PhD Dissertation

Use of Multivariate Generalised Linear Mixed Models in Life and Environmental Sciences

Jeanett Snitgaard Pelck

Department of Mathematics
Aarhus University
July, 2021

AARHUS UNIVERSITY

Use of Multivariate Generalised Linear Mixed Models in Life and Environmental Sciences

PhD Dissertation by
Jeanett Snitgaard Pelck

Department of Mathematics, Aarhus University
Ny Munkegade 118, 8000 Aarhus C, Denmark

Supervised by
Senior Researcher Rodrigo Labouriau, Aarhus University

Submitted to the Graduated School of Natural Science, Aarhus University, July 31, 2021

Preface

This dissertation is a result of my PhD studies in the period from August 1, 2018 to July 31, 2021 at the Laboratory for Applied Statistics at the Department of Mathematics, Aarhus University. All the work in this dissertation were supervised by Senior Researcher (with special qualification, MSK) Rodrigo Labouriau. The dissertation is a collection of six papers. Paper V is included as a draft whereas the remaining papers are included in their published format apart from layout and minor insignificant adjustments. The papers can be read independently but contains natural links to each other. The papers included in the thesis are:

Paper I Pelck, Labouriau: *Conditional Inference for Multivariate Generalized Linear Mixed Models.*

Paper II Pelck, Labouriau: *Multivariate Generalised Linear Mixed Models With Graphical Latent Covariance Structure.*

Paper III Pelck, Labouriau: *Using Multivariate Generalised Linear Mixed Models for Studying Roots Development.*

Paper IV Pelck, Maia, Pinhero & Labouriau: *A Multivariate Methodology for Analysing Students' Performance Using Register Data.*

Paper V Pelck, Labouriau: *Simultaneously Analysis of Time to Emergence of Different Weed Species.*

Paper VI Pelck, Luca, Holthusen, Edelenbos & Labouriau: *Multivariate Method for Detection of Rubbery Rot in Storage Apples by Monitoring Volatile Organic Compounds: An Example of Multivariate Generalised Linear Mixed Models.*

Parts of the material in Paper I, II and III were included in my progress report used for the qualifying exam that was held January 31, 2020, in accordance with the Graduate School of Natural Sciences rules (GSNS). The material used in the progress report was further developed with new ideas and results after the qualifying exam, which form the basis of the papers. Paper IV–VI are primarily a result of the last part of my studies after the qualifying exam. I have contributed comprehensively in both the writing as well as the research phase of all the papers.

This dissertation consists of an introductory section and six self-contained papers. The purpose of the introduction is threefold. Firstly, it provides a review section presenting an overview of relevant parts of the literature related to the content of the dissertation. Secondly, it contains one introductory section for each paper including a description of the research questions, the applied methodology and

summary of the results. Thirdly, it gives a short description of other work done during my PhD studies, not included in the dissertation.

Paper I and II are theoretical papers presenting a new inference method for an extended class of multivariate generalised linear mixed models with random intercepts, and a method for combining the theory of graphical models together with a statistical test to draw conclusions regarding the latent covariance structure in this class of models. Paper III–VI are all applications of the methods in Paper I and II, applied in very different contexts showing the high flexibility of the presented methodologies.

My three years of PhD studies have been an instructive experience both personally and academically but also challenging. I owe several people huge thanks for helping, guiding and supporting me through this journey. First of all, I would like to thank my supervisor Rodrigo Labouriau for giving me the opportunity of pursuing a PhD degree, and for his huge support throughout the studies. I feel honoured that Rodrigo has invested so much time and effort in me, which by far exceeds what could be expected from a supervisor. His high ambitions and trust in my abilities has definitely made me a better researcher.

I would also like to thank my co-authors Hildete P. Pinheiro and Rafael Pimentel Maia from the Department of Statistics, University of Campinas, Brazil. During Rafael's visit to the Laboratory for Applied Statistics, Aarhus University, in January 2020, we planned that I would visit the University of Campinas in June 2020, where we all would collaborate on an analysis of students' performance at the University of Campinas using the at that time developed parts of the methods described in Paper I and II. Unfortunately, the COVID-19 pandemic made this visit impossible, but instead, we managed to do the collaboration online which resulted in Paper IV. I am deeply grateful for their interest in my research and their willingness to adapt to the impossible situation we all found ourselves in. I hope that I will get the opportunity to visit them at the University of Campinas at some point. I would also like to thank my co-authors Hinrich H.F. Holthusen, Merete Edelenbos, Alexandru Luca from the department of FOOD Science, Aarhus university for an interesting collaboration on Paper I.

I thank as well my colleagues at the Department of Mathematics, Aarhus University, for creating a pleasant work environment. A particular thanks goes to my office mates during the first part of my PhD studies, Svend Vendelbo Nielsen and Johanna Bertl, for all the small talks and the friendly working atmosphere. Moreover, I would like to thank the local LaTeX expert Lars 'daleif' Madsen for helping me with the technical typesetting. I would also like to thank my friends Majka Cilleborg Bilde and Simone Fredin Mikkelsen for all the nice breakfasts, lunches and cake meetings during our time at the university. Due to the COVID-19 pandemic, the second part of my PhD studies was primarily spent working from home with my dog, Charlie, as office mate and sometimes also my husband. I would like to thank both of you for the nice company, and also, my friends Katrine Bødkergaard Nielsen and Kathrine Marie Graversen for the enjoyable (virtual) lunches.

Finally, my family and friends deserve a deep gratitude for their indispensable support and love. I am deeply grateful to my husband who always believe in me and encourage me in everything I do.

Jeanett Snitgaard Pelck
July, 2021

Abstract

This dissertation develops aspects of multivariate generalised linear mixed models including new methodologies for inference and analysis of the latent covariance structure. It is based on six self-contained papers, where the first two papers are methodology papers and the remaining papers are applications of the developed methods. These applications illustrate the great flexibility of the developed methods and the wide range of applications arising from very different fields of research.

Generalised linear mixed models (GLMMs) offer a flexible system of statistical models that can represent dependence in the data. These types of models include latent random variables which define the dependency structure in the model. Typically, these latent random variables are assumed to follow Gaussian distributions but in Paper I and II, we allow for other distributions satisfying minimal requirements. Moreover, the papers formulate multivariate versions of the univariate models defined in literature. These multivariate models apply to responses of very different nature, *e.g.,* discrete and continuous variables as in Paper VI.

First, in Paper I, we present an inference method to be used in (multivariate) GLMMs with random intercepts. This inference method differs from existing inference methods by allowing more general distributional assumptions, both for the responses and the random components. Next, in Paper II, we present a method to analyse the latent covariance structure that is present in the class of multivariate GLMMs introduced in Paper I, but with the restriction that the random components follow a multivariate elliptical contoured distribution. Moreover, this method allows us to characterise the correlation structure among the responses under the model. Paper III – VI are applications of the methodologies presented in Paper I and II, each contributing with new aspects of multivariate GLMMs. These applications derive from various fields of research and illustrate the high flexibility of the developed methods. Paper III characterises simultaneously the spatial and temporal distribution of the root system in a cultivated field using a multivariate GLMM. Paper IV presents a method for jointly modelling students' results in the university's admission exams and their performance in subsequent courses at the university. This multivariate analysis includes a marginal frailty discrete-time Cox proportional model combined with standard marginal Gaussian mixed models. Paper V consists of an analysis of a weed emergence follow up study, where the time to emergence of each propagule present in the soil for different species of weed were modelled simultaneously using a multivariate piecewise constant frailty Cox proportional model inferred by estimating a suitable multivariate GLMM on a specially constructed dataset. Paper VI describes a multivariate GLMM for screening potential chemical markers for early detection of post-harvest disease in storage fruit.

Resumé

Denne afhandling omhandler aspekter af multivariate generaliserede lineære mixede modeller, herunder en ny inferensmetode samt en metode til at analysere den latente kovariansstruktur, der er til stede i disse modeller. Afhandlingen er baseret på seks selvstændige artikler, hvor de to første artikler er metodiske artikler, mens de resterende artikler er anvendelser af de udviklede metoder. Anvendelserne illustrerer den høje fleksibilitet af de udviklede metoder samt de mange anvendelsesmuligheder inden for meget forskellige forskningsområder.

Generaliserede lineære mixede modeller (GLMMer) er statistiske modeller, der tilbyder en høj fleksibilitet samt giver mulighed for at repræsentere afhængige data. Denne type modeller inkluderer latente tilfældige variabler der typisk antages at følge en Gaussisk fordeling, men i Artikel I and II tillader vi andre fordelinger, der opfylder nogle minimale betingelser. Artiklerne definerer ydermere multivariate udgaver af de univariate modeller, der er defineret i litteraturen, som kan anvendes til at analysere responser af meget forskellig karakter, f.eks. diskrete og kontinuerte variabler som i Artikel VI.

Først introducerer vi en inferensmetode for (multivariate) GLMMer med tilfældig skæringskoefficienten i Artikel I. Denne inferensmetode adskiller sig fra eksisterende metoder ved at tillade mere generelle fordelingsantagelser, både for responsevariablerne og de tilfældige komponenter. Dernæst præsenterer vi i Artikel II en metode til at analysere den latente kovariansstruktur, der er tilstede i klassen af multivariate GLMMer defineret i Artikel I. Denne metode giver desuden mulighed for at karakterisere afhængighedsstrukturen mellem responserne i modellen. Artikel III–VI indeholder anvendelser af metoderne præsenteret i Artikel I og II, der hver især bidrager med nye aspekter inden for multivariate GLMMer. Disse anvendelser, der stammer fra meget forskellige forskingsområder, illustrerer den høje fleksibilitet af de udviklede metoder. Artikel III karakteriserer den rumlige og temporale fordeling af et rodsystem i en dyrket mark på samme tid ved at anvende en multivariat GLMM. Artikel IV præsenterer en metode til samtidig at modellere de studerendes resultater i adgangsprøver fra universitetet samt deres præstation i et efterfølgende kursus på universitet. Denne multivariate analyse inkluderer diskrete marginale frailty Cox proportionale modeller kombineret med standard marginale Gaussiske mixed modeller. Artikel V indeholder en analyse af et follow-up studie af forekomst af forskellige arter af ukrudt, hvor tiden til tilsynekomst for hvert frø tilstede i jorden modelleres simultant for hver art ved at anvende en multivariat stykvis konstant frailty cox proportionel model. Artikel VI beskriver en multivariat GLMM til at screene for potentielle kemiske markører for tidlig opdagelse af sygdom i lagret frugt efter høst.

Contents

Introduction

This introduction consists of three parts. First, it includes a review section, relating the included papers to the existing literature within the research field. Second, it presents several introductory sections, one for each paper included in the dissertation. Each introductory section describes the proposed research questions in the corresponding paper and summarises the key results. Thirdly, the introduction discusses other applications related to the content of the papers included in the dissertation, and shortly describes the developed R-package implementing the developed methods. This R-package was used for inference in the models presented in this dissertation.

All papers in this dissertation treat aspects of multivariate generalised linear mixed models which are of particular interest for applications in areas where we often find examples of several simultaneously observed characteristics requiring multivariate statistical models. The framework of generalised linear mixed models is very applicable in analyses with dependent observations. Therefore, many authors have worked on aspects of these models. However, in literature it is often only univariate models that are described and to the best of my knowledge there exist no methods connecting the theory of multivariate generalised linear mixed models with the theory of graphical models as presented in Paper II. Likewise, we are not aware of any inference methods equivalent to the introduced inference method in Paper I, however, there exists inference methods in the literature that are related.

Relating Developed Methods to Literature

The historical development of generalised linear mixed models can briefly be outlined by the following papers. Univariate generalised linear mixed models (GLMMs) were introduced in Breslow & Clayton (1993), where random components, representing unobservable random variables, were added to the linear predictor defined in generalised linear models as a scalar product of a vector of coefficients and a vector of explanatory variables (Nelder & Wedderburn (1972), McCullagh & Nelder (1989)). Generalised linear models are extensions of the linear models allowing for non-Gaussian distributions and non-linear link functions relating the linear predictors to the expectations of the responses under the model. In this class of models it is assumed that the responses follow an exponential dispersion model (Tweedie (1984), Jørgensen (1987)).

The model assumptions in a GLMM include distributional assumptions for the conditional distributions of the responses given the random components, typically exponential dispersion models with a link function connecting the conditional expectations to the linear predictors. Moreover, we assume Gaussian distributions for the random components (in the multivariate case, this is multivariate Gaussian distributions). In Paper I, we extend these assumptions by letting the conditional

distributions follow dispersion models (Jorgensen (1997), Jørgensen & Lauritzen (2000), Artes & Jørgensen (2000), Jørgensen & Labouriau (2012), Barndorff-Nielsen (2014), Labouriau (2020), Cordeiro et al. (2021)), and by letting the distributions of the random components belong to a larger family of distributions including for example the Gaussian and t-distribution. We call this model a (multivariate) extended (GLMM). In Paper II, the multivariate distribution of the random components are assumed to belong the the family of elliptically contoured distributions (Anderson 2003).

The likelihood function of GLMMs involves an integral integrating the conditional densities with respect to the distribution of the random components. Often, this integral cannot be evaluated in closed form. Therefore, Breslow & Clayton (1993) presents a Laplace approximation of the likelihood function for a GLMM in the case of a univariate model as a way of avoiding integration of conditional likelihood quantities. In McCulloch (1997) several ideas on maximum likelihood algorithms for inference in GLMMs were exposed. These algorithms are based on Monte-Carlo simulation either as a part of an EM algorithm or combined with a Newton-Raphson maximisation. Later, various research have originated from these historical papers, see for example Booth & Hobert (1999) and the references related to h-likelihood below. Inference in a GLMM can also be done using Gauss-Hermite quadrature approximation, restricted maximum likelihood (also multivariate), the h-likelihood described below and many others (see McCulloch & Searle (2001), Berridge & Crouchley (2011), Demidenko (2004) and references therein). There exists several implementations for different software programs, *e.g.,* the *lme4* package in R (Bates et al. (2015)) and the *GLIMMIX* procedure in SAS (SAS Institute Inc. (2017))

The inference method introduced in Paper I is related to the ideas behind h-likelihood (Lee & Nelder (1996), Lee & Nelder (2001), Lee & Nelder (2005), (Nelder et al. 2006)). The two inference methods share some of the same ideas, namely the idea to perform inference in a family of probability measures treating the values of the unobservable random variables as parameters to be estimated along with the other parameters in the model. The two inference procedures uses different families of probability measure to perform inference. Moreover, h-likelihood includes a change of scale of the random components that the conditional inference method do not. After the introduction of h-likelihood, several papers debating and further developing the theory of h-likelihood have been published. In the discussion part of Nelder et al. (2006) and in Lee et al. (2007), the authors of h-likelihood answer to some of the criticism there have been made of the theory of h-likelihood, specially when working with binary data. Meng (2009) discusses the developed theory of h-likelihood with a critical Bayesian approach.

Graphical models (Pearl (1988), Whittaker (1990), Lauritzen (1996), Edwards (2000) and Perl (2009)) are a combination between probability and graph theory which provide an intuitively way to model and understand the (conditional) dependence structure in a joint probability distribution. Abreu et al. (2010) and Edwards et al. (2010) show how graphical models can be inferred by minimising the AIC (Akaike information criterion) or BIC (Bayesian information criterion). The theory of graphical models can be combined with the methodology introduced in Paper I

to obtain a graphical model for the latent variables under a multivariate GLMM as described in Paper II. Below, we briefly discuss other multivariate approaches not directly connected to GLMM. These models have a very different structure and cannot simultaneously model responses of different nature requiring different marginal distributions which is the case in the applications in Paper III, IV and VI. In Jørgensen (2013) and Jørgensen & Lauritzen (2000), methods for constructing multivariate dispersion models are presented expanding the ideas of the univariate dispersion models (Jørgensen (1987) and Jorgensen (1997)). These models are an alternative to multivariate GLMMs which offer another structure but looses the marginal distributional assumptions for the responses which exists under the multivariate GLMMs. Therefore, we can not simultaneously model responses with different supports. Xue-Kun Song (2000) introduces a multivariate dispersion model generated from a Gaussian Copula. This is an extension of the multivariate dispersion model described in Jørgensen (1987) which possesses the property that the marginal distributions are closed. However, this cannot be directly combined with the methods presented in this dissertation because the distribution of the random components changes after applying the copula. Other multivariate models using copulas that are worth mentioning are Mikosch (2005), Song et al. (2009), Krupskii & Joe (2013) and Krupskii & Joe (2015). In Bonat & Jørgensen (2016) a framework called multivariate covariance generalised linear models for non-normal multivariate data analysis is presented. These models separate from other types of multivariate modelling frameworks by introducing a covariance link function combined with a matrix linear predictor involving known matrices.

Paper I

In this paper, we present an inference method for an extended class of (multivariate) GLMMs with random intercepts. In literature, it is often assumed that the random components in a GLMM are normally distributed with expectation zero and an unknown variance. In the class of models defined in this paper, we allow for other distributions for the random components as long as they satisfy some minimal requirements. Examples of distributions satisfying these requirements are the normal and the t-distribution. Another property of this class of models is the assumption regarding the conditional distributions of the responses given the random components. Usually in literature, these distributions are assumed to follow exponential dispersion models but in this class, we allow general dispersion models that satisfy some regularity assumptions. Thus, using the conditional inference method it is possible to perform inference in a class of (multivariate) GLMMs, and avoid the integration that enters the likelihood functions of these models, which often cannot be evaluated in closed form.

The conditional inference method is introduced for univariate models and then generalised to the multivariate case. This step is quite simple because of the structure of the inference method. For this reason, the paper fist describes the inference method for a one dimensional model followed by an extension to the multivariate case. The

reasoning behind the presented inference method can be explained by considering two families of probability measures denoted by $\mathcal{P}$ and $\mathcal{P}^*$ defined below. $\mathcal{P}^*$ contains the conditional distributions under the model, where the values of the unobservable random components are treated as parameters, denoted by $\boldsymbol{b}$, along with the fixed effects denoted by $\boldsymbol{\beta}$ and the dispersion parameter, λ. On the other hand, the family $\mathcal{P}$ contains all marginal distributions of the responses, that is, distributions including integration of conditional densities. These integrals are non-trivial which often cannot be evaluated in closed form as mentioned above.

We introduce inference functions for $\boldsymbol{\beta}$ and $\boldsymbol{b}$ that are equivalent to the score equations of the probability measures in $\mathcal{P}^*$. Moreover, we define an inference function for $\boldsymbol{\beta}$ under $\mathcal{P}$ as a function of the predicted values of the random components obtained as the roots of the inference functions under $\mathcal{P}^*$. The dispersion parameter is treated as a nuisance parameter in these inference functions, and the variance parameter can be estimated based on the predicted values of the random components. After finding the roots of the defined inference functions, the predicted values of the random components are projected onto the subspace of all vectors with mean zero to ensure an identifiable parametrisation.

In the paper, we show that the defined inference functions under $\mathcal{P}^*$ are regular. Moreover, we show that the sequence of roots (when the number of observations increases) of the inference functions are consistent under $\mathcal{P}^*$ and conditionally asymptotically Gaussian distributed under some regularity conditions. Moreover, we prove, under some regularity conditions, that the sequence of roots of the inference function under $\mathcal{P}$ are consistent and asymptotically Gaussian for small variances of the random components.Thus, the presented inference method preserve some of the desirable properties of classical likelihood-based inference methods.

A multivariate model is formulated by assuming a joint distribution of the random components associated with the same experimental unit across the different marginal models. The parameters of this distribution can be inferred using the predicted values of the random components obtained by estimating each marginal model separately.

We performed two simulation studies in which we simulated a two dimensional GLMM. Here, we study the distribution of the fixed effects for three different covariance matrices obtained by increasing the values of each entry by multiplying the matrix with different constants. Furthermore, we study the bias of the estimated parameters and compare with other inference methods for three different numbers of random components and observations. We concluded, that in this simulation study, the conditional inference method performed equally as good as other known inference methods, and that the multiplying constant had to be relatively high in order to lead to non-Gaussian distributed estimates.

Paper II

This paper introduces a method for representing the latent covariance structure in a multivariate GLMM via graphical models. A class of multivariate GLMMs, as defined in Paper I, is introduced with the additional assumption that the random

components follow a multivariate elliptically contoured distribution. This class of models is combined with the theory of graphical models by inferring a graphical model based on the predicted values of the random components under the model. Predictions of the random components can be obtained using the inference procedure presented in Paper I or another inference method, *e.g.,* the Laplace approximation described in the appendix of Paper I and in Breslow & Clayton (1993).

We present two types of graphical models both containing a set of vertices and a set of edges. An undirected graphical model satisfies that two vertices are connected by an edge if, and only if, they are not conditional independent given the remaining vertices in the model. In a directed acyclic graph the definition is the same but with the conditioning set modified. In this graph, the edges have arrows in the ends indicating which variable that caries information on another. Therefore, we only need to condition on the vertices that carries information on the two vertices in question either directly or indirectly through other vertices in the graph.

In an undirected graph, we say that there exists a path connecting two vertices if there exists a sequence of vertices, connected by edges, connecting the two vertices. Moreover, we say that a set of vertices, S, separates two disjoint subsets of vertices A and B in the graph when every path connecting a vertex in A to a vertex in B necessarily contains a vertex in S. An undirected graphical model satisfies the *separation principle*, which states that if a set of vertices S, separates two disjoint subsets of vertices A and B in the graph, then all variables in A are conditionally independent of all variables in B given S.

In the paper we define a combination of the two types of graphs which allow us to study the dependence structure among the random components and the responses. In the case of Gaussian distributed random components, we can interpret the graphical representation in terms of independence. However, when the random components are elliptically contoured distributed, independence should be interpret in terms of covariances equal to zero. In the paper, we formulate a principle, called the *induced separation principle*, stating that if A, B and S are subsets of random components satisfying the separation principle given above, then the responses corresponding to the random components in A are conditionally independent of the responses corresponding to the random components in B given S.

The graphical model can be inferred using adaptions of statistical tests presented in Anderson (2003). Here, an exact test of independence between sets of variables in a multivariate Gaussian distribution is presented. This test is adapted to the multivariate GLMM, introduced in the paper, in the case of Gaussian random components. When the random components are not normally distributed, Anderson (2003) describes an asymptotic test for (conditional) uncorrelation between sets of variables in an elliptically contoured distribution. An adaption of this test is also presented in the paper. Using the tests described above, we can construct a graphical model representing the covariance structure of the random components by testing if the conditional covariance between each pair of vertices is equal to zero given the remaining vertices (possibly correcting for multiple comparisons). Moreover, these tests allow us to test for covariances equal to zero for multiple sets of random components using only one test, and thus, we can reduce the number of tests if we

test for a specific graphical structure.

The statistical tests described above are based on an estimate of the covariance matrix. Under a model with Gaussian random components, the tests described above are only exact when using an estimator proportional to the maximum likelihood estimate. Using a consistent estimator of the covariance matrix yields in both the Gaussian and the elliptically contoured case an asymptotic test.

The paper includes a simulation study to examine the power of the tests as a function of the sample size in a four dimensional Gaussian- and t-distribution. As expected, we need more observations in a multivariate t-distribution than for a multivariate Gaussian distribution in order to reach the correct significance level. Furthermore, the paper includes another simulation study examining the power of the two tests when the off-diagonal entries in the covariance matrix in a two dimensional extended GLMM are varied. We simulate two models, one where the random components follow a multivariate Gaussian distribution, and another where the random components follow a multivariate t-distribution. In both models it is assumed that the conditional distributions are Gamma and Poisson distributed. This study indicated that when the random components are Gaussian distributed, the power curve for the test based on normality is steeper than the curve for the elliptical test when the off-diagonal values are close to zero. However, when the random components are multivariate t-distributed, the normality test rejects too often under the null hypothesis compared to the elliptical test.

Paper III

This paper presents a multivariate GLMM for studying roots' development based on minirhizotron observations, that is, observations obtained by inserting a special tube (minirhizotron) in the soil, where roots can be counted in special observation windows using a camera. The models described in this paper can be applied, with minor adaptions, in many experiments inside this field of research, *e.g.,* in Shanmugam et al. (2021) which was a result of a collaboration arranged through the Laboratory for Applied Statistic, Aarhus University. Here, Rodrigo Labouriau and I did the statistical analysis using models similar to the univariate models presented in the paper summarised in this Section.

The introduced multivariate model is combined with the theory of graphical models and the methods described in Paper II to characterise the dependence structure between the root scatter and intensity (described below) over three observed development stages of the culture in the field. The root scatter is a measure of the soil volume occupied by the root system in the field, whereas the root intensity is a measure of the root colonisation in the field. The root scatter is measured by counting the number of windows with a root present in each depth zone in each tube. The root intensity is measured by counting the number of roots crossing a reference line in each observation window in the minirhozotrons. The number of crossings can be used to obtain estimates of the length of the root system using a stochastic geometric argument.

A six dimensional model is formulated modelling the root scatter and root intensity at three different days (corresponding to the different development stages of the plants). The models for the root scatter at the three different observation days are all instances of GLMMs with a binomial distribution and a logit link function. The structure of the three models are assumed to be identical for each development stage, and therefore, we describe the model for a fixed development stage. Each model includes a random component taking the same value for all observations arising from the same tube. The responses consists of counts of the number of observation windows with a root present for the development stage in question for each combination of depth zone and tube. For each response, the depth zone and fertilisation of the observation are included in the model (as fixed effects) to adjust for the expected differences due to the different fertilisations and the soil depth zones. According to the model, the presence or not of roots are conditionally independent and Binomial distributed, given the random components, with probability parameters depending on the fixed effects and the given value of the random component representing the tube of the observation.

The root intensity at the three different observation days are modelled using GLMMs with a Poisson distribution and a logarithm link function. We only describe the model for one of the observation days since we assume that the structure of the models are identical. Analogues to the model for the root scatter, random components taking the same value for all observations arising from the same tube are included in the model. According to the model, the counts are conditionally independent and Poisson distributed given the random components with a conditional expectation depending on the fixed effects (soil depth zone and fertilisation), the given value of the random component representing the tube of the observation and an offset corresponding to the number of observational windows present in the tube and depth zone of the observation. The included offset plays an importing role in the interpretation of the expectation under the model. Because of the logarithm link function, we obtain a model were we can write the expectation as the mean number of crossings per window, which according to a stochastic geometric argument is proportional to the length of the roots that are visible in the observation windows.

Each marginal model is formulated as a GLMM with a random component taking the same value for all observations arising from the same tube. A multivariate model is constructed by assuming a multivariate Gaussian distribution, with expectation zero, of the random components representing the same tube in the six marginal models (the distributions of the different tubes are independent and identical). Observations from different tubes are assumed to be independent under the model. The multivariate model was inferred using the inference method described in Paper I. Moreover, the latent covariance structure was analysed as a graphical model by minimising the BIC (Bayesian Information Criterion) of a graphical model based on the predicted values of the random components and interpret using the methodology presented in Paper II.

The inferred graphical model indicates a first-order Markovianity dependency pattern between the random components related to models for the root intensity and scatter at the three development stages. This means that the random components

related to the models for the root intensity and scatter at the first and third observational days are conditionally independent given the random components associated to the models for the second observation day. At each observation day, we found that the random components related to the root scatter and intensity are positively correlated after conditioning on the random components included in the models for the other days. Therefore, these random components carry information on each other that is not contained in the other random components. Thus, there is evidence of specific common or cooperative underlying mechanisms determining the root scatter and intensity, specific for each day (development stage). This result rules out the possibility that the plants would be compensating a reduced occupation of the soil by increasing the intensity of the root system.

Using the extended separation theorem stated in Pelck & Labouriau (2021b), we were able to draw conclusions regarding the dependence structure of the responses based on the graphical model for the random components. The theorem states that if two sets of random components are conditional independent given a separating set, then the associated sets of responses are conditional independent given the same separating set (consisting of random components). For the inferred graphical model in this paper, this result implies that the root intensity and root scatter at the first development stage is conditional independent of the root intensity and root scatter at the third development stage given the random components associated to the models for the second development stage.

Paper IV

In this paper, a multivariate model that allows simultaneously analysis of the results in the university's admission exams and the performance in subsequent courses is presented. The introduced model is based on an example of data containing results of students enrolled at the University of Campinas, Brazil, in 2014 to evening studies programs in educational branches related to exact science. For these students, the results of seven admission exams together with the performance in the university course Geometry are analysed to characterise the information that the results of the admission exams carry on the performance in Geometry.

An affirmative action program implemented at the university gives students who went to a public high school (for all high school years) a bonus added to their final scores in the admission exams. Moreover, students who were self declared African or Indigenous Brazilian descendants received an additional bonus. We would expect, that these two different groups of students (either receiving a bonus or not) would present very different patterns, and therefore, the analysis were stratified into two separate analyses for the two groups of students.

The models used in the two analyses are similarly defined, and thus, we only specify the model for the students that received the bonus. We define a multivariate model with eight marginal models of which seven are marginal linear mixed models (Gaussian distributions), with identity link functions. The eighth marginal model describes the number of attempts needed to pass Geometry. This is modelled using a

frailty discrete-time Cox proportional model which can be represented using a GLMM, with a Binomial distribution and logarithm link function, on a transformed dataset. All models include two random components that accounted for the variation between different study branches and individual variation. Moreover, all models include fixed effects correcting for age and gender. In the model for the number of attempts to pass Geometry, we further add a fixed effect to the linear predictor, taking the same value for each number of trial possible. This implies, that the conditional hazard is formulated as a product of a baseline function, taking the same value for each number of attempts, and the exponential to the sum of the fixed effects and the given values of the random components.

The marginal models are combined into a multivariate model by assuming a joint multivariate Gaussian distribution for the random components representing the individuals. We assume that the random components representing different individuals are mutually independent, and independent of the random components representing the variation between the different study branches. Moreover, it is assumed that the random components representing the study branches are mutually independent. The multivariate models (one for each of the two populations of students analysed) were inferred using the methodology described in Paper I. The predicted values of the random components representing the individuals were used to infer graphical models by minimising the BIC for each analysis. The resulting graphical models were interpret using the methodology presented in Paper II. A summary of the results is given below.

The covariance structure of the random components representing individuals' variation after adjusting for differences in age, gender and educational branch was analysed for each of the two populations of students. We were able to draw conclusions regarding the dependence structure among the responses in the model by applying the induced separation principle stated in Paper II. The results are discussed shortly below. For a student receiving bonus, we found that the performance in Geometry is conditionally independent of the results in all admission exams except Mathematics, given the value of the random component corresponding to the entrance scores in Mathematics. Moreover, the conditional correlation between the random components corresponding to the performance in Geometry and the random components included in the model for the admission scores in Mathematics is positive, given the random components corresponding to the other responses. This result implies that there exist some characteristics or abilities that are individual to each student which are common to the result in the admission exam in Mathematics and the performance in Geometry but not shared by the other disciplines. Furthermore, this result shows that the prediction of the random components corresponding to the results in the admission exam in Mathematics suffices for predicting the performance in Geometry.

We obtain a different covariance structure for a student not receiving any bonus. Here, the performance in Geometry is conditional independent of the results in all the admission exams except Portuguese, given the predictions of the random components associated with the model for the results in the Portuguese admission exam. Thus, the interpretation of the results for a student not receiving the bonus is the same as for a student receiving a bonus just with Portuguese replacing Mathematics.

Comparing the results of the analyses for the two groups of students might indicate that among students receiving a bonus, the quality of their high school education in Mathematics has a much higher variability than in the group of students not receiving the bonus. Therefore, the result in the admission exam in mathematics has a stronger influence on the performance in Geometry for this group of students. On the other hand, one might speculate whether the results in the admission exam in Portuguese reflects the social-economic class of the students which play a key role in the performance in Geometry for the students not receiving the bonus.

Paper V

In this paper, we illustrate a multivariate method for analysing times to emergence of different species of weed. The times to emergence of the propagules present in the soil are modelled using a multivariate piecewise constant frailty cox proportional model with marginal models representing the species. A simultaneous modelling will allow us to make comparisons between the species and study the latent covariance structure of random components representing local characteristics of the locations in the field.

In each marginal model, we model the time to emergence for each propagule present in the soil given a random component representing the experimental unit of the propagule (the observation ring in the field), *i.e.,* each random component takes the same value for all observations of the same species arising from the same ring. A conditional hazard function for the time to emergence of each propagule, given the random components, is formulated in the paper. According to the model, the conditional hazard function is a product of a piecewise constant baseline function of time and the exponential of the given value of the random component representing the observation ring of the propagule in question. The piecewise constant baseline functions are assumed to be constant on intervals between different observation days.

The marginal models are combined into a multivariate model by assuming a joint multivariate Gaussian distribution for the random components. Under this model, the random components representing the different observation rings are independent. The likelihood function for this model coincides with the likelihood function for a GLMM applied to a transformed dataset. Therefore, the multivariate model was inferred using the inference method described in Paper I. Based on the predictions of the random components, a graphical model was estimated using the techniques described in Paper II.

The inferred graphical model showed that the random components associated with two of the species were isolated in the graph implying that they are independent of the other random components. Moreover, by the separation theorem stated in Paper II, the random components connected to three out of the remaining four species are conditional independent given the random components related to the fourth species. Therefore, the induced separation principle, presented in Paper II, indicates that that the times to emergence of two of the species are mutual independent, and independent from the times to emergence of the remaining species. Moreover,

knowing the values of the random components associated with the model for one specific species renders the times to emergence of all other species independent.

In the appendix of this paper we discuss the counting process related to defined survival model and present details on the coincidence of the likelihood function with the likelihood function of a multivariate GLMM applied to a constructed dataset. Moreover, we present model control of the inferred model.

Paper VI

This paper describes the non-standard statistical method applied in (Holthusen et al. 2021a). Infection of fruit in storage leads to serious losses in commercial production. Therefore, the analysis in this paper studies predictors of rubbery rot at an early stage of the infection development. We construct a multivariate method for screening potential chemical markers for early detection of post-harvest disease in storage fruit based on an example on detection of rubbery rot caused by *Phacidiopycnis washingtonensis* in apples.

In this experiment, a range of volatile organic compounds (VOCs) are measured simultaneously together with two measures of severity of disease infection; the number of fruit presenting visible symptoms and the lesion area. The analysed data contains observations from 10 glass jars (denoted by glasses below), each containing nine inoculated apples observed at three different observation times (given as the number of weeks post-inoculation). For each glass at each observation time, the concentrations of the VOCs were observed together with the number of infected apples out of the nine apples contained in the glass. Moreover, the total lesion area of the infected apples in each glass was observed (zero if none of the apples in the glass were infected).

We formulate a 16-dimensional multivariate GLMM representing concentrations of 14 VOCs and the two measures of severity of infection. All 16 marginal models are formulated as GLMMs including random components taking the same value for observations arising from the same glass in each marginal model. By including these random components, we take into account that observations within the same glass at different times might be correlated.

The marginal models for the VOCs are formulated as GLMMs with a Gamma distribution and logarithm link function modelling the positive VOC concentrations in each glass at each time. The marginal model representing the observed number of apples presenting symptoms in each glass is a GLMM with a binomial distribution (size 9) and a logit link function. The observation regarding the lesion area of infection in each glass is positive when at least one apple is infected and zero otherwise. Therefore, the marginal model representing these observations is formulated as a GLMM with a Gamma Compound Poisson distribution and logarithm link function, putting mass on zero and otherwise being a continuous positive distribution. The marginal models are combined into a multivariate model by assuming a multivariate Gaussian distribution, with expectation zero, of the random components representing the same glass in the different the marginal models (the distributions for the different glasses are independent and identical).

A graphical model based on the predicted values of the random components was inferred by minimising the BIC (Baysian Information Criterion). The predicted values of the random components were estimated using the Laplace approximation of the likelihood function. Based on the resulting graphical representation, we concluded that the random components associated with four of the VOC's carry all the information that the random components included in the models for the fourteen VOCs might hold on the random components included in the models for the two measures of severity of the fungal infection. According to the model and the induced separation principle stated in Paper I, this imply that knowing the values of the four random components referred above (corresponding to the responses anisole, 3-pentanone, 2-methyl-1-propanol and 2-phenylethanol) renders the measures of infection independent of the remaining VOCs.

Additional Work

During my PhD studies I have worked on other projects related to the content in the papers in this dissertation. As a part of my work for the Laboratory for Applied Statistics, Aarhus University, I analysed in collaboration with Rodrigo Labouriau an experiment regarding measure of physiological effects of sugar beet seed priming on different developmental stages (Salimi et al. (n.d.)). A part of these data contained right censored observations which were accounted for in the analysis using a frailty cox proportional model inferred by estimating a GLMM on a transformed dataset using the inference method discussed in Paper I.

Furthermore, I analysed in collaboration with Rodrigo Labouriau two experiments regarding local root intensity and root colonization in an field experiment using faba bean, grown as vegetable, and pointed cabbage grown in monocropping and intercropping systems (Shanmugam et al. (2021)). This analysis use the methodology described in Paper III, however, in this case it is only univariate analyses.

R Package Implementing the Conditional Inference Method and Methods for Analysing the Graphical Latent Covariance Structure in GLMM

Rodrigo Labouriau, and I have developed a R-package implementing the methods described in Paper I (including the (multivariate) Laplace approximation) and II. This package is a tool for inference in both univariate and multivariate generalised linear mixed models and only because of this implementation, we were able to do the analyses presented in Paper III-VI. At this stage, the package is only developed as a prototype for internal use at the Applied Statistic Laboratory, Aarhus University. However, we plan to further develop the package and publish a version in the future.

References

Abreu, G. C., Labouriau, R. & Edwards, D. (2010), 'High-dimensional graphical model search with gRapHD R package', *Journal of Statistical Software* **37**(1).

Anderson, T. W. (2003), *An introduction to multivariate statistical analysis*, Vol. 2, Wiley New York.

Artes, R. & Jørgensen, B. (2000), 'Longitudinal data estimating equations for dispersion models', *Scandinavian Journal of Statistics* **27**(2), 321–334.

Barndorff-Nielsen, O. (2014), *Information and exponential families: in statistical theory*, John Wiley & Sons.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015), 'Fitting linear mixed-effects models using lme4', *Journal of Statistical Software* **67**(1), 1–48.

Berridge, D. & Crouchley, R. (2011), *Multivariate Generalized Linear Mixed Models Using R*, CRC Press.

Bonat, W. H. & Jørgensen, B. (2016), 'Multivariate covariance generalized linear models', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65**(5), 649–675.

Booth, J. G. & Hobert, J. P. (1999), 'Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**(1), 265–285.

Breslow, N. E. & Clayton, D. G. (1993), 'Approximate inference in generalized linear mixed models', *Journal of the American statistical Association* **88**(421), 9–25.

Cordeiro, G. M., Labouriau, R. & Botter, D. (2021), 'An introduction to bent jørgensen's ideas', *Brazilian journal of Probability and Statistics* **35**(1), 2–20.

Demidenko, E. (2004), *Mixed Models: Theory and Applications (Wiley Series in Probability and Statistics)*, Wiley-Interscience, USA.

Edwards, D. (2000), *Introduction to Graphical Modelling*, Springer-Verlag New York.

Edwards, D., de Abreu, G. C. & Labouriau, R. (2010), 'Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests', *BMC Bioinformatics* **11**(1).

Holthusen, H., Luca, A., Pelck, J., Labouriau, R. & Edelenbos, M. (2021a), 'Detection of rubbery rot caused by *Phacidiopycnis washingtonensis* by use of volatile monitoring in apple storage.'. In preparation.

Jørgensen, B. (1987), 'Exponential dispersion models', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 127–162.

Jorgensen, B. (1997), *The theory of dispersion models*, CRC Press.

Jørgensen, B. (2013), 'Construction of multivariate dispersion models', *Brazilian Journal of Probability and Statistics* **27**(3), 285–309.

Jørgensen, B. & Labouriau, R. (2012), *Exponential Families and Theoretical Inference*, Vol. 52, 2 edn, Springer.

Jørgensen, B. & Lauritzen, S. L. (2000), 'Multivariate dispersion models', *Journal of Multivariate Analysis* **74**(2), 267–281.

Krupskii, P. & Joe, H. (2013), 'Factor copula models for multivariate data', *Journal of Multivariate Analysis* **120**, 85–101.

Krupskii, P. & Joe, H. (2015), 'Structured factor copula models: Theory, inference and computation', *Journal of Multivariate Analysis* **138**, 53–73.

Labouriau, R. (2020), 'Construction and extension of dispersion models'. arXiv:2008.05448.

Lauritzen, S. L. (1996), *Graphical models*, Vol. 17, Clarendon Press.

Lee, Y. & Nelder, J. A. (1996), 'Hierarchical generalized linear models', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(4), 619–656.

Lee, Y. & Nelder, J. A. (2001), 'Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions', *Biometrika* **88**(4), 987–1006.

Lee, Y. & Nelder, J. A. (2005), 'Conditional and marginal models: Another view (with discussion)', *Statistical Science* **19**(2), 219–238.

Lee, Y., Nelder, J. A. & Noh, M. (2007), 'H-likelihood: problems and solutions', **17**.

McCullagh, P. & Nelder, J. A. (1989), *Generalized linear models*, Vol. 37, CRC press.

McCulloch, C. E. (1997), 'Maximum likelihood algorithms for generalized linear mixed models', *Journal of the American statistical Association* **92**(437), 162–170.

McCulloch, C. & Searle, S. (2001), *Generalized, Linear, and Mixed Models*, John Wiley & Sons.

Meng, X.-L. (2009), 'Decoding the h-likelihood', *Statistical Science* **24**(3).

Mikosch, T. (2005), *Copulas: Tales and facts*, Laboratory of Actuarial Mathematics, University of Copenhagen.

Nelder, J. A., Pawitan, Y. & Lee, H. J. (2006), *Generalized linear models with random effects: unified analysis via H-likelihood*, Chapman and Hall/CRC.

Nelder, J. & Wedderburn, R. (1972), 'Generalized linear models. jr statist. soc. a 135, 370-384. nelder370135j. r', *Statist. Soc A* **1972**.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Pelck, J. S. & Labouriau, R. (2021b), Multivariate generalised linear mixed models with graphical latent covariance structure. arXiv:2107.14535.

Perl, J. (2009), *Causality: models, reasoning and inference*, second edition edn, Cambridge University Press.

Salimi, Z., Pelck, J. P., Labouriau, R. & Boelt, B. (n.d.), The phenological response of sugar beet (beta vulgaris) through different development stages to seed priming. In preparation.

SAS Institute Inc. (2017), *SAS/STAT®14.3 User's Guide*. The GLIMMIX Procedure.

Shanmugam, S., Hefner, M., Pelck, J., Labouriau, R. & Kristensen, H. (2021), 'Complementary resource use in intercropped faba bean and cabbage by increased root growth and nitrogen use in organic production'. Submitted.

Song, P. X.-K., Li, M. & Yuan, Y. (2009), 'Joint regression analysis of correlated data using gaussian copulas', *Biometrics* **65**(1), 60–68.

Tweedie, M. (1984), *Statistics: applications and new directions*, Indian Statistical Institute, Calcutta, chapter An index which distinguishes between some important exponential families, pp. 579–604.

Whittaker, J. (1990), *Graphical models in applied multivariate analysis*, Chichester New York et al: John Wiley & Sons.

Xue-Kun Song, P. (2000), 'Multivariate dispersion models generated from gaussian copula', *Scandinavian Journal of Statistics* **27**(2), 305–320.

Paper I

Conditional Inference for Multivariate Generalised Linear Mixed Models

Jeanett S. Pelck
Aarhus University

Rodrigo Labouriau
Aarhus University

Abstract. We propose a method for inference in generalised linear mixed models (GLMMs) and several extensions of these models. First, we extend the GLMM by allowing the distribution of the random components to be non-Gaussian, that is, assuming an absolutely continuous distribution with respect to the Lebesgue measure that is symmetric around zero, unimodal and with finite moments up to fourth-order. Second, we allow the conditional distribution to follow a dispersion model instead of exponential dispersion models. Finally, we extend these models to a multivariate framework where multiple responses are combined by imposing a multivariate absolute continuous distribution on the random components representing common clusters of observations in all the marginal models.

Maximum likelihood inference in these models involves evaluating an integral that often cannot be computed in closed form. We suggest an inference method that predicts values of random components and does not involve the integration of conditional likelihood quantities. The multivariate GLMMs that we studied can be constructed with marginal GLMMs of different statistical nature, and at the same time, represent complex dependence structure providing a rather flexible tool for applications.

I.1 Introduction

Generalised linear mixed models (GLMMs) form a flexible class of statistical models, which combines the capability to incorporate non-Gaussian distributions and non-linear link functions, inherited from standard generalised linear models, with the power of representing complex dependence structures using random components in the same fashion as classic (Gaussian) mixed models. Therefore, GLMMs appear as a natural tool in many applications (see Demidenko, 2004; McCulloch & Searle, 2001; Fahrmeir & Tutz, 2001 and Agresti, 2002). However, the power of GLMMs comes with a price: the required inference tools are more demanding than standard statistical models. For instance, the likelihood-based inference requires a non-trivial integration of conditional likelihood quantities. Moreover, some of the simplifications

of the integration used in the classic Gaussian mixed models (*e.g.,* the result of conditioning a Gaussian distribution on Gaussian random components yields a Gaussian marginal distribution) do not apply in general for GLMMs. For this reason, several inferential tools are discussed in the literature; see Breslow & Clayton (1993), McCulloch & Searle (2001); see also McCulloch (1997) for a comprehensive study comparing several methods ranging from simple numeric (quadrature) integration of the conditional likelihood to several versions of the EM algorithm.

In this paper we present an alternative method of inference for GLMMs, constructed using inference functions, which avoids integrating likelihood quantities while preserving some of the desirable properties of classic likelihood-based methods. Moreover, this new method applies to GLMMs with minimal requirements for the distribution of the random components, which are not necessarily assumed to be normally distributed, as in the standard setup of GLMMs. For instance, we will be able to consider models with heavy-tailed random components as the multivariate t-distribution.

The methods we expose allow us to construct natural extensions to multivariate GLMMs. The main idea is to construct one GLMM describing each response. It is assumed that there is a natural cluster of observations (*e.g.,* individuals or experimental units). Each of those GLMMs contains random components representing those clusters, *i.e.,* taking the same value for all the observations belonging to the same cluster. The multivariate GLMM is then constructed by assuming that the distributions of the random components representing the clusters are the marginal distributions of a multivariate distribution (*e.g.,* a multivariate normal distribution or a multivariate t-distribution). Note that the multivariate generalised linear mixed models (MGLMMs), that we obtain in this way, can have marginal models of different nature which might be defined with different distributions and different link functions. In this way, those multivariate models can simultaneously describe responses of varying nature in a way that is not possible to do with classic multivariate Gaussian models. Furthermore, since we defined the random components of the marginal GLMMs using minimal distributional assumptions, we will also obtain a MGLMM constructed with a flexible class of multivariate random components. For instance, the multivariate random components can be multivariate normally distributed or regular elliptical contoured distributed.

The paper is structured as follows. In Section I.2, we introduce an extension of GLMMs constructed using random components that are not normally distributed, and by extending the family of conditional distributions. We use a simple case, containing random components representing a grouping of the observations (denoted clusters) due to the observational scheme used in the experiment, to present the ideas behind the inference techniques we propose in Section I.2.2, and expose the basic asymptotic properties of those techniques in Section I.2.3. Section I.2.4 extends the inference techniques to the case of models with complex clustering structures. In Section I.3, we discuss the inference for multivariate versions of GLMMs. Section I.3.1 presents two simulation studies. The appendices I.A.1, I.A.2 and I.A.3 expose some technical details and involved calculations. Appendix I.A.4 presents a multivariate extension of the classical inference method based on a Laplace approximation for GLMMs.

I.2 Extended One Dimensional Generalised Linear Mixed Models

This section will study a one-dimensional extension of standard GLMMs defined with random intercepts, and discuss an estimation technique based on conditional inference for those models. The GLMMs that we consider contain random components that are not necessarily Gaussian distributed. Moreover, they allow the conditional distributions to follow a general dispersion model, and therefore, they enlarge the class of standard GLMMs. We extend the models and inferential techniques described here to a multivariate context in Section I.3.

I.2.1 Generalised Linear Mixed Models with Simple Random Components

Consider the situation where we observe the responses of n individuals or experimental units. Those responses are viewed as realisations of n random variables taking values in $\mathcal{Y} \subseteq \mathbb{R}$, which we denote by $Y_1, \ldots, Y_n$. Here $\mathcal{Y}$ is typically $\mathbb{R}$, $\mathbb{R}_+$, a compact real interval or $\mathbb{Z}_+$ (corresponding to models defined using for example the Normal, Gamma, von Mises or the Poisson distributions). Suppose that each individual belongs to one, and only one, of q groups of individuals, referred as *clusters*. We assume that there exist q independent unobservable random variables taking values in $\mathbb{R}$, say $B_1, \ldots, B_q$, termed the *random components*, that will be associated to the clusters as described below. Denote the random vector $(B_1, \ldots, B_q)$ by $\boldsymbol{B}$. According to the model, the responses $Y_1, \ldots, Y_n$ are conditionally independent given $\boldsymbol{B}$. Furthermore, for $i = 1, \ldots, n$ and each $\boldsymbol{b} \in \mathbb{R}^q$, Y_i is conditionally distributed according to a dispersion model (see Jorgensen, 1997 and Cordeiro et al., 2021, or equation (I.2)) given $\boldsymbol{B}$, with conditional expectation given by

$$g\left(\mathbb{E}\left[Y_i|\boldsymbol{B}=\boldsymbol{b}\right]\right) = \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{b}\,, \quad \text{for all } \boldsymbol{b} \in \mathbb{R}^q\,. \tag{I.1}$$

Here g is a given *link function*, $\boldsymbol{x}_i$ is a vector of k explanatory variables associated to the i^{th} individual and $\boldsymbol{\beta} \in \Omega \subseteq \mathbb{R}^k$ is a vector of coefficients, referred as the *fixed effects*. Furthermore, $\boldsymbol{z}_i$ is a q-dimensional allocation vector associating the i^{th} individual to one of the q clusters. The j^{th} entry of the vector $\boldsymbol{z}_i$ takes the value 1 if the i^{th} individual belongs to the j^{th} cluster and 0 otherwise. Other forms of allocation vectors are possible, but we restrict to the particular form above to simplify the exposition of ideas.

It is convenient to introduce the following nomenclature and notation for the right side of (I.1). The *linear predictor* and the *conditional mean response* for the i^{th} individual $(i = 1, \ldots, n)$ are defined by $\eta_i = \eta_i\left(\boldsymbol{\beta}, \boldsymbol{b}\right) \stackrel{\text{def}}{=} \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{b}$ and $\mu_i = \mu_i\left(\boldsymbol{\beta}, \boldsymbol{b}\right) \stackrel{\text{def}}{=} g^{-1}(\eta_i)$, respectively. The parameter space of the conditional means is denoted by $\mathcal{U} \subseteq \mathbb{R}$ and we write $\mu_i \in \mathcal{U}$. Additionally, denote the random vector of observations $(Y_1, \ldots, Y_n)$ by $\boldsymbol{Y}$, and the vector of observed responses $(y_1, \ldots, y_n)$ by $\boldsymbol{y}$.

The specification of the extended GLMM that we consider is completed by defining the distribution of the random components as follows. We assume that $B_1, \ldots, B_q$ are independent and identically distributed according to a distribution that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$, symmetric around zero, unimodal, and possessing finite moments up to the fourth-order. Note that the random components have expectation zero due to the symmetry. Denote the density of this distribution by $\varphi(\cdot, \sigma^2)$, where $\sigma^2 \in \mathcal{V} \stackrel{\text{def}}{=} \mathbb{R}_+$ is a parameter describing the dispersion of the distribution. Here a typical choice would be a normal or a regular absolute continuous one-dimensional elliptically contoured family of distributions and in this case σ^2 would be the variance parameter. [1]

Under the model defined above, the conditional distribution of the i^{th} observation Y_i given $\boldsymbol{B}$ (for $i = 1, \ldots, n$), has a density with respect to a dominating measure ν (defined on the measurable space $(\mathcal{Y}, \mathcal{A})$), taking the form of a dispersion model (see Jorgensen, 1997 and Cordeiro et al., 2021). Therefore, the referred density takes the form

$$\begin{aligned} f(y_i|\boldsymbol{B} = \boldsymbol{b}; \boldsymbol{\beta}, \lambda) &= a(y_i; \lambda) \exp\left[-\tfrac{1}{2\lambda}\, d\left\{y_i; g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{b})\right\}\right] \\ &= a(y_i; \lambda) \exp\left\{-\tfrac{1}{2\lambda}\, d\left(y_i; \mu_i\right)\right\}, \quad \forall\ y_i \in \mathcal{Y},\ \forall\ \boldsymbol{b} \in \mathbb{R}^q\,, \end{aligned} \tag{I.2}$$

where $\boldsymbol{\beta} \in \Omega$ and $\lambda \in \Lambda = \mathbb{R}_+$. The function $d : \mathcal{Y} \times \mathcal{U} \to \mathbb{R}_+$ is the *unit deviance* and, by definition, satisfies that $d(\mu, \mu) = 0$ and $d(y, \mu) > 0$ for all $(y, \mu) \in \mathcal{Y} \times \mathcal{U}$ such that $y \neq \mu$. The function $a : \mathcal{Y} \times \mathbb{R}_+ \to \mathbb{R}_+$ is a given normalising function. We assume that the unit deviance is regular, that is, d is twice continuously differentiable in $\mathcal{Y} \times \mathcal{U}$ and $\partial^2 d(\mu; \mu)/\partial\mu^2 > 0$ for all $\mu \in \mathcal{U}$. The function $V : \mathcal{U} \to \mathbb{R}_+$ given by $V(\mu) = 2/\{\partial^2 d(\mu, \mu)/\partial\mu^2\}$ for all μ in $\mathcal{U}$ is termed the *variance function* (Cordeiro et al. 2021). The conditional variance of Y_i given the random components is $V(\mu_i)/\lambda$. The following families of distributions are examples of dispersion models: Normal, Gamma, inverse Gaussian, von Mises, Poisson, and Binomial families.

We formally define the extended GLMM described above as the family

$$\mathcal{P} = \left\{P_{\boldsymbol{\beta},\lambda,\sigma^2} : \boldsymbol{\beta} \in \Omega, \quad \lambda \in \Lambda = \mathbb{R}_+, \quad \sigma^2 \in \mathcal{V} = \mathbb{R}_+\right\}$$

of probability measures defined on the product measurable space $(\mathcal{Y}^n, \mathcal{A}^n)$ (where $\mathcal{A}^n$ is the related product σ-algebra) corresponding to the probability measures defining the extended GLMM described above. Let $\boldsymbol{\nu}$ be the product measure induced by ν. The density of the distributions in $\mathcal{P}$, with respect to $\boldsymbol{\nu}$, are given by

$$p\left(\boldsymbol{y}; \boldsymbol{\beta}, \lambda, \sigma^2\right) \stackrel{\text{def}}{=} \frac{dP_{\boldsymbol{\beta},\lambda,\sigma^2}(\boldsymbol{y})}{d\boldsymbol{\nu}} = \int_{\mathbb{R}^q} \prod_{i=1}^{n} f(y_i|\boldsymbol{B} = \boldsymbol{b}; \boldsymbol{\beta}, \lambda) \prod_{j=1}^{q} \varphi(b_j; \sigma^2)\ d\boldsymbol{b}\,, \tag{I.3}$$

for all $\boldsymbol{y} \in \mathcal{Y}^n$, $\boldsymbol{\beta} \in \Omega$, $\lambda \in \Lambda$ and $\sigma^2 \in \mathcal{V}$.

We will use the following set of regularity conditions on the generalised linear mixed model $\mathcal{P}$:

[1] Here a one-dimensional elliptically contoured family of distributions is a location and scale family of distributions, with location and scale parameters μ and σ, for which the characteristic functions ϕ, satisfy the functional equation $\phi(t) = e^{i\mu t}\psi(-\frac{1}{2}t\sigma^2 t)$ for all $t \in \mathbb{R}$, for a given function ψ.

(i) The matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ have full rank (*i.e.,* rank k and q, respectively)

(ii) The link function is strictly monotone, invertible and twice continuously differentiable with bounded first order derivative

(iii) The unit deviance, $d(y, \mu)$, is twice continuous differentiable with respect to μ

(iv) The functions $\frac{\partial}{\partial \boldsymbol{\beta}} d\left\{ \cdot\,; g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{z}_i^T \boldsymbol{b}) \right\}$ and $\frac{\partial}{\partial \boldsymbol{b}} d\left\{ \cdot\,; g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{z}_i^T \boldsymbol{b}) \right\}$ are dominated by integrable functions (not necessarily the same dominating functions) for each $\boldsymbol{\beta} \in \mathbb{R}^k$ and $\boldsymbol{b} \in \mathbb{R}^q$.

These mild regularity conditions turn out to be minimal requirements for the inference theory that we construct.

Let $\boldsymbol{y} = (y_1, \ldots y_n)$ be a realisation of the random vector $\boldsymbol{Y} \stackrel{\mathsf{def}}{=} (Y_1, \ldots, Y_n)$ of responses. Under the model $\mathcal{P}$, the likelihood function for the parameters $\boldsymbol{\beta}, \lambda$ and σ^2, based on $\boldsymbol{y}$, is

$$L(\boldsymbol{\beta}, \lambda, \sigma^2; \boldsymbol{y}) = p\left(\boldsymbol{y}; \boldsymbol{\beta}, \lambda, \sigma^2\right) . \tag{I.4}$$

Usually, the integral in the right side of (I.3) involved in the calculation of the likelihood function in (I.4), cannot be evaluated in closed form. In Section I.2.2, we introduce an inference method that includes predictions of values of the random components, $B_1, \ldots, B_q$, and avoids the integration. This inferential procedure will be justified using asymptotic arguments in Section I.2.3.

We introduce below two families of probability measures related to $\mathcal{P}$, which will be convenient for presenting and discussing the conditional inference for the GLMMs under discussion. First, consider a statistical model, $\overline{\mathcal{P}}$, constructed on $\mathcal{Y}^n \times \mathbb{R}^q$, collecting the joint distributions of the n responses and the q random components. This model, called the *joint-model*, represents the hypothetical situation in which the random components would be observable. We will use the joint-model to introduce and motivate the inferential techniques we propose.

It is convenient to introduce also the following family of probability measures on $(\mathcal{Y}^n, \mathcal{A}^n)$, obtained by collecting the distributions constructed with the realisable values of the random components $B_1, \ldots, B_q$, in the following way

$$\mathcal{P}^* = \left\{ P^*_{\boldsymbol{\beta},\boldsymbol{b},\lambda} : \frac{dP^*_{\boldsymbol{\beta},\boldsymbol{b},\lambda}}{d\boldsymbol{\nu}}(\boldsymbol{y}) = f^*(\boldsymbol{y}; \boldsymbol{\beta}, \boldsymbol{b}, \lambda) \quad \text{for all } \boldsymbol{y} \in \mathcal{Y}^n, \boldsymbol{\beta} \in \Omega, \boldsymbol{b} \in \mathbb{R}^q, \lambda \in \Lambda \right\} . \tag{I.5}$$

The density of the probability measure referred above is given by

$$f^*\left(\boldsymbol{y}; \boldsymbol{\beta}, \boldsymbol{b}, \lambda\right) \stackrel{\mathsf{def}}{=} \prod_{i=1}^{n} f(y_i | \boldsymbol{B} = \boldsymbol{b}; \boldsymbol{\beta}, \lambda) = \prod_{i=1}^{n} a(y_i; \lambda) \exp\left\{ -\tfrac{1}{2\lambda}\, d\left[y_i; \mu_i\left(\boldsymbol{\beta}, \boldsymbol{b}\right)\right] \right\} ,$$

for all $\boldsymbol{y} \in \mathcal{Y}^n$, $\boldsymbol{\beta} \in \Omega$, $\lambda \in \Lambda$ and $\boldsymbol{b} \in \mathbb{R}^q$. We call the family $\mathcal{P}^*$ the *conditional model*. This family will be used for defining inference functions, and establishing the basic properties of the inference procedures we will propose.

I.2.2 Conditional Inference for Models with a Single Random Component

Under the joint model $\overline{\mathcal{P}}$, the log-likelihood function for estimating $\boldsymbol{\beta}$, λ and σ^2 based on realisations $\boldsymbol{y}$ and $\boldsymbol{b}$ of $\boldsymbol{Y}$ and $\boldsymbol{B}$, respectively, is

$$l\left(\boldsymbol{\beta}, \lambda, \sigma^2; \boldsymbol{y}, \boldsymbol{b}\right) \stackrel{\mathsf{def}}{=} \sum_{i=1}^{n} \log f(y_i|\boldsymbol{B} = \boldsymbol{b}; \boldsymbol{\beta}, \lambda) + \sum_{j=1}^{q} \log \varphi(b_j; \sigma^2)\,. \tag{I.6}$$

From this perspective, $\boldsymbol{b}$ is a S-sufficient statistic with respect to σ^2 (since the term of the likelihood function that contains σ^2 depends only on $\boldsymbol{b}$ and not on $\boldsymbol{y}$), and S-ancillary with respect to $\boldsymbol{\beta}$ and λ (since the term of the likelihood function that contains $\boldsymbol{\beta}$ and λ involves $\boldsymbol{b}$ only conditionally). See Barndorff-Nielsen (2014, page 50) or Jørgensen & Labouriau (2012, Section 3.2) for formal definitions.

The decomposition of the likelihood function of the joint model $\overline{\mathcal{P}}$, defined in (I.6), motivates that the inference on σ^2 should be performed using the term

$$\sum_{j=1}^{q} \log \varphi(b_j; \sigma^2)\,,$$

corresponding to base the inference on σ^2 on a sufficient statistic. Following the same line, the inference on $\boldsymbol{\beta}$ and λ should be performed only using the term

$$\sum_{i=1}^{n} \log f(y_i|\boldsymbol{B} = \boldsymbol{b}; \boldsymbol{\beta}, \lambda), \tag{I.7}$$

which corresponds to perform conditional likelihood-based inference given an ancillary statistic. Therefore, we propose to estimate $\boldsymbol{\beta}$ and λ by inserting a reasonable prediction of $\boldsymbol{b}$, say $\tilde{\boldsymbol{b}}$ as defined below, into (I.7) and maximising for $\boldsymbol{\beta}$ and λ. We argue in Section I.2.3 that the procedure informally defined here yields sensible estimates.

We turn now to the problem of predicting $\boldsymbol{b}$. Under the joint model $\overline{\mathcal{P}}$, it is natural to predict $\boldsymbol{b}$ by maximising $l\left(\boldsymbol{\beta}, \lambda, \sigma^2; \boldsymbol{y}, \boldsymbol{b}\right)$ given in (I.6), *i.e.,* by

$$\hat{\boldsymbol{b}}(\boldsymbol{\beta}, \lambda, \sigma^2; \boldsymbol{y}) = \underset{b_1,\ldots,b_q}{\arg\max} \left\{ \sum_{i=1}^{n} \log f(y_i|\boldsymbol{B} = \boldsymbol{b}, \boldsymbol{\beta}, \lambda) + \sum_{j=1}^{q} \log \varphi(b_j; \sigma^2) \right\}. \tag{I.8}$$

However, it is convenient, as we will demonstrate in Section I.2.3, to use the following approximation to $\hat{\boldsymbol{b}}$,

$$\tilde{\boldsymbol{b}}(\boldsymbol{\beta}, \lambda; \boldsymbol{y}) \stackrel{\mathsf{def}}{=} \Pi_{\mathcal{B}_0}\left(\underset{b_1,\ldots,b_q}{\arg\max} \sum_{i=1}^{n} \log f(y_i|\boldsymbol{B} = \boldsymbol{b}, \boldsymbol{\beta}, \lambda) \right), \tag{I.9}$$

where $\mathcal{B}_0 \stackrel{\mathsf{def}}{=} \{\boldsymbol{b} \in \mathbb{R}^q : \frac{1}{q}\sum_{j=1}^{q} b_j = 0\}$ is the subspace of the vectors in $\mathbb{R}^q$ with mean zero, and $\Pi_{\mathcal{B}_0} : \mathbb{R}^q \to \mathcal{B}_0$ is the projection function given by $\Pi_{\mathcal{B}_0}(\boldsymbol{y}) \stackrel{\mathsf{def}}{=} \boldsymbol{y} - 1/q \sum_{j=1}^{q} y_j$. Note, that $\tilde{\boldsymbol{b}}$ is an approximation of $\hat{\boldsymbol{b}}$, because the last term of the right side of (I.8) is maximised by setting $\boldsymbol{b}$ equal to zero. The approximation follows from the continuity of the function $\varphi(\cdot; \sigma^2)$, which has a unique mode at zero, and because $\tilde{\boldsymbol{b}}$ is in $\mathcal{B}_0$.

I.2.3 Asymptotic Properties of the Conditional Inference Method

In this section, we formulate the inferential techniques presented in Section I.2.2 using the theory of inference functions (Jørgensen & Labouriau (2012) and Barndorff-Nielsen (2014)). We show that the estimated value of $\boldsymbol{\beta}$ and the predicted values of $\boldsymbol{b}$ are asymptotically Gaussian distributed when the variance, σ^2, of the random component is small.

We consider below the inference functions

$$\psi^*_{\boldsymbol{\beta}} : \Omega \times \mathbb{R}^q \times \mathcal{Y} \to \mathbb{R}^k \text{ and } \psi^*_{\boldsymbol{b}} : \Omega \times \mathbb{R}^q \times \mathcal{Y} \to \mathbb{R}^q,$$

which are equivalent to the score functions for estimating $\boldsymbol{\beta}$ and $\boldsymbol{b}$, under $\mathcal{P}^*$, with λ treated as a nuisance parameter. The inference functions $\psi^*_{\boldsymbol{\beta}}$ and $\psi^*_{\boldsymbol{b}}$ referred above are defined by

$$\psi^*_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{y}) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}} d\left(y_i; g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{z}_i^T \boldsymbol{b})\right) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{\frac{\partial}{\partial \mu_i} d(y_i; \mu_i)}{g'(\mu_i)}, \tag{I.10}$$

$$\psi^*_{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{y}) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{b}} d\left(y_i; g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{z}_i^T \boldsymbol{b})\right) = \sum_{i=1}^{n} \boldsymbol{z}_i \frac{\frac{\partial}{\partial \mu_i} d(y_i; \mu_i)}{g'(\mu_i)}. \tag{I.11}$$

Note that the score functions for estimating $\boldsymbol{\beta}$ and $\boldsymbol{b}$ are given by $\psi^*_{\boldsymbol{\beta}}$ and $\psi^*_{\boldsymbol{b}}$ multiplied by $-\frac{1}{2\lambda}$. However, since λ is a positive number the solution to the score equations for $\boldsymbol{\beta}$ and $\boldsymbol{b}$ are exactly the roots of $\psi^*_{\boldsymbol{\beta}}$ and $\psi^*_{\boldsymbol{b}}$; in this sense they are equivalent. The inference function $\psi^* : \Omega \times \mathbb{R}^q \times \mathcal{Y} \to \mathbb{R}^{k+q}$ given by

$$\psi^*(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{y}) = \left\{ \left[\psi^*_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{y})\right]^T, \left[\psi^*_{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{y})\right]^T \right\}^T,$$

will be used for estimating $\boldsymbol{\beta}$ and predicting $\boldsymbol{b}$. We denote the sequences of roots of the inference functions $\psi^*_{\boldsymbol{\beta}}$ and $\psi^*_{\boldsymbol{b}}$ by $\{\widehat{\boldsymbol{\beta}}_n\}_{n\in\mathbb{N}}$ and $\{\widehat{\boldsymbol{b}}_n\}_{n\in\mathbb{N}}$, respectively, obtained when the number of observations, n, increases.

According to the classic theory of inference functions (see Jørgensen & Labouriau, 2012, Chapter 4), the estimating functions $\psi^*_{\boldsymbol{\beta}}$ and $\psi^*_{\boldsymbol{b}}$ yield consistent estimates under $\mathcal{P}^*$. Moreover, the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{b}$ are conditionally asymptotically normally distributed (see the details in Appendix I.A.2). However, our primary interest is on estimating $\boldsymbol{\beta}$ under the extended generalised linear mixed model $\mathcal{P}$. For this purpose, we define below the inference function $\psi_{\boldsymbol{\beta}} : \Omega \times \mathcal{Y} \to \mathbb{R}^k$ given by

$$\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}; \boldsymbol{y}) \stackrel{\mathsf{def}}{=} \psi^*_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \widehat{\boldsymbol{b}}; \boldsymbol{y}), \text{ for all } \boldsymbol{\beta} \in \Omega \text{ and all } \boldsymbol{y} \in \mathcal{Y}, \tag{I.12}$$

where $\widehat{\boldsymbol{b}}$ is obtained from the joint solution, $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{b}})$, of the estimating equation $\psi^*_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{y}) = 0$ and $\psi^*_{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{y}) = 0$. The theorem below shows that, under the assumed mild regularity conditions, the root of $\psi_{\boldsymbol{\beta}}$ are consistent and asymptotically Gaussian distributed when the variance of the random components converges to zero.

Theorem I.2.1. *Under the regularity conditions i-iv, the sequences* $\{\hat{\boldsymbol{\beta}}_n\}_{n\in\mathbb{N}}$ *and* $\{\widehat{\boldsymbol{b}}_n\}_{n\in\mathbb{N}}$ *are consistent (in probability) under* $\mathcal{P}^*$*. Moreover,* $\{\hat{\boldsymbol{\beta}}_n\}_{n\in\mathbb{N}}$ *is consistent (in probability) under* $\mathcal{P}$*. Both sequences are asymptotically Gaussian distributed, when* $n \to \infty$ *and* $\sigma^2 \downarrow 0$*.*

Proof. See Lemma I.A.4 for the consistency of $\{\widehat{\boldsymbol{\beta}}_n\}_{n\in\mathbb{N}}$ and $\{\widehat{\boldsymbol{b}}_n\}_{n\in\mathbb{N}}$ under $\mathcal{P}^*$. See Lemma I.A.5 for the consistency in probability of $\{\widehat{\boldsymbol{\beta}}_n\}_{n\in\mathbb{N}}$ under $\mathcal{P}$ and Theorem I.A.7 in Appendix I.A.2 for the asymptotic normality when the variance of the random components is sufficiently small. □

The parametrisation of the family $\mathcal{P}^*$ defined above is not identifiable. Note, that a natural parametrisation of $\mathcal{P}^*$ using the triplet $(\boldsymbol{\beta}, \boldsymbol{b}, \lambda) \in \Omega \times \mathbb{R}^q \times \Lambda$ is not identifiable. Indeed, according to the Lemma I.A.1 proved in the Appendix I.A.1, for any $i \in \{1, \ldots, n\}$ and any choice of $\boldsymbol{\beta}$, $\boldsymbol{b}$ and $\delta > 0$ there exists a $\boldsymbol{\beta}_\delta \in \Omega$ such that $\eta_i(\boldsymbol{\beta}, \boldsymbol{b}) = \eta_i(\boldsymbol{\beta}_\delta, \boldsymbol{b} - \delta)$. A convenient solution to this issue is to introduce a constraint and require that $\boldsymbol{b}$ takes values in $\mathcal{B}_0$ (*i.e.,* the sub-space of $\mathbb{R}^q$ of vectors with mean zero), which yields an identifiable parametrisation of $\mathcal{P}^*$. We adopt this parametrisation and re-write here (I.5) in the form

$$\mathcal{P}^* = \left\{ P^*_{\boldsymbol{\beta}^*, \boldsymbol{b}^*, \lambda} : \frac{dP^*_{\boldsymbol{\beta}^*, \boldsymbol{b}^*, \lambda}}{d\boldsymbol{\nu}}(\boldsymbol{y}) = f^*(\boldsymbol{y}; \boldsymbol{\beta}^*, \boldsymbol{b}^*, \lambda) \quad \text{for all } \boldsymbol{y} \in \mathcal{Y}^n, \boldsymbol{\beta}^* \in \Omega^*, \boldsymbol{b}^* \in \mathcal{B}_0, \lambda \in \Lambda \right\},$$

so the mapping from $\Omega \times \mathcal{B}_0 \times \Lambda$ to $\mathcal{P}^*$ given by $(\boldsymbol{\beta}^*, \boldsymbol{b}^*, \lambda) \mapsto P_{\boldsymbol{\beta}^*, \boldsymbol{b}^*, \lambda}$ is a bijection.

The sequences of estimates $\{\widehat{\boldsymbol{\beta}}^*_n\}_{n\in\mathbb{N}}$ and $\{\widehat{\boldsymbol{b}}^*_n\}_{n\in\mathbb{N}}$ obtained as roots to the inference functions defined as above but with the new identifiable parametrisation, yields the same maximum likelihood values as a consequence of Lemma I.A.1 proved in the Appendix I.A.1. By the law of large numbers and Lemma I.A.4, $(\widehat{\boldsymbol{\beta}}^*_n, \widehat{\boldsymbol{b}}^*_n)$ converges to $(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{b}}_n)$ in probability under $P^*_{\boldsymbol{\beta}, \boldsymbol{b}, \lambda}$ for q and n converging to infinity.

In Section I.3.2, we study the distribution of $\hat{\boldsymbol{\beta}}$ in a simulated example, where we assume that the random components follow a Gaussian distribution.

I.2.4 A Simple Algorithm for Conditional Inference

The following algorithm implements the inference method described above. The algorithm starts by setting the initial values $\boldsymbol{\beta}^{(0)}$ and $\lambda^{(0)}$ for the parameters $\boldsymbol{\beta}$ and λ. We used the estimated values of the corresponding parameters of a generalised linear model defined with the same distribution and link function as in the extended GLMM in study, and with the linear predictor given by the fixed effects of the extended GLMM in discussion. The algorithm repeats the following two steps, starting with $m = 0$, until convergence:

1. Let $\boldsymbol{\beta}^{(m)}$ and $\lambda^{(m)}$ be the current estimates of the parameters $\boldsymbol{\beta}$ and λ. Set

$$\boldsymbol{b}^{(m+1)} = \underset{b_1, \ldots, b_q}{\arg\max} \sum_{i=1}^{n} \log f(y_i | \boldsymbol{B} = \boldsymbol{b}, \boldsymbol{\beta}^{(m)}, \lambda^{(m)}),$$

and

$$\boldsymbol{b}^*_{(m+1)} = \tilde{\boldsymbol{b}}(\boldsymbol{\beta}^{(m)}, \lambda^{(m)}; \boldsymbol{y}) = \Pi_{\mathcal{B}_0}\left(\boldsymbol{b}^{(m+1)}\right),$$

with $\tilde{\boldsymbol{b}}$ is defined as in (I.9).

2. Given the latest predicted values of the random components denoted $\boldsymbol{b}^*_{(m+1)}$, $\boldsymbol{\beta}^{(m+1)}$ and $\lambda^{(m+1)}$ are estimated by maximising $\prod_{i=1}^n f^*(y_i; \boldsymbol{\beta}, \boldsymbol{b}^*_{(m+1)}, \lambda)$ with respect to $\boldsymbol{\beta}$ and λ.

After convergence has been obtained, we estimate the variance, finding the value of σ^2 that maximises the integral

$$\int_{\mathcal{B}_0} g(\hat{\boldsymbol{b}}; \boldsymbol{b}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{b}}}) \prod_{j=1}^{q} \varphi(b_j; \sigma^2) d\boldsymbol{b}, \ , \tag{I.13}$$

where $\hat{\boldsymbol{b}}$ denotes the value of $\boldsymbol{b}^{(m+1)}$ in the last round of the algorithm. Here, $g(\cdot; \boldsymbol{b}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{b}}})$ denotes the density of the predicted values from the final iteration, $\hat{\boldsymbol{b}}$, with expectation $\boldsymbol{b}$ and covariance $\boldsymbol{\Sigma}_{\hat{\boldsymbol{b}}}$. In the case where σ^2 is small enough and n is large enough, this density is close to the multivariate Gaussian density, see Theorem I.2.1 for details. In Appendix I.A.3, calculations of the above integral are given in the case where g and φ are densities of Gaussian distributions.

I.2.5 Conditional Inference for Models with Complex Random Components

This section extends the methods introduced in section I.2.3 to a context with complex random components. We first consider non-nested random components, and then we study a scenario where the random components are nested or a combination of the two cases.

When the random components are not nested, the values of the random components are easily predicted using the already described method. To simplify the notation, consider a one dimensional extended GLMM with two vectors of non-nested random components (each corresponding to a clustering of the observations), say $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ with length q_1 and q_2, respectively. We assume that $Y_1, \ldots, Y_n$ are conditional independent random variables given $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$, and conditionally distributed according to a dispersion model, with conditional density $f(\,\cdot\,|\boldsymbol{B}_1 = \boldsymbol{b}_1, \boldsymbol{B}_2 = \boldsymbol{b}_2, \boldsymbol{\beta}, \lambda)$, where f is defined in (I.2).

Recall, that values of the random components were predicted using Equation (I.9), which is equivalent to solving the inference functions in (I.10) and (I.11). This equation can easily be adapted to the situation with multiple non-nested random components. We replace $\mathcal{B}_0$ by $\widetilde{\mathcal{B}_0} \overset{\text{def}}{=} \{(\boldsymbol{b}_1, \boldsymbol{b}_2) \in \mathbb{R}^{q_1+q_2} : \boldsymbol{b}_1 \in \mathcal{B}_0(\mathbb{R}^{q_1}) \text{ and } \boldsymbol{b}_2 \in \mathcal{B}_0(\mathbb{R}^{q_2})\}$ (where $\mathcal{B}_0(\mathbb{R}^q)$ is the space of vectors of $\mathbb{R}^q$ with mean zero) and define

$$\tilde{\boldsymbol{b}}(\boldsymbol{\beta}, \lambda; \boldsymbol{y}) \overset{\text{def}}{=} \Pi_{\widetilde{\mathcal{B}_0}}\left[\underset{(\boldsymbol{b}_1, \boldsymbol{b}_2) \in \mathbb{R}^{q_1+q_2}}{\arg\max} \sum_{i=1}^{n} \log f(y_i | \boldsymbol{B}_1 = \boldsymbol{b}_1, \boldsymbol{B}_2 = \boldsymbol{b}_2; \boldsymbol{\beta}, \lambda)\right].$$

We turn now to the case of two nested vectors of random components $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$, where $\boldsymbol{B}_1$ is nested in $\boldsymbol{B}_2$, that is, the clusters corresponding to the entries in $\boldsymbol{B}_2$ groups multiple clusters associated with $\boldsymbol{B}_1$. Therefore, the variation in $\boldsymbol{B}_1$ should be interpreted as the remaining variation not explained by $\boldsymbol{B}_2$. In this case, we estimate the model including only the random component $\boldsymbol{B}_1$. After predicting (temporary) values for $\boldsymbol{B}_1$ denoted by $\bar{\boldsymbol{b}}_1$, we predict the final values of $\boldsymbol{b}_2$ by

$$\hat{\boldsymbol{b}}_2 = (\boldsymbol{Z}_2^T \boldsymbol{Z}_2)^{-1} \boldsymbol{Z}_2^T \bar{\boldsymbol{b}}_1,$$

where $\boldsymbol{Z}_2$ a $q_1 \times q_2$ dimensional matrix with the (i,j)'th entry equal to one if the cluster corresponding to the i^{th} entry of $\boldsymbol{B}_1$ is contained in the j^{th} cluster associated with the j^{th} entry of $\boldsymbol{B}_2$, and zero otherwise. Next, the predicted values of $\boldsymbol{b}_1$ is updated to the final values by

$$\hat{\boldsymbol{b}}_1 = \bar{\boldsymbol{b}}_1 - \boldsymbol{Z}_2 \hat{\boldsymbol{b}}_2.$$

These methods can easily be generalised to the multivariate case by using the approach described in Section I.3.

I.3 Multivariate Models

In this section, we extend the methods described so far in one dimension to a multivariate context. Consider d response vectors simultaneously observed, each of them following an GLMM described in Section I.2. Here the d responses might follow different dispersion models, use different link functions, but the d marginal extended GLMMs must have a common random component with the same clusters for each of the response vectors. The inference method presented in the Sections I.2.2– I.2.5 yields predicted values of the random components directly as an additional product of the estimation process.

I.3.1 Basic Setup

We introduce the following notation required for formally defining the multivariate model we have in mind. Let $\boldsymbol{Y} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_d\}$ be a $n \times d$ dimensional response variable matrix, and $\boldsymbol{B} = \{\boldsymbol{B}_{(1)}, \ldots, \boldsymbol{B}_{(d)}\} = \{\boldsymbol{B}_1, \ldots, \boldsymbol{B}_q\}^T$ a $q \times d$ dimensional matrix of random components. Each column of $\boldsymbol{Y}$ corresponds to n response variables in a univariate model. We assume, that the rows of $\boldsymbol{B}$ are independent and identical distributed according to a multivariate distribution which is absolute continuous with respect to the Lebesgue measure, symmetric around the vector of zeros, unimodal, and with finite moments up to fourth order. We will let $\boldsymbol{\Sigma}$ denote a covariance matrix of the distribution and $\varphi(\cdot, \boldsymbol{\Sigma})$ the density. Often, this distribution will be assumed to be multivariate Gaussian with expectation zero and covariance matrix given by $\boldsymbol{\Sigma}$.

For $i = 1, \ldots, n$ and $j = 1, \ldots, d$, we assume that Y_{ij} is conditional distributed according to a dispersion model given $\boldsymbol{B}_{(j)} = \boldsymbol{b}_{(j)}$. That is, $Y_{ij} | \boldsymbol{B}_{(j)} = \boldsymbol{b}_{(j)} \sim \mathrm{D}(\mu_{ij}, \lambda_j)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, d$, where $D(\mu; \lambda)$ denotes the dispersion model

distribution with expectation μ and dispersion λ. The conditional expectation, μ_{ij}, is connected to the linear predictor, η_{ij}, through the known link function denoted g_j, that is, $g_j(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_j + \boldsymbol{z}_i^T\boldsymbol{b}_{(j)}$, where $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_i$ denote the vector of explanatory variables and a location vector, respectively. Notice, that like in the one dimensional model, $\boldsymbol{z}_i$ has one entry equal to one and the remaining entries are equal to zero. Thus, $\boldsymbol{z}_i$ has a one in the entry corresponding to the cluster that the ith individual belongs to. The conditional density of Y_{ij} given $\boldsymbol{B}_{(j)}$ is denoted by f_j.

We assume, that Y_{ij} and $Y_{i'j}$ are conditionally independent given $\boldsymbol{B}_{(j)} = \boldsymbol{b}_{(j)}$ for $i \neq i'$ $(i, i' = 1, \ldots, n)$. Moreover, the structure of the model implies that Y_{ij} and $Y_{i'j'}$ are conditionally independent given $\boldsymbol{B}_{(j)}$ and $\boldsymbol{B}_{(j')}$ for all $i, i' = 1, \ldots, n$ and $j, j' = 1 \ldots, d$ such that $j \neq j'$.

I.3.2 Simulation Studies

In this section, we present results of two simulation studies illustrating basic properties of the proposed estimation procedure. Moreover, we compare the behaviour of the proposed estimates with two other inference methods: the multivariate Laplace approximation suggested by Breslow & Clayton (1993) (see Appendix I.A.4 for details) and a Hermite quadrature estimation procedure. Two simulation studies are presented to study the distribution of the estimates when the entries in the covariance matrix are varied, and the bias of the estimated parameters when we increase the numbers of clusters of the random component (and thereby the number of observations). In both simulation studies, we simulate a two dimensional generalised linear mixed model, where Y_{ij} for $i = 1, \ldots, n$ and $j = 1, 2$ denotes the response variables. We follow the notation introduced above and let $\boldsymbol{B}_{(1)}$ and $\boldsymbol{B}_{(2)}$ denote q-dimensional random vectors representing the random components in the model. We assume that $Y_{11}, \ldots, Y_{n2}$ are conditionally independent given $\boldsymbol{B}_{(1)}$ and $\boldsymbol{B}_{(2)}$. Moreover, we assume that given $\boldsymbol{B}_{(1)}$ and $\boldsymbol{B}_{(2)}$, Y_{i1} and Y_{i2} are conditionally distributed according to a Gaussian and a Poisson distribution, respectively, with conditional expectations given by

$$\begin{aligned}
\mathbb{E}[Y_{i1}|\boldsymbol{B}_{(1)} = \boldsymbol{b}] &= \boldsymbol{x}_{i1}^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{b} \quad \text{for } i = 1, \ldots, n, \\
\mathbb{E}[Y_{i2}|\boldsymbol{B}_{(2)} = \boldsymbol{b}] &= \exp\left(\boldsymbol{x}_{i2}^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{b}\right) \quad \text{for } i = 1, \ldots, n,
\end{aligned}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2) = (1.90, 0.21)$. The Gaussian conditional distribution is assumed to have a variance of 0.5 which is not varied in the simulations.

We assume that $\boldsymbol{B}^T = (\boldsymbol{B}_{(1)}^T, \boldsymbol{B}_{(2)}^T)$ is Gaussian distributed with expectation zero and covariance structure given by

$$\begin{aligned}
\text{Cov}(B_{(1)}^l, B_{(2)}^l) &= \boldsymbol{\Sigma} \quad \text{for } \ l = 1, ..., q, \\
\text{Cov}(B_{(1)}^l, B_{(2)}^k) &= 0 \quad \text{for } \ l, k = 1, ..., q \text{ such that } l \neq k,
\end{aligned}$$

where $B_{(j)}^l$ denotes the l^{th} entry in $\boldsymbol{B}_{(j)}$ for $j = 1, 2$, and

$$\boldsymbol{\Sigma} = \text{const} \begin{pmatrix} 0.28 & 0.09 \\ 0.09 & 0.12 \end{pmatrix}, \tag{I.14}$$

with the constant depending on the simulation study. That is,

$$\boldsymbol{B} \sim N_{2q}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n),$$

where $\boldsymbol{I}_m$ denotes a m-dimensional identity matrix.

In the first simulation study, we simulate the above described model for three different covariance matrices, corresponding to three different values of the constant in (I.14). In that way, we can examine the sensitivity in the normality of the estimates to an increase in the variance. Theorem I.2.1 states that under some regularity conditions, the estimated values of $\boldsymbol{\beta}$ should be Gaussian distributed when the variance of the random components goes to zero. That is, the lower the constant in (I.14) is, the closer is the distribution of $\boldsymbol{\beta}$ to a Gaussian distribution. In this simulation study, we used the following constants: $c_1 = 1$, $c_2 = 50$ and $c_3 = 100$. In each of the three simulation studies we simulate 500 datasets and estimate the above described model for each simulation. The results are presented in Figure I.1.

In the second simulation study, we fix the covariance matrix of the random components to $\boldsymbol{\Sigma}$ defined in (I.14) with the constant set to one. In this study, we vary the lengths of $\boldsymbol{B}_{(1)}$ and $\boldsymbol{B}_{(2)}$ between the values 10, 50 and 100, whereas the lengths was fixed to 60 in the above described simulation study. For each value of q (the length of each vector of random components), we simulate the model 500 times and estimate the bias and standard errors of the parameters.

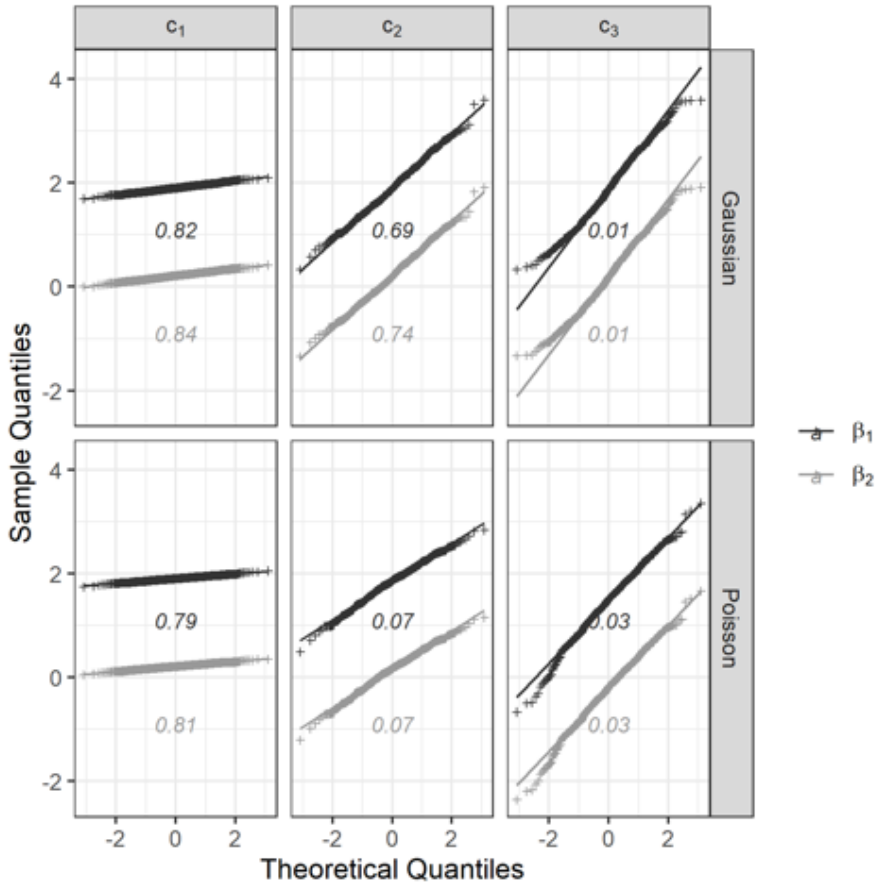


Figure I.1: QQ-plot of the theoretical Gaussian quantiles versus the sample quantiles of the estimated values of β_1 and β_2 in the described multivariate generalised linear mixed model for different sizes of $\mathbf{\Sigma}$. The numbers in the plots are the resulting p-values from Shapiro Wilk tests for normality.

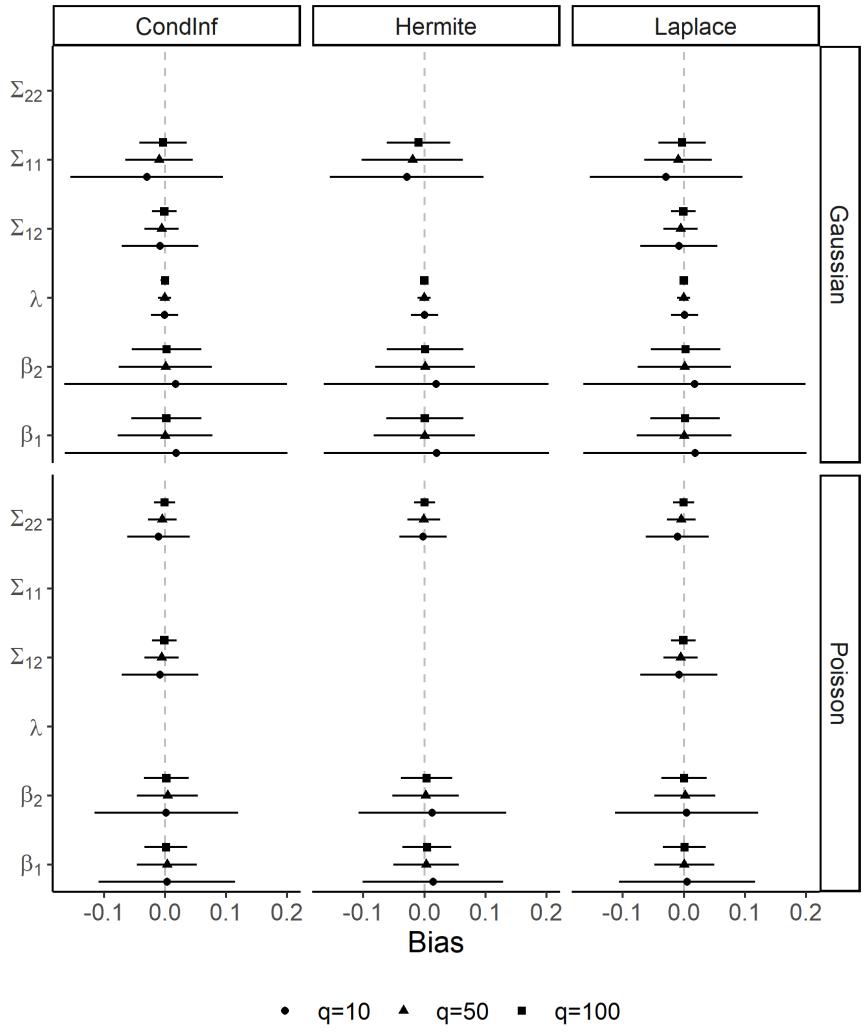


Figure I.2: Estimated bias calculated from simulations of the described model for three different lengths of the vectors of random components using three different inference methods. The error bars show the estimated bias plus/minus the estimated standard errors. The Hermite approximation was applied to each univariate marginal model; therefore, there are no estimates for covariances when using this method.

I.4 Discussion

The inference method introduced in this paper extends the applicability of standard GLMMs in two ways: first, it allows for defining and inferring multivariate GLMMs, provided there exist random components representing clusters of observations defined in the same way in each of the marginal GLMMs; second, it allows to use non-Gaussian distributions for the random components.

Remarkably, the marginal models of the defined MGLMMs can be of different statistical nature and at the same time represent complex dependence structures. Therefore, those models provide a rather flexible tool for applications. For instance, in Pelck & Labouriau (2020) the MGLMM contained marginal GLMMs for binomial and for Poisson distributed responses, which appeared naturally in the process of modelling a system for monitoring the development of roots over time. Moreover, the MGLMM used in Pelck & Labouriau (2020) could be used to detect and represent a first-order Markovian dependence induced by repeated measurements applied at the same experimental units over time (see also Shanmugam et al. (2021) for a similar application on roots development studies). Another example of MGLMMs including marginal GLMMs of different nature can be found in Pelck et al. (2021b), where marginal GLMMs defined with the Gamma, binomial and the compound Poisson families of distributions were used for simultaneously modelling the development of a fungal infection in apples and the concentration of a series of volatile organic compounds, observed along time. In a third study, Pelck et al. (2021a) used MGLMMs to simultaneously describe the students' marks obtained in different admission exams at the University (Gaussian distributed), and the performance in the course of geometry measured as the number of attempts required to pass the course (a Cox proportional model with discrete time). Those examples illustrate the usefulness of the MGLMMs studied in this paper.

The inference method proposed in this paper does not involve integration of conditional likelihood quantities, which might be advantageous with respect to naive integration based methods, as illustrated in the simulation study presented in Section I.3.2. The performance of the new introduced method is similar to the method introduced by Breslow & Clayton (1993), when we assume the random components to be Gaussian distributed. Indeed, when the random components are Gaussian distributed, the inference functions $\psi^*_{\boldsymbol{\beta}}$ and $\psi^*_{\boldsymbol{b}}$ are similar (but not the same) to the approximate score functions used in Breslow & Clayton (1993), which are based on a Laplace approximation of the likelihood function of the GLMM $\mathcal{P}$. In this case, the inference function in (I.10) is equivalent to the score equation of the fixed effects in Breslow & Clayton (1993), whereas the inference function in (I.11) differs from the score equation for the random effects by the additive term $\sigma^2 \boldsymbol{I}_q \boldsymbol{b}$, which has expectation zero. We extend the Laplace approximation method proposed by Breslow & Clayton (1993) to a multivariate context in the Appendix I.A.4.

The GLMMs and MGLMMs described in this paper are constructed using dispersion models instead of exponential dispersion models as usually done in the literature of GLMMs, see Breslow & Clayton (1993) and the literature referred there. We remark that the class of dispersion models defined in Jørgensen (1987), Jørgensen

et al. (1996) is much larger than the class of exponential dispersion models; see Cordeiro et al. (2021) and Labouriau (2020) for a list of examples and a discussion of the extension of the class of dispersion models.

Acknowledgement

The authors were partially financed by the Applied Statistics Laboratory (aStatLab) at the Department of Mathematics, Aarhus University.

References

Agresti, A. (2002), *Categorical Data Analysis*, John Wiley & Sons.

Barndorff-Nielsen, O. (2014), *Information and exponential families: in statistical theory*, John Wiley & Sons.

Breslow, N. E. & Clayton, D. G. (1993), 'Approximate inference in generalized linear mixed models', *Journal of the American statistical Association* **88**(421), 9–25.

Cordeiro, G. M., Labouriau, R. & Botter, D. (2021), 'An introduction to bent jørgensen's ideas', *Brazilian journal of Probability and Statistics* **35**(1), 2–20.

Demidenko, E. (2004), *Mixed Models: Theory and Applications (Wiley Series in Probability and Statistics)*, Wiley-Interscience, USA.

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

Jørgensen, B. (1987), 'Exponential dispersion models', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 127–162.

Jorgensen, B. (1997), *The theory of dispersion models*, CRC Press.

Jørgensen, B. & Labouriau, R. (2012), *Exponential Families and Theoretical Inference*, Vol. 52, 2 edn, Springer.

Jørgensen, B., Labouriau, R. & Lundbye-Christensen, S. (1996), 'Linear growth curve analysis based on exponential dispersion models', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 573–592.

Labouriau, R. (2020), 'Construction and extension of dispersion models'. arXiv:2008.05448.

McCulloch, C. E. (1997), 'Maximum likelihood algorithms for generalized linear mixed models', *Journal of the American statistical Association* **92**(437), 162–170.

McCulloch, C. & Searle, S. (2001), *Generalized, Linear, and Mixed Models*, John Wiley & Sons.

Pelck, J. S. & Labouriau, R. (2020), 'Using multivariate generalised linear mixed models for studying roots development: An example based on minirhizotron observations'. arXiv:2011.00546.

Pelck, J. S., Luca, A., Holthusen, H., Edelenbos, M. & Labouriau, R. (2021b), 'Multivariate method for detection of rubbery rot in storage apples by monitoring volatile organic compounds: An example of multivariate generalised linear mixed models'. arXiv:2107.11233.

Pelck, J. S., Maia, R. P., Pinheiro, H. P. & Labouriau, R. (2021a), 'A multivariate methodology for analysing students' performance using register data'. arXiv:2102.10565.

Shanmugam, S., Hefner, M., Pelck, J., Labouriau, R. & Kristensen, H. (2021), 'Complementary resource use in intercropped faba bean and cabbage by increased root growth and nitrogen use in organic production'. Submitted.

I.A Appendix

I.A.1 On the identifiability of the family of conditional densities $\mathcal{P}^*$

Here, we show that the family of conditional densities given by (I.5) is not identifiable parametrised by $(\boldsymbol{\beta}, \boldsymbol{b}, \lambda) \in \Omega \times \mathbb{R}^q \times \Lambda$.

Lemma I.A.1. *For any $i \in \{1, \ldots, n\}$ and any choice of $\boldsymbol{\beta}$, $\boldsymbol{b}$ and $\delta > 0$, there exist $\boldsymbol{\beta}_\delta \in \Omega$ such that $\eta_i(\boldsymbol{\beta}, \boldsymbol{b}) = \eta_i(\boldsymbol{\beta}_\delta, \boldsymbol{b} - \delta)$.*

Proof. Take arbitrary i, $(\boldsymbol{\beta}, \boldsymbol{b})$ and $\delta > 0$. Note that $\boldsymbol{z}_i^T(\boldsymbol{b} - \delta) = \boldsymbol{z}_i^T\boldsymbol{b} - \delta$ because, by construction, there is one entry of the allocation vector $\boldsymbol{z}_i$ that is equal to one and the other entries vanish. Assume, without loss of generality, that the first entry of the vector $\boldsymbol{x}_i$ is equal to 1 (*i.e.,* the fixed effect of the GLMM contains an intercept) so that $\boldsymbol{x}_i^T\boldsymbol{\beta} = \boldsymbol{\beta}_1 + \tilde{\boldsymbol{x}}_i^T\tilde{\boldsymbol{\beta}}$, where $\tilde{\boldsymbol{x}}_i$ and $\tilde{\boldsymbol{\beta}}$ are the $(k-1)$-dimensional vectors obtained by eliminating the first entry of $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$, respectively. Taking $\boldsymbol{\beta}_\delta = (\boldsymbol{\beta}_1 + \delta, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_k)$ we have that

$$\begin{aligned}\eta_i(\boldsymbol{\beta}, \boldsymbol{b}) &= \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{b} = \boldsymbol{\beta}_1 + \tilde{\boldsymbol{x}}_i^T\tilde{\boldsymbol{\beta}} + \boldsymbol{z}_i^T\boldsymbol{b} = (\boldsymbol{\beta}_1 + \delta) + \tilde{\boldsymbol{x}}_i^T\tilde{\boldsymbol{\beta}} + \boldsymbol{z}_i^T(\boldsymbol{b} - \delta) \\ &= \boldsymbol{x}_i^T\boldsymbol{\beta}_\delta + \boldsymbol{z}_i^T(\boldsymbol{b} - \delta) = \eta_i(\boldsymbol{\beta}_\delta, \boldsymbol{b} - \delta)\end{aligned}$$

The proof follows since i, $(\boldsymbol{\beta}, \boldsymbol{b})$ and δ were taken arbitrarily. □

I.A.2 Technical Proofs of the Asymptotic Distribution of the Conditional Inference Based Estimates

In this appendix, we present a sequence of lemmas and propositions that will culminate with the proof of the Theorem I.2.1, which establishes consistency and joint asymptotic normality of the proposed estimator of $\boldsymbol{\beta}$ and the predictor of $\boldsymbol{b}$ for small values of the variance of the random components.

Regular Inference Functions

We recall the definition of regular inference functions used in this appendix for the easy of the reader (see the details in Jørgensen & Labouriau, 2012, Chapter 4, from which we draw heavily). Consider a parametric family of distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$ and a σ-finite measure μ defined on a given measurable space $(\mathcal{X}, \mathcal{A})$. For each $P_\theta \in \mathcal{P}$, we chose a version of the Radon-Nikodym derivative (with respect to μ), denoted by

$$p(\cdot;\theta) = \frac{dP_\theta}{d\mu}(\cdot).$$

Definition 1. *A function $\Psi : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}^k$ is said to be a regular inference function when the following conditions are satisfied for all $\theta = (\theta_1, \ldots, \theta_k) \in \Theta$ and for $i, j = 1, \ldots, k$.*

(i) $\mathbb{E}_\theta[\Psi(\theta)] = 0$*;*

(ii) The partial derivative $\partial\Psi(x;\theta)/\partial\theta_i$ exists for μ-almost every $x \in \mathcal{X}$;

(iii) The order of integration and differentiation may be interchanged as follows:

$$\frac{\partial}{\partial\theta_i}\int_{\mathcal{X}} \Psi(x;\theta)p(x;\theta)d\mu(x) = \int_{\mathcal{X}} \frac{\partial}{\partial\theta_i}[\Psi(x;\theta)p(x;\theta)]d\mu(x)\,;$$

(iv) $\mathbb{E}\{\psi_i(\theta)\psi_j(\theta)\} \in \mathbb{R}$ *and the $k \times k$ matrix*

$$V_\psi(\theta) = \mathbb{E}\{\Psi(\theta)\Psi^T(\theta)\}$$

is positive definite;

(v) $\mathbb{E}\{\frac{\partial\psi_i}{\partial\theta_r}(\theta)\frac{\partial\psi_j}{\partial\theta_s}(\theta)\} \in \mathbb{R}$ *and the $k \times k$ matrix*

$$S_\psi(\theta) = \mathbb{E}\{\nabla_\theta\Psi(\theta)\}$$

is nonsingular.

Here ψ_i denoted the i^{th} component of the vector function

$$\Psi(\cdot) = (\psi_1(\cdot), \ldots, \psi_k(\cdot))^T\,,$$

and ∇_θ denotes the gradient operator relative to the vector θ, defined by

$$\nabla_\theta f(\theta) = \frac{\partial f}{\partial\theta^T}(\theta).$$

Some Key Lemmas

We denote the sequences of roots of the inference functions $\psi^*_{\boldsymbol{\beta}}$ and $\psi^*_{\boldsymbol{b}}$ by $\{\widehat{\boldsymbol{\beta}}_n\}_{n\in\mathbb{N}}$ and $\{\widehat{\boldsymbol{b}}_n\}_{n\in\mathbb{N}}$ respectively, obtained when the number of observations, n, increases. Moreover, define $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{b})$ and $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{b}}_n)$ (for each $n \in \mathbb{N}$). Recall, that the inference function $\psi^* : \Omega \times \mathbb{R}^q \times \mathcal{Y} \to \mathbb{R}^{k+q}$ for estimating $\boldsymbol{\theta}$ under $\mathcal{P}^*$, is defined by

$$\psi^*(\boldsymbol{\beta}, \boldsymbol{b}) = \left\{ \left[\psi^*_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{b})\right]^T, [\psi^*_{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{b})]^T \right\}^T,$$

for all $\boldsymbol{\beta} \in \Omega$ and $\boldsymbol{b} \in \mathbb{R}^q$.

Lemma I.A.2. *Under the regularity conditions i-iv, the partial inference functions $\psi^*_{\boldsymbol{\beta}}$ and $\psi^*_{\boldsymbol{b}}$ are unbiased, that is,*

$$\mathbb{E}_{P^*_{\boldsymbol{\beta},\boldsymbol{b},\lambda}}\left[\psi^*_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{Y}\right)\right] = \mathbf{0},$$
$$\mathbb{E}_{P^*_{\boldsymbol{\beta},\boldsymbol{b},\lambda}}\left[\psi^*_{\boldsymbol{b}}\left(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{Y}\right)\right] = \mathbf{0},$$

for all $\boldsymbol{\beta} \in \Omega$, $\boldsymbol{b} \in \mathbb{R}^q$ and $\lambda \in \Lambda$. Moreover, the partial inference functions $\psi^*_{\boldsymbol{\beta}}$ and $\psi^*_{\boldsymbol{b}}$, are regular.

Proof. We show that $\psi^*_{\boldsymbol{\beta}}$ is unbiased since the unbiasedness of $\psi^*_{\boldsymbol{b}}$ follows from the same arguments. Take arbitrarily $\boldsymbol{\beta} \in \Omega$, $\boldsymbol{b} \in \mathbb{R}^q$ and $\lambda \in \Lambda$. We aim to show that

$$0 = \int_{\mathcal{Y}} \psi^*_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{y}) f^*(\boldsymbol{y}; \boldsymbol{\beta}, \boldsymbol{b}, \lambda) d\boldsymbol{\nu}(\boldsymbol{y}).$$

The regularity conditions ensure that it is allowed to interchange the order of differentiation and integration in the following:

$$\begin{aligned}
&\int_{\mathcal{Y}} \psi^*_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{y}) f^*(\boldsymbol{y}; \boldsymbol{\beta}, \boldsymbol{b}, \lambda) d\boldsymbol{\nu}(\boldsymbol{y}) \\
&\qquad = \sum_{i=1}^{n} \int_{\mathcal{Y}} \tfrac{\partial}{\partial \boldsymbol{\beta}} \{d(y_i; g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta} + \tilde{\boldsymbol{z}}_i^T \boldsymbol{b}))\} f^*(y_i; \boldsymbol{\beta}, \boldsymbol{b}, \lambda) d\nu(y_i) \\
&\qquad = -2\lambda \sum_{i=1}^{n} \int_{\mathcal{Y}} \frac{\partial}{\partial \boldsymbol{\beta}} f^*(y_i; \boldsymbol{\beta}, \boldsymbol{b}, \lambda) d\nu(y_i) \\
&\qquad = -2\lambda \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}} \int_{\mathcal{Y}} f^*(y_i; \boldsymbol{\beta}, \boldsymbol{b}, \lambda) d\nu(y_i) = 0.
\end{aligned}$$

The proof follows since $\boldsymbol{\beta} \in \Omega$, $\boldsymbol{b} \in \mathbb{R}^q$ and $\lambda \in \Lambda$ are arbitrarily chosen.

The other regularity conditions for the inference functions follow straightforwardly from the assumed regularity conditions i-iv for the GLMM in play. □

We introduce some required notation before presenting the next lemma. Define the sensitivity block matrices

$$\begin{aligned}
S_{\boldsymbol{\beta b}} &= \mathbb{E}[\nabla_{\boldsymbol{b}} \psi^*_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{Y})], \quad S_{\boldsymbol{b\beta}} = \mathbb{E}[\nabla_{\boldsymbol{\beta}} \psi^*_{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{Y})], \\
S_{\boldsymbol{bb}} &= \mathbb{E}[\nabla_{\boldsymbol{b}} \psi^*_{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{Y})], \quad S_{\boldsymbol{\beta\beta}} = \mathbb{E}[\nabla_{\boldsymbol{\beta}} \psi^*_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{b}; \boldsymbol{Y})],
\end{aligned}$$

and the variability matrices

$$\begin{aligned} V_{\boldsymbol{\beta b}^*} &= \mathbb{E}[\psi^*_{\boldsymbol{b}}(\boldsymbol{\beta},\boldsymbol{b};\boldsymbol{Y})\psi^*_{\boldsymbol{\beta}}(\boldsymbol{\beta},\boldsymbol{b};\boldsymbol{Y})^T], \quad V_{\boldsymbol{b}^*\boldsymbol{\beta}} = \mathbb{E}[\psi^*_{\boldsymbol{\beta}}(\boldsymbol{\beta},\boldsymbol{b};\boldsymbol{Y})\psi^*_{\boldsymbol{b}}(\boldsymbol{\beta},\boldsymbol{b};\boldsymbol{Y})^T], \\ V_{\boldsymbol{bb}} &= \mathbb{E}[\psi^*_{\boldsymbol{b}}(\boldsymbol{\beta},\boldsymbol{b};\boldsymbol{Y})\psi^*_{\boldsymbol{b}}(\boldsymbol{\beta},\boldsymbol{b};\boldsymbol{Y})^T], \quad V_{\boldsymbol{\beta\beta}} = \mathbb{E}[\psi^*_{\boldsymbol{\beta}}(\boldsymbol{\beta},\boldsymbol{b};\boldsymbol{Y})\psi^*_{\boldsymbol{\beta}}(\boldsymbol{\beta},\boldsymbol{b};\boldsymbol{Y})^T]. \end{aligned}$$

Using these, we define

$$\begin{aligned} W &= D^{-1} = S_{\boldsymbol{bb}} - S_{\boldsymbol{\beta b}}S^{-1}_{\boldsymbol{\beta\beta}}S_{\boldsymbol{b\beta}}, \quad & A &= S^{-1}_{\boldsymbol{\beta\beta}} + S^{-1}_{\boldsymbol{\beta\beta}}S_{\boldsymbol{b\beta}}W^{-1}S_{\boldsymbol{\beta b}}S^{-1}_{\boldsymbol{\beta\beta}}, \\ E &= -S^{-1}_{\boldsymbol{\beta\beta}}S_{\boldsymbol{b\beta}}W^{-1}, & C &= -W^{-1}S_{\boldsymbol{\beta b}}S^{-1}_{\boldsymbol{\beta\beta}}. \end{aligned}$$

Lemma I.A.3. *The inverse Godambde information for the inference function ψ^* is the matrix-valued function $J^{-1}_{\psi^*} : \Omega \times \mathbb{R}^q \to \mathbb{R}^{(k+q)\times(k+q)}$ defined by*

$$J^{-1}_{\psi^*} = \begin{bmatrix} J^{-1}_{\psi^*_{\boldsymbol{\beta}}} & (J^{-1}_{\psi^*_{\boldsymbol{\beta b}}})^T \\ J^{-1}_{\psi^*_{\boldsymbol{\beta b}}} & J^{-1}_{\psi^*_{\boldsymbol{b}}} \end{bmatrix},$$

with

$$\begin{aligned} J^{-1}_{\psi^*_{\boldsymbol{\beta}}} &= AV_{\boldsymbol{\beta\beta}}A^T + EV_{\boldsymbol{\beta b}}A^T + AV_{\boldsymbol{b}^*\boldsymbol{\beta}}E^T + EV_{\boldsymbol{bb}}E^T \\ J^{-1}_{\psi^*_{\boldsymbol{\beta b}}} &= CV_{\boldsymbol{\beta\beta}}A^T + DV_{\boldsymbol{\beta b}}A^T + CV_{\boldsymbol{b\beta}}E^T + DV_{\boldsymbol{bb}}E^T \\ J^{-1}_{\psi^*_{\boldsymbol{b}}} &= CV_{\boldsymbol{\beta\beta}}C^T + DV_{\boldsymbol{\beta b}^*}C^T + CV_{\boldsymbol{b\beta}}D^T + DV_{\boldsymbol{bb}}D^T. \end{aligned}$$

for all $\boldsymbol{\beta} \in \Omega$ and $\boldsymbol{b} \in \mathbb{R}^q$ using the above introduced notation.

Proof. The result follows from the formulas in Chapter 4 in Jørgensen & Labouriau (2012) and inversion of block matrices. □

Lemma I.A.4. *Assume the regularity conditions i-iv. Then, for all $\boldsymbol{\beta} \in \Omega$, $\boldsymbol{b} \in \mathbb{R}^q$ and $\lambda \in \Lambda$, it is true that*

$$\widehat{\boldsymbol{\beta}}_n \xrightarrow[n\to\infty]{P^*_{\boldsymbol{\beta},\boldsymbol{b},\lambda}} \boldsymbol{\beta} \text{ and } \widehat{\boldsymbol{b}}_n \xrightarrow[n\to\infty]{P^*_{\boldsymbol{\beta},\boldsymbol{b},\lambda}} \boldsymbol{b}\,.$$

Moreover,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})|\boldsymbol{B} = \boldsymbol{b} \xrightarrow[n\to\infty]{\mathcal{D}} N_{k+q}(\boldsymbol{0}, J^{-1}_{\psi^*}(\boldsymbol{\beta},\boldsymbol{b})),$$

implying that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})|\boldsymbol{B} = \boldsymbol{b} \xrightarrow[n\to\infty]{\mathcal{D}} N_k(\boldsymbol{0}, J^{-1}_{\psi^*_{\boldsymbol{\beta}}}(\boldsymbol{\beta},\boldsymbol{b}))$$

and

$$\sqrt{n}(\widehat{\boldsymbol{b}}_n - \boldsymbol{b})|\boldsymbol{B} = \boldsymbol{b} \xrightarrow[n\to\infty]{\mathcal{D}} N_q(\boldsymbol{0}, J^{-1}_{\psi^*_{\boldsymbol{b}}}(\boldsymbol{\beta},\boldsymbol{b}))\,.$$

Proof. The proof follows from the results in Chapter 4 in Jørgensen & Labouriau (2012), and the fact that $\psi^*_{\boldsymbol{\beta}}$ and $\psi^*_{\boldsymbol{b}}$ are regular inference functions as a consequence of Lemma I.A.2. □

On the asymptotic variance of $\widehat{\boldsymbol{\theta}}_n$ under the family $\mathcal{P}$

Lemma I.A.5. *Assume the regularity conditions i-iv. The partial solution $\{\widehat{\boldsymbol{\beta}}_n\}_{n\in\mathbb{N}}$ of $\psi^*_{\boldsymbol{\beta}}$ is also a solution to $\psi_{\boldsymbol{\beta}} = 0$ defined in* (I.12)*, and the unconditionally asymptotic covariance matrices (for n converging to infinity and q fixed), denoted AV, of $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{b}}_n$ are given by*

$$AV\,(\widehat{\boldsymbol{\beta}}_n) = \mathbb{E}[J^{-1}_{\psi^*_{\boldsymbol{\beta}}}(\boldsymbol{\beta}, \boldsymbol{B})] + \mathbb{V}[\widehat{\boldsymbol{\beta}}_n(\boldsymbol{B})], \tag{I.15}$$

$$AV\,(\widehat{\boldsymbol{b}}_n) = \mathbb{E}[J^{-1}_{\psi^*_{\boldsymbol{b}}}(\boldsymbol{\beta}, \boldsymbol{B})] + \boldsymbol{I}_q\sigma^2, \tag{I.16}$$

with $\widehat{\boldsymbol{\beta}}_n(\boldsymbol{B})$ denoting the estimator of $\boldsymbol{\beta}$ as a function of $\boldsymbol{B}$ for all $n \in \mathbb{N}$, $\boldsymbol{\beta} \in \Omega$ and $\boldsymbol{B} \in \mathbb{R}^q$. Moreover,

$$\widehat{\boldsymbol{\beta}}_n \xrightarrow[n\to\infty]{P_{\boldsymbol{\beta},\lambda,\sigma^2}} \boldsymbol{\beta},$$

for all $\boldsymbol{\beta} \in \Omega$, $\lambda \in \Lambda$ and $\sigma^2 \in \mathbb{R}_+$.

Proof. If $\widehat{\boldsymbol{\beta}}_n$ is a solution to (I.12) then it is also a solution to (I.10) when inserting $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{b}}_n$ for a given $n \in \mathbb{N}$.

Take $\boldsymbol{\beta} \in \Omega$, $\lambda \in \Lambda$ and $\sigma^2 \in \mathbb{R}_+$ arbitrarily. The asymptotic covariance matrices follows from the law of total variance and Lemma I.A.4, which also implies that for all $\epsilon > 0$

$$\begin{aligned} &P_{\boldsymbol{\beta},\lambda,\sigma^2}(|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}| > \epsilon) \\ &\quad = \int_{\mathbb{R}^q} P^*_{\boldsymbol{\beta},\boldsymbol{b},\lambda}\Big(|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}| > \epsilon \;\Big|\; \boldsymbol{B} = \boldsymbol{b}\Big) \prod_{j=1}^{q} \varphi(b_j; \sigma^2) d\boldsymbol{b} \xrightarrow[n\to\infty]{} 0, \end{aligned}$$

since $P^*_{\boldsymbol{\beta},\boldsymbol{b},\lambda}(|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}| > \epsilon \,\big|\, \boldsymbol{B} = \boldsymbol{b}) \xrightarrow[n\to\infty]{} 0$ for all $\epsilon > 0$ and $\boldsymbol{b} \in \mathbb{R}^q$. By the regularity assumptions i-iv, we can interchange the order of limit and integration. The proof follows since $\boldsymbol{\beta} \in \Omega$, $\lambda \in \Lambda$ and $\sigma^2 \in \mathbb{R}_+$ are arbitrarily chosen. □

Often the distribution of the random components can be easily simulated in a computational efficient way (*e.g.*, when the random components are normally or t-distributed). In those cases, the expectations and variances referred in (I.15) and (I.16) can be easily obtained using Monte Carlo methods (this includes simulations of $\boldsymbol{B}$ and calculations of estimates of $\boldsymbol{\beta}$ as a function of the simulated values).

Proof of the Theorem I.2.1

The lemma below provides the calculation of the characteristic function of the asymptotic distribution of the sequence of estimated values of $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{b}}_n$, which will be crucial to prove Theorem I.2.1.

Lemma I.A.6. *Assume the regularity conditions i-iv. There exist two random vectors $\boldsymbol{Z}_{\boldsymbol{\beta}}$ and $\boldsymbol{Z}_{\boldsymbol{b}}$ with characteristic functions*

$$\mathbb{E}[\exp(i\boldsymbol{t}_1^T\boldsymbol{Z}_{\boldsymbol{\beta}})] = \mathbb{E}[\exp(-\tfrac{1}{2}\boldsymbol{t}_1^T J^{-1}_{\psi^*_{\boldsymbol{\beta}}}(\boldsymbol{\beta},\boldsymbol{B})\boldsymbol{t}_1)],\ \textit{for all } \boldsymbol{t}_1 \in \mathbb{R}^k,$$
$$\mathbb{E}[\exp(i\boldsymbol{t}_2^T\boldsymbol{Z}_{\boldsymbol{b}})] = \mathbb{E}[\exp(-\tfrac{1}{2}\boldsymbol{t}_2^T J^{-1}_{\psi^*_{\boldsymbol{b}}}(\boldsymbol{\beta},\boldsymbol{B})\boldsymbol{t}_2)],\ \textit{for all } \boldsymbol{t}_2 \in \mathbb{R}^q,$$

respectively, such that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow[n\to\infty]{\mathcal{D}} \boldsymbol{Z}_{\boldsymbol{\beta}} \textit{ and } \sqrt{n}(\widehat{\boldsymbol{b}}_n - \boldsymbol{b}) \xrightarrow[n\to\infty]{\mathcal{D}} \boldsymbol{Z}_{\boldsymbol{b}}\,.$$

Proof. By Lemma I.A.4 we have that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})|\boldsymbol{B} = \boldsymbol{b} \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}_k\Big((\boldsymbol{0}, J^{-1}_{\psi^*_{\boldsymbol{\beta}}}(\boldsymbol{\beta},\boldsymbol{b})\Big).$$

Let $\boldsymbol{Z}_{\boldsymbol{\beta}}$ denote a random variable distributed according to the above defined conditional asymptotically Gaussian distribution. By the Portmanteau theorem the above is equivalent to

$$\mathbb{E}\Big[h\Big(\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})\Big)|\boldsymbol{B} = \boldsymbol{b}\Big] \xrightarrow[n\to\infty]{} \mathbb{E}\Big[h(\boldsymbol{Z}_{\boldsymbol{\beta}})|\boldsymbol{B} = \boldsymbol{b}\Big]$$

for all continuous bounded functions h. Thus, we have that

$$\begin{aligned}\mathbb{E}\Big[h\Big(\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})\Big)\Big] &= \int_{\mathbb{R}^q} \mathbb{E}\Big[h\Big(\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})\Big)|\boldsymbol{B} = \boldsymbol{b}\Big] \prod_{j=1}^{q} \varphi(b_j;\sigma^2) d\boldsymbol{b} \xrightarrow[n\to\infty]{} \\ &\int_{\mathbb{R}^q} \mathbb{E}[h(\boldsymbol{Z}_{\boldsymbol{\beta}})|\boldsymbol{B} = \boldsymbol{b}] \prod_{j=1}^{q} \varphi(b_j;\sigma^2) d\boldsymbol{b} \\ &= \mathbb{E}[h(\boldsymbol{Z}_{\boldsymbol{\beta}})],\end{aligned}$$

since we can interchange the order of limit and integration due to the assumed regularity conditions. Therefore, we conclude that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow[n\to\infty]{\mathcal{D}} \boldsymbol{Z}_{\boldsymbol{\beta}}.$$

The characteristic function of $\boldsymbol{Z}_{\boldsymbol{\beta}}$ is given by:

$$\begin{aligned}\mathbb{E}[\exp(i\boldsymbol{t}_1^T\boldsymbol{Z}_{\boldsymbol{\beta}})] &= \mathbb{E}[\mathbb{E}[\exp(i\boldsymbol{t}_1^T\boldsymbol{Z}_{\boldsymbol{\beta}})|\boldsymbol{B}]] \\ &= \mathbb{E}[\exp(-\tfrac{1}{2}\boldsymbol{t}_1^T J^{-1}_{\psi^*_{\boldsymbol{\beta}}}(\boldsymbol{\beta},\boldsymbol{B})\boldsymbol{t}_1)],\ \text{for all } \boldsymbol{t}_1 \in \mathbb{R}^k.\end{aligned}$$

The proof for $\widehat{\boldsymbol{b}}_n$ follows by similar arguments by changing $\widehat{\boldsymbol{\beta}}_n$ to $\widehat{\boldsymbol{b}}_n$, and $\boldsymbol{Z}_{\boldsymbol{\beta}}$ to $\boldsymbol{Z}_{\boldsymbol{b}}$ (by changing $J^{-1}_{\psi^*_{\boldsymbol{\beta}}}(\boldsymbol{\beta},\boldsymbol{B})$ to $J^{-1}_{\psi^*_{\boldsymbol{b}}}(\boldsymbol{\beta},\boldsymbol{B})$) in the above. □

The theorem below corresponds to the second part of Theorem I.2.1.

Theorem I.A.7. *Under the regularity conditions i-iv, the sequences* $\{\hat{\boldsymbol{\beta}}_n\}_{n\in\mathbb{N}}$ *and* $\{\hat{\boldsymbol{b}}_n\}_{n\in\mathbb{N}}$ *are asymptotically Gaussian distributed, when* $n \to \infty$ *and* $\sigma^2 \to 0+$ *in the following way*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow[\substack{n\to\infty \\ \sigma^2\to 0+}]{\mathcal{D}} N_k(\mathbf{0}, J^{-1}_{\psi^*_{\boldsymbol{\beta}}}(\boldsymbol{\beta}, \mathbf{0})),$$

and

$$\sqrt{n}(\hat{\boldsymbol{b}}_n - \boldsymbol{b}) \xrightarrow[\substack{n\to\infty \\ \sigma^2\to 0+}]{\mathcal{D}} N_q(\mathbf{0}, J^{-1}_{\psi^*_{\boldsymbol{b}}}(\boldsymbol{\beta}, \mathbf{0}))\,.$$

Proof. Consider the characteristic function of $\boldsymbol{Z}_{\boldsymbol{\beta}}$ found in Lemma I.A.6:

$$\mathbb{E}[\exp(i\boldsymbol{t}^T\boldsymbol{Z}_{\boldsymbol{\beta}})] = \mathbb{E}[\exp(-\tfrac{1}{2}\boldsymbol{t}^T J^{-1}_{\psi^*_{\boldsymbol{\beta}}}(\boldsymbol{\beta}, \boldsymbol{B})\boldsymbol{t})], \text{ for all } \boldsymbol{t} \in \mathbb{R}^k. \tag{I.17}$$

Using a first order Taylor approximation, we find that

$$\begin{aligned}
\exp\Big(-\tfrac{1}{2}\boldsymbol{t}^T J^{-1}_{\psi^*_{\boldsymbol{\beta}}}(\boldsymbol{\beta}, \boldsymbol{B})\boldsymbol{t}\Big) &= \exp\Big(-\tfrac{1}{2}\sum_{i=1}^k\sum_{j=1}^k t_i t_j \{J^{-1}_{\psi^*_{\boldsymbol{\beta}}}(\boldsymbol{\beta}, \boldsymbol{B})\}_{ij}\Big) \\
&= \exp\Big(-\tfrac{1}{2}\sum_{i=1}^k\sum_{j=1}^k t_i t_j \{J^{-1}_{\psi^*_{\boldsymbol{\beta}}}(\boldsymbol{\beta}, \mathbf{0})\}_{ij}\Big) \\
&\quad + \exp(-\tfrac{1}{2}\sum_{i=1}^k\sum_{j=1}^k t_i t_j \{J^{-1}_{\psi^*_{\boldsymbol{\beta}}}(\boldsymbol{\beta}, \mathbf{0})\}_{ij}) \\
&\quad \times \Big(-\tfrac{1}{2}\Big)\boldsymbol{B}^T \sum_{i=1}^k\sum_{j=1}^k t_i t_j \frac{\partial\{J^{-1}_{\psi^*_{\boldsymbol{\beta}}}\}_{ij}}{\partial \boldsymbol{b}}(\boldsymbol{\beta}, \mathbf{0}) \\
&\quad + R(\boldsymbol{B}), \qquad \text{for all } \boldsymbol{t} \in \mathbb{R}^k,
\end{aligned}$$

where $R(\cdot)$ is the remainder term which converges to zero when $\boldsymbol{B}$ converges to zero. Thus, for σ^2 converging to zero, $\boldsymbol{B}$ converges to the expectation which is zero. This imply, that the remainder term converges to zero. Notice, that the second term has expectation zero since $\mathbb{E}[\boldsymbol{B}] = 0$, so inserting the above in (I.17) yields

$$\mathbb{E}_{\boldsymbol{Z}_{\boldsymbol{\beta}}}[\exp(i\boldsymbol{t}^T\boldsymbol{Z}_{\boldsymbol{\beta}})] = \exp(-\tfrac{1}{2}\boldsymbol{t}^T J^{-1}_{\psi^*_{\boldsymbol{\beta}}}(\boldsymbol{\beta}, \mathbf{0})\boldsymbol{t}) + R(\boldsymbol{B}) \underset{\sigma^2\to 0+}{\longrightarrow} \exp(-\tfrac{1}{2}\boldsymbol{t}^T J^{-1}_{\psi^*_{\boldsymbol{\beta}}}(\boldsymbol{\beta}, \mathbf{0})\boldsymbol{t}).$$

This proves that the asymptotically distribution of $\{\hat{\boldsymbol{\beta}}_n\}_{n\in\mathbb{N}}$ converges to a Gaussian distribution when σ^2 converges to zero. The argument for $\{\hat{\boldsymbol{b}}_n\}_{n\in\mathbb{N}}$ is equivalent and follows by changing $\hat{\boldsymbol{\beta}}_n$ to $\hat{\boldsymbol{b}}_n$ and $\boldsymbol{Z}_{\boldsymbol{\beta}}$ to $\boldsymbol{Z}_{\boldsymbol{b}}$ (changing $J^{-1}_{\psi^*_{\boldsymbol{\beta}}}(\boldsymbol{\beta}, \boldsymbol{B})$ to $J^{-1}_{\psi^*_{\boldsymbol{b}}}(\boldsymbol{\beta}, \boldsymbol{B})$) in the above. □

I.A.3 Variance Estimation in for Models with Gaussian Random Components

In this section, we calculate the integral in Equation (I.13) under the assumption that

$$g(\widehat{\boldsymbol{b}}; \boldsymbol{b}, \boldsymbol{\Sigma}_{\widehat{\boldsymbol{b}}}) = |2\pi\boldsymbol{\Sigma}_{\widehat{\boldsymbol{b}}}|^{-\frac{1}{2}} \exp\Big(-\tfrac{1}{2}(\widehat{\boldsymbol{b}} - \boldsymbol{b})^T \boldsymbol{\Sigma}_{\widehat{\boldsymbol{b}}}^{-1}(\widehat{\boldsymbol{b}} - \boldsymbol{b})\Big)$$
$$\varphi(b; \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\tfrac{1}{2\sigma^2} b^2).$$

Plugging into the integral yields

$$\begin{aligned}
&\int_{\mathbb{R}^q} g(\hat{\boldsymbol{b}}; \boldsymbol{b}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{b}}}) \varphi(\boldsymbol{b}; (\boldsymbol{I}_q - \tfrac{1}{q}\boldsymbol{E}_q)\sigma^2) d\boldsymbol{b} \\
&\quad = \int_{\mathbb{R}^q} |2\pi\boldsymbol{\Sigma}_{\widehat{\boldsymbol{b}}}|^{-\frac{1}{2}} \exp\Big(-\tfrac{1}{2}(\widehat{\boldsymbol{b}} - \boldsymbol{b})^T \boldsymbol{\Sigma}_{\widehat{\boldsymbol{b}}}^{-1}(\widehat{\boldsymbol{b}} - \boldsymbol{b})\Big) |2\pi\sigma^2 \boldsymbol{I}_q|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}\boldsymbol{b}^T \tfrac{1}{\sigma^2}\boldsymbol{I}_q \boldsymbol{b}) d\boldsymbol{b} \\
&\quad = |2\pi\boldsymbol{\Sigma}_{\widehat{\boldsymbol{b}}_k}|^{-\frac{1}{2}} (2\pi\sigma^2)^{-\frac{q}{2}} \exp\Big(-\tfrac{1}{2}\widehat{\boldsymbol{b}}^T \boldsymbol{\Sigma}_{\widehat{\boldsymbol{b}}}^{-1} \widehat{\boldsymbol{b}}\Big) \Big| 2\pi[\boldsymbol{\Sigma}_{\widehat{\boldsymbol{b}}}^{-1} + \tfrac{1}{\sigma^2}\boldsymbol{I}_q]^{-1} \Big|^{\frac{1}{2}} \\
&\qquad \times \exp\Big(\tfrac{1}{2}\widehat{\boldsymbol{b}}^T \boldsymbol{\Sigma}_{\widehat{\boldsymbol{b}}}^{-1} [\boldsymbol{\Sigma}_{\widehat{\boldsymbol{b}}}^{-1} + \tfrac{1}{\sigma^2}\boldsymbol{I}_q]^{-1} \boldsymbol{\Sigma}_{\widehat{\boldsymbol{b}}}^{-1} \widehat{\boldsymbol{b}}\Big).
\end{aligned}$$

In the case of multiple random components, we maximise the integral above for each random component. If the random components are nested, we only predict values for the random components with the highest number of clusters and then uses least squares to predict values for each random component, see Section I.2.5. Therefore, the calculations above are changed by replacing $\sigma^2\boldsymbol{I}_q$ with $\sum_{j=1}^K \sigma^2_{\boldsymbol{B}_j} \boldsymbol{C}_j \boldsymbol{C}_j^T$, where $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_K$ denotes the $K \in \mathbb{N}$ nested random components, and $\boldsymbol{C}_m$ the $q \times q_m$ dimensional matrix specifying for each level l (l^{th} row) which entry of $\boldsymbol{B}_m$ that enters the lth entry of $\widehat{\boldsymbol{b}}$. Here q_m is the dimension of the random vector $\boldsymbol{B}_m$.

In the multivariate model described in Section I.3.1, the above integral can be adapted by letting $\widehat{\boldsymbol{b}}^T = (\widehat{\boldsymbol{b}}_{(1)}^T, \ldots, \widehat{\boldsymbol{b}}_{(d)}^T)$ (and thus changing the dimension of $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{b}}}$) and replacing $\sigma^2\boldsymbol{I}_q$ with $\boldsymbol{\Sigma} \otimes \boldsymbol{I}_q$.

I.A.4 Multivariate Extension of the Laplace Approximation Method

We outline how the Laplace approximation in Breslow & Clayton (1993) can be extended to the multivariate model described in Section I.3.1, when the random components follow a multivariate Gaussian distribution. This extension follows directly from Breslow & Clayton (1993) by redefining some matrices and vectors. We shortly describe how this was done in the simulation study in Section I.3.2. The extension given below assumes that the marginal GLMMs are defined with exponential dispersion models (as in Breslow & Clayton (1993)) but this can easily be extended to include general dispersion models.

We assume that $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_q$ are i.i.d according to a d-dimensional Gaussian distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{B}_{(j)}$ denote a vector containing all the j^{th} entries of $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_q$ for $j = 1, \ldots, d$. The above distributional assumptions implies that $\tilde{\boldsymbol{B}}^T = \Big[(\boldsymbol{B}_{(1)})^T, \ldots, (\boldsymbol{B}_{(d)})^T\Big]$ is Gaussian distributed with

mean zero and covariance matrix $\boldsymbol{\Sigma} \otimes \boldsymbol{I}_q$, where $\boldsymbol{I}_q$ is the $q \times q$-dimensional identity matrix and $\otimes$ denotes the Kronecker product.

Recall that the i^{th} $(i = 1, \ldots, n_j)$ response in the j^{th} $(j = 1, \ldots, d)$ marginal model was denoted $y_i^{[j]}$. Define for $j = 1, \ldots, d$, the $n_j \times k_j$-dimensional matrix $\boldsymbol{X}^{[j]} = [\boldsymbol{x}_{1j}, \ldots, \boldsymbol{x}_{n_j j}]^T$, and likewise the $n_j \times q$ matrix $\boldsymbol{Z}^{[j]} = [\boldsymbol{z}_{1j}, \ldots, \boldsymbol{z}_{n_j j}]^T$. Based on these definitions, we define for $k = k_1 + \cdots + k_d$ and $n = n_1 + \cdots + n_d$, the $n \times k$-dimensional matrix $\boldsymbol{X} = \text{diag}[\boldsymbol{X}^{[1]}, \ldots, \boldsymbol{X}^{[d]}]$ and the $n \times dq$-dimensional matrix $\boldsymbol{Z} = \text{diag}[\boldsymbol{Z}^{[1]}, \ldots, \boldsymbol{Z}^{[d]}]$. Moreover, we define for each dimension $j = 1, \ldots, d$, the $n_j \times n_j$-dimensional diagonal glm weight matrix $\boldsymbol{W}^{[j]}$ with diagonal entries $w_{ii}^{[j]} = \frac{1}{2\lambda_j} \frac{2}{V_j(\mu_i^{[j]}) g_j'(\mu_i^{[j]})^2}$, and the $n \times n$ matrix $\boldsymbol{W} = \text{diag}[\boldsymbol{W}^{[1]}, \ldots, \boldsymbol{W}^{[d]}]$.

By redefining the matrices $\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{W}$, $\boldsymbol{D} = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_q$ and the vectors $\tilde{\boldsymbol{B}}$ and $\boldsymbol{y}^T = (y_1^{[1]}, \ldots, y_{n_d}^{[d]})$, we can use the Laplace approximation in Breslow & Clayton (1993) to estimate the multivariate model.

Paper II

Multivariate Generalised Linear Mixed Models With Graphical Latent Covariance Structure

Jeanett S. Pelck
Aarhus University

Rodrigo Labouriau
Aarhus University

Abstract. This paper introduces a method for studying the correlation structure of a range of responses modelled by a multivariate generalised linear mixed model (MGLMM). The methodology requires the existence of clusters of observations and that each of the several responses studied is modelled using a generalised linear mixed models (GLMM) containing random components representing the clusters. We construct a MGLMM by assuming that the distribution of each of the random components representing the clusters is the marginal distribution of a (sufficiently regular) multivariate elliptically contoured distribution. We use an undirected graphical model to represent the correlation structure of the random components representing the clusters of observations for each response. This representation allows us to draw conclusions regarding unknown underlying determining factors related to the clusters of observations. Using a combination of an undirected graph and a directed acyclic graph (DAG), we jointly represent the correlation structure of the responses and the related random components. Applying the theory of graphical models allows us to describe and draw conclusions on the correlation and, in some cases, the dependence between responses of different statistical nature (*e.g.,* following different distributions, different linear predictors and link functions). We present some simulation studies illustrating the proposed methodology.

II.1 Introduction

This paper introduces a method for studying the dependence structure of a range of responses modelled by a multivariate generalised linear mixed model (MGLMM) (see Pelck & Labouriau, 2021a, for details). The methodology we suggest requires the existence of clusters of observations (or experimental units) and that each of the responses studied is modelled using a GLMM containing random components representing the clusters. We will construct an MGLMM by assuming that the distribution of each of the random components representing the clusters is the

marginal distribution of a sufficiently regular multivariate elliptically contoured distribution (see Anderson, 2003). This choice of the distribution of the random components includes, as a particular case, the multivariate normal distribution used in standard generalised linear mixed models (GLMMs), as the models considered in Breslow & Clayton (1993), McCulloch & Searle (2001), McCulloch (1997).

We will use an undirected graphical model (see Lauritzen 1996, Whittaker 1990, Abreu et al. 2010) to represent the correlation structure of the random components representing the clusters of observations for each response. This representation will allow us to conclude on unknown underlying determining factors related to the clusters of observations. Furthermore, using a combination of an undirected graph and a directed acyclic graph (DAG), we jointly represent the correlation structure of the responses and the related random components. This representation arises naturally from the construction of the MGLMM we use and yields a known type of graphical model, namely a block chain independence graph (BCG) as defined in Whittaker (1990). Remarkably, this construction will allow us to describe and draw conclusions on the correlation between the responses of different statistical nature, *e.g.,* responses modelled with GLMMs defined using various combinations of distributions, linear predictors and link functions (not necessarily the same for each response).

We will base the inference for the proposed graphical models on variants of tests for correlation under multivariate normality or multivariate elliptically contoured distributions studied in detail in Anderson (2003), from which we draw heavily. In the particular case where the random components are multivariate normally distributed, non-correlation will imply independence, which makes the conclusions of the analysis stronger. When the random components are not normally distributed but follow an elliptically contoured distribution, we will obtain slightly weaker conclusions since, in that case, lack of correlation implies only mean independence. [1]

The paper is organised as follows. In Section II.2.1, we formulate a version of a multivariate generalised linear mixed model. For simplicity, we only specify one common clustering structure but the theory can easily be extended to include multiple clustering. In Section II.3.1, we introduce essential concepts of graphical models that we will use to study the covariance structure of the random components and the response variables. These concepts are connected to the introduced multivariate model in Section II.3.2. In Section II.3.3, we describe statistical tests adapted from Anderson (2003) and draw the connection to the theory of graphical models. In Section II.3.4, we perform a simulation study to study the distribution of the p-values in the simulated examples under the null hypothesis obtained using the statistical tests. Moreover, we study the power of the tests in the simulated examples on a grid consisting of values of the off-diagonal entry in the covariance matrix. Some concluding remarks are given in Section II.4. Appendix II.A.1 discusses how an estimate of the covariance matrix can be obtained based on consistent predictions

[1] Recall that a random vector X is mean independent of the random vector Y when $E(X|Y = y) = E(X)$ for all y in the support of the distribution of Y. It is well known that independence implies mean independence which implies non-correlation, but the reversed implications are in general not valid, see Wooldridge (2010).

of the random components. Appendix II.A.2 presents details on how the density of the introduced test statistic can be evaluated in the case of Gaussian random components.

II.2 Multivariate Generalised Linear Mixed Models

In this section, we formulate a version of the multivariate generalised linear mixed model described in Pelck & Labouriau (2021a). These models are based on marginal GLMMs that extend the standard GLMMs in two directions: we assume the random components to be distributed according to an elliptical contoured distribution, instead of following a multivariate Gaussian distribution, and we assume the conditional distributions of each response, given the random components, to belong to a dispersion model, instead of an exponential dispersion model. The MGLMMs we will define require, however, the existence of clusters of observations and the presence of random components representing those clusters in each of the marginal GLMMs representing the responses. The result of this process is a rather flexible class of models that can be used in many practical applications; see for example Pelck & Labouriau (2020, 2021c), Pelck et al. (2021a,b).

II.2.1 Model Definition

We define a d-dimensional multivariate generalised linear mixed model (*i.e.,* a MGLMM representing d responses) with n observations of the j^{th} marginal model, taking values in $\mathcal{Y}_j \subseteq \mathbb{R}$, for $j = 1, \ldots, d$. Here $\mathcal{Y}_j$ is typically $\mathbb{R}$, $\mathbb{R}_+$, a compact real interval or $\mathbb{Z}_+$. Denote by $Y_i^{[j]}$ the random variable representing the i^{th} observation of the j^{th} response, for $j = 1, \ldots, d$ and $i = 1, \ldots, n$. We assume that there exists a natural clustering of the observations causing dependence between observations arising from the same cluster (*e.g.,* grouping of observations within the same individual). We denote the cluster of the i^{th} observation, $Y_i^{[j]}$, by $c(i)$ taking one of the values $1, \ldots, q$. Moreover, we assume that the clustering of the responses is independent of d, that is, the clusters are represented in each marginal model. To ease the notation throughout the paper, we only consider one clustering mechanism but the methodology can be applied to a model with multiple clustering structures (*e.g.,* Pelck et al. (2021a)). Likewise, we assume the same number of observations for each response to simplify the notation. However, the methods described in this paper applies also to the case with n depending on j.

In each marginal model, we consider random components each taking the same value for all responses within the corresponding cluster. These random components are denoted by $B_{c(i)}^{[j]}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, d$. Define the d-dimensional random vectors of random components, taking values in $\mathbb{R}^q$, by $\boldsymbol{B}^{[j]} = (B_1^{[j]}, \ldots, B_q^{[j]})^T$, the vector of responses $\boldsymbol{Y}^{[j]} = (Y_1^{[j]}, \ldots, Y_n^{[j]})$ and the vector of realisations $\boldsymbol{Y}^{[j]}$ by $\boldsymbol{y}^{[j]}$ (denoted *observations*) of for $j = 1, \ldots, d$. Moreover, consider the d-dimensional

random vector $\boldsymbol{B}_l = (B_l^{[1]}, \ldots, B_l^{[d]})^T$ which we assume to be elliptically contoured distributed (Anderson 2003) satisfying the following regularity conditions, for $l = 1, \ldots, q$,

1. The moments up to fourth order of each marginal distribution exist
2. Each of the marginal distributions is absolute continuous with respect to the Lebesgue measure
3. All the conditional distributions exist and are elliptically contoured distributions also
4. The location parameter vector is equal to zero.

Furthermore, we assume that $\boldsymbol{B}_l$ is independent of $\boldsymbol{B}_{l'}$ for $l, l' = 1, \ldots, q$ such that $l \neq l'$, *i.e.,* we assume that the random vectors representing different clusters are independent. We define the density with respect to the Lebesgue measure of the elliptically contoured distribution by

$$\varphi(\boldsymbol{b}; \boldsymbol{\Lambda}) = |\boldsymbol{\Lambda}|^{-1/2} h\Big(\boldsymbol{b}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{b}\Big), \tag{II.1}$$

where $\boldsymbol{\Lambda}$ is a positive definite scatter matrix. The function $h(\cdot)$ is non-negative and satisfies that

$$\int_{\mathbb{R}^q} h(\boldsymbol{b}^T \boldsymbol{b})\, d\boldsymbol{b} = 1. \tag{II.2}$$

When the density exists, the covariance matrix, $\boldsymbol{\Sigma}$, is proportional to $\boldsymbol{\Lambda}$, *i.e.,* the correlation matrix can be equivalently calculated from both $\boldsymbol{\Sigma}$ and $\boldsymbol{\Lambda}$. An example of a commonly used distribution satisfying these regularity conditions is a multivariate Gaussian distribution with expectation zero and covariance matrix given by $\boldsymbol{\Sigma}$. Another example, that we will study later is the multivariate t-distribution. This distribution allows us to consider different degrees of tail heaviness. Note that because the moments of fourth order must exist in the multivariate t-distribution, the degrees of freedom should be larger than four.

According to the model, we assume that $Y_i^{[j]}$ is conditional distributed according to a dispersion model with dispersion parameter $\lambda_j \in \mathbb{R}_+$ given $B_{c(i)}^{[j]}$, and with conditional expectation

$$g_j\Big(\mu_i^{[j]}(b)\Big) \stackrel{\mathsf{def}}{=} g_j(\mathbb{E}[Y_i^{[j]} | B_{c(i)}^{[j]} = b]) = \boldsymbol{\beta}_j^T \boldsymbol{x}_i^{[j]} + b, \quad \forall\, b \in \mathbb{R},$$

for all $i = 1, \ldots, n$ and $j = 1, \ldots, d$. The vector $\boldsymbol{x}_i^{[j]}$ is a p_j dimensional vector of explanatory variables corresponding to the vector of coefficients, $\boldsymbol{\beta}_j$. The explanatory variables might differ for the different responses. The function $g_j(\cdot)$ is a given link function, which is assumed to be strictly monotone, invertible and continuously differentiable. Below, we will suppress the dependence in $\mu_i^{[j]}(b)$ of b to lighten the notation and denote the parameter space of the conditional means by $\mathcal{U}_j$. We define the conditional density corresponding to the conditional distribution of $Y_i^{[j]}$ given

$B^{[j]}_{c(i)} = b$ with respect to a domination measure ν_j (defined on the measurable space $(\mathcal{Y}_j, \mathcal{A})$) by

$$f(y_i^{[j]}|B^{[j]}_{c(i)} = b; \boldsymbol{\beta}^{[j]}, \lambda_j) \stackrel{\mathsf{def}}{=} p(y_i^{[j]}; \mu_i^{[j]}, \lambda_j) = a_j(y_i^{[j]}; \lambda_j) \exp[-\tfrac{1}{2\lambda_j} d_j(y_i^{[j]}; \mu_i^{[j]})]. \quad \text{(II.3)}$$

The function $d_j : \mathcal{Y}_j \times \mathcal{U}_j \to \mathbb{R}_+$ is the *unit deviance* and, by definition, satisfies that $d_j(\mu, \mu) = 0$ and $d_j(y, \mu) > 0$ for all $(y, \mu) \in \mathcal{Y}_j \times \mathcal{U}_j$ such that $y \neq \mu$. The function $a_j : \mathcal{Y}_j \times \mathbb{R}_+ \to \mathbb{R}_+$ is a given normalising function. We assume that the unit deviance is regular, that is, d is twice continuously differentiable in $\mathcal{Y}_j \times \mathcal{U}_j$ and $\partial^2 d(\mu; \mu)/\partial \mu^2 > 0$ for all $\mu \in \mathcal{U}_j$. The function $V_j : \mathcal{U}_j \to \mathbb{R}_+$ given by $V_j(\mu) = 2/\{\partial^2 d_j(\mu, \mu)/\partial \mu^2\}$ for all μ in $\mathcal{U}_j$ is termed the *variance function* (Cordeiro et al. 2021). The following families of distributions are examples of dispersion models: Normal, Gamma, inverse Gaussian, von Mises, Poisson, and Binomial families. This setup defines a version of the multivariate GLMM described in Pelck & Labouriau (2021a) with the additional assumption that the multivariate distribution of the random components follow an elliptical contoured distribution.

II.3 Representation of the Latent Covariance Structure via Graphical Models

In this section, we describe and illustrate how we can use the theory of graphical models to examine the latent covariance structure of the random components in the multivariate model described above, and how this covariance structure affects the correlation between the responses. First, we give a short account for the theory of graphical models. For a more comprehensively description see Lauritzen (1996) and Whittaker (1990).

II.3.1 Basic Theory of Graphical Models

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a graph defined with a set of vertices, $\mathcal{V}$, composed of random variables and a set of edges, $\mathcal{E} \in \mathcal{V} \times \mathcal{V}$. The set of edges, $\mathcal{E}$, consists of pairs of elements taken from $\mathcal{V}$. We distinguish between undirected independence graphs (UGs) and directed acyclic independence graphs (DAGs) but the two types of graphs can be combined as we will see below. The two types of graphs differ because of the underlying assumption of symmetry in the roles played by the variables in an UG, whereas in a DAG one variable can carry information on another without the converse being necessarily true. In the DAG we use an arrow from one variable pointing to another variable to indicate that the first variable carries information on the second. In an UG, two vertices are connected by an edge if, and only if, they are not conditionally independent given the remaining variables in $\mathcal{V}$. This is the same definition used for DAGs with the conditioning set modified from the remaining variables to a set containing all remaining variables that carry information on one of the two vertices either direct or through the other vertices in $\mathcal{V}$.

In an UG, we say that there is a path connecting two vertices, say v_1 and v_n, if there exists a sequence of vertices $v_1, \ldots, v_n$ such that, for $i = 1, \ldots, n-1$, the pair (v_i, v_{i+1}) is in $\mathcal{E}$. A set of vertices S, separates two disjoint sets of vertices A and B in the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ when every path connecting a vertex in A to a vertex in B necessarily contains a vertex in S. According to the theory of graphical models (see Lauritzen, 1996 and Perl, 2009), the UG defined above satisfies the *separation principle*, which states that if a set of vertices S, separates two disjoint subsets of vertices A and B in the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, then all variables in A are independent of all variables in B given S. Moreover, if the subsets A and B are isolated (*i.e.,* there are no paths connecting a vertex in A to a vertex in B), then the variables in A are independent of the variables in B.

A DAG possesses the Markov properties of its associated moral graph. Here the associated moral graph of a DAG is the UG obtained by the same vertex set but with a modified set of edges. The modified set of edges is formed by all the existing edges in the DAG replaced by undirected edges together with all edges necessary to eliminate forbidden Wermuth configurations. The latter means that for each vertex, we connect all vertices that have a directed edge towards the vertex in question with an undirected edge.

The two types of graphs can be combined into a block chain independence graph (BCG). In this graph, we assume that the vertex set $\mathcal{V}$ can be partitioned into subsets, called blocks, which are connected by directed edges but where all edges within the same block are undirected. As for the DAG, the BCG processes the same independence interpretation as its associated moral graph. For more information see Lauritzen (1996) and Whittaker (1990).

II.3.2 Connecting the Multivariate Model with the Theory of Graphical Models

We connect the model formulated in Section II.2.1 with the theory of graphical models by defining an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V} = \{\boldsymbol{B}^{[1]}, \ldots, \boldsymbol{B}^{[d]}\}$, where $\boldsymbol{B}^{[1]}, \ldots, \boldsymbol{B}^{[d]}$ are the vectors of random components in the multivariate model described in Section II.2.1. In this context, the edges can only be interpret in terms of independence when the random components are Gaussian distributed. In the case of a non-Gaussian elliptically contoured distribution, two vertices are connected by an edge if, and only if, they are conditionally correlated given the remaining variables, which in this context implies conditional mean independence. The set of vertices can also be formulated in terms of each variable in the model instead of vectors as above. In this case, the graphical representation will consist of q separated cliques each containing the respective entry of the vectors $\boldsymbol{B}^{[1]}, \ldots, \boldsymbol{B}^{[d]}$ due to the model assumptions. The choice of representation depends on the analysis and which choice that leads to the best discussion of the results. Note, that the results does not change only the visualisation. We will consider the vector representation below.

The graph defined above is interpret in terms of the random components as follows: if, for example, $\boldsymbol{B}^{[1]}$ and $\boldsymbol{B}^{[2]}$ are connected with an edge, then these two

random variables are conditionally correlated given $\{\boldsymbol{B}^{[3]}, \ldots, \boldsymbol{B}^{[d]}\}$. Therefore, $\boldsymbol{B}^{[1]}$ carries some information on $\boldsymbol{B}^{[2]}$ not contained in the other variables. For example if the random components represent variation between different blocks in a field experiment, this means that there are some latent factors affecting the blocks, could be some characteristics of the soil, which affect the first and second response differently than the other responses.

We introduce an extension of the separation principle below, which we call the *induced separation principle*. This can be used to draw general conclusions on the response variables. According to the model, the responses are independent given the random components. Therefore, conditional uncorrelation between, say, $\boldsymbol{B}^{[1]}$ and $\boldsymbol{B}^{[2]}$ given $\{\boldsymbol{B}^{[3]}, \ldots, \boldsymbol{B}^{[d]}\}$ imply that $\boldsymbol{Y}^{[1]}$ and $\boldsymbol{Y}^{[2]}$ are conditionally uncorrelated given $\{\boldsymbol{B}^{[3]}, \ldots, \boldsymbol{B}^{[d]}\}$. By including the random variables $Y_i^{[j]}$ (for $j = 1, \ldots, d$ and $i = 1, \ldots n$) in the set of vertices, and by taking the model assumptions into considerations, it is possible to formulate a block chain independence graph (BCG) that represents the covariance structure both among the random components but also within the response variables. The theory of BCG makes it possible to extend the separation principle to a version that applies to the total graph including both the random components and the response variables. That is, by looking at the moral graph, we can determine all conditional uncorrelations (Whittaker 1990, Theorem. 3.6.1).

We will describe how the BCG can be constructed in the multivariate model described in Section II.2.1. For simplicity we only consider one common clustering mechanism in this model, however, below we will argue how the BCG can be constructed in the case of multiple clustering mechanisms. We define a block chain independence graph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ (Whittaker 1990) by letting $\mathcal{V}' = \mathcal{V} \cup \{\boldsymbol{Y}^{[1]}, \ldots, \boldsymbol{Y}^{[d]}\}$ and $\mathcal{E}' = \mathcal{E} \cup \mathcal{E}_Y$, where $\mathcal{V}$ and $\mathcal{E}$ are defined as above. Here, $\mathcal{E}_Y$ includes directed edges from $\boldsymbol{B}^{[j]}$ to $\boldsymbol{Y}^{[j]}$ for $j = 1, \ldots, d$, whereas $\mathcal{E}$ only include undirected edges. Usually, the way to separate undirected and directed edges in $\mathcal{E}'$ is to use the notation that if there is a directed edge from V_i to V_j, the edge (i, j) is included in $\mathcal{E}'$. However, if there is an undirected edge from V_i to V_j both the edge (i, j) and (j, i) are included in $\mathcal{E}'$. The essential property of this graph is that by construction, any edge is undirected for intra-block vertices, and directed for inter-block vertices with direction from the random components to the response variables (the blocks are here defined by $\mathcal{V}$ and $\mathcal{V}_{\boldsymbol{Y}} = \{\boldsymbol{Y}^{[1]}, \ldots, \boldsymbol{Y}^{[d]}\}$). The induced separation principle implies that if $\mathcal{S}$ separates two disjoint subsets of vertices, $\mathcal{A}$ and $\mathcal{B}$ in $\mathcal{V}$, and $\mathcal{A}'$ and $\mathcal{B}'$ are the sets of the corresponding response variables, respectively, then all response variables in $\mathcal{A}'$ are conditionally uncorrelated of the variables in $\mathcal{B}'$ given the random components in $\mathcal{S}$.

In the case of multiple clustering mechanisms, we redefine $\mathcal{V}'$ to be the union of all sets of random components and the responses, that is, $\mathcal{V}' = \mathcal{V}_1 \cup \ldots \cup \mathcal{V}_b \cup \mathcal{V}_{\boldsymbol{Y}}$, where b is the total number of clustering mechanisms, and $\mathcal{V}_i$ is the set containing random vectors corresponding to the random components associated with the i^{th} clustering. The edges in this graph consist of the undirected edges inside each block together with directed edges from each random vector pointing towards the corresponding response variable (between the blocks). Under the model, we assume that each block

of random components is independent of the others. Therefore, we do not need to connect the blocks $\mathcal{V}_1, \ldots, \mathcal{V}_b$ with an edge. This structure is illustrated in Figure II.1.

The moral version of such a graph can be difficult to interpret in terms of conditional uncorrelation between the response variables. In that case, we suggest to either only consider the undirected graphs for the random components excluding the response vectors, or if one of the clustering mechanisms are of particular interest, we can restrict ourselves to only examining the graph that includes the random components and the response variables of interest. In the latter case, we are only able to interpret the graph on individual level. For example, in a study with two clustering mechanisms: one representing individual variation and another clustering the individuals in different groups, we might only be interested in examining the correlation between different responses caused by the individual clustering structure. Therefore, we can consider a graphical representation of the covariance structure of the individual variation for each individual and thus, avoid comparing individuals within the same group for which the corresponding responses will be correlated do to the random component grouping the individuals. Thus, in the complete block chain independence graph, many of the responses will only be conditional uncorrelated after conditioning on multiple clustering.

An example of a block chain independence graph representing a three dimensional model with two clustering and it's corresponding moral graph is presented in Figure II.2. In this example we observe from the moral graph that $\boldsymbol{Y}^{[1]}$ and $\boldsymbol{Y}^{[3]}$ are conditionally uncorrelated given $\boldsymbol{B}_1^{[2]}$ and $\boldsymbol{B}_2^{[2]}$. If there was an edge connecting $\boldsymbol{B}_1^{[2]}$ and $\boldsymbol{B}_3^{[2]}$, then $\boldsymbol{Y}^{[1]}$ and $\boldsymbol{Y}^{[3]}$ would only be conditionally independent given all the random components.

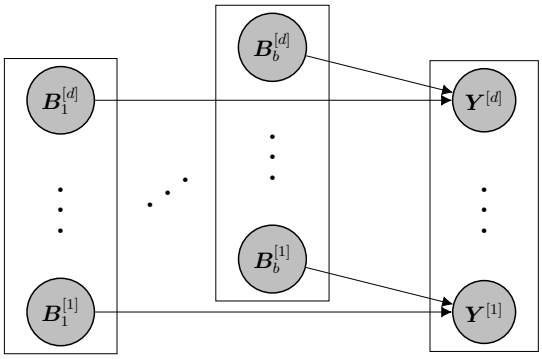


Figure II.1: Illustration of the structure of a block chain independence graph, ignoring the undirected edges inside each of the $(b+1)$ blocks.

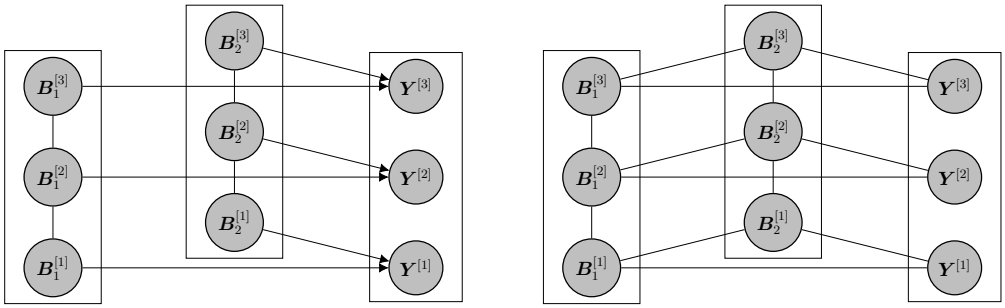


Figure II.2: Example of a BCG for a three dimensional multivariate generalised linear mixed model with two random components, and it's corresponding moral graph (to the right).

II.3.3 Testing the Covariance Structure

In this section, we formulate a statistical test based on the results in Anderson (2003). Using this test, it is possible to test for (conditional) uncorrelation between pairs or groups of variables.

We introduce some general notation that we will use to describe the statistical test in the case where the random components are assumed to be Gaussian distributed and the more general setup where we assume an elliptical contoured distribution. In both cases, we can test for uncorrelation between groups of variables either directly or conditional on a separating set. We show the conditional test but the approach is equivalent in the direct case.

Let $\boldsymbol{X} = (\boldsymbol{X}_{(1)}^T, \ldots, \boldsymbol{X}_{(k)}^T)^T$ be a d-dimensional random vector distributed according to an elliptically contoured distribution (including the special case of a Gaussian distribution) with location parameter equal to zero and a positive definite scatter matrix

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} & \cdots & \boldsymbol{\Lambda}_{1k} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} & \cdots & \boldsymbol{\Lambda}_{2k} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{\Lambda}_{k1} & \boldsymbol{\Lambda}_{k2} & \cdots & \boldsymbol{\Lambda}_{kk} \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{\Lambda}}^{(k-1)} & \tilde{\boldsymbol{\Lambda}}^{(k-1,k)} \\ \tilde{\boldsymbol{\Lambda}}^{(k,k-1)} & \boldsymbol{\Lambda}_{kk} \end{bmatrix},$$

which is proportional to the covariance matrix $\boldsymbol{\Sigma}$. Below, we let $d_1, \ldots, d_k$ denotes the length of the k sub-vectors of $\boldsymbol{X}$ such that $d = d_1 + \ldots + d_k$. We assume that the density of $\boldsymbol{X}$ exists with respect to the Lebesgue measure. Moreover, we assume that the conditional distribution of $\boldsymbol{X}^{(k-1)} = (\boldsymbol{X}_{(1)}^T, \ldots, \boldsymbol{X}_{(k-1)}^T)$ given $\boldsymbol{X}_{(k)}$ exists. The distribution of $\boldsymbol{X}^{(k-1)}|\boldsymbol{X}_{(k)}$ is also elliptically contoured distributed with scatter matrix

$$\boldsymbol{\Lambda}_{\cdot k} = \tilde{\boldsymbol{\Lambda}}^{(k-1)} - \tilde{\boldsymbol{\Lambda}}^{(k-1,k)} \boldsymbol{\Lambda}_{kk}^{-1} \tilde{\boldsymbol{\Lambda}}^{(k,k-1)}, \tag{II.4}$$

which is proportional to the covariance matrix in the conditional distribution (Anderson 2003). Consequently, the formulas which apply in the normal case apply in this more general setting as well.

We would like to test the null hypothesis that the subvectors $\boldsymbol{X}_{(1)}, \ldots, \boldsymbol{X}_{(k-1)}$ are independent given $\boldsymbol{X}_{(k)}$. This is equivalent to examining if $\boldsymbol{\Lambda}_{\cdot k}$ is on the form

$$\boldsymbol{\Lambda}_{\cdot k} = \begin{bmatrix} \boldsymbol{\Lambda}_{11 \cdot k} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_{22 \cdot k} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Lambda}_{(k-1)(k-1) \cdot k} \end{bmatrix}.$$

We first treat the special case where $\boldsymbol{X}$ is Gaussian distributed below. In this case, the statistical test will be exact. Second, we present an asymptotic test when the number of realisations of $\boldsymbol{X}$, denoted q, goes to infinity which is valid in the case of a general elliptically contoured distribution.

Normally Distributed Random Component

Here, we show a test for conditional independence for subsets of variables in a Gaussian distributed vector. Let $\boldsymbol{A}_{\cdot k}$ be the maximum likelihood estimate of $\boldsymbol{\Lambda}_{\cdot k}$ or another estimate proportional to the maximum likelihood estimate based on q observations ($\boldsymbol{A}_{\cdot k}$ can also be calculated from a maximum likelihood estimate of $\boldsymbol{\Sigma}$ using the formula in (II.4)).

The test statistic we will consider is given by

$$V = \frac{\det(\boldsymbol{A}_{\cdot k})}{\prod_{i=1}^{k-1} \det(\boldsymbol{A}_{ii \cdot k})},$$

that is, $V = \lambda^{q/2}$ where λ is the likelihood ratio statistic and q the number of observations (in the setup of multivariate GLMMs this is the number of groups of the random component).

It can be shown that under the null hypothesis (Anderson 2003) the distribution of V is given by

$$V \sim \prod_{i=2}^{k-1} \prod_{j=1}^{d_i} Z_{ij}, \tag{II.5}$$

where the random variables $Z_{21}, \ldots, Z_{(k-1)d_{k-1}}$ are independent and $Z_{ij} \sim \text{Beta}\Big(\frac{1}{2}[q - \bar{d}_i - j], \frac{1}{2}\bar{d}_i\Big)$ with $\bar{d}_i = d_1 + \ldots + d_{i-1}$ for $i = 2, \ldots, (k-1)$ and $j = 1, \ldots, d_i$.

The continuity of the determinant function implies that V remains constant for any estimated scatter matrix proportional to the maximum likelihood estimate, and thus the distribution is still exact. For a consistent estimator of the covariance matrix or scatter matrix, the distribution is only asymptotic.

Elliptical Contoured Distributed Random Component

Under the assumption of a general elliptical contoured distribution, Anderson (2003) shows that the following test statistic can be used to test asymptotically if the

correlation between groups of variables are zero for q going to infinity (either direct or conditioning on a separating set).

Let $\boldsymbol{A}_{\cdot k}$ denote the sample estimate of the covariance matrix of $\boldsymbol{X}^{(k-1)}$ given $\boldsymbol{X}_k$. This can be estimated directly or calculated using the formula in (II.4) on the sample covariance matrix given by

$$\boldsymbol{A} = \frac{1}{q-1}\sum_{i=1}^{q}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T.$$

Define $\tilde{\boldsymbol{A}}_{\cdot k}^{(i-1)}$ as the $\bar{d}_i \times \bar{d}_i$ dimensional sub-matrix of $\boldsymbol{A}_{\cdot k}$ with the first $\bar{d}_i = d_1 + \cdots + d_{i-1}$ rows and columns of $\boldsymbol{A}_{\cdot k}$. Moreover, let $\tilde{\boldsymbol{A}}_{\cdot k}^{(i,i-1)}$ denote the sub-matrix corresponding to the first $\bar{d}_i$ columns and the rows $\bar{d}_i + 1, \ldots, \bar{d}_i + d_i$ of $\boldsymbol{A}_{\cdot k}$. Define

$$\begin{aligned}
\boldsymbol{H}_i &= (q-1)\tilde{\boldsymbol{A}}_{\cdot k}^{(i,i-1)}\tilde{\boldsymbol{A}}_{\cdot k}^{(i-1)}(\tilde{\boldsymbol{A}}_{\cdot k}^{(i,i-1)})^T \\
\boldsymbol{G}_i &= (q-1)\boldsymbol{A}_{ii\cdot k} - \boldsymbol{H}_i \\
V_i &= \frac{|\boldsymbol{G}_i|}{|\boldsymbol{G}_i + \boldsymbol{H}_i|},
\end{aligned}$$

where $\boldsymbol{A}_{ii\cdot k}$ is the $(i,i)^{th}$ block matrix of $\boldsymbol{A}_{\cdot k}$.

The test statistic for the null hypothesis that $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1})$ are conditionally independent given $\boldsymbol{X}_k$ is formulated as

$$-q\sum_{i=2}^{k-1}\log V_i,$$

which converges in distribution to $(1+\kappa)\chi_f^2$ for q going to infinity, $f = \sum_{i=2}^{k-1}\bar{d}_i d_i$ and

$$\kappa = (\bar{d}_k(\bar{d}_k+2))^{-1}\mathbb{E}[(\boldsymbol{X}^{(k-1)})^T\boldsymbol{\Sigma}_{\cdot k}^{-1}\boldsymbol{X}^{(k-1)}]^2 - 1.$$

We can estimate the kurtosis parameter by

$$\hat{\kappa} = (\bar{d}_k(\bar{d}_k+2))^{-1}\frac{1}{q}\sum_{i=1}^{q}[(\boldsymbol{x}_i^{(k-1)})^T\boldsymbol{A}_{\cdot k}^{-1}\boldsymbol{x}_i^{(k-1)}]^2 - 1.$$

Simulation Study of Convergence Rate

In this section we study the power of the introduced tests as a function of q in a simulated example, that is, the probability of accepting the hypothesis of independence/uncorrelation when it is true. Working with a five percent significance level this should be close to 95 percent when q is large enough. For different values of q, we simulate 10 000 times q random variables from a 4 dimensional Gaussian and t-distribution, respectively, with expectation zero and covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix}
0.4083 & 0.000000 & 0.000000 & 0.000000 \\
0.0000 & 0.456510 & -0.451965 & 0.265170 \\
0.0000 & -0.451965 & 0.837030 & -0.491090 \\
0.0000 & 0.265170 & -0.491090 & 0.524365
\end{pmatrix}.$$

For each of the 10 000 simulations, we estimate the sample covariance matrix and apply the appropriate test described above. Based on the calculated p-values, we can examine how many times the estimated p-value is below five percent and divide this by the number of simulations in order to obtain an estimate of the probability of rejecting the hypothesis of independence/un-correlation when it is true (at a five percent significance level). If the distributional assumption of the test statistic is correct, the estimated value will be close to five percent when working with a five percent significance level.

We formulate the simulated model formally by letting $\boldsymbol{X} = (X_1, \ldots, X_4)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_4)$ denote random vectors distributed according to a multivariate Gaussian and t-distribution, respectively, with mean zero and covariance matrix $\boldsymbol{\Sigma}$. In the multivariate t-distribution, the degrees of freedom is assumed to be five. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_q$ denote i.i.d. copies of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, corresponding to the simulated random variables in each of the 10 000 rounds. Using the realizations of these variables denoted $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q$ and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q$, we estimate in each of the 10 000 rounds $\boldsymbol{\Sigma}$ by

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \tfrac{1}{q} \sum_{i=1}^{q} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$$
$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}} = \tfrac{1}{q} \sum_{i=1}^{q} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T,$$

where $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ are the estimated means. In each of the the 10 000 rounds, we test the hypothesis that

$$H : \boldsymbol{\Sigma}_{13} = \boldsymbol{\Sigma}_{31} = 0,$$

for both the multivariate Gaussian and t-distribution based on the estimated covariance matrices $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}}$, respectively, with $\boldsymbol{\Sigma}_{ij}$ denoting the $(i,j)^{\text{th}}$ entry in $\boldsymbol{\Sigma}$. Note, that vi also tested $\boldsymbol{\Sigma}_{12} = 0$ and $\boldsymbol{\Sigma}_{14} = 0$ but since the estimated power curves are similar, we only present one of them here. The estimated power curve (probability of rejecting the hypothesis when it is true) for H can be found in Figure II.3 and II.4 for the Gaussian and t-distribution, respectively. We conclude, that we need a much higher number of levels in the multivariate t-distribution, as expected. This is a result of the heavier tail and the fact that the distribution of the test statistic is only asymptotic where the distribution is exact in the Gaussian case.

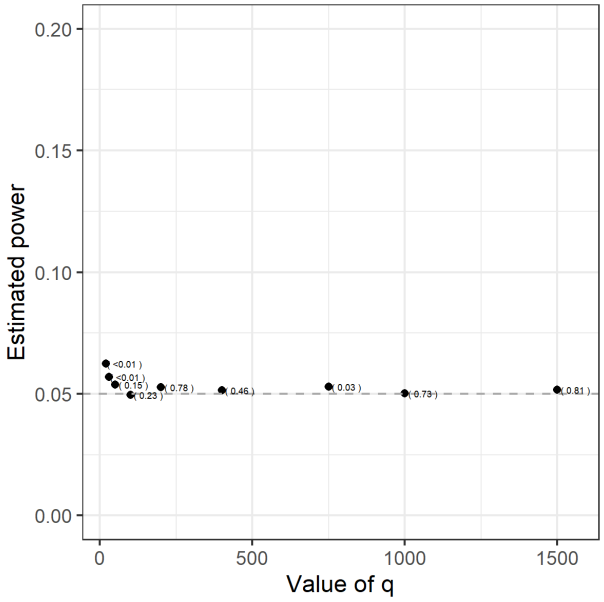


Figure II.3: Estimated power as a function of q based on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q$. The values in parenthesis is p-values from a Kolmogorov Smirnoff test for a Uniform distribution. When the hypothesis is true and the distributional assumption is correct, the p-values should be uniformly distributed on the interval from zero to one.

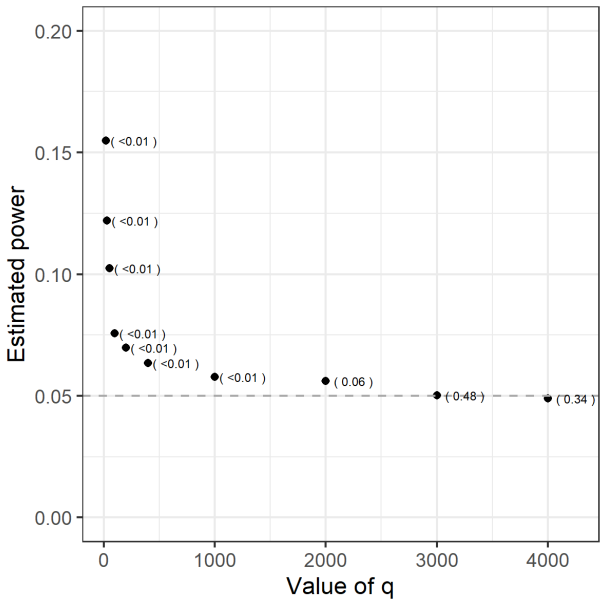


Figure II.4: Estimated power as a function of q based on $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q$. The values in parenthesis is p-values from a Kolmogorov Smirnov test for a Uniform distribution. When the hypothesis is true and the distributional assumption is correct, the p-values should be uniformly distributed on the interval from zero to one.

Graphical Representation of the Latent Covariance Structure in the Multivariate Model using a Statistical Test

We can draw conclusions regarding the covariance structure of the random components by applying the tests described above to the estimated covariance matrix of the random components in the multivariate model described in Section II.2.1, and by using the theory of graphical models as described in Section II.3.1.

Under the model formulated in Section II.2.1, we can estimate the covariance matrix of the random components consistently based on consistent predictions of the random components by applying Proposition 1. In this case, the distribution of the above-described tests will be asymptotically for the number of levels, q, and the number of observations, n, increasing. In the case where we use the asymptotically approximately maximum likelihood estimator of the covariance matrix as described in Section II.A.1, the estimator is also consistent. Therefore, the tests still apply asymptotically. The same applies to another consistent estimate of the covariance matrix.

We can examine the latent covariance structure in general by testing if the value of each off-diagonal entry in the conditional covariance matrix is equal to zero. If the p-value (possibly corrected for multiple testing) is below a given significance level, we connect the corresponding nodes by and edge. After constructing an undirected graph, we can combine the undirected independence graph with the responses as described in Section II.3.1. On the other hand, it might be of interest to test for a specific covariance structure of the latent variables. Here, the number of tests can be reduced using the structure of graphical models. It is possible to apply the test for independence between different groups of variables, without conditioning on a separating set, to test for independence between the isolated subgraphs in the graph (if any).

II.3.4 Simulation Studies

In this section, we perform a simulation study to examine the power of the two types of tests introduced in this paper under multivariate generalised linear mixed models in the case of Gaussian and t-distributed random components. We simulate a two dimensional generalized linear mixed model with the conditional distributions being Gamma and Poisson, respectively. We use a logarithm link function in both marginal models. Since our primary interest in these simulations study is the covariance structure of the random components, we will simulate a model only including a constant in the fixed effects (the value of this constant was set to 0.6). The data was simulated with the length of the vectors of random components being $q = 800$ (corresponding to 800 experimental units or clusters), and with 40 replicates for each unit giving 32 000 observations. Three models were simulated with different distributional assumptions for the random components (all having expectation zero), *i.e.,* a multivariate Gaussian, a multivariate t-distribution with 11 degrees of freedom, and a multivariate t-distribution with 7 degrees of freedom. We estimated the power (probability of rejecting the null hypothesis) of the tests for the different models on

a grid of values for the off-diagonal entry in the covariance matrix of the random components representing the same experimental unit. The covariance matrix is given by

$$\mathbf{\Sigma} = \begin{Bmatrix} 0.8166 & \sigma_{12} \\ \sigma_{12} & 0.91302 \end{Bmatrix},$$

where σ_{12} is varied on the grid $G = (0, 0.02, 0.04, 0.1)$.

In each round of the simulation study, we test the hypothesis

$$H_0 : \sigma_{12} = 0,$$

and the resulting p-values are used to estimate the power for each point in G. Notice, that in the Gaussian case, H_0 implies independence, whereas, in the elliptical case it implies un-correlation. We limit ourselves to a two dimensional model partly because of the computational time and the preference of a high dimension of the vector of random components (as we saw in II.3.3, we need a high number of levels of the random components when these are assumed to be multivariate t-distributed), but also because it is difficult to control that a high dimensional covariance matrix stay positive definite when changing the off-diagonal values.

We would expect that the probability of rejecting the null hypothesis increases when the corresponding entry in the covariance matrix is moved away from zero. For each grid point in G, the model was simulated 500 times and a p-value for testing H_0 was calculated for each simulation. Thus, for each grid point, the probability of rejecting the null hypothesis could be estimated based on the p-values. Figure II.5 shows the estimated probabilities of rejecting the hypothesis (at a significance level of five percent) as a function of the off-diagonal value in the covariance matrix for each combination of model and test. From the figure, we conclude that when the random components are Gaussian distributed both tests reach the correct significance levels under the null hypothesis. However, the curve for the Gaussian test is steeper than the elliptical test in the part close to zero meaning that the test has a higher power to detect small deviations from the null hypothesis. In the case of t-distributed random components, the test based on normality rejects too often under the null hypothesis which lead to a power curve with a higher intersection with the y-axis. Moreover, we see that the shape of the curve differs from the power curve for the elliptical test. This result imply that it would be preferable to use the elliptical test in cases where the normality of the random components are uncertain.

We would expect, that the p-values follow a uniform distribution on zero to one under the hypothesis H_0. In Figure II.6, we present a Q-Q plot of the observed quantiles of the calculated p-values versus the theoretical uniform quantiles based on 500 simulations for each model and for each test. Recall, that we simulated three different models, where the random components followed either a multivariate Gaussian, a multivariate t-distribution with 11 degrees of freedom or a multivariate t-distribution with 7 degrees of freedom. For each model, we compared two different tests: a test based on normality and a test based on a general elliptically contoured distribution. The number added to each plot is the resulting p-values from a Kolmorogov-Smirnov

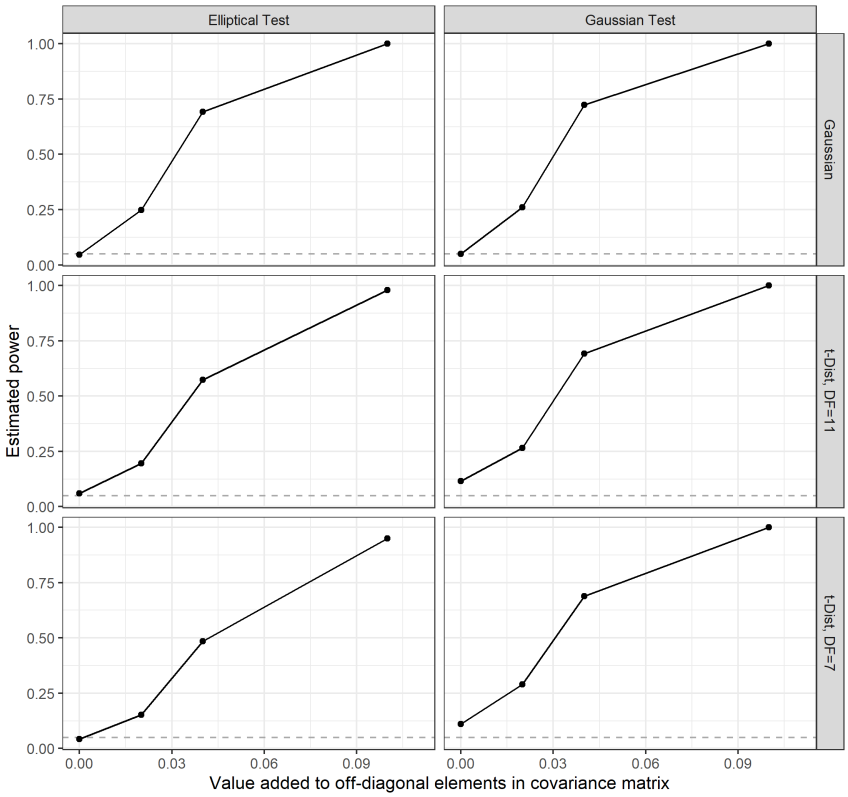


Figure II.5: Power curves showing the estimated probability of rejecting the hypotheses H_0 as a function of the off diagonal values in the covariance matrix for each test and distributional model for the random components.

test comparing the empirical distribution with the uniform distribution. As expected, the test based on an assumption of normality performs badly for the models where the random components follow a t-distribution.

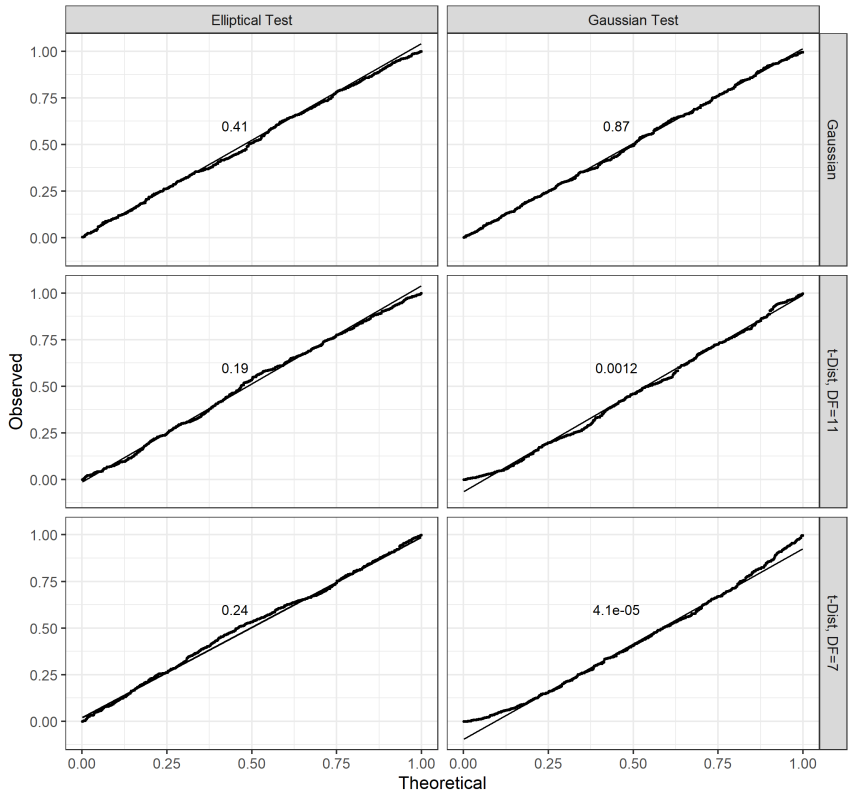


Figure II.6: QQ-plots showing the quantiles of the p-values under the null hypothesis against the theoretical quantiles of the uniform distribution for each test and distributional model for the random components. We used the Kolmogorov-Smirnov test comparing the empirical distribution with the uniform distribution. The resulting p-values are shown in the plots.

II.4 Discussion and Conclusion

The method for studying the dependence structure of multivariate responses described in this paper combines MGLMMs with the theory of graphical models and a variation of the tests for correlation and conditional correlation described in Anderson (2003). We constructed the MGLMMs used in this paper by joining marginal GLMMs that are based on weaker assumptions as compared to the literature (e.g., Breslow & Clayton, 1993, McCulloch & Searle, 2001, McCulloch, 1997). Indeed, we do not assume the random components to be multivariate normally distributed. Moreover, we use dispersion models (which includes exponential dispersion models as a particular case) to define the conditional distributions of the responses given the random components. While Pelck & Labouriau (2021a) developed techniques for estimating fixed effects and predicting random components of those MGLMMs, we concentrated here on the construction of methods for studying the correlation structure of multivariate responses. The nature of the tests we used here forced us to restrict the distribution of the random components to be regular elliptically contoured distributions (including the multivariate normal distribution), which is less general

than the class of distributions of the random components used in Pelck & Labouriau (2021a). Still, the assumptions on the distribution of the random components used here are weak and yield a flexible class of MGLMMs. For example, we can use models with multivariate t-distributed random components, which have heavier tails than Gaussian random components.

Remarkably, the proposed test for elliptically contoured distributed random components does not depend on the choice of the elliptically contoured distribution used. Indeed, the test statistic of those tests depends only on the estimate of the covariance matrix. Therefore, we might view this test as a semiparametric test since the class of regular elliptically contoured distributions is not finite-dimensional. Naturally, the test based on the multivariate normal distribution is advantageous relative to the generic test based on elliptically contoured distributions when the random components are Gaussian distributed. We illustrate this claim in the simulation studies presented.

Multivariate generalised linear models (and MGLMMs) can be constructed by connecting several marginal generalised linear models (or GLMMs) using copulas. For instance Song et al. (2009) use Gaussian copulas for constructing multivariate dispersion models. While this approach might be fruitful in some contexts, it cannot be directly applied in the type of analysis we discuss in this paper because the distribution of the random components after applying the copula transformation are in general not elliptically contoured and therefore the tests we use here are not applicable.

The inferential techniques described in this paper were applied in several fields recently. For instance, in Pelck & Labouriau (2020) the method described above was used in a study of a system for monitoring the development of roots over time, which involved binomial, and Poisson distributed responses. Another example is presented in Pelck et al. (2021b) where our methods were applied to study the dependence structure of responses representing the development of a fungal disease and the concentration of volatile organic compounds. Those responses were modelled by Pelck et al. (2021b) using Gamma, binomial and compound Poisson families of distributions. Furthermore, in a third study, Pelck et al. (2021a) used the methods studied here to discuss the covariance structure of the students' marks obtained in different admission exams at the University (Gaussian distributed) and the number of attempts required to pass the course of geometry (a Cox proportional model with discrete-time). Those examples illustrate the usefulness of the statistical tools studied in this paper.

Acknowledgement

The authors were partially financed by the Applied Statistics Laboratory (aStatLab) at the Department of Mathematics, Aarhus University.

References

Abreu, G. C., Labouriau, R. & Edwards, D. (2010), 'High-dimensional graphical model search with gRapHD R package', *Journal of Statistical Software* **37**(1).

Anderson, T. W. (2003), *An introduction to multivariate statistical analysis*, Vol. 2, Wiley New York.

Breslow, N. E. & Clayton, D. G. (1993), 'Approximate inference in generalized linear mixed models', *Journal of the American statistical Association* **88**(421), 9–25.

Cordeiro, G. M., Labouriau, R. & Botter, D. (2021), 'An introduction to bent jørgensen's ideas', *Brazilian journal of Probability and Statistics* **35**(1), 2–20.

Lauritzen, S. L. (1996), *Graphical models*, Vol. 17, Clarendon Press.

McCulloch, C. E. (1997), 'Maximum likelihood algorithms for generalized linear mixed models', *Journal of the American statistical Association* **92**(437), 162–170.

McCulloch, C. & Searle, S. (2001), *Generalized, Linear, and Mixed Models*, John Wiley & Sons.

Pelck, J. S. & Labouriau, R. (2020), 'Using multivariate generalised linear mixed models for studying roots development: An example based on minirhizotron observations'. arXiv:2011.00546.

Pelck, J. S. & Labouriau, R. (2021a), Conditional inference for multivariate generalised linear mixed models. arXiv:2107.11765.

Pelck, J. S. & Labouriau, R. (2021c), Simultaneously analysis of time to emergence of different weed species. In preparation.

Pelck, J. S., Luca, A., Holthusen, H., Edelenbos, M. & Labouriau, R. (2021b), 'Multivariate method for detection of rubbery rot in storage apples by monitoring volatile organic compounds: An example of multivariate generalised linear mixed models'. arXiv:2107.11233.

Pelck, J. S., Maia, R. P., Pinheiro, H. P. & Labouriau, R. (2021a), 'A multivariate methodology for analysing students' performance using register data'. arXiv:2102.10565.

Perl, J. (2009), *Causality: models, reasoning and inference*, second edition edn, Cambridge University Press.

Song, P. X.-K., Li, M. & Yuan, Y. (2009), 'Joint regression analysis of correlated data using gaussian copulas', *Biometrics* **65**(1), 60–68.

Tang, J. & Gupta, A. (1984), 'On the distribution of the product of independent beta random variables', *Statistics & Probability Letters* **2**(3), 165–168.

Whittaker, J. (1990), *Graphical models in applied multivariate analysis*, Chichester New York et al: John Wiley & Sons.

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, 2nd edn, London: The MIT Press.

II.A Appendix

II.A.1 Estimation of Covariance Matrix

The methods presented in this paper rely on either an estimate of the covariance matrix proportional to the maximum likelihood estimate or a consistent estimate. In this section, we discuss how such an estimate can be obtained based on consistent predictions of the random components. Such predictions can be obtained using the inference method described in Pelck & Labouriau (2021a).

A consistent estimate (for n and q increasing) of the covariance matrix can be found by calculating the sample covariance of the predicted values as we will see in Proposition 1. In the case where we only have few cluster, *i.e.,* q is small, we suggest a method to obtain an approximated maximum likelihood estimate of the covariance matrix. Here, we consider the general case were the random components follow an elliptically contoured distribution, and the special case where this distribution is assumed to be multivariate Gaussian separately.

Proposition 1. *Consider the model described in Section II.2.1. For $j = 1, \ldots, q$, let $\hat{\boldsymbol{b}}_j^n$ denote a d-dimensional vector of predicted values of the random components corresponding to the q^{th} cluster, $\boldsymbol{B_j}$, based on at least $n = \min\{n_1, \ldots, n_q\}$ observations. Moreover, assume that*

$$\hat{\boldsymbol{b}}_j^n \xrightarrow{P} \boldsymbol{B}_j \quad \textit{for } n \to \infty.$$

Then,

$$\hat{\boldsymbol{\Sigma}}_q = \frac{1}{q-1}\sum_{j=1}^{q}(\hat{\boldsymbol{b}}_j^n - \bar{\hat{\boldsymbol{b}}}_q)(\hat{\boldsymbol{b}}_j^n - \bar{\hat{\boldsymbol{b}}}_q)^T \xrightarrow{P} \boldsymbol{\Sigma} \quad \textit{for } q, n \to \infty,$$

where $\bar{\hat{\boldsymbol{b}}}_q^n = \frac{1}{q}\sum_{j=1}^{q}\hat{\boldsymbol{b}}_j^n$.

Proof. The proof follows from the fact that the predicted values of the random components are consistent, the continuity of the sample covariance mapping and that the average converges to the expectation for q increasing. □

We present below an approximation of the maximum likelihood function for estimating $\boldsymbol{\Sigma}$ based on the predicted values of the random components in the case of a multivariate Gaussian distribution, which can be used to estimate $\boldsymbol{\Sigma}$.

Approximated maximum likelihood for estimating $\boldsymbol{\Sigma}$ in the case of Gaussian random components

Consider the model described in Section II.2.1, where we assume that $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_q$ are i.i.d Gaussian distributed with expectation zero and covariance matrix $\boldsymbol{\Sigma}$. We let for $j = 1, \ldots, q$, $\hat{\boldsymbol{b}}_j^n$ denote a d-dimensional vector of predicted values of the random components corresponding to the j^{th} cluster, $\boldsymbol{B_j}$, based on at least $n = \min\{n_1, \ldots, n_q\}$ observations. Moreover, we assume that the predicted values are conditional asymptotically Gaussian distributed (as in Pelck & Labouriau (2021a)) for n increasing given $\boldsymbol{B_j} = \boldsymbol{b}_j$ with conditional expectation $\boldsymbol{b}_j$ and covariance matrix $\boldsymbol{V}_j$. Notice, that by the model assumptions, $\boldsymbol{V}_j$ is a diagonal matrix.

When q is small, we can maximise the following with respect to $\boldsymbol{\Sigma}$ by inserting an estimate $\hat{\boldsymbol{V}}_j$ of $\boldsymbol{V}_j$:

$$
\begin{aligned}
L(\boldsymbol{\Sigma}; \hat{\boldsymbol{b}}_1^n, \ldots, \hat{\boldsymbol{b}}_q^n) &= \prod_{j=1}^{q} \int_{\mathbb{R}^d} \varphi(\boldsymbol{b}_j; \boldsymbol{\Sigma}) h(\hat{\boldsymbol{b}}_j^n; \boldsymbol{b}_j, \hat{\boldsymbol{V}}_j) d\boldsymbol{b}_j \\
&= \prod_{j=1}^{q} \int_{\mathbb{R}^d} |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\tfrac{1}{2}\boldsymbol{b}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{b}_j\right) \\
&\qquad |2\pi\hat{\boldsymbol{V}}_j|^{-1/2} \exp\left(-\tfrac{1}{2}(\boldsymbol{b}_j - \hat{\boldsymbol{b}}_j^n)^T \hat{\boldsymbol{V}}_j^{-1} (\boldsymbol{b}_j - \hat{\boldsymbol{b}}_j^n)\right) d\boldsymbol{b}_j \\
&= \prod_{j=1}^{q} (2\pi)^{-d} |\boldsymbol{\Sigma}\hat{\boldsymbol{V}}_j|^{-1/2} \exp\left(-\tfrac{1}{2}(\hat{\boldsymbol{b}}_j^n)^T \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{b}}_j^n\right) \\
&\quad \int_{\mathbb{R}^d} \exp\left(\boldsymbol{b}_j^T \hat{\boldsymbol{V}}_j^{-1} \hat{\boldsymbol{b}}_j^n\right) \exp\left(-\tfrac{1}{2}\boldsymbol{b}_j^T (\hat{\boldsymbol{V}}_j^{-1} + \boldsymbol{\Sigma}^{-1}) \boldsymbol{b}_j\right) d\boldsymbol{b}_j \\
&= \prod_{j=1}^{q} (2\pi)^{-d} |\boldsymbol{\Sigma}\hat{\boldsymbol{V}}_j|^{-1/2} \exp\left(-\tfrac{1}{2}(\hat{\boldsymbol{b}}_j^n)^T \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{b}}_j^n\right) \\
&\quad |2\pi(\hat{\boldsymbol{V}}_j^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}|^{1/2} \exp\left(\tfrac{1}{2}(\hat{\boldsymbol{V}}_j^{-1} \hat{\boldsymbol{b}}_j^n)^T (\hat{\boldsymbol{V}}_j^{-1} + \boldsymbol{\Sigma}^{-1})^{-1} \hat{\boldsymbol{V}}_j^{-1} \hat{\boldsymbol{b}}_j^n\right) \\
&= \prod_{j=1}^{q} (2\pi)^{-d/2} |\boldsymbol{\Sigma}\hat{\boldsymbol{V}}_j|^{-1/2} |(\hat{\boldsymbol{V}}_j^{-1} + \boldsymbol{\Sigma}^{-1})|^{-1/2} \\
&\quad \exp\left(-\tfrac{1}{2}(\hat{\boldsymbol{b}}_j^n)^T [\boldsymbol{\Sigma}^{-1} - (\hat{\boldsymbol{V}}_j + \hat{\boldsymbol{V}}_j \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{V}}_j)^{-1}] \hat{\boldsymbol{b}}_j^n\right).
\end{aligned}
$$

Approximated maximum likelihood for estimating $\boldsymbol{\Sigma}$ in the case of general elliptical contoured random components

Consider the model described in Section II.2.1, where we assume that $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_q$ are i.i.d elliptically contoured distributed with expectation zero and covariance matrix $\boldsymbol{\Sigma}$ for a given choice of the function h in (II.1). We let for $j = 1, \ldots, q$, $\hat{\boldsymbol{b}}_j^n$ denote a d-dimensional vector of predicted values of the random components corresponding to the j^{th} cluster $\boldsymbol{B_j}$, based on at least $n = \min\{n_1, \ldots, n_q\}$ observations. Moreover, we assume that the predicted values are conditional asymptotically Gaussian distributed (as in Pelck & Labouriau (2021a)) for n increasing given $\boldsymbol{B_j} = \boldsymbol{b}_j$ with conditional expectation $\boldsymbol{b}_j$ and covariance matrix $\boldsymbol{V}_j$. Notice, that by the model assumptions, $\boldsymbol{V}_j$ is a diagonal matrix.

When q is small, we can maximise the following with respect to $\boldsymbol{\Sigma}$ by inserting an estimate $\hat{\boldsymbol{V}}_j$ of $\boldsymbol{V}_j$ and using a Gaussian Hermite approximation of the integral:

$$\begin{aligned}
L(\boldsymbol{\Sigma}; \hat{\boldsymbol{b}}_1^n, \ldots, \hat{\boldsymbol{b}}_q^n) &= \prod_{j=1}^{q} \int_{\mathbb{R}^d} \varphi(\boldsymbol{b}_j; \boldsymbol{\Sigma}) h(\hat{\boldsymbol{b}}_j^n; \boldsymbol{b}_j, \hat{\boldsymbol{V}}_j) d\boldsymbol{b}_j \\
&= \prod_{j=1}^{q} \int_{\mathbb{R}^d} \varphi(\boldsymbol{b}_j; \boldsymbol{\Sigma}) |2\pi \hat{\boldsymbol{V}}_j|^{-1/2} \exp\Big(-\tfrac{1}{2} (\boldsymbol{b}_j - \hat{\boldsymbol{b}}_j^n)^T \boldsymbol{V}_j^{-1} (\boldsymbol{b}_j - \hat{\boldsymbol{b}}_j^n) \Big) d\boldsymbol{b}_j \\
&= \pi^{-d/2} \prod_{j=1}^{q} \int_{\mathbb{R}^d} \varphi(\sqrt{2} \hat{\boldsymbol{V}}_j^{1/2} \tilde{\boldsymbol{b}}_j + \hat{\boldsymbol{b}}_j^n; \boldsymbol{\Sigma}) \exp\Big(-\tilde{\boldsymbol{b}}_j^T \tilde{\boldsymbol{b}}_j \Big) d\tilde{\boldsymbol{b}}_j \\
&\approx \pi^{-d/2} \prod_{j=1}^{q} \sum_{\boldsymbol{k} \in \mathcal{K}} (w_{k_1} \ldots w_{k_d}) \varphi(\sqrt{2} \hat{\boldsymbol{V}}_j^{1/2} \boldsymbol{x}_{\boldsymbol{k}} + \hat{\boldsymbol{b}}_j^n; \boldsymbol{\Sigma}),
\end{aligned}$$

with $\mathcal{K} = \{1, \ldots, l\}^d$, $\boldsymbol{w}_{\boldsymbol{k}} = (w_{k_1}, \ldots, w_{k_d})$, $\boldsymbol{x}_{\boldsymbol{k}} = (x_{k_1}, \ldots, x_{k_d})$, where x_{k_j} denotes the k_j^{th} root of the Hermite polynomial with l nodes and w_{k_j} is the associated weight.

II.A.2 Density of V in Case of Gaussian Random Components

In this section we present a formula for the density of the distribution of V defined in Equation (II.5). Recall, that V is distributed according to

$$V \sim \prod_{i=2}^{k-1} \prod_{j=1}^{d_i} Z_{ij},$$

where the random variables $Z_{21}, \ldots, Z_{(k-1)d_{k-1}}$ are independent and $Z_{ij} \sim \text{Beta}\Big(\frac{1}{2}[q - \bar{d}_i - j], \frac{1}{2}\bar{d}_i\Big)$ with $\bar{d}_i = d_1 + \ldots + d_{i-1}$ for $i = 2, \ldots, (k-1)$ and $j = 1, \ldots, d_i$.

Let $d = d_2 + \ldots + d_{k-1}$. We adapt the notation in Tang & Gupta (1984) to obtain the density of V. Define $t_1 = (2,1), \ldots, t_{d_2} = (2, d_2), t_{d_2+1} = (3,1), \ldots, t_{d_2+d_3} = (3, d_3), \ldots, t_{d_2+\ldots+d_{k-2}+1} = (k-1, 1), \ldots,$
$t_{d_2+\ldots+d_{k-1}} = (k-1, d_{k-1})$. The density of V can then be formulated as

$$f_V(v) = K_d v^{b(d)-1} (1-v)^{h(d)-1} \sum_{r=0}^{\infty} \sigma_r^{(d)} (1-v)^r \quad \text{for } 0 < v < 1, \tag{II.6}$$

where

$$\begin{aligned}
K_d &= \prod_{j=1}^{d} \frac{\Gamma[c(j)]}{\Gamma[b(j)]}, \\
h(d) &= \sum_{j=1}^{d} [c(j) - b(j)]
\end{aligned}$$

with $\Gamma(\cdot)$ denoting the Gamma function, $b(j) = \frac{1}{2}(q - \bar{d}_{t_j^{(1)}} - t_j^{(2)})$ and $c(j) = \frac{1}{2}\bar{d}_{t_j^{(1)}} + b(j)$ for $t_j = (t_j^{(1)}, t_j^{(2)})$. The term $\sigma_r^{(d)}$ can be calculated by the recursive relation:

$$\sigma_r^{(j)} = \frac{\Gamma[h(j-1)+r]}{\Gamma[h(j)+r]} \sum_{s=0}^{r} \left[\frac{c(j) - b(j-1)}{s!} \sigma_{r-s}^{(j-1)} \right], \; r = 0, 1, 2, \ldots, \; j = 2, 3, \ldots, d,$$

with initial values $\sigma_0^{(1)} = (\Gamma[h(1)])^{-1}$ and $\sigma_r^{(1)} = 0$ for $r = 1, 2, \ldots$. Notice, that

$$\lim_{j \to \infty} \frac{\Gamma[h(j-1)+r]}{\Gamma[h(j)+r]} = 0,$$

such that the infinite sum in (II.6) can be truncated after some point.

Paper III

Multivariate Generalised Linear Mixed Models for Studying Roots' Development

Jeanett S. Pelck
Aarhus University

Rodrigo Labouriau
Aarhus University

Abstract. The characterisation of the spatial and temporal distribution of the root system in a cultivated field depends on the soil volume occupied by the root systems (the scatter), and the local intensity of the root colonisation in the field (the intensity). We introduce a multivariate generalised linear mixed model for simultaneously describing the scatter and the intensity using data obtained with minirhizotrons (*i.e.,* tubes with observation windows, which are inserted in the soil, enabling to observe the roots directly). The presented models allow studying intricate spatial and temporal dependence patterns using a graphical model to represent the dependence structure of latent random components.

The scatter is described by a binomial generalised linear mixed model (presence of roots in observation windows). The number of roots crossing the reference lines in the observation windows of the minirhizotron is used to estimate the intensity through a specially defined Poisson generalised linear mixed model. We explore the fact that it is possible to construct multivariate extensions of generalised linear mixed models that allow to simultaneously represent patterns of dependency of the scatter and the intensity along with time and space.

We present an example where the intensity and scatter are simultaneously determined at three different time points. A positive association between the intensity and scatter at each time point was found, suggesting that the plants are not compensating a reduced occupation of the soil by increasing the number of roots per volume of soil. Using the general properties of graphical models, we identify a first-order Markovian dependence pattern between successively observed scatters and intensities. This lack of memory indicates that no long-lasting temporal causal effects are affecting the roots' development. The two dependence patterns described above cannot be detected with univariate models.

III.1 Introduction

The characterisation of the spatial and temporal distribution of the root system of a cultivated field depends, among other factors, on two key features: the volume occupied by the root systems in the field, here called the *scatter* (or root *frequency* in the terminology of Kristensen & Thorup-Kristensen 2004), and the local intensity of the root colonisation in the field, termed *intensity*. These two characteristics might vary with time and the use of different cultivation practices or treatments (see Kristensen & Thorup-Kristensen 2004, 2007, Kristensen & Stavridou 2017, Hefner & Labouriau 2019, Christensen et al. 2021). This article introduces and discusses a multivariate statistical model for simultaneously describing the root *scatter* and *intensity* using data obtained with a device called minirhizotron (briefly described below). Additionally, the models presented will allow us to study intricate spatial and temporal dependence patterns using a version of the so called *graphical model*, representing the dependence structure of latent random components.

A minirhizotron consists of a tube along with there are several transparent observation windows. According to the methodology, several tubes are inserted in the soil, allowing for observing the development of the roots at several depths, positions, and time points in the field. After a given growth period, the observation windows are examined using a camera introduced in the tube, and the presence or absence of roots in each window is registered. Each observation window has reference lines and the number of times the roots cross (if any) the reference lines are also recorded.

The main idea explored here is that the presence or not of roots in the observation windows is the result of sampling in the field and, therefore, can be used to characterise the volume occupied by the root system in the field, *i.e.,* to quantify the *scatter*. Furthermore, we will argue that the number of times the roots cross the reference lines can be used to estimate the *intensities*.

The *intensity* and *scatter* will both be modelled using suitable generalised linear mixed models, as described below. The *scatter* will be described by a binomial model (presence or not of roots in observation windows). The number of roots crossing the reference lines will be modelled using a specially defined Poisson model. A stochastic geometric argument will allow us to use the number of crosses to obtain estimates of the length of the root system in the region surrounding the observation windows of the minirhizotrons. The models used here will contain random components, allowing to represent the dependence structure induced by the experimental designs typically used in the applications we have in mind.

We will explore the fact that it is possible to construct multivariate extensions of generalised linear mixed models that allow to simultaneously represent, in a single model, patterns of dependence of the *scatter* and the *intensity* along with time and space. This achievement is remarkable since the nature of these two quantities is very different. We will show an example where the *intensity* and the *scatter* are simultaneously determined at three different time points. Jointly modelling these six quantities will allow us to identify a first-order Markovian dependence pattern between successively observed *scatters* and *intensities*. Moreover, we will show a positive association between the *intensity* and the *scatter*, quantify the magnitude of

those associations at the three different observation times, and show that there is a decay of the association between the *intensity* and the *scatter* at the last observation time. This type of characterisation of the time development of the root system cannot be obtained using only univariate analyses.

This article is organised as follows. Section III.2.1 presents a motivational real example. The basic models for the *scatter* and the *intensity* are discussed in the sections III.2.2 and III.2.3, respectively. In Section III.3, we introduce a multivariate model for describing the roots' colonisation at different developing stages simultaneously. After defining a multivariate generalised linear model connecting the univariate models in Section III.3.1, we model the covariance structure of those random components using a graphical model in Section III.3.2, and briefly describe techniques for inferring this graphical model in Section III.3.3. The motivational example presented in Section III.2.1 is analysed in Section III.4 using the multivariate model described. Section III.5 presents a brief discussion of the methods exposed.

III.2 Models for Scatter and Intensity of the Roots' Colonisation

III.2.1 A Motivational Reference Example

We consider below a real example arising from a study on the effects of different liming and phosphorous fertilisation techniques in a field experiment (see Christensen 2017 and Christensen et al. 2021). This example will be used to expose the modelling approach studied in this article. In this study, an experimental field cultivated with spring barley was split into three blocks containing four plots; in each block, four fertilisation treatments were randomly allocated to the plots. In each plot, two minirhizotron tubes were installed. Three soil depth zones were considered in the analyses below: the superficial layer (termed horizon A), the intermediate layer (called horizon B), and the subsoil (termed horizon C). The minirhizotron tubes had six observation windows in the superficial layer and twelve windows in the other two layers. The observation windows of all the 24 tubes were examined in three time points corresponding to different development stages of the culture of spring barley (see details in Christensen 2017). For simplicity of the exposition, we ignore the block and plot structure of the experiment.

The primary interest in the study referred above was to characterise the development of the root system in each soil depth zone when different fertilisation treatments are used. Here we approach a different question of characterising how the dependence between the *intensity* and *scatter* vary over the three observed development stages of the culture in the field. This problem involves studying the dependence of quantities of different stochastic nature. Indeed, while we will characterise the *intensity* using the counts of number of times the roots cross the reference lines in the observation windows, the *scatter* will be characterised examining the incidence of roots in observation windows.

The strategy we will adopt to analyse this example is to construct suitable

multivariate generalised linear mixed models describing the rooting *intensities* and the *scatters* at the three developmental stages (so the model will be six-dimensional). The one-dimensional generalised linear mixed models describing these two characteristics of the rooting system were first described in Labouriau (2019) and are presented in detail in Section III.2. The idea we will explore is that the fixed effects of the models will adjust for the expected differences due to the treatments and the soil depth zones. Each of these models will contain a Gaussian random component taking the same value for each observation arising from the same minirhizotron tube (here we interpret the tubes as the experimental units). Those random components represent the local variation of the *intensity* or the *scatter* present at each experimental unit after having corrected for the effects of the depth zones and the fertilisation treatments. The multivariate model we will consider will allow us to represent different covariance structure of the six Gaussian random components corresponding to the six observed responses. The covariance structure of the random components will determine the covariance structure of the responses, as we discuss below.

III.2.2 Modelling the Scatter

We model the *scatter* at a fixed development stage by studying the occurrence of roots in the different observation windows of the minirhizotron tubes. Denote by $Y^{[d]}_{tkz}$ the random variable representing the number of windows where a root is present in the z^{th} soil depth zone ($z = A, B, C$ representing the soil horizons) at the k^{th} tube ($k = 1, \ldots, 6$) exposed to the t^{th} treatment ($t = 1, \ldots, 4$), observed at the d^{th} development stage ($d = 1, 2, 3$). We keep the development stage fixed in this section and in Section III.2.3. Moreover, following the same convention for the sub-indices used above, denote the number of observation windows at the tkz^{th} observation by n_{tkz}. Note that by design, the number of observation windows does not change in the different observation times.

Denote, for $t = 1, \ldots, 4$ and $k = 1, \ldots, 6$, by $U^{[d]}_{tk}$ an unobservable random variable taking the same value for all observations arising from the tk^{th} tube. We assume that those random variables, corresponding to the 24 tubes used in the experiment, are independent and normally distributed with expectation zero and variance $\sigma^2_{U[d]}$. According to the model in discussion, the random variables $Y^{[d]}_{11A}, \ldots, Y^{[d]}_{46C}$, representing the observations, are conditionally independent given the random components $U^{[d]}_{11}, \ldots, U^{[d]}_{46}$. Moreover, we assume that, for $t = 1, \ldots, 4$, $k = 1, \ldots, 6$ and $z = A, B, C$, the random variable Y^{d}_{tkz} is conditionally binomial distributed given $U^{[d]}_{tk}$, with $Y^{[d]}_{tkz}|U^{[d]}_{tk} = u \sim \text{Bi}(n_{tkz}, p^{[d]}_{tkz})$, where

$$\text{logit}(p^{[d]}_{tkz}) = \beta_{tz,[d]} + u, \text{ for all } u \in \mathbb{R}\,. \tag{III.1}$$

The model described above coincides with a generalised linear mixed model (GLMM) defined with the binomial distribution, the logistic link function, a fixed effect representing the interaction of treatment and soil depth zone, and a random component representing the tubes.

The parameter $\beta_{tz,[d]}$ in (III.1) is clearly related to the *scatter*. Indeed, according to the model above, the probability of finding a root which is visible in an observation

window at the z^{th} soil depth of the plots that received the t^{th} treatment ($z = A, B, C$ and $t = 1, \ldots, 4$) at the d^{th} development stage is

$$E\left[\frac{Y_{tkz}^{[d]}}{n_{tkz}}\right] = \int_{\mathbb{R}} \frac{\exp(\beta_{tz,[d]} + u)}{1 + \exp(\beta_{tz,[d]} + u)} \varphi(u; 0, \sigma_{U[d]}^2) du \stackrel{\text{def}}{=} \alpha_{tz}(\beta_{tz,[d]}, \sigma_{U[d]}^2) \stackrel{\text{def}}{=} \tilde{\alpha}_{tz}^{[d]}\,. \quad \text{(III.2)}$$

Here $\varphi(\cdot; 0, \sigma_{U[d]}^2)$ denotes the density of a normal distribution with expectation 0 and variance $\sigma_{U[d]}^2$, which is the distribution of the random component $U_{tk}^{[d]}$. The quantity $\tilde{\alpha}_{tz}^{[d]}$ can easily be evaluated once we have estimated the parameters $\beta_{tz,[d]}$ and $\sigma_{U[d]}^2$ by numerically integrating the integral in (III.2) or using a straightforward Monte Carlo integration.

III.2.3 Modelling the Intensity

Let $C_{tkz}^{[d]}$ be a random variable representing the total number of times the roots cross the reference lines in all the observation windows at the z^{th} soil depth zones ($z = A, B, C$) at the k^{th} tube ($k = 1, \ldots, 6$) exposed to the t^{th} treatment ($t = 1, \ldots, 4$), observed at the d^{th} development stage ($d = 1, 2, 3$, fixed along this section).

Denote, for $t = 1, \ldots, 4$ and $k = 1, \ldots, 6$, by $V_{tk}^{[d]}$ an unobservable random variable taking the same value for all observations arising from the same tube. Those random variables are assumed to be independent and normally distributed with expectation zero and variance $\sigma_{V[d]}^2$. The random components defined above are analogous to the random components used for modelling the *scatter*. According to the model, the random variables $C_{11A}^{[d]}, \ldots, C_{46C}^{[d]}$, representing the observations of numbers of crosses, are conditionally independent given the random components $V_{11}^{[d]}, \ldots, V_{46}^{[d]}$. Moreover, we assume that, for $t = 1, \ldots, 4$, $k = 1, \ldots, 6$ and $z = A, B, C$, the random variable $C_{tkz}^{[d]}$ is conditionally Poisson distributed given $V_{tk}^{[d]}$, with conditional expectation given by

$$\log(\mathbb{E}[C_{tkz}^{[d]} | V_{tk}^{[d]} = v]) = \theta_{tz,[d]} + v + \log(n_{tkz}) \text{ for all } v \in \mathbb{R}\,. \quad \text{(III.3)}$$

The model above allows us to estimate the local length of the roots visible in the observation windows, characterising in this way the root *intensity*, as described below. Exponentiating both sides of (III.3) and taking expectations with respect to the distribution of the random components yields for $t = 1, \ldots, 4$, $k = 1, \ldots, 6$ and $z = A, B, C$, that

$$E\left[\frac{C_{tkz}^{[d]}}{n_{tkz}}\right] = \int_{\mathbb{R}} \exp(\theta_{tz,[d]}) \exp(v) \varphi(v; 0, \sigma_{V[d]}^2) dv = \exp(\theta_{tz,[d]}) \exp(\sigma_{V[d]}^2/2) \stackrel{\text{def}}{=} \omega_{tkz}^{[d]}\,. \quad \text{(III.4)}$$

The factor $\exp(\sigma_{V[d]}^2/2)$ in the right side of (III.4) is the expectation of the corresponding log-normal distribution (see Aitchison & Brown 1957).

The quantity $\omega_{tkz}^{[d]}$ defined in (III.4) is straightforwardly related to the *intensity* since the more intense the root colonisation process in a region around the tube is,

the more likely will be the occurrence of roots crossing the reference lines in the observation windows. Additionally, $\omega_{tkz}^{[d]}$ can be interpreted as an estimate of the length of the roots that are visible in an observation window using the argument sketched below. A classic argument for the Buffon's needle problem allows one to calculate the length of a rigid straight needle by randomly throwing the needle in a surface with parallel reference lines (see Klain & Rota 1997), the length of the needle being proportional to the probability of the needle cross a line. The Buffon's needle problem can be extended by dropping the assumption that the needle has a perfect straight form, yielding the so called Buffon's noodle problem. According to Ramaley (1969) the length of a one dimensional structure (the "noodle" replacing the needle) is proportional to the mean of the number of times the structure crosses the reference lines. Taking this approach, the left size of (III.4) is interpreted as the expected value of the Buffo's noodle estimate of the length of the roots that are visible in the observation windows.

The model described above coincides with a generalised linear mixed model (GLMM) defined with the Poisson distribution, the logarithm link function, a fixed effect representing the interaction of treatment and soil depth zone, an offset representing the logarithm of the number of observation windows and a random component representing the tubes.

III.3 Multivariate Simultaneous Models for the Scatter and the Intensity of the Roots' Colonisation

III.3.1 The Multivariate Construction

In Section III.2, we introduced GLMMs representing, separately, the *scatter* and the *intensity* of the root colonisation at a given development stage. Here, we construct a multivariate model for simultaneously describing these two characteristics at the three observed development stages. First, we define a GLMM for the *scatter* and for the *intensity* for each of the three development stages, as we explained above. In each of these models, there is a random component taking the same value for all the observations arising from the same experimental unit (*i.e.,* the same tube). Therefore, we might connect the models by assuming that the six random components are multivariate normally distributed. As we will argue below, the covariance structure of the multivariate distribution of the random components will allow us to characterise a type of association between the *scatters* and the *intensities* observed in the same or at different development stages. The details of this construction are given below.

According to the multivariate GLMM that we propose, for $t = 1, \ldots, 4$, $k =$

$1, \ldots, 6$, $z = A, B, C$, and $d = 1, 2, 3$,

$$\begin{cases} Y_{tkz}^{[1]}|U_{tk}^{[1]} = u_1 \sim \text{Bi}\left(n_{tkz}, \text{logit}^{-1}\left\{\beta_{tz,[1]} + u_1\right\}\right), \forall u_1 \in \mathbb{R} \\ Y_{tkz}^{[2]}|U_{tk}^{[2]} = u_2 \sim \text{Bi}\left(n_{tkz}, \text{logit}^{-1}\left\{\beta_{tz,[2]} + u_2\right\}\right), \forall u_2 \in \mathbb{R} \\ Y_{tkz}^{[3]}|U_{tk}^{[3]} = u_3 \sim \text{Bi}\left(n_{tkz}, \text{logit}^{-1}\left\{\beta_{tz,[3]} + u_3\right\}\right), \forall u_3 \in \mathbb{R} \\ C_{tkz}^{[1]}|V_{tk}^{[1]} = v_1 \sim \text{Po}\left(\exp\left\{\theta_{tz,[1]} + v_1\right\}\right), \forall v_1 \in \mathbb{R} \\ C_{tkz}^{[2]}|V_{tk}^{[2]} = v_2 \sim \text{Po}\left(\exp\left\{\theta_{tz,[2]} + v_2\right\}\right), \forall v_2 \in \mathbb{R} \\ C_{tkz}^{[3]}|V_{tk}^{[3]} = v_3 \sim \text{Po}\left(\exp\left\{\theta_{tz,[3]} + v_3\right\}\right), \forall v_3 \in \mathbb{R}. \end{cases} \tag{III.5}$$

Here $U_{tk}^{[d]}$ and $V_{tk}^{[d]}$ (for $d = 1, 2, 3$, $t = 1, \ldots, 4$ and $k = 1, \ldots, 6$) are Gaussian random components representing the k^{th} tube exposed to the t^{th} treatment belonging to a marginal model describing the *scatter* and the *intensity*, respectively, at the d^{th} development stage. We assume that $(U_{tk}^{[1]}, U_{tk}^{[2]}, U_{tk}^{[3]}, V_{tk}^{[1]}, V_{tk}^{[2]}, V_{tk}^{[3]})$ is multivariate normally distributed with expectation zero and covariance matrix $\mathbf{\Sigma}$. Furthermore, we assume that $(U_{tk}^{[1]}, U_{tk}^{[2]}, U_{tk}^{[3]}, V_{tk}^{[1]}, V_{tk}^{[2]}, V_{tk}^{[3]})$ and $(U_{t'k'}^{[1]}, U_{t'k'}^{[2]}, U_{t'k'}^{[3]}, V_{t'k'}^{[1]}, V_{t'k'}^{[2]}, V_{t'k'}^{[3]})$ are independent when $(t, k) \neq (t', k')$, *i.e.,* we assume that the random components corresponding to different tubes are independent.

III.3.2 Modelling the Covariance Structure of the Random Components

The next step in the construction of the multivariate model we have in mind is to model the covariance structure of the random components by a graphical model. Before embracing this project, we give a short account of the theory of graphical models. For a comprehensive description see Lauritzen (1996) and Whittaker (1990).

Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with a set of vertices, $\mathcal{V}$, composed of random variables. The set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is formed with the convention that two vertices are connected by an edge if, and only if, they are not conditionally independent given the remaining random variables in $\mathcal{V}$. We say that there is a path between two vertices, say V_1 and V_2, if there exist a sequence of pairs of vertices in $\mathcal{E}$ such that V_1 and V_2 belong to at least one vertex of the sequence. A set of vertices $\mathcal{S}$ is said to separate the sets of vertices $\mathcal{A}$ and $\mathcal{B}$ in the graph, if and only if, each path connecting an element of $\mathcal{A}$ to an element of $\mathcal{B}$ contains at least one element of $\mathcal{S}$. The separation principle (or global Markov property) is a crucial property of graphical models stating that if a set of vertices, $\mathcal{S}$, separates two disjoint subsets of vertices $\mathcal{A}$ and $\mathcal{B}$, then all variables in $\mathcal{A}$ are independent of all variables in $\mathcal{B}$ given the variables in $\mathcal{S}$, see Lauritzen (1996) and Perl (2009).

In the construction of the multivariate model in question here, we consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V} = \left\{U^{[1]}, U^{[2]}, U^{[3]}, V^{[1]}, V^{[2]}, V^{[3]}\right\}$ formed by the random variables representing the random components of the multivariate GLMM described in Section III.3.1 (with the obvious notational convention, *e.g.,* , $U^{[1]}$ is the random variable with the same distribution as $U_{tk}^{[1]}$, for $t = 1, \ldots, 4$, $k = 1, \ldots, 6$). Clearly, $\left(U^{[1]}, U^{[2]}, U^{[3]}, V^{[1]}, V^{[2]}, V^{[3]}\right) \sim N(\mathbf{0}, \mathbf{\Sigma})$ by construction. Here we define

the set of edges $\mathcal{E}$ using the conditional independence of pairs of random variables as in the paragraph above.

The interpretation of the graph above in terms of the random components of the multivariate generalised mixed model defined in Section III.3.1 is straightforward. For example, if there is an edge connecting $U^{[1]}$ and $V^{[1]}$, then these two random variables are not conditionally independent given the other random components; therefore, $U^{[1]}$ carries some information on $V^{[1]}$ and this information is not contained in the other random components. Note that $U^{[1]}$ and $V^{[1]}$ represent a latent variation on the *scatter* and the *intensity*, respectively, after having corrected for the effects of the treatment and the depth zone. Therefore, we would conclude that we have evidence that some latent mechanisms governing the *scatter* and *intensity* are related. The strength of this association can be estimated by inferring the entry of the precision matrix (*i.e.,* $\boldsymbol{\Sigma}^{-1}$) corresponding to the random components $U^{[1]}$ and $V^{[1]}$.

Note that the separation principle also holds in this graph, which is crucial for the interpretation of the model in terms of the covariance structure of the random components. Moreover, it is possible to extend the separation principle, obtaining what we call the *induced separation principle*, to draw some general conclusions on the response variables, as we explain below using a putative example. Consider the groups of random components $\mathcal{A} = \left\{U^{[1]}, V^{[1]}\right\}$ and $\mathcal{B} = \left\{U^{[3]}, V^{[3]}\right\}$ and $\mathcal{S} = \left\{U^{[2]}, V^{[2]}\right\}$ contained in $\mathcal{V}$. Define (for any choice of $t = 1, \ldots, 4$, $k = 1, \ldots, 6$, $z = A, B, C$) the sets of response variables $\widetilde{\mathcal{A}} = \left\{Y^{[1]}_{tkz}, C^{[1]}_{tkz}\right\}$ and $\widetilde{\mathcal{B}} = \left\{Y^{[3]}_{tkz}, C^{[3]}_{tkz}\right\}$. The sets $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ are the sets of response variables of the marginal models for which the elements of $\mathcal{A}$ and $\mathcal{B}$ are random components. According to the induced separation principle, if $\mathcal{S}$ separates $\mathcal{A}$ and $\mathcal{B}$ in the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, then the random variables in $\widetilde{\mathcal{A}}$ are conditionally independent of the random variables in $\widetilde{\mathcal{B}}$ given the random variables in $\mathcal{S}$. The proof of the induced separation principle can be done by using basic properties of conditional densities and using the factorisation of the joint densities of the distributions of $\mathcal{V}$, see the details in Pelck & Labouriau (2021b).

We stress that in the putative example above, conditioning on the responses in $\widetilde{\mathcal{S}} = \left\{Y^{[2]}_{tkz}, C^{[2]}_{tkz}\right\}$ (*i.e.,* the corresponding response variables to the elements of $\mathcal{S}$) does not necessarily render the random variables in $\widetilde{\mathcal{A}}$ independent of the random variables in $\widetilde{\mathcal{B}}$. Still in the putative example in discussion, the fact that the group of random components $\mathcal{S}$ separates $\mathcal{A}$ and $\mathcal{B}$ implies, in the present setup, that the responses at the first development stage are conditionally independent of the responses at the third development stage, given the random components related to the second development stage. We will see in Section III.4 that this indeed the case.

III.3.3 Inferring the Covariance Structure

The graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V} = \left\{U^{[1]}, U^{[2]}, U^{[3]}, V^{[1]}, V^{[2]}, V^{[3]}\right\}$ defining the covariance structure of the random components introduced in Section III.3.2 can be inferred by predicting the random components and using those predictors to infer a graphical model that minimises the BIC, as proposed and implemented in Abreu et al. (2010) and Edwards et al. (2010). The predictors of the random components of the

generalised linear mixed models can be obtained with inference procedures yielding consistent and normally distributed predictors. We used the procedure described in Pelck & Labouriau (2021a). Additionally, we tested each of the possible vertices using the conditional test for random components under multivariate generalised linear mixed models described in Pelck & Labouriau (2021b).

III.4 Analysing the Motivational Example

The example described in Section III.2.1 is analysed below. Figure III.1 displays a representation of the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V} = \left\{U^{[1]}, U^{[2]}, U^{[3]}, V^{[1]}, V^{[2]}, V^{[3]}\right\}$ estimated by minimising the BIC. Additionally, we tested whether the conditional covariances of each of the possible pairs of elements of $\mathcal{V}$, given the other elements of $\mathcal{V}$, is zero. The conditional covariances corresponding to the edges in the graph $\mathcal{G}$ displayed in Figure III.1 were all significantly different than zero (at a significance level of 0.05). Moreover, the conditional covariances corresponding to the pairs of elements of $\mathcal{V}$ that are not in $\mathcal{E}$ were not significantly different than zero (at a significance level of 0.05).

According to the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ displayed in Figure III.1, the *scatter* and the *intensity* are positively conditionally correlated at each of the three development stages studied, suggesting the presence of underlying mechanisms associated to the *scatter* and the *intensity* that are positively associated. The information that *scatter* carries on the *intensity* (and vice-versa), for each development stage, are not contained in the other random components, *i.e.,* there is evidence of specific common or cooperative underlying mechanisms determining the *scatter* and the *intensity*, specific for each development stage. This result rules out the possibility that the plants would be compensating a reduced occupation of the soil by increasing the intensity of the colonisation of the soil by the radicular system. Moreover, the strength of these associations is essentially the same in the first two development stages (testing the equality of the two conditional correlations yields the p-value 0.979), but decreases in the last development stage (p-values for comparing the first and the third stage and the second and the third stage are <0.001 and <0.0001, respectively).

The inferred graph $\mathcal{G}$ indicates the presence of a temporal Markovianity in the sense that the random components related to the first and the third development stages are conditionally independent given the random components associated to the second development stage. To see that, note that the set $\mathcal{S} = \{U^{[2]}, V^{[2]}\}$ separates the sets $\mathcal{A} = \{U^{[1]}, V^{[1]}\}$ and $\mathcal{B} = \{U^{[3]}, V^{[3]}\}$ in the graph $\mathcal{G}$. It follows then from the separation principle that the random components of $\mathcal{A}$ are conditionally independent of the random components of $\mathcal{B}$ given $\mathcal{S}$.

Using the induced separation principle, we conclude that for any choice of treatment $t = 1, \ldots, 4$ and tube $k = 1, \ldots, 6$, the group of responses $\widetilde{\mathcal{A}} = \{Y^{[1]}_{tkA}, C^{[1]}_{tkA}, Y^{[1]}_{tkB}, C^{[1]}_{tkB}, Y^{[1]}_{tkC}, C^{[1]}_{tkC}\}$ and $\widetilde{\mathcal{B}} = \{Y^{[3]}_{tkA}, C^{[3]}_{tkA}, Y^{[3]}_{tkB}, C^{[3]}_{tkB}, Y^{[3]}_{tkC}, C^{[3]}_{tkC}\}$ are conditional independent given the group of random components $\mathcal{S} = \{U^{[2]}_{tk}, V^{[2]}_{tk}\}$. That is, all the information that the response variables observed in the first development stage

(*i.e.,* the variables in $\widetilde{\mathcal{A}}$) might carry on the response variables in the third stage (*i.e.,* the variables in $\widetilde{\mathcal{B}}$) is entirely contained in the informational contents contained in the random components associated to the second development stage, and this conclusion holds for any combination of fertilisation treatment and depth zone. This lack of memory result described above indicates that there are no long lasting temporal causal effects affecting the roots' development (in terms of the *scatter* and the *intensity*), which is a non-trivial conclusion that cannot be reached with univariate models for describing the *scatter* and the *intensity*.

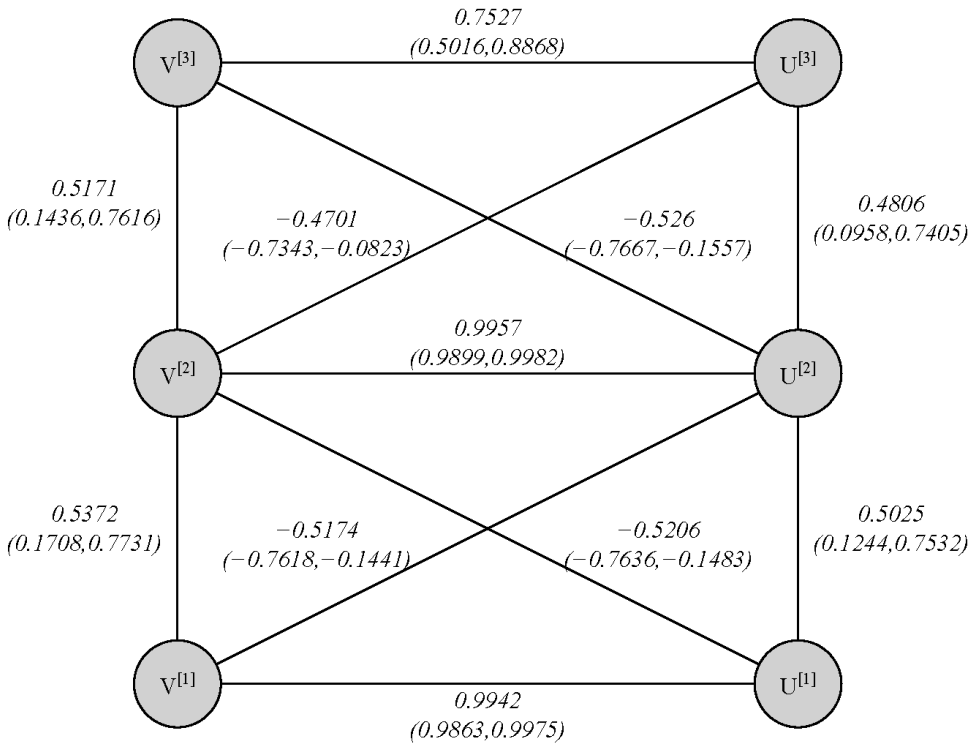


Figure III.1: Representation of the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V} = \left\{U^{[1]}, U^{[2]}, U^{[3]}, V^{[1]}, V^{[2]}, V^{[3]}\right\}$ formed by the random variables representing random components of the multivariate GLMM discussed in Section III.4. The graph was estimated by minimising the BIC of a covariance selection model inferred using the prediction of the random components. The super-indices in the vertices indicate in which development stage the response is observed. The letter V indicates that the random component correspond to the *intensity* while U represents the random components associated to the *scatter*. The numbers placed beside the edges are the estimated partial correlations with the corresponding confidence intervals (with coverage of 0.95).

III.5 Discussion and Conclusion

In this article, we combine multivariate GLMMs and graphical models (GMs) to characterise the spatial and temporal development of roots systems in a cultivation field. The use of graphical models in biological and agricultural research is not new, see Labouriau (2000), Labouriau & Amorim (2008, see the SOM), Holmstrup et al. (2011), Baral et al. (2017) and Taghizadeh-Toosi et al. (2019) were GMs are applied in different agricultural and biological contexts. In all these applications, the nature of the multivariate responses is essentially the same in the sense that whether the responses are continuous multivariate normally distributed or discrete multivariate multinomial distributed.

In Lamandé et al. (2011) and Azeez et al. (2020, using different data from the same experiment described here), the GM is of mixed type involving continuous multivariate normal distributed responses and discrete multivariate multinomial distributed responses. There, the GM uses the so-called CG (Conditional Gaussian) distributions as defined in Edwards et al. (2010) see also Abreu et al. (2010) for implementation and further discussions. However, in the current application, it is not possible to use multinomial based- graphical models since the supports of the distribution of counts (used to model the *intensity*) are not bounded and the binomial distributions (used for modelling the *scatter*) have different sizes for different observations. Therefore, one cannot do the analysis we discussed here using standard graphical models. The price we pay for modelling responses of varying nature (using graphical models that are not based on the CG distribution) is that we are forced to use random components, whether by directly interpreting the random components or by using the induced separation principle.

From the applied point of view, this article exemplifies the use of a combination of GLMMs and GMs, which allows exploring biological aspects such as the temporal development of two quantities of different nature. Here the close interface between Statistics and Biology played a crucial role in the modelling process.

Acknowledgements

We thank Hanne Lakkenborg Kristensen for having introduced us to the use of minirhizotron for characterising the development of the radicular system in fields. The statistical modelling presented here is a fruit of discussions with Hanne Lakkenborg Kristensen, Gitte Holton Rubæk, Lars Juhl Munkholm, Margita Hefner, Julie Therese Christensen, and Ellen Margrethe Wahlström, all of them from the Aarhus University. Gitte Holton Rubæk kindly supplied us the data of the motivational example. The first author was partially financed by the Applied Statistics Laboratory (aStatLab) at the Department of Mathematics, Aarhus University.

References

Abreu, G. C., Labouriau, R. & Edwards, D. (2010), 'High-dimensional graphical model search with gRapHD R package', *Journal of Statistical Software* **37**(1).

Aitchison, J. & Brown, J. A. (1957), *The lognormal distribution with special reference to its uses in economics*, Cambridge Univ. Press.

Azeez, M. O., Christensen, J. T., Ravnskov, S., Heckrath, G. J., Labouriau, R., Christensen, B. T. & Rubæk, G. H. (2020), 'Phosphorus in an arable coarse sandy soil profile after 74 years with different lime and p fertilizer applications', *Geoderma* **376**(114555).

Baral, K. R., Labouriau, R., Olesen, J. E. & Petersen, S. O. (2017), 'Nitrous oxide emissions and nitrogen use efficiency of manure and digestates applied to spring barley', *Agriculture, Ecosystems & Environment* **239**, 188–198.

Christensen, J. (2017), Long-term effects of liming and phosphorus application on the root growth conditions for spring barley on a sandy soil., Master's thesis, Aarhus University.

Christensen, J., Azeez, M., Labouriau, R., Ravnskov, S., Kristensen, H. & Munkholm, L. J., R. G. (2021), Responses of spring barley root growth, mycorrhizal colonisation, and grain yield to long-term lime and phosphate application strategies on sandy soil. (in press.

Edwards, D., de Abreu, G. C. & Labouriau, R. (2010), 'Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests', *BMC Bioinformatics* **11**(1).

Hefner, M. & Labouriau, R. (2019), 'Controlled traffic farming increased crop yield, root growth, and nitrogen supply at two organic vegetable farms', *Soil and Tillage Research* **191**, 117–130.

Holmstrup, M., Sørensen, J. G., Overgaard, J., Bayley, M., Bindesbøl, A.-M., Slotsbo, S., Fisker, K. V., Maraldo, K., Waagner, D., Labouriau, R. & Asmund, G. (2011), 'Body metal concentrations and glycogen reserves in earthworms (dendrobaena octaedra) from contaminated and uncontaminated forest soil', *Environmental Pollution* **159**(1), 190–197.

Klain, D. A. & Rota, G. (1997), *Introduction to geometric probability*, Cambridge University Press.

Kristensen, H. L. & Stavridou, E. (2017), 'Deep root growth and nitrogen uptake by rocket (diplotaxis tenuifolia l.) as affected by nitrogen fertilizer, plant density and leaf harvesting on a coarse sandy soil', *Soil Use and Management* **33**(1), 62–71.

Kristensen, H. L. & Thorup-Kristensen, K. (2004), 'Root growth and nitrate uptake of three different catch crops in deep soil layers', *Soil Science Society of America Journal* **68**(2), 529–537.

Kristensen, H. L. & Thorup-Kristensen, K. (2007), 'Effects of vertical distribution of soil inorganic nitrogen on root growth and subsequent nitrogen uptake by field vegetable crops', *Soil Use and Management* **23**(4), 338–347.

Labouriau, R. (2000), *Applying graphical models in plant genetics*, Institut National de la Recherche Agronomique (INRA), pp. 99–105.

Labouriau, R. (2019), 'Details on the modelling of the root growth data. supplying online material in hefner et al. 2019', *Soil and Tillage Research* **191**, 117–130.

Labouriau, R. & Amorim, A. (2008), 'Comment on "an association between the kinship and fertility of human couples"', *Science* **322**(5908), 1634–1634.

Lamandé, M., Labouriau, R., Holmstrup, M., Torp, S. B., Greve, M. H., Heckrath, G., Iversen, B. V., de Jonge, L. W., Moldrup, P. & Jacobsen, O. H. (2011), 'Density of macropores as related to soil and earthworm community parameters in cultivated grasslands', *Geoderma* **162**(3), 319–326.

Lauritzen, S. L. (1996), *Graphical models*, Vol. 17, Clarendon Press.

Pelck, J. S. & Labouriau, R. (2021a), Conditional inference for multivariate generalised linear mixed models. arXiv:2107.11765.

Pelck, J. S. & Labouriau, R. (2021b), Multivariate generalised linear mixed models with graphical latent covariance structure. arXiv:2107.14535.

Perl, J. (2009), *Causality: models, reasoning and inference*, second edition edn, Cambridge University Press.

Ramaley, J. F. (1969), 'Buffon's noodle problem.', *The American Mathematical Monthly* **76**(8).

Taghizadeh-Toosi, A., Elsgaard, L., Clough, T. J., Labouriau, R., Ernstsen, V. & Petersen, S. O. (2019), 'Regulation of n_2o emissions from acid organic soil drained for agriculture', *Biogeosciences* **16**(23), 4555–4575.

Whittaker, J. (1990), *Graphical models in applied multivariate analysis*, Chichester New York et al: John Wiley & Sons.

III.A Representation of the Covariance Structure in Terms of Direct Acyclic Graphs

The covariance structure of the responses and the random components at different developing stages can be represented in terms of a Direct Acyclic Graph (DAG) as follows. Here we use the terminology and properties of graphical models represented as DAGs exposed in Lauritzen (1996). This representation follows from the induced separation principle exposed in Section III.3.2 and the properties of the GLMM.

According to the inferred multivariate model, for any possible choice of treatment $t = 1, \ldots, 4$, tube $k = 1, \ldots, 6$, and depth zone $z = A, B, C$, the covariance structure of the responses and random components is expressed by the DAG represented in Figure III.2.

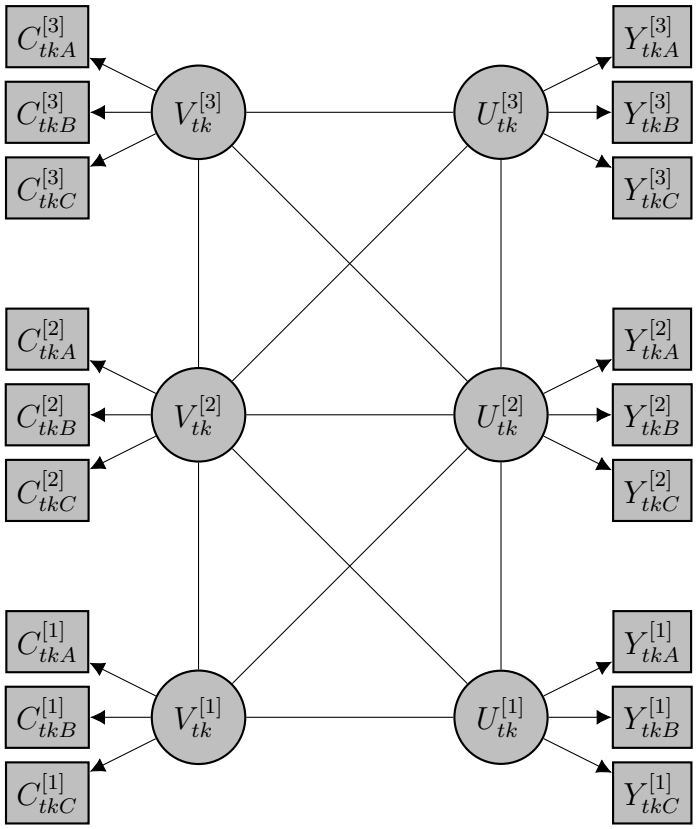


Figure III.2: DAG representation of the covariance structure of the responses and the random components for the tk^{th} tube at different developing stages induced by the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Edges without arrows represent arrows in both directions. The structure disputed holds for any choice of treatment t ($t = 1, \ldots, 4$) and tube k ($k = 1, \ldots, 6$).

Paper IV

A Multivariate Methodology for Analysing Students' Performance Using Register Data

Jeanett S. Pelck
Aarhus University

Rafael Pimentel Maia
University of Campinas

Hildete P. Pinheiro
University of Campinas

Rodrigo Labouriau
Aarhus University

Abstract. We present a new method for jointly modelling the students' results in the university's admission exams and their performance in subsequent courses at the university. The case considered involved all the students enrolled at the University of Campinas in 2014 to evening studies programs in educational branches related to exact sciences. We collected the number of attempts used for passing the university course of geometry and the results of the admission exams of those students in seven disciplines. The method introduced involved a combination of multivariate generalised linear mixed models (GLMM) and graphical models for representing the covariance structure of the random components. The models we used allowed us to discuss the association of quantities of very different nature. We used Gaussian GLMM for modelling the performance in the admission exams and a frailty discrete-time Cox proportional model, represented by a GLMM, to describe the number of attempts for passing Geometry.

The analyses were stratified into two populations: the students who received a bonus giving advantages in the university's admission process to compensate social and racial inequalities and those who did not receive the compensation. The two populations presented different patterns. Using general properties of graphical models, we argue that, on the one hand, the predicted performance in the admission exam of Mathematics could solely be used as a predictor of the performance in geometry for the students who received the bonus. On the other hand, the Portuguese admission exam's predicted performance could be used as a single predictor of the performance in geometry for the students who did not receive the bonus.

IV.1 Introduction

In this paper, we study the admission system to a Brazilian university and the bonus system for compensating social and racial inequalities. The data analysed below is based on the registers of entrance and performance at the University of Campinas, Brazil (UNICAMP). In 2005, UNICAMP implemented an affirmative action program giving extra bonus in the final entrance examination score for students who were enrolled for their entire high school years in the public system (with an additional bonus for those who self declared to be African / Indigenous Brazilian descendants). See Maia et al. (2016), Pedrosa et al. (2007), Pinheiro et al. (2019, 2020) for more details.

The main difficulty of the study of those registers is the multivariate nature of the characterisations of the object of interest, and the (unavoidable) presence of spurious associations. The responses observed in the data referred above are of very different nature but can be analysed in one multivariate model as we will describe below. The performance at the university is measured by the number of attempts required to pass the course of geometry, which is a key course in the beginning of the university education of the group of students we study. This response is typically right-censored, in the sense that there might be some students that have not passed the course when the data was collected, dropped out during the study or who's enrollment has been cancelled. The enrollment is cancelled if the student fails all the subjects in the first or second semester or reached the maximum number of semesters allowed without graduating (e.g. in Statistics this is 6 years). On the other hand, the performances at the admission exams are measured in a standardised scale using a scoring system. Those two types of responses are modelled differently: the time for passing a course is modelled using a variant of the frailty Cox proportional model with discrete-time; the scores at the entrance exam are described using a Gaussian mixed model. In both cases, the models can be represented as instances of generalised linear mixed models (GLMMs). The introduced GLMMs contain two common random components (one taking different values for each individual, and one taking the same value for individuals enrolled at the same branch of study). The random component related to the individuals allows us to connect the models describing the different responses, and in this way, characterise how much information each response carries on the other responses. This methodology differs from Pinheiro et al. (2020) where parts of the same data were analysed but in a different context.

The analyses performed were stratified into two populations: the students who received a bonus, and the students who did not receive the compensation. The two populations are different and presented different patterns, justifying the stratification.

This study aims to present statistical tools that allow to study the different facets of the type of data described above and to understand the associations between the different responses. These aims are fulfilled by using suitable multivariate versions of GLMMs and by using the theory of graphical models to describe the covariance structure of the common random component giving the variation between individuals (Pelck & Labouriau 2021b). We illustrate, in this way, a process of modelling responses of very different nature in a multivariate model that arises when working with

educational register data.

The paper is organised as follows. Section IV.2 describes the data used. The multivariate GLMM are introduced in Section IV.3, including details on the marginal Gaussian GLMMs and the frailty discrete-time Cox Proportional model. Section IV.4 describes the graphical model used for representing the covariance structure of the random components, and Section IV.5 presents and discusses the results. Appendix IV.A presents some model control, while some details of the representation of the graphical models are given in Appendix IV.B.

IV.2 Data Description

The data we used contain records on all the 299 students enrolled at the UNICAMP in 2014, in evening studies programs in one of the educational branches related to exact sciences listed in Table IV.1. Among those students, 151 received a bonus giving advantages in the university's admission process. The bonus group consists of students from a public high school and students from a public high school who are self-declared African or Indigenous Brazilian descendants.

Chemical engineering	Electrical engineering	Economical science
Mathematics	Physics	Computer science
Automation engineering	Technological chemistry	Bachelor in Chemistry
Medical Physics		

Table IV.1: Educational branches included in the population studied.

The data includes eight responses recorded for each student. The first seven responses correspond to the student's performance in the university admission exam in the disciplines: Mathematics, Physics, Chemistry, Biology, History, Geography, and Portuguese. The last response was the number of attempts the students used to pass the (first year) geometry course at the university. This response was right censored (with 24.08% of censoring) since some students used at least the observed number of attempts, but it was unknown whether the enrollment had been cancelled or the student passed the course after the data was collected. Additionally, the registers contain a range of information on each individual including gender and age.

IV.3 A Multivariate Model for for Simultaneously Describing the Admission Scores and the Performance in Geometry

The eight responses described above were jointly analysed using a multivariate generalised linear mixed model as described below. We performed separate parallel analyses for the students who received the bonus and those who did not receive the bonus, which, we anticipate, will yield contrasting results.

In each of the two separate analyses, we used a multivariate model combining seven Gaussian mixed models describing the seven admission exams, and a frailty Cox proportional model with discrete-time for modelling the number of attempts to pass the course of geometry. Each of the marginal models above included two random components: one accounting for the variation between the different study branches, and one representing each individual. The eight marginal models referred above were combined by assuming a joint multivariate Gaussian distribution for the random component representing the individuals. The precise model definition for the students who received bonus is given below. The models for the students that did not receive bonus are similarly defined.

In order to describe the models we will use, we index the individuals by i ($i = 1, \ldots, n$, with $n = 151$), the eight responses by j ($j = 1, \ldots, 8$) and the educational branches listed in Table IV.1 by k ($k = 1, \ldots, 10$). Moreover, the educational branch of the i^{th} student is denoted by $e(i)$. We describe below the covariance structure of the multivariate generalised linear mixed models we want to introduce using two random components. The random component representing the educational branches is defined by assuming that there exist 10 unobservable random variables $U_1^{[j]}, \ldots, U_{10}^{[j]}$ for each of the eight responses ($j = 1, \ldots, 8$) corresponding to the 10 educational branches. The random variable $U_k^{[j]}$ takes the same value for the j^{th} response for each student that is enrolled in the k^{th} educational branch ($k = 1, \ldots, 10$, $j = 1, \ldots, 8$). The random component representing the individuals are specified for the j^{th} response ($j = 1, \ldots, 8$) by defining the unobservable random variables $V_1^{[j]}, \ldots, V_n^{[j]}$ representing the n individuals.

According to the model, for $j, j' = 1, \ldots, 8$, the random vectors $\boldsymbol{U}^{[j]} \stackrel{\text{def}}{=} (U_1^{[j]}, \ldots, U_{10}^{[j]})^T$ and $\boldsymbol{V}^{[j']} \stackrel{\text{def}}{=} (V_1^{[j']}, \ldots, V_n^{[j']})^T$ are independent and multivariate Gaussian distributed as given below,

$$\begin{aligned} \boldsymbol{U}^{[j]} &\sim N_{10}\left(\boldsymbol{0}, \sigma^2_{U^{[j]}} \boldsymbol{I}_{10}\right) \\ \boldsymbol{V}^{[j']} &\sim N_n\left(\boldsymbol{0}, \sigma^2_{V^{[j']}} \boldsymbol{I}_n\right). \end{aligned}$$

Here $\boldsymbol{I}_m$ denotes a m-dimensional identity matrix (for $m \in \mathcal{N}$). Furthermore, for $j, j' = 1, \ldots, 8$, with $j \neq j'$, we assume that $\boldsymbol{U}^{[j]}$ is independent of $\boldsymbol{U}^{[j']}$, and that $V_i^{[j]}$ is independent $V_{i'}^{[j']}$, where $i, i' = 1, \ldots, n$ with $i \neq i'$. Additionally, we assume that for $i = 1, \ldots, n$,

$$\text{Cov}\left(V_i^{[1]}, \ldots, V_i^{[8]}\right) = \boldsymbol{\Sigma}_V, \tag{IV.1}$$

with $\text{diag}\left(\boldsymbol{\Sigma}_V\right) = \left(\sigma^2_{V^{[1]}}, \ldots, \sigma^2_{V^{[8]}}\right)$. Therefore, defining $\boldsymbol{V}^T \stackrel{\text{def}}{=} \left(\boldsymbol{V}^{[1]^T}, \ldots, \boldsymbol{V}^{[8]^T}\right)$ we have that

$$\text{Cov}\left(\boldsymbol{V}\right) = \boldsymbol{\Sigma}_V \otimes \boldsymbol{I}_n \,.$$

We will use the matrix $\boldsymbol{\Sigma}_V$ to characterise the dependence structure of the eight responses. In particular, in Section IV.4, we will use a graphical model structure by imposing zeroes in the inverse of $\boldsymbol{\Sigma}_V$, corresponding to assume that some pairs of the

random variable $V_i^{[1]}, \ldots, V_i^{[8]}$ are conditionally independent given the other random variables. For improving the readability of the discussion below, when relevant, we denote $V_i^{[1]}, \ldots, V_i^{[8]}$ by $V_i^{[Math]}, \ldots, V_i^{[Geom]}$,*i.e.,* we identify the superindices of the individual random components with a recognisable short form of the corresponding response.

We formulate the marginal generalised linear mixed models for the seven responses related to the admission exams by specifying the conditional expectation of the random variable $Y_i^{[j]}$, representing the i^{th} individual ($i = 1, \ldots, n$) in the j^{th} dimension ($j = 1, \ldots, 7$), given $(U_{e(i)}^{[j]}, V_i^{[j]})$ for $j = 1, \ldots, 7$, that is,

$$\mathbb{E}[Y_i^{[j]} | U_{e(i)}^{[j]} = u,\, V_i^{[j]} = v] = \boldsymbol{x}_i^T \boldsymbol{\beta}^{[j]} + u + v, \quad \forall\, u, v \in \mathbb{R}.$$

In all the models above, the term $\boldsymbol{x}_i^T \boldsymbol{\beta}^{[j]}$ ($j = 1, \ldots, 7$) adjusts for gender (female or male) and age (divided into two groups:under 21 or 21 and above). The conditional distributions are assumed to be Gaussian with identity link functions.

We formulate the discrete time frailty Cox proportional hazard model describing the number of attempts to pass the course of geometry. According to the model, for the student i^{th} ($i = 1, \ldots, n$), the discrete conditional hazard function given $(U_{e(i)}^{[8]}, V_i^{[8]})$ is

$$\begin{aligned}
&\lambda_i(t | U_{e(i)}^{[8]} = u,\, V_i^{[8]} = v) \\
&\quad \stackrel{\text{def}}{=} P(T_i = t | T_i \geq t, U_{e(i)}^{[8]} = u,\, V_i^{[8]} = v) \\
&\quad = \tilde{\lambda}_t \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}^{[8]}) \exp(u) \exp(v), \qquad \text{for } t = 1, 2, \ldots \text{ for all } u, v \in \mathbb{R}.
\end{aligned}$$

Here T_i is the random variable representing the number of attempts to pass the course of geometry used by the i^{th} individual and the term $\boldsymbol{x}_i^T \boldsymbol{\beta}^{[8]}$ adjusts for gender (female or male) and age (divided into two groups: under 21 or 21 and above). The model above coincides with a generalised linear mixed model defined with a binomial distribution and a logarithmic link function, applied to a specially constructed data representing the risk set of the related counting process. We used a Poisson approximation for avoiding numerical issues. See Maia et al. (2014).

IV.4 Modelling the Covariance Structure of the Random Components

We complete the specification of the multivariate generalised linear mixed model introduced in Section IV.3 by defining the covariance structure of the random components representing the individual's variation. Since the random components $V_i^{[1]}, \ldots, V_i^{[8]}$ have the same distribution for $i = 1, \ldots, n$, we suppress the subindex i from the notation and write $V^{[1]}, \ldots, V^{[8]}$ to denote $V_i^{[1]}, \ldots, V_i^{[8]}$ for an arbitrary individual.

The random components $V^{[1]}, \ldots, V^{[8]}$ represent the individual variation of the abilities of each of the n students affecting the performance in the seven admission

exams and in the course of geometry, respectively. Note that according to the model, the covariance of the random variables $V_i^{[1]}, \ldots, V_i^{[8]}$ is the same for all the individuals, namely $\boldsymbol{\Sigma}_V$, see (IV.1). Here, we will characterise this covariance structure common to all the individuals using graphical models, which will allow us to draw general conclusions on the interdependence between the eight responses studied. Before pursuing this task, we give a short account of the theory of graphical models; for a comprehensive description of this theory see Lauritzen (1996) and Whittaker (1990).

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a graph with a set of vertices, $\mathcal{V}$, and edges, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Each vertex represents a random variable, and two vertices are connected with an edge if, and only if, they are not conditionally independent given the remaining random variables. We say that there is a path between two vertices if there exist a sequence of pairs of vertices connected with an edge connecting the two vertices.

In the multivariate models described above, we consider a graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V} = \{V^{[1]}, \ldots, V^{[8]}\} = \{V^{[\text{Math}]}, \ldots, V^{[\text{Geom}]}\}$. The set of edges contains pairs of random variables which are not conditionally independent given the remaining random variables, and thus, they carry some information on each other that are not contained in the other random variables in $\mathcal{V}$. For example, suppose that there is an edge between $V^{[\text{Geom}]}$ and $V^{[\text{Math}]}$ associated to the individual responses related to the performance in the course of geometry and the admission exam in mathematics, respectively. This means that after correcting for differences in age, gender and education branch, the random variables $V^{[\text{Geom}]}$ and $V^{[\text{Math}]}$ are conditionally dependent given $\mathcal{V} \setminus \left\{V^{[\text{Geom}]}, V^{[\text{Math}]}\right\}$; therefore $V^{[\text{Math}]}$ carries information on $V^{[\text{Geom}]}$ that is not contained in the informational contents of the random variables $\mathcal{V} \setminus \left\{V^{[\text{Geom}]}, V^{[\text{Math}]}\right\}$. Conversely, the absence of an edge connecting two vertices indicates that the random variables associated to those vertices are conditionally independent given the other random variables in play; therefore, the knowledge of the other random variables renders the two random variables in question independent.

According to the theory of graphical models, a set of vertices, say $\mathcal{S}$, separates two sets of vertices $\mathcal{A}$ and $\mathcal{B}$ in the graph, if, and only if, each path connecting an element of $\mathcal{A}$ to an element of $\mathcal{B}$ contains at least one element of $\mathcal{S}$. A key result in the theory of graphical models is that if a set of vertices, $\mathcal{S}$, separates two disjoint subsets of vertices $\mathcal{A}$ and $\mathcal{B}$, then all the variables in $\mathcal{A}$ are independent of the variables in $\mathcal{B}$ given the variables in $\mathcal{S}$. This result is called the *separation principle* or *global Markov property* for undirected graphical models (Lauritzen 1996, page 32). For example, if the random variable $V^{[\text{Math}]}$ separates the random variable $V^{[\text{Geom}]}$ from the other vertices, then conditioning on solely $V^{[\text{Math}]}$ renders $V^{[\text{Geom}]}$ independent of the other vertices (*i.e.,* $\mathcal{V} \setminus \left\{V^{[\text{Geom}]}, V^{[\text{Math}]}\right\}$).

Note that the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ defined with $\mathcal{V} = \{V^{[\text{Math}]}, \ldots, V^{[\text{Geom}]}\}$ involves the unobservable random variables associated with the individual random components, not the observed responses. However, it is possible to extend the separation principle using the multivariate generalised linear mixed model's properties to discuss the interdependence of the observed responses. We explain this extended principle using an example. Suppose that the random variable $V^{[\text{Math}]}$ separates the random variable $V^{[\text{Geom}]}$ from the other vertices in $\mathcal{V}$, then, according to the induced separation

principle, each response corresponding to a vertex in $\mathcal{V} \setminus \left\{V^{[\text{Geom}]}, V^{[\text{Math}]}\right\}$, say $Y_i^{[*]}$, is independent of $T_i = T_i^{[\text{Geom}]}$ (*i.e.,* the number of attempts the i^{th} individual uses to pass the course of geometry) given $V^{[\text{Math}]}$. Here it is required to condition on the random variable $V^{[\text{Math}]}$ for obtaining independency of $Y_i^{[*]}$ and T_i; conditioning on the observable responses $Y_i^{[\text{Math}]}$ does not necessarily renders the variables $Y_i^{[*]}$ and T_i independent. See the Appendix IV.Band Pelck & Labouriau (2021b) for a general formulation of the induced separation principle using undirected graphical models.

The graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V} = \{V^{[1]}, \ldots, V^{[8]}\}$ was inferred using predicted values of the random variables to infer a graphical model that minimises the BIC. A method and implementation for minimising the BIC are described in Abreu et al. (2010), and an inference method for predicting values of the random components is presented in Pelck & Labouriau (2021a). Notice, that this method only yields normally distributed predictors in cases with small variance of the random components. However, in Labouriau (1998) it is shown that treating graphical models involving non Gaussian random variables as being normally distributed corresponds to using optimal inferential procedure for semiparametric models under mild regularity conditions.

The survival model presented was controlled using the methods described in Maia et al. (2014), Edwards et al. (2010). We found no indication of a lack of fit of the models. The marginal Gaussian mixed models where controlled by standard residual analyses. See Appendix IV.A.

IV.5 Results and Discussion

Figure IV.1 displays the graphs representing the estimated graphical models describing the covariance structure of the individual random components for the students that received bonus and the students that did not. The two populations of students presented different covariance structures which we discuss below. We stress that each of the individuals' random components represents latent individual abilities affecting the performance related to the respective responses. The covariance structures described in Figure IV.1 are obtained after adjusting for differences in age, gender and educational branch.

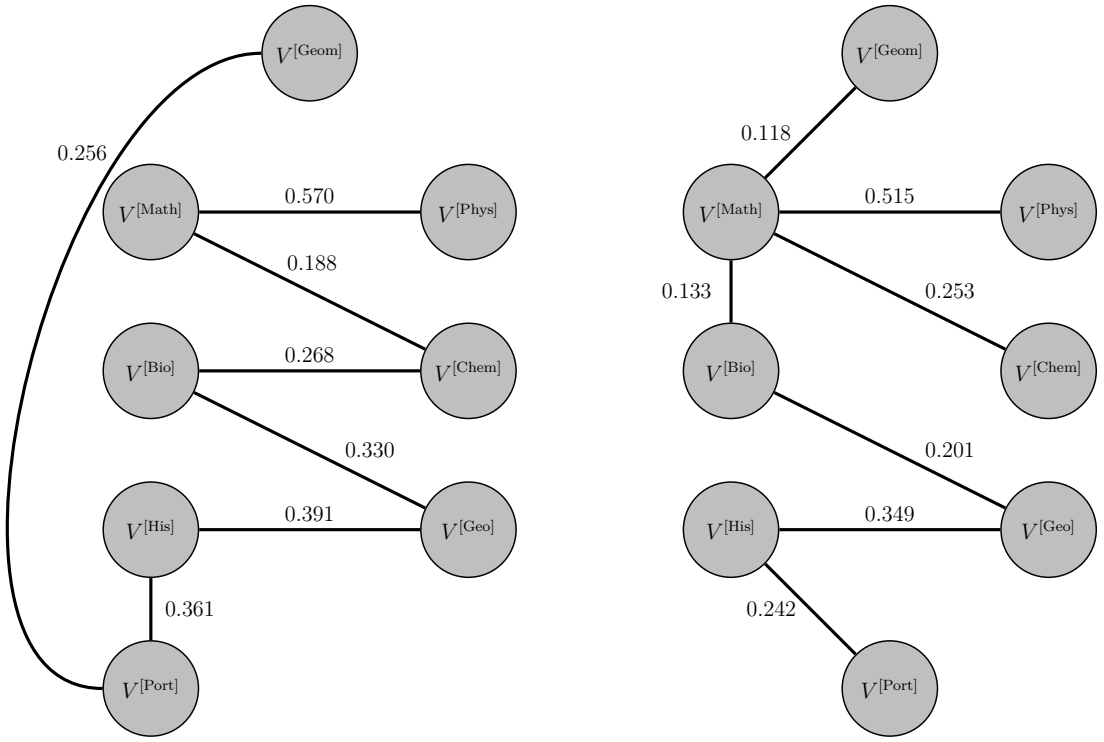
In the population of students that received bonus, the random component related to the performance in the course of geometry, $V^{[\text{Geom}]}$, is only connected to the random component related to the result in the admission exam of mathematics, $V^{[\text{Math}]}$ (see the left panel in Figure IV.1 and Figure IV.4). The conditional correlation between $V^{[\text{Geom}]}$ and $V^{[\text{Math}]}$ is positive. This result suggests the existence of common cognitive mechanisms associated with the latent abilities related to the performance in the admission exam of mathematics and in the course of geometry. Since $V^{[\text{Math}]}$ separates $V^{[\text{Geom}]}$ from the other individual random components, according to the separation principle, $V^{[\text{Geom}]}$ is conditionally independent of the other individual random components given $V^{[\text{Math}]}$. This conditional independence indicates that the putative common cognitive mechanisms referred above are specific to these

two disciplines and are not shared by the other disciplines' abilities. Regarding the results of the admission exams, according to the extended separation principle, the observed performance in the course of geometry is conditionally independent of the observed results of the admission exams given the individual random component $V^{[\text{Math}]}$. From the practical point of view, this result shows that the prediction of the random component $V^{[\text{Math}]}$ suffices for predicting the performance of the students that received the bonus in the course of geometry. After predicting $V^{[\text{Math}]}$, both the results of the other admission exams and their corresponding individual random components become uninformative concerning the performance in the course of geometry.

We obtain a different scenario for the population of the students that did not receive the bonus. There, the individual random component $V^{[\text{Port}]}$ (associated with the performance in the admission exam of Portuguese) is the only random component connected with the random component $V^{[\text{Geom}]}$; moreover, $V^{[\text{Port}]}$ separates $V^{[\text{Geom}]}$ from the other individual random components (see the right panel of Figure IV.1 and Figure IV.3). Therefore, using a similar rationale as above, we conclude that for the population of students that did not receive the bonus, the prediction of $V^{[\text{Port}]}$ suffices for predicting the performance of the students that received the bonus in the course of geometry.

The variance of the predictions of $V^{[\text{Math}]}$ is 43% larger in the population of students that received the bonus, as compared with the variance in the group that did not receive the bonus. A combination of two factors might cause this difference: in the population of students who received the bonus, there might be more considerable variability in the quality of the high school teaching in Mathematics; furthermore, the students with a lower level in Mathematics could enter the University by receiving a bonus. Therefore, the mathematics skills detected in the admission exam play an essential role in the performance in geometry among the students who received the compensation. One might speculate whether the random component $V^{[\text{Port}]}$ is representing social-economic class which plays a key role in the performance in the course of geometry population of students that did not receive the bonus.

There are also some similarities between the inferred covariance structures of the two populations of students studied. For example, in both populations, the individuals random component $V^{[\text{Bio}]}$ (related to the performance in the admission exam of biology) separates $V^{[\text{Math}]}$, $V^{[\text{Phys}]}$ and $V^{[\text{Chem}]}$ from $V^{[\text{His}]}$ and $V^{[\text{Geo}]}$. We let the reader explore further aspects of the results presented here. In this paper, we have exposed a new method based on a combination of multivariate generalised linear mixed models and graphical models for modelling and predicting students' performance using responses of different nature, namely, some Gaussian responses and a discrete right-censored response. Other response types can also be modelled by choosing different distributions and different link functions for constructing the marginal models.



(a) No bonus (b) Bonus

Figure IV.1: Independence graph representing the estimated graphical model describing the covariance structure of the individual random components $V_i^{[1]}, \ldots, V_i^{[8]}$ for an arbitrarily chosen individual ($i = 1, \ldots, n$, suppressing the index i in the graph). The estimated conditional correlations are reported for each edge.

Acknowledgement

We acknowledge the admission committee (CONVEST) from the University of Campinas (UNICAMP) for giving access to the data used. The first and the last authors were partially financed by the Applied Statistics Laboratory (aStatLab) from the Department of Mathematics, Aarhus University. The third author was partially financed by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Grant number: 310874/2018-1.

References

Abreu, G. C., Labouriau, R. & Edwards, D. (2010), 'High-dimensional graphical model search with gRapHD R package', *Journal of Statistical Software* **37**(1).

Edwards, D., de Abreu, G. C. & Labouriau, R. (2010), 'Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests', *BMC Bioinformatics* **11**(1).

Labouriau, R. (1998), Estimating Functions and Semiparametric Models, PhD thesis, Department of Theoretical Statistics, Aarhus University.

Lauritzen, S. L. (1996), *Graphical models*, Vol. 17, Clarendon Press.

Maia, R. P., Madsen, P. & Labouriau, R. (2014), 'Multivariate survival mixed models for genetic analysis of longevity traits', *Journal of Applied Statistics* **41**(6), 1286–1306.

Maia, R. P., Pinheiro, H. P. & Pinheiro, A. (2016), 'Academic performance of students from entrance to graduation via quasi u-statistics: a study at a brazilian research university', *Journal of Applied Statistics* **43**(1), 72–86.

Pedrosa, R. H., Dachs, J. N. W., Maia, R. P., Andrade, C. Y. & Carvalho, B. S. (2007), 'Academic performance, students' background and affirmative action at a brazilian university', *Higher education management and policy* **19**(3), 1–20.

Pelck, J. S. & Labouriau, R. (2021a), Conditional inference for multivariate generalised linear mixed models. arXiv:2107.11765.

Pelck, J. S. & Labouriau, R. (2021b), Multivariate generalised linear mixed models with graphical latent covariance structure. arXiv:2107.14535.

Pinheiro, H. P., Maia, R. P., Lima-Neto, E. A. & Rodrigues-Motta, M. (2019), 'Zero-one augmented beta and zero inflated discrete models with heterogeneous dispersion: an application to students' academic performance', *Stat. Methods Appl.* **28**, 749–767.

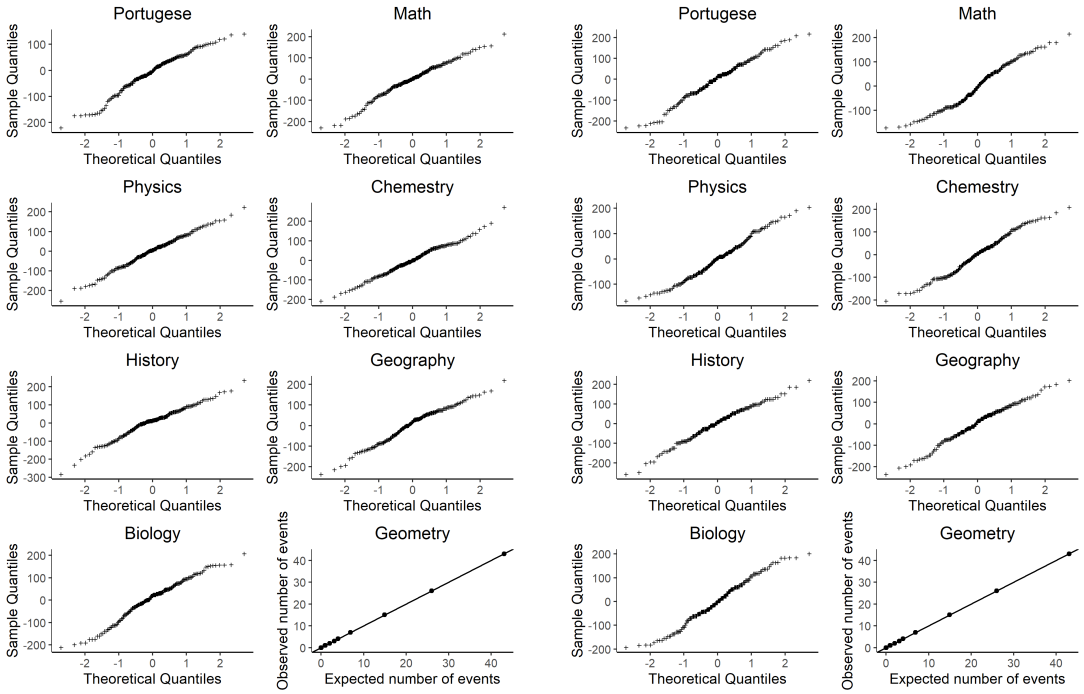
Pinheiro, H. P., Sen, P. K., Pinheiro, A. & Kiihl, S. F. (2020), 'A nonparametric approach to assess undergraduate performance', *Statistica Neerlandica* **74**, 538–588.

Whittaker, J. (1990), *Graphical models in applied multivariate analysis*, Chichester New York et al: John Wiley & Sons.

IV.A Some Model Control

We briefly discuss below the validity of the models used. The residual analyses in the marginal Gaussian mixed models, representing the the responses related to the seven admission exams, show that there is no indication of serious lack of fit, see Figure IV.2.

Comparing the observed and the expected number of students that passed the course of geometry at each time for each combination of age and gender allowed us to conclude that there is no evidence of lack of global adjustment of the survival model. More precisely, the predicted number of events at time t ($t = 1, \ldots, T$), denoted $n(t)$, is calculated as the number of individuals "at risk" (the number of students that



(a) No bonus (b) Bonus

Figure IV.2: Normal QQ-plot of the responses related to the seven admission exams and a scatter plot of the observed number of events versus expected number of events for each time and combination of age and gender.

have not passed the course yet and are still studying the course) times the average estimated hazard at time t. More precisely,

$$n(t) = |R_t| \sum_{i \in R_t} \hat{\lambda}_t \exp(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}) \exp(\hat{u}_{e(i)}) \exp(\hat{v}_i),$$

where $R_t = \{i \in \{1, \ldots, n\} : t_i \geq t\}$ denote the set of all individuals at risk at time t. Here t_i denotes the observed time the student passes the course or is censored, and $|R_t|$ the number of individuals in R_t. The results can be found in Figure IV.2.

IV.B Detailed Representation of the Graphical Models Involving the Random Components and the Response variables

For the reader acquaint with the theory of graphical models (see Whittaker, 1990), the extended separation principle can be formulated in general by defining an directed acyclic graph (DAG, *i.e.,* a graph formed by vertices and directed edges represented by arrows obtained by eliminating the symmetry property in the set of edges $\mathcal{E}$). Using basic properties of the generalised linear mixed models of the type discussed here and the factorisation of the joint densities of the distributions of the individual random component, it is possible to show that the interdependence of the the observable responses and the random components related to the individuals can be represented by an acyclic graph, where there is an arrow from each random component pointing to the random variables representing the corresponding observable responses. Additionally, the graphical representation referred above contains an undirected edge connecting the vertices that are not conditionally independent in the graph representing the individual random components (see Pelck & Labouriau, 2021b for the detailed construction). Noting that this acyclic graph satisfies the Wermuth condition (see Whittaker, 1990, page 75), which implies that the moral graph obtained, in this case, by making all the edges undirected, satisfies the separation principle (see Whittaker, 1990, theorem 3.5.2 on page 76). This construction yields the *Induced separation principle* which states that *"two sets of observable responses, say $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}$, are conditionally independent given a set $\mathcal{S}$ of individual random components, provided $\mathcal{S}$ separates the sets of individual random components $\mathcal{A}$ and $\mathcal{B}$ corresponding to the sets of observable responses $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}$"*.

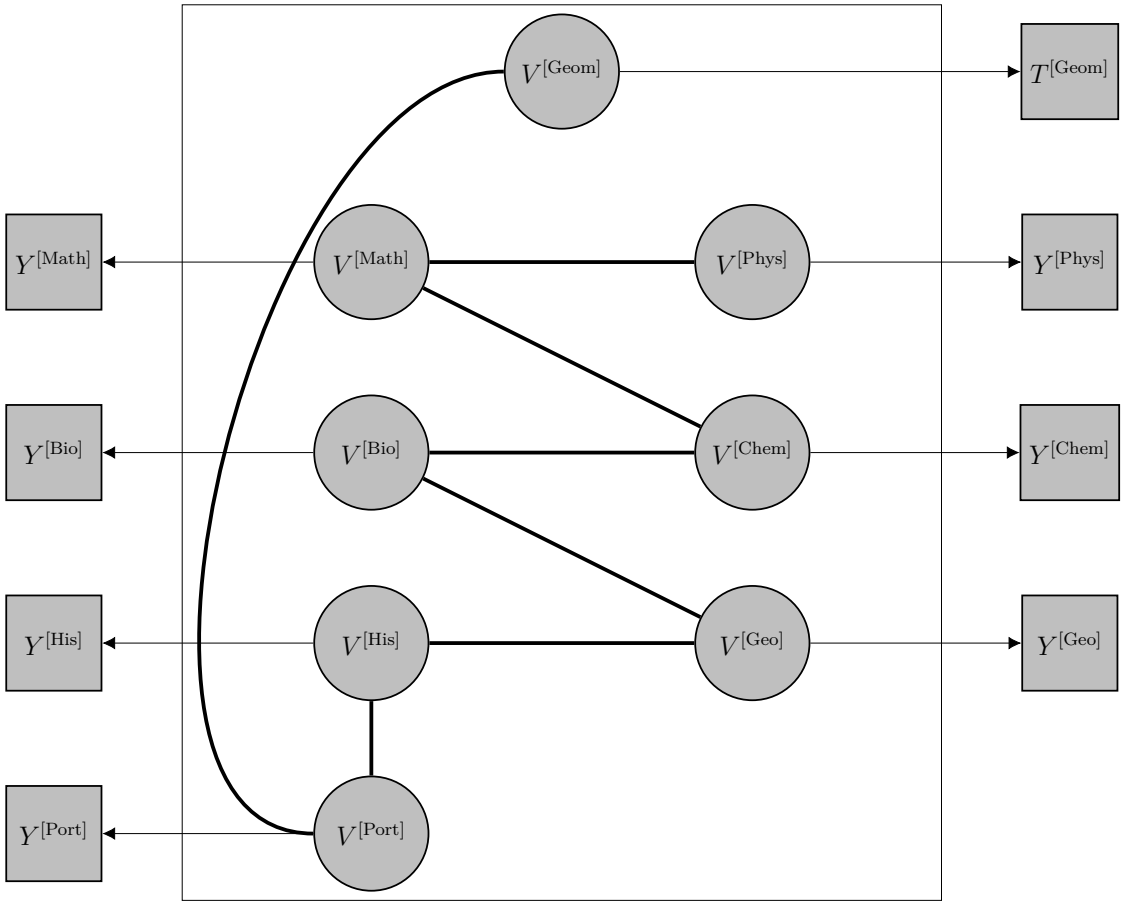


Figure IV.3: Independence graph (central rectangle) representing the estimated graphical model describing the covariance structure of the individual random components $V_i^{[1]}, \ldots, V_i^{[8]}$ for an arbitrarily chosen individual ($i = 1, \ldots, n$, suppressing the index i in the graph), for the students who did not receive bonus.

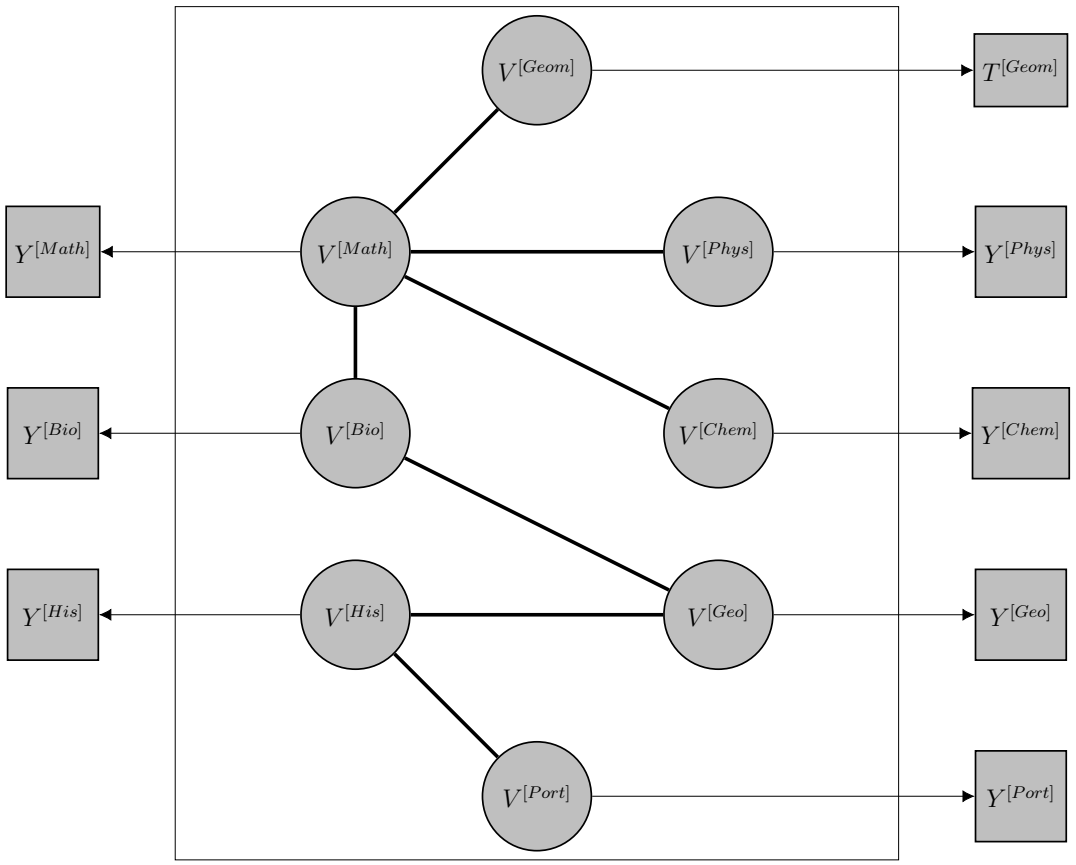


Figure IV.4: Independence graph (central rectangle) representing the estimated graphical model describing the covariance structure of the individual random components $V_i^{[1]}, \ldots, V_i^{[8]}$ for an arbitrarily chosen individual ($i = 1, \ldots, n$, suppressing the index i in the graph), for the students who received bonus.

Paper V

A Multivariate Survival Model for Studying Time to Emergence of Different Species of Weed

Jeanett S. Pelck
Aarhus University

Rodrigo Labouriau
Aarhus University

Abstract. This paper presents an analysis of the times to emergence of different species of weeds using a multivariate frailty piecewise constant Cox proportional model. In the multivariate model, we assume a joint Gaussian distribution on the frailty terms with expectation zero. We show how one can analyse the latent covariance structure of the frailty terms by inferring an undirected graphical model based on predicted values of the frailties. Moreover, we study how the dependence structure of the frailties influences the dependency structure between the emergence times of the several species. Furthermore, we show how we can estimate conditional expectations of the times to emergence given that the emergence is observed during the experiment period.

V.1 Introduction

We present a multivariate survival model for analysing the times to emergence of different species of weed, which takes into consideration that propagules emerging at the same physical location in the field might be correlated. This model is illustrated using part of a data from a field experiment performed at the Department of Agroecology in Flakkebjerg, Denmark, obtained as a part of a project of studying the cumulative emergence dynamics of annual weeds (see Scherner et al. (2017) for details).

The formulated model describes the time to emergence of each propagule known to be present in the soil for a given species using a frailty piecewise constant cox proportional model. A multivariate model is constructed by assuming a joint multivariate Gaussian distribution on the frailty terms. Since we do not know the number of propagules present in the soil, we only model the time to emergence for each propagule in the soil given that it emerges during the observation period.

The likelihood function of the formulated model coincides with the likelihood function of a multivariate generalised linear mixed model inferred on a specially

constructed dataset representing a related counting process. Therefore, we can use the methodology described in Pelck & Labouriau (2021a) and Pelck & Labouriau (2021b) to analyse the latent covariance structure between the random components and how this influences the dependence structure between the responses, that is, the times to emergence of the different species.

We infer an undirected graphical model based on predictions of the frailties to study the covariance structure in the multivariate Gaussian distribution. This representation can be used to study underlying unknown factors affecting the emergence times for the different species, and to eliminate spurious correlation (*i.e.,* the correlation between two variables is only present because of the presence of a third variable correlated with both variables). Combining the results of the undirected graphical model (UG) with the model assumptions, we construct a combination of an UG and a directed acyclic graph (DAG) to study how the dependence structure among the frailties affect the responses (Whittaker (1990) and Lauritzen (1996)).

In the analysis in this paper, we focus on the analysis of the latent covariance structure of the random components. However, we also show how we can estimate conditional expectations of the time to emergence given that the propagule in question emerges during the experiment period. These conditional expectations are estimated using the inferred conditional hazard function that is estimated according to the model.

The paper is structure as follows. In Section V.2, we give a description of the data for which we formulate a multivariate model in Section V.3. In Section V.4, we give a short introduction to the theory of graphical models and describe how we can draw non-trivial conclusions using a graphical representation of the latent covariance structure under the introduced model. We will see, that knowing the frailties included in the marginal survival model for one of the species renders all other responses (emergence times of the different species) independent. A more comprehensive description of the theory of graphical models can be found in Lauritzen (1996) and Whittaker (1990). Section V.5 presents and discusses the results of the analysis. The appendix includes some details on the counting process representing the survival model, the coincidence of the likelihood function with the likelihood of a multivariate generalised linear mixed model applied to a specially constructed dataset, some model controls and some details regarding approximation of the covariance of the estimated conditional expectations.

V.2 Data Description

We use a partial data of the project of studying the cumulative emergence dynamics of annual weeds Scherner et al. (2017) for illustrating how the techniques in Pelck & Labouriau (2021a) and Pelck & Labouriau (2021b) can be applied to model the time to emergence using a multivariate frailty cox proportional model. The data contains counts of emergence of different weed species in the agricultural year 2015 in a cultivated field with winter wheat. We study six species of weed which is given in Table V.1. The two species *Apera spica-venti* and *Vulpia myuros* were

registered together and counted together as one species. This was because they had demonstrated very similar germination in the same experimental area and because it is difficult to differentiate them at an early stage of development (Scherner et al. 2017).

Species	Short form
Apera spica-venti+ Vulpia myuros	AV
Poa annua	PA
Tripleurospermum maritimum	TM
Veronica persica Poir.	VS
Viola arvensis	VA
Papaver rhoeas L.	PR

Table V.1: The different species analysed.

The experimental design used was a split-plot with repetitions. The experiment consisted of four plots, and each plot was split into three sub-plots. Three types of tillage were randomly allocated to the three sub-plots of each plot. The three tillage used were: direct drilling(D), mouldboard plowing(P) to 20 cm soil depth and pre-sowing tine tillage(HW) to 8–10 cm soil depth (Scherner et al. 2017). In each sub-plot, eight physical locations were marked using metal rings. Those locations were kept fixed during the entire experiment and will be referred below as the rings.

The number of emerged plants of each species were registered at each of a range of observation days, not equally spaced over the experiment's period. After registering all plants were removed. The observation days will be indexed as the number of days after sowing.

V.3 Multivariate Model

We consider a multivariate model that takes into account that their might exist some unknown factors affecting the physical locations in the field differently which implies that counts might be correlated within the same physical location in the field. However, we assume that counts from different locations are independent. We analyse the data using a multivariate survival model that models the time to emergence of each propagule known to be present in the soil. That is, we only model the propagules that actually emerge during the observed period. At the end of the experiment only few propagules emerged, and thus, at some observational days only few or none species were represented (consequently not all rings were represented). Therefore, we analyse a shorter period of time and treat these observations as right censored.

Let $T_i^{(j)}$ be a random variable representing the time to emergence of the i^{th} propagule of the j^{th} species for $j = 1, \ldots, 6$ and $i = 1, \ldots, n_j$. We will use the notation that $b_i \in \{1, \ldots, 4\}$ denotes the block and $t_i \in \{1, 2, 3\}$ the tillage of the subplot that the i^{th} propagule belongs to for each species.

For each $j = 1, \ldots, 6$, we let $\boldsymbol{U}^{(j)} = (U_1^{(j)} \ldots, U_q^{(j)})^T$ be a random vector corresponding to the physical locations in the field affecting the j^{th} response ($q = 32$). We assume that $\boldsymbol{U}_{[l]} = (U_l^{(1)} \ldots, U_l^{(6)})^T$ is multivariate Gaussian distributed with expectation zero and covariance matrix given by $\boldsymbol{\Sigma}$ for $l = 1, \ldots, q$. Moreover, we assume that $\boldsymbol{U}_{[1]}, \ldots, \boldsymbol{U}_{[q]}$ are independent. That is, different physical locations in the field are independent for all responses, and the covariance structure within the same location is given by $\boldsymbol{\Sigma}$.

We assume that $T_i^{(j)}$ and $T_{i'}^{(j')}$ are conditionally independent given $\boldsymbol{U}^{(j)}$ and $\boldsymbol{U}^{(j')}$, for $j \neq j' \in \{1 \ldots 6\}$, $i' = 1, \ldots, n_{j'}$ and $i = 1, \ldots, n_j$. Moreover, we assume that $T_i^{(j)}$ and $T_{i'}^{(j)}$ are conditional independent given $\boldsymbol{U}^{(j)}$ for $i, i' = 1, \ldots, n_j$ such that $i \neq i'$ and $j = 1, \ldots, d$. We formulate the conditional distribution of $T_i^{(j)}$ given $\boldsymbol{U}^{(j)}$ in terms of the conditional hazard function given by

$$\begin{aligned}\lambda_i^{(j)}(t|\boldsymbol{U}^{(j)} = \boldsymbol{u}) &\stackrel{\text{def}}{=} \lim_{\Delta \to 0_+} \frac{P[t \leq T_i^{(j)} < t + \Delta | \boldsymbol{U}^{(j)} = \boldsymbol{u},\, T_i^{(j)} \geq t]}{\Delta} \\ &= \tilde{\lambda}_{b_i t_i}^{(j)}(\tau_k) \exp(\boldsymbol{u}^T \boldsymbol{z}_i^{(j)}) \quad \forall\, t \in (\tau_{k-1}^j, \tau_k^j] \text{ and } k = 1, \ldots, K, \quad \text{(V.1)}\end{aligned}$$

where $\boldsymbol{z}_i^{(j)}$ is the allocation vector giving the observed location of the i^{th} plant of species j, and $\tilde{\lambda}_{b_i t_i}^{(j)}(\tau_k)$ denoting a piecewise constant stratified baseline function, taking the same form for all plants in the same block that are exposed to the same tillage for the j^{th} species. In Appendix V.A.1, we describe how the conditional hazard described above enters the intensity process of a counting process describing the survival model defined in this paper. We let $0 = \tau_0 < \tau_1 < \ldots < \tau_K < \infty$ denote the observation days (as the number of days after sowing) and assume that the hazard function is constant on each interval $(\tau_{i-1}, \tau_i]$ for $i = 1, \ldots, K$.

The likelihood function of model defined above coincides with the likelihood function for a generalised linear mixed model defined with a Poisson distribution and a logarithmic link function applied to a specially constructed data representing the risk set of the related counting process, see Appendix V.A.2 for details. The model was inferred using the inference method described in Pelck & Labouriau (2021a).

V.4 Graphical Models

This section gives a short introduction to graphical models, for a more comprehensive description see Whittaker (1990) and Lauritzen (1996).

We define a graph $G = (V, E)$ as a mathematical object containing a set of vertices, V, and a set of edges, E. The set of vertices consists of random vectors, whereas the set of edges contains edges between the vertices indicating the dependency structure between the vectors.

We distinguish between two types of graphs and a combination of both. In an undirected graph (UG), two vertices are connected with an edge if, and only if, they are not conditionally independent given the remaining vertices in V. A directed acyclic graph (DAG) consist of directed edges with an arrow pointing to one of

the vertices indicating which vector that carries information on the other. An edge between two vertices might have an arrow in each end, implying that the vertices carry information on each other corresponding to an undirected edge. A directed edge between two vertices indicates that those two vertices are not conditional independent given a subset of the remaining vertices containing all vectors that carries information, directly or indirectly, on the vertices in question.

A DAG processes the same independence interpretation as there associated moral graph. This is a graph constructed from the originally graph by connecting, with an undirected edge, all vertices that have a directed edge towards a given vertex, and by replacing existing directed edges by undirected edge.

We say that there is a path between two vertices, say V_1 and V_2, if there exist a sequence of pairs of vertices in $\mathcal{E}$ such that V_1 and V_2 belong to at least one vertex of the sequence. According to the theory of graphical models, a set of vertices, say $\mathcal{S}$, separates two sets of vertices $\mathcal{A}$ and $\mathcal{B}$ in the graph, if, and only if, each path connecting an element of $\mathcal{A}$ to an element of $\mathcal{B}$ contains at least one element of $\mathcal{S}$. A key result in the theory of graphical models is that if a set of vertices, $\mathcal{S}$, separates two disjoint subsets of vertices $\mathcal{A}$ and $\mathcal{B}$, then all the variables in $\mathcal{A}$ are independent of the variables in $\mathcal{B}$ given the variables in $\mathcal{S}$. This result is called the *separation principle* or *global Markov property* for undirected graphical models (Lauritzen 1996, page 32). An undirected graph and moral graph satisfies the separation principle.

In Pelck & Labouriau (2021b) an extension of the separation principle is formulated called the *induced separation principle*. It states that if we denote $\mathcal{A}'$ and $\mathcal{B}'$ as the sets of responses corresponding to two sets of random components, $\mathcal{A}$ and $\mathcal{B}$, and let $\mathcal{S}$ denote a set of random components that separates $\mathcal{A}$ and $\mathcal{B}$ in the graph as described above, then $\mathcal{A}'$ and $\mathcal{B}'$ are conditionally independent given $\mathcal{S}$.

V.5 Results and Discussion

In this section we present and discuss the results of the analysis. We will focus on the graphical latent covariance structure but we will also estimate the expectation of the time to emergence for each propagule. The estimated expectations of the times to emergence of the propagules allow us to draw conclusions regarding the three types of tillage. This differs from the analysis of the latent covariance structure which makes it possible to examine how local characteristics represented by the observation rings affect the times to emergence of the species differently (some species might be more sensitive to some local characteristics than others). Model control for the estimated models can be found in Appendix V.A.3.

The graphical model in Figure V.1 shows the estimated dependency structure for the vectors of random components $\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(6)}$ and the responses $T_i^{(j)}$ for $i = 1, \ldots, n_j$ and $j = 1, \ldots, 6$. The structure inside the square represents the dependency structure between the random components (with estimated conditional correlations given the other random components), whereas the responses are drawn outside the square using circled vertices. In that way, we are able to study the latent covariance structure of the random components, but also how this affects

the dependency structure between the responses. The undirected graph inside the square can be interpret using the theory of undirected graphical models. However, to interpret the whole graph, we need the theory of DAGs and the induced separation principle.

The conditional correlation between the random components representing the species TM and PA, denoted $\boldsymbol{U}^{(\mathrm{TM})}$ and $\boldsymbol{U}^{(\mathrm{PA})}$, is significant positive given the other random components. This indicates that the intrinsic local environmental characteristics between those species are not the same but not antagonist either. The fact that the covariance is significant positive suggests that there are common intrinsic environmental characteristics affecting both species. The conditional correlations between $\boldsymbol{U}^{(\mathrm{PA})}$ and $\boldsymbol{U}^{(\mathrm{VA})}$ as well as between $\boldsymbol{U}^{(\mathrm{PA})}$ and $\boldsymbol{U}^{(\mathrm{VS})}$ are significant negative given the other random components. This suggests the existence of antagonist environmental local characteristics affecting the emergence of the two species in question. For example, there might exist environmental characteristics that positively affect the emergence intensity of VA but negatively affect the emergence intensity of PA. Note that it would be expected that these antagonist effects should be contained in the specific local environmental conditions that affect the emergence of PA but not the emergence of TM, otherwise we would have believed that the conditional correlation between VA and TM was significant different from zero.

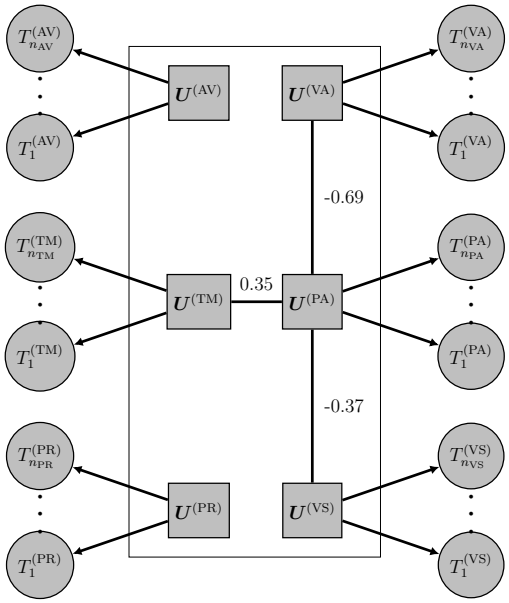


Figure V.1: The estimated graphical model in the multivariate model described in Section V.3. The squared vertices and edges inside the rectangle show an UG for the random components under the model. Estimated conditional correlations were added to the edges. The round vertices represents the responses and here the induced separation principle should be used to interpret the dependence structure.

In order to interpret the effect that the choice of tillage has on the expected times to emergence, we calculate the conditional expectations given that the propagules emerge before a time T. During the inference procedure, we estimate the parameters

entering the conditional hazard. However, using the below formulas, we are able to estimate the conditional expectations and perform approximate asymptotic tests comparing if two conditional expectations are significant different. Moreover, we will be able to estimate approximate asymptotic confidence intervals.

Based on the estimated values of $\tilde{\lambda}_{b_i t_i}^{(j)}(\tau_k)$ ($i = 1, \ldots, n_j$, $j = 1, \ldots, 6$, and $k = 1, \ldots, K$), we estimate the following conditional expectations given that the propagule emerges before the right censoring time T:

$$\begin{aligned}
\mathbb{E}[T_i^{(j)} | T_i^{(j)} \leq T] &= \int_0^T S_i^{(j)}(t) dt \\
&= \sum_{k=1}^{T} (\tau_k - \tau_{k-1}) S_i^{(j)}(\tau_k) \\
&= \sum_{k=1}^{T} (\tau_k - \tau_{k-1}) \exp\Big[- \sum_{l=1}^{k} (\tau_l - \tau_{l-1}) \lambda_i^{(j)}(\tau_l) \Big],
\end{aligned}$$

where

$$\begin{aligned}
\lambda_i^{(j)}(\tau_l) &= \int_{\mathbb{R}^q} \lambda_i^{(j)}(\tau_l | \boldsymbol{U}^{(j)} = \boldsymbol{u}) \varphi(\boldsymbol{u}; \boldsymbol{\Sigma}_{jj} \boldsymbol{I}_q) d\boldsymbol{u} \\
&= \tilde{\lambda}_{b_i t_i}(\tau_l) \int_{\mathbb{R}^q} \exp(\boldsymbol{u}^T \boldsymbol{z}_i^{(j)}) \varphi(\boldsymbol{u}; \boldsymbol{\Sigma}_{jj} \boldsymbol{I}_q) d\boldsymbol{u} \\
&= \tilde{\lambda}_{b_i t_i}(\tau_l) \exp(\tfrac{1}{2} \boldsymbol{\Sigma}_{jj}),
\end{aligned}$$

with $\varphi(\cdot; \boldsymbol{\Sigma}_{jj} \boldsymbol{I}_q)$ denoting the q-dimensional Gaussian density with expectation zero and covariance matrix $\boldsymbol{\Sigma}_{jj} \boldsymbol{I}_q$, where $\boldsymbol{\Sigma}_{jj}$ is the $(j, j)^{th}$ entry in $\boldsymbol{\Sigma}$.

The estimated conditional expectations given that the propagule emerge before day $T = 61$ are presented in Figure V.2. The letters in the figure indicate pairwise comparisons using a Wald test for each comparison based on an approximated covariance matrix for the estimated conditional expectations. This approximation is based on a first order Taylor approximation, and is also used to calculate 95 percent confidence intervals. The calculations of the approximated covariance matrix are given in Appendix V.A.4.

From the estimated conditional expectations, we conclude that there is an interaction between the blocks and treatments implying that there are some unknown factors affecting the emergence in the different blocks which were not accounted for in the model. In the fourth block, we see that the expected times to emergence for the different treatments are very close. Moreover, we observe similar patterns for the species TM and VS.

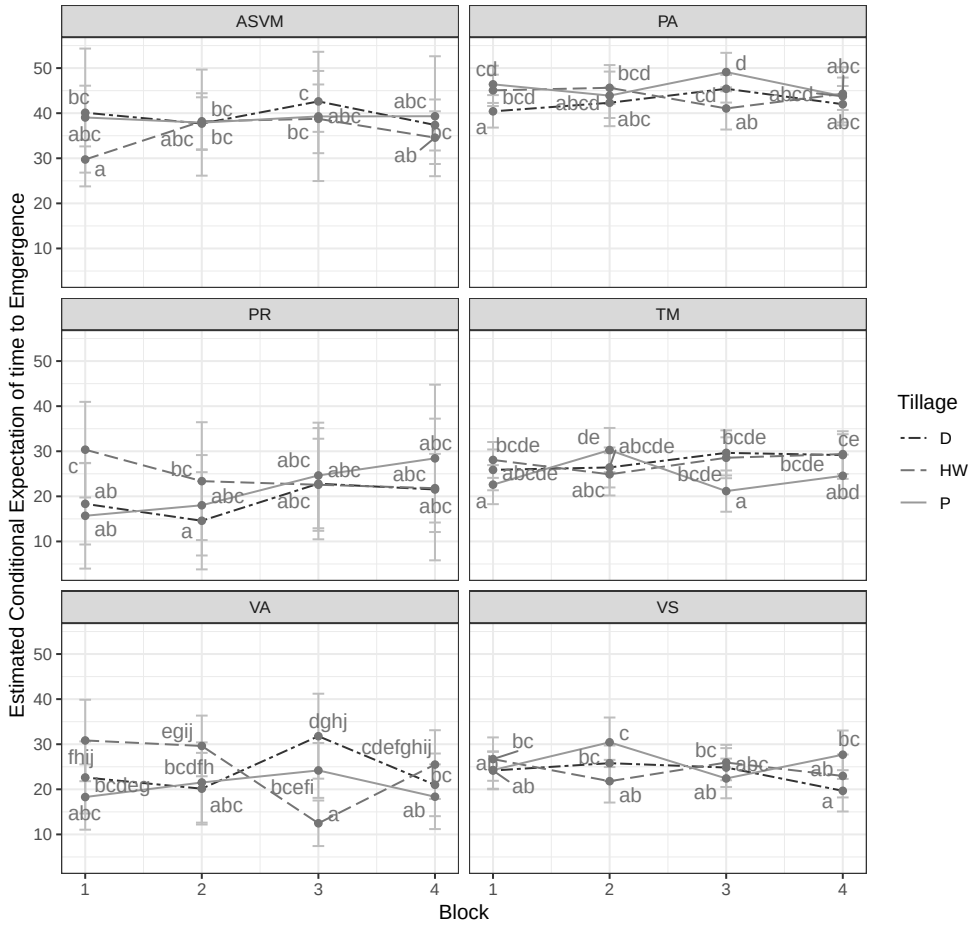


Figure V.2: Plot of the estimated conditional expectations of the times to emergence for each species and each combination of block and tillage given that the time is less or equal to the time of the end of the experiment.

Acknowledgement

The data analysed in this paper was kindly supplied by Dr. Ananda Scherner and the associate professor Bo Melander. The experiment was performed at the Department of Agroecology in Flakkebjerg, Denmark. The authors were partially financed by the Applied Statistics Laboratory (aStatLab) at the Department of Mathematics, Aarhus University.

References

Allison, P. D. (2014), *Event history and survival analysis: Regression for longitudinal event data*, Vol. 46, SAGE publications.

Ebner, B. & Henze, N. (2020), 'Tests for multivariate normality–a critical review with emphasis on weighted l^2-statistics', **29**.

Kalbfleisch, J. D. & Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data*, second edn, Wiley series in probability and statistics, New York.

Lauritzen, S. L. (1996), *Graphical models*, Vol. 17, Clarendon Press.

Maia, R. P., Madsen, P. & Labouriau, R. (2014), 'Multivariate survival mixed models for genetic analysis of longevity traits', *Journal of Applied Statistics* **41**(6), 1286–1306.

Martinussen, T. & Scheike, T. H. (2006), *Dynamic regression models for survival data*, Springer Science & Business Media.

Pelck, J. S. & Labouriau, R. (2021a), Conditional inference for multivariate generalised linear mixed models. arXiv:2107.11765.

Pelck, J. S. & Labouriau, R. (2021b), Multivariate generalised linear mixed models with graphical latent covariance structure. Submitted to arXiv.

Scherner, A., Melander, B., Jensen, P. K., Kudsk, P., Avila, L. A. & Riemens, M. (2017), 'Reducing tillage intensity affects the cumulative emergence dynamics of annual grass weeds in winter cereals', *Weed Research* **57**(5).

Whittaker, J. (1990), *Graphical models in applied multivariate analysis*, Chichester New York et al: John Wiley & Sons.

V.A Appendix

V.A.1 Survival Model Represented through a Counting Process

Here we show how the defined conditional hazards can be viewed as intensities of a counting process. Consider a given propagule, i, of a given species, j. Recall, that we observe $\tilde{T}_i^{(j)}$ which is a non-negative random variable representing either the observed time to emergence or censoring time of the i^{th} propagule of the j^{th} species. Let $D_i^{(j)}$ be an indicator variable taking the value one if the propagule in question emerged before the censoring time and zero otherwise. Define the counting process $\{N_i^{(j)}(t) : t \in \mathbb{R}_+\}$, where $N_i^{(j)}(t) = \mathbf{1}(\tilde{T}_i^{(j)} \leq t,\, D_i^{(j)} = 1)$. This process is equal to zero until emergence of propagule i of species j. If the propagule in question was censored, the process will stay at zero. Furthermore, define the risk process $\{Y_i^{(j)}(t) : t \in \mathbb{R}_+\}$, where the random variable $Y_i^{(j)}(t) = \boldsymbol{I}(t \leq \tilde{T}_i^{(j)})$ takes the value one when the propagule is at risk at time t, and zero otherwise. The counting process and the at risk process are both adapted to the filtration $\{\mathcal{F}_t^{i,(j)} : t \in \mathbb{R}_+\}$, where $\mathcal{F}_t^{i,(j)} = \sigma\{N_i^{(j)}(s), Y_i^{(j)}(s) : 0 \leq s \leq t\}$ is the σ-algebra representing the history of $N_i^{(j)}(\cdot)$ and $Y_i^{(j)}(\cdot)$ for each time up to and including time t.

Let $dN_i^{(j)}(t) = \lim_{\Delta \downarrow 0} \frac{N_i^{(j)}(t+\Delta) - N_i^{(j)}(t)}{\Delta}$. This is equal to one if the propagule in question emerges right after time t and zero otherwise. Conditional on the random component $\boldsymbol{U}^{(j)}$, the intensity of the counting process is given by

$$\begin{aligned}\mathbb{E}[dN_i^{(j)}(t)|U^{(j)} = \boldsymbol{u},\, \mathcal{F}_t^{i,(j)}] &= P[dN_i^{(j)}(t) = 1|U^{(j)} = \boldsymbol{u},\, \mathcal{F}_t^{i,(j)}] \\ &= Y_i^{(j)}(t)\lambda_i^{(j)}(t|\boldsymbol{U}^{(j)} = \boldsymbol{u}),\end{aligned}$$

where $\lambda_i^{(j)}(t|\boldsymbol{U}^{(j)} = \boldsymbol{u})$ is defined in (V.1). For more details see Maia et al. (2014), Martinussen & Scheike (2006) and Allison (2014).

V.A.2 Coincidence of Likelihood Function of Survival Model with the Likelihood Function of a Constructed GLMM

Consider the model described in Section V.3. Let $t_i^{(j)}$ denote the observed value of $T_i^{(j)}$, and define for $k = 1, \ldots, K$, $i = 1, \ldots, n_j$ and $j = 1, \ldots, d$ ($d = 6$), $\Delta_k = \tau_k - \tau_{k-1}$ and $\delta_{ijk} = \mathbf{1}(\tau_{k-1} < t_i^{(j)} \leq \tau_k)$. The contribution to the likelihood function for the propagules emerging in the l^{th} ring is formulated as

$$\begin{aligned}L_l(\boldsymbol{\theta}_l; \boldsymbol{\Sigma}) = \int_{\mathbb{R}^d} \prod_{i \in \mathcal{I}_l} \prod_{j=1}^{d} \prod_{k=1}^{t_i^{(j)}} &[\Delta_k \tilde{\lambda}_{b_i t_i}^{(j)}(\tau_k) \exp(u_l^{(j)})]^{\delta_{ijk}} \\ &\exp(-\Delta_k \tilde{\lambda}_{b_i t_i}^{(j)}(\tau_k) \exp(u_l^{(j)}))\varphi(\boldsymbol{u}_l; \boldsymbol{\Sigma}) d\boldsymbol{u}_l,\end{aligned}$$

where $\mathcal{I}_l$ denotes the set of propagules emerging in the l^{th} ring for $l = 1, \ldots, q$, $\boldsymbol{u}_l^T = (u_l^{(1)}, \ldots, u_l^{(d)})$ and $\boldsymbol{\theta}_l$ is the vector of parameters containing $\tilde{\lambda}_{b_i t_i}^{(j)}(\tau_k)$ for $i \in \mathcal{I}_l$, $k = 1, \ldots, t_i^{(j)}$ and $j = 1, \ldots, d$ (see Kalbfleisch & Prentice (2002) for more details on the likelihood function of the model).

The likelihood function for all the rings can then be formulated as

$$\begin{aligned}L(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q; \boldsymbol{\Sigma}) = \prod_{l=1}^{q} \int_{\mathbb{R}^d} \prod_{i \in \mathcal{I}_l} \prod_{j=1}^{d} \prod_{k=1}^{t_i^{(j)}} &[\Delta_k \tilde{\lambda}_{b_i t_i}^{(j)}(\tau_k) \exp(u_l^{(j)})]^{\delta_{ijk}} \qquad \text{(V.2)} \\ &\exp(-\Delta_k \tilde{\lambda}_{b_i t_i}^{(j)}(\tau_k) \exp(u_l^{(j)}))\varphi(\boldsymbol{u}_l; \boldsymbol{\Sigma}) d\boldsymbol{u}_l, \qquad \text{(V.3)}\end{aligned}$$

which coincides with the likelihood function for δ_{ijk} ($k = 1, \ldots, K$, $i = 1, \ldots, n_j$ and $j = 1, \ldots, d$) under the assumption of a multivariate generalised linear mixed model with a conditional Poisson distribution with conditional expectation

$$\mathbb{E}[\delta_{ijk}|U_{c(i,j)}^{(j)} = u_{c(i,j)}^{(j)}] = \Delta_l \tilde{\lambda}_{b_i t_i}^{(j)}(\tau_k) \exp(u_{c(i,j)}^{(j)}),$$

where $c(i,j)$ denotes the ring of the i^{th} propagule of the j^{th} species with $c(i,j) \in \{1, \ldots, q\}$ for $i = 1, \ldots, n_j$ and $j = 1, \ldots, d$. Therefore, we can perform statistical inference for the model by constructing a pseudo data with the variables δ_{ijk} for $k = 1, \ldots, K$, $i = 1, \ldots, n_j$ and $j = 1, \ldots, d$, and inferring a multivariate generalised linear mixed models, assuming a Poisson distribution with logarithm link function (Maia et al. (2014)).

V.A.3 Model Control

We will check the model assumptions by examining if the number of predicted events are consistent with the observed number of events for each species, for each observation day and for each combination of block and tillage. Moreover, we will check the multivariate normality assumption of the frailties.

We calculate the estimated and predicted number of events using the formulas presented in the supplementary material to Maia et al. (2014). Let $R^{(j)}(t)$ denote the set of propagules "at risk" of emerging at time t for the j^{th} species, that is, propagules that either emerged or were censored after time t. Let $m_j(t) = |R^{(j)}(t)|$ be the number of elements in $R^{(j)}(t)$. Moreover, let $\alpha_{btk}^{(j)}$ denote the number of predicted propagules of the j^{th} species ($j = 1, \ldots, 6$), emerging in the b^{th} block, exposed to tillage t at the k^{th} observation day under the model. Then $\alpha_{btk}^{(j)}$ is given by

$$\alpha_{btk}^{(j)} = R^{(j)}(t)|N_{btk}|^{-1} \sum_{l \in N_{tbk}} \hat{\hat{\lambda}}_{b(l)t(l)}^{(j)}(\tau_k) \exp(\hat{\boldsymbol{u}}^T \boldsymbol{z}_l^{(j)}),$$

with $N_{btk}^{(j)} = \{l : b(l) = b, t(l) = t, l \in R^{(j)}(\tau_k)\}$ and $\hat{\cdot}$ denotes the estimated value. Figure V.3 shows the predicted number of events against the observed for different values of b ($b = 1, \ldots, 4$), t ($t = 1, 2, 3$), k ($k = 1, \ldots, K$) and j ($j = 1, \ldots, 6$). We see that the number of predicted events and the number of observed events are very close but not equal. Therefore we found no evidence against this model.

We tested the multivariate Gaussian assumption of the predicted values of the frailties. First we considered marginal QQ-plots and used Shapiro-Wilk's test to test for deviations from marginal normality. We found no evidence against marginal normality. Moreover we used the Henze-Jimenes-Gamero-Meintanis test of multivariate normality suggested in Ebner & Henze (2020) which do not lead to any evidence against the multivariate normality.

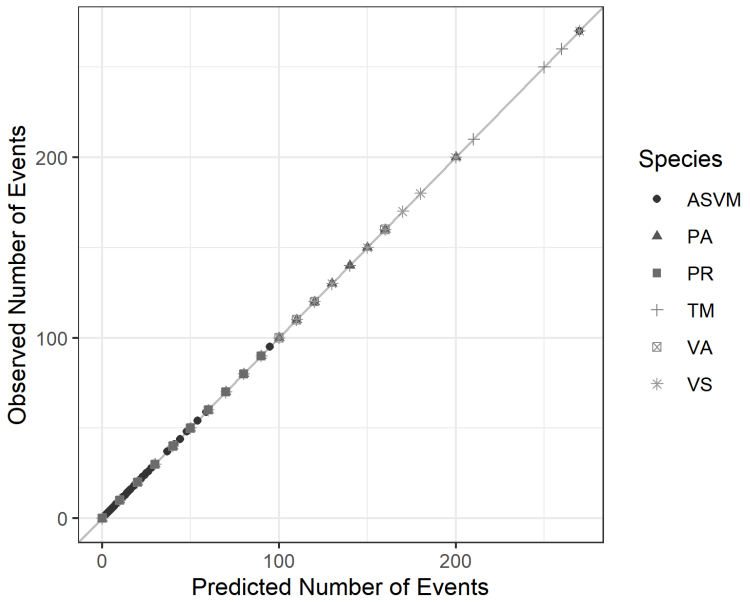


Figure V.3: The predicted number of events under the estimated model against the observed.

V.A.4 Covariance of Conditional Expectations

Here, we will show how we can approximate the covariance matrix of the calculated conditional expectations in Section V.5. Let $\hat{\tilde{\boldsymbol{\lambda}}}_{b_i t_i} = (\hat{\tilde{\lambda}}_{b_i t_i}(\tau_1), \ldots, \hat{\tilde{\lambda}}_{b_i t_i}(\tau_T))^T$ denote the estimated values of $\tilde{\boldsymbol{\lambda}}_{b_i t_i} = (\tilde{\lambda}_{b_i t_i}(\tau_1), \ldots, \tilde{\lambda}_{b_i t_i}(\tau_T))^T$. Define the estimated conditional expectation by

$$\begin{aligned} h(\hat{\tilde{\boldsymbol{\lambda}}}_{b_i t_i}) &\overset{\text{def}}{=} \hat{\mathbb{E}}[T_i^{(j)} | T_i^{(j)} \leq T] \\ &= \sum_{k=1}^{T} (\tau_k - \tau_{k-1}) \exp\Big[- \sum_{l=1}^{k} (\tau_l - \tau_{l-1}) \hat{\tilde{\lambda}}_{b_i t_i}(\tau_l) \exp(\tfrac{1}{2}\boldsymbol{\Sigma}_{jj}) \Big]. \end{aligned}$$

We will consider a first order Taylor approximation of h around $\tilde{\boldsymbol{\lambda}}_{b_i t_i}$:

$$h(\tilde{\boldsymbol{\lambda}}_{b_i t_i}) \approx h(\tilde{\boldsymbol{\lambda}}_{b_i t_i}) + \nabla h(\hat{\tilde{\boldsymbol{\lambda}}}_{b_i t_i})^T (\hat{\tilde{\boldsymbol{\lambda}}}_{b_i t_i} - \tilde{\boldsymbol{\lambda}}_{b_i t_i}),$$

with

$$\nabla h(\hat{\tilde{\boldsymbol{\lambda}}}_{b_i t_i}) = -h(\hat{\tilde{\boldsymbol{\lambda}}}_{b_i t_i}) \exp(\tfrac{1}{2}\boldsymbol{\Sigma}_{jj}) \hat{\tilde{\boldsymbol{\lambda}}}^{\boldsymbol{\tau}}_{b_i t_i},$$

where $\hat{\tilde{\boldsymbol{\lambda}}}^{\boldsymbol{\tau}}_{b_i t_i} = \Big((\tau_1 - \tau_0)\hat{\tilde{\lambda}}_{b_i t_i}(\tau_1), (\tau_2 - \tau_1)\hat{\tilde{\lambda}}_{b_i t_i}(\tau_2), \ldots, (\tau_T - \tau_{T-1})\hat{\tilde{\lambda}}_{b_i t_i}(\tau_T)\Big)^T$. Then we can approximate the covariance by

$$\begin{aligned}
&\mathrm{Cov}\Big(\hat{\mathbb{E}}[T_i^{(j)}|T_i^{(j)} \leq T],\, \hat{\mathbb{E}}[T_m^{(j)}|T_m^{(j)} \leq T]\Big) \\
&\qquad \approx \mathrm{Cov}\Big(\nabla h(\hat{\tilde{\boldsymbol{\lambda}}}_{b_i t_i})^T(\hat{\tilde{\boldsymbol{\lambda}}}_{b_i t_i} - \tilde{\boldsymbol{\lambda}}_{b_i t_i}), \nabla h(\hat{\tilde{\boldsymbol{\lambda}}}_{b_m t_m})^T(\hat{\tilde{\boldsymbol{\lambda}}}_{b_m t_m} - \tilde{\boldsymbol{\lambda}}_{b_m t_m})\Big) \\
&\qquad = \nabla h(\hat{\tilde{\boldsymbol{\lambda}}}_{b_i t_i})^T \mathrm{Cov}(\hat{\tilde{\boldsymbol{\lambda}}}_{b_i t_i}, \hat{\tilde{\boldsymbol{\lambda}}}_{b_m t_m}) \nabla h(\hat{\tilde{\boldsymbol{\lambda}}}_{b_i t_i}).
\end{aligned}$$

Paper VI

Multivariate Methods for Detection of Rubbery Rot in Storage Apples by Monitoring Volatile Organic Compounds:

An Example of Multivariate Generalised Mixed Models

Jeanett S. Pelck
Aarhus University

Hinrich H.F. Holthusen
Esteburg Fruit Research and Advisory Centre
Aarhus University

Merete Edelenbos
Aarhus University, Denmark

Alexandru Luca
Aarhus University, Denmark

Rodrigo Labouriau
Aarhus University

Abstract. This article is a case study illustrating the use of a multivariate statistical method for screening potential chemical markers for early detection of post-harvest disease in storage fruit. We simultaneously measure a range of volatile organic compounds (VOCs) and two measures of severity of disease infection in apples under storage: the number of apples presenting visible symptoms and the lesion area. We use multivariate generalised linear mixed models (MGLMM) for studying association patterns of those simultaneously observed responses via the covariance structure of random components. Remarkably, those MGLMMs can be used to represent patterns of association between quantities of different statistical nature. In the particular example considered in this paper, there are positive responses (concentrations of VOC, Gamma distribution based models), positive responses possibly containing observations with zero values (lesion area, Compound Poisson distribution based models) and binomially distributed responses (proportion of apples presenting infection symptoms). We represent patterns of association inferred with the MGLMMs using graphical models (a network represented by a graph), which allow us to eliminate spurious associations due to a cascade of indirect correlations between the responses.

VI.1 Introduction

Rubbery rot is a post-harvest disease in apples caused by the fungus *Phacidiopycnis washingtonensis*, leading to significant storage losses in commercial production (Ali et al. 2018). Therefore, it is interesting to find predictors of rubbery rots onset at early stages of the infection development under fruit storage. To this purpose, a comprehensive study involving the emission of a range of specially chosen volatile organic compounds (VOCs) under apple storage conditions was performed by Holthusen et al. (2021a). In this study, experimentally induced rubbery rots infections were monitored and contrasted with the concentration of 14 VOCs along the development of the disease, aiming to find chemical predictors for rubbery rot. This article exposes details of some non-standard statistical tools used in Holthusen et al. (2021a).

The experiment we will model can be shortly described in the following way. Ten glass jars (below referred as glasses), containing nine inoculated apples were observed at three fixed observation times (6, 12 and 18 weeks post-inoculation). The following quantities were determined at each observation time: the number of apples presenting visible symptoms, the area of lesions caused by the fungal infection, and the air concentration of 14 VOCs. The details of the experiment setup and the choice of the VOCs are exposed in Holthusen et al. (2021a), see also Holthusen et al. (2021b) and Holthusen & Weber (2021).

The proper statistical modelling of the complex of experiments referred to above presents several challenges. Indeed, the simultaneously observed responses are of different statistical nature. For example, while the number of apples showing visible symptoms (used to monitor the disease development) is naturally binomially distributed, the VOC concentrations follow continuous positive non-Gaussian distributions with high skewness. Furthermore, the area of lesions (characterising disease severity) presents many zero values (absence of infection) but otherwise follows a continuously skewed distributed and therefore is not adequately described by purely continuous distributions. Thus, the first challenge we encountered was to develop methods for establishing associations between these responses of different nature. We propose to solve this problem by using suitably constructed multivariate generalised linear mixed models (MGLMMs) simultaneously describing the responses referred to above. In this way, we will model the concentrations of the VOCs using Gamma distributions, representing positive valued responses with different degrees of skewness. Moreover, the lesion area will be modelled using Gamma compound Poisson distributions with positive mass at zero and otherwise continuous with variable degrees of skewness. All these families of distributions are particular cases of dispersion models, which are the families of distributions that form the basis of generalised linear mixed models (GLMMs).

The MGLMM we propose to use is composed of marginal GLMMs describing each of the responses studied. Each of those GLMMs will contain a random component representing the basic experimental unit (the glass), which we use to model the covariance structure of the different responses. We use the tools of graphical models to describe this covariance structure in a suitable compact form, which will allow us to draw valid general conclusions on the association between the responses, even

though they are of different statistical nature. For instance, we will eliminate spurious correlations between the responses, i.e., correlations between two responses that can be explained by a cascade of correlations between those responses and the other responses in play. In this way, we will identify a minimal group of VOCs sufficient to predict the rubbery rots onset, avoiding redundancy and the pitfalls of multicollinearity.

The MGLMMs we construct allow for incorporating corrections for determining factors know to have a strong influence in the responses (e.g., the observation week) and temporal correlation due to repeated observations at the same experimental unit.

This paper is structure as follow. Section VI.2 introduces the marginal GLMMs for each of the responses considered. Those models are used to construct a MGLMM in Section VI.3. The details of the construction are given in Section VI.3.1, and the graphical model representing the covariance structure of the random components is discussed in section VI.3.2. Section VI.4 briefly discusses the results obtained.

VI.2 Models for Several Responses with Different Nature

We introduce below a range of GLMMs describing each of the observed responses. Those models contain a random component representing the glass (which is viewed as the basic experimental unit) and a fixed effect representing the observation time (week). We used a GLMM defined with the binomial distribution and logistic link function for describing the number of apples presenting visible symptoms. The concentrations of VOCs were modelled using GLMMs defined with a Gamma distribution and the logarithmic link function. Finally, we used a GLMM defined with the family of Gamma compound Poisson distributions and a logarithmic link to describe the lesion area. We give the full details of those models below using a notation suitable for defining the multivariate model for simultaneously describing the 16 responses in play.

VI.2.1 Concentrations of Volatiles Organic Compounds - Positive Responses

We describe below the GLMM used for modelling the concentration of each of the 14 VOCs. We label those VOCs by the index j ($j = 1, \ldots, 14$), which is kept fixed along this section (referring to a choice of one of the VOCs). Denote by $X_{tg}^{[j]}$ the random variable representing the concentration of the value of the j^{th} VOC measured at the t^{th} week ($t = 6, 12, 18$) in the g^{th} glass ($g = 1, \ldots, 10$). According the GLMM we are defining, there exist 10 unobservable random variables, denoted by $U_1^{[j]}, \ldots, U_{10}^{[j]}$, such that for $t = 6, 12, 18$ and $g = 1, \ldots, 10$ the response $X_{tg}^{[j]}$ is conditional Gamma

distributed given $U_g^{[j]}$ with conditional expectation given by

$$\log(\mathbb{E}[X_{tg}^{[j]}|U_g^{[j]} = u]) = \theta_t^{[j]} + u \quad \text{for all } u \in \mathbb{R}.$$

Moreover, according to the GLMM the responses, $X_{gt}^{[j]}$ for $t = 6, 12, 18$ and $g = 1, \ldots, 10$, are conditionally independent given $U_1^{[j]}, \ldots, U_{10}^{[j]}$. The specification of the GLMM is completed by stating that $U_1^{[j]}, \ldots, U_{10}^{[j]}$ are independent and identically normally distributed with expectation 0. Here $\theta_6^{[j]}, \theta_{12}^{[j]}$ and $\theta_{18}^{[j]}$ are fixed effects describing the variation of the concentration of the j^{th} VOC at different observation times.

VI.2.2 Number of apples Presenting Symptoms - Binomial Counts

Let Y_{tg} be a random variable representing the number of apples presenting symptoms in the g^{th} glass ($g = 1, \ldots, 10$), at the at the t^{th} week ($t = 6, 12, 18$) out of the nine apples contained in each glass. We assume that there exist 10 unobservable random variables, denoted by $V_1, \ldots, V_{10}$ such that, for $t = 6, 12, 18$ and $g = 1, \ldots, 10$, Y_{tg} is conditionally binomially distributed given V_g with size 9 and conditional expectation given by

$$\text{logit}(\mathbb{E}[Y_{tg}|V_g = v]) = \alpha_t + v \quad \text{for all } v \in \mathbb{R}\,.$$

According to the model the responses, Y_{gt} for $t = 6, 12, 18$ and $g = 1, \ldots, 10$, are conditionally independent given $V_1, \ldots, V_{10}$. Moreover, the random components $V_1, \ldots, V_{10}$ are assumed to be independent and identically normally distributed with expectation zero.

VI.2.3 Lesion Area of Infection - Positive Responses with Zero Values

Denote by Z_{tg} the random variable describing the observed lesion area in the g^{th} glass ($g = 1, \ldots, 10$), at the t^{th} week ($t = 6, 12, 18$). We assume that there exist 10 independent and normally distributed random variables with expectation zero, denoted by $W_1, \ldots, W_{10}$, such that, for $t = 6, 12, 18$ and $g = 1, \ldots, 10$, the response Z_{tg} is distributed according to a Gamma-compound Poisson distribution with conditional expectation given by

$$\log(\mathbb{E}[Z_{tg}|W_g = w]) = \beta_t + w, \quad \text{for all } w \in \mathbb{R}.$$

Note that the Gamma-compound Poisson family of distributions is an exponential dispersion model (see Jørgensen,1987); therefore, the model we are defining is a genuine GLMM. Furthermore, the distributions in the Gamma-compound Poisson family have the peculiarity of attributing positive probability to the value zero

and otherwise being a continuous distribution taking positive values, making them suitable for describing the lesion area.

The Gamma-compound Poisson family has been known for a long time (see Tweedie, 1984; Jørgensen,1987, and Cordeiro et al.,2021), however, these distributions are not routinely used in applications of generalised linear models (or GLMMs). Therefore, we shortly describe the inference procedure we used (for a detailed study see Labouriau, 2021). The Gamma-compound Poisson family is characterised by having a power variance function (*i.e.,* a function expressing the variance as a function of the expectation) of the form $V(\mu) = k\mu^p$ for p in the open interval $(1, 2)$ (see Cordeiro et al.,2021), where different power indices p yield different Gamma-compound Poisson families. The probability of observing a zero value can be calculated as a function of the power index p and the expectation (see Jørgensen,1987). In the analysis described above, we calculated the probability of each observation taking the value zero using a grid of values of the power index p. We estimated the expected number of zeroes for each observation week by summing the probability of observing a zero for each observation made in this week. This process was repeated for each value of the power index p in a grid of possible values. We used in the analysis the value of the power index p that minimised the Euclidean distance between the vector containing the observed proportions of zeroes for each week and the vector of expected number of zeroes for each week.

VI.3 Multivariate Simultaneous Models for Responses of Different Statistical Nature

VI.3.1 A Multivariate Construction

We use now the marginal GLMMs described above to formulate a MGLMM that simultaneously describe the 16 responses observed in this experiment. The idea explored here is to combine the marginal models into a MGLMM by constructing a 16-dimensional multivariate Gaussian random component (corresponding to the 16 responses) for each glass. More precisely, define, for $g = 1, \ldots, 10$,

$$\boldsymbol{B}_g = \left(U_g^{[1]}, \ldots, U_g^{[14]}, V_g, W_g\right).$$

According to the MGLMM we define here, the random components $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_{10}$ are independent and multivariate normally distributed with distribution given by

$$\boldsymbol{B}_g \sim \mathrm{N}_{16}\left(\mathbf{0}, \boldsymbol{\Sigma}\right), \text{ for } g = 1, \ldots, 10\,.$$

Moreover, we assume that the multivariate vectors of responses, $(X_{tg}^{[1]}, \ldots X_{tg}^{[14]}, Y_{tg}, Z_{tg})$, for $t = 6, 12, 18$ and $g = 1, \ldots, 10$, are conditionally independent given the random components $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_{10}$. Moreover, according to the MGLMM, for $t = 6, 12, 18$ and $g = 1, \ldots, 10$, the vector of responses $(X_{tg}^{[1]}, \ldots X_{tg}^{[14]}, Y_{tg}, Z_{tg})$ is

conditionally distributed given $\boldsymbol{B}_g$ as specified below

$$\begin{cases} X_{tg}^{[1]}|U_g^{[1]} = u_1 \sim \text{Ga}\left(\exp\left\{\theta_t^{[1]} + u_1\right\}, \lambda_1\right), \ \forall u_1 \in \mathbb{R} \\ \vdots \\ X_{tg}^{[14]}|U_g^{[14]} = u_{14} \sim \text{Ga}\left(\exp\left\{\theta_t^{[14]} + u_{14}\right\}, \lambda_{14}\right), \ \forall u_{14} \in \mathbb{R} \\ Y_{tg}|V_g = v \sim \text{Bi}\left(9, \text{logit}^{-1}\left\{\alpha_t + v\right\}\right), \ \forall v \in \mathbb{R} \\ Z_{tg}|W_g = w \sim \text{ComPo}\left(\exp\left\{\beta_t + w\right\}, \lambda_Z\right), \ \forall w \in \mathbb{R}. \end{cases} \tag{VI.1}$$

Here the notation $X \sim \text{Ga}\ (\mu, \lambda)$ and $Z \sim \text{ComPo}\ (\mu, \lambda)$ indicates that X and Z are distributed according to the Gamma and Gamma-compound Poisson distributions with mean μ and dispersion parameter λ, respectively. $Y \sim \text{Bi}\ (n, p)$ denotes that Y is binomially distributed with size n and probability parameter p.

VI.3.2 The Covariance Structure of the Random Components

The covariance structure of the random components (given by the covariance matrix $\boldsymbol{\Sigma}$) will be characterised using the tools of graphical models. Before embracing this task, we give a short account of the basic theory of graphical models required for the exposition (see Whittaker, 1990 and Lauritzen, 1996 for details).

Let $\mathcal{G} = (V, E)$ denote an undirected graph with vertices composed of random variables. A pair of vertices belong to the set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ if, and only if, the corresponding variables are conditionally dependent given the remaining variables in the set of vertices $\mathcal{V}$. Usually, we represent the graph $\mathcal{G} = (V, E)$ by a set of points in the plane corresponding to the vertices in $\mathcal{V}$; an edge connecting two vertices is represented by a line connecting the two points corresponding to the vertices. The following basic definitions of graph theory will be necessary to characterise the covariance structure of the MGLMM we work with. We say that there is a path connecting two vertices, say v_1 and v_n, if there exists a sequence of vertices $v_1, \ldots, v_n$ such that, for $i = 1, \ldots, n-1$, the pair (v_i, v_{i+1}) is in $\mathcal{E}$. A set of vertices S, separates two disjoint sets of vertices A and B in the graph $\mathcal{G} = (V, E)$ when every path connecting a vertex in A to a vertex in B necessarily contains a vertex in S. According to the theory of graphical models (see Lauritzen, 1996 and Perl, 2009), the graph defined above satisfies the *separation principle*, which states that if a set of vertices S, separates two disjoint subsets of vertices A and B in the graph $\mathcal{G} = (V, E)$, then all variables in A are independent of all variables in B given S. Moreover, if the subsets A and B are isolated (*i.e.,* , there are no paths connecting a vertex in A to a vertex in B), then the variables in A are independent of the variables in B.

We characterise the covariance structure of the MGLMM defined in Section VI.3.1 by defining a graphical model constructed with its random components. In this way, for each g in $\{1, \ldots, 10\}$, we might construct the graph $\mathcal{G}_g = (\mathcal{V}_g, \mathcal{E}_g)$ with the set of vertices $\mathcal{V}_g = \{\ U_g^{[1]}, \ldots, U_g^{[14]}, V_g, W_g\}$ using the conventions defined above. Since the random vectors, $(U_g^{[1]}, \ldots, U_g^{[14]}, V_g, W_g)$ for $g = 1, \ldots, 10$, are per construction independent and identically distributed, the graphs $\mathcal{G}_1, \ldots \mathcal{G}_{10}$ are identical; therefore,

we suppress the subindex g in the discussion below and use the notation $\mathcal{G} = (V, E)$ to refer to a generic graph representing the (common) covariance structure of the random components.

Suppose that the multivariate random components of the MGLMM have a covariance structure encoded by the graph $\mathcal{G} = (V, E)$. This covariance structure allows us to draw conclusions on the dependence of the unobservable random components, which is not our primary interest. In order to extend those conclusions to the observed responses, we should use the *induced separation principle*, defined in Pelck & Labouriau (2021b), which states that if two disjoint sets of random components, for example $A = \{U^{[5]}, \ldots, U^{[14]}\}$ and $B = \{V, W\}$, are conditional independent given a separating set of random components $S = \{U^{[1]}, \ldots, U^{[4]}\}$, then the corresponding set of conditional responses $\tilde{A} = \{X^{[5]}, \ldots, X^{[14]}\}$ and $\tilde{B} = \{Y, Z\}$ are conditional independent given the set of random components S. This result implies that the knowledge of the random components in S renders the VOC's in $\tilde{A}$ uninformative with respect to the lesion area and the proportion of apples presenting visible symptoms.

In the analysis of the experiment described above, we adjusted the GLMMs introduced in Section VI.2 using the Laplace approximation method proposed by Breslow and Clayton (1993). We modelled the predicted values of the random components of those models by finding the graph which minimises the BIC (Bayesian Information Criterion) as exposed in Abreu et al. (2010) (see also Edwards et al., 2010).

VI.4 Results

We give below a brief description of the results obtained, see Holthusen et al. (2021a) for a full discussion. Figure VI.1 displays the representation of the estimated graph describing the covariance structure of the random components of the MGLMM we adjusted. It is remarkable that a group of only four random components related to VOCs (composed by anisole, 3-pentanone, 2-methyl-1-propanol and 2-phenylethanol) separates the random components associated to the two infection responses (*i.e.,* , the infection proportion and the lesion area) from the random components connected with the other VOCs. Therefore, by the extended separation principle, the knowledge of the random components corresponding to the concentration of anisole, 3-pentanone, 2-methyl-1-propanol and 2-phenylethanol renders the concentrations of the other VOCs independent of the infection proportion and the lesion area.

We verified the adequacy of the MGLMM described in the following way. First, we checked the marginal GLMMs by plotting the Pearson residuals against the fitted values (not shown). No anomalies were encountered. Moreover, we applied the cumulative distribution function of the putative distribution to each observation (with the estimated mean and dispersion) and verified whether the resulted transformed observations adhered to the uniform distribution in the interval between 0 and 1. All the p-values found were larger than 0.10.

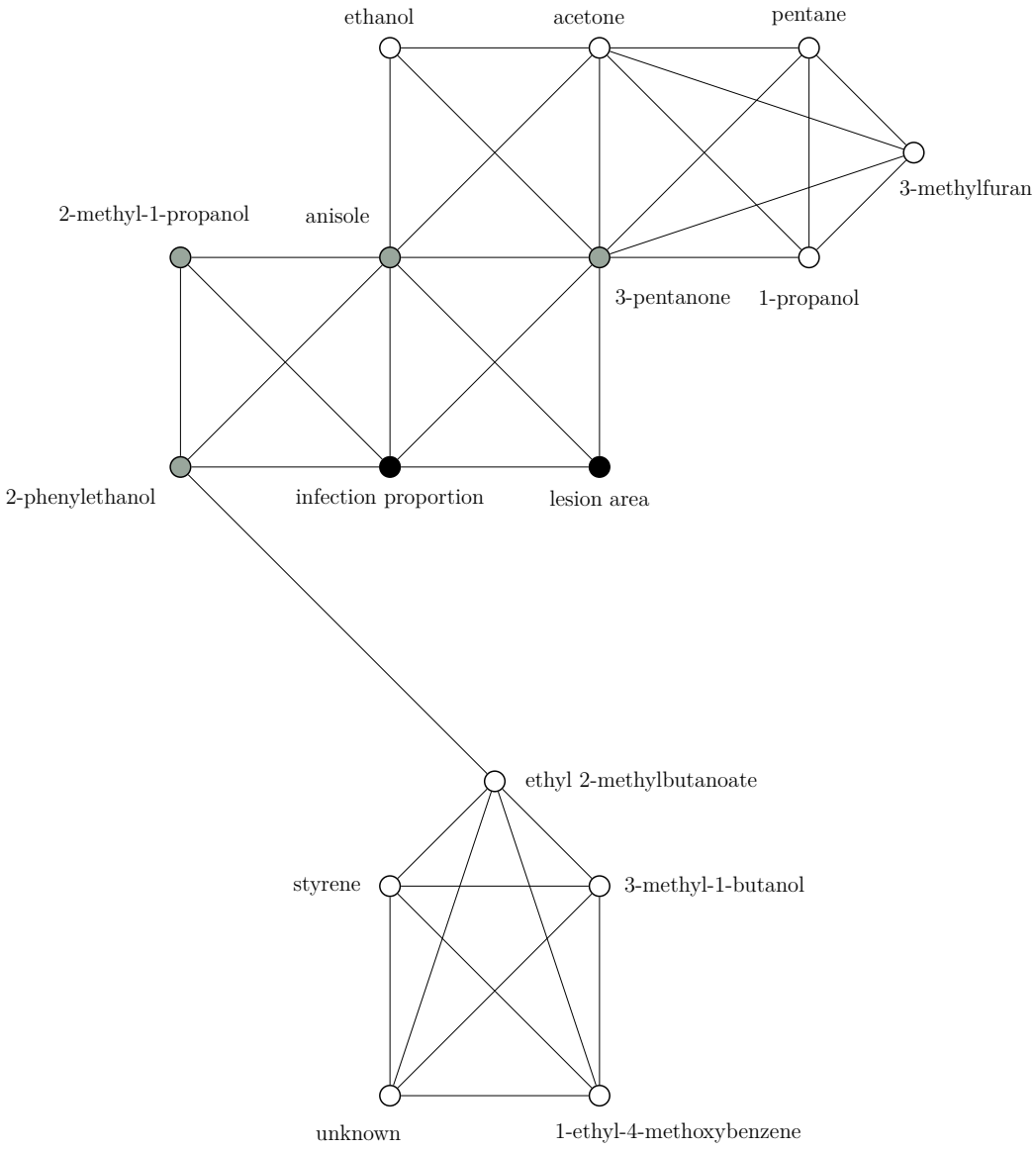


Figure VI.1: Graphical model representing the covariance structure of the random components related to the 14 VOCs, the lesion area and the proportion of infection. The vertices related to the infection responses (lesion area and proportion of infection) are represented as black circles. The vertices directly connected to the infection responses (depicted as grey circles) separate the vertices related to the infection responses from the vertices related to the other VOCs (represented as write circles).

Acknowledgements

The first and the last authors were partially financed by the Applied Statistics Laboratory (aStatLab) at the Department of Mathematics, Aarhus University.

References

Abreu, G. C., Labouriau, R. & Edwards, D. (2010), 'High-dimensional graphical model search with gRapHD R package', *Journal of Statistical Software* **37**(1).

Ali, E., Pandit, L., Mulvaney, K. & Amiri, A. (2018), 'Sensitivity of *Phacidiopycnis spp.* isolates from pome fruit to six pre- and postharvest fungicides.', *Plant Disease* **102**(3).

Cordeiro, G. M., Labouriau, R. & Botter, D. (2021), 'An introduction to bent jørgensen's ideas', *Brazilian journal of Probability and Statistics* **35**(1), 2–20.

Edwards, D., de Abreu, G. C. & Labouriau, R. (2010), 'Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests', *BMC Bioinformatics* **11**(1).

Holthusen, H., A., L., Weber, R. & Edelenbos, M. (2021b), 'Volatile markers emitted by *Phacidiopycnis washingtonensis* from agar culture and naturally infected apple fruit.'. In preparation.

Holthusen, H., Luca, A., Pelck, J., Labouriau, R. & Edelenbos, M. (2021a), 'Detection of rubbery rot caused by *Phacidiopycnis washingtonensis* by use of volatile monitoring in apple storage.'. In preparation.

Holthusen, H. & Weber, R. (2021), 'Infection conditions for *Neofabraea perennans* and *Phacidiopycnis washingtonensis* on developing apple fruit in the orchard.'. In preparation.

Jørgensen, B. (1987), 'Exponential dispersion models', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 127–162.

Labouriau, R. (2021), 'Inference and characterisation of the gamma-compound poisson family.'. In preparation.

Lauritzen, S. L. (1996), *Graphical models*, Vol. 17, Clarendon Press.

Pelck, J. S. & Labouriau, R. (2021b), Multivariate generalised linear mixed models with graphical latent covariance structure. arXiv:2107.14535.

Perl, J. (2009), *Causality: models, reasoning and inference*, second edition edn, Cambridge University Press.

Tweedie, M. (1984), *Statistics: applications and new directions*, Indian Statistical Institute, Calcutta, chapter An index which distinguishes between some important exponential families, pp. 579–604.

Whittaker, J. (1990), *Graphical models in applied multivariate analysis*, Chichester New York et al: John Wiley & Sons.