Technische Universität München
Zentrum Mathematik
Lehrstuhl für Mathematische Statistik

Modeling different dependence structures involving count data with applications to insurance, economics and genetics

Vinzenz Martin Erhardt

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. Rudi Zagst
Prüfer der Dissertation:	1.	Univ.-Prof. Claudia Czado, Ph.D.
	2.	Univ.-Prof. Dr. Ludwig Fahrmeir Ludwig-Maximilians-Universität München
	3.	Prof. Arnoldo Frigessi University of Oslo, Norwegen (nur schriftliche Beurteilung)

Die Dissertation wurde am 01.04.2010 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 14.06.2010 angenommen.

Zusammenfassung

In dieser Arbeit werden etliche Abhängigkeitsstrukturen für Zählvariablen, aber auch stetige Zielvariablen, untersucht. Diese Zählvariablen weisen typischerweise nicht nur Überdispersion auf, sondern haben auch einen hohen Anteil an Nullen; zwei Eigenschaften, die kaum von klassischen Verteilungen erklärt werden können. Regressionsmodelle für abhängige beschreibende Variablen werden ebenfalls untersucht. In einer Anwendung aus der Genetik werden verschiedene Ansätze verglichen, um mittels "QTL mapping" auf dem Genom nach signifikanten Regionen zu suchen, die ursächlich für bestimmte Phänotypen sind. Dabei werden überraschende Einblicke in die Ursachen von Überdispersion präsentiert. Zeitliche Abhängigkeit wird im Kontext von "generalized estimating equations" für verallgemeinerte Poisson Zielvariablen betrachtet. Damit soll das Outsourcingverhalten von Patentanmeldungen von 107 Firmen über acht Jahre beschrieben werden. Für die Jahresgesamtschäden in der Versicherung wird ein Abhängigkeitsmodell basierend auf Pair-Copula-Konstruktionen entwickelt. Die Herausforderung bei diesem Problem liegt darin, daß die Versicherungsschäden aus einigen der abhängigen Marginalien Null sein können, die marginalen Schadenhöhenverteilungen daher nicht in das klassische Copula-Konzept passen. Pair-Copula-Konstruktionen sind deshalb sehr attraktiv, da sie erlauben, eine hochdimensionale Dichtefunktion als Produkt bivariater Copulas und marginaler Dichten zu definieren. Zuletzt wird ein Verfahren zur Erzeugung hochdimensionaler Zählvariablen mit vorab spezifizierter Pearson-Korrelation entwickelt. Dieser neue Ansatz basiert ebenfalls auf Pair-Copula-Konstruktionen und hat eine höhere Genauigkeit als ein bekannter Vergleichs-Ansatz.

Abstract

In this thesis, several dependence structures for dependent count responses and continuous responses will be investigated. These count variables are typically not only overdispersed but also show a large share of zero observations which cannot be described by classical distributions. Therefore, zero-inflated generalized Poisson count regression and other regression models will be considered. Dependence in the responses as well as in the describing variables will be considered. In an application to genetics several methods of searching for causal genome regions for a certain trait will be compared. Surprising insights on another source of overdispersion will be presented. Temporal dependence will be addressed in the context of generalized estimating equations for generalized Poisson responses. We apply this approach to fit models for the outsourcing behavior of patent applications processes of 107 companies over eight years. In the field of dependent insurance claim totals, a dependence model based on pair-copula constructions will be developed. The challenge of this problem is that the insurance claims of some of the dependent margins may be zero, and a marginal claim size distribution will therefore not fit in the general framework of copula modeling. Pair-copula applications are especially appealing since they allow to define a high dimensional density function by a product of bivariate copulas and marginal densities. Finally this thesis will deal with an input modeling problem: a method for sampling from high-dimensional count random vectors with a specified Pearson correlation will be developed. For this challenging problem a novel approach also based on pair-copula constructions will be developed and prove to outperform a well-known benchmark approach. Software packages for *R* related to many of the topics have been developed.

Acknowledgment

I am greatly indebted to Prof. Claudia Czado for the perpetual and intensive supervision. This thesis has gained a lot from many fruitful discussions, her astute analysis and ongoing advice over the past four years. Likewise, I am very grateful for her encouragement to participate in many scientific conferences and to exchange with esteemed scientists.

It is a particular pleasure for me to thank Prof. Małgorzata Bogdan for the very fruitful collaboration and the many valuable debates. Also I would like to thank Prof. Ludwig Fahrmeir and Prof. Arnoldo Frigessi for acting as referees of this thesis. I would like to thank my colleagues at the Technische Universität München for a pleasant time during the last years.

Moreover, I want to express my explicit gratitude to Allianz Deutschland AG for the financial support. I would also like to thank my former supervisor at Allianz, Pierre Joos, both for giving me credit and for supporting me in the first year of the thesis. I am grateful to Dr. Beate Elfinger for helpful discussions and to Dr. Florian Beigel and his colleagues for giving me valuable advice.

Last but not least I am very grateful for the love of my family and especially of my wife Christine, and their support and patience throughout the years.

Contents

Introduction

> *"But he does not wear any clothes" said the little child in Hans Christian Andersen's "The Emperors's New Clothes."*
> Thomas Mikosch (2006) on copulas.

> *It is my personal belief that over the years to come, research will be able to put further garments on the poor man so that eventually in Hans Christian Andersen's words we can truly say "Goodness! How well they suit your Majesty! What a wonderful fit! What a cut! What colors! What sumptuous robes!".*
> Paul Embrechts (2009).

In the famous article by Mikosch (2006), Mikosch bashes the concept of copulas and continues to say that he wonders why more and more of his colleagues became immersed in copulas: "I suspect that some include the word *copula* in the title of their papers not because they contribute to the theory on copulas, but because they believe that one can publish easier."

Four years after Mikosch' article, nobody doubts that the concept of copulas is and will be one of the key concepts in dependence modeling. While a discussion of some of his criticism, such as questioning the justification of having uniform margins, has not been widely perpetuated, his article may be seen as an appeal for a correct and well-reflected utilization of copulas. David X. Li's pioneering model for the pricing of collateralized debt obligations (CDOs) proved to be an example of a model being adopted by the financial sector with too little reflection: the Gaussian copula model turned out to be over-simplistic for quantifying the dependence of risks. Mikosch (2006) and Embrechts (2009) name as major challenges in copula modeling

- copulas with discrete margins,
- difficulties arising in high dimension,
- the choice of a suitable copula class and
- difficulties in applying copulas in time series analysis.

This thesis will deal will dependent data analysis. Despite its focus on dependent data, it will not exclusively deal with copula models; the model formulation in the single chapters will rather be driven by the problem at hand. A major emphasis of the thesis will be on overdispersed

count data with many zeros. This thesis will touch several of the issues raised by Mikosch and Embrechts: it will deal with temporally clustered data, for which we will indeed not use a copula model. In fact, a specification based on generalized estimating equations will be developed. The difficulty of dealing with copulas with discrete margins will be addressed. In particular, a sampling approach for discrete data as well as trivariate copula models with binary margins will be developed. The problem of dimensionality will in general be addressed by using the concept of pair copula constructions. They allow to construct high-dimensional density functions based on a cascade of bivariate copulas. They also allow to cancel out from this construction all pairs of conditional margins who are close to independence, hence reducing the dimensionality of the model. Copula choice is addressed in the context of pair copula constructions as well as classical copula models. Graphical tools will be used and a choice criterion based on tests for non nested model selection will be illustrated.

There are multivariate representatives of several count distributions, such as a multivariate Poisson distribution (Kawamura (1979) or Karlis and Meligkotsidou (2005)), the multivariate Negative Binomial (Kopociński (1999)) or a multivariate generalization (Vernic (2000)) of the generalized Poisson distribution (Consul and Jain (1970)). They have, however, several shortcomings. Most of them either allow for exchangeable covariance structures only or the modeling of negative correlation is not possible. Due to the difficulty in calculating the required probabilities, their usefulness is limited in dimensions larger than two.

As starting point of this thesis consider the paper by Czado, Erhardt, Min, and Wagner (2007) which is based on my diploma thesis. In this paper the authors develop a flexible zero-inflated generalized Poisson regression model, which allows for regression effects not only on the mean but also on the dispersion and zero-inflation level.

In general there are two natural extensions of the Poisson distribution, which allow for modeling a difference between the mean and the variance: the Negative Binomial (or Poisson-Gamma) distribution and the generalized Poisson distribution. The generalized Poisson distribution $GP(\mu, \varphi)$ was first introduced by Consul and Jain (1970) and subsequently studied in detail by Consul (1989). We refer to its mean parametrization (see e.g. Consul and Famoye (1992)):

$$\text{for} \quad y \in \{0, 1, \ldots\} \quad P_{GP}(Y = y|\ \mu, \varphi) = \frac{\mu(\mu + (\varphi - 1)y)^{y-1}}{y!} \varphi^{-y} e^{-\frac{1}{\varphi}(\mu + (\varphi-1)y)},$$

where μ and φ are larger than 0. In the case of underdispersion ($\varphi < 1$), another additional condition needs to be fulfilled, i.e., $\varphi > \max(\frac{1}{2}, 1 - \frac{\mu}{m})$ where m is the largest natural number with $\mu + m(\varphi - 1) > 0$. The GP distribution does not belong to the exponential family even if the dispersion parameter φ is known. For $Y \sim GP(\mu, \varphi)$ mean and variance are given $E(Y) = \mu$ and $Var(Y) = \varphi^2 \mu$. This allows for modeling over- or underdispersion. However, in the case of underdispersion ($\varphi \in (0, 1)$), the support of the distribution depends on μ and φ, which is difficult to enforce when μ and φ need to be estimated. Therefore throughout the thesis we restrict to equi- and overdispersion, i.e., $\varphi \geq 1$.

When comparing to the Negative Binomial (NB) distribution, the GP distribution has several advantages. While the NB distribution with pmf

$$P(Y = y|\mu, \Psi) = \frac{\Gamma(y + \Psi)}{\Gamma(\Psi) y!} \left(\frac{\Psi}{\mu + \Psi}\right)^{\Psi} \left(\frac{\mu}{\mu + \Psi}\right)^{y},$$

and $E(Y) = \mu$, $Var(Y) = \mu(1 + \frac{\mu}{\Psi})$ contains the basic Poisson distribution only as a limiting case for $\Psi \to \infty$, the GP distribution contains the Poisson class for $\varphi = 1$. Second, unlike

the NB distribution the dispersion factor in GP is independent of the mean. Hence, in the NB distribution the statistical modeling of overdispersion is less transparent than in case of the GP. For a detailed comparison between GP and NB we refer the readers to Joe and Zhu (2005).

A zero-inflated generalized Poisson (ZIGP) distribution is a further extension of the GP distribution, which allows to model additional probability mass at zero. The ZIGP distribution is defined as a mixture of a distribution concentrated at 0 and the generalized Poisson distribution:

$$P_{ZIGP}(Y = y|\mu, \varphi, \omega) = \omega + (1 - \omega)\ P_{GP}(Y = y|\mu, \varphi),$$

where $\omega \in [0, 1]$ is the zero-inflation parameter. Mean and variance of $Y \sim ZIGP(\mu, \varphi, \omega)$ are given by

$$E(Y) = (1 - \omega)\mu \text{ and } Var(Y) = E(Y)\left(\varphi^2 + \mu\omega\right).$$

In some data sets a constant overdispersion and/or constant zero-inflation parameter might be too restrictive. Therefore, Czado, Erhardt, Min, and Wagner (2007) extend the ZIGP regression model of Famoye and Singh (2003) by allowing for regression on φ and ω and develop a $ZIGPR(\mu_i, \varphi_i, \omega_i)$ regression model. Most of the chapters of this thesis are based on the $ZIGPR(\mu_i, \varphi, \omega)$ regression model of Famoye and Singh (2003) which does not include additional regression effects on the dispersion and zero-inflation parameters. Generalized Poisson regression $GPR(\mu_i, \varphi)$ and zero-inflated Poisson regression $ZIPR(\mu_i, \omega)$ can be defined accordingly by assuming the responses in the regression model to follow a $GP(\mu_i, \varphi)$ or a $ZIP(\mu_i, \omega)$ distribution, respectively. Additionally, generalized Poisson regression specification $GPR(\mu_{it}, \varphi_{it})$ for temporally clustered data will be investigated. It will allow for regression on the dispersion parameter for which similar as in Czado et al. (2007) a shifted log link will be used. In order to specify appropriate regression models for the overdispersion and zero-inflation parameters, (Czado, Erhardt, Min, and Wagner 2007, Section 5) develop tools for an exploratory data analysis. The software for such models including exploratory tools has been implemented in the *R* package *"ZIGP"* available on CRAN.

Similar to generalized linear models all observations are assumed to be independent in these regression models. The data may, however, be multivariate, or temporally or spatially clustered. For the patent data investigated by Czado et al. (2007), the time series character is accounted for by allowing the regression designs to depend on the observation year. If however the temporal dependence is not completely captured by this specification, this might result in wrong parameter standard errors and hence model misspecification, i.e., some of the regressors in the model might actually be insignificant.

Multivariate analysis will be the central theme of this thesis. We extend the work by Czado, Erhardt, Min, and Wagner (2007) by investigating different dependence structures, i.e.,

- regression
- temporal dependence
- spatial dependence and
- multivariate models.

In what follows a general outline of the contribution of this thesis to the described topics will be given.

In **Chapter 1** we investigate a problem in genetics, i.e., the search for the location of significant genotype locations for certain disease traits. Such data is rather extreme since the number of covariates is a lot higher than the sample size and covariates are mutually correlated. This chapter is based on Erhardt, Bogdan, and Czado (2010). We consider the problem of locating multiple interacting quantitative trait loci (QTL) influencing traits which are measured in counts. In many applications the distribution of the count variable has a spike at zero, which e.g. may correspond to the absence of certain disease symptoms. Zero-inflated generalized Poisson regression (ZIGPR) allows for an additional probability mass at zero and hence allows for an improvement in model fit and the detection of significant trait loci. It has already been used successfully to locate QTL with interval mapping. ZIGPR can be also used to locate several interacting QTL. The most difficult part in this process is the estimation of the QTL number. As discussed in the literature of normal traits, the classical model selection criteria often have a strong tendency to overestimate the QTL number. To solve this problem modified versions of the Bayesian Information Criterion (mBIC and EBIC) were proposed and successfully used for QTL mapping. We apply these criteria for locating QTL based on ZIGPR as well as simpler models. We present the results of an extensive simulation study, which shows its good power of QTL detection, while controlling the false discovery rate at a reasonable level. The study also clearly demonstrates the advantage of using ZIGPR over some simpler statistical models. Another important finding is that the standard Poisson regression is not suited for QTL mapping since the number of QTL is dramatically overestimated. This behavior can be attributed to the inability of the Poisson regression to account for data over-dispersion and strongly discourages from its application for identifying factors influencing count data. The proposed method of QTL detection based on ZIGPR is used to analyze the mice gallstone data of Lyons et al. (2003), who investigate among other traits the highly zero-inflated number of gallstones, which a population of mice developed. In comparison to their analysis our results suggest the existence of a novel QTL on chromosome 4, which influences the number of gallstones by interaction interacting with another QTL, previously identified on chromosome 5. The *R* software is available from *www-m4.ma.tum.de/Papers/Erhardt/qtl-zigp-code.rar*.

Chapter 2 is based on Erhardt and Czado (2009a) and Erhardt and Czado (2009c) and deals with the problem of sampling of such count random vectors which have a certain specified Pearson correlation. We sample from dependent vectors using pair-copula constructions (PCC) and use C-vines, a graphical tool to organize such PCC. A major task is to determine the appropriate copula parameters to obtain the specified target correlation. We will introduce a sequential and very fast root finding routine to approximate them using bisection: for dimension T, $T(T-1)/2$ simple root finding step have to be carried out, usually done within seconds. Further we will illustrate that our sampling approach generates accurate results even in high dimensions and for relatively small sample sizes in several settings with Poisson, generalized Poisson, zero-inflated generalized Poisson and Negative Binomial margins in a variety of settings. We compare it to a simple "naive" sampling method and the NORTA (NORmal-To-Anything) method for discrete margins and illustrate that these methods are less accurate since the empirical correlation of the sample has a higher absolute deviation from the desired target correlation. Moreover, the input parameters obtained by NORTA may need a posteriori correction in order to result in a positive definite base correlation matrix, which is not necessary in our approach. The software is implemented in the *R* package *"corcounts"*.

A simulation study based on this sampling method will be included in **Chapter 3**. This chapter is based on Erhardt and Czado (2009b) and deals with temporal dependence. It may be seen as an extension of the model in Czado, Erhardt, Min, and Wagner (2007). The data consist of a time series of responses and explanatory variables of eight consequent years. We consider a specification based on generalized estimating equations (GEE, Liang and Zeger (1986)) which fit parameters based on sums of weighted residuals and may be applied for example to the Poisson distribution. We discuss generalized Poisson response data. Despite some advantages over the negative binomial distribution, this distribution has not been considered in the context of GEE. To fit the additional dispersion parameter of the GP distribution, second level estimating equations based on covariance residuals (Prentice and Zhao (1991)) 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. The specification used in this chapter allows for regression on the dispersion parameter of the GP distribution, i.e., the dispersion parameters may vary with covariates thus allowing to identify covariate combinations where one finds large and small overdispersion effects. In the real data example dealing with the outsourcing of patent filing processes, Czado et al. (2007) used time as a regressor in the model in an effort to obtain independent residuals. Exploratory data analysis tools developed in this paper are utilized to choose regression for the dispersion parameters. In extension of Czado et al. (2007), the GEE specification will allow to actually quantify the temporal dependence. For the given data, the GEE approach will outperform longitudinal Poisson regression and GP setups with only constant dispersion parameter.

Count regression models for spatially clustered data will be investigated in **Chapter 4** (based on Czado, Schabenberger, and Erhardt (2009)). The real data example is based on the number of ambulant treatments a patient received within the private health care system. We examine not only the Poisson distribution but also the generalized Poisson, the negative Binomial as well as the zero-inflated Poisson distribution as possible response distribution. We add random spatial effects for modeling spatial dependency and develop and implement MCMC algorithms for Bayesian estimation. In an application the presented models are used to analyze the number of benefits received per insured person in a German private health insurance company. Especially for health insurance benefits there is a significant spatial pattern in the utilization, i.e., urban versus rural areas or East-West differences. Model comparison between various non nested model classes is non standard. We utilize a test proposed by Vuong (1989) and the distribution-free test proposed by Clarke (2007) for non nested model comparison and illustrate how they may be applied in a Bayesian context. This is a novel approach since so far these two tests have only been used in classical estimation. Also, the comparison between spatial covariate and / or spatial effect specifications for count regression data has not been carried out elsewhere. The software is implemented in the *R* package *"spatcounts"*.

The insurance data investigated in Chapter 4 also includes inpatient and dental treatments as well as claim sizes for the same portfolio of insured persons. Health insurance claims in these fields are likely to be highly dependent since they will be influenced by the health status and age of the insured person. For this data we develop a joint model of the yearly claim totals based on pair copula constructions in **Chapter 5** (based on Erhardt and Czado (2010)). In many insurance applications there are numerous claim totals which are zero. However, the modeling of zero claims together with non-zero claims is often not necessary, since marginally one is interested in claim totals given that at least one claim occurred. On the other hand, as

soon as dependent claims in different coverage fields occur, one needs to consider all years for which in at least one policy field a claim occurred. A marginal claim distribution will then have an additional point mass at zero, hence this probability function will not be continuous at zero and the cdfs will not be uniform. Therefore using a copula approach to model dependency is not straightforward. We present a novel approach of modeling the joint density of total claims in the presence of many zero observations based on copulas for binary and continuous margins. We illustrate how pair copula constructions under marginals can be utilized for this problem. The zero claim events in this model will be discrete binaries which may be dependent. Since the dimension will only be three, we may explicitly write down and fit the joint probability obtained by a copula for discrete margins. Copula choice for such discrete margins is carried out by a novel approach.

Chapter 1

Locating multiple interacting quantitative trait loci with the zero-inflated generalized Poisson regression

1.1 Introduction

Despite a long history of QTL mapping (see e.g. Sax (1923)) this research field is still a very active area in which perpetually new statistical methodologies are developed. The majority of methods proposed in the literature, like classical interval mapping (Lander and Botstein (1989) and Haley and Knott (1992)), composite interval mapping (Zeng (1993), Zeng (1994)), multiple QTL mapping (Jansen (1993) and Jansen and Stam (1994)) or multiple interval mapping (Kao, Zeng, and Teasdale (1999)) are designed for the situation when the trait has a normal distribution. Since in many practical cases this assumption is violated, we observe lately a considerable effort to develop new methods, which could handle other trait distribution types. In this context we mention recent articles on the analysis of ordinal traits (see e.g., Yi, Xu, George, and Allison (2004), Yi, Banerjee, Pomp, and Yandell (2007), Coffman, Doerge, Simonsen, Nichols, and Duarte (2005) or Li, Wang, and Zeng (2006)), nonparametric methods based on ranks (see e.g., Kruglyak and Lander (1995), Zou, Yandell, and Fine (2003) or Zak, Baierl, Bogdan, and Futschik (2007)), extension of multiple interval mapping to generalized linear models Chen and Liu (2009) or specific methods which can handle a "spike" in the trait distribution (see e.g., Broman (2003) and Li and Chen (2009)). In case the trait is a count variable it often occurs that it has a "spike" at zero. A clear example of such a phenomenon is provided by the gallstone data of Lyons et al. (2003), where the number of gallstones is considered and a large proportion of mice did not develop any disease symptoms. As illustrated by Cui and Yang (2009), such data can be efficiently modeled using the zero-inflated generalized Poisson regression (ZIGPR, Famoye and Singh (2003)). In contrast to the generalized Poisson regression ZIGPR allows for excess zeros, which may be due to other than genetic reasons. The simulations and the real data analysis reported in Cui and Yang (2009) show that interval mapping based on ZIGPR can efficiently locate QTL influencing the count traits. Cui and Yang (2009) also suggest to apply ZIGPR in order to locate several interacting QTL, based on the multiple interval mapping

approach.

From the statistical point of view the most difficult part in fitting the multiple regression model lies in the estimation of the number of significant predictors. As discussed in Broman and Speed (2002) and Bogdan, Ghosh, and Doerge (2004), the classical model selection criteria have a strong tendency to overestimate the number of QTL when the number of markers is comparable to the sample size n. These experimental observations were confirmed by theoretical results in Bogdan, Ghosh, and Żak-Szatkowska (2008) and Chen and Chen (2008), which show that the classical Bayesian Information Criterion (BIC, Schwarz 1978) is not consistent when the number of potential regressors increases to infinity quicker then $\sqrt{n}$. To correct for this behavior of the BIC, several modifications of this criterion were proposed in the literature (e.g. see Ball (2001), Bogdan, Ghosh, and Doerge (2004), Manichaikul, Moon, Sen, Yandell, and Broman (2009)). Specifically, Bogdan, Ghosh, and Doerge (2004) propose to modify BIC by supplementing it with the Binomial prior distribution on the QTL number. If the expected value of this prior distribution does not depend on the number of markers, this leads to an additional "penalty" for the model dimension, which prevents overestimation. As illustrated by theoretical results in Bogdan, Ghosh, and Żak-Szatkowska (2008), mBIC controls the number of falsely detected QTL and has some asymptotic optimality properties in the context of selecting the best multiple regression model under sparsity. Recently, another interesting extension of the BIC, EBIC, was proposed by Chen and Chen (2008). In its standard form (e.g., see Li and Chen (2009)) EBIC uses a non informative uniform prior on the number of QTL. Chen and Chen (2008) support EBIC by showing its consistency.

In a sequence of papers Baierl, Bogdan, Frommlet, and Futschik (2006), Baierl, Futschik, Bogdan, and Biecek (2007), Zak, Baierl, Bogdan, and Futschik (2007) and Bogdan, Frommlet, Biecek, Cheng, Ghosh, and Doerge (2008) mBIC was successfully used to locate multiple interacting QTL. Specifically, Zak, Baierl, Bogdan, and Futschik (2007) proposed a nonparametric version of mBIC based on ranks, which can be used to analyze traits which do not have a normal distribution. However, the rank methods are only well justified if the trait has a continuous distribution. Therefore they have to be used with care when the trait has a "spiked" distribution, i.e., when some proportion of the trait data are concentrated at one point. Recently, a very interesting application of EBIC to the traits with "spiked" distributions was proposed in Li and Chen (2009). Li and Chen (2009) use the approach of Broman (2003) and model such traits with a mixture of a distribution concentrated at one point and a distribution from the general exponential family. They show that an appropriately modified BIC can be used successfully to locate QTL influencing such traits. Here we extend this approach and apply mBIC and EBIC for locating multiple interacting QTL based on the zero-inflated generalized Poisson regression. Note that this application goes beyond the framework of Li and Chen (2009), since the generalized Poisson distribution does not belong to the exponential family.

We illustrate the performance of mBIC and EBIC to a ZIGPR with an extensive simulation study. The results of this study show that the proposed methods allow for a good power of QTL detection, while keeping the false discovery rate at a reasonable level. They also clearly illustrate the superior performance of ZIGPR over other simplified methods analyzing count traits. Here, among other findings, we present the interesting phenomenon of overestimating the number of QTL by the standard Poisson regression. This behavior can be attributed to the inability of the Poisson regression to account for data over-dispersion and therefore it should not be applied for identifying QTL's based on count data. We also report results of the analysis of

the mice gallstone data of Lyons et al. (2003), which confirms the good performance of mBIC applied to ZIGPR. Specifically, our method confirms the existence of a QTL on a chromosome 5, influencing the number of gallstones, and additionally suggests a novel QTL on chromosome 4. The program in R, which can be used for future real data analyses, is available at *http://www-m4.ma.tum.de/Papers/Erhardt/qtl-zigp-code.rar*.

The outline of the chapter is as follows. In Section 1.2 we introduce and discuss our ZIGPR model for QTL mapping. In Section 1.3 we introduce the corresponding versions of mBIC and EBIC. In Section 1.4 we present results of the extensive simulation study comparing ZIGPR to simpler versions of Poisson regression as well as with a standard least squares regression with regard to the performance of mBIC. Section 1.5 contains the results of the analysis of mice gallstone data of Lyons et al. (2003) and Section 1.6 contains a summary as well as directions for further research.

1.2 Zero-inflated generalized Poisson regression

One of the simplest distributions which can be used to model count traits is the Poisson distribution. However, the range of applications of this distribution is very limited due to the lack of its flexibility. Specifically, the standard Poisson model assumes that the trait variance is equal to its mean. As discussed later in this chapter, this weakness becomes particularly disturbing when the Poisson distribution is used together with model selection tools for locating multiple interacting QTL.

There are two natural extensions of the Poisson distribution, which allow for modeling a difference between the mean and the variance: the Negative Binomial (or Poisson-Gamma) distribution and the generalized Poisson distribution. In this chapter we focus on the generalized Poisson distribution $GP(\mu, \varphi)$, which was first introduced by Consul and Jain (1970) and subsequently studied in detail by Consul (1989) (for details, see Introduction).

When comparing to the Negative Binomial (NB) distribution, the GP distribution has several advantages. While the NB distribution with pmf

$$P(Y = y|\mu, \Psi) = \frac{\Gamma(y + \Psi)}{\Gamma(\Psi)y!} \left(\frac{\Psi}{\mu + \Psi}\right)^{\Psi} \left(\frac{\mu}{\mu + \Psi}\right)^{y},$$

and $E(Y) = \mu$, $Var(Y) = \mu(1 + \frac{\mu}{\Psi})$ contains the basic Poisson distribution only as a limiting case for $\Psi \to \infty$, the GP distribution contains the Poisson class for $\varphi = 1$. Second, unlike the NB distribution the dispersion factor in GP is independent of the mean. Hence, in the NB distribution the statistical modeling of overdispersion is less transparent than in case of the GP. For a detailed comparison between GP and NB we refer the readers to Joe and Zhu (2005).

A zero-inflated generalized Poisson (ZIGP) distribution is a further extension of the GP distribution, which allows to model a "spike" at zero. Such a "spike" occurs quite often when the response variable counts disease symptoms (like e.g. the gallstones). As explained in Cui and Yang (2009), the over-excess of zeros may result from the fact that a certain fraction of a population was not exposed to the disease virus. Again we refer to the Introduction for a definition of the ZIGP distribution.

To model the dependence of the count response variable on explanatory variables Famoye and Singh (2006) introduced a zero-inflated generalized Poisson regression model for independent $Y_i \sim ZIGP(\mu_i, \varphi, \omega_i)$, where μ_i and ω_i are defined through the log-linear and logit link functions,

respectively. In this article we will restrict to the case when the zero-inflation parameter ω does not depend on genetic factors, while the dependency of μ_i on explanatory variables is given through the log-linear link function

$$\log \mu_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ji} \ .$$

The constant ω can be interpreted as the fraction of the population which was not exposed to the disease virus. Moreover, we observed that due to some confounding of μ and ω, a precise separation of regressors influencing these two parameters is hardly possible with the sample sizes typically used for QTL mapping. Therefore the extension of our model to include the dependency of ω on the genetic factors did not bring the expected benefits over the restricted version.

The class of considered ZIGPR models contains the subclasses of zero-inflated Poisson regression (ZIPR, $\varphi = 1$), generalized Poisson regression (GPR, $\omega = 0$) and standard Poisson regression (PoiR, $\varphi = 1$, $\omega = 0$).

1.3 mBIC and EBIC for ZIGPR

Consider the problem of locating multiple interacting QTL in experimental populations. In this case precise estimators of QTL positions can be obtained with the multiple interval mapping approach (see e.g. Li and Chen (2009)). However, due to the computational complexity, this method is rarely used in the genome-wide searches for interacting (epistatic) QTL. To reduce the computational burden, interesting genome regions can be initially chosen by identification of important marker-trait associations. In this situation regressor variables are defined by the genotypes of available markers. In case of a backcross design or recombinant inbred lines there are only two genotypes possible at every locus and each of the markers may be represented by just one dummy variable: $X_{ij} = \frac{1}{2}$ or $X_{ij} = -\frac{1}{2}$, depending on the number of alleles from the reference parental line present at marker j for the i^{th} individual. In case of an intercross design there are three possible genotypes and, according to the Cockerham's model (see Kao and Zeng (2002)), each of the markers can be represented by two dummy variables:

$$\text{Additive Effect for individual } i\text{:} \quad X_{aij} = \begin{cases} 1 & \text{if the } j^{th} \text{ marker has a genotype } g_{ij} = AA, \\ 0 & \text{if the } j^{th} \text{ marker has a genotype } g_{ij} = aA, \\ -1 & \text{if the } j^{th} \text{ marker has a genotype } g_{ij} = aa. \end{cases}$$

$$\text{Dominance Effect for individual } i\text{:} \quad X_{dij} = \begin{cases} 1/2 & \text{if } j^{th} \text{ marker has a genotype } g_{ij} = Aa, \\ -1/2 & \text{otherwise .} \end{cases}$$

Let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ denote the vector of values of some quantitative trait for n individuals and let $X_{n \times N_m}$ denote the corresponding design matrix, whose columns contain dummy variables corresponding to all available markers. Note that for the backcross and recombinant inbred lines $N_m = m$, where m is the number of available markers, while for the intercross $N_m = 2m$.

We assume that the relationship between QTL genotypes (coded as above) and the count trait can be described by a zero-inflated generalized Poisson regression model. As already discussed,

we will focus on identification of markers which are closest to the QTL. In our search, apart from main effects (additive and dominance), we may include two-way interactions (epistatic effects). Thus our task consists in choosing the best model of the form $Y_i \sim ZIGP(\mu_i, \varphi, \omega)$, with

$$\log(\mu_i) = \beta_0 + \sum_{j \in I} \beta_j X_{ij} + \sum_{(u,v) \in U} \gamma_{uv} X_{iu} X_{iv}, \tag{1.3.1}$$

where I is a subset of the set $N = \{1, \ldots, N_m\}$ and U is a subset of $N \times N$. Note that the total number of potential two-way interactions is equal to $N_e = N_m(N_m - 1)/2$.

Remark 1. *In principle the model (1.3.1) could be extended to include interactions of higher order. However, due to the increased multiple testing problem, the power for identification of such interactions is very limited for sample sizes typically used in QTL mapping. Therefore, genome-wide searches for high-order interactions are rarely carried out.*

Since we do not know the QTL number nor their locations, we use a model selection procedure for choosing the best regressors in model (1.3.1). One popular method for this purpose is the Schwarz Bayesian Information Criterion (BIC). However, when locating QTL with the standard least-squares regression, BIC was found to have a strong tendency to overestimate the QTL number (see e.g. Broman and Speed (2002)). As discussed in Bogdan, Ghosh, and Żak-Szatkowska (2008), this phenomenon is closely related to the well known multiple testing problem. Specifically, in Bogdan, Ghosh, and Żak-Szatkowska (2008) it is proved that under the orthogonal design the expected number of "false discoveries" produced by BIC converges to infinity if $\frac{N_m}{\sqrt{n}} \to \infty$. In Bogdan, Ghosh, and Doerge (2004) an alternative Bayesian explanation is provided. The Bayesian model selection suggests choosing the model M_j that has the highest posterior probability

$$P(M_j|\boldsymbol{Y}) \propto L(\boldsymbol{Y}|M_j)\pi(M_j) \ ,$$

where $L(\boldsymbol{Y}|M_j)$ is the likelihood of the data given the model M_j and $\pi(M_j)$ is a prior probability of M_j. The standard BIC neglects $\pi(M_j)$ and uses the Laplace approximation for $\log L(\boldsymbol{Y}|M_j)$ (see Ghosh, Delampady, and Samanta (2006)), which results in

$$BIC = \log(L(\boldsymbol{Y}|M_j, \hat{\delta}_j)) - \frac{1}{2} k_j \log(n) \ ,$$

where $\hat{\delta}_j$ is the maximum likelihood estimate of the parameter vector in model M_j and k_j denotes the dimension of δ_j.

In Bogdan, Ghosh, and Doerge (2004) it is observed that neglecting $\pi(M_j)$ corresponds to assigning the same prior probability to each model. It is easy to check that this leads to the implicit Binomial $B\left(N_m, \frac{1}{2}\right)$ prior on the number of main effects. This prior is concentrated mainly on the interval $\left(\frac{N_m - 3\sqrt{N_m}}{2}, \frac{N_m + 3\sqrt{N_m}}{2}\right)$ and assigns an unsuitably large prior probability to the event that the true number of QTL is close to $\frac{N_m}{2}$. This in turn causes the BIC to choose relatively large models. To solve this problem, in Bogdan, Ghosh, and Doerge (2004) a modified version of the BIC, called mBIC, has been proposed. The mBIC criteria allows to take prior information on the number of QTL into account. Let $E(k)$ and $E(r)$ denote the expected values of the prior distributions for the number of main and epistatic effects, respectively. In mBIC the parameter $p = \frac{1}{2}$ in the Binomial prior distribution for the number of true regressors is replaced with $p_a = \frac{E(k)}{N_m}$ for the main effects and $p_e = \frac{E(r)}{N_e}$ for the interactions.

After some simple algebra (for details see e.g., Bogdan, Ghosh, and Doerge (2004) or Zak-Szatkowska and Bogdan (2010)), we obtain that mBIC selects the model which maximizes the expression

$$mBIC := 2\log(L(\boldsymbol{Y}|M_j, \hat{\delta}_j))) - (k_j + r_j)\log(n) - 2k_j \log(l-1) - 2r_j \log(u-1), \quad (1.3.2)$$

where k_j and r_j are the numbers of main and interaction effects in the model M_j, $l = \frac{1}{p_a}$ and $u = \frac{1}{p_e}$. In the case of no prior information, Bogdan, Ghosh, and Żak-Szatkowska (2008) suggest using

$$l = \frac{N_m}{4} \ , \quad (1.3.3)$$

when the scan is restricted to main effects only and

$$l = \frac{N_m}{2.2} \quad \text{and} \quad u = \frac{N_e}{2.2} \ , \quad (1.3.4)$$

when epistatic effects are considered as well.

In comparison to BIC, the standard version of mBIC for detecting main effects and two-ways interactions contains the additional penalty term

$$2k_j \log\left(\frac{N_m}{2.2} - 1\right) + 2r_j \log\left(\frac{N_e}{2.2} - 1\right) \ ,$$

which depends on the number of markers used in the genome scan. As shown in Bogdan, Ghosh, and Żak-Szatkowska (2008), in case of the standard least squares regression this additional term allows to deal with the multiple testing problem and guarantees that the overall type I error does not exceed 0.08 for a sample size of 200 and more than 30 markers. Due to the consistency of mBIC, the probability of the type I error decreases when the sample size increases.

The choice of the same penalty constant (namely 2.2) for main and interaction effects results in dividing the probability of the overall type I error in two approximately equal parts: probability of detection of a "false" additive effect and probability of detection of a "false" interaction. Note that this choice implies a larger penalty for interaction terms than for main effects (since $N_e >> N_m$), so the power of detecting interaction effects is substantially smaller than the power of detecting main effects. This choice is a deliberate decision, justified by the fact that main effects are usually easier to interpret and more "important" than interactions and scientists are usually not interested in searching for interactions at the price of sacrificing the power of detecting main effects.

Calculations presented in Bogdan, Ghosh, and Doerge (2004), Bogdan, Frommlet, Biecek, Cheng, Ghosh, and Doerge (2008) and Bogdan, Ghosh, and Żak-Szatkowska (2008), which lead to the specific choices of l and u, are based on the assumption that the likelihood ratio statistics for testing the significance of specific explanatory variables have asymptotically the chi-square distribution. Since, under some mild regularity conditions, this assumption is satisfied for the Generalized Linear Models (see e.g. Shao (1999)), the proposed choices for l and u are appropriate also in this case. An extensive simulation study, confirming good properties of the mBIC in the context of logistic and Poisson regression, can be found in Zak-Szatkowska and Bogdan (2010). Note however that ZIGPR does not fit the general framework of GLM, since the ZIGP distribution does not belong to the exponential family. In this case a standard choice of l and u can be justified by theoretical results on the asymptotic normality of the maximum

likelihood estimate of $\boldsymbol{\delta} = ((\beta_j)_{j\in I}, (\gamma_{uv})_{(u,v)\in U}, \varphi, \omega)$, presented in Czado, Erhardt, Min, and Wagner (2007) based on Min and Czado (2010). This implies that under the null hypothesis the corresponding likelihood ratio test statistics have also asymptotically a chi-square distribution. The appropriateness of the standard choice of l and u for ZIGPR is confirmed by the simulation study, presented in the next section.

A similar modification of BIC was recently proposed by Chen and Chen (2008), who introduce an extended BIC (EBIC), based on the different prior choices for the model dimension. In comparison to mBIC, the priors used by EBIC substantially prefer models of larger dimensions. Specifically, the standard, most restrictive version of the EBIC, assumes that the prior distribution on the number of main effects is uniform on the set $\{0, 1, \ldots, N_m\}$ (see Li and Chen (2009)). After assigning the same prior probability to all models of the same dimension this results in $\pi(M_j) = \frac{1}{N_m+1}\binom{N_m}{k}^{-1}$. Interestingly, the same prior is proposed in Scott and Berger (2008), where it results from the application of a hierarchical model with a non informative, uniform prior on the proportion of true regressors p. The choice between mBIC and EBIC should depend on the prior expectations concerning the QTL number. As illustrated by theoretical results discussed in Bogdan, Frommlet, Biecek, Cheng, Ghosh, and Doerge (2008) and proved in Bogdan, Chakrabarti, and J.K.Ghosh (2008), mBIC has some asymptotic optimality properties in the context of selecting the best multiple regression model under sparsity. Therefore mBIC seems to be especially appropriate in case when one expects that the number of true predictors is much smaller than the number of columns in the "total" design matrix. To compare these two criteria, in the next section we present results of an extensive simulation study, in which we identify important main effects with the standard version of mBIC

$$mBIC := 2\log(L(Y|M_j, \hat{\delta}_j))) - k\log(n) - 2k\log\left(\frac{N_m}{4} - 1\right) \quad , \tag{1.3.5}$$

and the standard version of EBIC

$$EBIC := 2\log(L(Y|M_j, \hat{\delta}_j))) - k\log(n) - 2\log\binom{N_m}{k} \quad . \tag{1.3.6}$$

1.4 Simulation study

Simulations are carried out to investigate the performance of our proposed methods of QTL detection for a backcross design. We simulate genotypes of $N_m = 100$ markers located on 20 mice chromosomes. These marker positions are identical to the ones in the data set investigated by Lyons et al. (2003). The marker positions are supplemented by $k = 10$ fictional QTL's (not matching any of the markers) located on chromosomes 1 to 6. Figure 1.1 plots the marker and QTL positions on these 6 chromosomes.

Trait values are generated from the ZIGPR model, $Y_i \sim ZIGP(\mu_i(\boldsymbol{\beta}), \varphi, \omega))$, with

$$\mu_i(\boldsymbol{\beta}) := \exp\left\{2.05 + \boldsymbol{X}'_{Q,i}\boldsymbol{\beta}\right\}, \quad i = 1, \ldots, n \quad , \tag{1.4.1}$$

where $\boldsymbol{X}_{Q,i} = (X_{Q1,i}, \ldots, X_{Q10,i})'$ denotes the vector of 10 QTL genotypes coded as $-1/2$ and $1/2$ for homozygotes and heterozygotes, respectively, and parameter values are chosen as

$$\boldsymbol{\beta} = (-0.20, 1.00, 0.25, -0.60, 0.80, 1.20, 0.70, -0.15, -0.40, 1.50)' \quad .$$

Additionally, we choose $\varphi = 2$ and investigate small as well as medium sized zero-inflation of $\omega \in \{20\%, 40\%\}$.

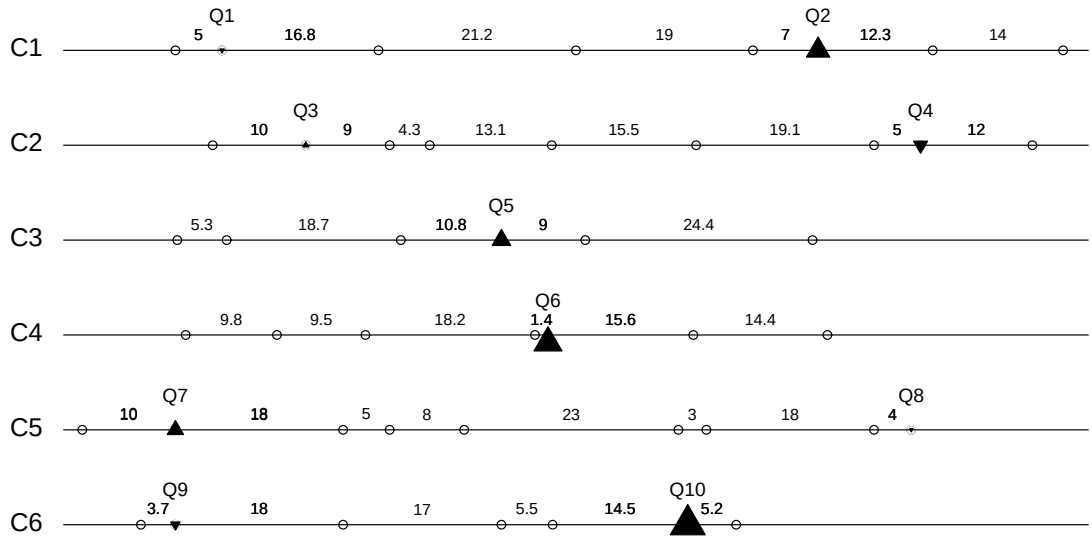

Figure 1.1: Marker positions and positions of the true QTL.

Our simulation results are based on $N = 1000$ replicates for the sample sizes $n = 200$ and $n = 500$. In each run new random markers and QTL genotypes are generated from the map, the coefficients $\boldsymbol{\beta}$, however, are kept identical. In each of these replications model selection is carried out using a forward selection procedure. At first we start with the Null model, i.e., we fit $ZIGPR(\mu_i(\beta_0), \varphi, \omega))$, where β_0 is the coefficient of an intercept. We sequentially add the marker which increases the standard version of mBIC (1.3.5) the most, as long as mBIC grows. Additionally, we carry out forward selection based on mBIC with a Gaussian linear model (LM), a Poisson regression (PoiR), generalized Poisson regression (GPR) and zero-inflated Poisson regression (ZIPR). We include the standard least squares regression LM, since it is capable of identifying correlations between explanatory and response variables, and may perform reasonably choosing important predictors for the ZIGP data. Also, due to the central limit theorem, we expect that mBIC with LM will control the number of false positives, even when the true data are generated according to ZIGPR. Additionally, for each model class we perform model selection based on the standard version of EBIC given in (1.3.6).

Results of the simulation study are compared for the five model classes. We consider the following statistics:

- true positives (TP): number of selected effects whose distance to the simulated QTL's was less or equal 20 cM; if more than one effect was caught in the interval around a certain QTL only one of them was counted
- false positives (FP): number of selected effects whose distance to the simulated QTL's was higher than 20 cM
- false positives (FP) + false negatives (FN), where $FN = 10 - TP$
- power: $TP/10$
- observed false discovery rate : $FDR = FP/(FP + TP)$

	n = 200									
	mBIC									
	$\omega = 20\%$					$\omega = 40\%$				
	LM	PoiR	ZIPR	GPR	ZIGPR	LM	PoiR	ZIPR	GPR	ZIGPR
FP	0.125	21.075	11.309	0.658	0.357	0.106	27.033	9.614	0.296	0.373
FP+FN	6.944	22.903	13.486	7.952	5.371	8.260	29.025	12.460	9.631	6.623
Power	0.318	0.817	0.782	0.271	0.499	0.185	0.801	0.715	0.066	0.375
FDR	0.036	0.711	0.575	0.188	0.062	0.050	0.764	0.558	0.282	0.088
	EBIC									
	$\omega = 20\%$					$\omega = 40\%$				
FP	0.149	40.869	19.746	0.880	0.604	0.090	53.568	15.354	0.281	0.614
FP+FN	6.879	41.808	21.377	7.943	5.196	8.368	54.276	17.671	9.633	6.397
Power	0.327	0.906	0.837	0.294	0.541	0.172	0.929	0.768	0.065	0.422
FDR	0.040	0.809	0.682	0.213	0.090	0.044	0.846	0.645	0.270	0.116
	n = 500									
	mBIC									
	$\omega = 20\%$					$\omega = 40\%$				
	LM	PoiR	ZIPR	GPR	ZIGPR	LM	PoiR	ZIPR	GPR	ZIGPR
FP	0.112	24.662	14.818	0.642	0.215	0.117	30.817	13.813	0.405	0.234
FP+FN	4.830	25.519	15.723	6.102	3.541	5.949	31.847	15.088	8.381	4.047
Power	0.528	0.914	0.909	0.454	0.667	0.417	0.897	0.873	0.202	0.619
FDR	0.019	0.723	0.607	0.112	0.028	0.025	0.770	0.599	0.142	0.033
	EBIC									
	$\omega = 20\%$					$\omega = 40\%$				
FP	0.144	40.428	26.274	0.984	0.466	0.120	48.520	23.765	0.465	0.435
FP+FN	4.662	40.936	26.878	6.145	3.397	5.815	49.110	24.665	8.490	3.940
Power	0.548	0.949	0.940	0.484	0.707	0.430	0.941	0.910	0.198	0.649
FDR	0.023	0.805	0.725	0.149	0.055	0.024	0.835	0.708	0.154	0.057

Table 1.1: Average number of false positives (FP), false positives + false negatives (FP+FN), power and false discovery rate (FDR) based on mBIC and EBIC for different model classes and $n = 200, 500$ and $\omega = 20\%, 40\%$

In Table 1.1 we will tabulate the averages of FP, power, $FP + FN$ and FDR. Figure 1.2 plots the estimated power against the magnitude of the true regression coefficients $\boldsymbol{\beta}$.

From Table 1.1 and Figure 1.2 we see that a higher number of observations substantially eases the detection of significant effects. On the other hand, higher zero-inflation makes the detection of correct effects more difficult even in the correctly specified ZIGPR model. Also, according to Figure 1.2, the power of detection clearly increases with the magnitude of the true regression coefficients.

Due to its lowest false discovery rates while maintaining high power rates, the ZIGP regression model is definitely the best of the regression models considered. Interestingly, the second best is the standard least squares regression model, LM. While LM clearly performs worse than the correct ZIGPR model, it substantially outperforms other misspecified models based on the Poisson regression. Specifically, LM offers a much larger power than the General Poisson (GPR) model without the zero inflation parameter. In case of the models without the overdispersion parameter, PoiR and ZIPR, we observe the opposite. These models offer a much higher power than LM, or even ZIGPR, but instead lead to the detection of a large number of false positives. FDR of the procedures based on PoiR and ZIPR systematically exceeds 50%, which means that the number of false positives usually exceeds the number of true discoveries. We believe that these models pick too many regressors in order to account for the data heterogeneity caused by overdispersion.

In Table 1.2 we report the results of a further simulation study, in which the data were

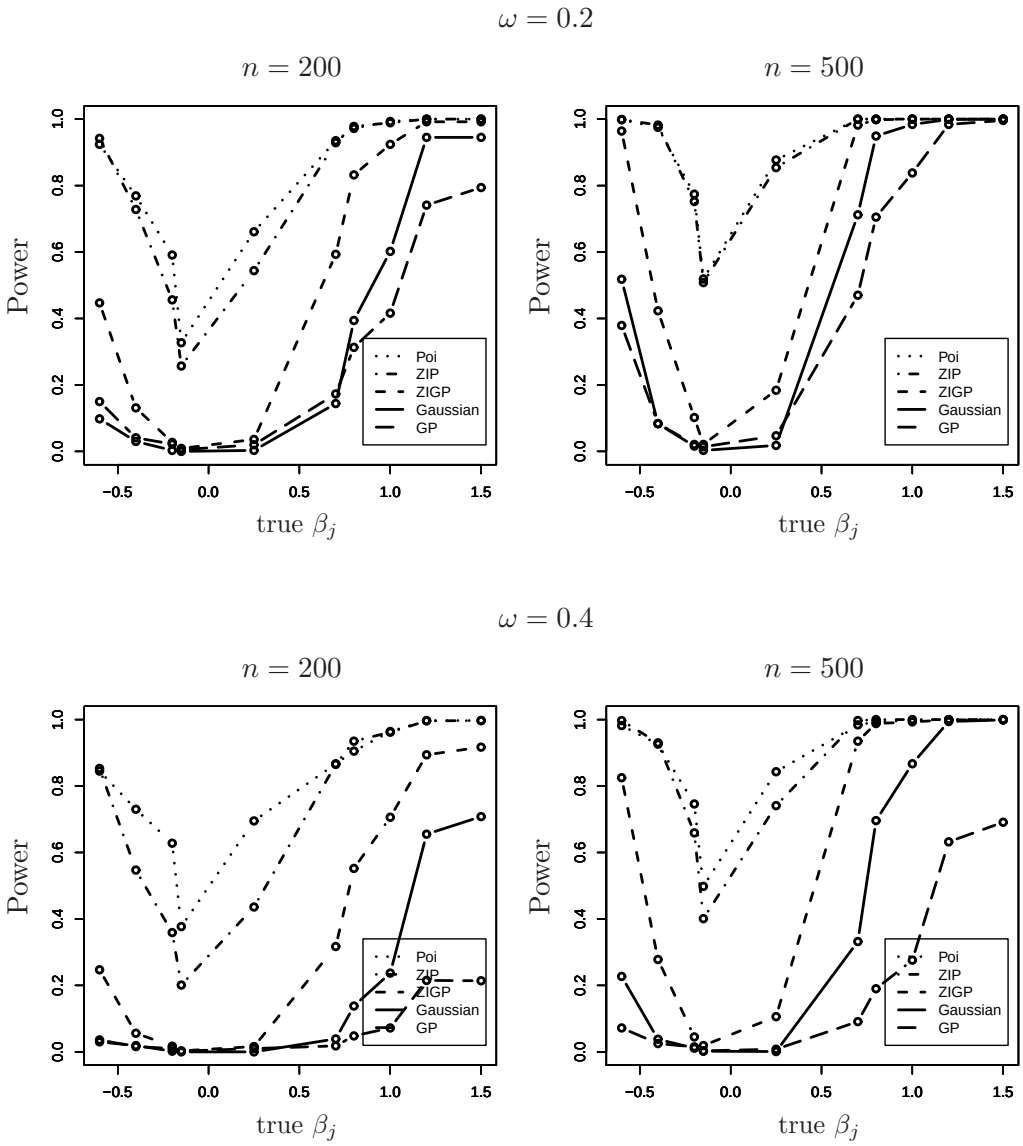

Figure 1.2: Power for different sizes of true regression coefficients based on several model classes. Note that the lines are linearly interpolated to increase visual comparability.

generated according to the standard Poisson regression model PoiR, with μ_i defined by (1.4.1). Since this is only meant to be an illustrative example, we restrict to the case of a scan based on mBIC for $n = 200$ mice. In this case PoiR and ZIPR perform similarly bad. The same holds for the performance of GPR and ZIGPR. The reason is simply that in the model classes allowing for excess zeros, the zero-inflation parameter for Poisson traits is estimated to be close to zero, hence the performance only depends on the underlying distribution, which is not inflated (i.e., Poisson and GP, respectively). Interestingly, also in this case PoiR and ZIPR substantially overestimate

	n = 200, mBIC				
	LM	PoiR	ZIPR	GPR	ZIGPR
FP	0.095	8.200	8.150	0.405	0.410
FP+FN	5.285	9.830	9.810	3.920	3.930
Power	0.481	0.837	0.834	0.648	0.648
FDR	0.018	0.476	0.475	0.053	0.053

Table 1.2: Average number of false positives (FP), false positives + false negatives (FP+FN), power and false discovery rate (FDR) based on mBIC and EBIC for different model classes when the traits come from a Poisson distribution

the number of QTL. The number of false positives produced by these methods is approximately equal to the number of true discoveries, with FDR close to 50%. At the same time procedures based on GPR work very well, maintaining a reasonable power and FDR at the level close to 5%. It turns out that the poor behavior of the method based on PoiR or ZIPR results from the model misspecification, caused by the discrepancy between the marker and QTL location. Here we give a simple illustrative example: we generate Poisson traits with 10 true effects $\boldsymbol{X}_i := (X_{i1}, \ldots, X_{i10})'$ and $\mu_i := \exp(2.05 + \boldsymbol{X}_i'\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is chosen as before. Then we fit two GPR models, one using $\boldsymbol{X}_i$ as regressors and one using misspecified $\boldsymbol{X}_i^{mis} := (X_{i1}^{mis}, \ldots, X_{i10}^{mis})'$, which are random and reflect genotypes referring to a recombination fraction with distance of $10cM$ to $\boldsymbol{X}_i$ in each component. In the left panel of Table 1.3 we see that in the first case φ is estimated to be 1.01. This illustrates that the GPR class contains the PoiR class and that the dispersion can be estimated with a very good precision. In the second case, however, the regressors are misspecified by not knowing the exact trait loci and the marker genotypes $\boldsymbol{X}^{mis}$ are used instead. Now φ is estimated to be 3.25 (see right panel of Table 1.3), i.e., the estimated variance exceeds the estimated mean by a factor of more than 10. As one can see in Table 1.2 this leads to a dramatic overfit when using PoiR since this model cannot reflect the additional overdispersion and picks too many regressors in order to account for the data heterogeneity. Zero-inflation also leads to overdispersion, however one can see in Table 1.1 for the ZIPR case that zero-inflation alone is insufficient to compensate the lack of the overdispersion parameter.

	Estimate	Std. Error	$Pr(>\|z\|)$		Estimate	Std. Error	$Pr(>\|z\|)$
Interc.	2.064	0.031	$< 2 \cdot 10^{-16}$	Interc.	2.197	0.081	$< 2 \cdot 10^{-16}$
X_1	−0.211	0.031	10^{-11}	X_1^{mis}	−0.097	0.097	0.316
X_2	0.920	0.038	$< 2 \cdot 10^{-16}$	X_2^{mis}	0.836	0.104	$7 \cdot 10^{-16}$
X_3	0.266	0.037	$5 \cdot 10^{-13}$	X_3^{mis}	0.211	0.103	0.041
X_4	−0.566	0.029	$< 2 \cdot 10^{-16}$	X_4^{mis}	−0.673	0.097	$5 \cdot 10^{-12}$
X_5	0.812	0.035	$< 2 \cdot 10^{-16}$	X_5^{mis}	0.626	0.105	$3 \cdot 10^{-9}$
X_6	1.228	0.049	$< 2 \cdot 10^{-16}$	X_6^{mis}	1.137	0.118	$< 2 \cdot 10^{-16}$
X_7	0.696	0.038	$< 2 \cdot 10^{-16}$	X_7^{mis}	0.379	0.102	$2 \cdot 10^{-4}$
X_8	−0.174	0.033	10^{-7}	X_8^{mis}	−0.191	0.097	0.049
X_9	−0.417	0.033	$< 2 \cdot 10^{-16}$	X_9^{mis}	−0.310	0.099	0.002
X_{10}	1.518	0.046	$< 2 \cdot 10^{-16}$	X_{10}^{mis}	1.199	0.114	$< 2 \cdot 10^{-16}$

Table 1.3: GP fit of Poisson data ($n = 200$) based on the 10 true effects $(X_1, \ldots, X_{10})'$ (left panel) with $\hat{\varphi} = 1.01$ (0.051). GP fit of the same data based on 10 misspecified effects correlated with $(X_1, \ldots, X_{10})'$ which are $10cM$ away from the true effects (right panel), $\hat{\varphi} = 3.25$ (0.477).

Comparing the performance of mBIC and EBIC under the most appropriate ZIGPR model

we observe that both these criteria perform very well and their results do not differ much. As expected, EBIC offers slightly larger power at the price of a larger, but still reasonable, FDR. Our simulations show that the power of these criteria increases and the expected number of false positives decreases as the samples size goes up, which strongly suggest that these criteria are consistent also under the ZIGPR model.

1.5 Real data analysis

The data by Lyons et al. (2003) considers different phenotypes related to gallstones. While Lyons et al. (2003) focus on the gallstone weight, a score for solid gallstones and the gallbladder volume, we will focus on the number of gallstones the 277 male mice developed. The data is publicly available at

http://phenome.jax.org/phenome/protodocs/QTL/QTL-Lyons3.xls

and refers to an intercross of CAST/Ei and 129S1/SvImJ inbred mice. Since the phenotypes considered in Lyons et al. (2003, Figure 5) are related to the number of gallstones the mice developed, we perform a preselection of interesting chromosomes based on this figure. Hence we restrict our search to eight chromosomes accounting for 41 markers, i.e., we consider the chromosomes 2, 3, 4, 5, 7, 17, 18 and 19. We replace missing genotypes by their expected values, given the flanking markers (see for instance Haley and Knott (1992)). Additive and dominance effects are added separately, according to the specification provided in Section 1.3, with a corresponding to the CAST/Ei allele. As a search method we used forward selection with mBIC based on ZIGPR. The reason for which we chose mBIC rather than EBIC, is that mBIC has been adapted for the search of interaction effects. In this case mBIC adjusts to the increased "multiple testing" problem by changing the penalty constant from 4 to 2.2 (see (1.3.2)). The adaptation of EBIC for the search of interactions is not obvious and we are not aware of existing solutions to this problem.

We performed two different analyses. At first we searched only for main effects with the standard version of mBIC (1.3.2) and a penalty constant provided in (1.3.3). In this case mBIC identifies one additive effect at D5Mit183 ("D5Mit183(a)"). This is in line with the result of Lyons et al. (2003), which found this marker to be significant for all three Gallstone related traits considered in their study. A model summary is given in the upper panel of Table 1.4. Note that asymptotic normality of the maximum likelihood estimates of the dispersion parameter φ and zero-inflation parameter ω have been shown in Czado, Erhardt, Min, and Wagner (2007, Theorem 1). Therefore we report the p-values of the Wald test also for these estimates. Additionally, we performed the search for both additive and interaction effects using mBIC (1.3.2) with constants provided in (1.3.4). In this search we detected an interaction term between two additive effects: D5Mit183 and a novel suggestive QTL, D4Mit42. A model summary is given in the middle panel of Table 1.4. Additionally, in the lower panel of Table 1.4 we provide the results of the analysis based on the model including additive effects of both D5Mit183 and D4Mit42 and their interaction. Interestingly, the p-value corresponding to the interaction term between D5Mit183 and D4Mit42 is substantially smaller than the p-values corresponding to the additive effects, which suggests that the interaction between D5Mit183 and D4Mit42 plays a very important role in determining the expected number of gallstones. This observation is confirmed by the graphical representation in Figure 1.3. In accordance

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
Intercept	0.067	0.983	0.068	0.946
D5Mit183(a)	−1.292	0.432	−2.991	0.003
φ	6.799	3.560	1.909	0.056
ω	0.631	0.362	1.743	0.081
Intercept	−0.156	0.572	−0.272	0.786
D5Mit183(a):D4Mit42(a)	−2.298	0.495	−4.647	$3.4 \cdot 10^{-6}$
φ	5.776	2.520	2.293	0.022
ω	0.575	0.167	3.437	0.001
Intercept	−0.864	0.573	−1.510	0.131
D5Mit183(a)	−1.244	0.442	−2.817	0.005
D4Mit42(a)	−0.215	0.476	−0.451	0.652
D5Mit183(a):D4Mit42(a)	−2.177	0.548	−3.973	$7.1 \cdot 10^{-5}$
φ	5.387	2.185	2.466	0.014
ω	0.458	0.163	2.809	0.005

Table 1.4: ZIGPR model summaries of forward selection based on mBIC for different regression designs.

with the results of the search for main effects this figure suggests that the expected number of gallstones decreases when the number of 129S1/SvImJ alleles at D5Mit183 increases. However, according to the bottom graph, the effect of D5Mit183 strongly depends on the genotype at D4Mit42, and is most pronounced for mice who are homozygous for 129S1/SvImJ allele at D4Mit42. Specifically, the average number of gallstones is decisively the largest in the group of mice with the combination of dummy variables equal to (-1,1), which corresponds to the mice homozygous for CAST/Ei allele at D5Mit183 and for 129S1/SvImJ allele at D4Mit42. Finally we add that we also carried out a scan based on the LM. Neither in the search over main effects nor in the search including epistatic effects a significant effect could be caught.

1.6 Discussion

We investigated the applicability of different versions of Poisson regression and the modified Bayesian Information Criterion for locating multiple interacting quantitative trait loci influencing count traits. Our research demonstrates very good properties of the zero-inflated generalized Poisson regression in this context. ZIGPR takes into account both the overdispersion and an over-excess of zeros and performs much better than simplified versions of Poisson regression in case when both these parameters play an important role. Moreover, we found out that the overdispersion parameter allows to compensate for a model misspecification due to the discrepancy between marker and QTL locations. Therefore, the search for markers associated with the count trait based on ZIGPR gives much better results than the one based on the standard Poisson regression, even when the data are generated according the latter. Also, our simulations illustrate very good properties of the modified versions of the Bayesian Information Criterion, mBIC and EBIC, as applied to select important predictors for ZIGPR. Both these criteria perform in a similar way and guarantee a good power of QTL detection, while keeping the false discovery rate at a low level. The reported real data analysis shows the possible gains, which can be obtained when ZIGPR with mBIC is used for detection of interacting QTL.

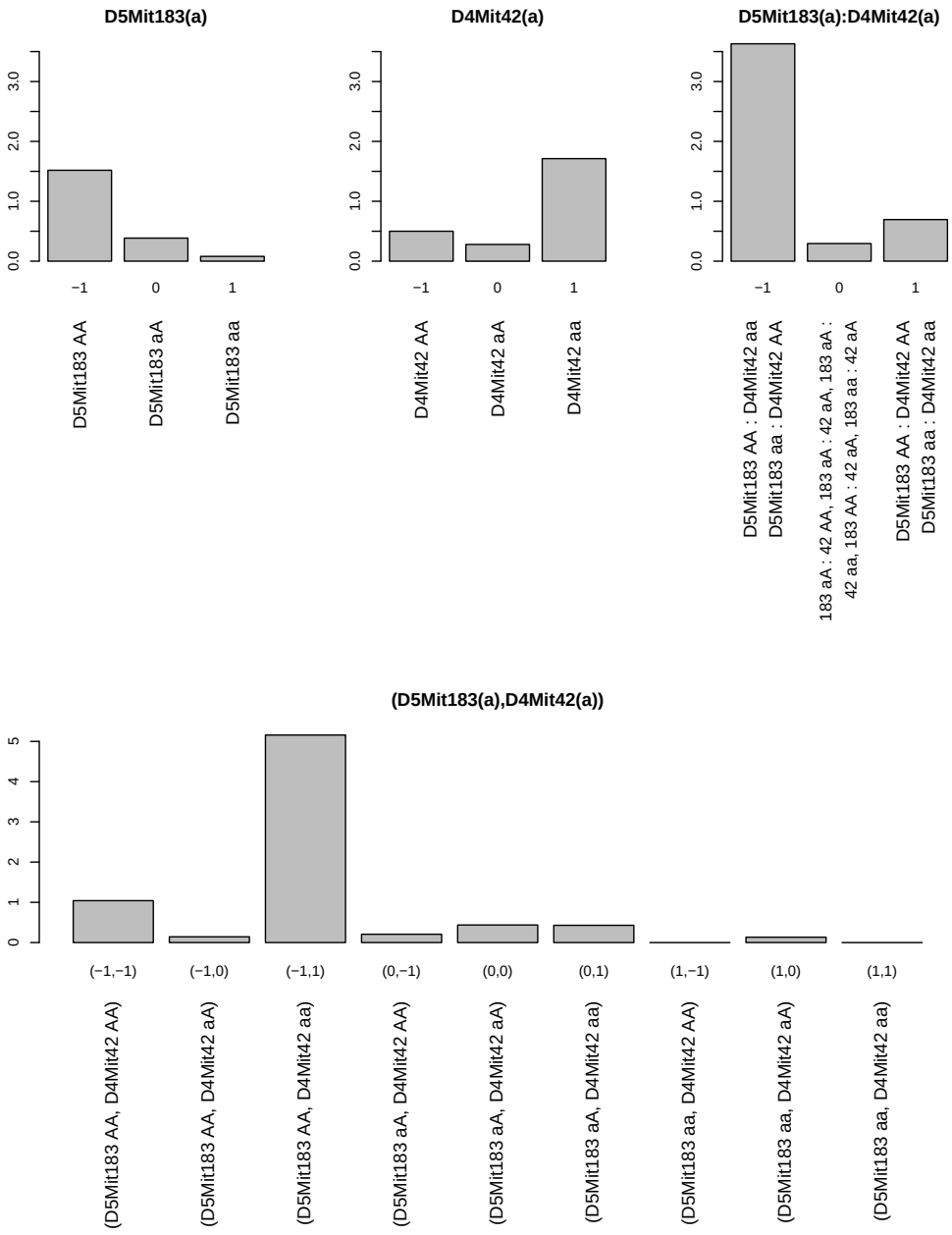

Figure 1.3: Average number of gallstones in different groups of mice, specified by dummy variables corresponding to the additive effects of D5Mit183 and D4Mit42.

Good properties of mBIC in the context of sparse orthogonal multiple regression were confirmed by the results on its asymptotic optimality, proved in Bogdan, Chakrabarti, and J.K.Ghosh (2008). Our preliminary results suggest that similar asymptotic optimality results can be proved for EBIC. However, the extension of these results to the nonorthogonal designs and ZIGPR models presents a major challenge and remains a topic for future research.

Due to the complexity of a large scale simulation study, whose main purpose was the comparison of different Poisson regression models, we reduced the attention to the search over markers. Note that the computational effort for the simulation study carried out in Table 1.1 was very high. We made quite some effort to optimize the R code, nevertheless the repeated search for

significant effects over the 100 main effects was running for more than 20 days on a parallelized 32-core cluster with 2.6 GHz processors. However, an extension of the proposed methodology to the multiple interval mapping is in general quite straightforward and, concerning the estimates of QTL effects and positions, goes along the line of an interval mapping for ZIGPR, as proposed in Cui and Yang (2009). Concerning the estimate of a QTL number, a successful application of EBIC for the multiple interval mapping with mixture General Linear Models was presented in Li and Chen (2009). Also, the results reported in Bogdan, Frommlet, Biecek, Cheng, Ghosh, and Doerge (2008) show that if markers are on the average distant by more than 5 cM then mBIC may be successfully used with the multiple interval mapping. However, the results reported in Bogdan, Frommlet, Biecek, Cheng, Ghosh, and Doerge (2008) show also that if markers are very densely spaced (less than 5 cM apart) then the neighboring marker genotypes are strongly correlated and the penalty in mBIC and EBIC should be substantially relaxed. We believe that the corresponding scaling coefficients provided in Bogdan, Frommlet, Biecek, Cheng, Ghosh, and Doerge (2008) would work well also for ZIGPR but an exact verification requires a very intensive simulation study and is out of the scope of the present chapter.

To reduce the complexity of our simulation study we identified the best regression model with a forward selection. Our simulations, as well as results reported in Broman (1997), Broman and Speed (2002), and Bogdan, Ghosh, and Doerge (2004), show that the forward selection usually performs very well in the context of QTL mapping. However, the real data analysis reported in Bogdan, Frommlet, Biecek, Cheng, Ghosh, and Doerge (2008) illustrates that in some situations this procedure may fail to identify the optimal model. The uncertainty related to the model choice can be well expressed within the Bayesian framework by the posterior model probabilities. The Bayesian approach for the analysis and comparison of ZIGPR models was investigated e.g. in Gschlößl and Czado (2008). However, the computational complexity of the full Bayes analysis by Markov Chain Monte Carlo (MCMC) substantially limits its range of applications in the context of localizing multiple interacting QTL. Note however that both mBIC and EBIC allow an approximation to the posterior probabilities of different models according to

$$P(M_i|Y) \approx \frac{\exp(xBIC(i)/2)}{\sum_j \exp(xBIC(j)/2)} \ , \tag{1.6.1}$$

where $xBIC$ denotes mBIC (1.3.2) or EBIC (1.3.6) and the sum in the denominator is over all possible ZIGPR models. Thus, to estimate the posterior probability of a given model by the modified BIC it is enough to visit each of the plausible models just once. This allows to substantially reduce the computational burden in comparison to the MCMC methods, which typically require multiple visits of each model, and then estimate the posterior probability by the frequency of such visits. However, the estimate of $P(M_i|Y)$ provided in (1.6.1) may be accurate only if the majority of plausible models is represented in the denominator. Therefore, to use mBIC or EBIC in a Bayesian context, a suitable, computationally efficient search strategy still needs to be developed.

Chapter 2

A method for approximately sampling high-dimensional count variables with prespecified Pearson correlation

2.1 Introduction

Input modeling comes into play when stochastic simulations are carried out where there is uncertainty in the simulated system. The input model represents the uncertainty. In the context of sampling from multivariate random variables it consists in the selection and fitting of a multivariate distribution whose behavior cannot be predicted with certainty. The task of sampling from a multivariate random vector becomes especially challenging when a multivariate distribution with the desired properties does not exist.

The goal of this chapter is to sample from count random variables (rv's) $Y_1, \ldots, Y_T$ with $Y_t \sim F_t$, $t = 1, \ldots, T$ with prespecified $\text{corr}(\boldsymbol{Y}) = \boldsymbol{\Sigma}^Y$, with (t, t^*)th element $\boldsymbol{\Sigma}^Y_{tt^*} = \rho_{tt^*}$ and $\rho_{11} = 1$. Kawamura (1979) defines a multivariate binomial distribution $B(N, p_i)$ and shows that for $N \to \infty$ under the condition that $Np_i = \lambda_i$ and a single common correlation term $\rho_{tt^*} = \rho > 0$, one can obtain a multivariate Poisson distribution with marginal parameters λ_i. Tsiamyrtzis and Karlis (2004) criticize the limited use of multivariate discrete models due to the difficulty in calculating the required probabilities and suggest a more efficient calculation of the joint probabilities. Karlis and Meligkotsidou (2005) generalize multivariate Poisson models and construct a model which allows for individual correlations for each pair of variables. This distribution, however, does not allow for negative correlation.

Kopociński (1999) develop a multivariate negative binomial distribution, whereas Vernic (2000) develops a multivariate generalization of the generalized Poisson (Consul and Jain (1970)) distribution both capable of modeling exchangeable covariance. It must be recognized that multivariate discrete distributions discussed in the literature have several shortcomings and hence are often not suited to represent the desired dependence structure.

Lurie and Goldberg (1998) introduce an "approximate method for sampling correlated random variables from partially-specified distributions" and state that it may be applied to continuous margins. This approach is based on Li and Hammond (1975), who use the multivariate

normal distribution for a sampling algorithm. They manipulate the correlations of the multivariate multinomial distribution by a constrained optimization approach using some distance measure between these correlations and the target correlations. Li and Hammond (1975) minimize over all ρ_{tt^*}, $1 \leq t < t^* \leq T$ simultaneously rather than - as in our approach - separately in a sequential order over correlations and partial correlations.

A widely used "naive" sampling approach is based on sampling from a Gaussian copula while assuming that the correlation parameters of the copula coincide with the desired target correlation. A more promising approximative solution to this problem is the NORTA method ('NORmal To Anything'), (see Cario and Nelson (1997) and Chen (2000)) which is based on the work of Marida (1970) and Li and Hammond (1975). Avramidis, Channouf, and L'Ecuyer (2009) apply it to discrete margins. A Gaussian copula is used to define a bivariate distribution. The copula parameter is approximated separately for each pair of margins of the high dimensional problem. In each case this is a problem of root finding and is solved using different optimization routines. Potentially, all pairwise copula parameters need to be corrected afterwards in order to obtain a positive definite correlation matrix. Based on the same approach, the ARTA method (Autoregressive-To-Anything) by Cario and Nelson (1996) allows to sample from univariate autoregressive time series. Further, Biller and Nelson (2003) combine these approaches. This Vector-Autoregressive-To-Anything technique (VARTA) extends the methodology applied by NORTA for margins with an autoregressive structure. Biller and Ghosh (2006) give an overview of a variety of input modeling approaches.

Building on the work on vines of Joe (1996), Bedford and Cooke (2001a), Bedford and Cooke (2001b) and Bedford and Cooke (2002), the work of Aas, Czado, Frigessi, and Bakken (2009) uses pair-copula decompositions of a general multivariate distribution and proposes a new method to perform inference. Our general idea is to use a conditional sampling approach based on pair-copula constructions, i.e., a decomposition of a T-dimensional distribution to a product of bivariate copulas. Since these bivariate copulas (Gaussian in our case) have only one copula parameter, we can use root finding based on bisection to determine optimal parameters for each pair-copula and use them to derive an approximate set of parameters for a copula in dimension T.

In order to account for a different upper and lower tail dependence Biller (2009) develops input models also based on such vine models. She suggests to select two-dimensional (conditional) copulas accounting for the desired behavior. For given data, maximum likelihood estimators of the copula parameters can be determined at data being sampled from the multivariate distributions obtained by these estimators. However the issue of determining the input parameters for other, hypothetical correlations is not subject of the chapter. A similar characterization for the NORTA distribution with a C-vine has been suggested in Biller and Gunes (2008), but with a different purpose: to account for parameter uncertainty in large-scale stochastic simulations with correlated inputs.

Our approach is innovative in the following context: it generates a sample from a multivariate specification where the empirical correlation of the sample comes very close to a hypothetical target correlation even when the sample size is small. It is not very numerically demanding: for dimension T, $T(T-1)/2$ simple root finding step have to be carried out, usually done within seconds. We illustrate that the naive method and NORTA are less suitable for this problem since the empirical correlation of the sample has a higher absolute deviation from the desired target correlation. Moreover, the input parameters obtained by NORTA may need a

posteriori correction in order to result in a positive definite base correlation matrix, which is not necessary in our approach. An implementation for R is available as package *corcounts* on 'The Comprehensive R Archive Network' (CRAN). A simulation study based on our sampling method will be carried out in the context of generalized estimating equations in Chapter 3 of this thesis.

This chapter is organized as follows: In Section 2.2, we will review basic properties of multivariate distributions and copulas and also will introduce additional building blocks needed. Section 2.3 consists of a high level version of the approach giving background insights. Afterwards we derive the sampling approach in detail for the trivariate case. An algorithm in pseudo-code will also be given. The general case in dimension T using a vine structure for a PCC will be discussed in Section 2.4. We summarize the naive and the NORTA methods in Section 2.5. For Negative-Binomial data in dimension 8 we illustrate the advantages of our approach comparing the C-vine sampling approach to the competing methods. An extensive simulation study is given in Section 2.6. We conclude with a summary and discussion in Section 2.7.

2.2 Copulas and multivariate Distributions

We aim to develop an approach for sampling from vectors of dependent random variables (rv's) with specified Pearson correlation. In general, our approach is applicable to continuous as well as discrete margins. We focus on the class of discrete and especially count distributions since they are the more challenging and hence more general class in the following sense: the efficiency of the empirical correlation as a correlation estimator is lower for discrete than for continuous margins, hence the manipulation of an empirical correlation matrix, which our approach is based on, is difficult for small sample sizes. Also, the inversion method is based on more general pseudo-inverses. For fitting such simulated count data with a different specification, i.e., generalized estimating equations, see the simulation study in Erhardt and Czado (2009c). We consider the Poisson distribution and also a generalized Poisson (GP) distribution which contains the Poisson class but additionally allows for extra dispersion. Moreover, in many applications it is desirable to allow for excess zeros, therefore we also consider the zero-inflated generalized Poisson (ZIGP) distribution containing the Poisson and the GP class as special cases. Finally we consider the popular Negative-Binomial (NB) distribution.

Consul and Jain (1970) introduced the GP distribution, which can model under- and overdispersion by an additional dispersion parameter φ and contains the Poisson distribution for $\varphi = 1$. The ZIGP distribution has an additional zero-inflation parameter ω allowing for excess zeros and hence a second source of overdispersion. For $\omega = 0$, the ZIGP distribution simplifies to the GP distribution. In order to allow for a comparison between these distribution families, the mean parameterization was chosen for all of them. The pmf of these distributions together with means and variances are given in Table 2.1.

We will use a PCC based on bivariate copulas to obtain multivariate multivariate uniform vectors and combine it with the inversion sampling method. A two-dimensional copula C is a bivariate cdf $C : [0,1]^2 \to [0,1]$ whose univariate margins are uniform on $[0,1]$, i.e., $C(u_1, 1) = u_1$ and $C(1, u_2) = u_2$. For two continuous rv's $\boldsymbol{Y} := (Y_1, Y_2)'$ with marginal distributions F_1, F_2, the rv $F_t(Y_t)$ is uniform on $[0,1]$, hence while F_t reflects the marginal distribution of Y_t, C

	$P(Y=y)$	$E(Y)$	$Var(Y)$
Poisson	$\frac{\mu}{y!}e^{-\mu}$	μ	μ
GP	$\frac{\mu(\mu+(\varphi-1)y)^{y-1}}{y!}\varphi^{-y}e^{-\frac{1}{\varphi}(\mu+(\varphi-1)y)}$, where $\varphi > \max(\frac{1}{2}, 1-\frac{\mu}{m})$ and m is the largest natural number with $\mu + m(\varphi-1) > 0$, if $\varphi < 1$.	μ	$E(Y)\varphi^2$
ZIGP	$\mathbb{1}_{\{y=0\}}\left[\omega + (1-\omega)e^{-\frac{\mu}{\varphi}}\right]$ $+\mathbb{1}_{\{y>0\}}\left[(1-\omega)\frac{\mu(\mu+(\varphi-1)y)^{y-1}}{y!}\varphi^{-y}e^{-\frac{1}{\varphi}(\mu+(\varphi-1)y)}\right]$, where in the case for $\varphi < 1$ the same condition as in the GP case must hold.	$(1-\omega)\mu$	$E(Y)\left(\varphi^2+\mu\omega\right)$
NB	$\frac{\Gamma(y+\Psi)}{\Gamma(\Psi)y!}\left(\frac{\Psi}{\mu+\Psi}\right)^{\Psi}\left(\frac{\mu}{\mu+\Psi}\right)^{y}$	μ	$\mu(1+\frac{\mu}{\Psi})$

Table 2.1: Pmf's of the Poisson, GP, ZIGP and NB distribution together with their means and variances

reflects the dependence. Sklar (1959) shows that these characteristics can be separated, i.e.,

$$F_{\boldsymbol{Y}}(y_1, y_2) = C(F_1(y_1 \mid \boldsymbol{\theta}_1), F_2(y_2 \mid \boldsymbol{\theta}_2)|\boldsymbol{\tau}), \tag{2.2.1}$$

where $\boldsymbol{\tau}$ are the corresponding copula parameters. If a multivariate cdf of $\boldsymbol{Y}$ exists, there is a copula C which separates the dependence structure from the marginal distributions. If the margins are continuous, C is unique. Vice versa, according to (2.2.1) we can construct a multivariate cdf from two marginal distributions using a bivariate copula C. For a more detailed introduction to copulas, see for instance Joe (1997) or Nelsen (2006). Our sampling approach is based on Gaussian copulas.

Definition 1 (Gaussian copula)**.** The bivariate Gaussian copula is a function $C : [0,1]^2 \to [0,1]$ with

$$C(u_1, u_2|\tau_{12}) := \Phi_2\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2)|\tau_{12}\right), \tag{2.2.2}$$

where $\Phi_2(\cdot,\cdot|\tau_{12})$ is the cdf of the bivariate normal distribution with mean $\boldsymbol{\mu} = \mathbf{0}_2$ and covariance τ_{12} and $\Phi^{-1}(\cdot)$ is the inverse of the univariate standard normal cdf.

The bivariate Gaussian copula density is defined as

$$c(u_1, u_2|\tau_{12}) \quad = \quad \phi_2\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2)|\tau_{12}\right)\prod_{t=1}^{2}\frac{1}{\phi(\Phi^{-1}(u_t))},$$

where ϕ_2 is the bivariate normal pdf with mean $\boldsymbol{\mu} = \mathbf{0}_2$ and covariance τ_{12}.

The following monotonicity property has been shown by Cario and Nelson (1997, Propositions 1, 2 and Theorem 1).

Proposition 1. For $\boldsymbol{Y} := (Y_1, Y_2)'$ count variables with parameter $\boldsymbol{\theta} := (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$ and joint cdf $F_{\boldsymbol{Y}}(y_1, y_2|\boldsymbol{\theta}) = C(F_1(y_1|\boldsymbol{\theta}_1), F_2(y_2|\boldsymbol{\theta}_2)|\tau_{12})$ with a Gaussian copula C, the Pearson correlation between Y_1 and Y_2 denoted by $\rho_{12}(\tau_{12}, \boldsymbol{\theta})$ is a strictly monotone function in τ_{12}. Further $\rho_{12}(\tau_{12}, \boldsymbol{\theta}) > 0$ (< 0) for $\tau_{12} > 0$ (< 0) and $\rho_{12}(0, \boldsymbol{\theta}) = 0$. The maximum and minimum boundaries of ρ_{12} are reached for $\rho_{12}(1, \boldsymbol{\theta}) \leq 1$ and $\rho_{12}(-1, \boldsymbol{\theta}) \geq -1$.

Another important building block for our construction will be partial correlations. Partial correlation is the correlation between two variables while controlling for a third or more other

variables. Let $\boldsymbol{Z} = (Z_1, Z_2, \boldsymbol{Z}_3')'$ be a standardized T-component random vector, where $\boldsymbol{Z}_3 = (Z_3, \ldots, Z_T)'$ is a $(T-2)$-dimensional random vector. Let its correlation matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \boldsymbol{\sigma}_{13}' \\ \sigma_{12} & \sigma_{22} & \boldsymbol{\sigma}_{23}' \\ \boldsymbol{\sigma}_{13} & \boldsymbol{\sigma}_{23} & \boldsymbol{\Sigma}_{33} \end{pmatrix}, \; \boldsymbol{\Sigma}^{-1} =: \begin{pmatrix} \sigma^{11} & \sigma^{12} & \boldsymbol{\sigma}^{13\prime} \\ \sigma^{12} & \sigma^{22} & \boldsymbol{\sigma}^{23\prime} \\ \boldsymbol{\sigma}^{13} & \boldsymbol{\sigma}^{23} & \boldsymbol{\Sigma}^{33} \end{pmatrix}.$$

According to Srivastava and Khatri (1979, p. 53f), partial correlation between Z_1 and Z_2 while controlling $\boldsymbol{Z}_3$ denoted by $\sigma_{12|3:T}$ is defined as the correlation between $Z_1 - \boldsymbol{\sigma}_{13}'\Sigma_{33}^{-1}\boldsymbol{Z}_3$ and $Z_2 - \boldsymbol{\sigma}_{23}'\Sigma_{33}^{-1}\boldsymbol{Z}_3$, which is the correlation between Z_1 and Z_2 after eliminating the best linear effects of $\boldsymbol{Z}_3$ from both variables, and that $\sigma_{12|3:T} = \frac{-\sigma^{12}}{\sqrt{\sigma^{11}\sigma^{22}}}$. For $I := \{1, \ldots, T\}$ and for any subset $I^* \subseteq I$, which contains at least i, j and k, Pearson (1916) derived the recurrence

$$\sigma_{ij|I^*\setminus\{i,j,k\}} = \frac{\sigma_{ij|I^*\setminus\{i,j,k\}} - \sigma_{ik|I^*\setminus\{i,j,k\}} \cdot \sigma_{jk|I^*\setminus\{i,j,k\}}}{\sqrt{(1-\sigma^2_{ik|I^*\setminus\{i,j,k\}})(1-\sigma^2_{jk|I^*\setminus\{i,j,k\}})}}, \tag{2.2.3}$$

i.e., partial correlations of the $(T-2)$nd order can be calculated from the $(T-3)$rd order.

We now illustrate the construction of a multivariate density function of uniform random variables based on a PCC. We restrict to the trivariate case, for general dimension we refer to Aas, Czado, Frigessi, and Bakken (2009). Let $U_t \sim G_t, t = 1, 2, 3$ continuous uniform margins with joint density

$$g(u_1, u_2, u_3) = g_1(u_1) g_{2|1}(u_2|u_1) g_{3|12}(u_3|u_1, u_2).$$

Since $g_t(u_t) \equiv 1$ and $G_t(u_t) = u_t$, we have according to Sklar

$$\begin{aligned} g(u_1, u_2) &= c_{12}(u_1, u_2), \text{ therefore } g_{2|1}(u_2|u_1) = c_{12}(u_1, u_2) \\ \text{and } g(u_2, u_3|u_1) &= c_{23|1}(G_{2|1}(u_2|u_1), G_{3|1}(u_3|u_1)) \cdot g_{2|1}(u_2|u_1)\; g_{3|1}(u_3|u_1), \\ \text{hence } g_{3|12}(u_3|u_1, u_2) &= c_{23|1}(G_{2|1}(u_2|u_1), G_{3|1}(u_3|u_1)) \cdot g_{3|1}(u_3|u_1). \end{aligned}$$

The conditional cdf's $G_{2|1}(u_2|u_1)$ and $G_{3|1}(u_3|u_1)$ can be calculated according to Joe (1996). He proved under regularity conditions for $\boldsymbol{u} := (u_2, \ldots, u_T)'$, $\boldsymbol{u}_{-i} := (u_2, \ldots, u_{i-1}, u_{i+1}, \ldots, u_T)'$ that the conditional cdf at u_1 can be calculated as

$$G(u_1|\boldsymbol{u}) = \frac{\partial C\left(F(u_1|\boldsymbol{u}_{-i}), F(u_i|\boldsymbol{u}_{-i})|\tau_{u_1 u_i|\boldsymbol{u}_{-i}}\right)}{\partial F(u_i|\boldsymbol{u}_{-i})}. \tag{2.2.4}$$

Similar to Aas, Czado, Frigessi, and Bakken (2009) we abbreviate the expression for $\boldsymbol{u} = u_2$ and unconditional cdf's as margins in (2.2.4) as a function h. Then $h(u_2, u_1, \tau_{12}) := G_{2|1}(u_2|u_1)$ and $h(u_3, u_1, \tau_{13}) := G_{3|1}(u_3|u_1)$. Later on we will also need the conditional cdf's

$$\begin{aligned} G_{1|2}(u_1|u_2) &= h(u_1, u_2, \tau_{12}) \\ G_{3|12}(u_3|u_1, u_2, \boldsymbol{\Sigma}^Z) &= h\Big(h(u_3, u_1, \tau_{13}), h(u_2, u_1, \tau_{12}), \tau_{23|1}\Big). \end{aligned} \tag{2.2.5}$$

In the case of a bivariate Gaussian copula, Aas, Czado, Frigessi, and Bakken (2009) show that $h(u_1, u_2, \tau_{12}) = \Phi\left(\frac{\Phi^{-1}(u_1) - \tau_{12}\Phi^{-1}(u_2)}{\sqrt{1-\tau_{12}^2}}\right)$. The inverse of this h-function with respect to the first argument is $h^{-1}(u_1, u_2, \tau_{12}) = \Phi\left(\Phi^{-1}(u_1)\sqrt{1-\tau_{12}^2} + \tau_{12}\Phi^{-1}(u_2)\right)$ which may be regarded as a conditional quantile function of U_2 given U_1.

Now let $U_t := \Phi(Z_t)$, $t = 1, 2, 3$ with $\Phi(\cdot)$ the standard normal cdf. Further let $\boldsymbol{Z} := (Z_1, Z_2, Z_3)'$ and $\boldsymbol{Z} \sim N_3(\boldsymbol{0}, \boldsymbol{\Sigma}^Z)$, $\boldsymbol{\Sigma}^Z = (\tau_{ij})_{i,j=1,\ldots,3}$, $\tau_{ii} = 1$. Using all these building blocks, the corresponding PCC of $\boldsymbol{U} := (U_1, U_2, U_3)'$ is

$$g(u_1, u_2, u_3) = c_{12}(u_1, u_2|\tau_{12}) \cdot c_{23|1}(h(u_2|u_1, \tau_{12}), h(u_3|u_1, \tau_{13})|\tau_{23|1}) \cdot c_{13}(u_1, u_3|\tau_{13}), \tag{2.2.6}$$

where the copula densities c_{tt^*} and h refer to the Gaussian copula. Further τ_{12} and τ_{13} are the correlation of (Z_1, Z_2) and of (Z_1, Z_3), respectively. Finally, $\tau_{23|1}$ represents the conditional correlation between Z_2 and Z_3 given Z_1, i.e., $\text{cov}(Z_2, Z_3|Z_1)/\sqrt{\text{var}(Z_2|Z_1)\text{var}(Z_3|Z_1)}$. According to Kurowicka and Cooke (2006, Proposition 3.29), $\tau_{23|1}$ coincides in this case of Gaussian copulas with the partial correlation between Z_2 and Z_3 given Z_1, for which according to (2.2.3),

$$\tau_{23|1} = \frac{\tau_{23} - \tau_{12}\tau_{13}}{\sqrt{(1 - \tau_{12}^2)(1 - \tau_{13}^2)}}. \tag{2.2.7}$$

More general, for any elliptical, multivariate hypergeometric, multivariate negative hypergeometric, multinomial and Dirichlet distribution, partial and conditional correlation coincide (Baba, Shibata, and Sibuya (2004)).
In the construction of our approach, there are three levels of correlated rv's:

Level	Distribution	Correlation
(i) **Multivariate normal**	$(Z_1, Z_2, Z_3)' \sim N_3\left(\boldsymbol{0}, \boldsymbol{\Sigma}^Z\right)$	$(\boldsymbol{\Sigma}^Z)_{tt^*} = \tau_{tt^*}$
(ii) **Uniform**	$U_1, U_2, U_3 \sim unif(0,1)$, $U_t := \Phi(Z_t)$, $t = 1, \ldots, 3$. Joint density $g(u_1, u_2, u_3)$ according to PCC in (2.2.6) with Gaussian copulas.	τ_{12}, τ_{13}, $\tau_{23\|1}$
(iii) **Count rv**	$\boldsymbol{Y} := (Y_1, Y_2, Y_3)'$ counts, $Y_t := F_t^{-1}(U_t\|\boldsymbol{\theta}_t)$, $t = 1, 2, 3$ and $\boldsymbol{\theta}_t$ parameters of margin t. Further, $F_t^{-1}(U_t\|\boldsymbol{\theta}_t)$ is the pseudo-inverse of F_t at U_t (i.e., $F_t^{-1}(U_t\|\boldsymbol{\theta}_t) := \inf(Q \in \mathbb{N} \mid F_t(Q\|\boldsymbol{\theta}_t) = U_t)$).	$(\boldsymbol{\Sigma}^Y)_{tt^*} = \rho_{tt^*}$

Remarks:

- Since $\tau_{tt^*} \neq \text{corr}(Y_t, Y_{t^*})$, τ_{tt^*} is sometimes referred to as "association parameter".
- For known $\boldsymbol{\Sigma}^Z$ one may sample "top-down" from levels (i) to (ii) and use continuous $U_1, \ldots, U_T$ as a correlated input for (iii). This may be regarded as a multivariate generalization of inversion sampling.
- The bivariate margins of $\boldsymbol{U} := (U_1, U_2, U_3)'$ are given by $G(u_1, u_2) = C_{12}(u_1, u_2|\tau_{12})$, $G(u_1, u_3) = C_{13}(u_1, u_3|\tau_{13})$ and $G(u_2, u_3) = C_{23}(u_2, u_3|\tau_{23})$ with τ_{23} determined by (2.2.7).
- If we specify the PCC by correlations and partial correlations with any arbitrary values in $(-1, 1)$, then the corresponding $\boldsymbol{\Sigma}^Z$ is positive definite (see Joe (2006)).

2.3 Sampling trivariate count RV's

After some general background we will derive the algorithm given in Section 2.3.2.

2.3.1 Background

We provide an overview on the methodology our approach is based on and on how we adapt it for our purpose.

The general approach used is conditional sampling (e.g., see Biller and Ghosh (2006)), i.e., the sampling from a rv U_t using the conditional cdf of U_t given $U_1, \ldots, U_{t-1}$. To generate $T = 3$ correlated rv's, one samples $u_1, u_{2|1}, u_{3|12} \sim unif(0,1)$ independent. These $u_{t|1:(t-1)}, t = 1,2,3$ may be regarded as realizations from G_1 and the conditional cdfs $G_{2|1}$ and $G_{3|12}$, respectively. Then our sampling scheme reads

Algorithm 1 Conditional sampling combined with inversion sampling

Conditional sampling from PCC for $\boldsymbol{U}$	Inversion
(i) u_1	$y_1 = F_1^{-1}(u_1\|\boldsymbol{\theta}_1)$
(ii) $u_2 = G_{2\|1}^{-1}(u_{2\|1}\|u_1) \ = h^{-1}(u_{2\|1}, u_1, \tau_{12})$,	$y_2 = F_2^{-1}(u_2\|\boldsymbol{\theta}_2)$
(iii) $u_3 = G_{3\|12}^{-1}(u_{3\|12}\|u_1, u_2) \ = h^{-1}(h^{-1}(u_{3\|12}, h(u_2, u_1, \tau_{12}), \tau_{23\|1}), u_1, \tau_{13})$,	$y_3 = F_3^{-1}(u_3\|\boldsymbol{\theta}_3)$.

Here, the conditional quantiles have been determined using the expressions in (2.2.5). Note that this is only one possible order for determining the margins. A different ordering can be chosen by selecting a different PCC for $\boldsymbol{U}$. For this order, our sampling approach will approximate τ_{12} followed by $\tau_{23|1}$ and τ_{13}. For known copula parameters, this approach in general dimension has been applied to a C-vine by Aas, Czado, Frigessi, and Bakken (2009, Algorithm 1). We will now adapt their algorithm and determine (conditional) correlation parameters $\tau_{tt^*}(\boldsymbol{\Sigma}^Y, \boldsymbol{\theta})$ by minimizing the absolute deviation of empirical correlation coefficients from specified target correlations ρ_{tt^*}. The adaption is carried out in the following way: we sample independent *vectors* $\boldsymbol{u}_{t|1:(t-1)} := (u^1_{t|1:(t-1)}, \ldots, u^n_{t|1:(t-1)})' \in [0,1]^N$, $t = 1,2,3$ with $u^n_{t|1:(t-1)} \sim unif(0,1)$ i.i.d. and manipulate sequentially the a priori unknown copula parameters of the PCC (2.2.6) such that the empirical correlation of the vectors $\boldsymbol{u}_t \in [0,1]^N$, $t = 1,2,3$ obtained from Algorithm 1 is close to the desired target correlation. These manipulations are carried out immediately, i.e., while the conditional sampling algorithm is being executed, hence limiting the computational time. The problem of sampling from $\boldsymbol{Y}$ is the problem of approximating suitable copula parameters. Regarding existence and uniqueness of a solution, Biller and Nelson (2003) stress that in general searching the "true, correct" input model for such a problem is neither a theoretical supportable nor practically useful paradigm. Instead input modeling needs to be viewed as customizing a highly flexible model that can capture the important features of interest.

We assume the conditional correlations to be $\text{corr}(Y_2|y_1, Y_3|y_1) = \rho_{23|1}$ with

$$\rho_{23|1} := \frac{\rho_{23} - \rho_{12}\rho_{13}}{\sqrt{(1-\rho_{12}^2)(1-\rho_{13}^2)}}$$

for $\text{corr}(Y_1, Y_2) = \rho_{12}$, $\text{corr}(Y_1, Y_3) = \rho_{13}$ and $\text{corr}(Y_2, Y_3) = \rho_{23}$. This recursive expression is in fact a property of the partial correlations (2.2.3) and would only apply to multivariate normal rv's. We assume that the count data $(Y_1, Y_2, Y_3)'$ has approximately the same conditional correlation structure, i.e., that $\text{corr}(Y_2|y_1, Y_3|y_1) \approx \rho_{23|1}$. Set $\boldsymbol{\theta} := (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \boldsymbol{\theta}_3')'$. Our aim is to determine $\tau_{12} = \tau_{12}(\rho_{12}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, $\tau_{13} = \tau_{13}(\rho_{13}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_3)$ and $\tau_{23|1} = \tau_{23|1}(\boldsymbol{\Sigma}^Y, \boldsymbol{\theta})$, such that $\text{corr}(F_1^{-1}(U_1|\boldsymbol{\theta}_1), F_2^{-1}(U_2|\boldsymbol{\theta}_2), F_3^{-1}(U_3|\boldsymbol{\theta}_3)) \approx \boldsymbol{\Sigma}^Y$ and $(\Phi^{-1}(U_1), \Phi^{-1}(U_2), \Phi^{-1}(U_3)) \sim$

$N_3(\mathbf{0}_3, \mathbf{\Sigma}^Z(\mathbf{\Sigma}^Y, \boldsymbol{\theta}))$, with

$$\begin{aligned}\mathbf{\Sigma}^Z(\mathbf{\Sigma}^Y, \boldsymbol{\theta}) &:= \mathbf{\Sigma}_3\left[\tau_{12}(\rho_{12}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \tau_{13}(\rho_{13}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_3), \tau_{23}(\mathbf{\Sigma}^Y, \boldsymbol{\theta})\right] \\ &:= \begin{pmatrix} 1 & \tau_{12}(\rho_{12}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) & \tau_{13}(\rho_{13}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_3) \\ \tau_{12}(\rho_{12}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) & 1 & \tau_{23}(\mathbf{\Sigma}^Y, \boldsymbol{\theta}) \\ \tau_{13}(\rho_{13}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_3) & \tau_{23}(\mathbf{\Sigma}^Y, \boldsymbol{\theta}) & 1 \end{pmatrix}.\end{aligned}$$

Here, $\tau_{23}(\mathbf{\Sigma}^Y, \boldsymbol{\theta}) := \tau_{23|1}(\mathbf{\Sigma}^Y, \boldsymbol{\theta})\sqrt{(1-\tau_{12}^2(\rho_{12}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2))(1-\tau_{13}^2(\rho_{13}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_3))} + \tau_{12}(\rho_{12}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ $\tau_{13}(\rho_{13}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_3)$ can be obtained according to (2.2.7). Using the monotonicity property of Proposition 1, we can express the correlation and conditional correlation among the discrete margins Y_t as a function of the correlation parameters of $\boldsymbol{Z}$ and the marginal parameters of Y_t, i.e., $\rho_{tt^*} = \rho_{tt^*}(\tau_{tt^*}, \boldsymbol{\theta}_t, \boldsymbol{\theta}_{t^*})$ and $\rho_{tt^*|1:(t-1)} = \rho_{tt^*|1:(t-1)}(\mathbf{\Sigma}^Z, \boldsymbol{\theta})$.

2.3.2 Derivation of the sampling algorithm for $T = 3$

The sampling algorithm for $T := 3$ will be derived first and a corresponding pseudo-code given later. As preparation, choose the target correlations ρ_{tt^*}, the marginal distributions $F_t(\cdot|\theta_t)$, $t = 1, \ldots, T$. In general, the sample size will be N, i.e., we sample a set of rv's of dimension $N \times T$. According to our experience the algorithm works well even for small $N \geq 500$. Throughout the algorithm $T(T-1)/2$ bisection searches will be carried out. We sample $u^n_{t|1:(t-1)} \sim unif(0,1)$ i.i.d. for $t = 1, 2, 3$ and $n = 1, \ldots, N$.

Algorithm 1, (ii): Set $u_2^n(\tau_{12}) := G^{-1}_{2|1}(u^n_{2|1}|u^n_1) = h^{-1}(u^n_{2|1}, u^n_1, \tau_{12})$, therefore $u^n_{2|1}$ can be expressed as $u^n_{2|1} = G_{2|1}(u^n_2|u^n_1) = h(u^n_2, u^n_1, \tau_{12})$. Then, $\{u^n_1, u^n_2(\tau_{12}), n = 1, \ldots, N\}$ is a sample from the joint distribution of (U_1, U_2), i.e., $G_{12}(u^n_1, u^n_2(\tau_{12})) = C_{12}(u^n_1, u^n_2(\tau_{12})|\tau_{12})$. Set

$$y^n_1(\boldsymbol{\theta}_1) := F_1^{-1}(u^n_1|\boldsymbol{\theta}_1) \text{ and } y^n_2(\boldsymbol{\theta}_2, \tau_{12}) := F_2^{-1}(u^n_2(\tau_{12})|\boldsymbol{\theta}_2). \tag{2.3.1}$$

With $\bar{y}_1(\boldsymbol{\theta}_1) := \frac{1}{N}\sum_{n=1}^N y^n_1(\boldsymbol{\theta}_1)$ and $\bar{y}_2(\boldsymbol{\theta}_2, \tau_{12}) := \frac{1}{N}\sum_{n=1}^N y^n_2(\boldsymbol{\theta}_2, \tau_{12})$, let

$$\hat{\rho}_{12}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \tau_{12}) := \frac{\sum_{n=1}^N (y^n_1(\boldsymbol{\theta}_1) - \bar{y}_1(\boldsymbol{\theta}_1))(y^n_2(\boldsymbol{\theta}_2, \tau_{12}) - \bar{y}_2(\boldsymbol{\theta}_2, \tau_{12}))}{\sqrt{\sum_{n=1}^N (y^n_1(\boldsymbol{\theta}_1) - \bar{y}_1(\boldsymbol{\theta}_1))^2 \sum_{n=1}^N (y^n_2(\boldsymbol{\theta}_2, \tau_{12}) - \bar{y}_2(\boldsymbol{\theta}_2, \tau_{12}))^2}} \tag{2.3.2}$$

the empirical correlation coefficient based on the sample $\{y^n_1(\boldsymbol{\theta}_1), y^n_2(\boldsymbol{\theta}_2, \tau_{12}),\ n = 1, \ldots, N\}$. Using bisection, find τ_{12} such that $|\hat{\rho}_{12}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \tau_{12}) - \rho_{12}| < \varepsilon$, where ρ_{12} is the desired target correlation. Denote the optimal value by $\tau^{(1)}_{12}$ and set $u^n_2 := u^n_2(\tau^{(1)}_{12})$, $n = 1, \ldots, N$. Lower and upper boundaries for the initiation of the bisection search over τ_{12} will be $[-1, 0]$ for $\rho_{12} < 0$ and $[0, 1]$ for $\rho_{12} \geq 0$. Each evaluation of $\hat{\rho}_{12}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \tau_{12})$ requires the calculation of the two quantiles (2.3.1) and of the empirical correlation coefficient (2.3.2). An anonymous referee pointed out that since $y^n_1(\boldsymbol{\theta}_1)$ and $y^n_2(\boldsymbol{\theta}_2, \tau_{12})$ are discrete, $\hat{\rho}_{12}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \tau_{12})$ will only take distinct values. Hence for a very small N, a solution of the bisection might not exist, for $N \to \infty$ the probability of no solution converges to zero for fixed $\varepsilon > 0$. For $N = 500$ and $\varepsilon = 10^{-3}$ we never came across such a case. Also, let $\tau^{(1)}_{13}$ and $\tau^{(1)}_{23}$ unspecified and $\tau^{(1)}_{23|1} := \frac{\tau^{(1)}_{23} - \tau^{(1)}_{12}\tau^{(1)}_{13}}{\sqrt{(1-(\tau^{(1)}_{12})^2)(1-(\tau^{(1)}_{13})^2)}}$ (therefore, (2.2.3) holds). Set $\mathbf{\Sigma}^{Z(1)} := \mathbf{\Sigma}_3\left[\tau^{(1)}_{12}, \tau^{(1)}_{13}, \tau^{(1)}_{23}\right]$. Now $G_{12}(u^n_1, u^n_2|\tau^{(1)}_{12})$ is the joint distribution of U_1 and U_2 with association $\tau^{(1)}_{12}$. Then $\{(u^n_1, u^n_2), n = 1, \ldots, N\}$ is a sample from $C_{12}(u_1, u_2|\tau^{(1)}_{12})$.

(iii): According to (2.2.5) we have

$$u^n_{3|12} = G_{3|12}(u^n_3|u^n_1, u^n_2, \mathbf{\Sigma}^{Z(1)}) = h\Big(\underbrace{h(u^n_3, u^n_1, \tau^{(1)}_{13})}_{=:u^n_{3|1}(\tau^{(1)}_{23|1})}, \underbrace{h(u^n_2, u^n_1, \tau^{(1)}_{12})}_{=u^n_{2|1}}, \tau^{(1)}_{23|1}\Big), \tag{2.3.3}$$

where both $\tau_{13}^{(1)}$ and $\tau_{23|1}^{(1)}$ are unknown. For unspecified $\tau_{23|1}^{(1)}$, $u_{3|1}^n(\tau_{23|1}^{(1)})$ can be expressed as

$$u_{3|1}^n(\tau_{23|1}^{(1)}) = h^{-1}(u_{3|12}^n, h(u_2^n, u_1^n, \tau_{12}^{(1)}), \tau_{23|1}^{(1)}) = h^{-1}(u_{3|12}^n, u_{2|1}^n(\tau_{12}^{(1)}), \tau_{23|1}^{(1)}). \tag{2.3.4}$$

Now we carry out a simplified transformation which we discuss in Remark 2 afterwards. Set

$$y_{2|1}^n(\boldsymbol{\theta}_2, \tau_{12}^{(1)}) := F_2^{-1}(u_{2|1}^n(\tau_{12}^{(1)})|\boldsymbol{\theta}_2) \text{ and } y_{3|1}^n(\boldsymbol{\theta}_3, \tau_{23|1}^{(1)}) := F_3^{-1}(u_{3|1}^n(\tau_{23|1}^{(1)})|\boldsymbol{\theta}_3). \tag{2.3.5}$$

Using the empirical correlation coefficient based on the sample $\{y_{2|1}^n(\boldsymbol{\theta}_2, \tau_{12}^{(1)}), y_{3|1}^n(\boldsymbol{\theta}_3, \tau_{23|1}^{(1)}), n = 1, \ldots, N\}$, i.e.,

$$\hat{\rho}_{23|1}(\boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \tau_{23|1}) := \frac{\sum_{n=1}^N (y_{2|1}^n(\boldsymbol{\theta}_2) - \bar{y}_{2|1}(\boldsymbol{\theta}_2))(y_{3|1}^n(\boldsymbol{\theta}_3, \tau_{23|1}) - \bar{y}_{3|1}(\boldsymbol{\theta}_3, \tau_{23|1}))}{\sqrt{\sum_{n=1}^N (y_{2|1}^n(\boldsymbol{\theta}_2) - \bar{y}_{2|1}(\boldsymbol{\theta}_2))^2 \sum_{n=1}^N (y_{3|1}^n(\boldsymbol{\theta}_3, \tau_{23|1}) - \bar{y}_{3|1}(\boldsymbol{\theta}_3, \tau_{23|1}))^2}}, \tag{2.3.6}$$

where $\bar{y}_{2|1}(\boldsymbol{\theta}_2) := \frac{1}{N}\sum_{n=1}^N y_{2|1}^n(\boldsymbol{\theta}_2)$ and $\bar{y}_{3|1}(\boldsymbol{\theta}_3, \tau_{23|1}) := \frac{1}{N}\sum_{n=1}^N y_{3|1}^n(\boldsymbol{\theta}_3, \tau_{23|1})$, find by bisection $\tau_{23|1}$ such that $|\hat{\rho}_{23|1}(\boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \tau_{23|1}) - \rho_{23|1}| < \varepsilon$. Denote the optimal value by $\tau_{23|1}^{(2)}$ and set $u_{3|1}^n := u_{3|1}^n(\tau_{23|1}^{(2)})$ for $n = 1, \ldots, N$.

Remark 2. *If $\tau_{13}^{(1)}$ was known and $\tau_{23|1}^{(1)}$ to be determined by bisection, the correct way to proceed after (2.3.4) would be a) to set*

$$\begin{aligned} u_2^n &:= h^{-1}(u_{2|1}^n, u_1^n, \tau_{12}^{(1)}), \qquad y_2^n(\boldsymbol{\theta}_2, \tau_{12}^{(1)}) := F_2^{-1}(u_2^n|\boldsymbol{\theta}_2) \text{ and} \\ u_3^n(\tau_{23|1}^{(1)}) &:= h^{-1}(u_{3|1}^n(\tau_{23|1}^{(1)}), u_1^n, \tau_{13}^{(1)}), \qquad y_3^n(\boldsymbol{\theta}_3, \tau_{13}^{(1)}) := F_3^{-1}(u_3^n(\tau_{23|1}^{(1)})|\boldsymbol{\theta}_3) \end{aligned}$$

and b) to eliminate the best linear effect of $y_1^n(\boldsymbol{\theta}_1)$ from both variables and hence determine the empirical partial correlation. Then we could manipulate $\tau_{23|1}^{(1)}$ such that this partial correlation matches $\rho_{23|1}$ (since $\rho_{23|1}$ is a partial correlation). Note that h^{-1} can be viewed as a conditional quantile function and that for $\tau_{12}^{(1)} = \tau_{13}^{(1)} = 0$ they simplify to quantile functions, i.e., the dependency on the conditioning variable is suppressed. On the other hand, in our transformation we set a) $\tilde{u}_2^n := h^{-1}(u_{2|1}^n, u_1^n, 0) = u_{2|1}^n$ and $\tilde{u}_3^n(\tau_{23|1}^{(1)}) := h^{-1}(u_{3|1}^n(\tau_{23|1}^{(1)}), u_1^n, 0) = u_{3|1}^n(\tau_{23|1}^{(1)})$ and compensate for the suppressed dependency on u_1^n by matching the classical correlation coefficient with the target partial correlation. Note also that we set $\tau_{12}^{(1)} = \tau_{13}^{(1)} = 0$ only in this intermediate step. For the following steps we will not keep $\tilde{u}_2^n$ or $\tilde{u}_3^n(\tau_{23|1}^{(1)})$, we only keep an "optimal" $\tau_{23|1}^{(2)}$.

Now define $\tau_{12}^{(2)} := \tau_{12}^{(1)}$, $\tau_{13}^{(2)}$ unspecified and
$\tau_{23}^{(2)} := \tau_{23|1}^{(2)} \cdot \sqrt{(1-(\tau_{12}^{(2)})^2)(1-(\tau_{13}^{(2)})^2)} + \tau_{12}^{(2)}\tau_{13}^{(2)}$. Set $\boldsymbol{\Sigma}^{Z(2)} := \boldsymbol{\Sigma}_3\left[\tau_{12}^{(2)}, \tau_{13}^{(2)}, \tau_{23}^{(2)}\right]$.

Since according to (2.3.3), $u_{3|1}^n$ can also be expressed as $h(u_3^n, u_1^n, \tau_{13})$ (where τ_{13} has to be determined), we need to define

$$\begin{aligned} u_3^n(\tau_{13}) &:= h^{-1}(u_{3|1}^n, u_1^n, \tau_{13}), \\ y_1^n(\boldsymbol{\theta}_1) &:= F_1^{-1}(u_1^n|\boldsymbol{\theta}_1) \text{ and } y_3^n(\boldsymbol{\theta}_3, \tau_{13}) := F_3^{-1}(u_3^n(\tau_{13})|\boldsymbol{\theta}_3) \end{aligned}$$

and determine the empirical correlation of $\{y_1^n(\boldsymbol{\theta}_1), y_3^n(\boldsymbol{\theta}_3, \tau_{13}), n = 1, \ldots, N\}$, i.e.,

$$\hat{\rho}_{13}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_3, \tau_{13}) := \frac{\sum_{n=1}^N (y_1^n(\boldsymbol{\theta}_1) - \bar{y}_1(\boldsymbol{\theta}_1))(y_3^n(\boldsymbol{\theta}_3, \tau_{13}) - \bar{y}_3(\boldsymbol{\theta}_3, \tau_{13}))}{\sqrt{\sum_{n=1}^N (y_1^n(\boldsymbol{\theta}_1) - \bar{y}_1(\boldsymbol{\theta}_1))^2 \sum_{n=1}^N (y_3^n(\boldsymbol{\theta}_3, \tau_{13}) - \bar{y}_3(\boldsymbol{\theta}_3, \tau_{13}))^2}}, \tag{2.3.7}$$

where $\bar{y}_1(\boldsymbol{\theta}_1) := \frac{1}{N}\sum_{n=1}^{N} y_1^n(\boldsymbol{\theta}_1)$ and $\bar{y}_3(\boldsymbol{\theta}_3, \tau_{13}) := \frac{1}{N}\sum_{n=1}^{N} y_3^n(\boldsymbol{\theta}_3, \tau_{13})$. Using bisection find τ_{13} such that $|\hat{\rho}_{13}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_3, \tau_{13}) - \rho_{13}| < \varepsilon$. Denote the optimal value by $\tau_{13}^{(3)}$ and set $u_3^n := u_3^n(\tau_{13}^{(3)})$. Also, let $\tau_{12}^{(3)} := \tau_{12}^{(2)}$, $\tau_{23|1}^{(3)} := \tau_{23|1}^{(2)}$ and $\tau_{23}^{(3)} := \tau_{23|1}^{(3)} \cdot \sqrt{(1-(\tau_{12}^{(3)})^2)(1-(\tau_{13}^{(3)})^2)} + \tau_{12}^{(3)}\tau_{13}^{(3)}$. Set $\boldsymbol{\Sigma}^{Z(3)} := \boldsymbol{\Sigma}_3\left[\tau_{12}^{(3)}, \tau_{13}^{(3)}, \tau_{23}^{(3)}\right]$.
Now define $y_t^n := F^{-1}(u_t^n|\boldsymbol{\theta}_t)$ for $t = 1, 2, 3$. Then for each n (y_1^n, y_2^n, y_3^n) is a realization from the joint count distribution with margins $F(\cdot|\boldsymbol{\theta}_t)$ and approximate correlation $\boldsymbol{\Sigma}^Y$.

We will store all quantities to a lower-triangular matrix $\boldsymbol{V}^n = (v_{t,j}^n)_{t=1,\ldots,T,\ j=1,\ldots,t}$ for $n = 1, \ldots, N$ in the algorithm (identical to the scheme in Aas, Czado, Frigessi, and Bakken (2009, Algorithm 1)). Row t of $\boldsymbol{V}^n$ is associated with margin t. The first column of $\boldsymbol{V}^n$ will be initiated with $u_{t|1:(t-1)}^n$ and will be overwritten by certain $u_{t|t^*}^n$ and finally u_t^n. We may overwrite these values for efficient storing reasons. The following $t-1$ columns of row t will store the arguments of h necessary for the next row, i.e., the next margin $t+1$. For example, $v_{2,2}^n$ will store $u_{2|1}^n = G_{2|1}(u_2^n|u_1^n, \tau_{12})$ because it is necessary in order to sample $u_{3|1}^n$ (see (2.3.4)). A corresponding pseudo-code is presented in Algorithm 2. For simplicity we drop the superscript iteration index for the τ's.

Algorithm 2 Sampling from a C-vine in dimension 3 (explicit, avoiding loops)

Initiation:

- Determine target correlations $(\rho_{12}, \rho_{13}, \rho_{23})$ and $\rho_{23|1}$, where $\rho_{23|1} := \frac{\rho_{23}-\rho_{12}\rho_{13}}{\sqrt{(1-\rho_{12}^2)(1-\rho_{13}^2)}}$;
- Sample $u_{t|1:(t-1)}^n$, $t = 1, 2, 3$, $n = 1, \ldots, N$ independent uniform on $[0, 1]$;
- Set $v_{1,1}^n \leftarrow u_1^n, n = 1, \ldots, N$;

Begin:

(i)
for $n = 1, \ldots, N$ **do**
$v_{2,1}^n \leftarrow u_{2|1}^n$;
Calculate $u_2^n = G_{2|1}^{-1}(u_{2|1}^n|u_1^n, \tau_{12}) = h^{-1}(u_{2|1}^n, u_1^n, \tau_{12})$*:*
Using bisection, find τ_{12} such that $|\hat{\rho}_{12}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \tau_{12}) - \rho_{12}| < \varepsilon$,
where $\hat{\rho}_{12}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \tau_{12})$ is defined in (2.3.2);
$v_{2,1}^n \leftarrow h^{-1}(v_{2,1}^n, v_{1,1}^n, \tau_{12})$;
Calculate $u_{2|1}^n = G_{2|1}(u_2^n|u_1^n, \tau_{12}) = h(u_2^n, u_1^n, \tau_{12})$ *(will be needed for (ii)):*
$v_{2,2}^n \leftarrow h(v_{2,1}^n, v_{1,1}^n, \tau_{12})$;
end for

(ii)
for $n = 1, \ldots, N$ **do**
$v_{3,1}^n \leftarrow u_{3|12}^n$;
Calculate $u_{3|1}^n = G_{23|1}^{-1}(u_{3|12}^n|u_{2|1}^n, \tau_{23|1}) = h^{-1}(u_{3|12}^n, u_{2|1}^n, \tau_{23|1})$ *and*
$u_3^n = G_{3|1}^{-1}(G_{23|1}^{-1}(u_{3|12}^n|u_{2|1}^n, \tau_{23|1}), u_1^n, \tau_{13}) = h^{-1}(u_{3|1}^n, u_1^n, \tau_{13})$*:*
Using bisection, find $\tau_{23|1}$ such that $|\hat{\rho}_{23|1}(\boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \tau_{23|1}) - \rho_{23|1}| < \varepsilon$,
where $\hat{\rho}_{23|1}(\boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \tau_{23|1})$ according to (2.3.6);

Using bisection, find τ_{13} such that $|\hat{\rho}_{13}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_3, \tau_{13}) - \rho_{13}| < \varepsilon$,
where $\hat{\rho}_{13}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_3, \tau_{13})$ is defined in (2.3.7);

$v_{3,1}^n \leftarrow h^{-1}(v_{3,1}^n, v_{2,2}^n, \tau_{23|1})$;
$v_{3,1}^n \leftarrow h^{-1}(v_{3,1}^n, v_{1,1}^n, \tau_{13})$;
end for

$y_1^n \leftarrow F_1^{-1}(v_{1,1}^n|\theta_1)$, $y_2^n \leftarrow F_2^{-1}(v_{2,1}^n|\theta_2)$ and $y_3^n \leftarrow F_3^{-1}(v_{3,1}^n|\theta_3)$;
Return: y_1^n, y_2^n, y_3^n.

2.4 Sampling T-variate count RV's

For high-dimensional distributions there are many possible pair-copula decompositions for the same multivariate distribution. Bedford and Cooke (2001b) introduced a graphical model called regular vine to help organize them. Decomposition (2.2.6) is a vine copula in dimension 3. Vines have been constructed and investigated by Kurowicka and Cooke (2006) while Aas, Czado, Frigessi, and Bakken (2009) study the statistical inference problem. In general, there are canonical vines (C-vines) and D-vines. We focus on C-vines but the approach can be applied to D-vines as well.

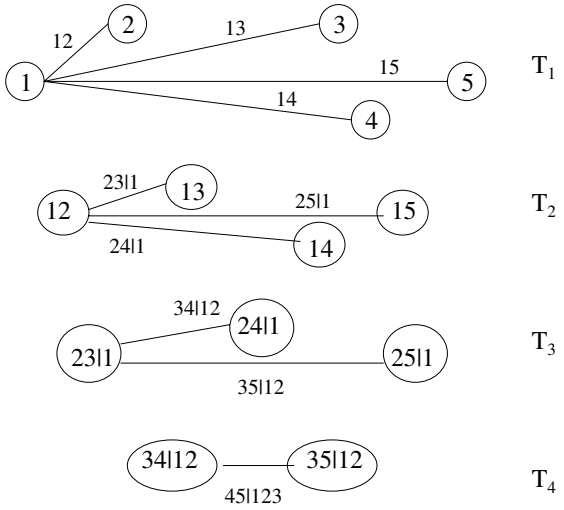

Figure 2.1: Five dimensional C-vine.

A T-dimensional regular vine is represented by $T-1$ trees. Tree j has $T+1-j$ nodes and $T-j$ edges. Each edge corresponds to a pair-copula density. Edges in tree T become nodes in tree $j+1$. Two nodes in tree $j+1$ are joined by an edge if the corresponding edges in tree j share a node. The complete decomposition is defined by the $\frac{T(T-1)}{2}$ edges (i.e., pair-copula densities) and the marginal densities. A *C-vine* in particular is a regular vine for which *each tree has a unique node that is connected to $T-j$ edges*. Note that for dimension 3, the C-vine and D-vine are identical. A graphical illustration of the five-dimensional C-vine is given

in Figure 2.1. Here, the unique node connected to the $T-j$ edges is chosen as 1 in T_1 and $t(t+1)|1:(t-1)$ in T_2 to T_4. Without loss of generality, we will base all our simulations on this decomposition. Once we have chosen the connecting node in every tree, the decomposition of the multivariate distribution and the order in which partial correlations will be determined in our algorithm, is fixed. One starts with margin 1 in T_1 and determines the correlation τ_{12} since the connecting node 12 in T_2 represents τ_{12}. Then in T_1, node 1 is connected by edge 12 to node 2, therefore we obtain margin 2. If we continue, we determine the partial correlation $\tau_{23|1}$, which is the connecting node in T_3, hence this edge in T_2 will connect 12 (τ_{12} is already determined) to 13. The recursive expression for the partial correlations (2.2.3) allows for a one-to-one relationship between the correlations and partial correlations of the multivariate normal distribution (see Joe (2006)). Thus the multivariate distribution defined by the Gaussian copula is fully specified as well. After determining τ_{13} we can sample margin 3 (see T_1). The next partial correlations to be obtained in order to sample margin 4 will be $\tau_{34|12}$, $\tau_{24|1}$ and τ_{14}. Finally, we will proceed with $\tau_{45|123}$, $\tau_{35|12}$, $\tau_{25|1}$ and τ_{15}. In general, we will generate samples $\{y^n_{k|1:(k-1)}(\boldsymbol{\theta}_k), y^n_{t|1:(k-1)}(\boldsymbol{\theta}_t, \tau_{kt|1:(k-1)}), n = 1, \ldots, N\}$, where $y^n_{k|1:(k-1)}(\boldsymbol{\theta}_k) := F_k^{-1}(u^n_{k|1:(k-1)}|\boldsymbol{\theta}_k)$ and $y^n_{t|1:(k-1)}(\boldsymbol{\theta}_t, \tau_{kt|1:(k-1)}) := F_t^{-1}(u^n_{t|1:(k-1)}(\tau_{kt|1:(k-1)})|\boldsymbol{\theta}_t)$. Similar to (2.3.4), for independent $u^n_1, \ldots, u^n_{T|1:(T-1)}$ we recursively define for $t = 2, \ldots, T$ and $k = 1, \ldots, t-1$

$$\begin{aligned} u^n_{t|1:(k-1)}(\tau_{kt|1:(k-1)}) &:= h^{-1}(u^n_{t|1:k}, u^n_{k|1:(k-1)}(\tau_{k(k-1)|1:(k-2)})|\tau_{kt|1:(k-1)}) \\ u^n_{k|1:(k-1)}(\tau_{k(k-1)|1:(k-2)}) &:= h(u^n_{k|1:(k-2)}(\tau_{k(k-1)|1:(k-2)}), u^n_{k-1|1:(k-2)}|\tau_{k(k-1)|1:(k-2)}). \end{aligned}$$

The approximating partial correlation coefficient of the sample is given by

$$\hat{\rho}_{kt|1:(k-1)}(\boldsymbol{\theta}_k, \boldsymbol{\theta}_t, \tau_{kt|1:(k-1)}) := \frac{\sum_{n=1}^N (y^n_{k|1:(k-1)}(\boldsymbol{\theta}_k) - \bar{y}_{k|1:(k-1)}(\boldsymbol{\theta}_k))(y^n_{t|1:(k-1)}(\boldsymbol{\theta}_t, \tau_{kt|1:(k-1)}) - \bar{y}_{t|1:(k-1)}(\boldsymbol{\theta}_t, \tau_{kt|1:(k-1)}))}{\sqrt{\sum_{n=1}^N (y^n_{k|1:(k-1)}(\boldsymbol{\theta}_k) - \bar{y}_{k|1:(k-1)}(\boldsymbol{\theta}_k))^2 \sum_{n=1}^N (y^n_{t|1:(k-1)}(\boldsymbol{\theta}_t, \tau_{kt|1:(k-1)}) - \bar{y}_{t|1:(k-1)}(\boldsymbol{\theta}_t, \tau_{kt|1:(k-1)}))^2}}, \tag{2.4.1}$$

where $\bar{y}_{k|1:(k-1)}(\boldsymbol{\theta}_k) := \frac{1}{N}\sum_{n=1}^N y^n_{k|1:(k-1)}(\boldsymbol{\theta}_k)$ and $\bar{y}_{t|1:(k-1)}(\boldsymbol{\theta}_t, \tau_{kt|1:(k-1)}) := \frac{1}{N}\sum_{n=1}^N y^n_{t|1:(k-1)}(\boldsymbol{\theta}_t, \tau_{kt|1:(k-1)})$.

2.4.1 Sampling algorithm in dimension T

Algorithm 3 Sampling from a C-vine in dimension T

Initiation:

- Determine target correlations;
- Sample $u^n_{t|1:(t-1)}$, $t = 1, \ldots, T$, $n = 1, \ldots, N$ independent uniform on $[0, 1]$;
- Set $v^n_{1,1} \leftarrow u^n_1, n = 1, \ldots, N$;

Begin:

for $t = 2, \ldots, T$ and $n = 1, \ldots, N$ **do**

$v^n_{t,1} \leftarrow u^n_{t|1:(t-1)}$;

for $k = t-1$ to 1 **do**

Using bisection, find $\tau_{kt|1:(k-1)}$ such that
$|\hat{\rho}_{kt|1:(k-1)}(\boldsymbol{\theta}_k, \boldsymbol{\theta}_t, \tau_{kt|1:(k-1)}) - \rho_{kt|1:(k-1)}| < \varepsilon$,
where $\hat{\rho}_{kt|1:(k-1)}(\boldsymbol{\theta}_k, \boldsymbol{\theta}_t, \tau_{kt|1:(k-1)})$ is defined in (2.4.1);

Set $v^n_{t,1} \leftarrow h^{-1}(v^n_{t,1}, v^n_{k,k}, \tau_{kt|1:(k-1)})$;
end for
$u^n_t \leftarrow v^n_{t,1}$;
if $t < T$ **then**
for $j = 1$ to $t-1$ **do**
$v^n_{t,j+1} \leftarrow h(v^n_{t,j}, v^n_{j,j}, \tau_{jt|1:(j-1)})$;
end for
end if
end for
$y^n_t \leftarrow F_t^{-1}(v^n_{t,1}|\theta_t)$, $t = 1, \ldots, T$, $n = 1, \ldots, N$;
Return: y^n_t, $t = 1, \ldots, T$.

In the *R* package *corcounts* the marginal distributions can be specified as Poisson, generalized Poisson, zero-inflated Poisson, zero-inflated generalized Poisson or negative binomial distribution with marginal parameters as well as the target correlation set by the users. Then correlated rv's from these distributions having approximately the specified correlation will be sampled. The runtime needed for the generation of one set of variables depends not only on the dimension of the problem but also on how fast quantiles of the desired marginal distribution can be calculated. Only the Poisson and the negative binomial quantile functions are part of the *R* base functions. For all other distributions, the quantile functions are part of the "corcounts" package and are not as fast. The runtime in seconds for generating one set of $T \times N$ variables, $T = 2, \ldots, 5$, $N = 500, 2000, 5000$ for different marginal distributions on a PC with Intel processor, 2GHz and 1GB RAM is given in Table 2.2.

	$T=2$			$T=3$			$T=4$			$T=5$		
N	500	2000	5000	500	2000	5000	500	2000	5000	500	2000	5000
Poi	0.03	0.04	0.06	0.06	0.19	0.42	0.15	0.43	0.96	0.18	0.79	1.96
NB	0.05	0.19	0.45	0.13	0.39	1.00	0.26	0.80	1.98	0.40	1.57	3.64
GP	0.48	2.58	5.46	1.70	5.24	14.59	3.08	12.45	27.71	4.71	23.66	52.05
ZIP	0.34	1.59	5.57	0.98	5.33	11.95	2.58	9.60	25.36	4.69	21.87	52.86
ZIGP	0.40	0.69	5.64	1.10	4.84	12.06	2.64	8.94	21.31	7.70	20.02	58.95

Table 2.2: Runtime in seconds for generating one set of $T \times N$ variables, $T = 2, \ldots, 5$, number of observations $N = 500, 2000, 5000$ in each cell

2.5 NORTA Sampling Method with Illustration to NB count Data

In this section we will compare our sampling approach to a naive approach and the NORTA approach for sampling count rv's. The naive approach is to use our desired target correlation

$\mathbf{\Sigma}^Y$ and generate for a sample of N subjects n-dimensional multivariate normal random vectors with covariance $\mathbf{\Sigma}^Y$, i.e. $\boldsymbol{Z}_k \sim N_n(\mathbf{0}, \mathbf{\Sigma}^Y)$, $k = 1, \ldots, N$. Next we transform the sample $\boldsymbol{z}_k = (z_{k1}, \ldots, z_{kn})'$ to the uniform level $\boldsymbol{u}_k := (\Phi(z_{k1}), \ldots, \Phi(z_{kn}))'$, $k = 1, \ldots, N$ and determine the sample correlation $\hat{\mathbf{\Sigma}}^U$ of $\{\boldsymbol{u}_k, k = 1, \ldots, N\}$. Then we generate outcomes according to the generalized Poisson distribution (see Table 2.1) with cdf F_i by determining the quantiles of the GP distribution with mean μ_i and variance $\mu_i\varphi_i^2$ at u_{ki}, $k = 1, \ldots, N$, $i = 1, \ldots, n$, i.e. $y_{ki}^{naive} := F_i^{-1}(u_{ki}|\mu_i, \varphi_i)$, and $\boldsymbol{y}_k^{naive} := (y_{k1}^{naive}, \ldots, y_{kn}^{naive})'$. The sample correlation of $\{\boldsymbol{y}_k^{naive}, k = 1, \ldots, N\}$ will be denoted by $\hat{\mathbf{\Sigma}}^{Y^{naive}}$.

As second benchmark, we will consider the NORTA method(see Cario and Nelson (1997) and Avramidis, Channouf, and L'Ecuyer (2009)): the Pearson correlation between Y_t and Y_{t^*} is given by $\rho_{tt^*} := \text{corr}(Y_t, Y_{t^*}) = \frac{E(Y_t,Y_{t^*})-E(Y_t)E(Y_{t^*})}{\sqrt{\text{var}(Y_t)\text{var}(Y_{t^*})}}$. If we define a bivariate cdf of (Y_t, Y_{t^*}) by a bivariate Gaussian copula with correlation parameter τ_{tt^*}, i.e. $F_2(y_t, y_{t^*}|\tau_{tt^*}) := C_{tt^*}(F_t(y_t \mid \boldsymbol{\theta}_t), F_{t^*}(y_{t^*} \mid \boldsymbol{\theta}_{t^*})|\tau_{tt^*})$, we may write $\rho_{tt^*}(\tau_{tt^*}) := \frac{g(Y_t,Y_{t^*}|\tau_{tt^*})-E(Y_t)E(Y_{t^*})}{\sqrt{\text{var}(Y_t)\text{var}(Y_{t^*})}}$, where

$$g(Y_t, Y_{t^*}|\tau_{tt^*}) = E(Y_tY_{t^*}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F_t^{-1}(\Phi(z_t))F_{t^*}^{-1}(\Phi(z_{t^*}))\phi_2(z_t, z_{t^*}|\tau_{tt^*})dz_tdz_{t^*}.$$

For $z_{t,i} := \Phi^{-1}(F_t(i))$ and $z_{t^*,j} := \Phi^{-1}(F_{t^*}(j))$

$$\begin{aligned} g(Y_t, Y_{t^*}|\tau_{tt^*}) &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} ij \left(\int_{z_{t,i-1}}^{z_{t,i}}\int_{z_{t^*,j-1}}^{z_{t^*,j}} \phi_2(z_t, z_{t^*}|\tau_{tt^*})dz_tdz_{t^*}\right) \\ &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} ij\Big[\Phi_2(z_{t,i-1}, z_{t^*,j-1}|\tau_{tt^*}) - \Phi_2(z_{t,i}, z_{t^*,j-1}|\tau_{tt^*}) \\ &\quad -\Phi_2(z_{t,i-1}, z_{t^*,j}|\tau_{tt^*}) + \Phi_2(z_{t,i}, z_{t^*,j}|\tau_{tt^*})\Big] \\ &\approx \sum_{i=0}^{K_t}\sum_{j=0}^{K_{t^*}} \Phi_2(z_{t,i}, z_{t^*,j}|\tau_{tt^*}), \end{aligned} \tag{2.5.1}$$

where K_t and K_{t^*} are defined as the quantiles of F_t and F_{t^*} at some value close to 1 (we use $1-10^{-6}$). According to Srivastava and Khatri (1979, p. 48), the derivative of $g(Y_t, Y_{t^*}|\tau_{tt^*})$ w.r.t. τ_{tt^*} is $\frac{\partial}{\partial\tau_{tt^*}}g(Y_t, Y_{t^*}|\tau_{tt^*}) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\phi_2(z_{t,i}, z_{t^*,j}|\tau_{tt^*})$, where we use $\frac{\partial}{\partial\tau_{tt^*}}\Phi_2(z_{t,i}, z_{t^*,j}|\tau_{tt^*}) = \frac{\partial^2}{\partial z_{t,i}\partial z_{t^*,j}}\Phi_2(z_{t,i}, z_{t^*,j}|\tau_{tt^*})$. An implementation for rank correlations has been implemented by Avramidis, Channouf, and L'Ecuyer (2009) in Java. It can easily be altered to deal with Pearson correlation. The implementation is available at

http://www.iro.umontreal.ca/ lecuyer/myftp/nortadisc/java/ .

In the Java implementation we use the NI2A method, one out of four suggested methods for finding the root of $f(\tau_{tt^*}) := g(Y_t, Y_{t^*}|\tau_{tt^*}) - E(Y_t)E(Y_{t^*}) - \rho_{tt^*}\sqrt{\text{var}(Y_t)\text{var}(Y_{t^*})}$ where the derivative of f is identical to that of g. NI2A finds a root of f by numerically integrating its derivative, for details see Avramidis, Channouf, and L'Ecuyer (2009, Subsection 3.1.3).

Now if $\mathbf{\Sigma} := (\tau_{tt^*})_{t,t^*=1,\ldots,T}$ is not positive definite one can perform a correction step to obtain a 'close' positive definite correlation matrix. We use the eigenvalue correction (see Ghosh and Henderson (2003)) where $\mathbf{\Sigma}$ is decomposed into eigenvalues and eigenvectors and all negative eigenvalues are set to some $\varepsilon > 0$ (in our case, $\varepsilon := 10^{-15}$). The corrected correlation matrix will be denoted by $\mathbf{\Sigma}^{NORTA}$. We use $\mathbf{\Sigma}^{NORTA}$ and generate for a sample of N subjects T-dimensional

multivariate normal random vectors with covariance $\mathbf{\Sigma}^{NORTA}$, i.e. $\boldsymbol{Z}_n \sim N_T(\mathbf{0}, \mathbf{\Sigma}^{NORTA})$, $n = 1, \ldots, N$. Next we transform the sample $\boldsymbol{z}_n = (z_{n1}, \ldots, z_{nT})'$ to the uniform level $\boldsymbol{u}_n := (\Phi(z_{n1}), \ldots, \Phi(z_{nT}))'$, $n = 1, \ldots, N$ and proceed as in the naive approach by determining the quantiles at $u_{nt} := \Phi(z_{nt})$, $n = 1, \ldots, N$, $t = 1, \ldots, T$, i.e. $y_{nt}^{NORTA} := F_t^{-1}(u_{nt}|\mu_t, \psi_t)$, and $\boldsymbol{y}_n^{NORTA} := (y_{n1}^{NORTA}, \ldots, y_{nT}^{NORTA})'$. The sample correlation of $\{\boldsymbol{y}_n^{NORTA}, n = 1, \ldots, N\}$ will be denoted by $\hat{\mathbf{\Sigma}}^{Y^{NORTA}}$.

For a first illustrative comparison of the algorithms we generate outcomes according to the Negative Binomial distribution (see Table 2.1) with cdf F_t, mean μ_t and variance $\mu_t(1 + \frac{\mu_t}{\psi_t})$. We choose as sampling setting $T = 8$ and $N = 100\ 000$. Also we use as target correlation matrix an exchangeable structure, i.e., $\mathbf{\Sigma}^Y = (\rho_{tt^*})$ with $\rho_{tt^*} = 0.6\ \forall t \neq t^*$ and $\rho_{tt} = 1$. Marginal means of the eight-dimensional NB distribution were set to $\boldsymbol{\mu} := (4, 25, 120, 2, 28, 7, 27, 5)'$, size parameters to $\boldsymbol{\psi} := (3.2, 2.22, 40, 0.38, 9.33, 0.88, 21.6, 0.95)'$.

Sampling based on the naive approach results in a sample of count variables whose correlation matrix is estimated to

$$\hat{\mathbf{\Sigma}}^{Y^{naive}} = \begin{pmatrix} 1.0000, 0.5789, 0.5813, 0.5081, 0.5822, 0.5523, 0.5787, 0.5567 \\ 0.5789, 1.0000, 0.5795, 0.5098, 0.5784, 0.5583, 0.5737, 0.5545 \\ 0.5813, 0.5795, 1.0000, 0.4888, 0.5991, 0.5450, 0.5957, 0.5483 \\ 0.5081, 0.5098, 0.4888, 1.0000, 0.4963, 0.5118, 0.4881, 0.5134 \\ 0.5822, 0.5784, 0.5991, 0.4963, 1.0000, 0.5535, 0.5935, 0.5535 \\ 0.5523, 0.5583, 0.5450, 0.5118, 0.5535, 1.0000, 0.5430, 0.5482 \\ 0.5787, 0.5737, 0.5957, 0.4881, 0.5935, 0.5430, 1.0000, 0.5452 \\ 0.5567, 0.5545, 0.5483, 0.5134, 0.5535, 0.5482, 0.5452, 1.0000 \end{pmatrix},$$

where the off-diagonal average absolute deviation from the target correlation is 0.0494 Sampling based on NORTA gives

$$\hat{\mathbf{\Sigma}}^{Y^{NORTA}} = \begin{pmatrix} 1.0000, 0.5925, 0.5979, 0.5285, 0.5964, 0.5673, 0.5942, 0.5714 \\ 0.5925, 1.0000, 0.5813, 0.5373, 0.5846, 0.5658, 0.5799, 0.5680 \\ 0.5979, 0.5813, 1.0000, 0.5111, 0.5992, 0.5491, 0.5970, 0.5554 \\ 0.5285, 0.5373, 0.5111, 1.0000, 0.5183, 0.5401, 0.5103, 0.5371 \\ 0.5964, 0.5846, 0.5992, 0.5183, 1.0000, 0.5575, 0.5960, 0.5645 \\ 0.5673, 0.5658, 0.5491, 0.5401, 0.5575, 1.0000, 0.5514, 0.5593 \\ 0.5942, 0.5799, 0.5970, 0.5103, 0.5960, 0.5514, 1.0000, 0.5578 \\ 0.5714, 0.5680, 0.5554, 0.5371, 0.5645, 0.5593, 0.5578, 1.0000 \end{pmatrix}.$$

The average absolute deviation from the target correlation is 0.0368. On the other hand, using the C-vine approach, we get

$$\hat{\mathbf{\Sigma}}^{Y} = \begin{pmatrix} 1.0000, 0.6026, 0.6095, 0.5977, 0.5992, 0.5921, 0.6005, 0.5917 \\ 0.6026, 1.0000, 0.6074, 0.6012, 0.6056, 0.6028, 0.6042, 0.6049 \\ 0.6095, 0.6074, 1.0000, 0.5599, 0.6149, 0.5882, 0.6225, 0.5888 \\ 0.5977, 0.6012, 0.5599, 1.0000, 0.5709, 0.6223, 0.5635, 0.6204 \\ 0.5992, 0.6056, 0.6149, 0.5709, 1.0000, 0.5896, 0.6228, 0.5928 \\ 0.5921, 0.6028, 0.5882, 0.6223, 0.5896, 1.0000, 0.5925, 0.6188 \\ 0.6005, 0.6042, 0.6225, 0.5635, 0.6228, 0.5925, 1.0000, 0.5941 \\ 0.5917, 0.6049, 0.5888, 0.6204, 0.5928, 0.6188, 0.5941, 1.0000 \end{pmatrix},$$

where the off-diagonal absolute deviations from the target correlations have an average value of 0.0121. This shows that we can expect gains of the C-vine sampling approach compared to the naive also also the NORTA one.

2.6 Simulation Study

We want to investigate the small sample performance of our proposed sampling method for a correlated count vector $\boldsymbol{Y} = (Y_1, \ldots, Y_T)$ with target correlation $\rho_{tt^*} = \text{corr}(Y_t, Y_{t^*})$, $1 \leq t < t^* \leq T$. For this we use the mean absolute deviation with respect to the target correlation: we generate i.i.d. samples from $\boldsymbol{Y}$ using either the C-vine, naive or NORTA sampling method. Based on the i.i.d. sample we assess the performance by the absolute deviation of the corresponding empirical correlations of the sampled values to the target correlations. To estimate this deviation we use R replications of the N dimensional sample of $\boldsymbol{Y}$. In particular we denote by $\boldsymbol{y}_n^r = (y_{n1}, \ldots, y_{nT})$ the nth sampled count random vector of replication r, $r = 1, \ldots, R$. Then the deviation is estimated by $\widehat{d}_{tt^*} := \frac{1}{R}\sum_{r=1}^{R} |\hat{\rho}_{tt^*}^r - \rho_{tt^*}|$, where $\hat{\rho}_{tt^*}^r := \frac{\sum_{n=1}^{N}(y_{nt}^r - \bar{y}_t^r)(y_{nt^*}^r - \bar{y}_{t^*}^r)}{\sqrt{\sum_{n=1}^{N}(y_{nt}^r - \bar{y}_t^r)^2}\sqrt{\sum_{n=1}^{N}(y_{nt^*}^r - \bar{y}_{t^*}^r)^2}}$ and $\bar{y}_t^r := \frac{1}{N}\sum_{n=1}^{N} y_{nt}^r \ \forall t = 1, \ldots, T$. We consider the mean absolute value of $\widehat{d}_{tt^*}$, i.e. $MEANAD := \frac{1}{T(T-2)/2}\sum_{1\leq t<t^*\leq T} |\widehat{d}_{tt^*}|$ as an overall performance measure. Note that the deviation estimates $\widehat{d}_{tt^*}$ are dependent since the components of $\boldsymbol{y}_n^r$ are correlated. Therefore the standard error of $MEANAD$ cannot be estimated easily, hence we do not consider it. Nevertheless, the average of $\widehat{d}_{tt^*}$ is still a consistent estimator of its mean.

In order to get a first insight on the accuracy of the C-vine algorithm we choose a trivariate example, i.e., we choose three Poisson margins with mean 10 and an exchangeable target correlation with $\rho = 0.5$ and choose a random N uniform from $[500, 10000]$. For $\varepsilon = 10^{-2}$ we repeat the sampling $R = 1000$ times and report empirical quantiles of $\hat{d}_{12}$, $\hat{d}_{13}$ and $\hat{d}_{23}$. We see in Table 2.3 that even for the pair of margins $(2, 3)$ for which τ_{23} is determined via partial correlations $\tau_{23|1}$ the median absolute deviation $\hat{d}_{23}$ is below ε which demonstrates the accuracy of our approach. Now we proceed with a general simulation study investigating many parameter

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\hat{d}_{12}$	0.0000	0.0016	0.0039	0.0042	0.0064	0.0100
$\hat{d}_{13}$	0.0000	0.0019	0.0041	0.0043	0.0065	0.0100
$\hat{d}_{23}$	0.0000	0.0031	0.0080	0.0165	0.0162	0.0955

Table 2.3: Empirical quantiles of $\hat{d}_{12}$, $\hat{d}_{13}$ and $\hat{d}_{23}$ for C-vine sampling in a trivariate setting with $Poi(10)$ margins and target exchangeable correlation of 0.5, $R = 1000$, $\varepsilon = 10^{-2}$ and N random uniform between 500 and 10000.

settings. While again we set $R = 1000$ we now fix $N = 1000$. We consider the four distributions introduced in Section 2.2. Marginal parameters $\boldsymbol{\theta}_t$ are μ_t in the Poisson case, (μ_t, φ_t) in the GP case, $(\mu_t, \varphi_t, \omega_t)$ in the ZIGP case and (μ_t, ψ_t) in the NB case. Variances $Var(Y_{rit})$ will be equal in the GP and NB case if we set $\varphi_t^2 = 1 + \frac{\mu_t}{\psi_t}$ or equivalently $\psi_t = \frac{\mu_t}{\varphi_t^2 - 1}$.

We investigate the influence the dimension T and the size of the correlation in an exchangeable and an AR(1) target correlation structure, i.e., $\rho_{tt^*} = \rho$ and $\rho_{tt^*} = \rho^{|t-t^*|}$, respectively. The settings were $\rho \in \{0.1, 0.4, 0.7\}$, $T \in \{2, 5, 10\}$. Medium sized marginal parameters according to Table 2.4 were used. Results are summarized in Table 2.5. According to Table 2.5, in all settings chosen, $MEANAD$ is lower for C-vine sampling than for the other approaches. Whereas for NORTA $MEANAD$ increases slowly with the dimension T it increases faster for C-vine sampling. For NORTA the explanation is given by Ghosh and Henderson (2003) who show that the probability of a negative definite correlation matrix produced by NORTA in-

	T	Parameters
Poi	2	$\boldsymbol{\mu} := (8, 20)'$
	5	$\boldsymbol{\mu} := (8, 20, 11, 9, 13)'$
	10	$\boldsymbol{\mu} := (8, 20, 11, 9, 13, 19, 5, 27, 12, 10)'$
GP		$\boldsymbol{\mu}$ as in Poisson case
	2	$\boldsymbol{\varphi} := (1.5, 1.5)'$
	5	$\boldsymbol{\varphi} := (1.5, 1.5, 2, 3.5, 1.5)'$
	10	$\boldsymbol{\varphi} := (1.5, 1.5, 2, 3.5, 1.5, 2.5, 3, 2, 1.5, 2)'$
ZIGP		$\boldsymbol{\mu}$ and $\boldsymbol{\varphi}$ as in GP case
	2	$\boldsymbol{\omega} := (0.2, 0.25)'$
	5	$\boldsymbol{\omega} := (0.2, 0.25, 0.15, 0.3, 0.1)'$
	10	$\boldsymbol{\omega} := (0.2, 0.25, 0.15, 0.3, 0.1, 0.2, 0.15, 0.05, 0.24, 0.1)'$
NB		$\boldsymbol{\mu}$ as in Poisson case
	2	$\boldsymbol{\psi} := (6.4, 16)'$
	5	$\boldsymbol{\psi} := (6.4, 16, 3\frac{2}{3}, 0.8, 10.4)'$
	10	$\boldsymbol{\psi} := (6.4, 16, 3\frac{2}{3}, 0.8, 10.4, 3.62, 0.625, 9, 9.6, 3\frac{1}{3})'$

Table 2.4: Marginal parameter choices for $T = 2$, 5 and 10 and exchangeable correlation structure for different marginal distributions (marginal variances for GP and NB margins are chosen to be equal)

creases with dimension and is around 85% for $T = 10$. In our approach, however, the reason for this behavior is simply error propagation, since for larger T a greater share of association parameters needs to be determined indirectly via partial correlations. While a higher target correlation ρ leads to an increase of $MEANAD$ for C-vine sampling, NORTA sampling even faintly improves with growing ρ. For all approaches, the $AR(1)$ settings perform slightly worse than the exchangeable ones. For C-vine sampling overdispersed settings perform worse than equidispersed ones. Zero-inflation also increases overdispersion and hence worsens the results. For zero-inflated margins NORTA basically fails to find the appropriate input parameters. Even the naive approach performs better. For other settings the naive approach performs worse than NORTA in the settings where ρ is larger than 0.1. In general, we confirm the insights of Biller and Ghosh (2006) who stress that the efficiency of NORTA depends on the target correlation and the distributional class.

2.7 Summary and Discussion

This chapter develops an applied approach in the difficult field input modeling for high-dimensional correlated count data. It is an approximating sampling method since we a) use a partial correlation property for conditional correlations and b) carry out a simplified intermediate transformation in order to determine partial correlation parameters. Nevertheless, since this simplified approach is only used to determine partial correlation parameters, the margins have the desired distributions, i.e., an error only occurs as far as the correlation of the outcomes is concerned. Despite these simplifications we show via simulation that our approach generates count variables with precise correlation. We compared our approach to NORTA, the most famous competitor for the problem at hand. A shortcoming of both our and the NORTA approach is that not all correlations can be sampled. As Cario and Nelson (1997) pointed out (Proposition 1) the upper

		exchangeable				AR(1)			
ρ		Poisson	GP	ZIGP	NB	Poisson	GP	ZIGP	NB
0.1	2	**0.0044**	**0.0043**	**0.0044**	**0.0042**	**0.0044**	**0.0042**	**0.0043**	**0.0043**
		0.0253	***0.0267***	***0.0350***	***0.0247***	***0.0248***	***0.0274***	***0.0349***	***0.0255***
		0.0252	*0.0247*	*0.0256*	*0.0250*	*0.0263*	*0.0265*	*0.0253*	*0.0263*
	5	**0.0075**	**0.0073**	**0.0074**	**0.0071**	**0.0072**	**0.0069**	**0.0072**	**0.0069**
		0.0256	***0.0256***	***0.0333***	***0.0260***	***0.0252***	***0.0258***	***0.0290***	***0.0257***
		0.0250	*0.0271*	*0.0275*	*0.0262*	*0.0252*	*0.0259*	*0.0263*	*0.0262*
	10	**0.0103**	**0.0101**	**0.0105**	**0.0101**	**0.0098**	**0.0097**	**0.0101**	**0.0098**
		0.0254	***0.0259***	***0.0319***	***0.0259***	***0.0254***	***0.0255***	***0.0265***	***0.0255***
		0.0251	*0.0272*	*0.0276*	*0.0267*	*0.0253*	*0.0257*	*0.0257*	*0.0256*
0.4	2	**0.0047**	**0.0045**	**0.0045**	**0.0046**	**0.0046**	**0.0046**	**0.0045**	**0.0047**
		0.0222	***0.0220***	***0.1173***	***0.0219***	***0.0210***	***0.0215***	***0.1148***	***0.0216***
		0.0209	*0.0230*	*0.0298*	*0.0228*	*0.0217*	*0.0228*	*0.0299*	*0.0232*
	5	**0.0096**	**0.0109**	**0.0114**	**0.0104**	**0.0104**	**0.0101**	**0.0105**	**0.0100**
		0.0215	***0.0231***	***0.1042***	***0.0229***	***0.0238***	***0.0250***	***0.0654***	***0.0245***
		0.0217	*0.0348*	*0.0452*	*0.0319*	*0.0236*	*0.0313*	*0.0365*	*0.0291*
	10	**0.0124**	**0.0139**	**0.0145**	**0.0133**	**0.0148**	**0.0148**	**0.0155**	**0.0150**
		0.0213	***0.0235***	***0.0887***	***0.0228***	***0.0246***	***0.0253***	***0.0411***	***0.0250***
		0.0216	*0.0364*	*0.0440*	*0.0331*	*0.0243*	*0.0288*	*0.0306*	*0.0280*
0.7	2	**0.0045**	**0.0046**	**0.0046**	**0.0046**	**0.0045**	**0.0045**	**0.0047**	**0.0046**
		0.0128	***0.0139***	***0.2044***	***0.0139***	***0.0133***	***0.0141***	***0.2042***	***0.0132***
		0.0142	*0.0164*	*0.0373*	*0.0154*	*0.0147*	*0.0165*	*0.0373*	*0.0162*
	5	**0.0108**	**0.0150**	**0.0180**	**0.0135**	**0.0150**	**0.0166**	**0.0166**	**0.0155**
		0.0130	***0.0143***	***0.1362***	***0.0139***	***0.0185***	***0.0194***	***0.0958***	***0.0193***
		0.0145	*0.0371*	*0.0630*	*0.0313*	*0.0185*	*0.0358*	*0.0547*	*0.0319*
	10	**0.0127**	**0.0177**	**0.0194**	**0.0150**	**0.0194**	**0.0197**	**0.0209**	**0.0194**
		0.0131	***0.0148***	***0.1243***	***0.0143***	***0.0210***	***0.0226***	***0.0647***	***0.0224***
		0.0143	*0.0378*	*0.0567*	*0.0332*	*0.0215*	*0.0351*	*0.0431*	*0.0324*

Table 2.5: Mean absolute deviation ($MEANAD$) based on $R = 1000$ replications of $N = 1000$ samples of size T for exchangeable target correlation ρ and different count margins and parameters as in Table 2.4 (bold: C-vine sampling, bold italics: NORTA sampling, italics: naive sampling)

and lower boundaries of feasible absolute correlations may be smaller than 1. An advantage of C-vine sampling over NORTA is that the resulting input correlation is already positive definite (Joe (2006)). Throughout all settings we investigated, our sampling approach had a lower absolute deviation than NORTA.

Chapter 3

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

3.1 Introduction

This chapter considers longitudinal setups for generalized Poisson (GP) data using GEE. The GP distribution has first been introduced by (Consul and Jain 1970). GP regression models (GPR) were discussed by (Consul and Famoye 1992) and (Famoye 1993). (Famoye, J.T., and K.P. 2004) apply generalized Poisson regression to accident data, whereas (Famoye and 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 and Gupta 1984). A Bayesian analysis is carried out by (Scollnik 1995) and (Gschlößl and Czado 2008). The interest in the class of GPR models is driven by the fact that it can handle under- and overdispersion, which count data very often exhibits. We allow for regression effects not only on the mean but on the dispersion parameter as well. In this case we will illustrate that we need to focus on the case of overdispersion, which most data exhibits. This regression specification allows to model overdispersion by individual characteristics (e.g. by a company's industry) and to improve model fit when constant dispersion is insufficient. The fact that the GP distribution is a hyper model of the Poisson distribution allows for nested model comparison if the mean specifications are hierarchical. The variance function of the GP distribution can be written as a product of the mean and an independent dispersion parameter, which eases regression specifications of the dispersion parameter. In contrast to the negative binomial distribution exploratory tools for suitable choice and transformation of describing variables for the dispersion parameter have already been developed by (Czado, Erhardt, Min, and Wagner 2007).

GEE have been introduced by (Liang and Zeger 1986) which are also known by GEE1. Second level GEE (GEE2, (Prentice and Zhao 1991)) allow to determine variance parameters as well. (Yan and Fine 2004) consider generalized estimating equations for the Poisson distribution.

An implementation can be found in the *R* package 'geepack' (see (Yan 2002)). (Hall and Severini 1998) develop extended generalized estimating equations which are based on a Taylor series expansion of an extended quasi-likelihood function. According to (Hilbe 2007, p. 119) they are less accurate for the estimation of the dispersion than GEE2, which which our approach is based on. For the conditional fixed-effects negative binomial distribution, generalized estimating equations are implemented in Stata ((StataCorp 2007)). (Hilbe 2007, Section 10.4) emphazises that in this setup the dispersion parameter is not estimated as a separate parameter, it is apportioned across panels. So far, no approach for fitting the dispersion parameters of the negative binomial distribution, even without regression effects, has been developed.

A comparison of three models starting with the regular Poisson GEE extended by dispersion designs will be carried out in this chapter. 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, Erhardt, Min, and Wagner 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 chapter. 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 chapter. The implementation is carried out in *R*.

The chapter 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 chapter is organized as follows: Section 3.2 introduces our GPR regression setup. In Section 3.3, we show how the GEE approach by (Liang and Zeger 1986) and the extensions by (Prentice and Zhao 1991) can be applied to estimate parameters in our setup. A simulation study investigating small sample properties will be given in Section 3.4 showing a satisfactory behavior for medium sample sizes. Subsection 3.5.1 reviews the variable selection criterion for panel data by (Pan 2001) while in Subsection 3.5.2 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 3.6 applies our findings to patent outsourcing data and interprets the results of our 'best' model. We conclude with a summary and discussion section.

3.2 A GEE setup for longitudinal count data

Let $Y_{it} \sim GP(\mu_{it}, \varphi_{it})$. For the GP distribution we refer to the Introduction chapter. 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:

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

Here $\boldsymbol{Y}_i := \boldsymbol{Y}_{i\sim} = (Y_{i1}, \dots, Y_{iT})'$ summarizes the T dimensional vector of dependent variables for subject i. Observations from different subjects are assumed to be independent. Similarly $\boldsymbol{Y}_{\sim t} := (Y_{1t}, \dots, Y_{Kt})' \in \mathbb{N}_0^K$ 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

$$\text{corr}(\boldsymbol{Y}) = \begin{pmatrix} \boldsymbol{R}_1(\lambda_1) & \boldsymbol{0}_{T\times T} & \dots & \boldsymbol{0}_{T\times T} \\ \boldsymbol{0}_{T\times T} & \boldsymbol{R}_1(\lambda_1) & & \boldsymbol{0}_{T\times T} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{0}_{T\times T} & \boldsymbol{0}_{T\times T} & \dots & \boldsymbol{R}_1(\lambda_1) \end{pmatrix} \in \mathbb{R}^{KT\times KT}. \tag{3.2.1}$$

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

$$\text{var}(\boldsymbol{Y}_{\sim t}) = \begin{pmatrix} \mu_{1t}\varphi_{1t}^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mu_{Kt}\varphi_{Kt}^2 \end{pmatrix}. \tag{3.2.2}$$

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 $\varphi_{it} := \sqrt{\frac{\text{var}(Y_{it})}{E(Y_{it})}}$ which might fall below 1 for some observations. According to the definition of the underdispersed GP distribution, 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 and 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, Erhardt, Min, and Wagner 2007, Section 5), which will be illustrated in Section 3.6.1. 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 $t = 1, \ldots, T$.

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:

$$
\begin{aligned}
&\text{(i) } \textit{Mean level} \\
E(Y_{it} \mid \boldsymbol{\beta}) = \mu_{it}(\boldsymbol{\beta}) &:= E_{it} e^{\boldsymbol{x}'_{it}\boldsymbol{\beta}} = e^{\boldsymbol{x}'_{it}\boldsymbol{\beta} + log(E_{it})} > 0 \\
\Leftrightarrow \eta^{\mu}_{it}(\boldsymbol{\beta}) &= \log(\mu_{it}(\boldsymbol{\beta})) - \log(E_{it}) \text{ (log link)}, \qquad (3.2.3) \\
&\text{(ii) } \textit{Overdispersion level} \\
\varphi_{it}(\boldsymbol{\alpha}) &:= 1 + e^{\boldsymbol{w}'_{it}\boldsymbol{\alpha}} > 1 \\
\Leftrightarrow \eta^{\varphi}_{it}(\boldsymbol{\alpha}) &= \log(\varphi_{it}(\boldsymbol{\alpha}) - 1)) \text{ (shifted log link)}. \qquad (3.2.4)
\end{aligned}
$$

This setup for longitudinal count regression data $\{Y_{it}, i = 1, \ldots, K; t = 1, \ldots, T\}$ we denote by $GPR(\mu_{it}, \varphi_{it}, \boldsymbol{R}_1(\lambda_1))$. 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:

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

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 $H_1 :\ \lambda_1 \neq 0$.

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

Generalized estimating equations have first been introduced by (Liang and 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 and 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 $\text{var}(Y_{it}) = \sigma^2_{it}(\boldsymbol{\delta})$, a working covariance matrix for $\boldsymbol{Y}_i$ can be constructed by

$$\boldsymbol{V}_{i1}(\boldsymbol{\delta}) := \boldsymbol{A}_i^{1/2}(\boldsymbol{\delta})\boldsymbol{R}_1(\lambda_1(\gamma))\boldsymbol{A}_i^{1/2}(\boldsymbol{\delta}) \ \in \mathbb{R}^{T\times T}, \tag{3.3.1}$$

where $\boldsymbol{A}_i(\boldsymbol{\delta}) := \text{diag}\{\sigma^2_{i1}(\boldsymbol{\delta}), \ldots, \sigma^2_{iT}(\boldsymbol{\delta})\}$. Covariances will be denoted by $\sigma^2_{itt^*}(\boldsymbol{\delta}) := \text{cov}(Y_{it}, Y_{it^*})$ and $\boldsymbol{\sigma}^2_i(\boldsymbol{\delta}) := \big(\sigma^2_{itt^*}(\boldsymbol{\delta}); t \le t^*; t, t^* = 1, \ldots, T\big)' \in \mathbb{R}^{T(T+1)/2}$ will be the vector of covariances of subject i. Further, let $\boldsymbol{S}_i(\boldsymbol{\beta}) = (S_{itt^*}(\boldsymbol{\beta}); t \le t^*; t, t^* = 1, \ldots, T\big)' \in \mathbb{R}^{T(T+1)/2}$ 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 $\tau^2_{itt^*}(\boldsymbol{\delta}) := \text{var}(S_{itt^*}(\boldsymbol{\beta}) \mid \boldsymbol{\delta})$, we can again construct a working covariance

$$\begin{aligned}\boldsymbol{V}_{i2}(\boldsymbol{\delta}, \lambda_2) := \text{cov}(\boldsymbol{S}_i(\boldsymbol{\beta}) \mid \boldsymbol{\delta}, \lambda_2) \ &= \ \text{diag}\big(\tau_{i11}(\boldsymbol{\delta}), \tau_{i12}(\boldsymbol{\delta}), \ldots, \tau_{iTT}(\boldsymbol{\delta})\big)\boldsymbol{R}_2(\lambda_2) \\ &\quad \times \text{diag}\big(\tau_{i11}(\boldsymbol{\delta}), \tau_{i12}(\boldsymbol{\delta}), \ldots, \tau_{iTT}(\boldsymbol{\delta})\big).\end{aligned} \tag{3.3.2}$$

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

$$K^{-1/2}\sum_{i=1}^{K} \boldsymbol{D}'_{i1}(\boldsymbol{\beta})\boldsymbol{V}^{-1}_{i1}(\boldsymbol{\delta})(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \boldsymbol{0}_{p+1}, \tag{3.3.3}$$

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

$$K^{-1/2}\sum_{i=1}^{K} \boldsymbol{D}'_{i2}(\boldsymbol{\delta})\boldsymbol{V}^{-1}_{i2}(\boldsymbol{\delta}, \lambda_2)\big(\boldsymbol{S}_i(\boldsymbol{\beta}) - \boldsymbol{\sigma}^2_i(\boldsymbol{\delta})\big) = \boldsymbol{0}_{q+2}, \tag{3.3.4}$$

where $\boldsymbol{D}_{i2}(\boldsymbol{\delta}) := \frac{\partial \boldsymbol{\sigma}^2_i(\boldsymbol{\delta})}{\partial(\boldsymbol{\alpha}', \gamma)'} \in \mathbb{R}^{[T(T+1)/2]\times[q+2]}$. Additionally, $\boldsymbol{D}_{i12}(\boldsymbol{\delta}) := \frac{\partial \boldsymbol{\sigma}^2_i(\boldsymbol{\delta})}{\partial \boldsymbol{\beta}} \in \mathbb{R}^{[T(T+1)/2]\times[p+1]}$ will be calculated since we hope to gain information on the mean parameters $\boldsymbol{\beta}$ also from $\boldsymbol{\sigma}^2_i(\boldsymbol{\delta})$. 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 $\rho_{tt^*}\big(\lambda_1(\gamma)\big) = \text{corr}(Y_{it}, Y_{it^*})$, $i = 1, \ldots, K$. Their derivatives $\boldsymbol{D}_{i1}(\boldsymbol{\beta})$, $\boldsymbol{D}_{i2}(\boldsymbol{\delta})$ and $\boldsymbol{D}_{i12}(\boldsymbol{\delta})$ are given in the Appendix of this chapter. To solve (3.3.3) and (3.3.4) simultaneously, let

$$\boldsymbol{D}_i(\boldsymbol{\delta}) := \begin{pmatrix} \frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} & \frac{\partial \boldsymbol{\sigma}^2_i(\boldsymbol{\delta})}{\partial \boldsymbol{\beta}} \\ \boldsymbol{0} & \frac{\partial \boldsymbol{\sigma}^2_i(\boldsymbol{\delta})}{\partial(\boldsymbol{\alpha}', \gamma)'} \end{pmatrix} = \begin{pmatrix} \boldsymbol{D}_{i1}(\boldsymbol{\delta}) & \boldsymbol{D}_{i12}(\boldsymbol{\delta}) \\ \boldsymbol{0} & \boldsymbol{D}_{i2}(\boldsymbol{\delta}) \end{pmatrix} \in \mathbb{R}^{[T(T+3)/2]\times[p+q+3]}, \tag{3.3.5}$$

$$\boldsymbol{V}_i(\boldsymbol{\delta}, \lambda_2) := \begin{pmatrix} \boldsymbol{V}_{i1}(\boldsymbol{\delta}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_{i2}(\boldsymbol{\delta}, \lambda_2) \end{pmatrix} \in \mathbb{R}^{[T(T+3)/2]\times[T(T+3)/2]}, \tag{3.3.6}$$

$$\boldsymbol{f}_i(\boldsymbol{\delta}) := \begin{pmatrix} \boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}) \\ \boldsymbol{s}_i(\boldsymbol{\beta}) - \boldsymbol{\sigma}_i^2(\boldsymbol{\delta}) \end{pmatrix} \in \mathbb{R}^{T(T+3)/2} \tag{3.3.7}$$

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 $K^{-1/2} \sum_{i=1}^K \boldsymbol{D}_i'(\boldsymbol{\delta}) \boldsymbol{V}_i^{-1}(\boldsymbol{\delta}, \lambda_2) \boldsymbol{f}_i(\boldsymbol{\delta}) = \boldsymbol{0}_{p+q+3}$. While $\boldsymbol{\delta}$ will be updated by a Fisher-Scoring step, residuals $\hat{r}_{ilm}(\hat{\boldsymbol{\delta}}) := s_{ilm}(\hat{\boldsymbol{\beta}}) - \sigma_{ilm}^2(\hat{\boldsymbol{\delta}})$ may be used to estimate λ_2. For example, for an exchangeable matrix $\boldsymbol{R}_2(\lambda_2)$, define $I^* := \{(lm, l^*m^*) : l \leq l^* \wedge m \leq m^*\}$. Then according to (Liang and Zeger 1986, p. 18, Example 3), an estimate of λ_2 is given by

$$\hat{\lambda}_2(\hat{\boldsymbol{\delta}}) = \frac{\sum_{i=1}^K \sum_{(lm,l^*m^*) \in I^*} \hat{r}_{ilm}(\hat{\boldsymbol{\delta}}) \hat{r}_{il^*m^*}(\hat{\boldsymbol{\delta}})}{K \left[\frac{T(T+1)}{2} \left(\frac{T(T+1)}{2} - 1 \right) / 2 \right] - (p+q+3)}. \tag{3.3.8}$$

According to (Prentice and 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 $\boldsymbol{0}_{p+q+3}$ and covariance

$$\begin{aligned} \operatorname{cov}(\boldsymbol{Z}) \;=\; & K^{-1} \boldsymbol{\Gamma}^{-1}(\boldsymbol{\delta}, \lambda_2) \\ & \cdot \left(\sum_{i=1}^K \boldsymbol{D}_i'(\boldsymbol{\delta}) \boldsymbol{V}_i^{-1}(\boldsymbol{\delta}, \lambda_2) \operatorname{cov}\big((\boldsymbol{Y}_i', \boldsymbol{S}_i'(\boldsymbol{\beta}))' \big) \boldsymbol{V}_i^{-1}(\boldsymbol{\delta}, \lambda_2) \boldsymbol{D}_i(\boldsymbol{\delta}) \right) \\ & \cdot \boldsymbol{\Gamma}^{-1}(\boldsymbol{\delta}, \lambda_2) \end{aligned} \tag{3.3.9}$$

with $\operatorname{cov}\big((\boldsymbol{Y}_i', \boldsymbol{S}_i'(\boldsymbol{\beta}))' \big)$ unknown. A consistent 'sandwich' estimator of (3.3.9) is

$$\begin{aligned} \boldsymbol{\Omega}_{sw}(\hat{\boldsymbol{\delta}}, \hat{\lambda}_2) \;:=\; & K^{-1} \boldsymbol{\Gamma}^{-1}(\hat{\boldsymbol{\delta}}, \hat{\lambda}_2) \\ & \cdot \left(\sum_{i=1}^K \boldsymbol{D}_i'(\hat{\boldsymbol{\delta}}) \boldsymbol{V}_i^{-1}(\hat{\boldsymbol{\delta}}, \hat{\lambda}_2) \boldsymbol{f}_i(\hat{\boldsymbol{\delta}}) \boldsymbol{f}_i'(\hat{\boldsymbol{\delta}}) \boldsymbol{V}_i^{-1}(\hat{\boldsymbol{\delta}}, \hat{\lambda}_2) \boldsymbol{D}_i(\hat{\boldsymbol{\delta}}) \right) \\ & \cdot \boldsymbol{\Gamma}^{-1}(\hat{\boldsymbol{\delta}}, \hat{\lambda}_2) \end{aligned} \tag{3.3.10}$$

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

$$\boldsymbol{\Omega}_{mb}(\hat{\boldsymbol{\delta}}, \hat{\lambda}_2) := K^{-1} \boldsymbol{\Gamma}^{-1}(\hat{\boldsymbol{\delta}}, \hat{\lambda}_2). \tag{3.3.11}$$

An issue still open is how to determine $\tau_{itt^*}^2(\boldsymbol{\delta}) := \operatorname{var}\big((Y_{it} - \mu_{it}(\boldsymbol{\beta}))(Y_{it^*} - \mu_{it^*}(\boldsymbol{\beta})) \big)$ in (3.3.2). Here $\tau_{itt^*}^2(\boldsymbol{\delta})$ is a function of higher mixed moments $E\left[Y_{it}^l Y_{it^*}^{l^*} \right]$, $l, l^* = 1, 2$ for which a closed form is unknown. They can be determined only 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 $(Z_1, Z_2) \sim N_2(\boldsymbol{0}, \Sigma(\rho))$ where $\Sigma(\rho)$ is diagonal with elements $g(\rho)$. This $g(\rho)$ is unknown and is approximately determined using the approach suggested by (Erhardt and Czado 2009a).

To approximate $\tau^2(\boldsymbol{\theta}) := \operatorname{var}\left((Y_1 - \mu_1)(Y_2 - \mu_2) \right)$ 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 $GPR(\mu_i, \varphi_i)$ regression model according to (Czado, Erhardt, Min, and Wagner 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

$$\begin{aligned}
\hat{E}(\log(\hat{\tau}_j^2)) \;=\; & -0.454 \cdot +1.027 \cdot \log(\mu_{1j}) + 2.615 \cdot \log(\varphi_{1j}) + 0.974 \cdot \log(\mu_{2j}) \\
& +2.650 \cdot \log(\varphi_{2j}) + 1.186 \cdot \rho_j - 0.135 \cdot \log(\mu_{1j}) \cdot \log(\varphi_{1j}) \\
& +0.010 \cdot \log(\mu_{1j}) \cdot \log(\mu_{2j}) - 0.028 \cdot \log(\mu_{1j}) \cdot \log(\varphi_{2j}) - 0.110 \cdot \log(\mu_{1j}) \cdot \rho_j \\
& +0.086 \cdot \log(\varphi_{1j}) \cdot \log(\varphi_{2j}) + 0.913 \cdot \log(\varphi_{1j}) \cdot \rho_j - 0.116 \cdot \log(\mu_{2j}) \cdot \log(\varphi_{2j}) \\
& +0.804 \cdot \log(\varphi_{2j}) \cdot \rho_j - 0.058 \cdot \log(\mu_{1j}) \cdot \log(\varphi_{1j}) \cdot \rho_j \\
& -0.056 \cdot \log(\varphi_{1j}) \cdot \log(\mu_{2j}) \cdot \rho_j - 0.087 \cdot \log(\mu_{2j}) \cdot \log(\varphi_{2j}) \cdot \rho_j
\end{aligned} \tag{3.3.12}$$

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 $\tau^2_{itt^*}(\boldsymbol{\delta})$ by the analytical expression $\exp(\hat{E}(\log(\hat{\tau}^2)|\boldsymbol{\theta}^*_{itt^*}))$, where $\boldsymbol{\theta}^*_{itt^*} := \big(\mu_{it}(\boldsymbol{\beta}), \mu_{it^*}(\boldsymbol{\beta}), \varphi_{it}(\boldsymbol{\alpha}), \varphi_{it^*}(\boldsymbol{\alpha}), \rho_{itt^*}(\lambda_1(\gamma))\big)$.

3.4 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 and Czado 2009a).

'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

$$\log(\mu_{it}) = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot t/T \tag{3.4.1}$$

$$\log(\varphi_{it} - 1) = \alpha_0 + \alpha_1 \cdot w_i. \tag{3.4.2}$$

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 3.1 were used to assess the asymptotic normality of the estimates. The parameters on the mean level have

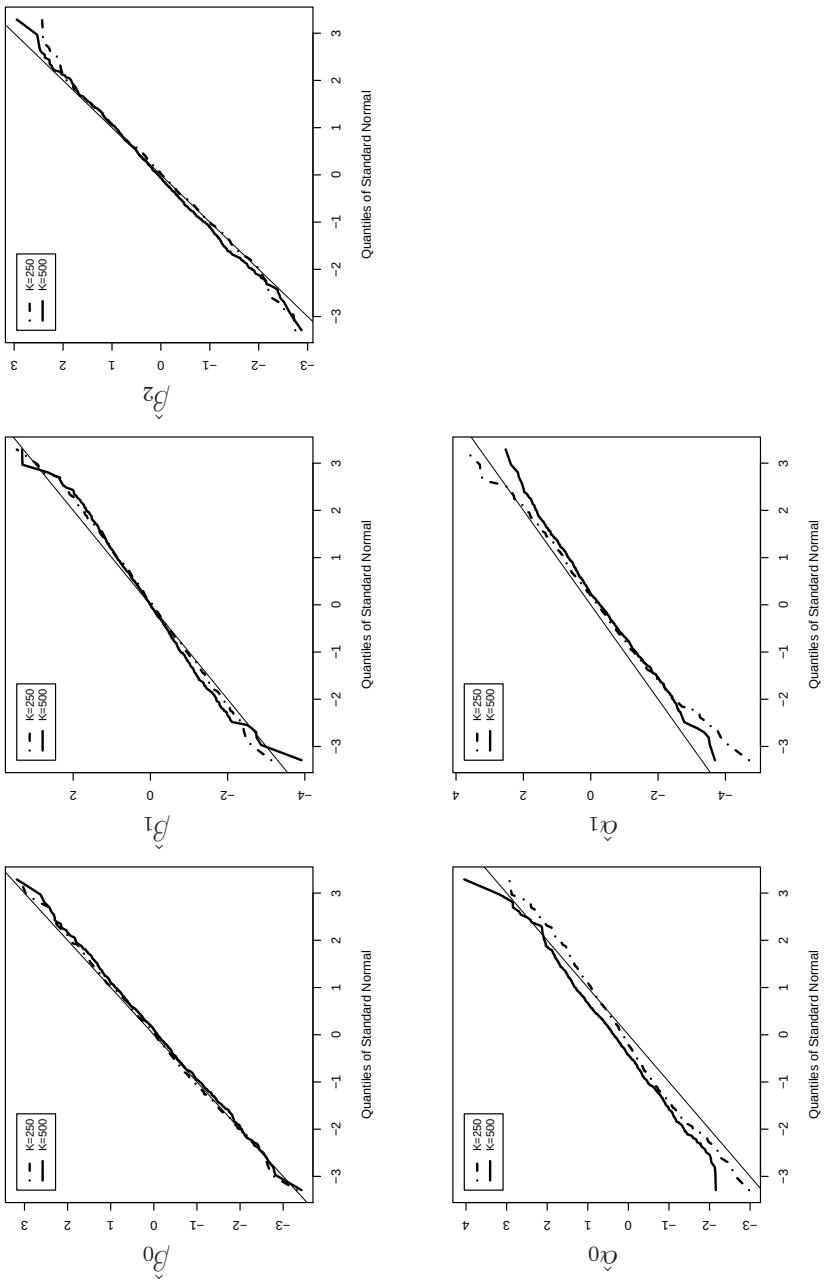

Figure 3.1: QQ plots of centered and standardized estimates based on $N = 1000$ replications ($K = 250, 500$, $T = 8$, $\lambda = 0.5$, $\boldsymbol{\beta} = (1.32, 0.70, 0.70)'$, $\boldsymbol{\alpha} = (0.25, 0.31, 0.31)'$)

approximately a normal distribution already for $K = 250$, while this is only approximately true for $K = 500$ for the parameters on the dispersion level.

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 3.1). 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

$$\hat{b}(\theta, \hat{\boldsymbol{\theta}}) = \frac{1}{N\theta} \sum_{i=1}^{N} \hat{\theta}_i - 1. \tag{3.4.3}$$

The estimated variance of the estimated relative bias is given by $\widehat{\text{var}}\big(\hat{b}(\theta, \hat{\theta}_i)\big) := 1/\theta^2 \widehat{\text{var}}(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}} := (\hat{\theta}_1, \ldots, \hat{\theta}_N)'$, and $\widehat{\text{var}}(\hat{\boldsymbol{\theta}}) := \frac{1}{N-1} \sum_{i=1}^{N} \left(\hat{\theta}_i - \frac{1}{N} \sum_{k=1}^{N} \hat{\theta}_k\right)^2$. The mean squared error (MSE) is given by

$$R(\theta, \hat{\theta}) := E([\hat{\theta} - \theta]^2) = \text{var}(\hat{\theta}) + b^2(\hat{\theta}, \theta). \tag{3.4.4}$$

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 (3.3.10). This shows that the accuracy of the estimations is satisfactory for medium sample sizes. The absolute values of the relative bias in Table 3.1 are smaller for the mean effects than for the dispersion effects, hence the mean coefficients are estimated better than the dispersion coefficients. 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.

3.5 Variable selection and model comparison

3.5.1 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 and Nelder 1989, p. 325)). In the GP context, we have $E(Y_{it}) = \mu_{it}(\boldsymbol{\beta})$ and $\text{var}(Y_{it}) = \varphi^2_{it}(\boldsymbol{\alpha})\mu_{it}(\boldsymbol{\beta})$, i.e., $V(\mu_{it}(\boldsymbol{\beta})) = \mu_{it}(\boldsymbol{\beta})$ and $\phi = \varphi^2_{it}(\boldsymbol{\alpha})$, and obtain

	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 3.1: Average coefficients, relative bias (see 3.4.3) and mean squared error (see 3.4.4) together with estimated 'sandwich' standard deviations in round brackets according to (3.3.10) for $N = 1000$ fitted samples

$$\begin{aligned} QL(\mu_{it}(\boldsymbol{\beta}), \varphi_{it}(\boldsymbol{\alpha}), y_{it}) &= \int_{y_{it}}^{\mu_{it}(\boldsymbol{\beta})} \frac{y_{it} - t}{\varphi_{it}^2(\boldsymbol{\alpha}) t} dt \\ &= \frac{1}{\varphi_{it}^2(\boldsymbol{\alpha})} \left(y_{it} \log(\mu_{it}(\boldsymbol{\beta})) - \mu_{it}(\boldsymbol{\beta}) \right) \\ &\quad + \text{constants ind. of } (\boldsymbol{\beta}, \boldsymbol{\alpha}). \end{aligned} \tag{3.5.1}$$

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

$$\begin{aligned} Q(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{y}) &:= \sum_{i=1}^{K} \sum_{t=1}^{T} QL(\mu_{it}(\boldsymbol{\beta}), \varphi_{it}(\boldsymbol{\alpha}), y_{it}) \\ &= \sum_{i=1}^{K} \sum_{t=1}^{T} \frac{1}{\varphi_{it}^2(\boldsymbol{\alpha})} \left(y_{it} \log(\mu_{it}(\boldsymbol{\beta})) - \mu_{it}(\boldsymbol{\beta}) \right) \\ &\quad + \text{constants ind. of } (\boldsymbol{\beta}, \boldsymbol{\alpha}). \end{aligned} \tag{3.5.2}$$

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 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

$$QIC(\hat{\boldsymbol{\delta}}, \hat{\lambda}_2) := -2Q(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \boldsymbol{I}, \boldsymbol{y}) + 2 \text{ trace}(\boldsymbol{\Omega}_{mb}^{-1}(\hat{\boldsymbol{\delta}}, \hat{\lambda}_2)\boldsymbol{\Omega}_{sw}(\hat{\boldsymbol{\delta}}, \hat{\lambda}_2)). \tag{3.5.3}$$

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

3.5.2 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 $GPR(\mu_{it}, \varphi_{it}, \boldsymbol{R})$.

Similar we denote by $PoiR(\mu_{it}, \boldsymbol{R})$ 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 $GPR(\mu_{it}, \varphi, \boldsymbol{R})$. The corresponding model hierarchy is given in Figure 3.2. A covariate being significant in terms of the Wald test (e.g. in the mean design of $PoiR(\mu_{it}, \boldsymbol{R})$) can be insignificant in a different model (say $GPR(\mu_{it}, \varphi, \boldsymbol{R})$). 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.

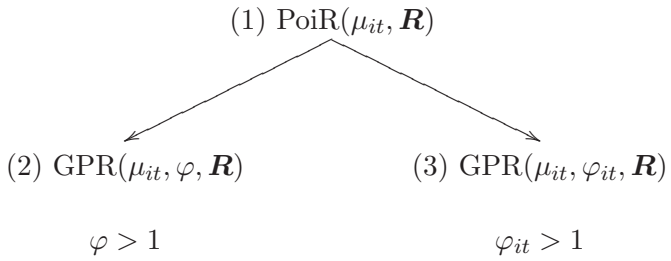

Figure 3.2: Investigated setup hierarchy

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 $E(T) = \frac{2n_pn_n}{n_p+n_n} + 1$ and $\text{var}(T) = \frac{2n_pn_n(2n_pn_n-n_p-n_n)}{(n_p+n_n)^2(n_p+n_n-1)}$. Then $W_Z := \frac{T-E(T)}{\sqrt{\text{var}(T)}}$ is approximately standard normal. A α level test can be constructed as

$$\text{Reject } H_0 \text{ if } |W_Z| > q_{1-\alpha/2} \tag{3.5.4}$$

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.

3.6 Application: Outsourcing of patent applications

3.6.1 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, Erhardt, Min, and Wagner 2007) for this data. A more detailed description of this model will be given

in Section 3.6.2. 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 3.2.

(Czado, Erhardt, Min, and Wagner 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, Erhardt, Min, and Wagner 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 3.3. We see that for smaller **EMP** the dispersion is lower whereas for higher values it is high.

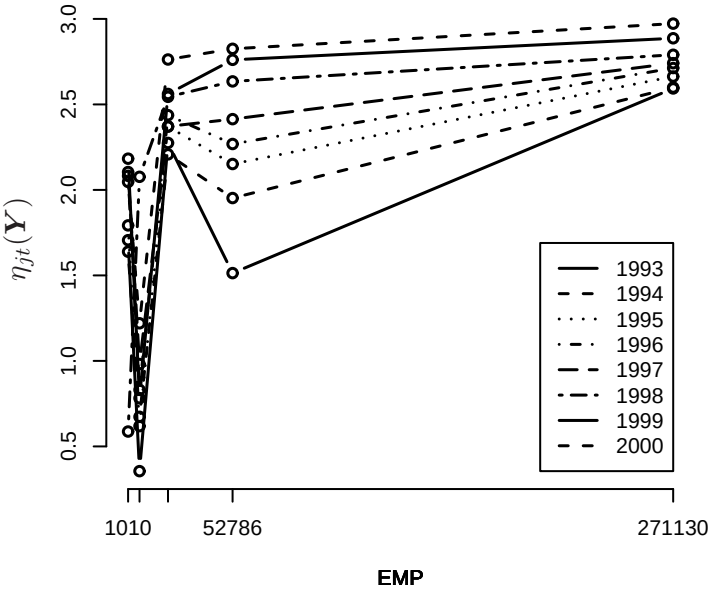

Figure 3.3: Influence of **EMP** on the overdispersion parameter

As a working correlation matrix for $\text{corr}(\boldsymbol{Y}_i(\boldsymbol{\beta}))$ 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, Erhardt, Min, and Wagner 2007) strongly suggests that it has this structure. For $\text{corr}(\boldsymbol{S}_i(\boldsymbol{\delta}))$ we choose the identity matrix $\boldsymbol{I}_{T(T+1)/2}$. 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 the covariates **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 (3.5.3), 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)	W_Z (p)
			$p+q+1$	$p+q+1$
(1) $PoiR(\mu_{it}, \boldsymbol{R})$	-215813.10	-216270.49	-1.34 (0.18)	-2.02 (0.04)
			17	12
(2) $GPR(\mu_{it}, \varphi, \boldsymbol{R})$	-2409.02	-2546.67	-0.56 (0.58)	-2.53 (0.01)
			18	12
(3) $GPR(\mu_{it}, \varphi_{it}, \boldsymbol{R})$	-3945.78	-4581.72	-0.46 (0.64)	-0.48 (0.63)
			25	14

Table 3.2: QIC (see (3.5.3)) and results of the Wald-Wolfowitz test (see (3.5.4)) of full and reduced designs for the three model classes specified in Figure 3.2

Note that it make only sense to compare the QIC for nested settings. Poisson and GPR models are not nested since for a dispersion of 1 in a GPR 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) $GPR(\mu_{it}, \varphi, \boldsymbol{R})$ and (3) $GPR(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ are nested. The QIC statistics according to (3.5.3) and the result of the Wald-Wolfowitz test according to (3.5.4) can be found in Table 3.2. 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 3.3.

Model	Mean	Dispersion	$p+q+1$
$PoiR(\mu_{it}, \boldsymbol{R})$ *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
$PoiR(\mu_{it}, \boldsymbol{R})$ *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
$GPR(\mu_{it}, \varphi, \boldsymbol{R})$ *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
$GPR(\mu_{it}, \varphi, \boldsymbol{R})$ *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
$GPR(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ *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
$GPR(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ *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 3.3: Model equations of the models shown in Figure 3.2 using backward selection by QIC (3.5.3)

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) $PoiR(\mu_{it}, \boldsymbol{R})$ to (2) $GPR(\mu_{it}, \varphi, \boldsymbol{R})$, 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) $GPR(\mu_{it}, \varphi, \boldsymbol{R})$ to (3) $GPR(\mu_{it}, \varphi_{it}, \boldsymbol{R})$, 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) $GPR(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ is to be preferred over all other classes discussed (see Table 3.2). However, this goodness-of-fit criterion does not account for the model complexity. Since the reduced design for (3) $GPR(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ 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.

3.6.2 Model interpretation

The paper by (Czado, Erhardt, Min, and Wagner 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 experienced that the inclusion of zero-inflation in the model adds too much flexibility to the dispersion specification and that the Newton Raphson updates often did not converge.

In the ZIGPR model of (Czado, Erhardt, Min, and Wagner 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 ZIGPR model in (Czado, Erhardt, Min, and Wagner 2007).

We will now briefly interpret the reduced setup (3) $GPR(\mu_{it}, \varphi_{it}, \boldsymbol{R})$. Note that some of the covariates in the $GPR(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ setup in Table 3.4 are insignificant according to the Wald test.

In contrast to the ZIGPR model mentioned, variable selection has been done using backward selection with respect to QIC. In the ZIGPR 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 GPR model, **RDmiss** still appears, it is, however, insignificant according to Wald: the p-value is 93.9% (Table 3.4). Obviously, the lack of modeling zero-inflation in the GPR model is reflected in the higher overdispersion range of $[2.20, 11.04]$ as compared to $[2.41, 10.15]$ in the ZIGPR model. Also, there is an additional interaction between the Chemical / Pharmaceutical industry dummy and the square root of the number of employees

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

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 ZIGPR 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, Erhardt, Min, and Wagner 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 ZIGPR fit, we will look at the influence of **EMP** on the outsourcing rate in Figure 3.4 (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, Erhardt, Min, and Wagner 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 3.4 (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.

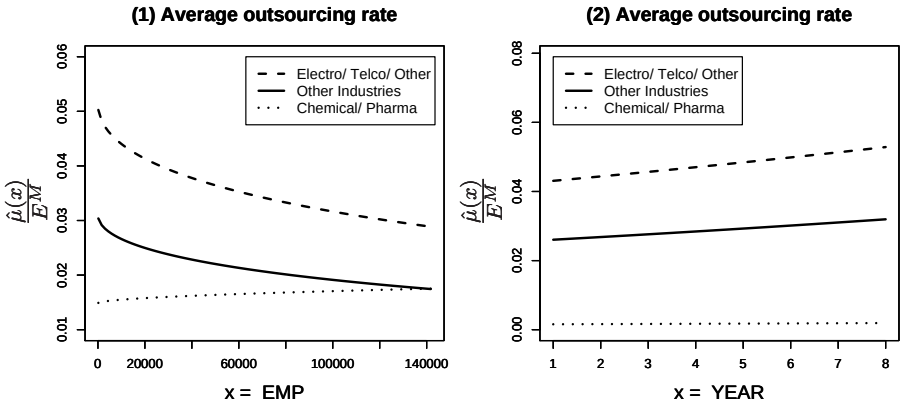

Figure 3.4: Influence of **EMP** and **YEAR** on the outsourcing rate while fixing other covariates by their empirical modes

We define the overdispersion factor of a random variable $Y_{it} \sim GPR(\mu_{it},\, \varphi_{it})$ as $V_{it} := \frac{\mathrm{var}(Y_{it})}{E(Y_{it})} = \varphi_{it}^2$. There are only categorical covariates for overdispersion: $\boldsymbol{w}$:=(**1**, **CAR.SUPP.OTHER**, **EMP.11291**). Using (3.2.4), 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 3.5 lists $\hat{V}(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w})$ depending on the settings arising from the categorical dispersion designs. As in (Czado, Erhardt, Min, and Wagner 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 ZIGPR 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 3.5: Estimated overdispersion factor in the 'best' model (Table 3.4) 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, Erhardt, Min, and Wagner 2007).

3.7 Conclusions and Discussions

We introduced a $GPR(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ 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 $GPR(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ 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 $GPR(\mu_{it}, \varphi_{it}, \boldsymbol{R})$ 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.

Appendix of Chapter 3

Covariance derivatives of the GP distribution

The derivatives of $\sigma^2_{itt^*}(\boldsymbol{\delta})$, i.e., $\boldsymbol{D}_{i1}(\boldsymbol{\beta})$, $\boldsymbol{D}_{i2}(\boldsymbol{\delta})$ and $\boldsymbol{D}_{i12}(\boldsymbol{\delta})$, are given by

$$\boldsymbol{D}_{i1}(\boldsymbol{\beta}) = \Big[\mu_{it}(\boldsymbol{\beta}) x_{itr}\Big]_{\ t=1,\dots,T,\ r=1,\dots,p+1}, \tag{3.7.1}$$

$$\boldsymbol{D}'_{i2}(\boldsymbol{\delta}) = \begin{pmatrix} \left[\begin{array}{c} \rho_{tt^*}(\lambda_1(\gamma))\sqrt{\mu_{it}(\boldsymbol{\beta})\mu_{it^*}(\boldsymbol{\beta})}\times \\ \{b_{it}(\boldsymbol{\alpha}) w_{itr} \varphi_{it^*}(\boldsymbol{\alpha}) + \\ b_{it^*}(\boldsymbol{\alpha}) w_{it^*r} \varphi_{it}(\boldsymbol{\alpha})\} \end{array}\right]_{\substack{(t,t^*)\in I,\\ r=1,\dots,q+1}} \\ \left[\frac{\partial \rho_{tt^*}(\lambda_1(\gamma))}{\partial \lambda_1(\gamma)} \frac{4e^{2\gamma}}{(e^{2\gamma}+1)^2} \sigma_{it}(\boldsymbol{\delta})\sigma_{it^*}(\boldsymbol{\delta})\right]_{(t,t^*)\in I} \end{pmatrix}, \tag{3.7.2}$$

$$\boldsymbol{D}'_{i12}(\boldsymbol{\delta}) = \left[\begin{array}{c} \frac{1}{2}\rho_{tt^*}(\lambda_1(\gamma))\ \varphi_{it}(\boldsymbol{\alpha})\varphi_{it^*}(\boldsymbol{\alpha})\times \\ \sqrt{\mu_{it}(\boldsymbol{\beta})\mu_{it^*}(\boldsymbol{\beta})}\,\{x_{itr} + x_{it^*r}\} \end{array}\right]_{(t,t^*)\in I,\ r=1,\dots,p+1} \tag{3.7.3}$$

where $I := \{(t, t^*) \mid t \le t^*\}$.

Chapter 4

Non nested model selection for spatial count regression models with application to health insurance

4.1 Introduction

We speak of count data when the data values are contained in the natural numbers. A common distribution for count data is the Poisson (Poi) distribution, which is rather restrictive since variance and mean are equal. But often in observed count data the sample variance is considerably larger than the sample mean - a phenomenon called over-dispersion. In such cases the Poisson assumption is not appropriate for analyzing this data.

Frequently the negative Binomial (NB) distribution instead of the Poisson distribution is used to model over-dispersed data. Another possibility for modeling over-dispersion is the generalized Poisson (GP) distribution introduced by Consul and Jain (1973) which allows for a more flexible variance function than the Poisson distribution by an additional parameter (see e.g. Consul and Famoye (1992) and Famoye (1993)).

Over-dispersion may also be caused by a large proportion of zero counts in the data. Yip and Yau (2005) stress that especially claim numbers often exhibit a large number of zeros and hence traditional distributions may be insufficient. In addition to the zeros arising from the count data model, zero-inflated models (see for example Winkelmann (2008)) also allow for excess zeros. Zero-inflated models can be used in combination with any count data distribution. We consider in this chapter the zero-inflated Poisson regression (ZIPR) (see e.g. Lambert (1992)) and the zero-inflated generalized Poisson (ZIGPR) model. ZIGPR models have been investigated by Famoye and Singh (2003), Gupta, Gupta, and Tripathi (2004), Bae, Famoye, Wulu, Bartolucci, and Singh (2005), Joe and Zhu (2005) and Famoye and Singh (2006).

The variability in over-dispersed data can also be interpreted as unobserved heterogeneity which is not sufficiently explained by the covariates. Especially for simple models with few parameters, theoretical model predictions may not match empirical observations for higher moments. When information on the location of the individuals is known, the data is spatially indexed. For count regression models, Gschlößl and Czado (2007) include spatial random effects using a proper conditional autoregressive (CAR) model based on Pettitt, Weir, and Hart (2002). In other words, one assumes random effects associated with geographic areas rather than individuals and presumes that the effects in neighboring regions are similar. In contrast to Gschlößl

and Czado (2007), however, we also include covariates with spatial information, e.g. measures for the degree of urbanity at a certain location. We carry out a comparison investigating whether one of these two spatial specifications or both fit our data better.

Altogether, in this chapter we account for extra variability not only by addressing distributions capable of handling over-dispersion and over-dispersion caused by an excessive number of zeros, we also take extra spatial variability in the data into account.

Since in these spatial models maximum likelihood estimation and confidence interval estimation is not tractable we consider the models in a Bayesian context. Thus, for parameter estimation Markov Chain Monte Carlo (MCMC) methods are used.

Model comparison between different model classes is non standard. For nested models, i.e., when one of the two models is a super model of the other, model comparison may be carried out using tools like Akaike's information criterion or likelihood ratio tests. This condition may be violated when the distribution on which the two models are based, are different. Even within such a class of regression models, two models may be non nested when they use different link functions or when linear predictors are non hierarchical. We utilize a test proposed by Vuong (1989) and the distribution-free test proposed by Clarke (2007) for non nested model comparison and illustrate how they may be applied in a Bayesian context.

This is a novel approach since so far these two tests have only been used in classical estimation. Also, the comparison between spatial covariate and / or spatial effect specifications for count regression data has not been carried out elsewhere.

In our application we consider health insurance policies in the following context: for more than 35000 policyholders, the data contain the number of benefits received by the patients in the ambulant (i.e., outpatient) setting as well as several covariates like the total of all deductibles, age, gender, number of physicians per inhabitants, number of inhabitants per square kilometer and buying power. Further, we quantify the best fitted model according to DIC as well as Vuong and Clarke test.

This chapter proceeds as follows. In Section 4.2 an overview on spatial count regression models as well as the modeling of spatial effects is given, where we introduce a proper Gaussian conditional autoregressive prior based on Pettitt, Weir, and Hart (2002). The necessary background to Bayesian inference and MCMC methods is briefly summarized in Section 4.3. This includes the deviance information criterion of Spiegelhalter, Best, Carlin, and van der Linde (2002) as a model selection criterion. The test proposed by Vuong (1989) and the distribution-free test utilized in a Bayesian framework are presented in Section 4.4. An application to private health insurance data for policyholders in Germany is presented in Section 4.5.

4.2 Spatial count regression models

4.2.1 Spatial effects

Spatial covariates

Spatial variation may sometimes be explained by covariates which vary spatially. Such covariates we call 'spatial covariates'. Examples in our data set are the number of physicians per inhabitant in a certain district, the number of inhabitants per square kilometer or the buying power per district.

CAR

In order to account for spatial heterogeneity we will incorporate, in addition to covariate information, spatial random effects in the regression models. Therefore we consider the Gaussian Conditional Autoregressive (CAR) formulation introduced by Pettitt, Weir, and Hart (2002) which permits the modeling of spatial dependence and dependence between multivariate random variables at irregularly spaced regions. Assume that J regions $\{1, \ldots, J\}$ are given and let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)^t$ the vector of spatial effects for each region. Let $\boldsymbol{\gamma}$ be multivariate normal distributed with

$$\boldsymbol{\gamma} \sim N_J(0, \sigma^2 Q^{-1}) \tag{4.2.1}$$

where the precision matrix $Q = (Q_{ij})_{i,j=1,\ldots,J}$ is given by

$$Q_{ij} = \begin{cases} 1 + |\psi| \cdot N_i & i = j \\ -\psi & i \sim j \\ 0 & \text{otherwise} \end{cases} . \tag{4.2.2}$$

Here the notation $i \sim j$ indicates that the regions i and j are neighbors and N_i denotes the number of neighbors of region i. Thus the full conditional distribution of γ_i given all the other values $\boldsymbol{\gamma}_{-i}$, $i = 1, \ldots, J$ is

$$\gamma_i | \boldsymbol{\gamma}_{-i} \sim N\left(\frac{\psi}{1 + |\psi| \cdot N_i} \sum_{j \sim i} \gamma_j,\ \sigma^2 \frac{1}{1 + |\psi| \cdot N_i} \right) . \tag{4.2.3}$$

Parameter ψ determines the overall degree of spatial dependence. If all regions are spatially independent, i.e., $\psi = 0$, the precision matrix Q (see (4.2.2)) reduces to the identity matrix, whereas for $\psi \to \infty$ the degree of dependence increases. The multivariate normal distribution (4.2.1) is a proper distribution since Pettitt, Weir, and Hart (2002) show that the precision matrix Q is symmetric and positive definite. Another convenient feature of this CAR model is that according to Pettitt, Weir, and Hart (2002) the determinant of Q, which is needed for the update of ψ in a MCMC algorithm, can be computed efficiently.

4.2.2 Count regression models

The count regression models considered in this chapter will be the Poisson (PoiR), the negative Binomial (NBR), the generalized Poisson (GPR), the zero-inflated Poisson (ZIPR) and the zero-inflated generalized Poisson (ZIGPR) regression. In order to allow for a comparison between these distributions, we choose a mean parameterization for all of them. Their probability mass functions (pmf) together with means and variances are given in Table 4.1. Regression models for these considered distributions can be constructed similar to generalized linear models (GLM) (McCullagh and Nelder (1989)). We denote the regression model with response Y_i and (known) explanatory variables $x_i = (1, x_{i1}, \ldots, x_{ip})^t$ for the mean $i = 1, \ldots, n$ by $ZIGPR(\mu_i, \varphi, \omega)$. For individual observation periods, we allow exposure variables t_i, which satisfy $t_i > 0\ \forall i$ and in case without individual exposure $t_i = 1\ \forall i$.

1. **Random component:**
   $\{Y_i, 1 \leq i \leq n\}$ are independent with response distribution $Poi(\mu_i)$, $NB(\mu_i, r)$, $GP(\mu_i, \varphi)$, $ZIP(\mu_i, \omega)$ or $ZIGP(\mu_i, \varphi, \omega)$.

	$P(Y=y)$	$E(Y)$	$Var(Y)$	Parameter restriction
$Poi(\mu)$	$\frac{\exp\{-\mu\}\mu^y}{y!}$	μ	μ	$\mu \in \mathbb{R}$
$NB(\mu, r)$	$\frac{\Gamma(y+r)}{\Gamma(r)y!}\left(\frac{r}{\mu+r}\right)^r\left(\frac{\mu}{\mu+r}\right)^y$	μ	$\mu(1+\frac{\mu}{r})$	$r>0$
$GP(\mu, \varphi)$	$\frac{\mu(\mu+(\varphi-1)y)^{y-1}}{y!}\varphi^{-y}e^{-\frac{1}{\varphi}(\mu+(\varphi-1)y)}$	μ	$\varphi^2\mu$	$\varphi>0$
$ZIP(\mu, \omega)$	$\omega\cdot\mathbb{1}_{\{y=0\}}+(1-\omega)\cdot\frac{\exp(-\mu)\mu^y}{y!}$	$(1-\omega)\mu$	$(1-\omega)\mu(1+\omega\mu)$	$\omega\in(0,1)$
$ZIGP(\mu, \varphi, \omega)$	$\omega\cdot\mathbb{1}_{\{y=0\}}+(1-\omega)\cdot$ $\cdot\frac{\mu(\mu+(\varphi-1)y)^{y-1}}{y!}\varphi^{-y}e^{-\frac{1}{\varphi}(\mu+(\varphi-1)y)}$	$(1-\omega)\mu$	$(1-\omega)\mu(\varphi^2+\omega\mu)$	$\varphi>0, \omega\in(0,1)$

Table 4.1: Pmf's of the Poisson, NB, GP, ZIP and ZIGP distribution together with their means and variances in mean parameterization

2 **Systematic component:**
The linear predictor is $\eta_i^\mu(\boldsymbol{\beta}) = \boldsymbol{x}_i^t\boldsymbol{\beta}+\gamma_i$ which influence the response Y_i. Here, $\boldsymbol{\beta} = (\boldsymbol{\beta}^{NS}, \boldsymbol{\beta}^S)$ are the unknown regression parameters with $\boldsymbol{\beta}^{NS} = (\beta_0, \beta_1, \ldots, \beta_r)^t$ the non-spatial explanatory factors, $\boldsymbol{\beta}^S = (\beta_{r+1}, \beta_{r+2}, \ldots, \beta_p)^t$ the spatial covariates and γ_i the spatial random effects (not included in our base models). The matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^t$ is called design matrix.

3 **Parametric link component:**
To get a positive mean the linear predictor $\eta_i^\mu(\boldsymbol{\beta})$ is related to the parameters $\mu_i(\boldsymbol{\beta})$, $i = 1, \ldots, n$ as follows:

$$\begin{aligned} E(Y_i|\boldsymbol{\beta}) = \mu_i(\boldsymbol{\beta}) &:= t_i\ \exp\left\{\boldsymbol{x}_i^t\beta+\gamma_i\right\} = \exp\left\{\boldsymbol{x}_i^t\beta+\gamma_i+\log(t_i)\right\} \\ \Leftrightarrow \eta_i^\mu(\boldsymbol{\beta}) &= \log(\mu_i(\boldsymbol{\beta}))-\log(t_i) \quad \text{(log - link)} \end{aligned}$$

4.3 MCMC including model selection

In order to incorporate spatial random effects we consider the models in a Bayesian context which allows the modeling of a spatial dependency pattern. The determination of the posterior distributions require high dimensional integrations. MCMC will be used for parameter estimation, in particular we use the Metropolis Hastings sampler introduced by Metropolis et al. (1953) and Hastings (1970). For more information on Bayesian data analysis and MCMC methods see Gilks, S., and D. (1996) and Gelman, Carlin, Stern, and Rubin (2003). Throughout this chapter, an independence MH sampler using the Student's t-distribution with $\nu = 20$ degrees of freedom will be used. For details on the MCMC algorithms see Gschlößl and Czado (2008) and Schabenberger (2009b).

The DIC (Spiegelhalter, Best, Carlin, and van der Linde (2002)) is a popular information criterion which was designed to compare hierarchical models, and can easily be computed using the available MCMC output. Let $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^T$ be a sample from the posterior distribution of the

model. The calculation of the DIC is based on two quantities. On one hand this is the so called *unstandardized deviance* $D(\boldsymbol{\theta}) = -2\log\left(p\left(\boldsymbol{y}|\boldsymbol{\theta}\right)\right)$ where $p\left(\boldsymbol{y}|\boldsymbol{\theta}\right)$ is the observation model and on the other hand the so called effective number of parameters p_D defined by

$$p_D := \overline{D(\boldsymbol{\theta}|\boldsymbol{y})} - D(\bar{\boldsymbol{\theta}}).$$

Here $\overline{D(\boldsymbol{\theta}|\boldsymbol{y})} := \frac{1}{T}\sum_{t=1}^{T} D\left(\boldsymbol{\theta}^t\right)$ is the estimated posterior mean of the deviance and $D(\bar{\boldsymbol{\theta}})$ is the deviance of the estimated posterior means $\bar{\boldsymbol{\theta}} := \frac{1}{T}\sum_{t=1}^{T} D\left(\boldsymbol{\theta}^t\right)$. Finally the DIC determined as

$$\text{DIC } = \overline{D(\boldsymbol{\theta}|\boldsymbol{y})} + p_D = 2\overline{D(\boldsymbol{\theta}|\boldsymbol{y})} - D(\bar{\boldsymbol{\theta}}).$$

The preferred model is the one which has the smallest DIC. DIC depends on the specific values obtained in an MCMC run, thus it is difficult to assess how different DIC values have to be for different models to select among these models. For exponential family models DIC approximates the Akaike information criterion (AIC).

4.4 Non nested model selection

We use tests proposed by Vuong (1989) and Clarke (2003) to compare regression models which need not to be nested. These tests are based on the Kullback-Leibler information criterion (KLIC). According to Vuong (1989) the Kullback-Leibler distance is defined as

$$KLIC := E_0[\log h_0(Y_i|\boldsymbol{x}_i)] - E_0[\log f(Y_i|\boldsymbol{\upsilon}_i, \hat{\boldsymbol{\delta}})],$$

where $h_0(\cdot|\cdot)$ is the true conditional density of Y_i given $\boldsymbol{x}_i$, that is, the true but unknown model. Let E_0 denote the expectation under the true model, $\boldsymbol{\upsilon}_i$ are the covariates of the estimated model and $\hat{\boldsymbol{\delta}}$ are the pseudo-true values of $\boldsymbol{\delta}$ in model with $f(Y_i|\boldsymbol{\upsilon}_i, \boldsymbol{\delta})$, which is not the true model. Generally, the model with minimal $KLIC$ is the one that is closest to the true, but unknown, specification.

4.4.1 Vuong test

Consider two models, $f_1 = f_1(Y_i|\boldsymbol{\upsilon}_i, \hat{\delta}^1)$ and $f_2 = f_2(Y_i|\boldsymbol{\omega}_i, \hat{\delta}^2)$ then if model 1 is closer to the true specification, we have

$$\begin{aligned} E_0[\log h_0(Y_i|\boldsymbol{x}_i)] - E_0[\log f_1(Y_i|\boldsymbol{\upsilon}_i, \hat{\boldsymbol{\delta}}^1)] &< E_0[\log h_0(Y_i|\boldsymbol{x}_i)] - E_0[\log f_2(Y_i|\boldsymbol{\omega}_i, \hat{\boldsymbol{\delta}}^2)] \\ \Leftrightarrow E_0\left[\log \frac{f_1(Y_i|\boldsymbol{\upsilon}_i, \hat{\boldsymbol{\delta}}^1)}{f_2(Y_i|\boldsymbol{\omega}_i, \hat{\boldsymbol{\delta}}^2)}\right] &> 0 \end{aligned} \tag{4.4.1}$$

Vuong defines the statistics

$$m_i := \log\left(\frac{f_1(y_i|\boldsymbol{\upsilon}_i, \hat{\boldsymbol{\delta}}^1)}{f_2(y_i|\boldsymbol{\omega}_i, \hat{\boldsymbol{\delta}}^2)}\right), \quad i = 1, \ldots, n. \tag{4.4.2}$$

If h_0 is the true probability mass function, then $\mathbf{m} = (m_1, \ldots, m_n)^t$ is a random vector with mean $\boldsymbol{\mu}_0^m = (\mu_1^m, \ldots, \mu_n^m) := E_0(\mathbf{m})$. Hence, we can test the null hypothesis

$$H_0 : \boldsymbol{\mu}_0^m = \mathbf{0} \text{ against } H_1 : \boldsymbol{\mu}_0^m \neq \mathbf{0}.$$

The mean $\boldsymbol{\mu}_0^m$ in the above hypothesis is unknown. With convenient standardization and the central limit theorem Vuong (1989) shows that under H_0

$$\nu := \frac{\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^n m_i\right]}{\sqrt{\frac{1}{n}\sum_{i=1}^n (m_i - \bar{m})^2}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \quad \text{as } n \to \infty$$

where $\bar{m} := \frac{1}{n}\sum_{i=1}^n m_i$. This allows to construct an asymptotic α-level test of $H_0 : \boldsymbol{\mu}_0^m = \mathbf{0}$ versus $H_1 :$ not H_0. It rejects H_0 if and only if $|\nu| \geq z_{1-\frac{\alpha}{2}}$, where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution. The test chooses model 1 over 2, if $\nu \geq z_{1-\frac{\alpha}{2}}$. This is reasonable since according to the equivalence given in (4.4.1), significantly high values of ν indicate a higher $KLIC$ of model 1 as compared to model 2. Similarly, model 2 is chosen if $\nu \leq -z_{1-\frac{\alpha}{2}}$. No model is preferred for $-z_{1-\frac{\alpha}{2}} < \nu < z_{1-\frac{\alpha}{2}}$. According to Clarke (2007, p. 349) the Vuong test must be corrected if the number of estimated coefficients in each model is different. Vuong (1989) suggests to use the Schwarz correction, which is given by

$$\left[\left(\frac{p}{2}\log n\right) - \left(\frac{q}{2}\log n\right)\right]. \tag{4.4.3}$$

Here p and q are the number of estimated coefficients in models f_1 and f_2, respectively (Clarke (2003, p. 78)). Thus the Vuong test statistic ν with Schwarz correction is defined as:

$$\tilde{\nu} := \frac{\sqrt{n}\left(\left[\frac{1}{n}\sum_{i=1}^n m_i\right] - \left[\left(\frac{p}{2}\log n\right) - \left(\frac{q}{2}\log n\right)\right]/n\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^n (m_i - \bar{m})^2}}.$$

4.4.2 Clarke test

An alternative to the Vuong test is a distribution-free test (see Clarke (2007)) which applies a modified paired sign test to the differences in the individual log-likelihoods from two non nested models. The null hypothesis of the distribution-free test is

$$H_0 : P_0\left[\log\frac{f_1(Y_i|\boldsymbol{\upsilon}_i, \hat{\boldsymbol{\delta}}^1)}{f_2(Y_i|\boldsymbol{\omega}_i, \hat{\boldsymbol{\delta}}^2)} > 0\right] = 0.5. \tag{4.4.4}$$

Under the null hypothesis (4.4.4) the log-likelihood ratios should be symmetrically distributed around zero. That means that about half the log-likelihood ratios should be greater and half less than zero. Using m_i as defined in (4.4.2), Clarke considers the test statistic

$$B = \sum_{i=1}^n \mathbb{1}_{\{0,+\infty\}}(m_i), \tag{4.4.5}$$

where $\mathbb{1}_A$ is the indicator function which is 1 on the set A and 0 elsewhere. The quantity B is the number of positive differences and follows a Binomial distribution with parameters n and probability 0.5 under H_0. If B is, under the null hypothesis, significantly larger than its expected value, model f_1 is "better" than model f_2. This allows to construct the following distribution-free test.

First let $m_i(Y_i)$ correspond to the random variable with value m_i, then the null hypothesis (4.4.4) is equivalent to

$$H_0^{DF} : P_0\left[m_i(Y_i) > 0\right] = 0.5 \quad \forall i = 1, \ldots, n.$$

For the test problem $H_0^{DF} : P_0 [m_i (Y_i) > 0] = 0.5 \quad \forall i = 1, \ldots, n$ versus $H_{1+}^{DF} : P_0 [m_i (Y_i) > 0] > 0.5,\ i = 1, \ldots, n$, the corresponding α - level upper tail test rejects H_0^{DF} versus H_{1+}^{DF} if and only if $B \geq c_{\alpha+}$, where $c_{\alpha+}$ is the smallest integer such that $\sum_{c=c_{\alpha+}}^{n} \binom{n}{c} 0.5^n \leq \alpha$. If the upper tail test rejects H_0^{DF} then we decide that model 1 is preferred over model 2. For the alternative $H_{1-}^{DF} : P_0 [m_i (Y_i) > 0] < 0.5,\ i = 1, \ldots, n$, the α - level lower tail test rejects H_0^{DF} versus H_{1-}^{DF} if and only if $B \leq c_{\alpha-}$, where $c_{\alpha-}$ is the largest integer such that $\sum_{c=0}^{c_{\alpha-}} \binom{n}{c} 0.5^n \leq \alpha$ (compare to Clarke (2007, p. 349)). If H_0^{DF} versus H_{1-}^{DF} is rejected, then model 2 is preferred over model 1. If H_0^{DF} cannot be rejected, no model is preferred.

Like the Vuong test this test is sensitive to the number of estimated coefficients in each model. Once again, we need a correction for the degrees of freedom.

Since the distribution-free tests work with the individual log-likelihood ratios, we cannot apply the Schwarz correction as in the Vuong test with the "summed" log-likelihood ratio. Clarke (2003) suggests to apply the average correction to the individual log-likelihood ratios. So we correct the individual log-likelihoods for model f_1 by a factor of $\left[\left(\frac{p}{2n} \log n\right)\right]$ and the individual log-likelihoods for model f_2 by a factor of $\left[\left(\frac{q}{2n} \log n\right)\right]$.

In the Bayesian approach we can quantify the uncertainty of the test decisions for the Vuong and Clarke test accordingly. For this we utilize the sampled parameter values from the MCMC output and determine the test decision for each sampled value. This allows to estimate the posterior percentages of how many times model 1 (model 2) was chosen over model 2 (model 1) and the percentage of no test decision.

All MCMC algorithms for model fit and the model comparison are implemented in package `spatcounts` (Schabenberger (2009a)) in *R*, which is available on CRAN.

4.5 Application

We now apply the models described in Section 4.3 to a large portfolio of a German health insurer. Before the parametric models are fitted, a basic exploratory analysis is carried out. At the end of this Section, all fitted models are compared using the DIC as well as the Vuong and the Clarke tests described in Section 4.4.

4.5.1 Data description and exploration

The data set considers 37751 insured persons of a private health insurance company in 2007. The response variable is the number of benefits received per patient for ambulant treatments. In the German private health care system, the policyholders may opt to cover a part of each invoice themselves, this amount is called deductible. Depending on the policy type and the treatment setting, deductibles can be either an annual total or a percentage of each invoice. If no bill is reimbursed throughout the whole year, the policyholder receives a refund. A variable description including the response variable and the explanatory variables is given in Table 4.2. Germany has 439 districts. The data includes patients from all districts.

Around 76% of the insured persons are male, which is typical for the policy line considered. To obtain a first overview of the dependent variable Y_i, a histogram of the observed count frequencies is given in Figure 4.1). For a better graphical illustration, outliers $Y_i > 50$ are not displayed. The histogram shows that we have a high variation in Y_i and a rather large number of zeros. In particular 43% of the response data is equal to zero. The covariates can be split up into two groups. The first group of the covariates depend on the patient like the total of

Variable	Type	Description
Y_i	discrete	Number of outpatient benefits received by patient i.
DED_i	continuous	Total of all deductibles of patient i.
AGE_i	discrete	Age of patient i.
SEX_i	binary	Indicator for gender of patient i. (0 = female, 1 = male)
ZIP_i	categorical	ZIP Code of the home address of patient i.
$D(i)$	categorical	Indicates the home district for patient i.
$PHYS.INH_j$	continuous with spatial information	Number of physicians per inhabitant in district j multiplied by 100.
$URBAN_j$	continuous with spatial information	Number of inhabitants per square kilometer in district j.
BP_j	continuous with spatial information	Buying power in district j.

Table 4.2: Variable description for the analyzed health insurance data set

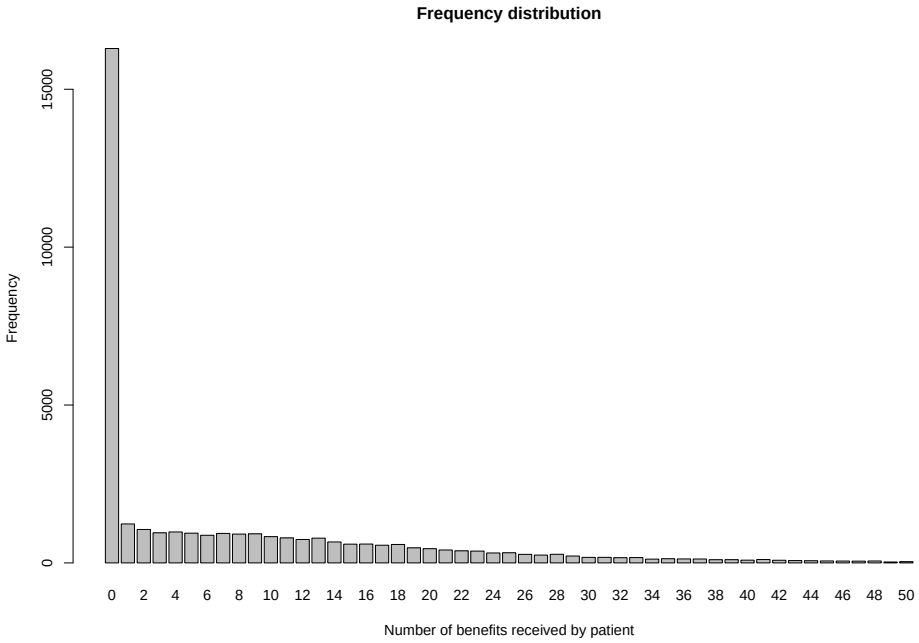

Figure 4.1: Frequency distribution for the response variables ($\boldsymbol{Y} \in [0, 705]$ without outliers $Y_i > 50$).

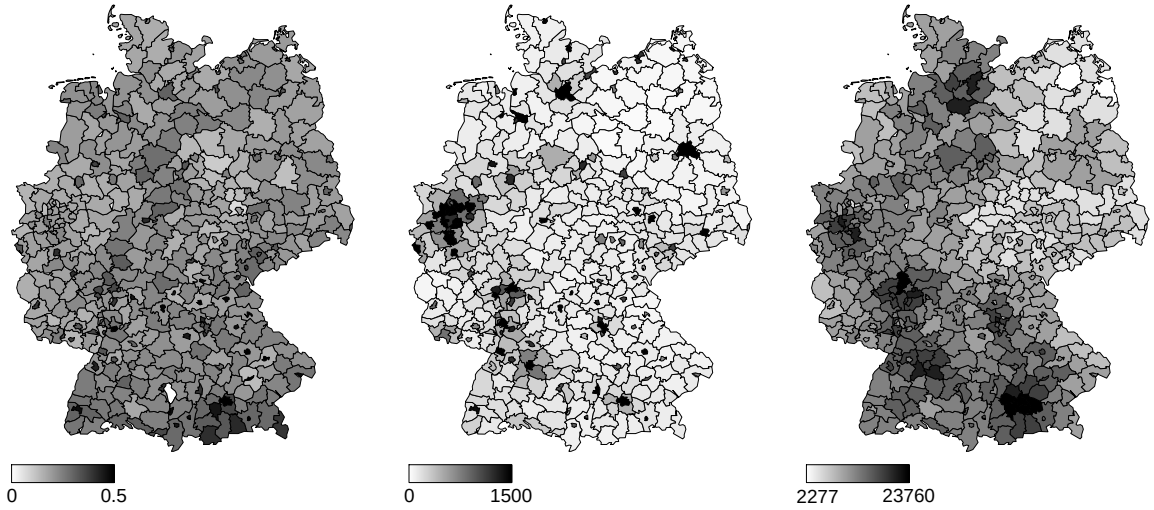

Figure 4.2: Exploratory maps of the spatial covariates **PHYS.INH** (left panel), **URBAN** (middle panel) and **BP** (right panel).

all deductibles with values $\mathbf{DED} \in [0, 1821]$, the age with $\mathbf{AGE} \in [3, 88]$ or the gender dummy **SEX**. The second group of covariates are spatial covariates like the number of physicians per inhabitant with $\mathbf{PHYS.INH} \in [0, 0.5622]$, the number of inhabitants per square kilometer with value $\mathbf{URBAN} \in [39.28, 4060]$ or the average buying power $\mathbf{BP} \in [12277.4, 23760.38]$ in Euro. The maps in Figure 4.2 show the spatial distribution of the spatial covariates. The number of physicians per inhabitant in Germany seem to be distributed very uniformly (left panel of Figure 4.2) whereas the most inhabitants per square kilometer can be found in the larger cities, e.g. Berlin, Bremen, Hamburg, Munich or the Ruhr area (middle panel of Figure 4.2). West Germany has higher buying power with a peak around Munich compared to East Germany (former German Democratic Republic) (see right panel of Figure 4.2).

A natural next step is to look at scatter plots of the dependent variable $\boldsymbol{Y}$ against each of the regressors. The LOWESS (solid line) and the GAM (dashed line) smoothing curves of the scatter plot in Figure 4.3 indicates that the variable **AGE** has to be transformed, i.e., we allow a quadratic influence on the response. In health insurance this is not unusual since in general children and older people need more medical attendance. For numerical stability we use standardized (sometimes called autoscaled) covariates for the variables **DED**, **PHYS.INH**, **URBAN** and **BP** denoted with ".s".

4.5.2 Identification of base models

To establish base models we first analyze the data set in the statistical program "R" without spatial effects. We allow for an intercept, the covariates gender (**SEX**), the standardized covariates **DED.s**, **PHYS.INH.s**, **URBAN.s** and **BP.s** as well as the orthogonal polynomial transformed covariates **AGE.p1** (polynomial transformation of degree 1), **AGE.p2** (polynomial transformation of degree 2). For maximum likelihood parameter estimates we use the function `est.zigp()` in the `R` package `ZIGP` developed by Erhardt (2009) for all models except the negative Binomial regression model which is estimated with the basic `R` library `MASS` using the function `glm.nb()`.

In a next step sequential elimination according to a Wald test with 5% α-level of significance

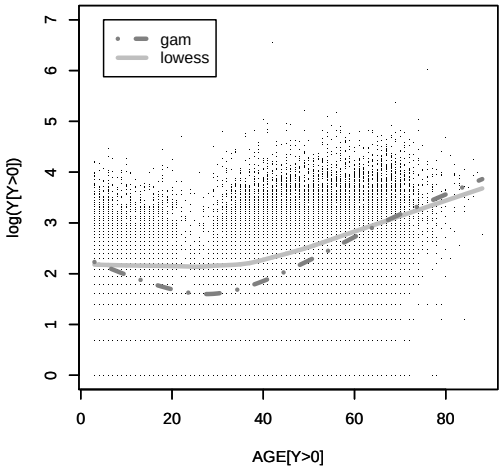

Figure 4.3: Scatter plot (including gam (dashed) and lowess (solid) smoothing lines) and box plot of the number of benefits received per patient against age of the patient.

is conducted. In Table 4.3 the full and reduced regression specifications are given for every model class considered in Section 4.2. The penalty term in the AIC statistic includes parameters which are estimated (such as φ in $GP(\mu_i, \varphi)$) and does not include them if they are fix (such as $\varphi = 1$ in $Poi(\mu_i)$). We stress that the comparison of different models based on AIC is only possible within one model class, that is when the distribution of the responses are the same and designs are hierarchical. If the models are non nested, the test decisions should be based on the Vuong test or the distribution-free test (Clarke test).

Table 4.4 displays for the models NBR, GPR, ZIPR and ZIGPR (defined as model (**I**) and PoiR, NBR, GPR and ZIPR (defined as model **II**) the entries of the Vuong and Clarke tests for each combination of model (**I**) and model (**II**). We choose an α-level of 5%, i.e., $z_{1-\frac{\alpha}{2}} = 1.96$. In the first line of each cell, the Vuong test statistic ν is given. In the second and third line the decision of the Vuong test (V) and the Clarke test (C) is shown, i.e., if model (**I**) or (**II**) is better. The corresponding p-values for each test are given in parentheses. For example V: (I) $\left(< 2 \cdot 10^{-16}\right)$ means that the Vuong test prefers model (**I**) with p-value smaller than $2 \cdot 10^{-16}$. We now discuss the conclusions to be drawn from Table 4.4. Since the Poisson model is not preferred over any of the other model classes, we see evidence that the data is in fact overdispersed. Overdispersion may be explained either by a dispersion parameter as in the GPR or the NBR model, by excess zeros as in the ZIPR model, or both. Since the GPR model outperforms the NBR model, we consider zero-inflation jointly with the GPR distribution, i.e., we also fit a ZIGPR model. In general, the tests by Vuong and Clarke are suitable for pairwise model comparison, thus they do not have to lead to an overall decision between all model classes, much less do both test necessarily decide equivalently. In our case, however, the pairwise decisions given in Table 4.4 are identical, and we can sort the models in a unique ranking: the GPR model outperforms all other models and is followed downward by ZIGPR, NBR, ZIPR and the PoiR model. The comparison of the $ZIGPR(\mu_i, \varphi, \omega)$ model to all other model classes gives almost identical results as the comparison of the $GPR(\mu_i, \varphi)$ model to these classes. The reason is that the zero-inflation parameter in the ZIGPR model is estimated almost to zero (see Table 4.3) and therefore the ZIGPR fit is almost identical to the GPR fit. In the comparison between the GPR and ZIGPR model, the GPR model by far outperforms the ZIGPR model. This can be explained by the nature of the two test: even if the likelihood contributions per observation in

Model	Model equation μ	Dispersion (SE)	Zero-inflation (SE)	$l(\hat{\boldsymbol{\theta}})$	Parameters	AIC
$PoiR(\mu_i)$	**1 + DED.s + AGE.p1 + AGE.p2 + SEX + PHYS.INH.s + URBAN.s + BP.s**	$\varphi = 1$ (not estimated)	$\omega = 0$ (not estimated)	-218 486.8	8	436 990
$NBR(\mu_i, r)$ (full)	**1 + DED.s + AGE.p1 + AGE.p2 + SEX + PHYS.INH.s + URBAN.s + BP.s**	$\hat{r} = 0.5811$ (0.0062)		-99 552.8	9	199 124
$NBR(\mu_i, r)$ (reduced)	**1 + DED.s + AGE.p1 + AGE.p2 + SEX + BP.s**	$\hat{r} = 0.5811$ (0.0062)		-99 553.3	7	199 121
$GPR(\mu_i, \varphi)$ (full)	**1 + DED.s + AGE.p1 + AGE.p2 + SEX + PHYS.INH.s + URBAN.s + BP.s**	$\hat{\varphi} = 4.6369$ (0.0397)	$\omega = 0$ (not estimated)	-96 849.1	9	193 716
$GPR(\mu_i, \varphi)$ (reduced)	**1 + DED.s + AGE.p1 + AGE.p2 + SEX + URBAN.s + BP.s**	$\hat{\varphi} = 4.6893$ (0.0410)	$\omega = 0$ (not estimated)	-96 850.5	8	193 717
$ZIPR(\mu_i, \omega)$ (full)	**1 + DED.s + AGE.p1 + AGE.p2 + SEX + PHYS.INH.s + URBAN.s + BP.s**	$\varphi = 1$ (not estimated)	$\hat{\omega} = 0.4312$ (0.0026)	-161 674.1	9	323 366
$ZIPR(\mu_i, \omega)$ (reduced)	**1 + DED.s + AGE.p1 + AGE.p2 + SEX**	$\varphi = 1$ (not estimated)	$\hat{\omega} = 0.4312$ (0.0026)	-161 675.4	6	323 363
$ZIGPR(\mu_i, \phi, \omega)$	**1 + DED.s + AGE.p1 + AGE.p2 + SEX + PHYS.INH.s + URBAN.s + BP.s**	$\hat{\varphi} = 4.7010$ (0.0414)	$\hat{\omega} = 10^{-6}$ (0.0007)	-96 454.5	10	192 929

Table 4.3: Model specifications and AIC for each of the models after sequential elimination of insignificant covariates according to a Wald test with $\alpha = 5\%$

(I) \ **(II)**	$PoiR(\mu_i)$	$NBR(\mu_i, r)$	$GPR(\mu_i, \varphi)$	$ZIPR(\mu_i, \omega)$
$NBR(\mu_i, r)$	$\nu = 30.2$ V: (I) $\left(< 2 \cdot 10^{-16}\right)$ C: (I) $\left(< 2 \cdot 10^{-16}\right)$			
$GPR(\mu_i, \varphi)$	$\nu = 34.7$ V: (I) $\left(< 2 \cdot 10^{-16}\right)$ C: (I) $\left(< 2 \cdot 10^{-16}\right)$	$\nu = 4.2$ V: (I) $\left(2.26 \cdot 10^{-5}\right)$ C: (I) $\left(< 2 \cdot 10^{-16}\right)$		
$ZIPR(\mu_i, \omega)$	$\nu = 21.4$ V: (I) $\left(< 2 \cdot 10^{-16}\right)$ C: (I) $\left(< 2 \cdot 10^{-16}\right)$	$\nu = -24.7$ V: (II) $\left(< 2 \cdot 10^{-16}\right)$ C: (II) $\left(< 2 \cdot 10^{-16}\right)$	$\nu = -25.0$ V: (II) $\left(< 2 \cdot 10^{-16}\right)$ C: (II) $\left(< 2 \cdot 10^{-16}\right)$	
$ZIGPR(\mu_i, \varphi, \omega)$	$\nu = 34.7$ V: (I) $\left(< 2 \cdot 10^{-16}\right)$ C: (I) $\left(< 2 \cdot 10^{-16}\right)$	$\nu = 4.3$ V: (I) $\left(2.14 \cdot 10^{-5}\right)$ C: (I) $\left(< 2 \cdot 10^{-16}\right)$	$\nu = -137$ V: (II) $\left(< 2 \cdot 10^{-16}\right)$ C: (II) $\left(< 2 \cdot 10^{-16}\right)$	$\nu = 25.0$ V: (I) $\left(< 2 \cdot 10^{-16}\right)$ C: (I) $\left(< 2 \cdot 10^{-16}\right)$

Table 4.4: Model comparison using the Vuong and the Distribution-Free (Clarke) test; test statistic ν of the Vuong test together with decision according to Vuong (V) and Clarke (C) and their p-values, respectively.

both of these models are almost identical, there is a minimal correction toward the GPR model by virtue of the larger Schwarz penalty term, which corrects for the additional zero-inflation parameter in the ZIGPR model. Notwithstanding this application, overdispersion explained by both a dispersion parameter and zero-inflation simultaneously is present in many other applications, e.g. the ZIGPR model considered by Czado, Erhardt, Min, and Wagner (2007) to analyze patent filing processes.

By including a random spatial effect for each region extra heterogeneity in the data might be taken into account by assuming a finer geographic resolution. The CAR prior presented in Section 4.2 will be assumed for these spatial effects.

4.5.3 Bayesian inference using MCMC

The MCMC algorithms for the PoiR, NBR, GPR, ZIPR and ZIGPR models are run for 50000 iterations. The mean parameter μ_i, $i = 1, \ldots, n$ has the general form

$$\mu_i = t_i \cdot \exp\left(\boldsymbol{x}_i^t \boldsymbol{\beta} + \gamma_{D_{(i)}}\right)$$

with the observation specific exposure t_i fixed to 1. We fit models with spatial covariates only (denoted by SC), models with spatial random effects only (denoted by CAR) and models with both spatial random effects and spatial covariates (denoted by CAR+SC). Recall that we have the spatial covariates: number of physicians per inhabitants (**PHYS.INH.s**), number of inhabitants per square kilometer (**URBAN.s**) and buying power (**BP.s**).

The starting values for each parameter of the four models are taken from the regression without spatial effect. That means we use the results of the `R` functions `est.zigp()` and `glm.nb()` for all models with all covariates for SC and CAR+SC and without the spatial covariates for the CAR model. The posterior means and 80% credible intervals for the model specific parameters r, φ and ω in the different models are shown in Table 4.5 (the posterior means and 80% credible intervals for the regression parameter vector $\boldsymbol{\beta}$ can be found in Schabenberger (2009b, p. 59)). As in the base models in Section 4.5.2, the zero-inflation parameter in the ZIGPR model is very close to zero for the SC, CAR and SC+CAR specifications. Note that only positive zero-inflation is allowed, therefore the credible intervals cannot contain the zero. Since the ZIGPR model becomes a GPR model when there is no zero-inflation present, we will no longer consider the ZIGPR model for the remainder of this chapter.

Estimation of the regression parameter slightly differs between the models and also changes when spatial effects are added, especially for the GPR models where large spatial effects are observed. Although there are some insignificant covariates we do not reduce the models to compare whether SC, CAR or CAR+SC is preferred. Estimation of the specific parameters is rather similar in all models SC, CAR and CAR+SC. The range of the estimated spatial effects in all of the models is roughly the same in each model even though the Poisson model captures unexplained heterogeneity only by spatial effects. In the ZIP model the proportion of extras zeros ω is estimated as 43%.

In Figure 4.4 we present map plots of the estimated posterior means. In Figure 4.5 the 80% credible intervals of the spatial effects in the Poisson, negative Binomial, generalized Poisson and zero-inflated Poisson models are given. In each Figure the model specification SC is shown in the left panel, the CAR model specification in the middle panel and the CAR+SC model specification in the right panel. Here we see that the spatial effects of all four regression models are almost the same. The spatial covariates have nearly no influence but according to the 80%

Parameter	Model	Mean	(10%, 90%)
	SC	0.5808	(0.5723, 0.5887)
r **in NBR**	CAR	0.5912	(0.5831, 0.5995)
	CAR+SC	0.5910	(0.5830, 0.5993)
	SC	4.6840	(4.6271, 4.7412)
φ **in GPR**	CAR	4.4492	(4.3994, 4.4999)
	CAR+SC	4.4488	(4.3985, 4.4994)
	SC	0.4312	(0.4278, 0.4346)
ω **in ZIPR**	CAR	0.4310	(0.4276, 0.4345)
	CAR+SC	0.4310	(0.4276, 0.4344)
	SC	4.6825	(4.6544, 4.7110)
φ **in ZIGPR**	CAR	4.4514	(4.4219, 4.4805)
	CAR+SC	4.4518	(4.4214, 4.4792)
	SC	$2.4 \cdot 10^{-4}$	$(2.0 \cdot 10^{-5}, 5.5 \cdot 10^{-4})$
ω **in ZIGPR**	CAR	$1.6 \cdot 10^{-4}$	$(1.9 \cdot 10^{-5}, 4.0 \cdot 10^{-4})$
	CAR+SC	$1.5 \cdot 10^{-4}$	$(1.3 \cdot 10^{-5}, 3.4 \cdot 10^{-4})$

Table 4.5: Estimated posterior means and 80% credible intervals for the model specific parameters in the considered SC, CAR, CAR+SC models

credible interval they have a negative spatial effect. According to the 80% credible intervals the CAR and the CAR+SC models have small significant spatial effect.

Unfortunately, the estimated empirical autocorrelations in some of the models decrease very slow. Therefore to compare the different models we decide to thin the 50000 MCMC output by choosing every 200th value.

In order to compare these models, the DIC, defined in Section 4.3, is considered. In Table 4.6, the DIC, the posterior mean of the deviance and the effective number of parameters are listed for each model. $E\left[D(\boldsymbol{\theta}|\boldsymbol{y})\right]$ is based only on the unscaled deviance (see Section 4.3) which cannot be interpreted directly as an overall goodness of fit measure of one specific model. However, $E\left[D(\boldsymbol{\theta}|\boldsymbol{y})\right]$ can be used for comparing the model fit of several models when the number of parameters is roughly the same.

For each regression model the model SC has the highest DIC value. The DIC for the CAR and CAR+SC model is roughly the same. For SC models the effective number of parameters p_D is close to the true number, which is eight for the Poisson regression model and nine for the NBR, GPR and ZIPR model. This is to be expected, since these models do not include random effects. When spatial effects are added, the number of effective parameters increases rapidly. The DIC and the posterior mean of the deviance, $E\left[D(\boldsymbol{\theta}|\boldsymbol{y})\right]$, for CAR are the smallest in all regression models except for the Poisson model. Here the DIC value of CAR+SC is slightly lower than the one of CAR.

Note that the DIC must be used with care, since strictly speaking the DIC is defined for distributions of the exponential family only. Additionally, if two models have similar DIC values it is possible that the model decision varies for different MCMC runs. Therefore we make another comparison using the Vuong and the Clarke test discussed in Section 4.4.

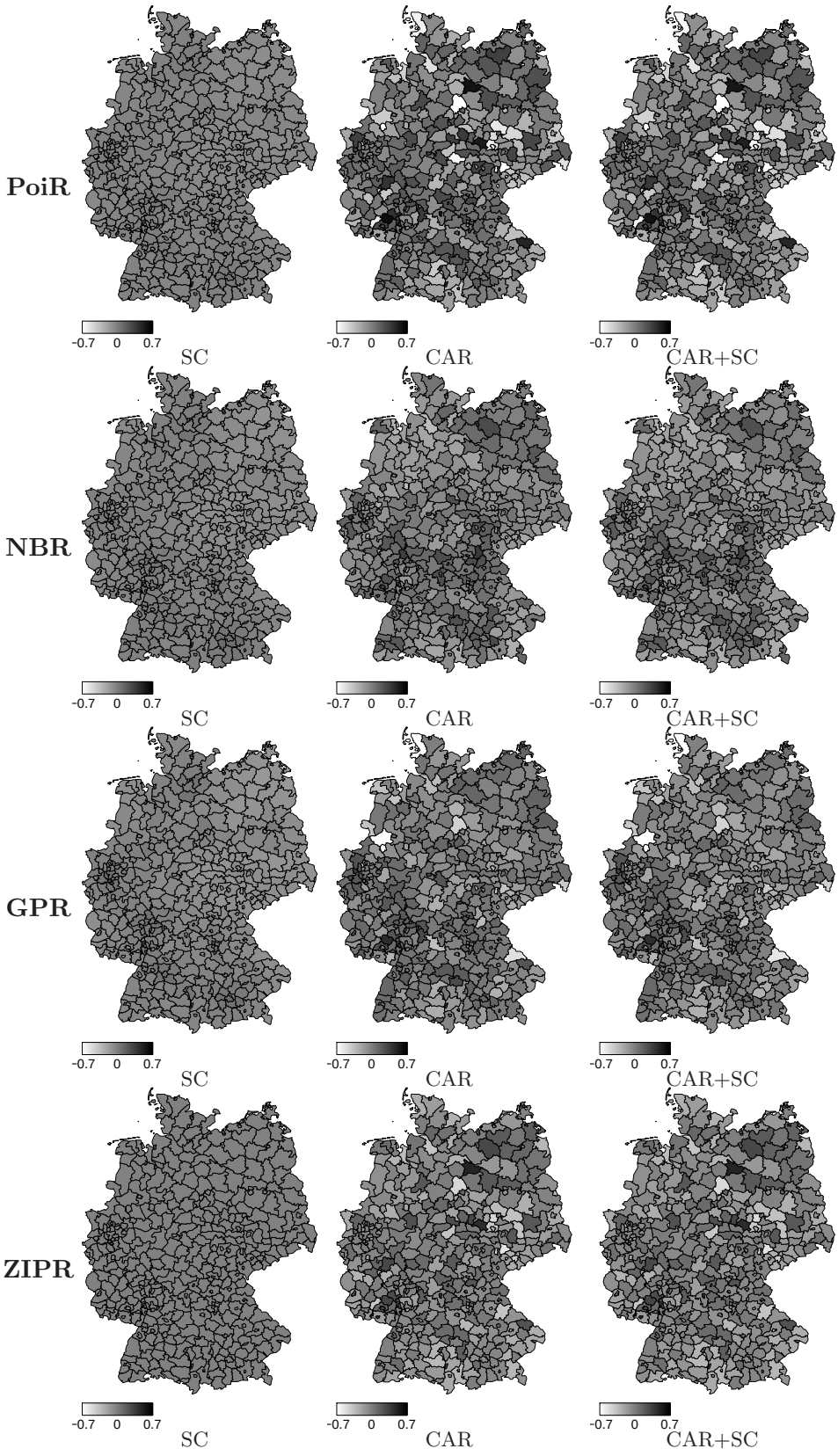

Figure 4.4: Maps of the estimated posterior means (top panels) of the spatial effects in the PoiR, NBR, GPR and ZIPR models SC, CAR and CAR+SC

Figure 4.5: Maps of the 80% credible intervals (white: strictly positive, black: strictly negative, gray: zero is contained in 80% credible interval) of the spatial effects in the PoiR, NBR, GPR and ZIPR models SC, CAR and CAR+SC

	Model	DIC	$E\left[D(\boldsymbol{\theta}\vert\boldsymbol{y})\right]$	p_D
	SC	436990.8	436982.9	7.89
PoiR	CAR	426818.5	426390.3	428.21
	CAR+SC	426817.2	426388.9	428.24
	SC	199124.9	199115.8	9.10
NBR	CAR	198867.8	198651.8	216.00
	CAR+SC	198868.1	198650.8	217.29
	SC	192927.9	192918.5	9.33
GPR	CAR	190764.4	190461.8	302.64
	CAR+SC	190764.7	190461.4	303.30
	SC	323367.0	323357.7	9.36
ZIPR	CAR	318740.5	318364.8	375.66
	CAR+SC	318742.2	318366.2	376.03

Table 4.6: DIC, $E\left[D(\boldsymbol{\theta}\vert\boldsymbol{y})\right]$ and effective number of parameters p_D for the different models

4.5.4 Model selection

Selecting spatial models

First of all we compare SC, CAR and CAR+SC for each regression model $PoiR(\mu_i)$, $NBR(\mu_i, r)$, $GPR(\mu_i, \varphi)$ and $ZIPR(\mu_i, \omega)$. Table 4.7 shows the percentage of 250 Vuong and Clarke test decisions between model (I) and model (II). For the Vuong test we use the statistic ν and choose again an α-level of 5%, i.e., the decision border is $z_{1-\frac{\alpha}{2}} = 1.96$ For the Clarke test we report B/n. The number of parameters p and q of model (**I**) and (**II**), neccessary for the corrections, are taken from the DIC calculations, i.e., we use the effective number of parameters p_D. The decisions of the Vuong and Clarke tests given in Table 4.7 are not consistent. For the Poisson regression models the SC specification performs poorly, however for the comparison between the CAR and CAR+SC specifications only the Clarke test slightly prefers CAR. Since this model has less covariates than CAR+SC, we choose this design as the preferred one within the Poisson class. For the negative Binomial model there is no distinct decision between CAR and CAR+SC, however the SC model is preferred over both of them. The same holds for the ZIPR class. For the generalized Poisson regression models the test by Vuong prefers none of the models in all three comparisons. Therefore we only consider the Clarke test, which slightly decides toward the CAR model. Since this model also has the smallest DIC value (see Table 4.6), we choose the CAR specification within the GPR class.

Selecting count distribution

Now we want to compare the observed preferred models PoiR CAR, NBR SC, GPR CAR and ZIPR SC to get the overall favored one. Therefore we use again the Vuong and the Clarke test like in the section before. The results are shown in Table 4.8. The first value in the round brackets favors model (**I**), the second one stands for no decision taken and the right one prefers model (**II**) (all in percent). For example (100%, 0%, 0%) means the test perfers model (**I**) over model (**II**) in 100% of the sampled MCMC posterior parameter values based on 250 iterations.

	Model (**I**)/ (**II**)	Test	Decision f. model (**I**)	No decision	Decision f. model (**II**)
PoiR	SC/CAR+SC	Vuong	0.0%	6.0%	94.0%
		Clarke	0.0%	1.2%	98.8%
	CAR/CAR+SC	Vuong	0.0%	100.0%	0.0%
		Clarke	40.8%	19.2%	40.0%
	CAR/SC	Vuong	93.6%	6.4%	0.0%
		Clarke	98.8%	1.2%	0.0%
NBR	SC/CAR+SC	Vuong	100.0%	0.0%	0.0%
		Clarke	100.0%	0.0%	0.0%
	CAR/CAR+SC	Vuong	1.2%	98.8%	0.0%
		Clarke	47.6%	4.0%	48.4%
	CAR/SC	Vuong	0.0%	0.0%	100.0%
		Clarke	0.0%	0.0%	100.0%
GPR	SC/CAR+SC	Vuong	0.0%	100.0%	0.0%
		Clarke	14.4%	35.2%	50.4%
	CAR/CAR+SC	Vuong	0.4%	99.6%	0.0%
		Clarke	44.0%	18.4%	37.6%
	CAR/SC	Vuong	0.0%	100.0%	0.0%
		Clarke	55.2%	30.0%	14.8%
ZIPR	SC/CAR+SC	Vuong	0.0%	100.0%	0.0%
		Clarke	100.0%	0.0%	0.0%
	CAR/CAR+SC	Vuong	0.0%	100.0%	0.0%
		Clarke	51.6%	0.0%	48.4%
	CAR/SC	Vuong	0.0%	100.0%	0.0%
		Clarke	0.0%	0.0%	100.0%

Table 4.7: Decision of the Vuong and Clarke tests between model (I) and model (II) as a percentage

The generalized Poisson regression model CAR seems to fit our data in terms of the Vuong test and the Clarke test the best. This model is preferred over all other models discussed (see Table 4.8).

(I) \ (II)		PoiR CAR	NBR SC	GPR CAR
NBR SC	V	(100%, 0%, 0%)		
	C	(100%, 0%, 0%)		
GPR CAR	V	(100%, 0%, 0%)	(100%, 0%, 0%)	
	C	(100%, 0%, 0%)	(100%, 0%, 0%)	
ZIPR SC	V	(100%, 0%, 0%)	(0%, 0%, 100%)	(0%, 0%, 100%)
	C	(100%, 0%, 0%)	(0%, 0%, 100%)	(0%, 0%, 100%)

Table 4.8: Selection of the response distributions ((I)>(II),(I)=(II),(I)<(II)) based on the Vuong (V) and Clarke (C) tests

4.6 Conclusions

For count regression data we have presented several models. In order to model over-dispersion we used models with an additional parameter as in the NBR and GPR model or models with an extra proportion of zero observations like the zero-inflated model ZIPR.

Further, in order to account for unobserved spatial heterogeneity in the data we included spatial random effects which allow for spatial correlations between observations and / or spatially varying covariates.

These models were applied to analyze the number of ambulant benefits received per patient in 2007. The DIC, the Vuong and the Clarke tests were used for model comparison. Models allowing for over-dispersion showed a significantly better fit than an ordinary non spatial Poisson regression model. For the NBR and the ZIPR model the inclusion of spatial effects did not improve the model fit. For the Poisson model which does not allow for over-dispersion, and the GPR model, the inclusion of spatial effects led to an improved model fit. According to the considered criteria the GPR model with spatial random CAR effects but no spatial covariates is to be preferred to all other models. However, the fitted spatial model shows no smooth surface structure. Rather it indicates isolated specific regions where the covariates provide no adequate fit.

There are several interesting avenues for further research. For instance, instead of analyzing the number of ambulant benefits received by patient for one year only, it might be interesting to include data over several years in order to examine whether the spatial pattern changes over the years. Another interesting possibility is to extend the regression models by allowing for regression on φ and ω in order to find a better model fit and to address heterogeneity on a more differentiated basis.

Chapter 5

Modeling dependent yearly claim totals including zero-claims in private health insurance

5.1 Introduction

Dependencies in insurance data may occur in many fields. Claim frequencies and sizes are likely to be dependent. A copula approach to this issue applied to car insurance data has been developed by Kastenmeier (2008). Specifically in the field of health insurance, dependencies between inpatient and outpatient treatments are considered by Frees, Gao, and Rosenberg (2007). Pitt, Chan, and Kohn (2006) discuss multi-dimensional measures of health care utilization. They model the dependency between six measures of medical care demand, which are categorized numbers of visits to physicians. Zimmer and Trivedi (2006) use a copula for three simultaneously determined outcomes, i.e., the health insurance status for married couples and their individual health care demand. Dependencies between the number of visits of insured and uninsured persons per year have been considered by Deb, Munkin, and Trivedi (2006). Spatial clustering is investigated by Brezger and Lang (2006) where treatment costs are assumed to be influenced by time, age, sex and spatial effects. A longitudinal model for normalized patient days per year in Wisconsin nursing homes from 1995 through 2001 using copulas was developed by Sun, Frees, and Rosenberg (2008).

The aim of this chapter is to develop a collective model of yearly claim totals capable of reflecting the dependency between different coverage fields. This will allow to appropriately predict which yearly total amount an insurer needs to reserve in order to cover expenditures for an insured person depending on age, sex and other attributes. This is crucial for the pricing of premiums and for risk management. Neglecting dependency between dependent fields may result for example in a misspecification of significant policy characteristics or in false reserving calculus, since diversification effects are neglected. Applications for such a model abound: in operational risk, losses of different dependent types occur very seldom, hence many loss totals are zero. Whenever policies cover different risks claim totals may be zero for some risk and positive for other risks: a car insurance contract may cover vehicle damages of different subclasses or third-party liabilities.

Claim frequency and claim size models are standard tools in non-life actuarial science. For claim frequencies often many zeros are observed caused for example by deductibles. Specifically

in private health insurance there is an additional incentive for excess zeros: the policyholder receives a premium refund by the end of the year when not claiming a single reimbursement. Based on claim frequency and claim size models one can construct models for the yearly total claim which will be a continuous random variable only given that at least one claim occurred. The interest, however, often lies in modeling claims in general. If one allows for claim frequencies of zero, the yearly total claim distribution will have an additional point mass at zero. Using a copula approach to model dependency of the yearly total claims thus requires the use of discrete as well as continuous copula properties. We will develop such an approach and estimate parameters using maximum likelihood.

In this chapter, we also utilize pair-copula constructions (PCC's) of general multivariate distributions. We model multivariate data using a cascade of pair-copulas, acting on two variables at a time. Pair-copula decompositions build on the work on vines of Joe (1996), Bedford and Cooke (2001a), Bedford and Cooke (2001b) and Bedford and Cooke (2002). For high-dimensional distributions there are many possible pair-copula decompositions for the same multivariate distribution. Bedford and Cooke (2001b) introduced a graphical model called regular vine to help organize them. They also identified two important subclasses of regular vines, which they called C- and D-vines. Pair-copula decomposed models also represent a very flexible class of higher-dimensional copulas. While Kurowicka and Cooke (2006) considered nonstandard estimation methods, Aas, Czado, Frigessi, and Bakken (2009) used maximum likelihood for statistical inference and explored the flexibility to model financial time series. There are several advantages of using PCC's: a T-dimensional multivariate density of continuous margins will be expressed as a product of marginal densities and bivariate copulas with individual parameters each. Therefore, in high dimensions T the numerical evaluation of the joint density is very tractable. Each pair of margins can be modeled separately, i.e., the copula class and hence tail dependence properties can be chosen individually. Also, since Archimedean copulas (see e.g. Nelsen (2006, Chapter 4)) are capable only of modeling exchangeable correlation structures, PCC's provide a possibility for generalizing the correlation structure. Finally, model selection in the sense of eliminating weakly correlated copula densities from the joint density can be facilitated.

The chapter is innovative with regard to the following aspects: first of all, we present a novel opportunity for modeling the joint density of total claims including zero claims based on copulas for binary and continuous margins. We illustrate how PCC's can be utilized under marginals. Finally we present a novel approach to choose the copula when the margins are discrete. Our model will allow to model the dependency of large claim portfolios in the presence of zero observations.

This chapter is organized as follows: in Section 5.2 we will give a short review of the concept of copulas and illustrate how multivariate distributions can be constructed using pair-copula constructions. In Section 5.3 an appropriate model for dependent yearly total claims including the zero will be developed: while Subsection 5.3.1 deals with the aggregation to yearly totals, Subsection 5.3.2 addresses the problem of specifying a copula based model dependent for claim totals and zero-claim events. An application to health insurance including a detailed illustration of how to deal with the copula choice problem will be given in Section 5.4. We conclude with a summary and discussion.

5.2 Copulas and multivariate distributions

A J-dimensional copula C_J is a multivariate cdf $C_J : [0,1]^J \to [0,1]$ whose univariate margins are uniform on $[0,1]$, i.e., $C_J(1,\ldots,1,u_j,1,\ldots,1) = u_j \ \forall j \in \{1,\ldots,J\}$. For J continuous random variables (rv) $\boldsymbol{X} := (X_1,\ldots,X_J)'$ with marginal distributions $F_1,\ldots,F_J$ and densities $f_1,\ldots,f_J$, all transformed rv's $U_j := F_j(X_j)$, $j = 1,\ldots,J$ are uniform on $[0,1]$, hence while F_j reflects the marginal distribution of X_j, C_J reflects the dependency. Sklar (1959) shows that

$$F_{\boldsymbol{X}}(x_1,\ldots,x_J) = C_J(F_1(x_1),\ldots,F_J(x_J)|\boldsymbol{\zeta}), \tag{5.2.1}$$

where $\boldsymbol{\zeta}$ are the corresponding copula parameters. If a multivariate cdf of $\boldsymbol{X}$ exists, there also exists a copula C_J which separates the dependency structure from the marginal distributions. If the margins are continuous, C_J is unique. Vice versa, according to (5.2.1) we can construct a multivariate cdf from J marginal distributions using a J-dimensional copula C_J. For a more detailed introduction to copulas, see for instance Joe (1997) or Nelsen (2006). Definitions of some elliptical and Archimedean copulas together with their bivariate densities can be found in Appendix 5.4.5.

While in this chapter we use J dimensional copulas to model dependent discrete margins, a pair-copula construction (PCC) of the joint density will be utilized to describe the dependence of continuous margins. A PCC consists of a cascade of pair-copulas, acting on two variables at a time. In high dimensions there are many different PCC's possible. Bedford and Cooke (2001b) and Bedford and Cooke (2002) show that they can represent such a PCC in a sequence of nested trees with undirected edges, which they call regular vine. One distinguishes between the classes of C and D vines where in the trivariate case these classes coincide. In the following we will illustrate the construction of a C-vine: a multivariate density can be expressed as a product of conditional densities, i.e.

$$f(x_1,...,x_J) = f(x_J|x_1,\cdots,x_{J-1})f(x_1,\cdots,x_{J-1}) = \prod_{j=2}^{J} f(x_j|x_1,\cdots,x_{J-1})\cdot f(x_1). \tag{5.2.2}$$

Here $F(\cdot|\cdot)$ and $f(\cdot|\cdot)$ denote conditional cdf's and densities, respectively. Using Sklar's theorem applied to conditional bivariate densities we can express $f(x_J|x_1,\cdots,x_{J-1})$ as

$$\begin{aligned} f(x_J|x_1,\cdots,x_{J-1}) &= \frac{f(x_{J-1},x_J|x_1,\cdots,x_{J-2})}{f(x_{J-1}|x_1,\cdots,x_{J-2})} \\ &= c_{J-1,J|1,\cdots,J-2}\cdot f(x_J|x_1,\cdots,x_{J-2}). \end{aligned} \tag{5.2.3}$$

Here we use for arbitrary distinct indices $i,j,i_1,\cdots,i_k$ with $i<j$ and $i_1<\cdots<i_k$ the following abbreviation for a bivariate conditional copula density evaluated at conditional cdf's:

$$c_{i,j|i_1,\cdots,i_k} := c_{i,j|i_1,\cdots,i_k}(F(x_i|x_{i_1},\cdots,x_{i_k}),F(x_j|x_{i_1},\cdots,x_{i_k})).$$

Joe (1996) showed that for a d-dimensional vector $\boldsymbol{\nu}$ and a reduced vector $\boldsymbol{\nu}_{-j}$ equal to $\boldsymbol{\nu}$ but without component j the conditional cdf can be obtained recursively by

$$F\left(x|\boldsymbol{\nu}\right) = \frac{\partial C\left(F(x|\boldsymbol{\nu}_{-j}),F(\nu_j|\boldsymbol{\nu}_{-j})\right)}{\partial F(x|\boldsymbol{\nu}_{-j})}. \tag{5.2.4}$$

A detailed proof of this can be found for example in Czado, Min, Baumann, and Dakovic (2009). For the special case where $\boldsymbol{\nu} = \{\nu\}$ it follows that $F\left(x|\nu\right) = \frac{\partial C(F(x),F(\nu))}{\partial F(\nu)}$. For the

uniform margins $U := F(x)$ and $V := F(v)$ we define a function

$$h(u|v) := \frac{\partial C(u,v)}{\partial v}. \tag{5.2.5}$$

This h function has been derived explicitly for many copulas by Aas, Czado, Frigessi, and Bakken (2009). A summary of the ones used in this chapter is given in Table 5.12 in Appendix 5.4.5. By recursive use of (5.2.3) one can express the product of conditional densities (5.2.2) by

$$\begin{aligned} f(x_1,...,x_J) &= f(x_1)\cdot\prod_{j=2}^{J}\prod_{k=1}^{j-1} c_{j-k,j|1,\cdots,j-k-1}\cdot f(x_j) \\ &= \prod_{r=1}^{J} f(x_r)\cdot\prod_{j=2}^{J}\prod_{k=1}^{j-1} c_{j-k,j|1,\cdots,j-k-1}. \end{aligned} \tag{5.2.6}$$

For $k = j-1$ the conditioning set in $c_{j-k,j|1,\cdots,j-k-1}$ is empty, i.e., we set $c_{1,j|10} := c_{1,j}$.

In the trivariate there are only three theoretical decompositions of $f(x_1,x_2,x_3)$ (ignoring the possibility of choosing different bivariate copula classes), whereas in higher dimensions, there are many more. On possible decomposition is obtained by using $f(x_1,x_2,x_3) = f_1(x_1)f_{2|1}(x_2|x_1)f_{3|12}(x_3|x_1,x_2)$. Then the PCC is given by

$$\begin{aligned} f(x_1,x_2,x_3) &= c_{12}(F_1(x_1),F_2(x_2))\; c_{23|1}(F_{2|1}(x_2|x_1),F_{3|1}(x_3|x_1)) \\ &\quad \cdot c_{13}(F_1(x_1),F_3(x_3))\prod_{j=1}^{3} f_j(x_j). \end{aligned} \tag{5.2.7}$$

The joint density of pairs of margins corresponding to the PCC in (5.2.7) can be written as

$$\begin{aligned} f(x_1,x_2) &= \int f_1(x_1)f_{2|1}(x_2|x_1)f_{3|12}(x_3|x_1,x_2)dx_3 = f_1(x_1)f_{2|1}(x_2|x_1) \\ &= c_{12}(F_1(x_1),F_2(x_2))\prod_{j=1}^{2} f_j(x_j), \end{aligned}$$

and similarly $f(x_1,x_3) = c_{13}(F_1(x_1),F_3(x_3))\cdot f_1(x_1)f_3(x_3)$. The final margin requires integration, i.e.

$$\begin{aligned} f(x_2,x_3) &= \int f_1(x_1)f_{2|1}(x_2|x_1)f_{3|12}(x_3|x_1,x_2)dx_1 \\ &= \int c_{12}(F_1(x_1),F_2(x_2))\cdot c_{23|1}(h(F_2(x_2)|F_1(x_1)),h(F_3(x_3)|F_1(x_1))) \\ &\quad \cdot c_{13}(F_1(x_1),F_3(x_3))\prod_{j=1}^{3} f_j(x_j)dx_1 \\ &= \int_0^1 c_{12}(u_1,u_2)\cdot c_{23|1}(h(u_2|u_1),h(u_3|u_1))\cdot c_{13}(u_1,u_3) \\ &\quad \cdot\prod_{j=2}^{3} f_j(F_j^{-1}(u_j))du_1, \end{aligned} \tag{5.2.8}$$

where we substitute $u_j = F_j(x_j)$ and transform $dx_1 = \frac{1}{f_1(x_j)}du_1$. For most copula choices, the integral in (5.2.8) can only be calculated numerically.

5.3 A model for dependent yearly claim totals

In this section we aim to develop a joint model for yearly dependent total claims including zero claims. One possible approach to this has been developed by Frees and Valdez (2008) and has been applied to car accident claims where payments may occur in three different correlated random classes. They point out that this is a nonstandard problem since all three claim types are rarely observed simultaneously. In their approach, the combination of claims and zero claims is modeled by a multinomial logit model. We will model yearly total claims for a certain claim type and utilize a copula to obtain a joint model. In general, we assume that we have J dependent yearly claims available. For a population of insured individuals this may be J different treatment fields. A zero claim may arise for different reasons: first of all, a healthy patient simply had no need to see a physician. Secondly, the invoice may be below a deductible. Thirdly, the insured person will get a premium refund when not recovering a single bill throughout the year and opts for this when expecting the refund to be higher than the invoice. In health insurance we consider the dependence between ambulant, inpatient and dental treatments. Zero claim events will certainly be dependent due to the health status of an insured person. The deductible will not have an impact on the dependency of the zero claim events since they apply separately for the three fields. The premium refund, however, will only be paid if no reimbursement is claimed in either of these fields. Therefore, it will also influence the dependence.

5.3.1 Aggregation of claim frequencies and sizes to yearly totals

We will express the yearly total claim T_j in field j as

$$T_j := W_j \cdot 0 + (1 - W_j) N_j \bar{S}_j = (1 - W_j) T_j^+ \geq 0.$$

Here W_j is a binary indicator for the zero claim event, i.e., $W_j = 1$ if the claim is zero and $W_j = 0$ else. Also, $\boldsymbol{W} := (W_1, \ldots, W_T)'$ and $\boldsymbol{W}_{-j} := (W_1, \ldots, W_{j-1}, W_{j+1}, \ldots, W_T)'$. Further, N_j is the positive number of claims and $\bar{S}_j$ the average claim size (also strictly positive). In general we will observe only $\bar{S}_j|\{W_j = 0\}$ but we assume that $\bar{S}_j|\{W_j = 1\}$ exists. Let the total claim $T_j^+ := N_j \bar{S}_j > 0$ if $W_j = 0$ and unknown but positive if $W_j = 1$.

A more general case is given when the single claims S_{kj}, $k = 1, \ldots, N_j$, which contribute to the yearly total, are known and are i.i.d., i.e., if $T_j^+ := \sum_{k=1}^{N_j} S_{kj}$. Then the distribution of T_j^+ will be obtained by convolution and can be approximated for example using the methods summarized in the R package 'actuar' (see Dutang, Goulet, and Pigeon (2008)). This chapter, however, will focus on average claims but apart from the expression for total claims all approaches carried out in this chapter apply in a similar way.

We denote probability mass functions (pmf) by p and their cumulative distribution functions (cdf) by P. Probability density functions (pdf) and cdf of continuous random variables are denoted by f and F, respectively. The following distributions for the zero claim event, claim frequencies and sizes are assumed:

$$\begin{aligned}
W_j &\sim \textit{binary}(p_{W_j}(1)), \text{ with pmf } p_{W_j}, \text{ cdf } P_{W_j}, \\
N_j|\{W_j = 0, \boldsymbol{W}_{-j}\} &\sim P_{N_j|\{W_j=0,\boldsymbol{W}_{-j}\}}, \text{ positive no. of claims, pmf } p_{N_j|\{W_j=0,\boldsymbol{W}_{-j}\}}, \\
\bar{S}_j|\{N_j, W_j = 0, \boldsymbol{W}_{-j}\} &\sim F_{\bar{S}_j|\{N_j,W_j=0,\boldsymbol{W}_{-j}\}}, \text{ average claim size, pdf } f_{\bar{S}_j|\{N_j,W_j=0,\boldsymbol{W}_{-j}\}}.
\end{aligned}$$

For data following these distributions regression models may be fitted. Then the realizations $\boldsymbol{w}_{-j}$ of $\boldsymbol{W}_{-j}$ are used as regressors in the latter two models and additionally the realizations

n_j of N_j in the last one. It follows that $T_j^+ := N_j \bar{S}_j$ is conditionally independent of $\boldsymbol{W}_{-j}$. Note that one only fits these two regression models to data with $W_j = 0$. The distribution of N_j may be modeled using a zero-truncated count distribution (see for example Zuur, Leno, Walker, Saveliev, and Smith (2009, Chapter 11)) which can be constructed based on any count distribution. For example, let N_c follow some count distribution (Poisson, Negative-Binomial etc.) with pmf p_c, then $N \sim P_N$ with pmf

$$p_N(n) := \frac{p_c(n)}{1 - p_c(0)},\ n = 1, 2, \ldots$$

will be the zero-truncated representative of this count distribution.

Example 1. For the Negative Binomial (NB) distribution with mean parameter $\mu > 0$, shape parameter $r > 0$ and variance $\mu\left(1 + \frac{\mu}{r}\right)$ a zero-truncated Negative Binomial (ZTNB) distribution has the pmf

$$p_N(n) \quad := \quad \frac{1}{1 - \left(\frac{r}{\mu+r}\right)^r} \cdot \frac{\Gamma(n+r)}{\Gamma(r)\Gamma(n!)} \cdot \left(\frac{r}{\mu+r}\right)^r \cdot \left(\frac{\mu}{\mu+r}\right)^n,\ n = 1, 2, \ldots.$$

We can now model and quantify the dependency of the vectors $\boldsymbol{W}$ and $\boldsymbol{T}$. On the one hand, the number of zero claims $\boldsymbol{W}$ reflects the impact of the health insurer's incentives for not having a single claim throughout the year, which the insurer wants to know about in order to arrange its deductibles and premium refund policy. On the other hand, the dependence of $\boldsymbol{T}$ can be used for premium and risk capital calculation.

Let F_{T_j} the cdf of T_j, then a derivative at $t = 0$ does not exist and therefore only the derivative conditional on $\{W_j = 0\}$ may be called a density function. Similar to Heller, Stasinopoulos, Rigby, and De Jong (2007) we simply refer to f_{T_j} as the probability function (pf) of T_j. Jørgensen and de Souza (1994) and Smyth and Jørgensen (2002) consider models for continuous claim sizes including zero claims. These are based on the class of Tweedie distributions (Tweedie (1984)), which are members of the exponential family. In particular they use a compound Poisson-gamma distribution which is contained in the class of the Tweedie distributions. Belasco and Ghosh (2008) develop a model based on the Tobit model (Tobin (1958)) in which a zero outcome arises from left-censoring. The marginal distributions of T_j^+ given N_j, $W_j = 0$ and $\boldsymbol{W}_{-j}$ will be denoted by

$$T_j^+|\{N_j, W_j = 0, \boldsymbol{W}_{-j}\} \quad \sim \quad F_{T_j^+|\{N_j, W_j=0, \boldsymbol{W}_{-j}\}},\ \text{with pdf } f_{T_j^+|\{N_j, W_j=0, \boldsymbol{W}_{-j}\}}.$$

For the moment, we drop the index j for field j and also the dependency on $\boldsymbol{W}_{-j}$.

Lemma 1. For average claims $\bar{S}$, cdf and pdf of $T^+|\{W = 0\}$ are given by

$$\begin{aligned} F_{T^+|\{W=0\}}(t^+) &= \sum_{k=1}^{\infty} F_{\bar{S}|\{N,W=0\}}\left(\frac{t^+}{k}|\{N = k\}\right) p_{N|\{W=0\}}(k) \\ f_{T^+|\{W=0\}}(t^+) &= \sum_{k=1}^{\infty} f_{\bar{S}|\{N,W=0\}}\left(\frac{t^+}{k}|\{N = k\}\right) p_{N|\{W=0\}}(k). \end{aligned}$$

Proof. See Appendix. □

Lemma 2. The cdf of the yearly total claim T at t is

$$F_T(t) = p_W(0) + \mathbb{1}_{\{t>0\}}(1 - p_W(0))F_{T^+|\{W=0\}}(t). \tag{5.3.1}$$

Proof. See Appendix. □

5.3.2 A joint distribution of yearly total claims based on copulas

In this section we develop a joint distribution of $\boldsymbol{T} := (T_1, \ldots, T_J)'$. Utilizing copulas in order to model dependency between $\boldsymbol{T} := (T_1, \ldots, T_J)'$ is nonstandard since according to Lemma 2, $U_j := F_{T_j}(T_j)$ will have a point mass at 0 and hence U_j will not be uniform on $[0, 1]$. Nevertheless, we will develop a joint distribution of T_j, $j = 1, \ldots, J$, based on two copula constructions, one with discrete margins $\boldsymbol{W} := (W_1, \ldots, W_J)'$ and one with continuous margins $\boldsymbol{T}^+ := (T_1^+, \ldots, T_J^+)'$. We allow $\boldsymbol{W}$ and $\boldsymbol{T}^+$ to be dependent random vectors and use the conditional independence between $\boldsymbol{W}$ and $\boldsymbol{T}|\boldsymbol{W}$, i.e., we use

$$P(T_j \leq t_j, W_j = w_j, \forall j) \;=\; P((1 - W_j)T_j^+ \leq t_j | \{W_j = w_j\}, \forall j) \cdot P(W_j = w_j, \forall j).$$

Here $p_{\boldsymbol{W}} := P(W_j = w_j, \forall j = 1, \ldots, J)$ can be obtained by constructing $P_{\boldsymbol{W}}$ by a J dimensional copula and using the formula of Song (2007)[p. 128] to obtain the joint pmf. For $\boldsymbol{T}^+$ a joint pdf $f_{\boldsymbol{T}^+|\boldsymbol{W}}$ and hence a joint cdf $F_{\boldsymbol{T}^+|\boldsymbol{W}}$ may be constructed using a PCC. We stress that the PCC is utilized for $\boldsymbol{T}^+$, which is unobserved for some observations of t_j^+ but nevertheless we use the conditional independence of $\boldsymbol{W}$ and $\boldsymbol{T}|\boldsymbol{W}$. The joint distribution of yearly total claims $\boldsymbol{T}$ and zero claim events $\boldsymbol{W}$ will be given in Proposition 2.

Proposition 2. Let $T_j = (1 - W_j) \cdot T_j^+$ and $J^0(\boldsymbol{w}) := \big\{j \in \{1, \ldots, J\} : W_j = 0\big\} = \{j_{1(\boldsymbol{w})}, \ldots, j_{n(\boldsymbol{w})}\}$ with $n(\boldsymbol{w})$ the cardinality of $J^0(\boldsymbol{w})$. Then the joint probability function of $\boldsymbol{T}$ and $\boldsymbol{W}$ is given by

$$f_{\boldsymbol{T},\boldsymbol{W}}(\boldsymbol{t}, \boldsymbol{w}) \;=\; p_{\boldsymbol{W}}(\boldsymbol{w}) f_{T^+_{1(\boldsymbol{w})}, \ldots, T^+_{n(\boldsymbol{w})}|\boldsymbol{W}}(t^+_{1(\boldsymbol{w})}, \ldots, t^+_{n(\boldsymbol{w})}|\boldsymbol{w})$$

where $f_{T^+_{1(\boldsymbol{w})}, \ldots, T^+_{n(\boldsymbol{w})}|\boldsymbol{W}}(t^+_{1(\boldsymbol{w})}, \ldots, t^+_{n(\boldsymbol{w})}|\boldsymbol{w})$ is the joint pdf of $\boldsymbol{T}^+$ where all margins with $W_j = 1$, $j = 1, \ldots, J$ are integrated out.

Proof. We consider

$$\begin{aligned} P(T_j \leq t_j, j = 1, \ldots, J, \boldsymbol{W} = \boldsymbol{w}) &= P((1 - W_j)T_j^+ \leq t_j, \forall j | \boldsymbol{W} = \boldsymbol{w}) \cdot p_{\boldsymbol{W}}(\boldsymbol{w}) \\ &= P(T_{j_k}^+ \leq t_{j_k}, j_k \in J^0(\boldsymbol{w}) | \boldsymbol{W} = \boldsymbol{w}) \cdot p_{\boldsymbol{W}}(\boldsymbol{w}). \end{aligned}$$

The joint probability function is obtained by deriving for t_{j_k}, $j_k \in J^0(\boldsymbol{w})$, i.e.,
$f_{T^+_{1(\boldsymbol{w})}, \ldots, T^+_{n(\boldsymbol{w})}|\boldsymbol{W}}(t^+_{1(\boldsymbol{w})}, \ldots, t^+_{n(\boldsymbol{w})}|\boldsymbol{w}) = \frac{\partial}{\partial t_{1(\boldsymbol{w})}} \cdots \frac{\partial}{\partial t_{n(\boldsymbol{w})}} P(T_{j_k}^+ \leq t_{j_k}, j_k \in J^0(\boldsymbol{w}) | \boldsymbol{W} = \boldsymbol{w})$. □

So whenever an observation $T_j^+ | \{W_j = 1\}$ is unknown, the margin in the corresponding PCC is integrated out. Hence the distribution of the vector $\boldsymbol{T}|\boldsymbol{W}$ is defined for strictly positive numbers.

Example 2. For $J = 3$,

$$\begin{aligned} f_{\boldsymbol{T},\boldsymbol{W}}(t_1, t_2, t_3, w_1, w_2, w_3) \;=\; & p_{\boldsymbol{W}}(w_1, w_2, w_3) \cdot \Big[\mathbb{1}_{\{\boldsymbol{w}=(1,1,1)'\}} + \mathbb{1}_{\{\boldsymbol{w}=(1,1,0)'\}} f_{T_3^+}(t_3) \\ & + \mathbb{1}_{\{\boldsymbol{w}=(1,0,1)'\}} f_{T_2^+}(t_2) + \mathbb{1}_{\{\boldsymbol{w}=(0,1,1)'\}} f_{T_1^+}(t_1) \\ & + \mathbb{1}_{\{\boldsymbol{w}=(1,0,0)'\}} f_{T_2^+, T_3^+}(t_2, t_3) + \mathbb{1}_{\{\boldsymbol{w}=(0,1,0)'\}} f_{T_1^+, T_3^+}(t_1, t_3) \\ & + \mathbb{1}_{\{\boldsymbol{w}=(0,0,1)'\}} f_{T_1^+, T_2^+}(t_1, t_2) + \mathbb{1}_{\{\boldsymbol{w}=(0,0,0)'\}} f_{\boldsymbol{T}^+}(t_1, t_2, t_3)\Big]. \end{aligned}$$

We define $P_{\boldsymbol{W}}(w_1,\ldots,w_J|\boldsymbol{\zeta}^W) := C_J(P_{W_1}(w_1),\ldots,P_{W_J}(w_J)|\boldsymbol{\zeta}^W)$ by a copula cdf C_J in dimension J with copula parameters $\boldsymbol{\zeta}^W$. For binary margins,

$$\begin{aligned} p_{\boldsymbol{W}}(w_1,\ldots,w_J|\boldsymbol{\zeta}^W) &= \sum_{j_1=0}^{w_1}\cdots\sum_{j_J=0}^{w_J}(-1)^{\sum_{k=1}^J(j_k+w_k)} \\ &\quad \cdot C_J(P_{W_1}(j_1),\ldots,P_{W_J}(j_J)|\boldsymbol{\zeta}^W). \end{aligned} \tag{5.3.2}$$

Proof. According to Song (2007)[p. 128]

$$p_{\boldsymbol{W}}(w_1,\ldots,w_J|\boldsymbol{\zeta}^W) = \sum_{k_1=0}^{1}\cdots\sum_{k_J=0}^{1}(-1)^{k_1+\ldots+k_J}C_J\big(u_{1k_1}(w_1),\ldots,u_{Jk_J}(w_J)|\boldsymbol{\zeta}^W\big),$$

where $u_{t0}(w_t) := P_{W_t}(w_t)$ and $u_{t1}(w_t) := P_{W_t}(w_t-1)$. Now $u_{t0}(1) = P_{W_t}(1) = 1$, $u_{t1}(1) = P_{W_t}(0)$ and $u_{t0}(0) = P_{W_t}(0)$. Since $u_{t1}(0) = P_{W_t}(-1) = 0$ and $C_J(\ldots,0,\ldots|\boldsymbol{\zeta}^W) = 0$ we only need to consider $k_t \le w_t$. By transforming $j_t := w_t - k_t \ge 0$ we obtain the required result. □

Note that $P_{W_j}(1) = 1, j = 1,\ldots,J$. In this case and given we use an elliptical or Archimedean copula, the copula in (5.3.2) at such a marginal probability will be of the same class only with decreased dimension. On the other hand, for the continuous random vector $\boldsymbol{T}^+$ we define $F_{\boldsymbol{T}^+}(t_1^+,\ldots,t_J^+|\boldsymbol{\zeta}^{T^+})$ by a PCC introduced in (5.2.7).

We return to the trivariate case ($J = 3$). The bivariate marginal distributions are defined according to (5.2.8), where $f_{T_2^+,T_3^+}(t_{i2}^+,t_{i3}^+)$ requires numerical integration. Let $\boldsymbol{\zeta}^W$ the parameters of the copula of $\boldsymbol{W}$ and $\boldsymbol{\zeta}^{T^+} := (\boldsymbol{\zeta}_{12}^{T^+},\boldsymbol{\zeta}_{13}^{T^+},\boldsymbol{\zeta}_{23|1}^{T^+})'$ the parameters of the PCC of $\boldsymbol{T}^+$. Since the expression depending on $\boldsymbol{\zeta}^W$ is independent of the expression depending on $\boldsymbol{\zeta}^{T^+}$, i.e., for $\boldsymbol{\zeta} := (\boldsymbol{\zeta}^W,\boldsymbol{\zeta}^{T^+})'$, the log-likelihood is $l(\boldsymbol{\zeta}) = l(\boldsymbol{\zeta}^W) + l(\boldsymbol{\zeta}^{T^+})$, which in a maximum likelihood context can be fitted separately over those two parameter sets. For observations $i = 1,\ldots,I$,

$$\begin{aligned} l(\boldsymbol{\zeta}^W) &= \sum_{i=1}^I \log(p_{\boldsymbol{W}}(w_{i1},w_{i2},w_{i3})) \\ &= \sum_{i=1}^I \log\left(\sum_{j_1=0}^{w_{i1}}\sum_{j_2=0}^{w_{i2}}\sum_{j_3=0}^{w_{i3}}(-1)^{\sum_{k=1}^3(j_k+w_{ik})}C_J(P_{W_1}(j_1),P_{W_2}(j_2),P_{W_3}(j_3)|\boldsymbol{\zeta}^W)\right), \\ l(\boldsymbol{\zeta}^{T^+}) &= \sum_{i=1}^I\Big[\mathbb{1}_{\{w_{i1}=1,w_{i2}=0,w_3=0\}}\cdot\log(f_{T_2^+,T_3^+}(t_{i2}^+,t_{i3}^+|\boldsymbol{\zeta}^{T^+})) \\ &\quad +\mathbb{1}_{\{w_{i1}=0,w_{i2}=1,w_{i3}=0\}}\cdot\log(c_{13}(F_{T_1^+}(t_{i1}^+),F_{T_3^+}(t_{i3}^+)|\boldsymbol{\zeta}_{13}^{T^+})) \\ &\quad +\mathbb{1}_{\{w_{i1}=0,w_{i2}=0,w_{i3}=1\}}\cdot\log(c_{12}(F_{T_1^+}(t_{i1}^+),F_{T_2^+}(t_{i2}^+)|\boldsymbol{\zeta}_{12}^{T^+})) \\ &\quad +\mathbb{1}_{\{w_{i1}=0,w_{i2}=0,w_{i3}=0\}}\cdot\Big[\log(c_{12}(F_{T_1^+}(t_{i1}^+),F_{T_2^+}(t_{i2}^+)|\boldsymbol{\zeta}_{12}^{T^+})) \\ &\quad +\log(c_{23|1}(h(F_{T_2^+}(t_{i2}^+)|F_{T_1^+}(t_{i1}^+),\boldsymbol{\zeta}_{12}^{T^+}),h(F_{T_3^+}(t_{i3}^+)|F_{T_1^+}(t_{i1}^+),\boldsymbol{\zeta}_{13}^{T^+})|\boldsymbol{\zeta}_{23|1}^{T^+})) \\ &\quad +\log(c_{13}(F_{T_1^+}(t_{i1}^+),F_{T_3^+}(t_{i3}^+)|\boldsymbol{\zeta}_{13}^{T^+}))\Big]\Big] + \text{ const. independent of } \boldsymbol{\zeta}^{T^+}. \end{aligned}$$

5.4 Application to health insurance data

We will consider data from a German private health insurer. Each record represents one out of 37 819 insured persons. Claim frequencies will be the number of benefits received by an

Variable	Description
Responses	
W_{ijt}	Zero claim event (1 if zero claim) by patient i in treatment field j and year t.
$N_{ijt}\|\{W_{ijt}=0\}$	Total positive number of benefits received by patient i in treatment field j and year t.
T_{ijt}	Total invoice for patient i in treatment field j and year t (including deductibles).
$\bar{S}_{ijt}\|\{W_{ijt}=0\}$	T_{ijt}/N_{ijt}, average invoice of patient i in treatment field j and year t.
Covariates	
	CATEGORICAL
SEX_i	Dummy for gender of patient i.
	DISCRETE
AGE_{it}	Age of patient i at December, 31 in year t.
	CONTINUOUS
DED_{ijt}	Total of all deductibles of $\bar{S}_{ijt}$ of patient i in treatment field j and year t.
$\overline{DED}_{ijt}$	Average deductible of patient i in treatment field j and year t
	SPATIAL
ZIP_i	ZIP code of the home address of patient i as of Dec. 31, 2007.
$D(i)$	Dummy for home district of patient i as of Dec. 31, 2007. There are 439 German districts. Individuals are spread over all districts.
	CONTINUOUS WITH SPATIAL INFORMATION
$PHYS.INH_{D(i)}$	(number of physicians in district $D(i)$ listed in the yellow pages as of April 15, 2008 divided by the number of inhabitants in district $D(i)$ in 2007) $\cdot 100$.
$URBAN_{D(i)}$	Number of inhabitants per square kilometer in district $D(i)$ in 2007
$BP_{D(i)}$	Average buying power in Euro in 2007 in district $D(i)$ on a scale of nine scoring levels. Buying power has been determined as the average net income per district + public transfer payments.

Table 5.1: Description of variables considered for claim frequencies and claim size models for the health insurance data

insured person, where a benefit may be any treatment or prescription balanced to a patient, i.e., a patient usually gets charged for several benefits during one visit. Claim sizes will be the average invoice, i.e., the yearly total amount divided by the number of benefits. Responses as well as explanatory variables have been observed in the ambulant (i.e., outpatient), inpatient and dental field over three years from 2005 to 2007. We will abbreviate the treatment fields by 'A' for ambulant, 'I' for inpatient and 'D' for dental or indices $j = 1, \ldots, 3$, respectively. Around 76% of the insured persons are male, which is typical for the policy line considered. All policyholders in the population are covered in all three fields. The private German health care system allows for deductibles, which - depending on policy type and treatment - may be a specific amount for a certain benefit or a percentage of the amount invoiced. Policyholders not handing in a single bill for a whole year in any of the three fields will get a premium refund. Therefore, we might not see the actual treatment numbers and amounts invoiced in the data. A variable description including responses and explanatory variables will be given in Table 5.1.

The data has been supplemented by data from different sources:

- a mapping of ZIP codes to 439 districts not including corporate ZIP codes (*http://www.manfrin-it.com/postleitzahlen/plz.html*), completed by single queries for missing ZIP codes from *http://w3logistics.com/infopool/plz/index.php*,
- number of physicians per ZIP code listed in the yellow pages from 8233 automated web requests searching for 'Arzt' (physician) to *http://web2.cylex.de*,
- number of inhabitants and area in square kilometers of each of the 439 German districts according to the GfK GeoMarketing GmbH (*http://www.gfk-geomarketing.de/marktdaten/samples.php*),
- data transcribed from a map displaying the buying power per district by the GfK GeoMarketing GmbH *http://www.gfk-geomarketing.de/presse/bdm/html/01_2007.html*, reference *Grafik: GfK GeoMarketing.*

Fitting marginal distributions first and fitting the copula parameters for fixed margins afterwards is known as inference functions for margins or the IFM method (see for example Joe (1997, Section 10.1)). In the following subsections we will briefly summarize the regression models chosen for W_{jt}, N_{jt} and $\bar{S}_{jt}$, $j, t = 1, 2, 3$.

5.4.1 Marginal zero claim event models

Consider a logistic regression model for W_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, J$, i.e.

$$W_{ij} \sim binary\left(\frac{\exp(\boldsymbol{x}_{ij}^{Wt}\boldsymbol{\beta}_j)}{1+\exp(\boldsymbol{x}_{ij}^{Wt}\boldsymbol{\beta}_j)}\right).$$

We choose variables by backward selection based on the Wald test with a 5% significance level. The model equations of the reduced designs are given in Table 5.2. An exemplary summary of the regression model for W_{A7} is given in Table 5.3.

Model equations
$W_{A5} \sim 1 + \mathbb{1}_{DED_{A5}\leq 100} + \mathbb{1}_{AGE_5\leq 32} \cdot AGE_5 + \mathbb{1}_{AGE_5>32} \cdot AGE_5 + SEX + BP$
$W_{A6} \sim 1 + \mathbb{1}_{DED_{A6}\leq 100} + \mathbb{1}_{AGE_6\leq 32} \cdot AGE_6 + \mathbb{1}_{AGE_6>32} \cdot AGE_6 + SEX + BP$
$W_{A7} \sim 1 + \mathbb{1}_{DED_{A7}\leq 100} + \mathbb{1}_{AGE_7\leq 32} \cdot AGE_7 + \mathbb{1}_{AGE_7>32} \cdot AGE_7 + SEX + BP$
$W_{I5} \sim 1 + \mathbb{1}_{DED_{I5}=0} + \mathbb{1}_{AGE_5\leq 32} \cdot AGE_5 + \mathbb{1}_{AGE_5>32} \cdot AGE_5 + SEX$
$W_{I6} \sim 1 + \mathbb{1}_{DED_{I6}=0} + \mathbb{1}_{AGE_6\leq 32} \cdot AGE_6 + \mathbb{1}_{AGE_6>32} \cdot AGE_6 + SEX$
$W_{I7} \sim 1 + \mathbb{1}_{DED_{I7}=0} + \mathbb{1}_{AGE_7\leq 32} \cdot AGE_7 + \mathbb{1}_{AGE_7>32} \cdot AGE_7 + SEX + PHYS.INH$
$W_{D5} \sim 1 + \mathbb{1}_{DED_{D5}=0} + \mathbb{1}_{AGE_5>32} \cdot AGE_5 + SEX + PHSY.INH + BP$
$W_{D6} \sim 1 + \mathbb{1}_{DED_{D6}=0} + \mathbb{1}_{AGE_6\leq 32} \cdot AGE_6 + \mathbb{1}_{AGE_6>32} \cdot AGE_6 + SEX + BP$
$W_{D7} \sim 1 + \mathbb{1}_{DED_{D7}=0} + \mathbb{1}_{AGE_7\leq 32} \cdot AGE_7 + \mathbb{1}_{AGE_7>32} \cdot AGE_7 + SEX + PHYS.INH + BP$

Table 5.2: Reduced model equations for each of the nine logistic regression models for W_{jt}, $j = 1, 2, 3 = A, I, D$; $t = 5, 6, 7$ after applying sequential backward selection based on the Wald test

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.4914	0.9951	-9.54	$< 2 \cdot 10^{-16}$
$\mathbb{1}_{DED_{A7} \leq 100}$	10.9182	0.9945	10.98	$< 2 \cdot 10^{-16}$
$\mathbb{1}_{AGE_7 \leq 32} \cdot AGE_7$	0.3480	0.0284	12.26	$< 2 \cdot 10^{-16}$
$\mathbb{1}_{AGE_7 > 32} \cdot AGE_7$	-0.6850	0.0354	-19.35	$< 2 \cdot 10^{-16}$
SEX	0.1965	0.0425	4.63	$3.7 \cdot 10^{-6}$
BP	-0.0795	0.0169	-4.71	$2.5 \cdot 10^{-6}$

Table 5.3: Model summary for the reduced logistic regression model of W_{A7}.

5.4.2 Marginal claim frequency models

Let N_{ij}, $i \in \mathcal{I}_j := \{i = 1, \ldots, n, W_{ij} = 0\}$, $j = 1, \ldots, J$ follow the zero-truncated negative binomial distribution (ZTNB) defined in Example 1. Further let $\boldsymbol{w}_{i(-j)} := (w_{i1}, \ldots, w_{i(j-1)}, w_{i(j+1)}, \ldots, w_{iJ})'$. Then a ZTNB regression model (see e.g. Cruyff and van der Heijden (2008)) is given by

$$\begin{aligned} N_{ij} | \{\boldsymbol{x}_{ij}^N, \boldsymbol{w}_{i(-j)}\} &\sim ZTNB\left(\mu_{ij}(\boldsymbol{x}_{ij}^N, \boldsymbol{w}_{i(-j)}), r_j\right), \\ \mu_{ij}(\boldsymbol{x}_{ij}^N, \boldsymbol{w}_{i(-j)}) &= \exp(\boldsymbol{x}_{ij}^{Nt} \boldsymbol{\gamma}_j^1 + \boldsymbol{w}_{i(-j)} \boldsymbol{\gamma}_j^2). \end{aligned}$$

We utilize the Wald test for backward selection. Thereby we use the observed Fisher information

Model equations
$N_{A5} \sim 1 + DED_{A5} + poly(AGE_5)[,1] + poly(AGE_5)[,2] + SEX + DED_{A5} : poly(AGE_5)[,1] + DED_{A5} : poly(AGE_5)[,2] + SEX : poly(AGE_5)[,1] + SEX : poly(AGE_5)[,2] + W_{I5} + W_{D5}$ $N_{A6} \sim 1 + DED_{A6} + poly(AGE_6)[,1] + poly(AGE_6)[,2] + SEX + URBAN + DED_{A6} : poly(AGE_6)[,1] + DED_{A6} : poly(AGE_6)[,2] + SEX : poly(AGE_6)[,1] + SEX : poly(AGE_6)[,2] + W_{I6} + W_{D6}$ $N_{A7} \sim 1 + DED_{A7} + poly(AGE_7)[,1] + poly(AGE_7)[,2] + SEX + URBAN + BP + DED_{A7} : poly(AGE_7)[,1] + DED_{A7} : poly(AGE_7)[,2] + SEX : poly(AGE_7)[,1] + SEX : poly(AGE_7)[,2] + URBAN : BP + W_{I7} + W_{D7}$
$N_{I5} \sim DED_{I5} + AGE_5 + SEX + URBAN + BP + SEX : URBAN + W_{A5} + W_{D5}$ $N_{I6} \sim 1 + DED_{I6} + AGE_6 + BP + W_{A6}$ $N_{I7} \sim DED_{I7} + AGE_7 + BP + DED_{I7} : AGE_7 + W_{A7}$
$N_{D5} \sim 1 + \log(DED_{D5}) + poly(AGE_5)[,1] + poly(AGE_5)[,2] + SEX + URBAN + \log(DED_{D5}) : poly(AGE_5)[,1] + \log(DED_{D5}) : poly(AGE_5)[,2] + SEX : poly(AGE_5)[,2] + poly(AGE_5)[,1] : URBAN + W_{A5} + W_{I5}$ $N_{D6} \sim 1 + \log(DED_{D6}) + poly(AGE_6)[,1] + poly(AGE_6)[,2] + SEX + \log(DED_{D6}) : poly(AGE_6)[,1] + \log(DED_{D6}) : poly(AGE_6)[,2] + SEX : poly(AGE_6)[,1] + SEX : poly(AGE_6)[,2] + W_{A6} + W_{I6}$ $N_{D7} \sim 1 + \log(DED_{D7}) + poly(AGE_7)[,1] + poly(AGE_7)[,2] + SEX + URBAN + BP + log(DED_{D7}) : poly(AGE_7)[,1] + \log(DED_{D7}) : poly(AGE_7)[,2] + SEX : poly(AGE_7)[,1] + SEX : poly(AGE_7)[,2] + poly(AGE_7)[,1] : URBAN + W_{A7}$

Table 5.4: Reduced model equations for each of the nine ZTNB claim frequency models after applying sequential backward selection based on the Wald test

based on the numerical Hessian matrix obtained by the *R* routine *optim*. The reduced model

equations are given in Table 5.4. For N_{A7}, a summary of the reduced design is given in Table 5.5.

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
Intercept	2.216	0.015	151.123	$< 2 \cdot 10^{-16}$
DED_{A7}	0.469	0.006	72.379	$< 2 \cdot 10^{-16}$
$poly(AGE_7)[,1]$	52.296	1.705	30.672	$< 2 \cdot 10^{-16}$
$poly(AGE_7)[,2]$	9.455	1.908	4.955	$7.2 \cdot 10^{-7}$
SEX	−0.286	0.012	−23.794	$< 2 \cdot 10^{-16}$
$URBAN$	0.007	0.005	1.307	0.191
BP	0.003	0.005	0.641	0.521
$DED_{A7} : poly(AGE_7)[,1]$	−8.003	1.148	−6.972	$3.1 \cdot 10^{-12}$
$DED_{A7} : poly(AGE_7)[,2]$	−6.585	1.202	−5.477	$4.3 \cdot 10^{-8}$
$SEX : poly(AGE_7)[,1]$	−14.489	1.887	−7.680	$1.6 \cdot 10^{-14}$
$SEX : poly(AGE_7)[,2]$	21.686	1.992	10.884	$< 2 \cdot 10^{-16}$
$URBAN : BP$	−0.009	0.004	−2.359	0.018
W_{I7}	0.618	0.014	45.614	$< 2 \cdot 10^{-16}$
W_{D7}	0.237	0.011	21.081	$< 2 \cdot 10^{-16}$

Table 5.5: Model summary for the reduced ZTNB regression model of the claim frequencies for treatment field ambulant in 2007 with dispersion parameter θ is estimated to be 2.15.

5.4.3 Marginal claim size models

As marginal models for the claim sizes $\bar{S}_{ij}$, $i \in \mathcal{I}_j := \{i = 1, \ldots, n, W_{ij} = 0\}$, $j = 1, \ldots, J$ we aim to use weighted log normal models given by

$$\bar{S}_{ij} | \{\boldsymbol{x}_{ij}^{\bar{S}}, \boldsymbol{w}_{i(-j)}, n_j\} \quad \sim \quad lognormal(\boldsymbol{x}_{ij}^{\bar{S}t}\boldsymbol{\alpha}_j^1 + \boldsymbol{w}_{i(-j)}\boldsymbol{\alpha}_j^2 + n_j\alpha_j^3, \sigma_j, \text{ weights } \omega_{ij}^{\bar{S}}). \quad (5.4.1)$$

Since we model average claims rather than actual claim sizes we observe high heteroscedasticy in $\bar{S}_{ij}$ which will depend on the number of claims per year for each observation. As for the logarithmic transformation of the responses in the linear model the exact theoretical influence of N on the heteroscedasticy cannot be determined. We perform a three step approach based on ordinary least square (OLS) regression and weighted least square (WLS) regression (DeMaris (2004, p.201)) with unknown weights. First we fit a log normal OLS regression model based on $\{\boldsymbol{x}_{ij}^{\bar{S}}, \boldsymbol{w}_{i(-j)}, n_j\}$. Now we want to allow for heteroscedasticy using a WLS approach. In order to determine weights $\omega_{ij}^{\bar{S}}$, we regress the OLS squared residuals (as responses) on the model's predictors in another lognormal OLS regression model and use fitted values from this run as variance estimates (see DeMaris (2004, p.205)). In a third step, we replace the OLS model from the first step by the weighted regression (5.4.1). Variable selection is carried out using backward selection based on the Wald test. For every design which we consider new weights are determined, i.e., we update the estimated coefficients in order to predict variances. The choice of regressors for determining the weights, however, is not changed throughout the backward selection procedure. The model equations of the reduced fitted models are given in Table 5.6. A model summary for $\bar{S}_{A7}$ is given in Table 5.7.

5.4.4 Results of fitting copulas to the binary and continuous margins

We model the dependency between the three treatment fields ambulant, inpatient and dental. The years 2005 to 2007 will be investigated separately.

Model equations
$\bar{S}_{A5} \sim 1 + \overline{DED}_{A5} + poly(AGE_5)[,1] + poly(AGE_5)[,2] + SEX + URBAN + BP + \overline{DED}_{A5} : SEX + URBAN : BP + W_{I5} + W_{D5} + N_{A5}$ $\bar{S}_{A6} \sim 1 + \overline{DED}_{A6} + poly(AGE_6)[,1] + poly(AGE_6)[,2] + SEX + PHYS.INH + URBAN + BP + URBAN : BP + W_{I6} + W_{D6} + N_{A6}$ $\bar{S}_{A7} \sim 1 + \overline{DED}_{A7} + poly(AGE_7)[,1] + poly(AGE_7)[,2] + SEX + PHYS.INH + URBAN + BP + \overline{DED}_{A7} : SEX + PHYS.INH : URBAN + W_{I7} + W_{D7} + N_{A7}$
$\bar{S}_{I5} \sim 1 + poly(\overline{DED}_{I5})[,1] + poly(\overline{DED}_{I5})[,2] + poly(AGE_5)[,1] + poly(AGE_5)[,2] + BP + W_{A5}$ $\bar{S}_{I6} \sim 1 + poly(\overline{DED}_{I6})[,1] + poly(\overline{DED}_{I6})[,2] + poly(AGE_6)[,1] + SEX + BP + SEX : BP + W_{A6} + N_{I6}$ $\bar{S}_{I7} \sim 1 + poly(\overline{DED}_{I7})[,1] + poly(\overline{DED}_{I7})[,2] + poly(AGE_7)[,1] + SEX + PHYS.INH + URBAN + BP + SEX : BP$
$\bar{S}_{D5} \sim 1 + \log(\overline{DED}_{D5}) + poly(AGE_5)[,1] + poly(AGE_5)[,2] + SEX + PHYS.INH + URBAN + BP + W_{A5} + N_{D5}$ $\bar{S}_{D6} \sim 1 + \log(\overline{DED}_{D6}) + poly(AGE_6)[,1] + poly(AGE_6)[,2] + BP + W_{A6} + N_{D6}$ $\bar{S}_{D7} \sim 1 + \log(\overline{DED}_{D7}) + poly(AGE_7)[,1] + poly(AGE_7)[,2] + PHYS.INH + URBAN + BP + URBAN : BP + W_{A7} + W_{I7} + N_{D7}$

Table 5.6: Reduced model equations for each of the nine average claim size models after applying sequential backward selection based on the Wald test

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
Intercept	3.4401	0.0093	369.23	$< 2 \cdot 10^{-16}$
$\overline{DED}_{A7}$	0.1445	0.0057	25.51	$< 2 \cdot 10^{-16}$
$poly(AGE_7)[,1]$	19.5573	0.6597	29.65	$< 2 \cdot 10^{-16}$
$poly(AGE_7)[,2]$	−6.4546	0.6748	−9.56	$< 2 \cdot 10^{-16}$
SEX	0.0263	0.0084	3.13	0.0018
$PHYS.INH$	−0.0057	0.0042	−1.36	0.1732
$URBAN$	0.0315	0.0038	8.26	$< 2 \cdot 10^{-16}$
BP	0.0291	0.0037	7.92	$< 2 \cdot 10^{-16}$
$\overline{DED}_{A7} : SEX$	−0.0140	0.0067	−2.09	0.0369
$PHYS.INH : URBAN$	−0.0057	0.0025	−2.34	0.0195
W_{I7}	0.0952	0.0101	9.40	$< 2 \cdot 10^{-16}$
W_{D7}	−0.0166	0.0069	−2.39	0.0167
N_{A7}	0.0066	0.0003	22.36	$< 2 \cdot 10^{-16}$

Table 5.7: Model summary for the reduced ZTNB regression model of the claim frequencies for treatment field ambulant in 2007, θ estimated to be 2.41.

Binary margins

The distribution of eight combinations of zero-claims over the three fields in 2005 to 2007 is listed in Table 5.8. More than 40% of the insured persons in every year did not claim any reimbursement whatsoever. Recall that $\{W_j = 0\}$ refers to *not* having a zero claim. The copula arguments for (5.3.2) will be determined using predicted cdf $\hat{P}(W_{ij} \leq 0|\boldsymbol{x}_{ij}^W) = \frac{1}{1+\exp(\boldsymbol{x}_{ij}^{Wt}\hat{\boldsymbol{\beta}}_j)}$ and $\hat{P}(W_{ij} \leq 1|\boldsymbol{x}_{ij}^W) = 1$. In Table 5.9 the fitted copula parameters for the independence copula as well as the trivariate Gaussian, Student t, Clayton and Gumbel copulas are given. Note that we are not using a PCC for modeling the dependency between the binary margins. This would imply multiple integration of the PCC with different upper boundaries in order to obtain joint

A	I	D	2005	2006	2007
1	1	1	44.12%	41.27%	40.22%
1	1	0	2.49%	2.54%	2.55%
1	0	1	0.46%	0.39%	0.46%
0	1	1	13.73%	14.60%	13.73%
1	0	0	0.05%	0.04%	0.04%
0	1	0	30.20%	31.84%	33.40%
0	0	1	3.22%	3.35%	3.28%
0	0	0	5.74%	5.97%	6.32%

Table 5.8: Distribution of outcomes of $\boldsymbol{W}$ in the data for 2005 - 2007

cdfs of these margins, which would then be used in (5.3.2) for calculating the joint pmf. In order to compare those fits we utilize a test proposed by Vuong (1989) and the distribution-free test (Clarke (2007)) for nonnested model comparison. Both tests are described in Section 4.4. The test decisions applied to our data are given in Table 5.10. Note that we also apply the

		Year	MLE
Gaussian	$(\hat{\tau}^W_{AI}, \hat{\tau}^W_{AD},$	2005	$(0.373, 0.886, 0.420)'$
	$\hat{\tau}^W_{ID})'$	2006	$(\mathbf{0.319}, \mathbf{0.816}, \mathbf{0.384})'$
		2007	$(\mathbf{0.382}, \mathbf{0.886}, \mathbf{0.410})'$
Student t	$(\hat{\psi}^W_{AI}, \hat{\psi}^W_{AD},$	2005	$(\mathbf{0.329}, \mathbf{0.908}, \mathbf{0.366}, \mathbf{19.86})'$
	$\hat{\psi}^W_{ID}, \hat{\nu}^W)'$	2006	$(0.408, 0.759, 0.382, 18.73)'$
		2007	$(0.405, 0.771, 0.387, 18.84)'$
Clayton	$\hat{\theta}^W$	2005	0.642
		2006	0.623
		2007	0.640
Gumbel	$\hat{\lambda}^W$	2005	1.917
		2006	1.837
		2007	1.838

Table 5.9: Fitted copula parameters for different trivariate copula families with binary margins in 2005 - 2007. The preferred models according to Vuong and Clarke tests (see Table 5.10) are highlighted in boldtype.

Schwarz correction described in these papers for the number of parameters. In each cell the decisions toward model (I) labeled row-wise or (II) labeled column-wise are given. The decision of the Vuong test together with its p value is given in the first row of each cell. The decision of the Clarke test with the p value in brackets are given in the second row. We that see the independence copula is not preferred over any other copula for both the Vuong and the Clarke test in any year. Also the Clayton and Gumbel are not preferred over the Gaussian and Student t copula fit. Between these two classes the Student t copula is preferred according to the Clarke test in 2005, whereas the Vuong test decision is less significant. For 2006 and 2007 the Clarke test chooses the Gaussian model with very low p-value. For all three years we see in Table 5.9 strong correlation between the binary margins. It is driven not only by the health status of the insured person but also by the incentive the insurer sets: if no bill is refunded in any of the three fields throughout the year, the policyholder will receive a premium refund. The

more policyholders can "optimize" their medical treatment patterns, the higher the correlation between these fields will be. This explains the high correlation between ambulant and dental treatments. Since the policyholders' influence on whether or not they have to go to a hospital (inpatient treatments) will be very low, the correlations between the ambulant/ dental field and the inpatient field are relatively low.

(II) / (I)	2005			
	Gaussian	**Student t**	Clayton	Gumbel
Indep.	V: (II) $< 2 \cdot 10^{-16}$	(II) $< 2 \cdot 10^{-16}$	(II) $< 2 \cdot 10^{-16}$	(II) $< 2 \cdot 10^{-16}$
	C: (II) $< 2 \cdot 10^{-16}$	(II) $< 2 \cdot 10^{-16}$	(II) $< 2 \cdot 10^{-16}$	(II) $< 2 \cdot 10^{-16}$
Gaussian		(II) 0.0004	(I) $< 2 \cdot 10^{-16}$	(I) $< 2 \cdot 10^{-16}$
		(II) $< 2 \cdot 10^{-16}$	(I) $< 2 \cdot 10^{-16}$	(I) $< 2 \cdot 10^{-16}$
Student t			(I) $< 2 \cdot 10^{-16}$	(I) $< 2 \cdot 10^{-16}$
			(I) $< 2 \cdot 10^{-16}$	(I) $< 2 \cdot 10^{-16}$
Clayton				(II) $< 2 \cdot 10^{-16}$
				(II) $< 2 \cdot 10^{-16}$
(II) / (I)	2006			
	Gaussian	Student t	Clayton	Gumbel
Indep.	V: (II) $< 2 \cdot 10^{-16}$	(II) $< 2 \cdot 10^{-16}$	(II) $< 2 \cdot 10^{-16}$	(II) $< 2 \cdot 10^{-16}$
	C: (II) $< 2 \cdot 10^{-16}$	(II) $< 2 \cdot 10^{-16}$	(II) $< 2 \cdot 10^{-16}$	(II) $< 2 \cdot 10^{-16}$
Gaussian		(II) 0.2063	(I) $< 2 \cdot 10^{-16}$	(I) $< 2 \cdot 10^{-16}$
		(I) $< 2 \cdot 10^{-16}$	(I) $< 2 \cdot 10^{-16}$	(I) $< 2 \cdot 10^{-16}$
Student t			(I) $< 2 \cdot 10^{-16}$	(I) $< 2 \cdot 10^{-16}$
			(I) $< 2 \cdot 10^{-16}$	(I) $< 2 \cdot 10^{-16}$
Clayton				(II) $< 2 \cdot 10^{-16}$
				(II) $< 2 \cdot 10^{-16}$
(II) / (I)	2007			
	Gaussian	Student t	Clayton	Gumbel
Indep.	V: (II) $< 2 \cdot 10^{-16}$	(II) $< 2 \cdot 10^{-16}$	(II) $< 2 \cdot 10^{-16}$	(II) $< 2 \cdot 10^{-16}$
	C: (II) $< 2 \cdot 10^{-16}$	(II) $< 2 \cdot 10^{-16}$	(II) $< 2 \cdot 10^{-16}$	(II) $< 2 \cdot 10^{-16}$
Gaussian		(II) 0.0001	(I) $< 2 \cdot 10^{-16}$	(I) $< 2 \cdot 10^{-16}$
		(I) $< 2 \cdot 10^{-16}$	(I) $< 2 \cdot 10^{-16}$	(I) $< 2 \cdot 10^{-16}$
Student t			(I) $< 2 \cdot 10^{-16}$	(I) $< 2 \cdot 10^{-16}$
			(I) $< 2 \cdot 10^{-16}$	(I) $< 2 \cdot 10^{-16}$
Clayton				(I) $< 2 \cdot 10^{-16}$
				(I) $< 2 \cdot 10^{-16}$

Table 5.10: Preferred model according to the tests proposed by Vuong ("V", first row of each cell) and Clarke ("C", second row) followed by p-values for different copula choices modeling the dependence structure of the binary margins $\boldsymbol{W}$. The preferred models are highlighted in boldtype.

Continuous margins

The arguments of the PCC model given in (5.2.7) will be estimated using Lemma 1, i.e., for $i \in \mathcal{I}_j$

$$\hat{F}_{T_{ij}^+}(t_{ij}^+|\boldsymbol{x}_{ij}^N, \boldsymbol{x}_{ij}^{\bar{S}}, \boldsymbol{w}_{i(-j)}, n_{ij}) := \sum_{k=1}^{\infty} \hat{F}_{\bar{S}}\left(\frac{t_{ij}^+}{k}|\boldsymbol{x}_{ij}^N, \boldsymbol{x}_{ij}^{\bar{S}}, \boldsymbol{w}_{i(-j)}, n_{ij}\right) \hat{p}_N(k). \tag{5.4.2}$$

Two additional choices have to be made in order to fully specify the PCC. First one needs to determine which pairs of margins will be modeled by the unconditional copulas c_{12} and c_{13} and which by $c_{23|1}$, i.e., the problem of choosing a good permutation of the margins. Further one needs to pick appropriate copula families to describe the dependency structure between pairs of margins. The first problem may be addressed for example by performing a simple a priori fit. Thereby we fit three arbitrary but identical bivariate Gaussian copulas on the subset of the data, where all observations with at least on zero claim in either of the two margins have been taken out. The two pairs of margins with the strongest fitted correlation parameter will be modeled by c_{12} and c_{13}. For the data at hand, there is no permutation necessary for any of the three years, i.e., we choose treatment fields A, I and D to be the margins $1, 2$ and 3, respectively, hence $c_{12} = c_{AI}$, $c_{13} = c_{AD}$ and $c_{23|1} = c_{ID|A}$. The second problem may be addressed by looking at scatterplots for the same reduced data subsets. Since it is hard to detect typical copula structures from scatterplots based on marginally transformed uniform margins $u_{ij} := \hat{F}_{T_{ij}^+}(t_{ij}^+|\boldsymbol{x}_{ij}^N, \boldsymbol{x}_{ij}^{\bar{S}}, \boldsymbol{w}_{i(-j)}, n_{ij})$, $j = 1, 2, 3$, $i \in \mathcal{I}_j$, we consider scatterplots of $z_{ij} := \Phi^{-1}(u_{ij}))$, where $\Phi^{-1}(\cdot)$ is the quantile of the standard normal distribution. We will compare these plots to contour plots of the corresponding theoretical copulas with standard normal margins at the maximum likelihood estimate of the empirical data. In Figure 5.1 scatterplots of Z_{j_1} and Z_{j_2} are plotted for the pairs of margins AI and AD in 2005 to 2007. Additionally kernel density estimates are added to these scatterplots. The theoretical contour plots for an appropriate copula choice are plotted to the right of each scatterplot. The copula parameters are the MLE obtained from the corresponding data conditional on $\{W_{j_1} = 0\}$ and $\{W_{j_2} = 0\}$. Based on these copula choices the conditional arguments of $c_{23|1}$ can be calculated. For example, for 2005 we have to determine $u_{iI|A5} := h^C(u_{iI5}|u_{iA5}, \hat{\theta} = 0.33)$ and $u_{iD|A5} := h^{Ga}(u_{iD5}|u_{iA5}, \hat{\rho} = 0.10)$, $i \in \mathcal{I}_A \cap I_I \cap I_D$, where h^C and h^{Ga} are the h functions w.r.t. to the Clayton and the Gaussian copula, respectively (see Appendix 5.4.5), and 0.33 and 0.10 are the MLE of these copulas determined in the previous step. We will plot $z_{iI|A5} := \Phi^{-1}(u_{iI|A5})$ and $z_{iD|A5} := \Phi^{-1}(u_{iD|A5})$ in Figure 5.2 and proceed similarly as before in order to choose appropriate copulas. The maximum likelihood estimates when jointly estimating the copula parameters for the PCC's are given in Table 5.11. Since in 2006 the parameter of $c_{23|1}$ of the Gumbel copula is close to 1 and the parameter of the Gaussian copula for $c_{23|1}$ in 2007 is close to 0, we replace these copulas by the independence copulas. The optimal model choices are typed bold. For these copulas there is a one-to-one relationship to Kendall's τ, i.e., we can determine theoretical Kendall's τ corresponding to the ML copula parameters and compare them to the empirical Kendall's τ. For the Gaussian and the Student t copulas we transform $\tau := 2/\pi \cdot \sin^{-1}(\rho)$, for the Clayton we need to calculate $\tau := \theta/(2+\theta)$ and for the Gumbel we have $\tau := 1 - 1/\lambda$ (see for instance Frees and Valdez (1998, Appendix B)). The empirical Kendall's τ is based on the uniformely transformed margins. The results concerning Kendall's τ are given in the lower panel of Table 5.11: the theoretical and empirical Kendall's τ are quite close which confirms the results of our fitting approach. There is a positive correlation between ambulant and inpatient as well as for ambulant and dental

Copulas	Year	$\hat{\zeta}^{T^+}_{AI}$	$\hat{\zeta}^{T^+}_{AD}$	$\hat{\zeta}^{T^+}_{ID\|A}$
C / C / Ga	2005	**0.333**	**0.096**	**−0.041**
C / C / Gu	2006	0.346	0.176	1.010
C / C / Ind		**0.345**	**0.176**	
t / C / Ga	2007	0.272, $df = 20.9$	0.144	−0.005
t / C / Ind		**0.272**, **df = 20.9**	**0.144**	
corresponding Kendall's τ				
theor.	2005	0.143	0.061	−0.026
empir.		*0.171*	*0.058*	*−0.027*
theor.	2006	0.147	0.081	0.010
theor.		0.147	0.081	
empir.		*0.178*	*0.082*	*0.027*
theor.	2007	0.175	0.067	−0.003
theor.		0.175	0.067	
empir.		*0.173*	*0.066*	*−0.013*

Table 5.11: Maximum likelihood estimates of the copula parameters for the Gaussian (Ga), Student t (t), Clayton (C) and Gumbel copula (Gu). Corresponding theoretical Kendall's τ and empirical Kendall's τ of copula data. Updated fit using the independence copula for $ID|A$ in 2006 and 2007.

treatments for all three years, which is driven by the health status of the insured person. Given ambulant treatments, the correlation between inpatient and dental treatments is close to zero and is set to zero for 2006 and 2007.

5.4.5 Model interpretation

For the year 2007 we aim to investigate the influence of AGE on the predicted probability of a refund $\hat{P}_W(1,1,1|\boldsymbol{x}_j^W)$. Thereby we fix all other covariates, i.e., we fix the applied deductible DED_{A7} at its median value 34.85, whereas DED_{I7} and DED_{D7} will be fixed at 0. The buying power will be fixed at its mode 19499.40 and the urbanity at its median 396.35. The number of physicians per inhabitants we set to its mode 0.223. Modes are estimated using kernel density estimates of histograms of the covariates. For men and women, the influence of AGE on $\hat{P}_W(1,1,1|\boldsymbol{x}_j^W)$ both under the joint model and assuming independence are graphed in the left panel of Figure 5.3.

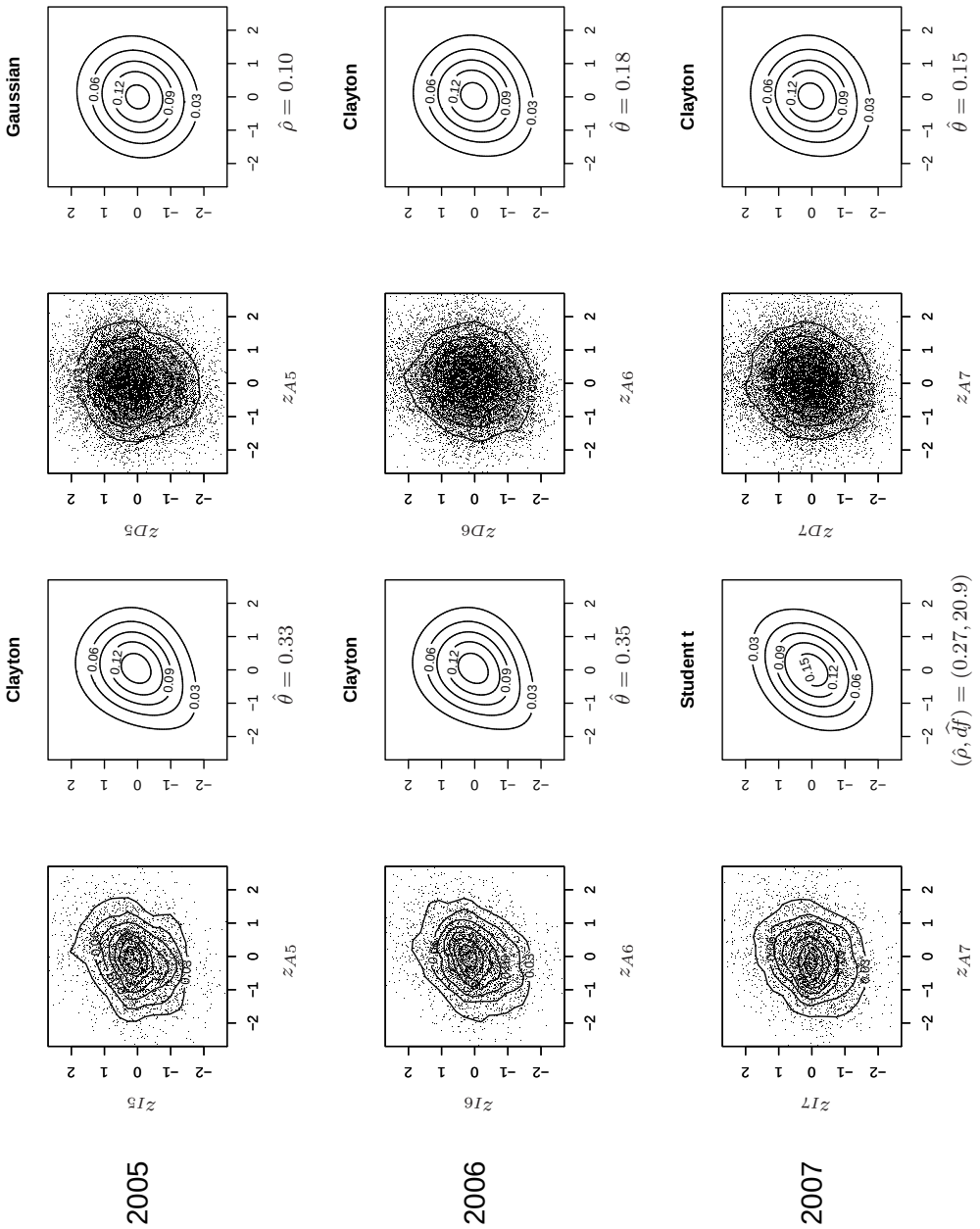

Figure 5.1: Scatterplots of pairs of $z_{ij} := \Phi^{-1}\left(\hat{F}_{T_{ij}^+}(t_{ij}^+|\boldsymbol{x}_{ij}^N, \boldsymbol{x}_{ij}^{\bar{S}}, \boldsymbol{w}_{i(-j)}, n_{ij})\right)$, $j = 1, 2, 3$ with contour plots of bivariate kernel density estimates for ambulant / inpatient margins (first column) and for ambulant / dental margins (third column). In column two (four) we show theoretical contour plots based on a chosen pair copula family for ambulant / hospital (ambulant / dental) margins.

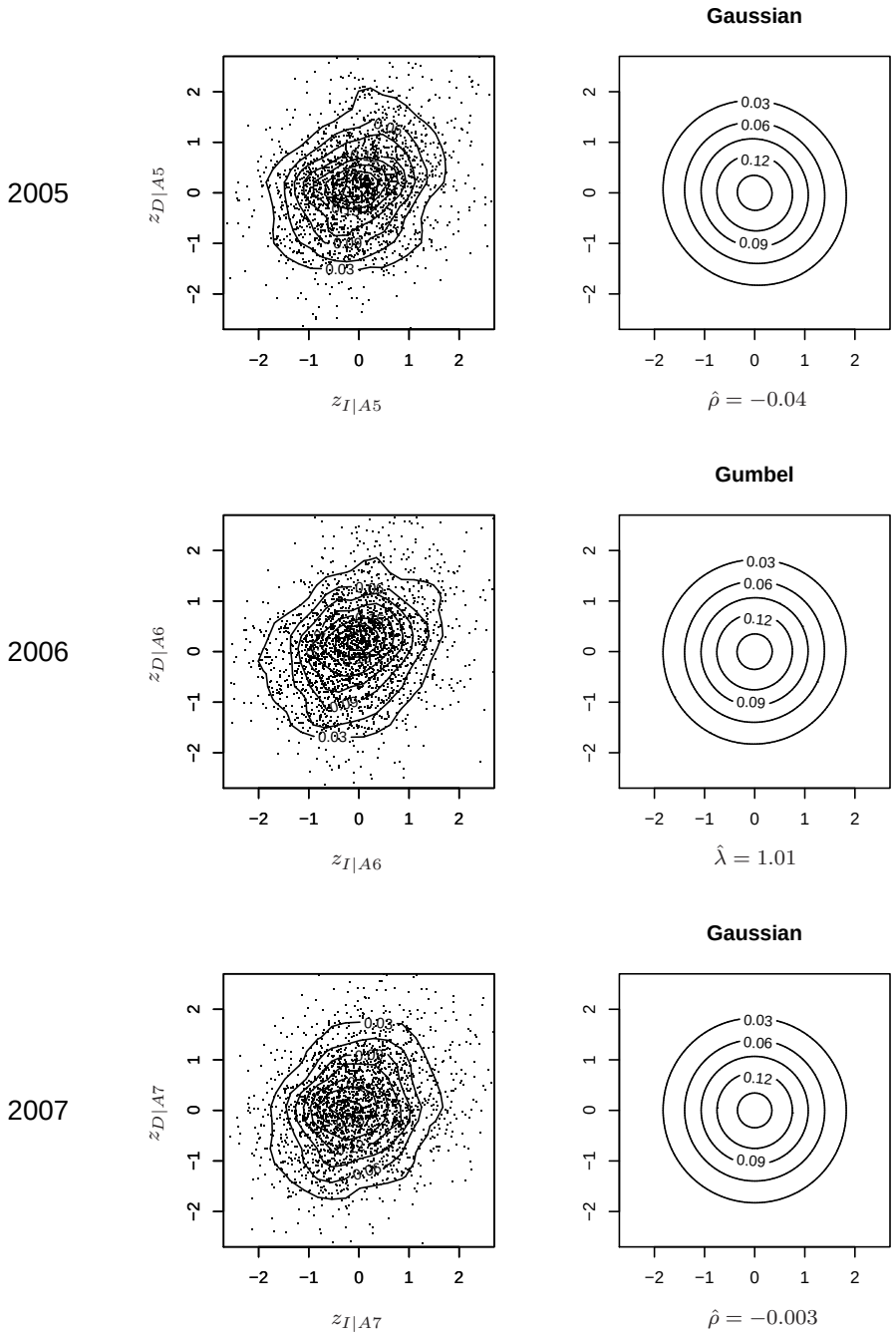

Figure 5.2: Scatterplots of conditional pairs of $z_{ij_1|j_2} := \Phi^{-1}\left(h(u_{ij_1}|u_{ij_2})\right)$, $j_k = 1, 2, 3$ with contour plots of bivariate kernel density estimates for inpatient / dental margins given the ambulant margin (first column). In column two we show theoretical contour plots based on a chosen pair copula family for each year.

Male insured persons have a higher refund probability in general. Since AGE was taken into our models as a piecewise linear function there is a jump at 32. Whereas earlier than 32 the refund probability slightly increases, it rapidly falls when the person gets older, hence it becomes increasingly difficult to get the premium refund. In a second step we are interested in estimating the density of $T_1^+ + T_2^+ + T_3^+$, therefore we additionally fix AGE at its mode of

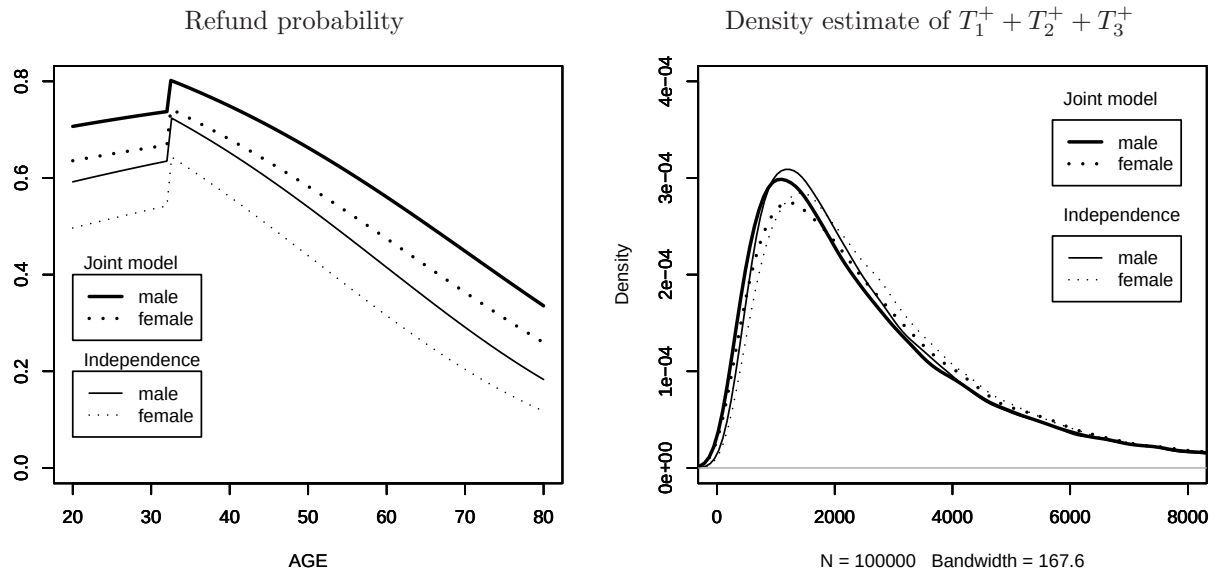

Figure 5.3: Influence of AGE on the refund probability when assuming independence and using the joint fitted probability $p_{\boldsymbol{W}}(\boldsymbol{w})$ while fixing other covariates; density estimates of sums of claims.

40.79. Further we assume we have a rather sick person and set $\boldsymbol{W} := (0,0,0)'$, i.e., we assume a claim occurred in each treatment field. The arguments of the PCC will be predictive cdfs of T_j^+, $j = 1,2,3$ determined according to (5.4.2). We approximate quantile functions for T_j^+ using the R function "approxfun" in package *stats*. Then we proceed by sampling $(t_{r1}^+, t_{r2}^+, t_{r3}^+)'$, $r = 1, \ldots, 100\,000$ from $\boldsymbol{T}$. Sampling from a C-vine is straightforward, we refer to Aas, Czado, Frigessi, and Bakken (2009) for details. Finally we compute $t_r^+ := t_{r1}^+ + t_{r2}^+ + t_{r3}^+$ and plot its density estimate using the *stats* function "density". On the right panel of Figure 5.3 we see that the highest predicted density of $T_1^+ + T_2^+ + T_3^+$ when using the joint model for males lies around 1600 Euro (1750 Euro for females). Under the independence assumption the peaks of the estimated densities are even higher, therefore the joint model also reflects diversification effects.

Summary and Discussion

For the first time, a multivariate analysis of claims including zero claims based on PCC's was carried out. We have fitted separately a joint distribution for total claims given zero claim events and for the zero claims. The total claims given zero claim events can be expressed as a PCC under margins. Whatever combination of zero claims occurs one gains knowledge in terms of a likelihood contribution either on the correlation of the total positive claims or on the correlation of the zero claims. Even if the percentage of positive claims in one or more margins is very low, our approach yet allows to fit these data. In higher dimensional problems, however, the computational effort of numerically integrating margins out, increases. Other approaches for approximating high-dimensional integrals may be more efficient for the problem at hand and may decrease the computational time. Such approaches might also allow to efficiently approximate the joint cdf of the PCC and hence to model the dependency of the binary margins also based on PCC's. The choice of the bivariate copula families of such a PCC with binary margins is still

an open question.

In an application to health insurance we saw that the zero claim events between ambulant and dental treatments show a large positive correlation. There is a positive correlation also for the positive claims fitted by the pair-copula construction. The correlation is driven by the health status of the insured person. Given ambulant treatments, the correlation between inpatient and dental treatments is very low and needs not be fitted by a copula for 2006 and 2007, i.e., we may assume independence between the conditional margins.

Appendix of Chapter 5

Definition of selected copulas

Definition 2 (Gaussian copula)**.** The J-dimensional Gaussian copula with association matrix $\mathbf{\Sigma} = (\tau_{ij})_{i,j=1,\dots,J}$ is given by

$$C_J^G(u_1,\dots,u_J|\mathbf{\Sigma}) := \Phi_J\left(\Phi^{-1}(u_1),\dots,\Phi^{-1}(u_J)|\mathbf{\Sigma}\right), \tag{5.4.3}$$

where $\Phi_J(\cdot|\mathbf{\Sigma})$ is the cdf of the J-dimensional normal distribution with mean $\boldsymbol{\mu} = \mathbf{0}_J$ and covariance $\mathbf{\Sigma}$, $\phi_J(\cdot|\mathbf{\Sigma})$ its density and $\Phi^{-1}(\cdot)$ is the quantile of the standard normal distribution.

In the special case of $J = 2$ we use notation $C_{12}^G(u_1,u_2|\tau_{12}) = \Phi_2(\Phi^{-1}(u_1),\Phi^{-1}(u_2)|\tau_{12})$ instead of (5.4.3).

Definition 3 (Student t copula)**.** The J-dimensional t copula with parameters ν and $\mathbf{\Psi} = (\psi_{ij})_{i,j=1,\dots,J}$ is given by

$$\begin{aligned} C_J^t(u_1,\dots,u_J|\nu,\mathbf{\Psi}) &:= F_J(t_\nu^{-1}(u_1),\dots t_\nu^{-1}(u_J)|\nu,\mathbf{\Psi}) \\ &= \int_{-\infty}^{t_\nu^{-1}(u_1)}\cdots\int_{-\infty}^{t_\nu^{-1}(u_J)} \frac{\Gamma(\frac{\nu+J}{2})}{\Gamma(\frac{\nu}{2})\sqrt{(\pi\nu)^J|\mathbf{\Psi}|}}\left(1+\frac{\boldsymbol{x}'\mathbf{\Psi}^{-1}\boldsymbol{x}}{\nu}\right)^{-\frac{\nu+J}{2}} d\boldsymbol{x}, \end{aligned} \tag{5.4.4}$$

where $F_J(\cdot|\nu,\mathbf{\Psi})$ is the joint cdf of a t distributed random vector with mean $\mathbf{0}$, covariance $\mathbf{\Psi}$ and ν degrees of freedom, $f_J(\cdot|\nu,\mathbf{\Psi})$ its density, and t_ν^{-1} denotes the quantile function of a standard univariate t_ν distribution.

For $J = 2$ we write $C_{12}^t(u_1,u_2|\nu,\psi_{12})$ instead of (5.4.4).

Definition 4 (Archimedean copula)**.** Archimedean copulas are defined as

$$C_J(u_1,\dots,u_J|\theta) = \varphi^{-1}\big(\sum_{j=1}^{J}\varphi(u_j)\big), \tag{5.4.5}$$

where function φ is called generator. Further $\varphi : [0,1] \to [0,\infty)$ is a continuous, strictly monotonic decreasing convex function with $\varphi(1) = 0$.

We consider in particular the Clayton and the Gumbel copula. The generator for the J-dimensional Clayton copula with parameter $\theta > 0$ is $\varphi^C(u) := \frac{1}{\theta}(u^{-\theta}-1)$. For the J-dimensional Gumbel copula with parameter $\lambda \geq 1$ is $\varphi^{Gu}(u) := (-\log(u))^\lambda$. The bivariate copula densities (for the Clayton and Gumbel see Venter (2001)) together with h functions defined in 5.2.4 (see Aas, Czado, Frigessi, and Bakken (2009)) are given in Table 5.12.

	Bivariate copula density	$h(u_1\|u_2)$
Gaussian	$\phi_2\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\|\tau_{12}\right) \cdot \prod_{j=1}^{2} \frac{1}{\phi(\Phi^{-1}(u_j))}$	$\Phi\left(\frac{\Phi^{-1}(u_1)-\tau_{12}\Phi^{-1}(u_2)}{\sqrt{1-\tau_{12}{}^2}}\right)$
Student t	$f_2(t_\nu^{-1}(u_1), t_\nu^{-1}(u_2)\|\nu,\psi_{12}) \cdot \prod_{j=1}^{2} \frac{1}{f_\nu(t_\nu^{-1}(u_j))}$	$t_{\nu+1}\left(\frac{t_\nu^{-1}(u_1)-\psi_{12}t_\nu^{-1}(u_2)}{\sqrt{\frac{\nu+\left(t_\nu^{-1}(u_2)\right)^2(1-(\psi_{12})^2)}{\nu+1}}}\right)$
Clayton	$(1+\theta)(u_1u_2)^{-1-\theta}\left(u_1^{-\theta}+u_2^{-\theta}-1\right)^{-1/\theta-2}$	$u_2^{-\theta-1}\left(u_1^{-\theta}+u_2^{-\theta}-1\right)^{-1-1/\theta}$
Gumbel	$C_{12}(u_1,u_2) \quad (u_1u_2)^{-1} \quad ((-\log u_1)^\lambda + (-\log u_2)^\lambda)^{-2+2/\lambda}(\log u_1 \log u_2)^{\lambda-1}$ $\left[1+(\lambda-1)((-\log u_1)^\lambda+(-\log u_2)^\lambda)^{-1/\lambda}\right]$, where $C_{12}(u_1,u_2) = \exp\left(-\left[(-\log u_1)^\lambda + (-\log u_2)^\lambda\right]^{1/\lambda}\right)$	$C_{12}(u_1,u_2)\frac{1}{u_2}(-\log u_2)^{\lambda-1}$ $\left[(-\log u_1)^\lambda+(-\log u_2)^\lambda\right]^{1/\lambda-1}$

Table 5.12: Bivariate copula densities and h functions for selected copulas

5.5 Proofs of Lemmas and Propositions

Proof. (Lemma 1)

$$\begin{aligned}
F_{T^+|\{W=0\}}(t^+) &= P(N\bar{S} \le t^+|\{W=0\}) \\
&= \sum_{k=1}^{\infty} P(N\bar{S} \le t^+|\{N=k, W=0\})P(N=k|\{W=0\}) \\
&= \sum_{k=1}^{\infty} P(\bar{S} \le \frac{t^+}{k}|\{N=k, W=0\})P(N=k|\{W=0\}) \\
&= \sum_{k=1}^{\infty} F_{\bar{S}|\{N,W=0\}}\left(\frac{t^+}{k}|\{N=k\}\right) p_{N|\{W=0\}}(k).
\end{aligned}$$

□

Proof. (Lemma 2) For $t \ge 0$

$$\begin{aligned}
F_T(t) &= P(T \le t) \\
&= P(T \le t|\{W=1\}) \cdot P(W=1) + P(T \le t|\{W=0\}) \cdot P(W=0) \\
&= P((1-W)T^+ \le t|\{W=1\}) \cdot P(W=1) \\
&\quad + P(W=0) \cdot P((1-W)T^+ \le t|\{W=0\}) \\
&= P(0 \le t|\{W=1\}) \cdot P(W=1) + P(W=0) \cdot P(T^+ \le t|\{W=0\}) \\
&= P(W=1) + \mathbb{1}_{\{t>0\}} P(W=0) \cdot F_{T^+|\{W=0\}}(t) \\
&= p_W(0) + \mathbb{1}_{\{t>0\}}(1-p_W(0)) \cdot F_{T^+|\{W=0\}}(t).
\end{aligned}$$

□

Bibliography

Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009). Pair-copula constructions of multiple dependence. *Insurance: Math. and Econom. 44*(2), 182–198.

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

Avramidis, A. N., N. Channouf, and P. L'Ecuyer (2009). Efficient Correlation Matching for Fitting Discrete Multivariate Distributions with Arbitrary Marginals and Normal-Copula Dependence. *INFORMS J. Comput. 21*, 88–106.

Baba, K., R. Shibata, and M. Sibuya (2004). Partial correlation and conditional correlation as measures of conditional independence. *Aust. & N.Z. J. Statist. 46*, 657–664(8).

Bae, S., F. Famoye, J. T. Wulu, A. A. Bartolucci, and K. P. Singh (2005). A rich family of generalized Poisson regression models. *Math. Comput. Simulation 69*(1-2), 4–11.

Baierl, A., M. Bogdan, F. Frommlet, and A. Futschik (2006). On Locating Multiple Interacting Quantitative Trait Loci in Intercross Designs. *Genet. 173*(3), 1693–1703.

Baierl, A., A. Futschik, M. Bogdan, and P. Biecek (2007). Locating multiple interacting quantitative trait loci using robust model selection. *Comp. Stat. and Dat. Anal. 51*, 6423–6434.

Ball, R. (2001). Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. *Genet. 159*(3), 1351–1364.

Bedford, T. and R. M. Cooke (2001a). *Monte Carlo simulation of vine dependent random variables for applications in uncertainty analysis*. 2001 Proceed. of ESREL2001, Turin, Italy.

Bedford, T. and R. M. Cooke (2001b). Probability Density Decomposition for Conditionally Dependent Random Variables Modeled by Vines. *Ann. Math. Artif. Intell. 32*(1-4), 245–268.

Bedford, T. and R. M. Cooke (2002). Vines -a new graphical model for dependent random variables. *Ann. Statist 30*, 1031–1068.

Belasco, E. J. and S. K. Ghosh (2008). Modeling censored data using mixture regression models with an application to cattle production yields. 2008 Annual Meeting, Orlando, Florida 6341, Agricultural and Applied Economics Association.

Biller, B. (2009). Copula-Based Multivariate Input Models for Stochastic Simulation. *Oper. Res. 57*(4), 878–892.

Biller, B. and S. Ghosh (2006). *Multivariate input processes*. In: Handbooks in Operations Research and Management Science: Simulation, ed. B. L. Nelson and S. G. Henderson. Elsevier Science, Amsterdam.

Biller, B. and C. Gunes (2008). Accounting for multivariate input model uncertainty in large-scale stochastic simulations. Technical report, Tepper Working Paper, Carnegie Mellon University, Pittsburgh, PA.

Biller, B. and B. L. Nelson (2003). Modeling and generating multivariate time-series input processes using a vector autoregressive technique. *ACM Trans. Model. Comput. Simul. 13*(3), 211–237.

Bogdan, M., A. Chakrabarti, and J.K.Ghosh (2008). Optimal rules for multiple testing and sparse multiple regression. *Tech. Rep. I-18/08/P-003*. Institute of Mathematics and Computer Science, Wroclaw University of Technology, *www.im.pwr.wroc.pl/∼mbogdan/papers*.

Bogdan, M., F. Frommlet, P. Biecek, R. Cheng, J. Ghosh, and R. Doerge (2008). Extending the Modified Bayesian Information Criterion (mBIC) to dense markers and multiple interval mapping. *Biometrics 64*(8), 1162–1169.

Bogdan, M., J. Ghosh, and R. Doerge (2004). Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitative trait loci. *Genet. 167*(2), 989–999.

Bogdan, M., J. Ghosh, and M. Żak-Szatkowska (2008). Selecting explanatory variables with the modified version of Bayesian Information Criterion. *Qual. Reliab. Eng. Int. 24*, 627–641.

Brezger, A. and S. Lang (2006, February). Generalized structured additive regression based on bayesian p-splines. *Comput. Stat. Data Anal. 50*(4), 967–991.

Broman, K. (1997). Identifying quantitative trait loci in experimental crosses. Master's thesis, PhD dissertation. Department of Statistics, University of California, Berkeley, CA.

Broman, K. (2003). Mapping quantitative trait loci in the case of a spike in the phenotype distribution. *Genet. 163*(3), 1169–1175.

Broman, K. and T. Speed (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. Roy. Stat. Soc. B 64*, 641–656.

Cario, M. C. and B. L. Nelson (1996). Autoregressive to anything: Time-series input processes for simulation. *Oper. Res. Lett.* (19), 51–58.

Cario, M. C. and B. L. Nelson (1997). Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Technical report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL.

Chang, Y.-C. (2000). Residuals analysis of the generalized linear models for longitudinal data. *Stat. Med. 19*(10), 1277–1293.

Chen, H. (2000). Initialization for norta: Generation of random vectors with specified marginals and correlations. *INFORMS J. Comput.* (13), 312–331.

Chen, J. and Z. Chen (2008). Extended Bayesian Information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771.

Chen, Z. and J. Liu (2009). Mixture Generalized Linear Models for Multiple Interval Mapping of Quantitative Trait Loci in Experimental Crosses. *Biometrics 65*(2), 470–477.

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

Clarke, K. A. (2003). Nonparametric model discrimination in international relations. *J. Conflict Resolution 47*, 72–93.

Coffman, C., R. Doerge, K. Simonsen, K. Nichols, and C. Duarte (2005). Model selection in binary trait locus mapping. *Genet. 170*(3), 1281–1297.

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

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

Consul, P. C. and G. C. Jain (1970). On the generalization of Poisson distribution. *Ann. Math. Statist. 41*, 1387.

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

Cruyff, M. J. L. F. and P. G. M. van der Heijden (2008). Point and Interval Estimation of the Population Size Using a Zero-Truncated Negative Binomial Regression Model. *Biom. J. 50*(6), 1035–1050.

Cui, Y. and W. Yang (2009). Zero inflated generalized Poisson regression mixture model for mapping quantitative trait loci underlying count trait with many zeros. *J. Theor. Biol. 256*(2), 276–285.

Czado, C., V. Erhardt, A. Min, and S. Wagner (2007). Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. *Stat. Modell. 7*(2), 125–153.

Czado, C., A. Min, T. Baumann, and R. Dakovic (2009). Pair-copula constructions for modeling exchange rate dependence. *Submitted*. Preprint at *http://www-m4.ma.tum.de/Papers/index.html*.

Czado, C., H. Schabenberger, and V. Erhardt (2009). Nonnested model selection for spatial count regression models with application to health insurance. *Stat. Pap., accepted.*.

Deb, P., M. K. Munkin, and P. K. Trivedi (2006). Private Insurance, Selection, and Health Care Use: A Bayesian Analysis of a Roy-Type Model. *J. Bus. & Econ. Stat. 24*(4), 403–415.

DeMaris, A. (2004). *Regression with social data: modeling continuous and limited response variables*. Hoboken, N.J. : John Wiley & Sons, Inc.

Dutang, C., V. Goulet, and M. Pigeon (2008). actuar: An R Package for Actuarial Science. *J. Stat. Software*. Preprint available from: http://www.cran.r-project.org/web/packages/actuar/.

Embrechts, P. (2009). Copulas: A personal view. *J. Risk Insur. 76*(3), 639–650.

Erhardt, V. (2009). ZIGP: Zero-Inflated Generalized Poisson (ZIGP) Models. R package version 3.8.

Erhardt, V., M. Bogdan, and C. Czado (2010). Locating multiple interacting quantitative trait loci with the zero-inflated generalized Poisson regression. *Statistical Applications in Genetics and Molecular Biology, to appear*.

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

Erhardt, V. and C. Czado (2009b). Generalized estimating equations for longitudinal generalized Poisson count data with regression effects on the mean and dispersion level. *Submitted*. Preprint at *http://www-m4.ma.tum.de/Papers/index.html*.

Erhardt, V. and C. Czado (2009c). Sampling Count Variables with specified Pearson Correlation - a Comparison between a naive and a C-vine Sampling Approach. *In: Kurowicka, D., Joe, H. (Ed.) Dependence Modeling - Handbook on Vine Copulae*.

Erhardt, V. and C. Czado (2010). Modeling dependent claim totals including zero claims in private health care insurance. *Scandinavian Actuarial Journal, to appear*.

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

Famoye, F., W. J.T., and S. K.P. (2004). On the Generalized Poisson Regression Model with an Application to Accident Data. *J. Dat. Sci. 2*, 287–295.

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

Famoye, F. and K. P. Singh (2006). Zero-inflated generalized Poisson model with an application to domestic violence data. *J. Dat. Sci. 4*(1), 117–130.

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

Frees, E. and E. A. Valdez (1998). Understanding Relationships Using Copulas. *N. Amer. Actuarial J. 2*, 1–25.

Frees, E. W., J. Gao, and M. A. Rosenberg (2007). Predicting the frequency and amount of health care expenditures. *http://research3.bus.wisc.edu/file.php/129/Papers/AggLossExpenditures24Aug2007.pdf*.

Frees, E. W. and E. A. Valdez (2008). Hierarchical Insurance Claims Modeling. *J. Amer. Stat. Assoc. 103*(484), 1457–1469.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2003). *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC.

Ghosh, J., M. Delampady, and T. Samanta (2006). *An introduction to Bayesian analysis-theory and methods*. Springer, Berlin / Heidelberg.

Ghosh, S. and S. G. Henderson (2003). Behavior of the norta method for correlated random vector generation as the dimension increases. *ACM Trans. Model. Comput. Simul. 13*(3), 276–294.

Gilks, W., R. S., and S. D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC.

Gschlößl, S. and C. Czado (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scand. Actuar. J. 2007*(3), 202–225.

Gschlößl, S. and C. Czado (2008). Modelling count data with overdispersion and spatial effects. *Statist. Pap. 49*(3), 531–552.

Gupta, P. L., R. C. Gupta, and R. C. Tripathi (2004). Score test for zero inflated generalized Poisson regression model. *Comm. Statist. Theory Methods 33*(1), 47–64.

Haley, C. and S. Knott (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity 69*, 315–324.

Hall, D. B. and T. A. Severini (1998). Extended generalized estimating equations for clustered data. *J. Amer. Statist. Assoc. 93*(444), 1365–1375.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika 57*(1), 97–109.

Heller, G. Z., D. M. Stasinopoulos, R. A. Rigby, and P. De Jong (2007). Mean and dispersion modelling for policy claims costs. *Scand. Actuarial J. 2007*(4), 281–292.

Hilbe, J. M. (2007). *Negative Binomial Regression*, Volume 1. New York: Cambridge Univ. Press.

Jansen, R. (1993). Interval mapping of multiple quantitative trait loci. *Genet. 135*(1), 205–211.

Jansen, R. and P. Stam (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genet. 136*(4), 1447–1455.

Joe, H. (1996). *Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters*. In L. Rüschendorf and B. Schweizer and M. D. Taylor (Ed.), Distributions with Fixed Marginals and Related Topics.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. London: Chapman and Hall.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *J. Multivar. Anal. 97*(10), 2177–2189.

Joe, H. and R. Zhu (2005). Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. *Biom. J. 47*(2), 219–229.

Jørgensen, B. and M. C. P. de Souza (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scand. Actuarial J.*, 69–93.

Kao, C. and Z. Zeng (2002). Modeling Epistasis of Quantitative Trait Loci Using Cockerham's Model. *Genet. 160*(3), 1243–1261.

Kao, C., Z. Zeng, and R. Teasdale (1999). Multiple interval mapping for quantitative trait loci. *Genet. 152*(3), 1203–1216.

Karlis, D. and L. Meligkotsidou (2005, October). Multivariate poisson regression with covariance structure. *Stat. Comp. 15*(4), 255–265.

Kastenmeier, R. (2008). Joint Regression Analysis of Insurance Claims and Claim Sizes. Master's thesis, Technische Universität München (*www-m4.ma.tum.de/Diplarb/*).

Kawamura, K. (1979). The structure of multivariate Poisson distribution. *Kodai Math. J. 2*, 337–345.

Kopociński, B. (1999). Multivariate negative binomial distributions generated by multivariate exponential distributions. *Appl. Math. 25*(4), 463–472.

Kruglyak, L. and E. Lander (1995). A nonparametric approach for mapping quantitative trait loci. *Genet. 139*(3), 1421–1428.

Kurowicka, D. and R. Cooke (2006). *Uncertainty analysis with high dimensional dependence modelling*. Chichester, England: Wiley: Wiley series in probability and statistics.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics 34*(1), 1–14.

Lander, E. and D. Botstein (1989). Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genet. 121*(1), 185–199.

Li, J., S. Wang, and Z.-B. Zeng (2006). Multiple interval mapping for ordinal traits. *Genet. 173*(3), 1649–1663.

Li, S. and J. Hammond (1975). Generation of Pseudo-Random Numbers with Specified Univariate Distributions and Correlation Coefficients. *IEEE Trans. on Systems, Man and Cybernetics 5*, 557–561.

Li, W. and Z. Chen (2009). Multiple interval mapping for quantitative trait loci with a spike in the trait distribution. *Genet. 182*(2), 337–342.

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

Lurie, P. and M. Goldberg (1998). An Approximate Method for Sampling Correlated Random Variables from Partially-Specified Distributions. *Manage. Sci. 44*(2), 203–218.

Lyons, M. A., H. Wittenburg, R. Li, K. A. Walsh, M. R. Leonard, G. A. Churchill, M. C. Carey, and B. Paigen (2003). New quantitative trait loci that contribute to cholesterol gallstone formation detected in an intercross of CAST/Ei and 129S1/SvImJ inbred mice. *Physiol. Genomics 14*(3), 225–239.

Manichaikul, A., J. Moon, S. Sen, B. Yandell, and K. Broman (2009). A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. *Genet. 181*(3), 1077–1086.

Marida, K. V. (1970). A translation family of bivariate distributions and fréchet's bounds. *Sankhya 32*, 119–122.

McCullagh, P. and J. Nelder (1989). *Generalized linear models* (Second ed.). London: Chapman & Hall.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys. 21*, 1087–1091.

Mikosch, T. (2006). Copulas: Tales and facts. *Extremes 9*, 3–20(18).

Min, A. and C. Czado (2010). Testing for zero-modification in count regression models. *Stat. Sin. 20*, 323–341.

Nelsen, R. B. (2006). *An introduction to copulas. 2nd ed.* Springer Series in Statistics. New York, NY: Springer. xiii, 269 p.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics 57*(1), 120–125.

Pearson, K. (1916). On Some Novel Properties of Partial and Multiple Correlation Coefficients in a Universe of Manifold Characteristics. *Biometrika 11*(3), 231–238.

Pettitt, A. N., I. S. Weir, and A. G. Hart (2002). A conditional autoregressive gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Stat. and Comput. 12*(4), 353–367.

Pitt, M., D. Chan, and R. Kohn (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika 93*(3), 537–554.

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

Sax, K. (1923). The association of size difference with seed-coat pattern and pigmentation in phaseolus vulgaris. *Genet. 8*(6), 552–560.

Schabenberger, H. (2009a). spatcounts: Spatial count regression. *Submitted*. R package version 1.1.

Schabenberger, H. (2009b). Spatial count regression models with applications to health insurance data. Master's thesis, Technische Universität München (*www-m4.ma.tum.de/Diplarb/*).

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

Scott, J. and J. Berger (2008). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Discussion Paper 2008-10, Duke University Department of Statistical Science.

Shao, J. (1999). *Math. Stat.* Springer-Verlag, New York.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris 8*, 229–231.

Smyth, G. K. and B. Jørgensen (2002). Fitting tweedie's compound Poisson model to insurance claims data: dispersion modelling. *ASTIN Bull. 32*(1), 143–157.

Song, P. X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics and Applications*, Volume 1. New York: Springer-Verlag.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B 64*(4), 583–639.

Srivastava, M. and C. Khatri (1979). *An introduction to multivariate statistics.* New York, Oxford: North Holland, New York. XVII, 350 p. $ 19.50 .

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

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

Sun, J., E. W. Frees, and M. A. Rosenberg (2008, April). Heavy-tailed longitudinal data modeling using copulas. *Insurance: Math. and Econom. 42*(2), 817–830.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica 26*(1), 24–36.

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

Tsiamyrtzis, P. and D. Karlis (2004). Strategies for efficient computation of multivariate Poisson probabilities. *Commun. Stat., Simulation Comput. 33*(2), 271–292.

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. In J. K. Ghosh and J. Roy (Eds.), *Statistics: Applications and New Directions*. Calcutta: Indian Statistical Institute: Proceedings of the Indian Statistical Institute Golden Jubilee International Conference.

Venter, G. G. (2001). Tails of copulas. In *Proceed. ASTIN Washington, USA*, pp. 68–113.

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

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

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

Wagner, S. M. (2006b). *Economic Analyses of the European Patent System*, Volume 1. Deutscher Univ.verlag.

Winkelmann, R. (2008). *Econometric Analysis of Count Data* (5th ed.). Berlin: Springer-Verlag.

Yan, J. (2002). geepack: Yet another package for generalized estimating equations. *R-News 2/3*, 12–14.

Yan, J. and J. P. Fine (2004). Estimating equations for association structures. *Stat. in Med. 23*, 859–880.

Yi, N., S. Banerjee, D. Pomp, and B. Yandell (2007). Bayesian mapping of genomewide interacting quantitative trait loci for ordinal traits. *Genet. 176*(3), 1855–1864.

Yi, N., S. Xu, V. George, and D. Allison (2004). Mapping multiple quantitative trait loci for complex ordinal traits. *Behav. Genet. 34*, 3–15.

Yip, K. C. and K. K. Yau (2005). On modeling claim frequency data in general insurance with extra zeros. *Insurance: Math. and Econom. 36*(2), 153–163.

Zak, M., A. Baierl, M. Bogdan, and A. Futschik (2007). Locating multiple interacting quantitative trait loci using rank-based model selection. *Genet. 176*(3), 1845–1854.

Zak-Szatkowska, M. and M. Bogdan (2010). Applying generalized linear models for identifying important factors in large data bases. Technical Report I-18/2010/P-001, Inst. of Math. and Comp. Sci., Wroclaw University of Technology, *www.im.pwr.wroc.pl/∼mbogdan/papers*.

Zeng, Z. B. (1993). Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. *Proc. Natl. Acad. Sci. USA 90*, 10972–10976.

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genet. 136*(4), 1457–1468.

Zimmer, D. M. and P. K. Trivedi (2006, January). Using Trivariate Copulas to Model Sample Selection and Treatment Effects: Application to Family Health Care Demand. *J. Bus. Econ. Statist. 24*, 63–76.

Zou, F., B. Yandell, and J. Fine (2003). Rank based statistical methodologies for QTL mapping. *Genet. 165*(3), 1599–1605.

Zuur, A. F., E. N. Leno, N. Walker, A. Saveliev, and G. M. Smith (2009). *Mixed effects models and extensions in ecology with R (in: Stat. Biol. Health)*. Springer New York.