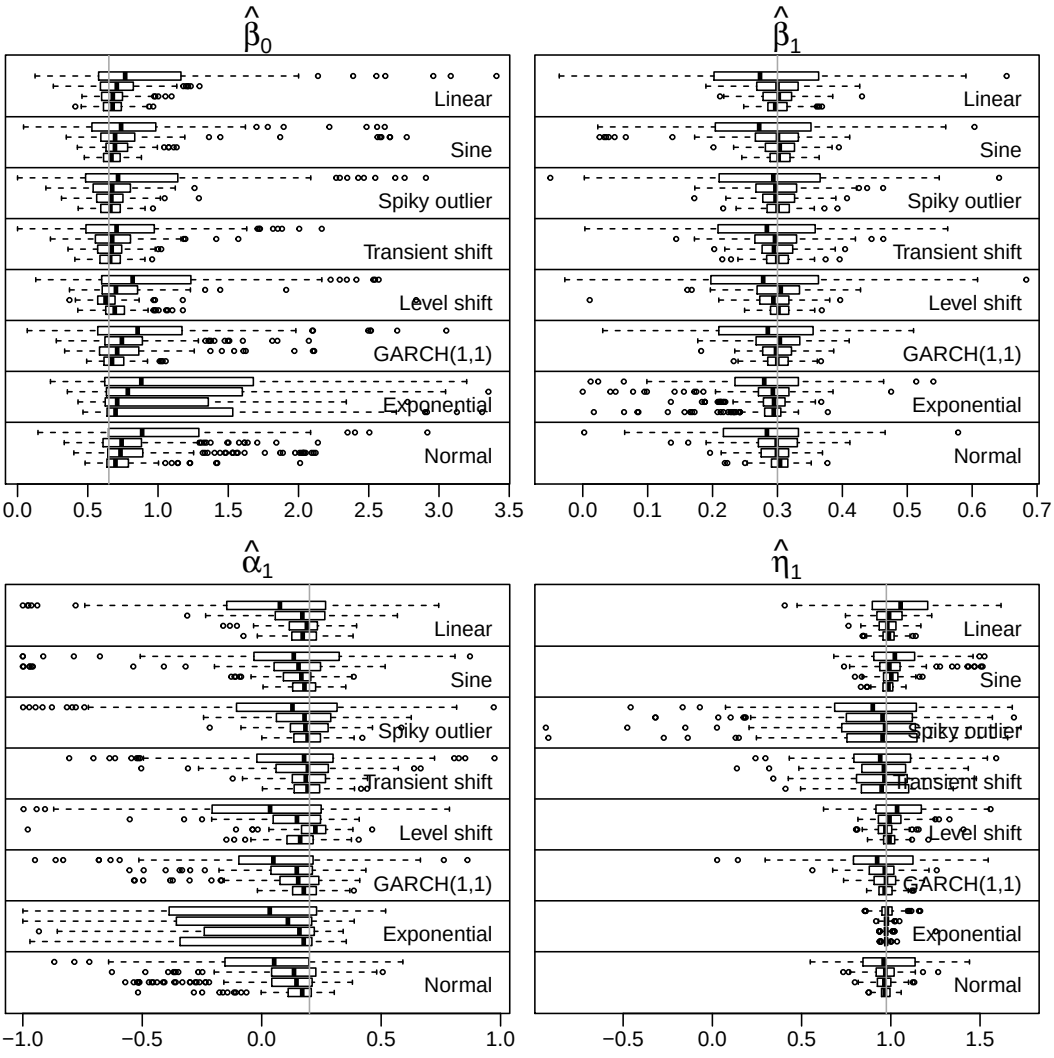


Figure 7: Simulation results equivalent to those shown in Figure 6 but for a log-linear model.

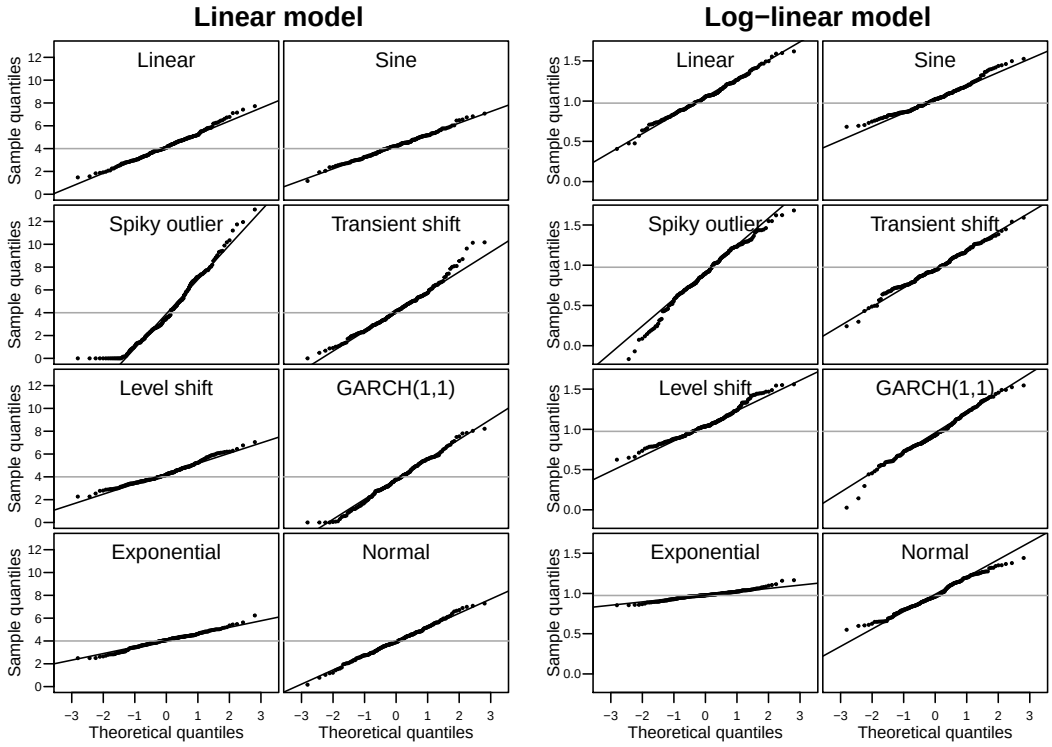


Figure 8: Normal QQ-plots for the estimated covariate coefficient $\widehat{\eta}_1$ in a linear (left) respectively log-linear (right) model of order $p = q = 1$ with an additional covariate of the given type. The time series of length 100 are simulated from the respective model with the true coefficient marked by a grey horizontal line. Each plot is based on 200 replications.

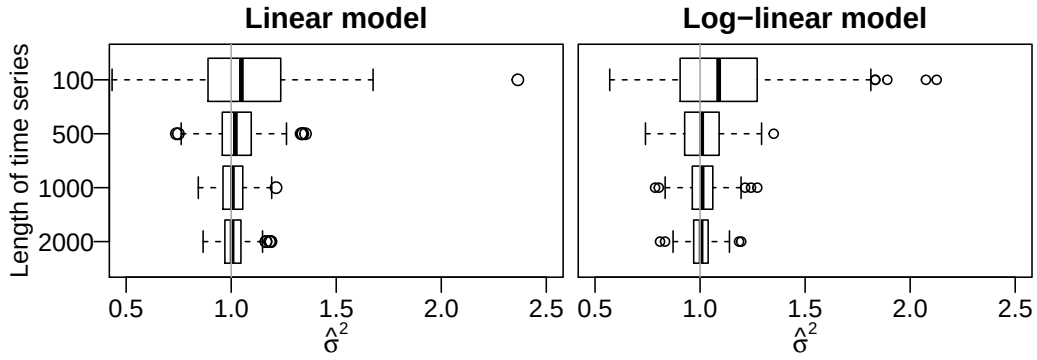


Figure 9: Estimated overdispersion coefficient $\widehat{\sigma}^2$ of the Negative Binomial distribution for a linear (left) respectively log-linear (right) model of order $p = q = 1$. The time series are simulated from the respective model with the true overdispersion coefficient marked by a grey vertical line. Each boxplot is based on 200 replications.