Generalized estimating equations for longitudinal generalized Poisson count data with regression effects on the mean and dispersion level

Vinzenz Erhardt[*]   Claudia Czado

Technische Universität München
Lehrstuhl für Mathematische Statistik
Boltzmannstr. 3
D-85747 Garching, Germany
E-mail: erhardt@ma.tum.de, cczado@ma.tum.de
Tel: +49 89 289-17434, Fax: +49 89 289-17435

Abstract

Generalized estimating equations (GEE) fit parameters based on sums of weighted residuals, which may be applied for example to the Poisson distribution. We discuss Generalized Poisson (GP) response data. This distribution has a more flexible variance function than the Poisson distribution and has an additional dispersion parameter. To fit this parameter, second level estimating equations based on covariance residuals are necessary. This requires knowledge of variances of empirical covariances, which for most discrete distributions except the binary cannot be derived from first level GEE. We approximate them by a novel approach. We allow for regression on mean and overdispersion parameters. In an application we deal with the outsourcing of patent filing processes. Exploratory data analysis tools developed earlier by the authors are utilized to choose regression for the dispersion parameters. For the given data, our approach will outperform longitudinal Poisson regression and GP setups with constant dispersion.

Keywords: generalized estimating equations; generalized Poisson regression; longitudinal count data; make-or-buy decision; overdispersion

[*] corresponding author

1 Introduction

This paper considers longitudinal setups for generalized Poisson (GP) data using GEE. The GP distribution has first been introduced by Consul & Jain (1970). GP regression models were discussed by Consul & Famoye (1992) and Famoye (1993). Famoye et al. (2004) apply generalized Poisson regression to accident data, whereas Famoye & Singh (2003) develop a zero-inflated generalized Poisson regression model. A multivariate generalization of the generalized Poisson distribution capable of modeling exchangeable covariance structures has been developed by Vernic (2000) and is applied to insurance. Statistical inference regarding the generalized Poisson distribution is done by Tripathi & Gupta (1984). A Bayesian analysis is carried out by Scollnik (1995) and Gschlößl & Czado (2008). The interest in the class of GP models is driven by the fact that it can handle overdispersion, which count data very often exhibits. Here we allow for regression effects not only on the mean but on the overdispersion parameter as well. This allows to model overdispersion by individual characteristics (e.g. by a company's industry) and improve model fit when constant dispersion is insufficient. The GP distribution is a hyper model of the Poisson distribution which allows for nested model comparison if the mean specifications are hierarchical. Its variance function can be written as a product of the mean and an independent dispersion parameter, which allows for simple second moment regression specifications. In contrast to the GP distribution the variance of the negative binomial distribution is a product of the mean and a dispersion factor, which depends on both the mean and a dispersion parameter.

GEE have been introduced by Liang & Zeger (1986). Second level GEE (Prentice & Zhao (1991)) allow to determine variance parameters as well. Yan & Fine (2004) consider generalized estimating equations for the Poisson distribution. An implementation can be found in the *R* package 'geepack' (see Yan (2002)). For the conditional fixed-effects negative binomial distribution, generalized estimating equations are implemented in Stata (StataCorp (2007)). Hilbe (2007, Section 10.4) emphasizes that in this setup the dispersion parameter is not estimated as a separate parameter, it is apportioned across panels.

A comparison of three models starting with the regular Poisson GEE extended by dispersion designs will be carried out in this paper. Since these models might be nonnested, partial deviance, likelihood ratio tests or AIC are not applicable. Instead we use the 'quasilikelihood under independence criterion' (QIC) introduced by Pan (2001) for variable selection and the Wald-Wolfowitz run test (Chang (2000)) for assessing the goodness-of-fit.

The usefulness of our extensions will be demonstrated in an application to make-or-buy decision drivers in the field of patent filing processes. This data has already been examined by

Wagner (2006b), who used negative binomial panel regression to fit the data. Wagner (2006a) applies Transaction Cost Economics and a resource based view on make-or-buy decisions of patent related services. Czado et al. (2007) apply zero-inflated generalized Poisson (ZIGP) regression to this data and present tools for an exploratory data analysis to select covariates on the dispersion level, which will also be used in this paper. While in the ZIGP paper the observation year was conditioned on by considering it as a covariate, this temporal dependency will actually be quantified in this paper.

The paper is innovative with regard to the following aspects: first of all, despite its advantages over the negative binomial distribution, the GP distribution has not been considered in the context of GEE. Thereby we suggest an approach to approximate higher mixed moments for second level estimating equations. Secondly the GP distribution allows to let the dispersion parameter vary with covariates thus to identify covariate combinations where one finds large and small overdispersion effects. The dispersion coefficients will be estimated using second-level estimating equations. Thirdly, a closer look at the patent data including a quantification of the time dependency will be taken.

The paper is organized as follows: Section 2 gives a short review of the GP distribution. Section 3 introduces our GP regression setup. In Section 4, we show how the GEE approach by Liang & Zeger (1986) and the extensions by Prentice & Zhao (1991) can be applied to estimate parameters in our setup. A simulation study investigating small sample properties will be given in Section 5 showing a satisfactory behaviour for medium sample sizes. Subsection 6.1 reviews the variable selection criterion for panel data by Pan (2001) while in Subsection 6.2 an overview of our extensions to GEE techniques applied to longitudinal Poisson data is given and the goodness-of-fit will be compared for the different setups. Section 7 applies our findings to patent outsourcing data and interprets the results of our 'best' model. We conclude with a summary and discussion section.

2 The Generalized Poisson distribution

The generalized Poisson distribution $GP(\mu, \varphi)$ was first introduced by Consul & Jain (1970) and subsequently studied in detail by Consul (1989). Here we will utilize the mean parametrization (see e.g. Consul & Famoye (1992)). For $Y \sim GP(\mu, \varphi)$ we have $\text{Var}(Y) > (=, <)\ E(Y) \Leftrightarrow \varphi > (=, <)\ 1$. This allows for modeling over-, equi- and underdispersion. Its probability mass function (pmf) is given by

$$P(Y = y|\ \mu, \varphi) \ = \ \begin{cases} \frac{\mu(\mu+(\varphi-1)y)^{y-1}}{y!}\varphi^{-y}e^{-\frac{1}{\varphi}(\mu+(\varphi-1)y)} & y = 0, 1, 2... \\ 0 & y > m, \text{ if } \varphi < 1 \end{cases} \tag{2.1}$$

where $\varphi > \max(\frac{1}{2}, 1 - \frac{\mu}{m})$ and m is the largest natural number with $\mu + m(\varphi - 1) > 0$, if $\varphi < 1$. Hence, in the case of underdispersion ($\varphi < 1$), the support of the distribution depends on μ and φ, which is difficult to enforce when μ and φ need to be estimated. In the regression context this fact implies that the support of a link function for φ depends on μ. Therefore, we restrict to the case of overdispersion. Mean and variance of $Y \sim GP(\mu, \varphi)$ are given by $E(Y|\mu, \varphi) = \mu$ and $\text{Var}(Y|\mu, \varphi) = \mu\varphi^2$. The GP distribution does not belong to the exponential family even if the dispersion parameter φ is known.

3 A GEE setup for longitudinal count data

Assume we have longitudinal responses Y_{it} for $t = 1, \ldots, T$ time points and $i = 1, \ldots, K$ subjects, which we arrange as follows:

$$\begin{array}{ccc|c} Y_{11} & \ldots & Y_{1T} & \boldsymbol{Y}_{1\sim} \in \mathbb{N}_0^T \\ \vdots & & \vdots & \vdots \quad \text{(independent random vectors)} \\ Y_{K1} & \ldots & Y_{KT} & \boldsymbol{Y}_{K\sim} \in \mathbb{N}_0^T \\ \hline \boldsymbol{Y}_{\sim 1} \in \mathbb{N}_0^K & \ldots & \boldsymbol{Y}_{\sim T} \in \mathbb{N}_0^K & \end{array} .$$

Here $\boldsymbol{Y}_i := \boldsymbol{Y}_{i\sim} = (Y_{i1}, \ldots, Y_{iT})'$ summarizes the T dimensional vector of dependent variables for subject i. Observations from different subjects are assumed to be independent. Similarly $\boldsymbol{Y}_{\sim t} := (Y_{1t}, \ldots, Y_{Kt})' \in \mathbb{N}_0^K$ collects the i.i.d. marginal data at time point t. Moreover let $\boldsymbol{\mu}_i(\boldsymbol{\beta}) := E(\boldsymbol{Y}_i \mid \boldsymbol{\beta}) \in \mathbb{R}^T$ denote the vector of means of subject i. Variances are given by $\sigma_{it}^2(\boldsymbol{\delta}) := \text{Var}(Y_{it}|\boldsymbol{\delta})$ with $\boldsymbol{\delta}$ being a vector summarizing all parameters which influence the variance. Correlations are modeled by a 'working correlation matrix' $\boldsymbol{R}_1(\lambda_1) = (\rho_{tt^*}(\lambda_1)) \in [-1, 1]^{T\times T}$ for $\boldsymbol{Y}_i$, which will be equal for all subjects. Without loss of generality, assume a scalar $\lambda_1 \in [-1, 1]$, which allows for the most common correlation structures used in the literature. We investigate two specifications for $\boldsymbol{R}_1(\lambda_1)$, i.e.

- exchangeable: $\rho_{tt^*}(\lambda_1) = \lambda_1$ and $\rho_{tt}(\lambda_1) = 1$, $\lambda_1 \in (-1, 1)$,
- first-order autoregressive $AR(1)$: $\rho_{tt^*}(\lambda_1) = \lambda_1^{|t-t^*|}$ and $\rho_{tt}(\lambda_1) = 1$, $\lambda_1 \in (-1, 1)$.

Collecting all observations in a vector $\boldsymbol{Y} := (\boldsymbol{Y}_1', \dots, \boldsymbol{Y}_K')'$ the correlation matrix of $\boldsymbol{Y}$ is

$$\operatorname{Corr}(\boldsymbol{Y}) = \begin{pmatrix} \boldsymbol{R}_1(\lambda_1) & \boldsymbol{0}_{T\times T} & \dots & \boldsymbol{0}_{T\times T} \\ \boldsymbol{0}_{T\times T} & \boldsymbol{R}_1(\lambda_1) & & \boldsymbol{0}_{T\times T} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{0}_{T\times T} & \boldsymbol{0}_{T\times T} & \dots & \boldsymbol{R}_1(\lambda_1) \end{pmatrix} \in \mathbb{R}^{KT\times KT}. \tag{3.1}$$

The advantage of using a GEE approach is that one does not need to specify the joint distribution of $\boldsymbol{Y} \in \mathbb{R}^{K\times T}$ but it is enough to specify the first two moments of the distribution. For $t = 1, \dots, T$ we assume the following marginal specification for $\boldsymbol{Y}_{\sim t} \sim GP(\boldsymbol{\mu}_{\sim t}, \boldsymbol{\varphi}_{\sim t})$ where $\boldsymbol{\mu}_{\sim t} := (\mu_{1t}, \dots, \mu_{Kt})'$ and $\boldsymbol{\varphi}_{\sim t} := (\varphi_{1t}, \dots, \varphi_{Kt})'$, i.e. we have $E(\boldsymbol{Y}_{\sim t}) = \boldsymbol{\mu}_{\sim t}$ and

$$\operatorname{Var}(\boldsymbol{Y}_{\sim t}) = \begin{pmatrix} \mu_{1t}\varphi_{1t}^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mu_{Kt}\varphi_{Kt}^2 \end{pmatrix}. \tag{3.2}$$

Since for some data a constant overdispersion parameter might be too restrictive, we allow for regression on both mean and overdispersion parameters. Thereby, we use (known) explanatory variables $\boldsymbol{x}_{it} = (1, x_{it1}, \dots, x_{itp})'$ for the mean and $\boldsymbol{w}_{it} = (1, w_{it1}, \dots, w_{itq})'$ for overdispersion, $i = 1, \dots, K$, $t = 1, \dots, T$.

Another possibility for specifying the influence of regressors on the distribution's heterogeneity would be to regress on the variances directly. However, this would imply that we would have to set $\varphi_{it} := \sqrt{\frac{\operatorname{Var}(Y_{it})}{E(Y_{it})}}$ which might fall below 1 for some observations. According to the definition of the underdispersed GP distribution in Section 2, in this case $\varphi_{it} > \max(\frac{1}{2}, 1 - \frac{\mu_{it}}{m_{it}})$ needs to be fulfilled and the cumulative sum of probabilities needs not be 1 (see Consul & Jain (1970, p. 4)). Therefore we prefer to regress on the overdispersion parameter itself. In order to specify appropriate regression models for the overdispersion parameter, we utilize tools for an exploratory data analysis suggested by Czado et al. (2007, Section 5), which will be illustrated in Section 7.1. Finally we allow individual (known) exposure variables $E_{it} > 0$. The complete specification is given by:

1. *Random components:*

   Let $Y_{it} \sim GP(\mu_{it}, \varphi_{it})$, where $\{Y_{it},\ 1 \le i \le K, 1 \le t \le T\}$ are independent over all i and dependent with correlation matrix $\boldsymbol{R}_1(\lambda_1)$ for $t = 1, \dots, T$.

2. *Systematic components:*

   Two linear predictors $\eta^{\mu}_{it}(\boldsymbol{\beta}) = \boldsymbol{x}'_{it}\boldsymbol{\beta}$ and $\eta^{\varphi}_{it}(\boldsymbol{\alpha}) = \boldsymbol{w}'_{it}\boldsymbol{\alpha}$, $i = 1, \ldots, K$, $t = 1, \ldots, T$ influence the response Y_{it}. Here, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_q)'$ are unknown regression parameters. The matrices $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT})'$ and $\boldsymbol{W}_i = (\boldsymbol{w}_{i1}, \ldots, \boldsymbol{w}_{iT})'$ are the corresponding design matrices.

3. *Parametric link components:*

   The linear predictors $\eta^{\mu}_{it}(\boldsymbol{\beta})$ and $\eta^{\varphi}_{it}(\boldsymbol{\alpha})$ are related to $\mu_{it}(\boldsymbol{\beta})$ and $\varphi_{it}(\boldsymbol{\alpha})$, $i = 1, \ldots, K$, $t = 1, \ldots, T$ as follows:

$$\begin{aligned} &\text{(i) } \textit{Mean level} && \\ E(Y_{it} \mid \boldsymbol{\beta}) = \mu_{it}(\boldsymbol{\beta}) &:= E_{it} e^{\boldsymbol{x}'_{it}\boldsymbol{\beta}} = e^{\boldsymbol{x}'_{it}\boldsymbol{\beta} + log(E_{it})} > 0 && \\ \Leftrightarrow \eta^{\mu}_{it}(\boldsymbol{\beta}) &= \log(\mu_{it}(\boldsymbol{\beta})) - \log(E_{it}) \text{ (log link)}, && (3.3) \\ &\text{(ii) } \textit{Overdispersion level} && \\ \varphi_{it}(\boldsymbol{\alpha}) &:= 1 + e^{\boldsymbol{w}'_{it}\boldsymbol{\alpha}} > 1 && \\ \Leftrightarrow \eta^{\varphi}_{it}(\boldsymbol{\alpha}) &= \log(\varphi_{it}(\boldsymbol{\alpha}) - 1)) \text{ (shifted log link)}. && (3.4) \end{aligned}$$

This setup for longitudinal count regression data $\{Y_{it}, i = 1, \ldots, K; t = 1, \ldots, T\}$ we denote by $GP(\mu_{it}, \varphi_{it}, \boldsymbol{R}_1(\lambda_1))$. To be precise this is not a complete statistical formulation, since only the margins and the covariance structure are specified. This however is sufficient for estimation using a GEE approach. The following abbreviations will be used:

$$\begin{aligned} \rho_{tt^*}\big(\lambda_1(\gamma)\big) &:= [\boldsymbol{R}_1(\lambda_1(\gamma))]_{tt^*} = \text{Corr}(Y_{it}, Y_{it^*}),\ t \neq t^*, \\ \lambda_1(\gamma) &:= \tfrac{e^{2\gamma}-1}{e^{2\gamma}+1} = \tanh(\gamma) \in (-1, 1),\ \gamma \in \mathbb{R}, \text{ where} \\ & \quad\ \lambda_1(\gamma) \text{ is the parameter of the working correlation matrix,} \\ \boldsymbol{\delta} &:= (\boldsymbol{\beta}', \boldsymbol{\alpha}', \gamma)' \in \mathbb{R}^{p+q+3},\ \boldsymbol{\beta} \in \mathbb{R}^{p+1},\ \boldsymbol{\alpha} \in \mathbb{R}^{q+1}, \\ E_{it} &:= \text{known exposure of observation } i \text{ at time } t, \\ \mu_{it}(\boldsymbol{\beta}) &:= e^{\boldsymbol{x}'_{it}\boldsymbol{\beta} + log(E_{it})}, \\ \varphi_{it}(\boldsymbol{\alpha}) &:= 1 + e^{\boldsymbol{w}'_{it}\boldsymbol{\alpha}} = 1 + b_{it}(\boldsymbol{\alpha}),\ \ b_{it}(\boldsymbol{\alpha}) := e^{\boldsymbol{w}'_{it}\boldsymbol{\alpha}} \end{aligned}$$

The Fisher Z-transformation $\lambda_1(\gamma) := \tanh(\gamma)$ (Fisher (1921)) will be used to allow for unconstrained optimization over γ (instead of constrained optimization over λ_1 on $(-1, 1)$). Also, this will allow to estimate the variance of $\hat{\gamma}$ along with the variances of the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$. Since $\lambda_1(0) = 0$, using this transformation for testing $H_0: \ \gamma = 0$ versus $H_1: \ \gamma \neq 0$

will correspond to testing $H_0: \ \lambda_1 = 0$ versus $H_1: \ \lambda_1 \neq 0$.

4 A GEE approach for $GP(\mu_{it}, \varphi_{it}, \boldsymbol{R}_1(\lambda_1))$

Generalized estimating equations have first been introduced by Liang & Zeger (1986) and will be denoted by GEE1. Since GEE1 are based on weighted residuals, only parameters influencing the means (i.e. $\boldsymbol{\beta}$) can be estimated. In the GEE1 context, the correlation has to be estimated separately using for instance estimators based on residuals. Prentice & Zhao (1991) extend generalized estimating equations (GEE2). These extensions allow to estimate the correlation parameter γ simultaneously with $\boldsymbol{\beta}$. The additional variance parameters $\boldsymbol{\alpha}$ are estimated by a second set of estimating equation based on covariance residuals. For $E(Y_{it}) = \mu_{it}(\boldsymbol{\beta})$ and $\mathrm{Var}(Y_{it}) = \sigma^2_{it}(\boldsymbol{\delta})$, a working covariance matrix for $\boldsymbol{Y}_i$ can be constructed by

$$\boldsymbol{V}_{i1}(\boldsymbol{\delta}) := \boldsymbol{A}_i^{1/2}(\boldsymbol{\delta})\boldsymbol{R}_1(\lambda_1(\gamma))\boldsymbol{A}_i^{1/2}(\boldsymbol{\delta}) \ \in \mathbb{R}^{T\times T}, \tag{4.1}$$

where $\boldsymbol{A}_i(\boldsymbol{\delta}) := \mathrm{diag}\{\sigma^2_{i1}(\boldsymbol{\delta}), \ldots, \sigma^2_{iT}(\boldsymbol{\delta})\}$. Covariances will be denoted by $\sigma^2_{itt^*}(\boldsymbol{\delta}) := \mathrm{Cov}(Y_{it}, Y_{it^*})$ and $\boldsymbol{\sigma}^2_i(\boldsymbol{\delta}) := \big(\sigma^2_{itt^*}(\boldsymbol{\delta}); t \leq t^*; t, t^* = 1, \ldots, T\big)' \in \mathbb{R}^{T(T+1)/2}$ will be the vector of covariances of subject i. Further, let $\boldsymbol{S}_i(\boldsymbol{\beta}) = \big(S_{itt^*}(\boldsymbol{\beta}); t \leq t^*; t, t^* = 1, \ldots, T\big)' \in \mathbb{R}^{T(T+1)/2}$ be empirical covariances with entries $S_{itt^*}(\boldsymbol{\beta}) := (Y_{it} - \mu_{it}(\boldsymbol{\beta}))(Y_{it^*} - \mu_{it^*}(\boldsymbol{\beta}))$. Finally, let $\boldsymbol{R}_2(\lambda_2) \in \mathbb{R}^{[T(T+1)/2]\times[T(T+1)/2]}$ be a working correlation matrix for $\boldsymbol{S}_i(\boldsymbol{\beta})$ and λ_2 its parameter. With $\tau^2_{itt^*}(\boldsymbol{\delta}) := \mathrm{Var}(S_{itt^*}(\boldsymbol{\beta}) \mid \boldsymbol{\delta})$, we can again construct a working covariance

$$\begin{aligned}\boldsymbol{V}_{i2}(\boldsymbol{\delta}, \lambda_2) := \mathrm{Cov}(\boldsymbol{S}_i(\boldsymbol{\beta}) \mid \boldsymbol{\delta}, \lambda_2) \ &= \ \mathrm{diag}\big(\tau_{i11}(\boldsymbol{\delta}), \tau_{i12}(\boldsymbol{\delta}), \ldots, \tau_{iTT}(\boldsymbol{\delta})\big)\boldsymbol{R}_2(\lambda_2) \\ &\quad \times \mathrm{diag}\big(\tau_{i11}(\boldsymbol{\delta}), \tau_{i12}(\boldsymbol{\delta}), \ldots, \tau_{iTT}(\boldsymbol{\delta})\big).\end{aligned} \tag{4.2}$$

We will address the problem of determining analytical expressions for $\tau^2_{itt^*}(\boldsymbol{\delta})$ later. The estimating equation according to GEE1 is

$$K^{-1/2}\sum_{i=1}^{K} \boldsymbol{D}'_{i1}(\boldsymbol{\beta})\boldsymbol{V}^{-1}_{i1}(\boldsymbol{\delta})(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \boldsymbol{0}_{p+1}, \tag{4.3}$$

where $\boldsymbol{D}_{i1}(\boldsymbol{\beta}) = \frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \in \mathbb{R}^{T\times[p+1]}$ and $\boldsymbol{0}_{p+1}$ is a $(p+1)$-dimensional vector of zeros. Parameter $\boldsymbol{\alpha}$ together with γ will be estimated using GEE2 by solving

$$K^{-1/2}\sum_{i=1}^{K} \boldsymbol{D}'_{i2}(\boldsymbol{\delta})\boldsymbol{V}^{-1}_{i2}(\boldsymbol{\delta}, \lambda_2)\big(\boldsymbol{S}_i(\boldsymbol{\beta}) - \boldsymbol{\sigma}^2_i(\boldsymbol{\delta})\big) = \boldsymbol{0}_{q+2}, \tag{4.4}$$

where $\boldsymbol{D}_{i2}(\boldsymbol{\delta}) := \frac{\partial \boldsymbol{\sigma}_i^2(\boldsymbol{\delta})}{\partial(\boldsymbol{\alpha}',\gamma)'} \in \mathbb{R}^{[T(T+1)/2]\times[q+2]}$. Additionally, we calculate $\boldsymbol{D}_{i12}(\boldsymbol{\delta}) := \frac{\partial \boldsymbol{\sigma}_i^2(\boldsymbol{\delta})}{\partial \boldsymbol{\beta}} \in \mathbb{R}^{[T(T+1)/2]\times[p+1]}$ since we hope to gain information on the mean parameters $\boldsymbol{\beta}$ also from $\boldsymbol{\sigma}_i^2(\boldsymbol{\delta})$. According to our experience, setting $\boldsymbol{D}_{i2}(\boldsymbol{\delta})$ to $\boldsymbol{0}$ gives similar results and decreases the required computational time. For the GP distribution, covariances are $\sigma^2_{itt^*}(\boldsymbol{\delta}) = \rho_{tt^*}\big(\lambda_1(\gamma)\big)\sqrt{\mu_{it}(\boldsymbol{\beta})\varphi^2_{it}(\boldsymbol{\alpha})}\sqrt{\mu_{it^*}(\boldsymbol{\beta})\varphi^2_{it^*}(\boldsymbol{\alpha})}$, where $\rho_{tt^*}\big(\lambda_1(\gamma)\big) = \mathrm{Corr}(Y_{it}, Y_{it^*})$, $i = 1, \dots, K$. Then,

$$\boldsymbol{D}_{i1}(\boldsymbol{\beta}) = \Big[\mu_{it}(\boldsymbol{\beta})x_{itr}\Big]_{t=1,\dots,T,\; r=1,\dots,p+1}, \tag{4.5}$$

$$\boldsymbol{D}'_{i2}(\boldsymbol{\delta}) = \begin{pmatrix} \begin{bmatrix} \rho_{tt^*}(\lambda_1(\gamma))\sqrt{\mu_{it}(\boldsymbol{\beta})\mu_{it^*}(\boldsymbol{\beta})}\times \\ \{b_{it}(\boldsymbol{\alpha})w_{itr}\varphi_{it^*}(\boldsymbol{\alpha}) + \\ b_{it^*}(\boldsymbol{\alpha})w_{it^*r}\varphi_{it}(\boldsymbol{\alpha})\} \end{bmatrix}_{\substack{(t,t^*)\in I, \\ r=1,\dots,q+1}} \\ \left[\frac{\partial\rho_{tt^*}(\lambda_1(\gamma))}{\partial\lambda_1(\gamma)}\frac{4e^{2\gamma}}{(e^{2\gamma}+1)^2}\sigma_{it}(\boldsymbol{\delta})\sigma_{it^*}(\boldsymbol{\delta})\right]_{(t,t^*)\in I} \end{pmatrix}, \tag{4.6}$$

$$\boldsymbol{D}'_{i12}(\boldsymbol{\delta}) = \begin{bmatrix} \frac{1}{2}\rho_{tt^*}(\lambda_1(\gamma))\ \varphi_{it}(\boldsymbol{\alpha})\varphi_{it^*}(\boldsymbol{\alpha})\times \\ \left\{\sqrt{\mu_{it}(\boldsymbol{\beta})}x_{itr} + \sqrt{\mu_{it^*}(\boldsymbol{\beta})}x_{it^*r}\right\} \end{bmatrix}_{(t,t^*)\in I,\; r=1,\dots,p+1} \tag{4.7}$$

and $I := \{(t,t^*) \mid t \le t^*\}$. Now (4.3) and (4.4) can be solved simultaneously. One defines

$$\boldsymbol{D}_i(\boldsymbol{\delta}) := \begin{pmatrix} \frac{\partial\boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} & \frac{\partial\boldsymbol{\sigma}_i^2(\boldsymbol{\delta})}{\partial\boldsymbol{\beta}} \\ \boldsymbol{0} & \frac{\partial\boldsymbol{\sigma}_i^2(\boldsymbol{\delta})}{\partial(\boldsymbol{\alpha}',\gamma)'} \end{pmatrix} = \begin{pmatrix} \boldsymbol{D}_{i1}(\boldsymbol{\delta}) & \boldsymbol{D}_{i12}(\boldsymbol{\delta}) \\ \boldsymbol{0} & \boldsymbol{D}_{i2}(\boldsymbol{\delta}) \end{pmatrix} \in \mathbb{R}^{[T(T+3)/2]\times[p+q+3]}, \tag{4.8}$$

$$\boldsymbol{V}_i(\boldsymbol{\delta},\lambda_2) := \begin{pmatrix} \boldsymbol{V}_{i1}(\boldsymbol{\delta}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_{i2}(\boldsymbol{\delta},\lambda_2) \end{pmatrix} \in \mathbb{R}^{[T(T+3)/2]\times[T(T+3)/2]}, \tag{4.9}$$

$$\boldsymbol{f}_i(\boldsymbol{\delta}) := \begin{pmatrix} \boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}) \\ \boldsymbol{s}_i(\boldsymbol{\beta}) - \boldsymbol{\sigma}_i^2(\boldsymbol{\delta}) \end{pmatrix} \in \mathbb{R}^{T(T+3)/2} \tag{4.10}$$

and $\boldsymbol{\Gamma}(\boldsymbol{\delta},\lambda_2) := K^{-1}\sum_{i=1}^K \boldsymbol{D}'_i(\boldsymbol{\delta})\boldsymbol{V}_i^{-1}(\boldsymbol{\delta},\lambda_2)\boldsymbol{D}_i(\boldsymbol{\delta})$. The overall set of estimating equations is $K^{-1/2}\sum_{i=1}^K \boldsymbol{D}'_i(\boldsymbol{\delta})\boldsymbol{V}_i^{-1}(\boldsymbol{\delta},\lambda_2)\boldsymbol{f}_i(\boldsymbol{\delta}) = \boldsymbol{0}_{p+q+3}$. Updating $\boldsymbol{\delta}$ by a Fisher-Scoring step yields

$$\hat{\boldsymbol{\delta}}_{j+1} = \hat{\boldsymbol{\delta}}_j + \left\{\sum_{i=1}^K \boldsymbol{D}'_i(\hat{\boldsymbol{\delta}}_j)\boldsymbol{V}_i^{-1}(\hat{\boldsymbol{\delta}}_j,\hat{\lambda}_{2j})\boldsymbol{D}_i(\hat{\boldsymbol{\delta}}_j)\right\}^{-1}\left\{\sum_{i=1}^K \boldsymbol{D}'_i(\hat{\boldsymbol{\delta}}_j)\boldsymbol{V}_i^{-1}(\hat{\boldsymbol{\delta}}_j,\hat{\lambda}_{2j})\boldsymbol{f}_i(\hat{\boldsymbol{\delta}}_j)\right\}, \tag{4.11}$$

where $\hat{\lambda}_{2j} := \hat{\lambda}_2(\hat{\boldsymbol{\delta}}_j)$. Residuals $\hat{r}_{ilm}(\hat{\boldsymbol{\delta}}) := s_{ilm}(\hat{\boldsymbol{\beta}}) - \sigma^2_{ilm}(\hat{\boldsymbol{\delta}})$ may be used to estimate λ_2. For example, for an exchangeable matrix $\boldsymbol{R}_2(\lambda_2)$, define $I^* := \{(lm, l^*m^*) : l \le l^* \wedge m \le m^*\}$. Then according to Liang & Zeger (1986, p. 18, example 3), an estimate of λ_2 is given by

$$\hat{\lambda}_2(\hat{\boldsymbol{\delta}}) = \frac{\sum_{i=1}^{K} \sum_{(lm,l^*m^*)\in I^*} \hat{r}_{ilm}(\hat{\boldsymbol{\delta}})\hat{r}_{il^*m^*}(\hat{\boldsymbol{\delta}})}{K\left[\frac{T(T+1)}{2}\left(\frac{T(T+1)}{2}-1\right)/2\right]-(p+q+3)}. \tag{4.12}$$

According to Prentice & Zhao (1991, Appendix 1), $\boldsymbol{Z} := K^{1/2}\big((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})', (\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})', (\hat{\gamma}-\gamma)'\big)'$ is asymptotically normal for $K \to \infty$ with mean $\mathbf{0}_{p+q+3}$ and covariance

$$\mathrm{Cov}(\boldsymbol{Z}) = K^{-1}\boldsymbol{\Gamma}^{-1}(\boldsymbol{\delta},\lambda_2)\left(\sum_{i=1}^{K} \boldsymbol{D}_i'(\boldsymbol{\delta})\boldsymbol{V}_i^{-1}(\boldsymbol{\delta},\lambda_2)\mathrm{Cov}\big((\boldsymbol{Y}_i', \boldsymbol{S}_i'(\boldsymbol{\beta}))'\big)\boldsymbol{V}_i^{-1}(\boldsymbol{\delta},\lambda_2)\boldsymbol{D}_i(\boldsymbol{\delta})\right)\boldsymbol{\Gamma}^{-1}(\boldsymbol{\delta},\lambda_2). \tag{4.13}$$

Note that $\mathrm{Cov}\big((\boldsymbol{Y}_i', \boldsymbol{S}_i'(\boldsymbol{\beta}))'\big)$ is unknown. A consistent 'sandwich' estimator of (4.13) is

$$\begin{aligned} \boldsymbol{\Omega}_{sw}(\hat{\boldsymbol{\delta}},\hat{\lambda}_2) &:= \widehat{\mathrm{Cov}}(\boldsymbol{Z})_{sw} && (4.14)\\ &= K^{-1}\boldsymbol{\Gamma}^{-1}(\hat{\boldsymbol{\delta}},\hat{\lambda}_2)\left(\sum_{i=1}^{K} \boldsymbol{D}_i'(\hat{\boldsymbol{\delta}})\boldsymbol{V}_i^{-1}(\hat{\boldsymbol{\delta}},\hat{\lambda}_2)\boldsymbol{f}_i(\hat{\boldsymbol{\delta}})\boldsymbol{f}_i'(\hat{\boldsymbol{\delta}})\boldsymbol{V}_i^{-1}(\hat{\boldsymbol{\delta}},\hat{\lambda}_2)\boldsymbol{D}_i(\hat{\boldsymbol{\delta}})\right)\boldsymbol{\Gamma}^{-1}(\hat{\boldsymbol{\delta}},\hat{\lambda}_2) \end{aligned}$$

(see Prentice & Zhao (1991, p. 828)). Alternatively, a model-based estimator of the variance of $\boldsymbol{Z}$ is obtained by replacing $\mathrm{Cov}((\boldsymbol{Y}_i', \boldsymbol{S}_i'(\boldsymbol{\beta}))')$ by $\boldsymbol{V}_i(\hat{\boldsymbol{\delta}},\hat{\lambda}_2)$ yielding

$$\boldsymbol{\Omega}_{mb}(\hat{\boldsymbol{\delta}},\hat{\lambda}_2) := \widehat{\mathrm{Cov}}(\boldsymbol{Z})_{mb} = K^{-1}\boldsymbol{\Gamma}^{-1}(\hat{\boldsymbol{\delta}},\hat{\lambda}_2). \tag{4.15}$$

An issue still open is how to determine

$$\begin{aligned} \tau_{itt^*}^2(\boldsymbol{\delta}) &:= \mathrm{Var}\big((Y_{it}-\mu_{it}(\boldsymbol{\beta}))(Y_{it^*}-\mu_{it^*}(\boldsymbol{\beta}))\big)\\ &= E\left[Y_{it}^2Y_{it^*}^2\right] - 2E\left[Y_{it}^2Y_{it^*}\right]\mu_{it^*}(\boldsymbol{\beta}) - 2E\left[Y_{it}Y_{it^*}^2\right]\mu_{it}(\boldsymbol{\beta})\\ &\quad +4E\left[Y_{it}Y_{it^*}\right]\mu_{it}(\boldsymbol{\beta})\mu_{it^*}(\boldsymbol{\beta}) + E[Y_{it}^2]\mu_{it^*}^2(\boldsymbol{\beta}) + E[Y_{it^*}^2]\mu_{it}^2(\boldsymbol{\beta}) - 3\mu_{it}^2(\boldsymbol{\beta})\mu_{it^*}(\boldsymbol{\beta})^2\\ &\quad -\left[\rho_{tt^*}\big(\lambda_1(\gamma)\big)\sqrt{\mathrm{Var}(Y_{it})\mathrm{Var}(Y_{it^*})}\right]^2. \end{aligned} \tag{4.16}$$

in (4.2). So $\tau_{itt^*}^2(\boldsymbol{\delta})$ is a function of higher mixed moments $E\left[Y_{it}Y_{it^*}\right]$, $E\left[Y_{it}^2Y_{it^*}\right]$, $E\left[Y_{it}Y_{it^*}^2\right]$ and $E\left[Y_{it}^2Y_{it^*}^2\right]$ depending on $\boldsymbol{\delta}$ for which a closed form is unknown - except for the first one, for which we have an expression based on our working correlation. The remaining mixed moments can be determined if $t = t^*$, since in this case moments up to order 4 are needed, which exist for the GP distribution (see Consul (1989, p. 50)).

However, if $t < t^*$ a different approach is necessary. For this consider the general bivariate specification $\boldsymbol{Y} = (Y_1, Y_2)$ with $Y_1 \sim GP(\mu_1, \varphi_1)$, $Y_2 \sim GP(\mu_2, \varphi_2)$ and correlation ρ. Here we abbreviate $Y_1 := Y_{it}$, $Y_2 := Y_{it^*}$, $\mu_1 := \mu_{it}(\boldsymbol{\beta})$, $\mu_2 := \mu_{it^*}(\boldsymbol{\beta})$, $\varphi_1 := \varphi_{it}(\boldsymbol{\alpha})$, $\varphi_2 := \varphi_{it^*}(\boldsymbol{\alpha})$ and $\rho :=$

$\rho_{tt^*}\big(\lambda_1(\gamma)\big)$. We would like to simulate from such a specification by using a bivariate Gaussian copula, i.e. by first simulating $(Z_1, Z_2) \sim N_2(\mathbf{0}, \Sigma(\rho))$ where $\Sigma(\rho)$ is diagonal with elements $g(\rho)$. Then we consider the probability integral transformation $(U_1, U_2) := (\Phi(Z_1), \Phi(Z_2))$ and utilize the inversion method (Famoye (1997, p. 222)) to sample count random variables, i.e. we calculate the quantiles $y_1 := F_{GP}^{-1}(u_1|\mu_1, \varphi_1)$ and $y_2 := F_{GP}^{-1}(u_2|\mu_2, \varphi_2)$. Here $F_{GP}(\cdot|\mu, \varphi)$ denotes the cdf of a $GP(\mu, \varphi)$ random variable. We need to determine $g(\rho)$ such that $\text{Corr}(Y_1, Y_2) = \rho$. This is approximately accomplished using the approach suggested by Erhardt & Czado (2009).

To approximate $\tau^2(\boldsymbol{\theta}) := \text{Var}\left((Y_1 - \mu_1)(Y_2 - \mu_2)\right)$ with $\boldsymbol{\theta} := (\mu_1, \mu_2, \varphi_1, \varphi_2, \rho)$ we generate a sample of $\boldsymbol{Y}^r(\boldsymbol{\theta}) = (Y_1^r(\boldsymbol{\theta}), Y_2^r(\boldsymbol{\theta}))$, $r = 1, \ldots, R$ using the above sampling approach. Now we approximate $\hat{\tau}^2(\boldsymbol{\theta}) := \frac{1}{R}\sum_{r=1}^R \Big[(y_1^r(\boldsymbol{\theta}) - \mu_1)(y_2^r(\boldsymbol{\theta}) - \mu_2)\Big]^2 - \Big[\frac{1}{R}\sum_{r=1}^R (y_1^r(\boldsymbol{\theta}) - \mu_1)(y_2^r(\boldsymbol{\theta}) - \mu_2)\Big]^2$.

Since we are interested in an approximate analytical expression for $\tau^2(\boldsymbol{\theta})$ for arbitrary values of $\boldsymbol{\theta}$, we use a log-normal regression approach to express $\hat{\tau}^2(\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ over a grid of values $(\mu_1, \mu_2, \varphi_1, \varphi_2, \rho)$. In particular we use grid values $\boldsymbol{\theta}_j = (\mu_{1j}, \mu_{2j}, \varphi_{1j}, \varphi_{2j}, \rho_j)$, $j = 1, \ldots, 6^3 \cdot 5^2 = 5\ 400$ constructed by

1. $\{2, 8, 25, 50, 150, 400\}$ for μ_{1j} and μ_{2j}, respectively,

2. $\{1, 2, 3, 6, 9\}$ for φ_{1j} and φ_{2j}, respectively,

3. $\{-0.8, -0.5, -0.25, 0.25, 0.5, 0.8\}$ for ρ_j.

In order to specify such a grid we started by fitting a $GP(\mu_i, \varphi_i)$ regression model according to Czado et al. (2007) using the *R* software package 'ZIGP' (Erhardt (2009)) available on CRAN. Thereby we ignored the clustered structure of the data and assumed all observations to be independent. Then we chose as smallest and largest grid points for μ_{1j} and μ_{2j}, φ_{1j} and φ_{2j}, values not far outside the range of fitted means and overdispersion parameters, respectively. The remaining grid points were chosen such that they were more dense at the lower part of the chosen range where most of the fitted values could be found. The grid points for ρ_j were chosen symmetric around 0 and also more close to 0.

Let $\hat{\tau}_j^2 := \hat{\tau}^2(\boldsymbol{\theta}_j)$ and consider the log-normal regression of response $\hat{\tau}_j^2$ with covariates μ_{1j}, μ_{2j}, φ_{1j}, φ_{2j} and ρ_j for $j = 1, \ldots, 5400$. From an exploratory data analysis we see that we need to distinguish the cases $\rho_j < 0$ and $\rho_j \geq 0$. For both cases we use as explanatory variables an intercept, $\log(\mu_{1j})$, $\log(\mu_{2j})$, $\log(\varphi_{1j})$, $\log(\varphi_{2j})$ and ρ_j and all three-dimensional interactions. Then by backward selection we eliminate nonsignificant effects according to the Wald test. The fitted mean function $\hat{E}(\log(\hat{\tau}_j^2))$ for $\log(\hat{\tau}_j^2)$ for the case $\rho_j \geq 0$ is given by

$$\begin{aligned}
\hat{E}(\log(\hat{\tau}_j^2)) &= -0.454 \cdot +1.027 \cdot \log(\mu_{1j}) + 2.615 \cdot \log(\varphi_{1j}) + 0.974 \cdot \log(\mu_{2j}) + 2.650 \cdot \log(\varphi_{2j}) + \\
&\quad 1.186 \cdot \rho_j - 0.135 \cdot \log(\mu_{1j}) \cdot \log(\varphi_{1j}) + 0.010 \cdot \log(\mu_{1j}) \cdot \log(\mu_{2j}) - \\
&\quad 0.028 \cdot \log(\mu_{1j}) \cdot \log(\varphi_{2j}) - 0.110 \cdot \log(\mu_{1j}) \cdot \rho_j + 0.086 \cdot \log(\varphi_{1j}) \cdot \log(\varphi_{2j}) + \\
&\quad 0.913 \cdot \log(\varphi_{1j}) \cdot \rho_j - 0.116 \cdot \log(\mu_{2j}) \cdot \log(\varphi_{2j}) + 0.804 \cdot \log(\varphi_{2j}) \cdot \rho_j - \\
&\quad 0.058 \cdot \log(\mu_{1j}) \cdot \log(\varphi_{1j}) \cdot \rho_j - 0.056 \cdot \log(\varphi_{1j}) \cdot \log(\mu_{2j}) \cdot \rho_j - \\
&\quad 0.087 \cdot \log(\mu_{2j}) \cdot \log(\varphi_{2j}) \cdot \rho_j \qquad (4.17)
\end{aligned}$$

with an adjusted R^2 of 99.2%. For $\rho_j \leq 0$ we get a similar expression (adjusted $R^2 = 99.96\%$. Finally for $\boldsymbol{\delta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \gamma)$ and $\gamma = \tanh^{-1}(\rho) = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$ we approximate $\tau_{itt^*}^2(\boldsymbol{\delta})$ by the analytical expression $\exp(\hat{E}(\log(\hat{\tau}^2)|\boldsymbol{\theta}_{itt^*}^*))$, where $\boldsymbol{\theta}_{itt^*}^* := \big(\mu_{it}(\boldsymbol{\beta}), \mu_{it^*}(\boldsymbol{\beta}), \varphi_{it}(\boldsymbol{\alpha}), \varphi_{it^*}(\boldsymbol{\alpha}), \rho_{itt^*}(\lambda_1(\gamma))\big)$.

5 Small sample properties of the GEE estimates

In a simulation study we generated $N = 1000$ samples from $\{Y_{it}, i = 1, \ldots, K; t = 1, \ldots T\}$ counts with $Y_{it} \sim GP(\mu_{it}, \varphi_{it})$ independent for $i = 1, \ldots, K$ and correlation $\boldsymbol{\Sigma}$ for $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iT})$. As correlation matrix $\boldsymbol{\Sigma}$ we chose an autoregressive $AR(1)$ structure, i.e. $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\lambda)$, where $[\boldsymbol{\Sigma}(\lambda)]_{tt^*} = \lambda^{|t-t^*|}$. Again this is facilitated using the approximate approach suggested by Erhardt & Czado (2009).

'Small' and 'large' number of subjects of $K = 250$ and $K = 500$ were taken into consideration. As test setting, we chose $T = 8$ and $\lambda_1 = 0.5$. The design matrix for the mean level contains an intercept, subject-specific and a time specific covariate, while the one for the dispersion level contains an intercept and a subject-specific one. In particular we use

$$\begin{aligned}
\log(\mu_{it}) &= \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot t/T \qquad (5.1) \\
\log(\varphi_{it} - 1) &= \alpha_0 + \alpha_1 \cdot w_i. \qquad (5.2)
\end{aligned}$$

Here x_i is distributed equidistantly on $[-1, 1]$ and w_i on $[-2, 2]$ over all subjects. Choosing $\beta_1 = \beta_2$, the parameter values were chosen to be $\boldsymbol{\beta} = (1.32, 0.70, 0.70)'$ and $\boldsymbol{\alpha} = (0.21, 0.90)'$ to yield $\mu_{it}(\boldsymbol{\beta}) \in [2, 15]$ and $\varphi_{it}(\boldsymbol{\alpha}) \in [1.5, 4]$, respectively. QQ plots shown in Figure 1 were used to assess the asymptotic normality of the estimates. The parameters on the mean level have approximately a normal distribution already for $K = 250$, while this is only approximately true for $K = 500$ for the parameters on the dispersion level.

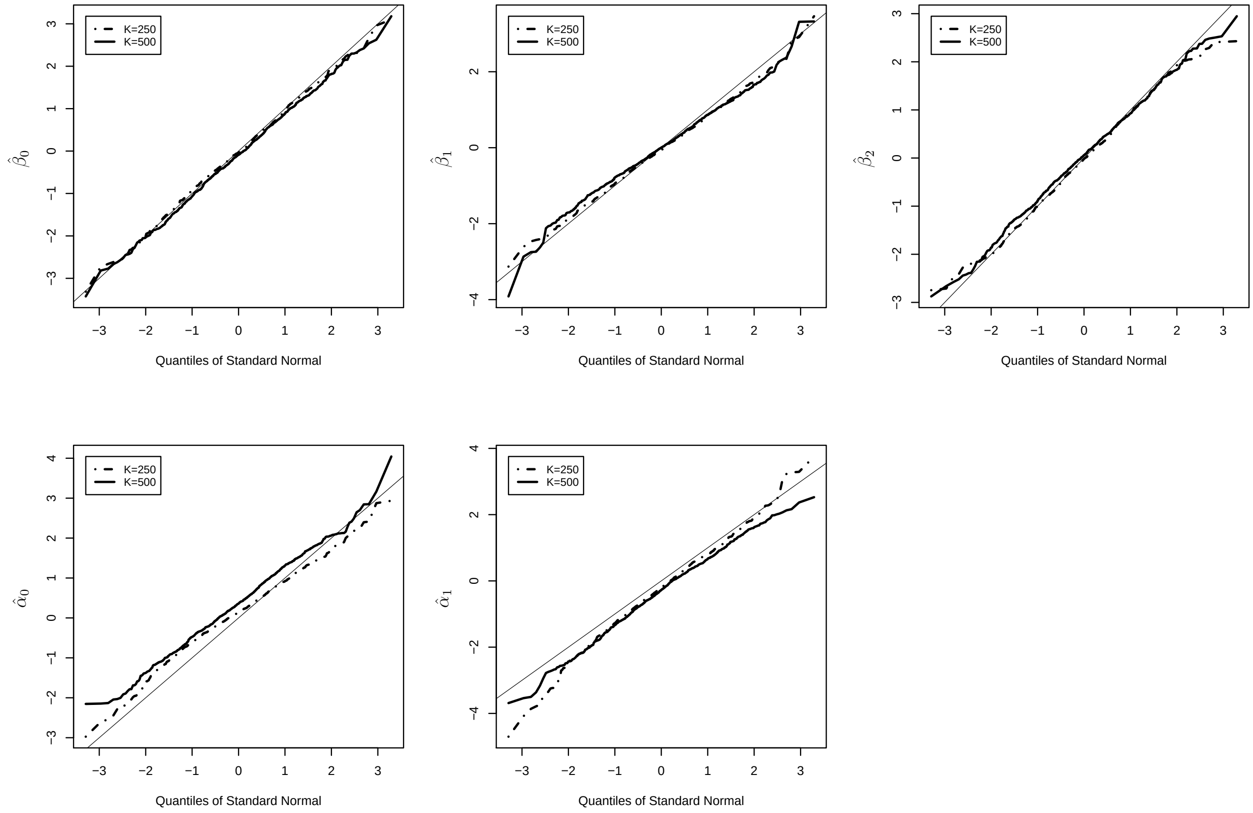

Figure 1: QQ plots of centered and standardized estimates based on $N = 1000$ replications ($K = 250, 500$, $T = 8$, $\lambda = 0.5$, $\boldsymbol{\beta} = (1.32, 0.70, 0.70)'$, $\boldsymbol{\alpha} = (0.25, 0.31, 0.31)'$)

By considering the mean of the estimated parameters and estimated mean squared errors (MSE) together with standard errors for both statistics, the predictive quality of the estimation method will be assessed (see Table 1). The relative bias of an estimate $\hat\theta$ of θ is given by $b(\theta, \hat\theta) = \frac{E(\hat\theta)-\theta}{\theta}$ and for a sample of N independent estimates it will be estimated by

$$\hat b(\theta, \hat{\boldsymbol{\theta}}) = \frac{1}{N\theta}\sum_{i=1}^{N}\hat\theta_i - 1. \tag{5.3}$$

The estimated variance of the estimated relative bias is given by $\widehat{\mathrm{Var}}\big(\hat b(\theta, \hat\theta_i)\big) := 1/\theta^2\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}} := (\hat\theta_1, \ldots, \hat\theta_N)'$, and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}) := \frac{1}{N-1}\sum_{i=1}^{N}\left(\hat\theta_i - \frac{1}{N}\sum_{k=1}^{N}\hat\theta_k\right)^2$. The mean squared error (MSE) is given by

$$R(\theta, \hat\theta) := E([\hat\theta - \theta]^2) = \mathrm{Var}(\hat\theta) + b^2(\hat\theta, \theta). \tag{5.4}$$

Its variance can be estimated by $\widehat{Var}(R(\theta, \hat\theta)) = \frac{1}{N}(m_4 - 4\theta m_3 + 4\theta^2 m_2 - m_2^2 + 4\theta m_1 m_2 - 4\theta^2 m_2)$, where m_k is the kth moment estimate of θ, so $m_k := \frac{1}{N}\sum_{i=1}^{N}\theta_i^k$ is an estimate of

$\mu_k = E(\theta^k)$ (Stekeler (2004, p. 126)). The standard deviations of the parameter estimates $\hat{\boldsymbol{\delta}}$ will be calculated using 'sandwich' estimates given in (4.14).

	Parameter	True value	T	K	Estimate	Relative Bias	MSE
$\mu_{it} \in [2, 15]$	β_0	1.32	8	250	1.319 (0.067)	0.001 (0.051)	0.004 $(3 \cdot 10^{-8})$
				500	1.316 (0.048)	0.004 (0.036)	0.002 $(9 \cdot 10^{-9})$
	β_1	0.7	8	250	0.696 (0.092)	0.004 (0.132)	0.009 $(8 \cdot 10^{-8})$
				500	0.701 (0.068)	-0.001 (0.097)	0.005 $(2 \cdot 10^{-8})$
	β_2	0.7	8	250	0.698 (0.063)	0.002 (0.090)	0.004 $(2 \cdot 10^{-8})$
				500	0.702 (0.044)	-0.002 (0.063)	0.002 $(5 \cdot 10^{-9})$
$\varphi_{it} \in [1.5, 4]$	α_0	0.21	8	250	0.215 (0.108)	-0.005 (0.514)	0.012 $(1 \cdot 10^{-7})$
				500	0.223 (0.082)	-0.013 (0.388)	0.007 $(3 \cdot 10^{-8})$
	α_1	0.9	8	250	0.873 (0.181)	0.027 (0.201)	0.033 $(2 \cdot 10^{-6})$
				500	0.865 (0.139)	0.035 (0.154)	0.021 $(9 \cdot 10^{-7})$
$\lambda = 0.5$	λ	0.5	8	250	0.489 (0.101)	0.011 (0.202)	0.010 $(8 \cdot 10^{-9})$
				500	0.504 (0.080)	-0.004 (0.160)	0.006 $(4 \cdot 10^{-9})$

Table 1: Average coefficients, relative bias (see 5.3) and mean squared error (see 5.4) together with estimated 'sandwich' standard deviations in round brackets according to (4.14) for $N = 1000$ fitted samples.

This shows that the accuracy of the estimations is satisfactory for medium sample sizes. The absolute values of the relative bias in Table 1 are smaller for the mean effects than for the dispersion effects, hence the mean coefficients are estimated better than the dispersion coefficents. This is due to the approximating approach for determining $\tau^2_{itt^*}(\boldsymbol{\delta})$.

Several alternative setups have also been investigated. The main results of these additional simulations are that increasing the range of means $\mu_{it}(\boldsymbol{\beta})$ leads to even better results. The reason is that a larger range of $\mu_{it}(\boldsymbol{\beta})$ covers a larger and steeper interval of the inverse link function which implies larger absolute derivatives of the link functions and larger absolute true values. These circumstances improve parameter estimation. Increasing overdispersion results in worse estimates of the mean parameters. The reason is simply higher data heterogeneity in the counts. Understandably, dispersion parameters are estimated better in this setting because again, a larger and steeper interval of the inverse dispersion link is covered. Moreover, higher correlated data improves the estimation of time-specific covariates. For all other covariates, highly correlated data seems to carry less information over time than weakly correlated data. Finally, increasing the number of time points T has a positive impact on the estimation quality of the mean parameters. This is in line to what one would expect from longer time series.

6 Variable selection and model comparison

6.1 A variable selection criterion for nested models

Standard approaches for variable selection such as the Akaike Information Criterion (AIC) (Akaike (1974)) require a fully specified likelihood. Pan (2001) introduces a criterion for GEE which uses only the quasi-likelihood. For a r.v. Y with $E(Y) = \mu$ and $\text{Var}(Y) = \phi V(\mu)$, where ϕ is a dispersion parameter, the quasi-likelihood function is defined as $QL(\mu, \phi, y) = \int_y^\mu \frac{y-t}{\phi V(t)} dt$ (McCullagh & Nelder (1989, p. 325)). In the GP context, we have $E(Y_{it}) = \mu_{it}(\boldsymbol{\beta})$ and $\text{Var}(Y_{it}) = \varphi_{it}^2(\boldsymbol{\alpha})\mu_{it}(\boldsymbol{\beta})$, i.e. $V(\mu_{it}(\boldsymbol{\beta})) = \mu_{it}(\boldsymbol{\beta})$ and $\phi = \varphi_{it}^2(\boldsymbol{\alpha})$, and obtain

$$\begin{aligned} QL(\mu_{it}(\boldsymbol{\beta}), \varphi_{it}(\boldsymbol{\alpha}), y_{it}) &= \int_{y_{it}}^{\mu_{it}(\boldsymbol{\beta})} \frac{y_{it} - t}{\varphi_{it}^2(\boldsymbol{\alpha}) t} dt \qquad (6.1)\\ &= \frac{1}{\varphi_{it}^2(\boldsymbol{\alpha})} \left(y_{it} \log(\mu_{it}(\boldsymbol{\beta})) - \mu_{it}(\boldsymbol{\beta}) \right) + \text{ constants ind. of } (\boldsymbol{\beta}, \boldsymbol{\alpha}). \end{aligned}$$

If overall independence across times and subjects is assumed, the overall quasi-likelihood under independence becomes

$$\begin{aligned} Q(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{y}) &:= \sum_{i=1}^K \sum_{t=1}^T QL(\mu_{it}(\boldsymbol{\beta}), \varphi_{it}(\boldsymbol{\alpha}), y_{it}) \qquad (6.2)\\ &= \sum_{i=1}^K \sum_{t=1}^T \frac{1}{\varphi_{it}^2(\boldsymbol{\alpha})} \left(y_{it} \log(\mu_{it}(\boldsymbol{\beta})) - \mu_{it}(\boldsymbol{\beta}) \right) + \text{ constants ind. of } (\boldsymbol{\beta}, \boldsymbol{\alpha}). \end{aligned}$$

A model with parameter vector $\boldsymbol{\theta} \in \mathbb{R}^k$ and estimate $\hat{\boldsymbol{\theta}}$ is compared by $AIC(\hat{\boldsymbol{\theta}}) := -2L(\hat{\boldsymbol{\theta}}) + 2k$. Pan (2001) replaces the log-likelihood by the quasi-likelihood and the penalty term $2k$ by 2 $\text{trace}(\boldsymbol{\Omega}_{mb}^{-1}(\hat{\boldsymbol{\delta}}, \hat{\lambda}_2)\boldsymbol{\Omega}_{sw}(\hat{\boldsymbol{\delta}}, \hat{\lambda}_2))$. With $\hat{\boldsymbol{\delta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\gamma})$ and the working correlations matrix $\boldsymbol{R}_1$ being a function of γ, a 'quasi-likelihood under independence model criterion' (QIC) is

$$QIC(\hat{\boldsymbol{\delta}}, \hat{\lambda}_2) := -2Q(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \boldsymbol{I}, \boldsymbol{y}) + 2 \text{ trace}(\boldsymbol{\Omega}_{mb}^{-1}(\hat{\boldsymbol{\delta}}, \hat{\lambda}_2)\boldsymbol{\Omega}_{sw}(\hat{\boldsymbol{\delta}}, \hat{\lambda}_2)). \qquad (6.3)$$

As for the AIC, the smaller the QIC, the better the model.

6.2 Assessing model fit for nonnested models

Recall that we denote the GEE setup for correlated count data $\boldsymbol{Y} = (Y_{it}, i = 1, \ldots, K; t = 1, \ldots, T)$ with $GP(\mu_{it}, \varphi_{it})$ margins for Y_{it} and working correlation matrix $\boldsymbol{R}$ by $GP(\mu_{it}, \varphi_{it}, \boldsymbol{R})$. Similar we denote by $Poi(\mu_{it}, \boldsymbol{R})$ a setup with margin $Y_{it} \sim Poi(\mu_{it})$. A GEE setup where the overdispersion parameter for Y_{it} is constant, we denote by $GP(\mu_{it}, \varphi, \boldsymbol{R})$. The corresponding model hierarchy is given in Figure 2. A covariate being significant in terms of the Wald test (e.g. in the mean design of $Poi(\mu_{it}, \boldsymbol{R})$) can be insignificant in a different model (say $GP(\mu_{it}, \varphi, \boldsymbol{R})$). The same holds for dispersion designs. Therefore, a pool of covariates chosen in an exploratory data analysis will be reduced by backward selection using the QIC in each one of our setup classes. Since design matrices may thus be different, their designs need not be nested.

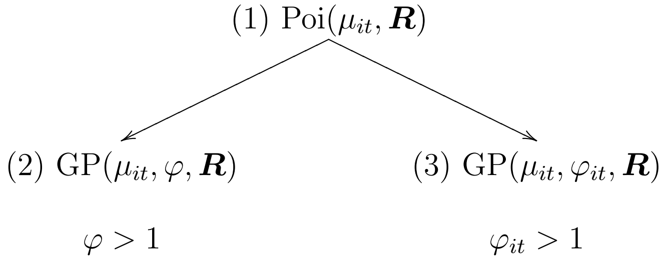
(1) Poi($\mu_{it}, \boldsymbol{R}$)
(2) GP($\mu_{it}, \varphi, \boldsymbol{R}$)
(3) GP($\mu_{it}, \varphi_{it}, \boldsymbol{R}$)
$\varphi > 1$
$\varphi_{it} > 1$

Figure 2: Investigated setup hierarchy

There exists a test proposed by Vuong (1989) which can be used to compare models with nonnested settings. The test statistic, however, is based on the Kullback-Leibler information criterion (KLIC), which requires a fully specified likelihood. Therefore, this approach cannot by applied here. The same holds for a distribution-free test proposed by Clarke (2007).

We will use the Wald-Wolfowitz run test for testing the goodness-of-fit as proposed by Chang (2000) and also described in Hilbe (2007, Section 4.2.1f). The residuals will be sorted by the corresponding fitted means. We define an indicator whether the residual is positive ('1') or negative ('-1') in the same ordering. Further n_p will be the number of positive, n_n the number of negative residuals. Let T the number of runs in the sequence of indicators. Under the null hypothesis that the signs of the residuals are distributed in a random sequence, the expected value and variance of T are given as $E(T) = \frac{2n_p n_n}{n_p+n_n} + 1$ and $\mathrm{Var}(T) = \frac{2n_p n_n(2n_p n_n - n_p - n_n)}{(n_p+n_n)^2(n_p+n_n-1)}$. Then $W_Z := \frac{T-E(T)}{\mathrm{Var}(T)}$ is approximately standard normal. A α level test can be constructed as

$$\text{Reject } H_0 \text{ if } |W_Z| > q_{1-\alpha/2} \tag{6.4}$$

where $q_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. Note that this criterion does not account for the model complexity, for the choice between competing setups one has to consider the number of model parameters as well.

7 Application: Outsourcing of patent applications

7.1 Data description and model comparison

The data consists of patent information of the European Patent Office. It has been examined and completed with corporate information by Wagner (2006b). A zero-inflated generalized Poisson regression model assuming independent observations has been considered by Czado et al. (2007) for this data. A more detailed description of this model will be given in Section 7.2. The survey considers 107 European companies over eight years (1993 to 2000). There are two ways of filing a patent application: a company's internal patent department can undergo the application process itself or the company may delegate it to an external patent attorney. Wagner (2006b) examines make-or-buy decision drivers using negative binomial panel regression. We will consider the three classes illustrated in Figure 2.

Czado et al. (2007, Table 1) gives an overview of all influential variables. For more details see Wagner (2006b, pp. 119-121). We used standard exploratory data analysis tools to investigate main effects and two-dimensional interactions on the mean level. The four strongest two-dimensional interactions were **LN.COV** $*$ **BREADTH**, **CHEM.PHA** $*$ **LN.COV**, **CHEM.PHA** $*$ **SQRT.EMP** and **RDmiss** $*$ **CHEM.PHA**. To find covariates which have a significant influence on the overdispersion parameter, we apply the approach by Czado et al. (2007, Section 5).

A covariate's influence on the overdispersion parameter can be quantified by comparing sample mean to sample variances. For a level j of a categorical covariate w_{it} or a class of a discretized continuous covariate with n_{jt} observations, let δ_{itj} be a dummy indicating if observation $w_{i,t}$ falls in class j, i.e. $\delta_{itj} = 1$ if $w_{i,t} \in$ class j and 0 else. Sample mean and sample variance for j will be $\hat{\mu}_{jt}(\boldsymbol{Y})$ and $\hat{\sigma}^2_{jt}(\boldsymbol{Y})$. For overdispersed GP data we have $\varphi_{it} = \sqrt{\frac{\sigma^2_{it}}{\mu_{it}}}$. Therefore, in a regression context using the shifted log link $\varphi_{it} = 1 + e^{w_{it}\alpha}$, we obtain $w_{it}\alpha = \log\left(\sqrt{\frac{\hat{\sigma}^2_{jt}(\boldsymbol{Y})}{\hat{\mu}_{jt}(\boldsymbol{Y})}} - 1\right) =: \eta_{jt}(\boldsymbol{Y})$. If the data was not overdispersed, mean and variance would coincide and the fraction $\frac{\hat{\sigma}^2_{jt}}{\hat{\mu}_{jt}}$ would be around 1 in every class. The values inside the logarithm would be close to zero. High values, however, indicate overdispersion. A value larger than 0 indicates that the estimated variance exceeds the estimated mean already more than four times. For the covariate **EMP**, the values of $\eta_{jt}(\boldsymbol{Y})$ are plotted in Figure 3. We see that for smaller **EMP** the dispersion is lower whereas for higher values it is high.

As a working correlation matrix for $\mathrm{Corr}(\boldsymbol{Y}_i(\boldsymbol{\beta}))$ we choose $AR(1)$, i.e. $\rho_{tt^*}(\lambda_1) = \lambda_1^{|t-t^*|}$, since the matrix of empirical correlations of residuals based on the model from Czado et al. (2007) strongly suggests that it has this structure. For $\mathrm{Corr}(\boldsymbol{S}_i(\boldsymbol{\delta}))$ we choose the identity matrix

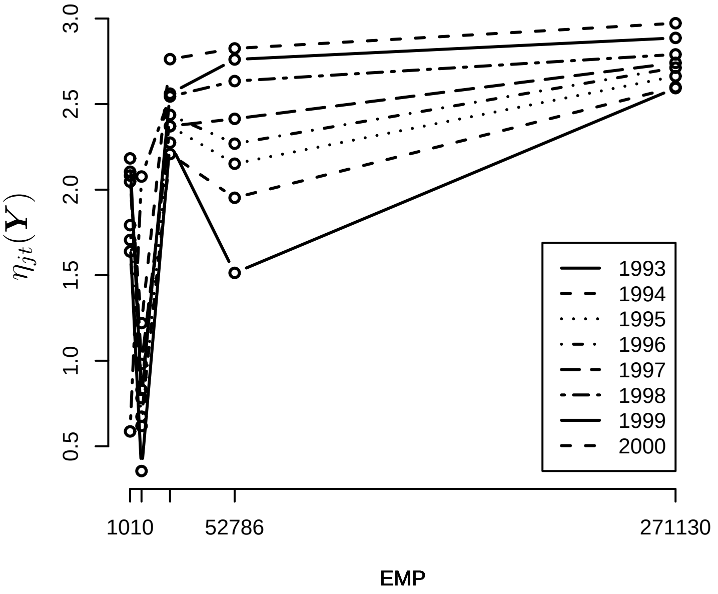

Figure 3: Influence of **EMP** on the overdispersion parameter

$\boldsymbol{I}_{T(T+1)/2}$. For mean regression we select the covariates **1**, **LN.COV**, **BREADTH**, **SQRT.EMP**, **INV.RDP**, **RDE1**, **RDE2**, **RDE3**, **RDmiss**, **CHEM.PHA**, **ELEC.TEL.OTHER**, **YEAR**, **LN.COV** $*$ **BREADTH**, **CHEM.PHA** $*$ **LN.COV**, **CHEM.PHA** $*$ **SQRT.EMP** and **RDmiss** $*$ **CHEM.PHA**. For overdispersion we select **1**, **ENGINEER**, **CAR.SUPP.OTHER**, **MED.BIOT**, **YEAR**, **BREADTH.49.72**, **EMP.11291** and **RDE.63**. All covariates have been centered and standardized for numerical stability. We apply backward selection using the QIC (6.3), i.e. sequentially eliminate the covariate from the full model which decreases the QIC the most (as long as QIC shrinks).

	QIC		Wald-Wolfowitz	
	full	reduced	full	reduced
			W_Z (p) $p+q+1$	W_Z (p) $p+q+1$
(1) $Poi(\mu_{it}, \boldsymbol{R})$	-215813.10	-216270.49	-1.34 (0.18) 17	-2.02 (0.04) 12
(2) $GP(\mu_{it}, \varphi, \boldsymbol{R})$	-2409.02	-2546.67	-0.56 (0.58) 18	-2.53 (0.01) 12
(3) $GP(\mu_{it}, \varphi_{it}, \boldsymbol{R})$	-3945.78	-4581.72	-0.46 (0.64) 25	-0.48 (0.63) 14

Table 2: QIC (see (6.3)) and results of the Wald-Wolfowitz test (see (6.4)) of full and reduced designs for the three model classes specified in Figure 2.

Note that it make only sense to compare the QIC for nested settings. Poisson and GP models are not nested since for a dispersion of 1 in a GP setting an infinitesimal small predictor would be required. 'Full' and 'reduced' models within each model class, however, are nested. Also, the 'full' settings of (2) $GP(\mu_{it}, \varphi, \boldsymbol{R})$ and (3) $GP(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ are nested. The QIC statistics according to (6.3) and the result of the Wald-Wolfowitz test according to (6.4) can be found in Table 2. We report W_Z together with the p-value and the number of parameters. Note that a p-value of more than 5% indicates that one cannot reject H_0 on the 5% level and hence the residuals indicate a good fit. A summary of the resulting model equations is given in Table 3.

Model	Mean	Dispersion	$p + q + 1$
$Poi(\mu_{it}, \boldsymbol{R})$ *full*	offset(**E**) + **1** + **LN.COV** + **BREADTH** + **SQRT.EMP** + **INV.RDP** + **RDE1** + **RDE2** + **RDE3** + **RDmiss** + **CHEM.PHA** + **ELEC.TEL.OTHER** + **YEAR** + **LN.COV.BREADTH** + **CHEM.PHA.LN.COV** + **CHEM.PHA.SQRT.EMP** + **RDmiss.CHEM.PHA**	**1** (not estimated)	17
$Poi(\mu_{it}, \boldsymbol{R})$ *reduced*	offset(**E**) + **1** + **LN.COV** + **BREADTH** + **SQRT.EMP** + **INV.RDP** + **RDmiss** + **CHEM.PHA** + **ELEC.TEL.OTHER** + **YEAR** + **CHEM.PHA.LN.COV** + **RDmiss.CHEM.PHA**	**1** (not estimated)	12
$GP(\mu_{it}, \varphi, \boldsymbol{R})$ *full*	offset(**E**) + **1** + **LN.COV** + **BREADTH** + **SQRT.EMP** + **INV.RDP** + **RDE1** + **RDE2** + **RDE3** + **RDmiss** + **CHEM.PHA** + **ELEC.TEL.OTHER** + **YEAR** + **LN.COV.BREADTH** + **CHEM.PHA.LN.COV** + **CHEM.PHA.SQRT.EMP** + **RDmiss.CHEM.PHA**	**1**	18
$GP(\mu_{it}, \varphi, \boldsymbol{R})$ *reduced*	offset(**E**) + **1** + **LN.COV** + **BREADTH** + **SQRT.EMP** + **INV.RDP** + **RDmiss** + **CHEM.PHA** + **ELEC.TEL.OTHER** + **LN.COV.BREADTH** + **CHEM.PHA.SQRT.EMP**	**1**	12
$GP(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ *full*	offset(**E**) + **1** + **LN.COV** + **BREADTH** + **SQRT.EMP** + **INV.RDP** + **RDE1** + **RDE2** + **RDE3** + **RDmiss** + **CHEM.PHA** + **ELEC.TEL.OTHER** + **YEAR** + **LN.COV.BREADTH** + **CHEM.PHA.LN.COV** + **CHEM.PHA.SQRT.EMP** + **RDmiss.CHEM.PHA**	**1** + **ENGINEER** + **CAR.SUPP.OTHER** + **MED.BIOT** + **YEAR** + **BREADTH.49.72** + **EMP.11291** + **RDE.63**	25
$GP(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ *reduced*	offset(**E**) + **1** + **LN.COV** + **BREADTH** + **SQRT.EMP** + **RDmiss** + **CHEM.PHA** + **ELEC.TEL.OTHER** + **YEAR** + **LN.COV.BREADTH** + **CHEM.PHA.SQRT.EMP**	**1** + **CAR.SUPP.OTHER** + **EMP.11291**	14

Table 3: Model equations of the models shown in Figure 2 using backward selection by QIC (6.3).

For these designs we now discuss the consequences of our suggested enhancements.

Adding a dispersion parameter

Adding a dispersion parameter to the Poisson setup has a positive impact on model fit. Comparing (1) $Poi(\mu_{it}, \boldsymbol{R})$ to (2) $GP(\mu_{it}, \varphi, \boldsymbol{R})$, the p-value for rejecting H_0 in the Wald-Wolfowitz test increases from 0.18 to 0.58 in the full settings. In the reduced settings, however, both setups having the same number of parameters show no good fit on the 5% level.

Regression on the dispersion parameter

Comparing model (2) $GP(\mu_{it}, \varphi, \boldsymbol{R})$ to (3) $GP(\mu_{it}, \varphi_{it}, \boldsymbol{R})$, the p-values of the Wald-Wolfowitz run test increases from 0.58 to 0.64 (full settings) and from 0.01 to 0.63 (reduced settings). This indicates the usefulness of allowing for regression on the dispersion parameter. Since the full settings of these two models are nested, the QIC can be used for model comparison here as well. There is a large decrease from -2409.02 to -3945.78, which reinforces the conclusion from above.

In terms of the Wald-Wolfowitz test, the full model (3) $GP(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ is to be preferred over all other classes discussed (see Table 2). However, this goodness-of-fit criterion does not account for the model complexity. Since the reduced design for (3) $GP(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ shows a comparable test result (p-value of 0.63 instead of 0.64) but has only 14 parameters instead of 25 we choose this setup to be our best.

7.2 Model interpretation

The paper by Czado et al. (2007) considers a zero-inflated generalized Poisson regression model (among others) for this data. In the context of GEE, however, we will not consider zero-inflation. We encountered numerical problems when fitting a ZIGP specification with the means of GEE and learned that the model flexibility of this distribution is too large.

In the ZIGP model of Czado et al. (2007) the observation year is allowed to be included as a covariate and is found to be highly significant in the dispersion level. Hence the autocorrelation of the ZIGP residuals was very low. Due to this fact and due to the different distributional assumptions, we stress that these two models cannot be compared with respect to the panel correlation. However, as for the regression designs on the mean and dispersion levels which we found to be most suitable, both models have a great deal in common. There is a detailed graphical evaluation of the ZIGP model in Czado et al. (2007).

We will now briefly interpret the reduced setup (3) $GP(\mu_{it}, \varphi_{it}, \boldsymbol{R})$. Note that some of the covariates in the $GP(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ setup in Table 4 are insignificant according to the Wald test.

		Estimate	Std. Error	z-value	$Pr(>\|z\|)$
	μ REGRESSION				
b0	**1**	-1.174	0.141	-8.344	$< 2 \cdot 10^{-16}$
b1	**LN.COV**	0.036	0.035	1.035	0.301
b2	**BREADTH**	0.043	0.029	1.459	0.145
b3	**SQRT.EMP**	-0.222	0.050	-4.444	$8.8 \cdot 10^{-6}$
b4	**RDmiss**	0.004	0.046	0.076	0.939
b5	**CHEM.PHA**	-0.403	0.384	-1.048	0.295
b6	**ELEC.TEL.OTHER**	0.504	0.157	3.198	0.001
b7	**YEAR**	0.067	0.033	2.046	0.041
b8	**LN.COV.BREADTH**	-0.002	0.028	-0.073	0.942
b9	**CHEM.PHA.SQRT.EMP**	0.286	0.374	0.765	0.444
	φ REGRESSION				
a0	**1**	2.346	0.087	26.906	$< 2 \cdot 10^{-16}$
a1	**CAR.SUPP.OTHER**	-0.961	0.099	-9.706	$< 2 \cdot 10^{-16}$
a2	**EMP.11291**	-1.096	0.106	-10.360	$< 2 \cdot 10^{-16}$
	CORRELATION				
	γ	1.499	0.188	7.963	$1.7 \cdot 10^{-15}$

QIC	-4581.72	
Range $\boldsymbol{\mu}$	[0.22,	568.44]
Range $\boldsymbol{\varphi}$	[2.33,	11.44]
$\hat{\lambda}_1(\gamma)$	0.90	

Table 4: Summary of the fitted $GP(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ model obtained by backward selection using QIC

In contrast to the ZIGP model mentioned, variable selection has been done using backward selection with respect to QIC. In the ZIGP model, **RDmiss** is insignificant and therefore does not appear in the final model instead of AIC. This is a desirable result since **RDmiss** is a dummy for missing R & D data. In our GP model, **RDmiss** still appears, it is, however, insignificant according to Wald: the p-value is 93.9% (Table 4). Obviously, the lack of modeling zero-inflation in the GP model is reflected in the higher overdispersion range of $[2.20, 11.04]$ as compared to $[2.41, 10.15]$ in the ZIGP model. Also, there is an additional interaction between the Chemical / Pharmaceutical industry dummy and the square root of the number of employees **SQRT.EMP**. Further, the observation year remains in the mean design. On the other hand, **RDE1** and **INV.RDP** are not appearing any longer. On dispersion level, the engineering industry dummy as well as **YEAR** and **RDE.63** are eliminated in addition to effects already taken out of the ZIGP model. Neglecting correlation between the counts within each subject leads to an underestimation of the predicted variances of the parameter estimates. Thus, the z values calculated tend to be too large and therefore effects may be regarded as significant although they are not. Comfortingly, the signs of the coefficients of common covariates in both models compared do not change. Hence there is no turnaround in how means and dispersion are affected by the chosen descriptive variables. Similar to Czado et al. (2007) we will look at patent outsourcing rates for the interpretation. In order to obtain outsourcing rates as functions of the covariates, we will fix the exposure by its mode $E^M = 13.36$. Then, we can define functions $\frac{\hat{\mu}(x_k)}{E^M}$, where x_k is the kth covariate. All other covariates will be fixed by their mode as well, where for interacting covariates, their common mode will be used. Since there is an additional interaction between **CHEM.PHA** and **SQRT.EMP** as compared to the ZIGP fit, we will look at the influence of **EMP** on the outsourcing rate in Figure 4 (1) since here there might crop up a considerable difference of **EMP**'s influence on the outsourcing rates. For the Chemical / Pharmaceutical industry, the interaction leads to an inverted influence of the number of employees (compare to Czado et al. (2007, Figure 4 (1))). While in all remaining industries large companies in terms of employees tend to have their own patent departments, large Chemical / Pharmaceutical companies are likely to contract out. As one can see in Figure 4 (2), there is a positive time trend. The share of outsourced patent applications was increasing in all industries. This reflects the general tendency to decrease economic risk by the outsourcing of services in recent years.

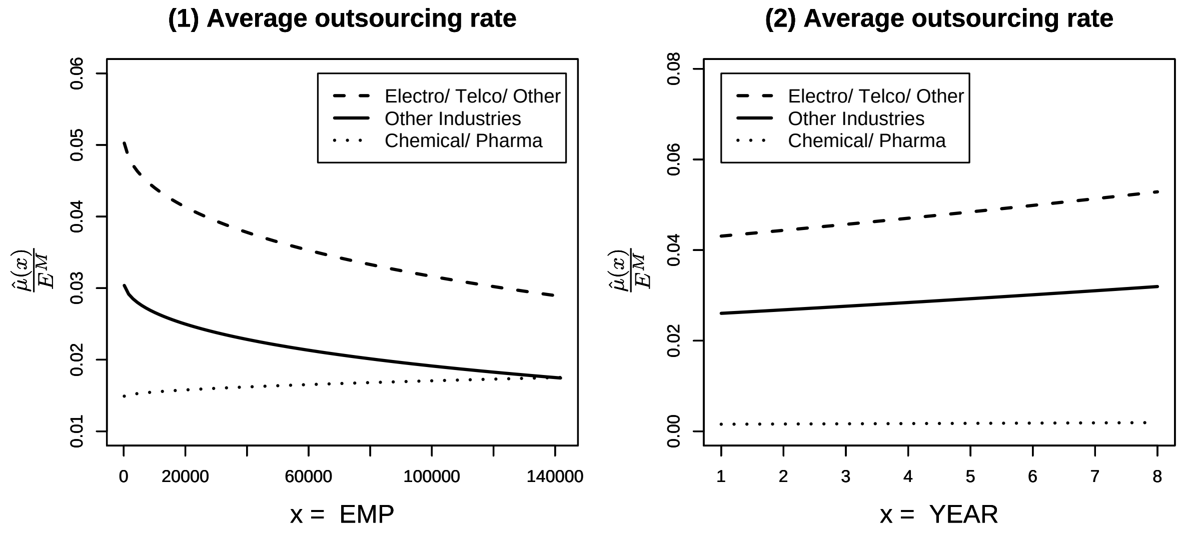

Figure 4: Influence of **EMP** and **YEAR** on the outsourcing rate while fixing other covariates by their empirical modes

We define the overdispersion factor of a random variable $Y_{it} \sim GP(\mu_{it}, \varphi_{it})$ as $V_{it} := \frac{\text{Var}(Y_{it})}{E(Y_{it})} = \varphi_{it}^2$. There are only categorical covariates for overdispersion: $\boldsymbol{w}$:=(**1**, **CAR.SUPP.OTHER**, **EMP.11291**). Using (3.4), we define $\hat{\varphi}(\boldsymbol{w}) := 1 + \exp\Big(\hat{\alpha}_0 + w_1 \cdot \hat{\alpha}_1 + w_2 \cdot \hat{\alpha}_2\Big)$. We use this overdispersion function to estimate $V(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}) = \varphi^2$ by $\hat{V}(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}) := \hat{\varphi}(\boldsymbol{w})^2$. Table 5 lists $\hat{V}(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w})$ depending on the settings arising from the categorical dispersion designs. As in Czado et al. (2007, Table 6), companies in the Cars / Suppliers / 'Others' sector are predicted to have lower overdispersion than companies in other industries. Large companies show higher overdispersion, which in line with the ZIGP model as well.

Industry	Employees	$\hat{V}(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w})$
Cars / Suppl. / Other	$\geq$ 11 291	24.9
Cars / Suppl. / Other	$\leq$ 11 291	5.5
Remaining industries	$\geq$ 11 291	130.9
Remaining industries	$\leq$ 11 291	20.2

Table 5: Estimated overdispersion factor in the 'best' model (Table 4) depending on categorical overdispersion covariates

For more graphical evaluations, for example the effect of the interacting covariates **LN.COV** and **BREADTH** on the outsourcing rate, see Czado et al. (2007).

8 Conclusions and Discussions

We introduced a $GP(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ setup for longitudinal count data, which not only extends the known Poisson GEE by overdispersion but also allows for regression on this parameter. We estimate variances of empirical covariances by a log normal regression model using a data designed grid. This grid can be adjusted when other data sets are considered.

We carried out a comparison of different setups extending Poisson GEE using data dealing with the determinants of patent outsourcing. We illustrated that every extension incorporated in our $GP(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ setup improved model fit in terms of the QIC for nested comparisons and the Wald-Wolfowitz run test for assessing the goodness-of-fit. Both QIC and the Wald-Wolfowitz test chose the introduced $GP(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ setup as the one fitting our data best.

A short model interpretation confirmed insights of former work on the given data from an economic point of view. We added some analytical and economic interpretation for mean and overdispersion drivers in our 'best' model. The correlation between outcomes of two subsequent years is estimated to be 90%.

It would be interesting to compare the GEE approach to other estimating techniques such as MCMC, maximization by parts or composite likelihood. Also, including zero-inflation in these models will be subject of further research.

Acknowledgement

We thank Stefan Wagner from INNO-tec for making the data available to us. V. Erhardt is supported by a grant from Allianz Deutschland AG. C. Czado is supported by DFG (German Science Foundation) grant CZ 86/1-3.

References

H. Akaike (1974). 'A new look at the statistical model identification'. *IEEE Trans. Automat. Control* **19**:716–723.

Y.-C. Chang (2000). 'Residuals analysis of the generalized linear models for longitudinal data'. *Statistics in Medicine* **19**(10):1277–1293.

K. Clarke (2007). 'A Simple Distribution-Free Test for Nonnested Model Selection'. *Political Analysis 2007* **15**(3):347–363.

P. C. Consul (1989). *Generalized Poisson distributions*, vol. 99 of *Statistics: Textbooks and Monographs*. Marcel Dekker Inc., New York. Properties and applications.

P. C. Consul & F. Famoye (1992). 'Generalized Poisson regression model'. *Comm. Statist. Theory Methods* **21**(1):89–109.

P. C. Consul & G. C. Jain (1970). 'On the generalization of Poisson distribution'. *Annals of Mathematical Statistics* **41**(4):1387.

C. Czado, et al. (2007). 'Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates'. *Statistical Modelling* **7**(2):125–153.

V. Erhardt (2009). 'ZIGP: Zero-Inflated Generalized Poisson (ZIGP) Models' R package version 3.6.

V. Erhardt & C. Czado (2009). 'A method for approximately sampling high-dimensional count variables with prespecified Pearson correlation'. *Submitted for publication* Preprint available at *http://www-m4.ma.tum.de/Papers/index.html*.

F. Famoye (1993). 'Restricted generalized Poisson regression model'. *Comm. Statist. Theory Methods* **22**(5):1335–1354.

F. Famoye (1997). 'Generalized poisson random variate generation'. *Amer. J. Math. Management Sci.* **17**(3-4):219–237.

F. Famoye, et al. (2004). 'On the Generalized Poisson Regression Model with an Application to Accident Data'. *Journal of Data Science* **2**:287–295.

F. Famoye & K. P. Singh (2003). 'On inflated generalized Poisson regression models'. *Adv. and Appl. Stat.* **3**(2):145–158.

R. Fisher (1921). 'On the 'probable error' of a coefficient of correlation deduced from a small sample'. *Metron* **1**:3–32.

S. Gschlößl & C. Czado (2008). 'Modelling count data with overdispersion and spatial effects'. *Statist. Papers* **49**(3):531–552.

J. M. Hilbe (2007). *Negative Binomial Regression*, vol. 1. Cambridge University Press, New York.

K.-Y. Liang & S. L. Zeger (1986). 'Longitudinal data analysis using generalized linear models.'. *Biometrika* **73**:13–22.

P. McCullagh & J. Nelder (1989). *Generalized linear models*. Chapman & Hall, London, second edn.

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

R. L. Prentice & L. P. Zhao (1991). 'Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses.'. *Biometrics* **47**(3):825–839.

D. P. M. Scollnik (1995). 'Bayesian Analysis of Two Overdispersed Poisson Models'. *Biometrics* **51**(3):1117–1126.

StataCorp (2007). 'Stata Statistical Software: Release 10'. College Station, TX: StataCorp LP.

D. Stekeler (2004). 'Verallgemeinerte Poissonregression und daraus abgeleitete Zero-Inflated und Zero-Hurdle Regressionsmodelle'. Master's thesis, Technische Universität München (*www-m4.ma.tum.de/Diplarb/*).

R. C. Tripathi & R. C. Gupta (1984). 'Statistical Inference regarding the Generalized Poisson Distribution'. *Sankhya, Series B* **46**(2):166–173.

R. Vernic (2000). 'A multivariate generalization of the generalized Poisson distribution'. *Astin Bull.* **30**(1):57–67.

Q. H. Vuong (1989). 'Likelihood ratio tests for model selection and nonnested hypotheses'. *Econometrica* **57**(2):307–333.

S. Wagner (2006a). ‘Make-or-Buy Decisions in Patent Related Services’. *Münchener Wirtschaftswissenschaftliche Beiträge (VWL) 2006-16, http://epub.ub.uni-muenchen.de/archive/00001264/* .

S. M. Wagner (2006b). *Economic Analyses of the European Patent System*, vol. 1. Deutscher Universitätsverlag.

J. Yan (2002). ‘geepack: Yet Another Package for Generalized Estimating Equations’. *R-News* **2/3**:12–14.

J. Yan & J. P. Fine (2004). ‘Estimating Equations for Association Structures’. *Stat. in Med.* **23**:859–880.