Modeling techniques and statistical inference for multidimensional effects

Doctoral thesis
for the award of the doctoral degree
of the Faculty of Mathematics and Natural Sciences
of the University of Cologne

submitted by

Niklas Hagemann

accepted in the year 2025

Acknowledgements

First of all, I would like to thank my supervisor, Kathrin Möllenhoff, who encouraged and supported me in every phase of my doctoral studies. She always maintained the perfect balance between providing guidance and offering freedom, creating an environment in which I could grow academically and personally.

My special thanks also go to all the members of RTG 2624, especially Holger Schwender, who also kindly agreed to serve as second reviewer. I would also like to thank my colleague Lucia as well as all other doctoral researchers in the RTG, I am grateful for the academic exchange as well as the mutual support and encouragement during our time together.

Additionally, I am thankful for the fruitful discussions on various statistical topics that I still have with many of my former fellow students from the Master's program.

I would also like to thank all my friends, who have enriched my life outside the university and always found motivating words at the right time.

And most of all, I want to thank my parents, Karin and Gerhard, for their constant support and unconditional trust in me. Without you, I would not be the person I am.

Danke für alles.

Abstract

The increasing complexity of biostatistical research questions requires statistical methods that can effectively address multidimensional problems. This thesis addresses issues arising from multidimensionality in statistical testing and modeling, with a focus on model-based equivalence tests and hazard regression models. Specifically, it examines three directions of multidimensionality: multivariate outcomes, model uncertainty, and multidimensional covariate effects. Four contributions discuss the necessity of adapting model-based equivalence tests and hazard regression models to account for these three aspects of multidimensionality.

The first contribution extends model-based equivalence tests to multivariate, potentially mixed-scale outcomes using generalized joint regression models. This approach overcomes the limitations of a previous approach that is only capable of bivariate binary outcomes and relies on the intersection-union principle leading to an overly conservative test, particularly for small sample sizes. In contrast, a new maximum of maxima approach is used to increase the power in finite samples while maintaining asymptotic validity.

The second contribution addresses model uncertainty, a common issue in applied research where often the true model is unknown. By incorporating model averaging to model-based equivalence tests and deploying a confidence interval-based testing approach, the proposed method offers a robust and numerically feasible alternative that retains the asymptotic properties.

The third and fourth contributions shift the focus to multidimensional covariate effects. In the third article, functional random coefficients are introduced to model heterogeneously time-varying covariate effects. Such coefficients are not only capable of time-varying and subgroup-specific covariate effects but also of covariate effects in which the time-variation itself is heterogeneous. The functional random coefficients are constructed as tensor product interactions of heterogeneity and time. While the third contribution introduces these effects to generalized additive models, the fourth article discusses their applicability in survival analysis by incorporating such effects into hazard regression models.

The methods are evaluated through simulations outlining their flexibility while either retaining the asymptotic properties of the model-based equivalence test or the prevention of overfitting of the regression models. The practical relevance of the proposed methods is demonstrated using case studies from pharmacology, toxicology, and oncology. This thesis thus contributes novel approaches that enhance the flexibility and applicability of statistical methods in multidimensional biostatistical research.

Contents

Acknowledgment **II**

Abstract **III**

1 Introduction **1**

2 Preliminaries **5**

2.1 Model-based equivalence tests 5

2.2 Copulae 10

2.3 Survival analysis 11

2.4 Generalized additive models 12

3 Articles **15**

3.1 Testing for similarity of multivariate mixed outcomes using generalized joint regression models with application to efficacy-toxicity responses . . . 16

3.2 Overcoming model uncertainty – how equivalence tests can benefit from model averaging 43

3.3 Dynamic Heterogeneity in Discrete Choice Experiments 63

3.4 Capturing heterogeneous time-variation in covariate effects in non-proportional hazard regression models 104

4 Discussion **123**

References **126**

Erklärung zur Dissertation **131**

1 Introduction

In recent years, multidimensional and multivariate research questions have become increasingly relevant in the field of biostatistics (Rahnenführer et al., 2023). However, numerous statistical methods require further adaptation to be applicable in multidimensional settings. This dissertation focuses on the development, enhancement, and application of regression methods and regression-based tests capable of accommodating multidimensionality. Concerning regression-based testing approaches, *model-based equivalence tests* (Dette et al., 2018) will be considered, while in the field of regression modeling *hazard regression models* will be investigated. Hence, the overarching topic of this thesis consists of two subtopics: multidimensionality in model-based equivalence tests and multidimensionality in hazard regression models.

In contrast to model-based equivalence tests, which will be investigated for classical data types, such as binary and continuous variables, hazard regression models are designed for a specific data type: time-to-event data, also referred to as survival data, which measures the time until a specific event of interest, such as death, occurs. Time-to-event data is different from other forms of data since it exhibits a time-dependent structure and usually includes censored observations, i.e. cases for which either the event is not observed or the exact event time is unknown (Klein and Moeschberger, 2005). Consequently, survival data requires specialized statistical techniques, such as hazard regression models, to account for censoring and the dynamic relationship between time and event risk.

There are several directions in which a modeling problem can be multidimensional and this thesis investigates the following three directions:

1. The response variable can be multivariate, which leads to special requirements for the modeling approach. Although modeling techniques for multivariate outcomes, particularly *generalized joint regression models* (Radice and Marra, 2016; Filippou et al., 2017; Marra and Radice, 2017; Klein et al., 2019), have already been developed, the adaptation of model-based tests remains underexplored.

2. The multidimensionality can also be inherent in the model itself. In applied research, model uncertainty is often present, i.e. the true underlying model is unknown. In addition, for a single modeling problem, there is often more than one reasonable modeling approach and combining several models (e.g. via model averaging) might be necessary. Model averaging has been extensively researched (e.g. Buckland et al., 1997; Wasserman, 2000; Hjort and Claeskens, 2003; Schorning et al., 2016) but its incorporation into model-based tests needs further research.

3. The effects of explanatory variables can be multidimensional. Here, multidimensional varying coefficients are of special interest, i.e. coefficients that do not only vary with respect to one variable but with respect to (non-linear) interactions of variables. In this context, capturing heterogeneous time-variation in hazard regression models is of particular interest.

For each of these directions of multidimensionality, its effect on a statistical method will be investigated and the necessary adaptations will be developed. While the first two directions of multidimensionality will be investigated in the context of model-based equivalence tests, the third one will be examined with regard to hazard regression models.

Demonstrating equivalence between two groups is an important research question in biostatistics, most prominently used in bioequivalence studies (Hauschke et al., 2007; Möllenhoff et al., 2022) but also in other applications (see, e.g. Cade, 2011; MacKenzie and Kendall, 2002; Dixon and Pechmann, 2005). The underlying null hypothesis is that the difference of a parameter of interest between two groups is outside an equivalence region, which is determined by a pre-specified threshold value ε. In other words, this means that the absolute value of this difference is larger or equal to ε. If the null hypothesis is rejected, equivalence can be concluded. Hence, equivalence tests reverse the burden of proof compared to a standard significance test making them promising in regulatory settings (Hauschke et al., 2007).

Classical approaches (e.g. Schuirmann, 1987; Lakens, 2017) are based on testing the equivalence of single quantities, e.g. the mean, the area under the curve (AUC), or other values of interest. However, when differences depending on a particular covariate are observed, e.g. dose-response relations, these approaches may not be very accurate (Dette et al., 2018). Instead, considering the entire covariate range, describing for instance a time window or a dose range, has recently been proposed by testing for equivalence of whole regression curves. Such tests are typically based on the principle of confidence interval inclusion (Liu et al., 2009; Gsteiger et al., 2011; Bretz et al., 2018). However, a more direct approach applying various distance measures has been introduced by Dette et al. (2018), which was observed to be more powerful. Based on this, many further developments e.g. for different outcome distributions or specific model structures have been introduced (see, e.g. Möllenhoff et al., 2020, 2021, 2024).

However, two topics merit further research: First, regarding the first direction of multidimensionality, the issue arises that some studies involve a joint comparison of more than one response variable. This is particularly relevant for the comparison of two drugs whenever both – efficacy and toxicity – need to be investigated. Möllenhoff et al. (2021) developed an adaptation of the test of Dette et al. (2018) for the special case of bivariate binary outcomes. However, this can necessitate the transformation of continuous variables to binary ones based on thresholds, which can result in a loss of information (see the case study of Möllenhoff et al. (2021) for an explicit example). Therefore, this dissertation aims to introduce a more flexible approach allowing for other scales of measures of the outcome variables, including mixed outcomes. This is achieved by deploying *generalized joint regression models* (GJRMs; Radice and Marra, 2016; Filippou et al., 2017; Marra and Radice, 2017; Klein et al., 2019), which allow for the joint modeling of multivariate outcomes with arbitrary marginal distributions, as underlying regression models. This also ensures that the proposed approach directly generalizes for outcome variables with more than two dimensions. The approach of Möllenhoff et al. (2021) relies on the intersection-union principle (Berger, 1982) to construct the test statistic, which results in a test that is overly conservative for smaller sample sizes. Therefore, this thesis uses an alternative test statistic, called *maximum of maxima*, in order to increase the power in finite samples while retaining the same asymptotic properties.

Second, another issue arises from the second direction of multidimensionality: the test of Dette et al. (2018) as well as all proposed adaptations of this test (e.g. Möllenhoff et al., 2020, 2021, 2024) have one thing in common: they assume the true underlying regression model to be known. In applied research, this assumption is frequently not fulfilled. Usually, only a set of plausible models, also called candidate models, is known leading to model uncertainty. The model is then chosen either manually or based on a model selection procedure (see Möllenhoff et al. (2018) for an example of the latter). However, in both cases accidentally misspecifying the model can invalidate the test and cause severe type I error inflation. Therefore, this dissertation aims to introduce model averaging to the test of Dette et al. (2018) to overcome the model uncertainty. This can lead to the issue that the test algorithm of Dette et al. (2018) may become numerically infeasible due to the increased model complexity. Hence, an alternative testing procedure is suggested which is similar to the one proposed by Bastian et al. (2024) and makes use of the duality between confidence intervals and hypothesis testing.

While the first part of this thesis investigates these two directions of multidimensionality in model-based equivalence tests, the second part addresses multidimensionality in hazard regression models, which are specifically designed for time-to-event data. Hazard regression models play a crucial role in analyzing the effects of covariates on the survival time by modeling the hazard function, which represents the instantaneous rate of occurrence of the event at a given time, conditional on survival up to that time, as a function of time and explanatory variables. Hence, they allow to assess the impact of various factors on the hazard rate and enable the identification of significant effects. Besides the widely used standard model, the *Cox proportional hazards model* (Cox, 1972), more flexible approaches (e.g. Kneib and Fahrmeir, 2007; Hennerfeind et al., 2006; Bender et al., 2018) have been developed more recently. While these approaches already introduce several flexible effects including frailty and time-varying effects, heterogeneously time-varying covariable effects remain underexplored. Such effects occur if the effect of a covariate is subgroup-specific, time-varying and its time-variation is also subgroup-specific. This dissertation aims to close this research gap by introducing *functional random coefficients* to hazard regression models. These functional random coefficients are constructed as *tensor products* (Kneib et al., 2019), i.e. the non-linear interaction of the two main effects – heterogeneity and time-variation. The proposed approach allows for non-linear time effects due to being based on penalized splines, which also prevents overfitting in case of the absence of such effects, and uses an efficient random effects basis to model the heterogeneity.

An empirical example of a hazard regression model with functional random coefficients is given in Figure 1, where the survival times of patients with brain tumors are investigated. The patients are grouped with regard to their specific diagnoses. Here, the effect of one of the explanatory variables – FGA – is diagnosis-specific, time-varying and this time-variation is diagnosis-specific, too. Therefore, its effect is modeled by a functional random coefficient leading to one non-linear smooth time effect for each of the groups. This example will be discussed in more detail in Section 3.4.

Functional random coefficients are introduced to hazard regression models utilizing the framework of *piecewise exponential additive mixed models* (PAMMs Bender et al., 2018). The main advantage of PAMMs is that the model parameters can be estimated from a *generalized additive model* (Hastie and Tibshirani, 1986) using standard estimation techniques.

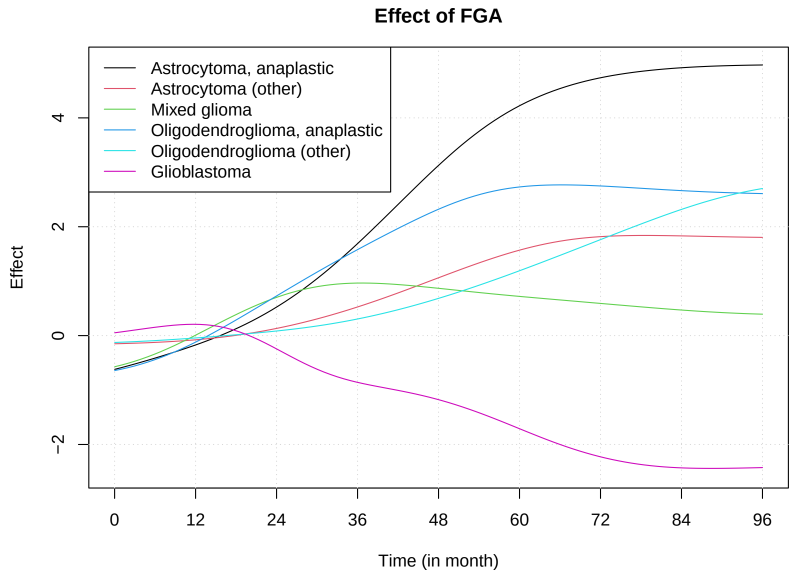


Figure 1: The heterogeneously time-varying effect of FGA is modeled by a functional random coefficient resulting in one smooth non-linear time effect for each of the six diagnosis groups.

The remainder of this thesis is structured as follows: in Section 2, the methodological background is introduced. Here, model-based equivalence testing is presented, copulae as the theoretical foundation of GJRMs are outlined, survival analysis, including a general introduction to time-to-event data, is discussed and generalized additive models are formally introduced. In Section 3, four articles discussing the three research problems are presented. In the first part of this thesis, adaptations of model-based equivalence tests are discussed with regard to multivariate (mixed) outcomes in Section 3.1 and model uncertainty in Section 3.2. In the second part, Section 3.3 introduces functional random coefficients as an approach to model heterogeneously time-varying covariable effects, before they are applied to hazard regression models in 3.4. While the application in Section 3.3 is not from the field of biostatistics, it provides the theoretical basis for 3.4 and introduces Bayesian estimation as an alternative to the frequentist inference in Section 3.4. Finally, Section 4 closes with a discussion.

2 Preliminaries

This section provides the theoretical foundations for the methods which will be discussed in Section 3. In Section 2.1, model-based equivalence tests are introduced, and in Section 2.2 copulae as the theoretical foundation of GJRMs are outlined. With regard to the second part of the thesis, Section 2.3 discusses survival analysis including a general introduction to time-to-event data and generalized additive models are formally introduced in Section 2.4.

2.1 Model-based equivalence tests

Classical significance tests, e.g. the t-test, aim to show significant differences by rejecting the null hypothesis of equality of some quantity μ between two groups versus the alternative of inequality, i.e.

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_1 : \mu_1 \neq \mu_2.$$

However, when the null hypothesis cannot be rejected, equality can not be concluded as the *absence of evidence is not evidence of absence* (Altman and Bland, 1995) of inequality. In contrast, equivalence tests aim to show the equivalence of the quantities rather than the difference. Here, equivalence is defined as equality up to a pre-specified threshold which is usually interpreted as the threshold of practical irrelevance. The corresponding hypothesis is given by

$$H_0 : |\mu_1 - \mu_2| \geq \varepsilon \quad \text{vs.} \quad H_1 : |\mu_1 - \mu_2| < \varepsilon, \tag{1}$$

where ε is the equivalence threshold. The null hypothesis in (1) is usually tested using the two one-sided tests procedure proposed by Schuirmann (1987). However, this approach is only capable of comparing scalar quantities. Therefore, when investigating differences that depend on a particular covariate, such as dose-response relations, these approaches require summarizing such relations to one single quantity, e.g. the mean or the area under the curve (AUC). Therefore, they may not be very accurate for such studies.

Instead, considering the entire covariate range, describing for instance a time window or a dose range, has recently been proposed by testing for equivalence of whole regression curves. Such tests are typically based on the principle of confidence interval inclusion (Liu et al., 2009; Gsteiger et al., 2011; Bretz et al., 2018). However, a more direct approach applying various distance measures has been introduced by Dette et al. (2018), which appeared to be particularly more powerful.

In order to introduce this test approach formally, first the underlying models need to be defined. For compatibility with existing literature, especially Dette et al. (2018), the underlying models will be introduced based on the typical notation of dose-response models. This is motivated by the fact that comparing dose-response models is a typical application of model-based equivalence tests in applied research. In pharmaceutical research, dose-response models (see, e.g. Pinheiro et al., 2006) model the influence of the dose on the response (e.g. the efficacy) based on a parametric regression model. Although simple linear models can be used as well, most dose-response models are non-linear. Frequently used examples are the Emax model, the exponential model, or the sigmoid Emax model (Pinheiro et al., 2006; Duda et al., 2022). Figure 2 shows an overview of dose-response

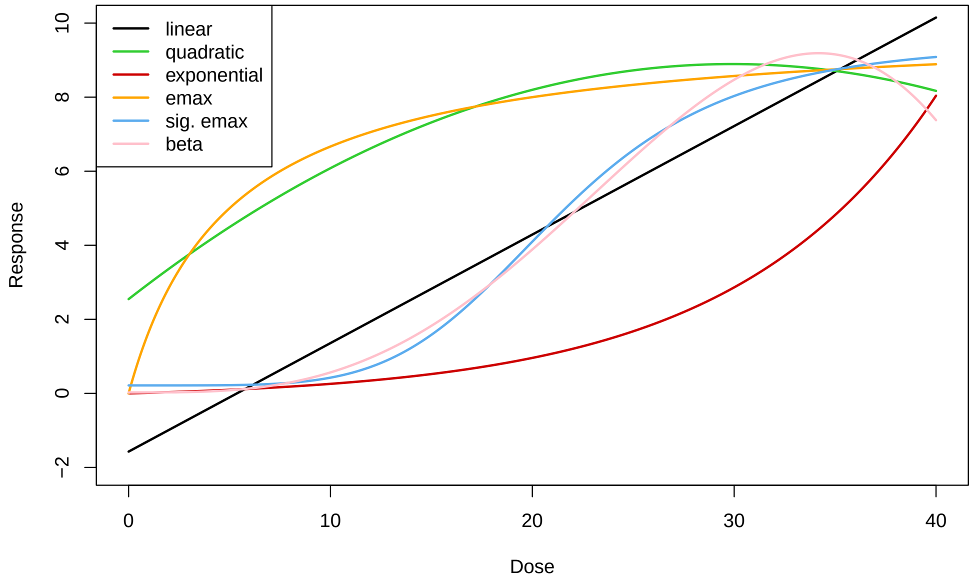


Figure 2: Common dose response models according to Duda et al. (2022).

models that Duda et al. (2022) consider to be common. For the two groups $\ell = 1, 2$, the response variable is given as $y_{\ell ij}$, whre the index $i = 1, ..., I_\ell$ denotes the dose levels, $j = 1, ..., n_{\ell i}$ denotes the observation index within each dose level and $n = n_1 + n_2$ is the overall number of observations with $n_\ell = \sum_{i=1}^{I} n_{\ell i}$, $\ell = 1, 2$. A flexible dose-response model is then given by

$$y_{\ell ij} = m_\ell(x_{\ell i}, \boldsymbol{\theta}_\ell) + e_{\ell ij}, \quad j = 1, ..., n_{\ell i}, \quad i = 1, ..., I_\ell, \quad \ell = 1, 2,$$

where $x_{\ell i} \in \mathcal{X} \subset \mathbb{R}$ is the dose level, i.e. the value of the deterministic explanatory variable. The error terms $e_{\ell ij}$ are assumed to be independent, have expectation zero and finite variance σ_ℓ^2. The function m_ℓ models the effect of $x_{\ell i}$ on $y_{\ell ij}$ via a potentially non-linear regression curve with $\boldsymbol{\theta}_\ell$, $\ell = 1, 2$ being its parameter vector. It should be noted that there is an alternative notation omitting the index for the dose level which is used in the first article, i.e. in Section 3.1.

Using this notation, the hypotheses of the test of Dette et al. (2018) are given as

$$H_0 : d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \geq \varepsilon \quad \text{vs.} \quad H_1 : d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) < \varepsilon, \tag{2}$$

where $d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) := d(m_1(x, \boldsymbol{\theta}_1), m_2(x, \boldsymbol{\theta}_2))$ is some distance measure of the difference curve $\Delta(x, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = m_1(x, \boldsymbol{\theta}_1) - m_2(x, \boldsymbol{\theta}_2)$. Dette et al. (2018) propose to use either the maximum absolute deviation, also known as L^∞ norm, leading to

$$d_\infty = d_\infty(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \max_{x \in \mathcal{X}} |m_1(x, \boldsymbol{\theta}_1) - m_2(x, \boldsymbol{\theta}_2)|, \tag{3}$$

or the (squared) L^2 norm, leading to

$$d_2 = d_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \int_{\mathcal{X}} (m_1(x, \boldsymbol{\theta}_1) - m_2(x, \boldsymbol{\theta}_2))^2 dx. \tag{4}$$

In addition, Bastian et al. (2024) introduced the L^1 norm

$$d_1 = d_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \int_{\mathcal{X}} |m_1(x, \boldsymbol{\theta}_1) - m_2(x, \boldsymbol{\theta}_2)| dx \tag{5}$$

as an alternative. However, subsequent research (e.g. Möllenhoff et al., 2018, 2020, 2021) predominately uses the maximum absolute deviation due to it being the most intuitive measure and having a relatively simple interpretation. For the remainder of this section, d is used as a general term for d_∞, d_2, and d_1. The threshold value ε must be specified depending on the choice of distance measure, as the measures differ in scale.

The test statistic $\widehat{d}$ is given by calculating the chosen distance, i.e. (3), (4) or (5), of the observed difference curve $\Delta(x, \widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2)$ resulting from the estimated models $m_1(x, \widehat{\boldsymbol{\theta}}_1)$ and $m_2(x, \widehat{\boldsymbol{\theta}}_2)$. For $\widehat{d}_\infty$, the calculation of the test statistic is exemplarily shown in Figure 3. Dette et al. (2018) propose to estimate these models using *ordinary least squares* (OLS) optimization and show that the OLS parameter estimates converge in distribution to a normal distribution. However, subsequent research (e.g. Möllenhoff et al., 2021) deploys maximum likelihood (ML) estimation. This is justified by the fact that under regulatory conditions the ML estimator also converges in distribution to a normal distribution (see, e.g. Theorem 3.3 of Newey and McFadden, 1994). In addition, for linear models with i.i.d. normally distributed responses, the OLS and ML estimators are identical as they optimize the same objective function (see, e.g. Section 3.2.1 of Fahrmeir et al., 2022).

Dette et al. (2018) develop two test approaches, one conducting the test decision based on the asymptotic distribution of the test statistic and one using a bootstrap approach. Under regulatory assumptions (see the Appendix of Dette et al. (2018) for details), the asymptotic distribution of the test statistic $\widehat{d}_2$ is a normal distribution, i.e.

$$\sqrt{n}(\widehat{d}_2 - d_2) \xrightarrow{D} N(0, \mathrm{Var}_{d_2}(\theta_1, \theta_2)),$$

with a closed form equation for the variance $\mathrm{Var}_{d_2}(\theta_1, \theta_2)$. Under the same assumptions, the distribution of $\widehat{d}_\infty$ converges towards a normal distribution if the maximum of $|\Delta(x, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)|$ is unique. However, the variance depends on the location of this maximum, which in practice necessitates its precise estimation. In contrast, if the set of extremal points has a cardinality larger than one, the asymptotic distribution is the distribution of

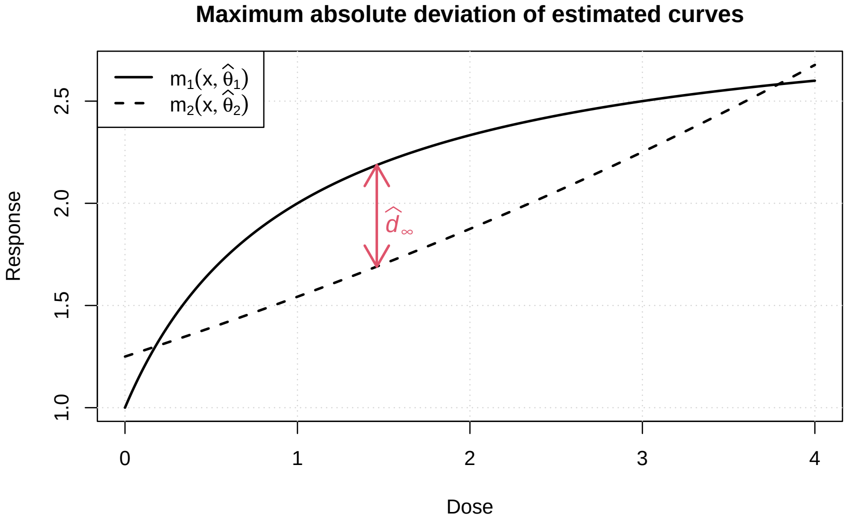


Figure 3: Maximum absolute deviation $\widehat{d}_\infty$ of the estimated curves $m_1(x, \widehat{\boldsymbol{\theta}}_1)$ and $m_2(x, \widehat{\boldsymbol{\theta}}_2)$. The shown example is taken from the simulation study of Dette et al. (2018), where m_1 is an Emax model and m_2 is an exponential model.

a maximum of dependent normally distributed random variables whose variances and dependence structure depend on the location of the extremal points. Therefore, Dette et al. (2018) suggest using a bootstrap-based test instead. In addition, for scenarios where the asymptotic test is applicable, i.e. if the maximum of $|\Delta(x, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)|$ is unique, their simulation study indicates that the asymptotic test is overly conservative even for relatively large sample sizes.

When using the test of Bastian et al. (2024), i.e. the test statistic $\widehat{d}_1$, the asymptotic distribution of the test statistic is an integral over Gaussian processes which depend on the values of the set $\{x \in \mathcal{X} | \Delta(x, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = 0\}$. It should be noted that here the regulatory assumptions (see the Appendix of Bastian et al. (2024) for details) are slightly stricter than the ones of Dette et al. (2018). Due to the test performance being strongly dependent on the precise estimation of the set $\{x \in \mathcal{X} | \Delta(x, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = 0\}$, the bootstrap test is also recommended here.

Except for the test statistic, the bootstrap test is identical across the different distance measures. It is based on generating a random sample from the distribution of the test statistic under the null hypothesis. In the first step, a set of parameters is estimated that optimizes the OLS or ML criterion under the side constraint of being on the edge of the null hypothesis, i.e. $d = \varepsilon$. Then, a parametric bootstrap is conducted where data is sampled using this set of parameters. Estimating the test statistic from the bootstrap data leads to a random sample of the distribution of the test statistic under the null hypothesis. Hence, the empirical quantile of the ordered sample can be used as an approximate quantile of this distribution. The bootstrap procedure is outlined in detail in Algorithm 1, which is similar to Algorithm 1 of Dette et al. (2018) but uses a notation closer to the one in Sections 3.1 and 3.2 and provides more technical details.

Naturally, optimization under an equality side constraint is numerically challenging, especially for complex models. Therefore, for the L^1 norm, Bastian et al. (2024) suggests an alternative test approach which is based on the *duality of tests and confidence intervals* and deploys a percentile bootstrap (for full details, see Algorithm 1 of Bastian et al., 2024). In Section 3.2, this idea will be transferred to the L^∞ norm-based test. Here, an additional approach will be introduced that combines a bootstrap variance estimator with the asymptotic normality.

Algorithm 1: Parametric bootstrap algorithm of Dette et al. (2018).

1. Calculate parameter estimates $\widehat{\boldsymbol{\theta}}_\ell$, $\ell = 1, 2$, either via OLS or ML estimation, and estimate the variance using a consistent estimator $\widehat{\sigma}_\ell$, $\ell = 1, 2$.

2. Calculate the test statistic $\widehat{d} = d(m_1(x, \widehat{\boldsymbol{\theta}}_1), m_2(x, \widehat{\boldsymbol{\theta}}_2))$.

3. To approximate the null distribution, define estimators for parameter vectors $\boldsymbol{\theta}^{(l)}$, $\ell = 1, 2$, so that the corresponding curves fulfill the null hypothesis in (2). That is,

$$\widehat{\widehat{\boldsymbol{\theta}}}_\ell = \begin{cases} \widehat{\boldsymbol{\theta}}_\ell & \text{if } \widehat{d} \geq \varepsilon \\ \overline{\boldsymbol{\theta}}_\ell & \text{if } \widehat{d} < \varepsilon \end{cases} \qquad \ell = 1, 2,$$

   where $\overline{\boldsymbol{\theta}}_\ell$, $\ell = 1, 2$ maximizes the same objective function as $\widehat{\boldsymbol{\theta}}_\ell$, $\ell = 1, 2$, but under the constraint

$$d = \varepsilon.$$

   Technically, the range $\mathcal{X}$ of the explanatory variable is discretized to make the optimization feasible. The constrained problem can be solved using the augmented Lagrangian minimization algorithm (Hestenes, 1969).

4. Execute the following steps:

   (a) Obtain bootstrap samples under the null hypothesis in (2) by generating data according to the model parameters $\widehat{\widehat{\boldsymbol{\theta}}}_\ell$, $l = 1, 2$. Under the assumption of normality that is

$$y^*_{\ell ij} \sim N(\widehat{\mu}_{\ell i}, \widehat{\sigma}^2_\ell), \quad j = 1, ..., n_{\ell i},\ i = 1, ..., I_\ell,\ \ell = 1, 2,$$

   where

$$\widehat{\mu}_{\ell i} = m_\ell(x_{\ell i}, \widehat{\boldsymbol{\theta}}_\ell), \quad i = 1, ..., I_\ell,\ \ell = 1, 2.$$

   Alternative distributions, e.g. the Bernoulli distribution for binary data, can be used as well.

   (b) From the bootstrap samples, calculate parameter estimates $\widehat{\boldsymbol{\theta}}^*_\ell$ as in step 1 and the test statistic

$$\widehat{d}^* = d(m_1(x, \widehat{\boldsymbol{\theta}}^*_1), m_2(x, \widehat{\boldsymbol{\theta}}^*_2)).$$

   (c) Repeat steps (a) and (b) B times to generate replicates $\widehat{d}^*_1, \ldots, \widehat{d}^*_B$ of $\widehat{d}^*$ and let $\widehat{d}^*_{(1)} \leq \ldots \leq \widehat{d}^*_{(B)}$ denote the corresponding order statistic. The estimator of the α-quantile of the distribution of $\widehat{d}^*$ is given by $\widehat{d}^*_{(\lfloor B\alpha \rfloor)}$.

5. At a significance level α, reject the null hypothesis in (2) and assess similarity if

$$\widehat{d} < \widehat{d}^*_{(\lfloor B\alpha \rfloor)}.$$

   Alternatively, obtain the p-value $\widehat{F}_B(\widehat{d}) = \frac{1}{B} \sum_{i=1}^B \mathbb{1}(\widehat{d}^*_i \leq \widehat{d})$ and reject the null hypothesis in (2) if $\widehat{F}_B(\widehat{d}) < \alpha$, where $\widehat{F}_B$ denotes the empirical cumulative distribution function of the bootstrap sample.

2.2 Copulae

Understanding and modeling dependencies between random variables is a fundamental challenge in (bio-) statistics, particularly regarding the joint distribution of dependent random variables. A typical example is the joint investigation of the efficacy and toxicity of a drug, which are typically not independent of each other. Therefore, this dependency needs to be taken into account when applying the dose-response models introduced in Section 2.1 to such outcomes.

Copulae (Sklar, 1959) provide a flexible framework allowing the construction of multivariate models with arbitrary marginals and complex dependencies. In addition, they allow to separate the effect resulting from the dependence of the variables from the effect of the marginal distributions in a joint distribution. Sklar's theorem (Sklar, 1959) states that for any d-dimensional multivariate cumulative distribution function (CDF) $\boldsymbol{F}(y_1, \ldots, y_d)$ with marginal CDFs $F_1(y_1), \ldots, F_d(y_d)$, a copula C exists such that

$$F(y_1, y_2, \ldots, y_d) = C(F_1(y_1), \ldots, F_d(y_d)),$$

where $C : [0,1]^d \to [0,1]$ is a multivariate CDF with uniform marginals and that C is unique if $F_1(y_1), \ldots, F_d(y_d)$ are continuous. This decomposition allows to specify the marginal distributions independently from the dependence structure, facilitating model construction in high-dimensional settings (Joe, 2015).

Almost all commonly used copulae belong to one of two families: Archimedean and elliptical copulae. Following the definition of McNeil and Nešlehová (2009), a copula is considered to be an Archimedean copula if and only if it can be expressed as

$$C(u_1, \ldots, u_d) = \varphi\left(\sum_{i=1}^{d} \varphi^{-1}(u_i)\right), \quad (u_1, \ldots, u_d) \in [0,1]^d,$$

where the continuous, convex and decreasing generator function $\varphi : [0, \infty) \to [0,1]$ satisfies $\varphi(0) = 1$ and $\lim_{x\to\infty} \varphi(x) = 0$ and is strictly decreasing on $[0, \inf\{x|\varphi(x) = 0\})$. For $d > 2$, φ must additionally be d-monotone, which is satisfied if φ is $d-2$ times differentiable, each of the derivatives $\varphi^{(\tilde{d})}, \tilde{d} = 1, ..., d-2$ satisfies $(-1)^{\tilde{d}}\varphi^{(\tilde{d})}(z) \geq 0$ $\forall\, \tilde{d} \in \{1, ..., d-2\}, z \in (0, \infty)$ and the function $(-1)^{d-2}\varphi^{(d-2)}(z)$ is decreasing and convex on $(0, \infty)$ (McNeil and Nešlehová, 2009). It should be noted that some authors (e.g. Joe, 2015) define Archimedean copulae in terms of φ^{-1} rather than φ.

In contrast, elliptical copulae are derived from elliptical distributions, such as the multivariate normal and multivariate t distributions. Formally, an elliptical copula is defined as

$$C(u_1, \ldots, u_d) = \boldsymbol{F}(F^{-1}(u_1), \ldots, F^{-1}(u_d), \boldsymbol{R}), \quad (u_1, \ldots, u_d) \in [0,1]^d,$$

where $\boldsymbol{F}$ is the CDF of a multivariate elliptical distribution with correlation matrix $\boldsymbol{R}$ and F^{-1} is a univariate quantile function (Fang et al., 2002; Mai and Scherer, 2017). The most well-known representative of the class of elliptical copulae might be the Gaussian copula, which is given by

$$C(u_1, \ldots, u_d) = \boldsymbol{\Phi}_d(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d), \boldsymbol{R}), \quad (u_1, \ldots, u_d) \in [0,1]^d,$$

where $\boldsymbol{\Phi}_d$ is the CDF of the d-dimensional multivariate Gaussian distribution and Φ^{-1} is the quantile function of the univariate Gaussian distribution (Mai and Scherer, 2017).

In applied statistical modeling, copulae enable the construction of flexible multivariate distributions from known marginals. A prominent application are generalized joint regression models (GJRMs; Radice and Marra, 2016; Filippou et al., 2017; Marra and Radice, 2017; Klein et al., 2019), which provide regression methods to simultaneously model dependent multivariate response variables by explicitly incorporating the dependencies. GJRMs will be discussed in Article 1, i.e. in Section 3.1.

2.3 Survival analysis

The second part of this thesis discusses multidimensionality in the context of hazard regression models, which are specifically developed for time-to-event data. This type of data, also known as survival data, plays a crucial role in various research areas, particularly in medical research. The primary interest is the time until an event of interest occurs, which is often the participant's death (hence, the name survival data) but could also be any other event, e.g. disease recurrence or full recovery (Klein and Moeschberger, 2005). As any other continuous distribution, the distribution of continuous event times t can be characterized by the CDF $F(t)$, which can be interpreted as the probability that the event has occurred by time t, or the density $f(t)$, which gives the instantaneous likelihood of the event occurring at time t. However, the distribution of continuous event times is frequently expressed in terms of the survival function $S(t)$, the hazard function $\lambda(t)$ or the cumulative hazard function $\Lambda(t)$, rather than in terms of the CDF or the density (Klein and Moeschberger, 2005). The survival function

$$S(t) = P(T > t) = 1 - F(t)$$

represents the probability that a subject survives beyond a given time t. It is monotonically decreasing and starts at $S(0) = 1$. The hazard function

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{6}$$

represents the instantaneous rate of occurrence of the event at a given time t, conditional on survival up to that time. The cumulative hazard function

$$\Lambda(t) = \int_0^t \lambda(u)\, du$$

provides a measure of total risk accumulated over time, where an increasing cumulative hazard leads to exponentially decreases survival, i.e. $S(t) = e^{-H(t)}$.

Unlike other data types, time-to-event data is unique as it often includes censored observations, meaning that for some subjects the event is not observed or the exact time of the event is unknown. The three primary types of censoring are right-censoring, left-censoring, and interval-censoring (Klein and Moeschberger, 2005). Right-censoring is the most common type and results from subjects who do not experience an event before either leaving the study or the study ends (Collett, 2015). Left-censoring occurs when the event of interest already happened before the subject enters the study but the exact timing is unknown (Kalbfleisch and Prentice, 2011). Interval-censoring results from events for which only a time interval in which the event happened but not the exact time is observed, often due to event types that can only be observed by a medical examination or a laboratory test (Sun, 2006).

The presence of censoring necessitates specialized statistical techniques to conduct valid inference. Most methods, e.g. the *Kaplan-Meier estimator* (Kaplan and Meier, 1958) or the *Cox proportional hazards model* (Cox, 1972), assume the censoring to be non-informative and random, which means that the censoring times result from some random variable and that the distribution of survival times provides no information about the distribution of censoring times and vice versa. However, methods for survival data that is subject to informative censoring have been developed as well (see, e.g. Ibrahim et al., 2014).

One of the major topics in survival analysis is analyzing the effect covariates have on survival time. This is often encountered by hazard regression models, which estimate the hazard function (6). By incorporating covariates, hazard regression models allow to assess the impact of various factors on the hazard rate and, hence, to identify significant effects. The widely used standard model is the Cox proportional hazards model (Cox, 1972), where the hazard rate of an observation $i \in \{1, ..., n\}$ with corresponding covariate vector $\boldsymbol{x}_i$ is given by

$$\lambda_i(t) = \lambda_0(t) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}),$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients. The assumption of proportionality of the hazards results from the model being strictly split into the time-dependent baseline hazard $\lambda_0(t)$ and the time-constant covariate effects $\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$. In addition, the Cox model assumes the covariate effects to be (exp-transformed) linear effects.

These strict assumptions are often not fulfilled in practice (Li et al., 2015; Jachno et al., 2019), which is often caused by non-proportional hazards, i.e. the covariates or their effects are time-dependent. In addition, assuming all effects to be linear might also be oversimplifying.

Therefore, several flexible extensions to the Cox model (see, e.g. Zucker and Karr, 1990; Murphy and Sen, 1991; Gray, 1992; Hess, 1994; Vaupel et al., 1979; Ripatti and Palmgren, 2000; Therneau et al., 2003) have been introduced. In addition, several authors (see, e.g. Kneib and Fahrmeir, 2007; Hennerfeind et al., 2006; Bender et al., 2018) propose new flexible hazard regression frameworks in order to capture models of the form

$$\lambda_i(t|\boldsymbol{x}_i) = \lambda_0(t) \exp\left(\sum_{k=1}^{K} f_k(\boldsymbol{x}_i, t)\right) = \exp\left(\tilde{\lambda}_0(t) + \sum_{k=1}^{K} f_k(\boldsymbol{x}_i, t)\right),$$

where $\tilde{\lambda}_0(t)$ is the log-baseline hazard and f_k can resemble different types of effects, including time-varying, non-linear, and interaction effects. A comprehensive discussion of these extensions and alternative frameworks will be given in Section 3.4. In contrast to other frameworks, which require specific inference techniques, Bender et al. (2018) propose *piecewise exponential additive mixed models* (PAMMs) whose parameters can be estimated using standard *generalized additive models* (Hastie and Tibshirani, 1986) inference.

2.4 Generalized additive models

Generalized additive models (GAMs; Hastie and Tibshirani, 1986) provide a flexible regression framework that is capable of modeling response variables other than continuous real-valued, e.g. binary or count data. This is achieved by relating the expected value of the response variable to explanatory variables through a link function g, analogous to generalized linear models (GLMs). For an observation $i \in \{1, ..., n\}$ with corresponding covariate vector $\boldsymbol{x}_i \in \mathbb{R}^P$ this results in the model

$$g(\mathbb{E}(y_i)) = \eta(\boldsymbol{x}_i) \quad \Leftrightarrow \quad \mathbb{E}(y_i) = h(\eta(\boldsymbol{x}_i)) \tag{7}$$

where $\eta(\boldsymbol{x}_i)$ is the predictor term and $h = g^{-1}$ is called the response function (Fahrmeir et al., 2022; Wood, 2017). Common link functions are for example the logit or probit function for binary responses or the logarithm for count data.

In contrast to GLMs, where $\eta(\boldsymbol{x}_i)$ is a linear predictor, i.e. $\eta(\boldsymbol{x}_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$), GAMs are able to account for complex covariable effects, such as non-linear, multivariate or random effects which is achieved by deploying an additive predictor term in (7), i.e.

$$\eta(\boldsymbol{x}_i) = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i),$$

consisting of $k = 1, ..., K$ effects terms $f_k(\boldsymbol{x}_i)$ each being potentially non-linear (Fahrmeir et al., 2022; Wood, 2017). Each effect can depend on any subset of the covariate vector $\boldsymbol{x}_i$ since one covariable can be contained in more than one effect (e.g. main and interaction effect) and one effect can consist of more than one variable (e.g. interaction effects).

In GAMs, covariable effects are often modeled as smooth functions, e.g. as B-splines, which allows for the discovery of patterns in data without imposing strict parametric forms, thereby also reducing the risk of model misspecification. Univariate smooth effects are usually implemented in terms of basis function expansion as

$$f_k(\boldsymbol{x}_i) = \sum_{d_k=1}^{D_k} \gamma_{kd_k} B_{kd_k}(x_{ip})$$

where x_{ip} is the i-th observation of the p-th covariable, $B_{kd_k}(x_{ip})$ are the basis functions, γ_{kd_k} are the basis coefficients, and D_k is the corresponding dimension (Fahrmeir et al., 2022; Wood, 2017). Based on this basis function expansion, multivariate effects can be construed as tensor product interactions (Kneib et al., 2019; Wood et al., 2013). Smooth effects, basis function expansion, and tensor product interactions will be discussed in more detail in Articles 3 and 4, i.e. in Sections 3.3 and 3.4.

Since the high flexibility can also lead to overfitting, usually penalized approaches, e.g. *penalized B-splines* (*P-splines*; Eilers and Marx, 1996; Lang and Brezger, 2004), are deployed. In frequentist estimation, the penalization is usually implemented in terms of first or second order differences (Eilers and Marx, 1996). For Bayesian approaches, the corresponding analogs are first or second order random walk priors (Lang and Brezger, 2004). Bayesian penalization is introduced in more detail in Section 3.3 while frequentist penalization is extensively discussed in Section 3.4. In both cases, the extent of penalization is controlled by a *smoothing parameter*.

In contrast to Bayesian approaches (e.g. Kneib et al., 2019), where for each smoothing parameter a (hyper-) prior distribution is assigned and it is then estimated naturally during the Markov Chain Monte Carlo estimation, their estimation is a challenging task in frequentist inference. There are essentially three different approaches to this:

1. A smoothness selection criteria, e.g. the Akaike information criterion (AIC) or, a little more sophisticated, a generalized cross-validation can be deployed. Choosing the optimal smoothing parameter is then either based on a grid search or a nested iterative procedure. For more details see Section 6.2 of Wood (2017).

2. A mixed-model representation can be chosen where the smooth components are estimated as if they were random effects in a generalized linear mixed model (GLMM). The inference can then be conducted based on well-known restricted maximum likelihood methods for GLMMs. For more details see Section 6.8 of Wood (2017) and Section 9.6.2 of Fahrmeir et al. (2022).

3. Wood (2011) proposes a more direct method that avoids the formal mixed-model framework while still utilizing the random-effects perspective. The smoothing parameters are estimated directly from the restricted likelihood function without requiring the specification of a full mixed-model structure. This is achieved by using a direct Laplace approximation that integrates out the random effects, i.e. the spline coefficients. Hence, this method optimizes a well-defined likelihood function directly with respect to the smoothing parameters. Therefore, it bypasses the need to solve the mixed-model equations.

Section 3.3 extensively discusses Bayesian estimation using Hamiltonian Monte Carlo simulation (Hoffman and Gelman, 2014; Duane et al., 1987). Frequentist estimation deploying the method of Wood (2011) is presented in Section 3.4.

3 Articles

This section addresses the three identified research gaps, each resulting from one of the three directions of multidimensionality. Sections 3.1 - 3.4 present four articles that will develop the necessary adaptations. The first two articles focus on model-based equivalence tests, while the latter two discuss heterogeneously time-varying covariable effects and their application in hazard regression models.

Section 3.1 investigates the issues that arise when attempting to compare multidimensional response variables through a model-based equivalence test. In previous research, only the special case of bivariate binary outcomes has been addressed by Möllenhoff et al. (2021). However, their approach does not directly generalize to other outcome distributions. Section 3.1 introduces a more flexible approach based on generalized joint regression models, allowing for other scales of measures of the outcome variables, including mixed outcomes. This approach also ensures direct generalizability for outcomes with more than two dimensions. In contrast to the approach of Möllenhoff et al. (2021), whose test statistic relies on the intersection-union principle, an alternative test statistic, called maximum of maxima, is used, which resolves the problem of the test being overly conservative for smaller sample sizes.

The applicability of model-based equivalence tests in the presence of model uncertainty is discussed in Section 3.2. The test of Dette et al. (2018) as well as all methods based thereon assume the true underlying regression model to be known. However, in applied research often only a set of plausible models, also called candidate models, is known. Section 3.2 proposes to overcome the model uncertainty by introducing model averaging to the test of Dette et al. (2018). This prevents model misspecification, which can otherwise invalidate the test.

Section 3.3 focuses on heterogeneously time-varying covariable effects, which occur if the effect of a covariable is subgroup-specific, time-varying and its time-variation is also subgroup-specific. To model such effects, functional random coefficients based on tensor product interactions are proposed. The subsequent section, Section 3.4, discusses the applicability of these functional random coefficients in survival analysis by incorporating such effects into hazard regression models.

3.1 Testing for similarity of multivariate mixed outcomes using generalized joint regression models with application to efficacy-toxicity responses

In this article, the adaptation of model-based equivalence tests for the first direction of multidimensionality, i.e. for multivariate potentially mixed-scale outcomes, is discussed. An approach based on the generalized joint regression framework exploiting the Gaussian copula is introduced. Compared to existing methods, this approach accommodates various outcome variable scales including mixed outcomes in multi-dimensional spaces. Finite sample properties are investigated through a simulation study and an efficacy-toxicity case study highlights the practical relevance.

Authorship:	First author
Coauthors:	Giampiero Marra, Frank Bretz and Kathrin Möllenhoff
Contribution statement:	My contribution to this project was developing the method, programming it in R, implementing the simulation study, analyzing the case study data, preparing the figures and tables, and writing the first draft of the manuscript. The initial idea for this work came from Kathrin Möllenhoff, who also supported the development of the method. The programming was supported by Giampierro Marra to a minor extent. Frank Bretz provided the case study dataset and supported its analysis. Discussions and revising were done together with all coauthors.
Status:	Published
Journal:	Biometrics
DOI:	10.1093/biomtc/ujae077

Biometrics, 2024, 80(3), ujae077
https://doi.org/10.1093/biomtc/ujae077
Biometric Methodology

OXFORD

Testing for similarity of multivariate mixed outcomes using generalized joint regression models with application to efficacy-toxicity responses

Niklas Hagemann [1,2], Giampiero Marra [3], Frank Bretz [4,5], Kathrin Möllenhoff [2,*]

[1]Mathematical Institute, Heinrich Heine University Düsseldorf, Düsseldorf, 40225, Germany, [2]Institute of Medical Statistics and Computational Biology (IMSB), Faculty of Medicine, University of Cologne, Cologne, 50923, Germany, [3]Department of Statistical Science, University College London, London, WC1E 6BT, United Kingdom, [4]Statistical Methodology, Novartis Pharma AG, Basel, 4056, Switzerland, [5]Section for Medical Statistics, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, 1090, Austria

[*]Corresponding author: Kathrin Möllenhoff, Institute of Medical Statistics and Computational Biology (IMSB), Faculty of Medicine, University of Cologne, Cologne, 50923, Germany (kathrin.moellenhoff@uni-koeln.de).

ABSTRACT

A common problem in clinical trials is to test whether the effect of an explanatory variable on a response of interest is similar between two groups, for example, patient or treatment groups. In this regard, similarity is defined as equivalence up to a pre-specified threshold that denotes an acceptable deviation between the two groups. This issue is typically tackled by assessing if the explanatory variable's effect on the response is similar. This assessment is based on, for example, confidence intervals of differences or a suitable distance between two parametric regression models. Typically, these approaches build on the assumption of a univariate continuous or binary outcome variable. However, multivariate outcomes, especially beyond the case of bivariate binary responses, remain underexplored. This paper introduces an approach based on a generalized joint regression framework exploiting the Gaussian copula. Compared to existing methods, our approach accommodates various outcome variable scales, such as continuous, binary, categorical, and ordinal, including mixed outcomes in multi-dimensional spaces. We demonstrate the validity of this approach through a simulation study and an efficacy-toxicity case study, hence highlighting its practical relevance.

KEYWORDS: bootstrap; copula; dose-response models; model-based equivalence tests.

1 INTRODUCTION

A common challenge in applied research, especially in clinical trials, is determining whether an explanatory variable's effect on a response variable is equivalent or similar across different groups (see, eg, Jhee et al., 2004; Otto et al., 2008). In this context, similarity is defined as equivalence up to a *similarity threshold* value. Equivalence tests are widely used in various fields, particularly to determine if a treatment has comparable effects in different groups, based, for instance, on gender, age, or treatments, just to mention a few. Moreover, they are commonly used to investigate whether two formulations of a drug have nearly the same effect and are hence considered to be interchangeable, the key question of bioequivalence studies (eg, Möllenhoff et al., 2022).

One usually assesses the question of similarity by testing whether the (marginal) effects of covariates on a response variable are similar among the groups, either based on confidence interval inclusion (Liu et al., 2009; Gsteiger et al., 2011; Bretz et al., 2018) or using various distance measures as test statistics (Dette et al., 2018; Möllenhoff et al., 2018). These approaches assume a univariate continuous outcome variable, which, as outlined by Möllenhoff et al. (2021), might not be appropriate in many applications. On the one hand, the outcome might be, for example, binary, categorical, or ordinal. On the other hand, multivariate (often bivariate) outcomes arise, such as when analyzing the efficacy and toxicity of a drug (eg, Jhee et al., 2004), which cannot be assumed to be independent of each other and, therefore, need to be modeled jointly.

There are different approaches to jointly model multivariate outcome variables based on copulae (Sklar, 1959). Tao et al. (2013) proposed to use Archimedean copulae for such models, while Möllenhoff et al. (2021) suggested the Gumbel model (Murtaugh and Fisher, 1990; Heise and Myers, 1996) based on the Farlie-Gumbel-Morgenstern copula, which also belongs to the class of Archimedean copulae. In contrast, other authors employed elliptical copulae, especially the Gaussian, which was adopted by de Leon and Wu (2011) for regression models with bivariate mixed outcomes, and Chiu and Crump (2012) for bivariate binary outcomes. The Gaussian copula is rather flexible for practical modeling: although it assumes linear dependence, it easily generalizes to more than two dimensions and neatly characterizes multivariate dependence through the covariance matrix (Joe, 2015). It also makes it possible to combine several types of variables (eg, continuous, binary, categorical, continuous non-negative, and ordinal) following various distributions. In different applied contexts, Radice and Marra (2016) introduced bivariate models with binary margins, which were

Received: January 18, 2024; Revised: June 5, 2024; Accepted: July 24, 2024

then generalized by Filippou et al. (2017) to the multivariate (specifically, trivariate) case. Marra and Radice (2017) introduced bivariate copula models with continuous margins and Klein et al. (2019) additionally developed models for mixed responses (binary and continuous). The aforementioned models belong to the class of *generalized joint regression models* implemented in the R-package GJRM (Marra and Radice, 2023). Note that GJRM allows for many more modeling options than those mentioned here (eg, Marra and Radice, 2020; Marra et al., 2020).

To generalize a distance-based similarity test for associated bivariate binary responses, Möllenhoff et al. (2021) adopted a copula approach to jointly model the efficacy and toxicity of a drug. However, the method proposed in this paper is more flexible, allowing for arbitrary dimensions and various types (including mixed) of outcome variables. The proposal is based on the generalized joint regression framework with Gaussian copula and:

- can be applied to multivariate responses of any size;
- accommodates various outcome types, including continuous, binary, and ordinal;
- adopts another type of test statistic that leads to a higher statistical power; and
- addresses the problem of increasing type I error rates with increasing sample size, observed by Möllenhoff et al. (2021).

The paper is structured as follows: In Section 2, the modeling framework, based on generalized joint regression models with Gaussian copula, is succinctly discussed. In Section 3, the new testing approach is introduced. Type I and II error rates for three relevant applied cases (bivariate binary, bivariate continuous, and bivariate mixed outcomes) are studied in Section 4. Section 5 illustrates the method using clinical trial data.

2 COPULA REGRESSION MODELS

2.1 Regression structures

Let $i = 1, \ldots, n$ be the observation index, where n denotes the sample size. For a univariate outcome, the adopted modeling approach relies on the flexible regression structure

$$\mu_i = m(x_i, \boldsymbol{\theta}),$$

where $\mu_i = \mathbb{E}(y_i)$, $y_i \in \mathcal{Y} \subseteq \mathbb{R}$ denotes the response variable within the set $\mathcal{Y}$ of all possible outcomes, $x_i \in \mathcal{X} \subseteq \mathbb{R}$ is a deterministic explanatory variable, $m(\cdot)$ is a function modeling the effect of x_i on y_i via a regression curve, and $\boldsymbol{\theta} \in \mathbb{R}^{\dim(\boldsymbol{\theta})}$ the related parameter vector. The function $m(\cdot)$ can be linear or nonlinear, as illustrated in later sections. We assume that $m(\cdot)$ is continuous, consequently resulting in continuous distances among model curves.

In the following, we are interested in comparing the effect of the variable x_i on y_i for two separate groups. This requires an additional group index $l = 1, 2$. Consequently, we observe outcomes $y_i^{(l)}$, $i = 1, \ldots, n^{(l)}$, $l = 1, 2$, and regression curves

$$\mu_i^{(1)} = m^{(1)}(x_i^{(1)}, \boldsymbol{\theta}^{(1)}) \quad \text{and} \quad \mu_i^{(2)} = m^{(2)}(x_i^{(2)}, \boldsymbol{\theta}^{(2)}).$$

For a multivariate outcome $\boldsymbol{y}_i^{(l)} = (y_{i1}^{(l)}, \ldots, y_{iK}^{(l)})$, this generalizes to

$$\begin{aligned} \boldsymbol{\mu}_i^{(l)} &= \boldsymbol{m}^{(l)}(x_i^{(l)}, \boldsymbol{\theta}^{(l)}) \\ \Leftrightarrow \begin{pmatrix} \mu_{i1}^{(l)} \\ \vdots \\ \mu_{iK}^{(l)} \end{pmatrix} &= \begin{pmatrix} m_1^{(l)}(x_i^{(l)}, \boldsymbol{\theta}_1^{(l)}) \\ \vdots \\ m_K^{(l)}(x_i^{(l)}, \boldsymbol{\theta}_K^{(l)}) \end{pmatrix}, \quad l = 1, 2, \end{aligned}$$

with outcome dimension $K \in \mathbb{N}^+$ and group index l. In dose-response studies with efficacy-toxicity outcomes, we have that $K = 2$, the outcomes $y_{i1}^{(l)}$ and $y_{i2}^{(l)}$ express the efficacy and the toxicity, respectively, and the explanatory variable $x_i^{(l)}$ describes the dose for patient i in group l, $i = 1, \ldots, n^{(l)}$, $l = 1, 2$. We thus have regression structures $m_1^{(1)}(x_i^{(1)}, \boldsymbol{\theta}_1^{(1)})$ and $m_2^{(1)}(x_i^{(1)}, \boldsymbol{\theta}_2^{(1)})$ modeling the efficacy and toxicity for group 1, and $m_1^{(2)}(x_i^{(2)}, \boldsymbol{\theta}_1^{(2)})$ and $m_2^{(2)}(x_i^{(2)}, \boldsymbol{\theta}_2^{(2)})$ for group 2, respectively.

In general, since the K responses are assumed to be dependent, the models have to be estimated jointly as described in the following section.

2.2 Copulae

Copulae can be used to characterize the multivariate distribution of the response variables $y_{i1}^{(l)}, \ldots, y_{iK}^{(l)}$, $l = 1, 2$. Specifically, for a K-dimensional distribution with cumulative distribution function (cdf) $\boldsymbol{F}$ and univariate marginal cdfs $F_1, \ldots, F_K$ following uniforms on [0,1], the copula $C : [0, 1]^K \to [0, 1]$ and $\boldsymbol{F}$ are linked as follows (Sklar, 1959)

$$\boldsymbol{F}(y_1, \ldots, y_k) = C\{F_1(y_1), \ldots, F_K(y_K)\},$$

where group index l has been omitted for simplicity. We refer to Joe (2015) and Trivedi and Zimmer (2007), for comprehensive introductions to copulae.

Commonly used classes of copulae include the Archimedean copulae (which encompasses the Gumbel, Frank, and Clayton) and the meta-elliptical copulae (which includes the Gaussian). The choice of copula often depends on the specific application and its modeling requirements. For the purpose of the present work, we adopt the Gaussian copula, which, as mentioned in the Introduction, offers the required flexibility and generality in modeling the multivariate dependence structure of the variables of interest (Joe, 2015). The Gaussian copula can be generically defined as

$$\begin{aligned} \boldsymbol{F}(y_{i1}, \ldots, y_{iK}) &= C\{F_1(y_{i1}), \ldots, F_K(y_{iK})\} \\ &= \Phi_K[\Phi^{-1}\{F_1(y_{i1})\}, \ldots, \Phi^{-1}\{F_K(y_{iK})\}, \boldsymbol{\Gamma}], \end{aligned}$$

or, in the bivariate case, as

$$\boldsymbol{F}(y_{i1}, y_{i2}) = \Phi_2[\Phi^{-1}\{F_1(y_{i1})\}, \Phi^{-1}\{F_2(y_{i2})\}, \rho],$$

where Φ_K is the cdf of the K dimensional multivariate Gaussian distribution, Φ^{-1} is the quantile function of the univariate Gaussian, $F_1(y_{i1}), \ldots, F_K(y_{iK})$ are the cdfs of the marginal distributions, $\boldsymbol{\Gamma} = \mathbf{Cor}(y_{i1}, \ldots, y_{iK})$ and $\rho = \mathrm{Cor}(y_{i1}, y_{i2})$. Note that in the above, we have suppressed $\boldsymbol{\theta}_k$ from the related marginal cdf for notational convenience.

2.3 Log-likelihoods

The structure of the log-likelihood function to employ in model fitting depends on the marginals considered in the analysis. The general likelihood theory of the K-dimensional case is given by Song et al. (2009). However, for simplicity of exposition, we report the functions considered in the applied cases of this paper: bivariate binary, bivariate continuous, and bivariate mixed outcomes. For the same reason, we drop index l and $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ from the marginal cdfs.

The log-likelihood for bivariate continuous outcomes is given by (Marra and Radice, 2017)

$$\ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho) = \sum_{i=1}^{n} (\log[c\{F_1(y_{i1}), F_2(y_{i2})\}] + \log\{f_1(y_{i1})\} + \log\{f_2(y_{i2})\}),$$

where the copula density is defined as

$$c\{F_1(y_{i1}), F_2(y_{i2})\} = \frac{\partial^2 C\{F_1(y_{i1}), F_2(y_{i2})\}}{\partial F_1(y_{i1}) \partial F_2(y_{i2})}$$

and f_1 and f_2 are marginal densities. For bivariate mixed outcomes, with y_{i1} binary and y_{i2} continuous, the log-likelihood is (Klein et al., 2019)

$$\ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho) = \sum_{i=1}^{n} \big[(1 - y_{i1}) \log\{F_{1|2}(0|y_{i2})\} + y_{i1} \log\{1 - F_{1|2}(0|y_{i2})\} + \log\{f_2(y_{i2})\}\big],$$

where

$$F_{1|2}(0|y_{i2}) = \frac{\partial C\{F_1(0), F_2(y_{i2})\}}{\partial F_2(y_{i2})}$$

and $F_1(0) = \mathbb{P}(y_{i1} = 0)$.

When both outcomes are binary, the log-likelihood is (Radice and Marra, 2016)

$$\ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \rho) = \sum_{i=1}^{n} \{y_{i1}y_{i2} \log p_{11i} + y_{i1}(1 - y_{i2}) \log p_{10i} + (1 - y_{i1})y_{i2} \log p_{01i} + (1 - y_{i1})(1 - y_{i2}) \log p_{00i}\},$$

where

$$p_{11i} = \mathbb{P}(y_{i1} = 1, y_{i2} = 1) = C(\mathbb{P}(y_{i1} = 1), \mathbb{P}(y_{i2} = 1)),$$
$$p_{10i} = \mathbb{P}(y_{i1} = 1) - \mathbb{P}(y_{i1} = 1, y_{i2} = 1),$$
$$p_{01i} = \mathbb{P}(y_{i2} = 1) - \mathbb{P}(y_{i1} = 1, y_{i2} = 1) \text{ and}$$
$$p_{00i} = 1 - \{\mathbb{P}(y_{i1} = 1) + \mathbb{P}(y_{i2} = 1) - \mathbb{P}(y_{i1} = 1, y_{i2} = 1)\}.$$

As explained earlier, for this work, we specify function C using the Gaussian copula. Regarding the margins, we employ the Bernoulli distribution (with logit, probit, or c-log link) when the outcome is binary, whereas the normal, logistic, or another distribution can be utilized for a continuous response.

We achieve model fitting via the R-package GJRM (Marra and Radice, 2023) whose parameter estimation is based on an efficient and stable implementation of the trust region algorithm.

3 TESTING FOR SIMILARITY OF MULTIVARIATE NON-INDEPENDENT RESPONSES

3.1 Hypotheses

One approach to assess similarity of two curves, $m^{(1)}$ and $m^{(2)}$ in the univariate case is based on the maximum absolute deviation between them. In this case, the hypotheses are

$$H_0 : \max_{x\in\mathcal{X}} |m^{(1)}(x, \boldsymbol{\theta}^{(1)}) - m^{(2)}(x, \boldsymbol{\theta}^{(2)})| \geq \varepsilon \quad \text{vs.}$$
$$H_1 : \max_{x\in\mathcal{X}} |m^{(1)}(x, \boldsymbol{\theta}^{(1)}) - m^{(2)}(x, \boldsymbol{\theta}^{(2)})| < \varepsilon, \tag{1}$$

where ε is a prespecified threshold for similarity (Dette et al., 2018; Möllenhoff et al., 2021). Rejecting the null hypothesis suggests that, for a given significance level, the curves are similar since their distance is lower than the threshold value.

For multivariate responses, there are several possibilities to formulate hypotheses for (joint) similarity. An intuitive approach, which generalizes the approach of Möllenhoff et al. (2021) for the bivariate case, would be testing for (joint) similarity of all the curves associated with the K outcomes. Formally, this leads to testing the hypotheses

$$H_0 : d_k \geq \varepsilon_k \text{ for at least one } k \in \{1, \ldots, K\} \quad \text{vs.}$$
$$H_1 : d_k < \varepsilon_k \;\forall\; k \in \{1, \ldots, K\}, \tag{2}$$

where

$$d_k = \max_{x\in\mathcal{X}} |m_k^{(1)}(x, \boldsymbol{\theta}_k^{(1)}) - m_k^{(2)}(x, \boldsymbol{\theta}_k^{(2)})|, \; k = 1, \ldots, K,$$

denotes the maximum absolute deviation between the curves describing the kth response and ε_k the corresponding similarity threshold. Since the alternative hypothesis stated in Equation 2 is expressed as an intersection of sub-hypotheses of similarity for each of the K outcomes, Möllenhoff et al. (2021) suggested to test all of these K sub-hypotheses

$$\begin{array}{ccc} H_0^{(1)} : d_1 \geq \varepsilon_1 & \text{vs.} & H_1 : d_1 < \varepsilon_1 \\ & \vdots & \\ H_0^{(K)} : d_K \geq \varepsilon_K & \text{vs.} & H_1 : d_K < \varepsilon_K \end{array} \tag{3}$$

individually. According to the intersection union principle (Berger, 1982), the global null hypothesis in Equation 2 is then rejected if all of these individual hypotheses are rejected. This procedure guarantees an α-level test if all individual tests are of size α. However, such an approach is known to be quite conservative, especially for small sample sizes or a large number of individual tests, ie for a large K (Möllenhoff et al., 2021).

To this end, we propose an alternative testing procedure, which we will call the *maximum of maxima* approach. This is based on a different type of test statistic, that is $d_{max} := \max(d_1, \ldots, d_K)$. The basic idea of this approach is that if the largest difference is sufficiently small then all the other differences are small enough too. This approach is similar to the one used by Möllenhoff et al. (2024) to jointly test for similarity of more than one transition intensity in a competing risks model. The corresponding hypotheses are then given by

$$H_0 : d_{max} \geq \varepsilon \quad \text{vs.} \quad H_1 : d_{max} < \varepsilon, \tag{4}$$

where ε now represents a global similarity threshold, denoted by $\varepsilon = \varepsilon_1 = \ldots = \varepsilon_K$. The necessity of a global ε is due to the construction of the new test statistic, which bundles the maximum distances $d_1, \ldots, d_K$ into one single value d via taking their maximum. In general, the individual thresholds ε_k, $k = 1, \ldots, K$, may vary across the K outcomes.

However, by adding a data transformation step to the analysis, it is still possible to incorporate unequal individual thresholds in many cases. For continuous outcomes, one can achieve this, for example, by linear rescaling. For $K = 2$, an example of this is given by $y_1^{(l)}$ being binary, $y_2^{(l)}$ being continuous and measured in milligrams (mg), $\varepsilon_1 = 0.1$ on a probability scale and $\varepsilon_2 = 0.02$ mg. We set $\varepsilon = \varepsilon_1$ and rescale $y_2^{(l)}$ by multiplying its values with $\frac{\varepsilon_1}{\varepsilon_2} = \frac{0.1}{0.02} = 5$. Correspondingly, its units change by a factor of $\frac{\varepsilon_2}{\varepsilon_1} = \frac{1}{5}$. As a result, $\varepsilon = \varepsilon_1 = \varepsilon_2 = 0.1$ and $y_2^{(l)}$ is now measured in fifths of milligrams $(\frac{\text{mg}}{5})$. This procedure directly generalizes to the K-dimensional case as long as at most one response variable is non-continuous.

3.2 Testing procedure

To test the hypotheses in Equation 4, we propose a parametric bootstrap approach similar to Algorithm 1 of Möllenhoff et al. (2021) and to the method of Dette et al. (2018). Following these papers, we approximate the distribution under H_0 rather than obtaining confidence intervals for d_{max} in order to make use of the asymptotic properties outlined at the end of this section.

Conducting parametric bootstrap requires the simulation of multivariate correlated outcomes. For multivariate binary outcomes, data generation is based on the algorithm of Emrich and Piedmonte (1991). For the continuous case, one can just sample from a multivariate normal distribution. For multivariate mixed outcomes, we employ the algorithm of Demirtas and Doganay (2012).

The test presented in Algorithm 1 has an asymptotic level α and is consistent. That is, under the null hypothesis in Equation 4, $\limsup_{n\to\infty} \mathbb{P}\big(\widehat{d}_{max} \le \widehat{d}^*_{max,(\lfloor n_{boot}\alpha\rfloor)}\big) \le \alpha$ and, under the alternative, $\lim_{n\to\infty} \mathbb{P}\big(\widehat{d}_{max} \le \widehat{d}^*_{max,(\lfloor n_{boot}\alpha\rfloor)}\big) = 1$ for any $\alpha \in (0, 0.5)$. A formal proof of this can be directly obtained by transferring the proof given in Möllenhoff et al. (2024), who investigate the same type of test statistic. Precisely, it is based on the fact that the MLE $\widehat{\boldsymbol{\theta}}_k^{(l)}$, $k = 1, \ldots, K, l = 1, 2$, obtained by maximizing the log-likelihoods given in Section 2.3, converge weakly to a normal distribution, such that the proof of the general procedure of a constrained bootstrap given in Dette et al. (2018) can be adapted as described by Möllenhoff et al. (2024). We investigate these properties for finite sample sizes in the following section.

4 FINITE SAMPLE PROPERTIES

In this section, we investigate type I error rates and power of the proposed approach using Algorithm 1. Therefore, we simulate bivariate efficacy-toxicity outcomes as a function of dose, modeled by dose-response curves.

For comparability with existing studies, particularly those in Möllenhoff et al. (2021), the simulation setup for the bivariate

Algorithm 1:

(1) Obtain, via MLE, $\widehat{\boldsymbol{\theta}}_k^{(l)}$, $l = 1, 2,\ k = 1, \ldots, K$, by maximizing for each group the relevant log-likelihood (see Section 2.3). The test statistic is calculated as

$$\widehat{d}_{max} = \max(\widehat{d}_1, \ldots, \widehat{d}_K),$$

where

$$\widehat{d}_k = \max_{x\in\mathcal{X}} |m_k^{(1)}(x, \widehat{\boldsymbol{\theta}}_k^{(1)}) - m_k^{(2)}(x, \widehat{\boldsymbol{\theta}}_k^{(2)})|,$$

$$k = 1, \ldots, K.$$

(2) To approximate the null distribution, define estimators for parameter vectors $\boldsymbol{\theta}_k^{(l)}$, $l = 1, 2,\ k = 1, \ldots, K$, so that the corresponding curves fulfill the null hypothesis in Equation 4. That is,

$$\widehat{\widehat{\boldsymbol{\theta}}}_k^{(l)} = \begin{cases} \widehat{\boldsymbol{\theta}}_k^{(l)} & \text{if} \quad \widehat{d}_{max} \ge \epsilon \\ \overline{\boldsymbol{\theta}}_k^{(l)} & \text{if} \quad \widehat{d}_{max} < \epsilon \end{cases} \quad l = 1, 2,\ k = 1, \ldots, K,$$

where $\overline{\boldsymbol{\theta}}_k^{(l)}$ maximises the same objective function as $\widehat{\boldsymbol{\theta}}_k^{(l)}$, $l = 1, 2,\ k = 1, \ldots, K$ does, but under the constraint

$$d_{max} = \epsilon. \tag{5}$$

Technically, we discretize the range, $\mathcal{X}$, of the explanatory variable to make the optimization feasible. The constrained problem is solved using the augmented Lagrangian minimization algorithm via function `auglag()` in the R package alabama (Varadhan, 2022).

(3) Execute the following steps:

(a) Obtain bootstrap samples under the null hypothesis in Equation 4 by generating data according to the model parameters $\widehat{\widehat{\boldsymbol{\theta}}}_k^{(l)}$, $l = 1, 2,\ k = 1, \ldots, K$. This is achieved by obtaining parameter estimates for the marginal distributions and correlations and then feeding them into the data generation algorithms introduced above.

(b) From the bootstrap samples, calculate the MLE $\widehat{\boldsymbol{\theta}}_k^{(l)*}$ as in step (1) and the test statistic

$$\widehat{d}^*_{max} = \max(\widehat{d}^*_1, \ldots, \widehat{d}^*_K), \tag{6}$$

where

$$\widehat{d}^*_k = \max_{x\in\mathcal{X}} |m_k^{(1)}(x, \widehat{\boldsymbol{\theta}}_k^{(1)*}) - m_k^{(2)}(x, \widehat{\boldsymbol{\theta}}_k^{(2)*})|,$$

$$k = 1, \ldots, K.$$

(c) Repeat steps (a) and (b) n_{boot} times to generate replicates $\widehat{d}^*_{max,1}, \ldots, \widehat{d}^*_{max,(n_{boot})}$ of $\widehat{d}^*_{max}$. Let $\widehat{d}^*_{max,(1)} \le \ldots \le \widehat{d}^*_{max,(n_{boot})}$ denote the corresponding order statistic. The estimator of the α-quantile of the distribution of $\widehat{d}^*_{max}$ is given by $\widehat{d}^*_{max,(\lfloor n_{boot}\alpha\rfloor)}$.

Reject the null hypothesis in (4) and assess similarity based on

$$\widehat{d}_{max} < \widehat{d}^*_{max,(\lfloor n_{boot}\alpha\rfloor)}. \tag{7}$$

```
Algorithm 1: Continued.
        Alternatively, obtain the $p$-value based on
        $\widehat{F}_{n_{boot}}(\widehat{d}_{max}) = \frac{1}{n_{boot}} \sum_{i=1}^{n_{boot}} \mathbb{1}(\widehat{d}^*_{max,i} \leq \widehat{d}_{max})$ and reject
        the null hypothesis in (4) if $\widehat{F}_{n_{boot}}(\widehat{d}_{max}) < \alpha$ for a
        pre-specified significance level $\alpha$, where $\widehat{F}_{n_{boot}}$ denotes
        the empirical cumulative distribution function of the
        bootstrap sample.
```

binary case closely follows their scenarios. This includes the data generation routine and the levels of the explanatory variable $x_i^{(l)}$, set at specific dose levels 0, 0.1, 0.2, 0.5, 1, 1.5 and 2. The simulation involves seven dose groups ($g = 1, \ldots, 7$), with equal sample sizes $n_g^{(l)} \in \{7, 14, 21, 28, 50\}$, resulting in total sample sizes from 49 to 350 per group. Note that we generate the data for each of the seven dose levels separately due to the algorithms' limitations in handling varying marginal probabilities/means, as detailed in Section 3.2.

We keep the correlation parameter ρ constant within each group g, leading to a different global correlation of the combined data (see Dunlap, 1937, for details of the relationship between group-wise and global correlations). Furthermore, we assume $\rho^{(1)} = \rho^{(2)} = \rho$, and employ the two different similarity threshold values, $\varepsilon = 0.15$ and 0.2, introduced by Möllenhoff et al. (2021). Given the computational costs related to the augmented Lagrangian minimization algorithm and the data generation process, the study comprises 1000 simulation replicates and 300 bootstrap repetitions.

4.1 Bivariate binary outcome

We adopt the same configurations as Möllenhoff et al. (2021), employing Bernoulli marginals with logit links for both efficacy and toxicity, and

$$\boldsymbol{\theta}^{(l)} = (\boldsymbol{\theta}_1^{(l)}, \boldsymbol{\theta}_2^{(l)}, \rho) = (\beta_{01}^{(l)}, \beta_{11}^{(l)}, \beta_{02}^{(l)}, \beta_{12}^{(l)}, \rho), \quad l = 1, 2,$$

where $\boldsymbol{\theta}^{(1)} = (-1, 2, -3, 3, \rho)$. To simulate type I error rates, we investigate $(d_1, d_2) \in \{(\varepsilon, \varepsilon), (0, \varepsilon)\}$ for both $\varepsilon = 0.15$ and $\varepsilon = 0.2$, hence leading to four scenarios. Regarding the power, we investigate the three scenarios $(d_1, d_2) = (0.1, 0.1)$, $(0.05, 0.05)$ and $(0, 0)$. The latter choice simulates the maximum power of the testing approach. The exact details of parameter combinations considered are shown in Web Table 1. Finally, we investigate three different levels of group-wise correlations, $\rho = 0.1$, 0.2 and 0.3.

Table 1 shows the simulated type I error rates of the test implemented via Algorithm 1. We observe that, for $d_1 = d_2 \approx \varepsilon$, type I error rates are well below or very close to the significance level of $\alpha = 0.05$. For the scenarios with $\min(d_1, d_2) = 0$, we observe slightly inflated type I error rates, up to a maximum of 0.106 for the smallest group size of $n_g^{(l)} = 7$ and $\varepsilon = 0.2$. However, as the group size increases, the type I error rates decrease and approach the desired level of 0.05. Of note, the value of ρ does not seem to be that influential in this regard.

In comparison to Möllenhoff et al. (2021), where type I error rates were predominantly close to zero, the results for our approach align more closely with the nominal level. This is in line with the theoretical arguments of Section 3.1: the proposal is less conservative compared to testing based on the intersection union principle. However, for some configurations with high correlation, Möllenhoff et al. (2021) observed an inflation of the type I error rates as the sample size increased up to a value of 12.7%. In contrast, the type I error rates decrease for increasing sample sizes when using our approach. In addition, the maximum type I error rate is 10.6% for our approach, which is considerably smaller than the 12.7% observed in Möllenhoff

TABLE 1 Simulated type I error rates of the test proposed in Algorithm 1 for bivariate binary outcomes and two different similarity thresholds ε.

ε	$\boldsymbol{\theta}^{(2)}$	$(\boldsymbol{d_1}, \boldsymbol{d_2})$	$\boldsymbol{n_g^{(l)}}$	$\boldsymbol{\rho = 0.1}$	$\boldsymbol{\rho = 0.2}$	$\boldsymbol{\rho = 0.3}$
0.2	$(-2.4, 3.4, -1.8, 2.51, \rho)$	(0.2, 0.2)	7	0.031	0.037	0.038
0.2	$(-2.4, 3.4, -1.8, 2.51, \rho)$	(0.2, 0.2)	14	0.012	0.012	0.018
0.2	$(-2.4, 3.4, -1.8, 2.51, \rho)$	(0.2, 0.2)	21	0.013	0.006	0.012
0.2	$(-2.4, 3.4, -1.8, 2.51, \rho)$	(0.2, 0.2)	28	0.007	0.006	0.005
0.2	$(-2.4, 3.4, -1.8, 2.51, \rho)$	(0.2, 0.2)	50	0.006	0.009	0.004
0.2	$(-1, 2, -1.8, 2.51, \rho)$	(0, 0.2)	7	0.072	0.106	0.106
0.2	$(-1, 2, -1.8, 2.51, \rho)$	(0, 0.2)	14	0.084	0.100	0.089
0.2	$(-1, 2, -1.8, 2.51, \rho)$	(0, 0.2)	21	0.082	0.088	0.078
0.2	$(-1, 2, -1.8, 2.51, \rho)$	(0, 0.2)	28	0.064	0.070	0.078
0.2	$(-1, 2, -1.8, 2.51, \rho)$	(0, 0.2)	50	0.054	0.066	0.058
0.15	$(-2, 3.4, -2, 2.51, \rho)$	(0.15, 0.15)	7	0.057	0.051	0.058
0.15	$(-2, 3.4\text{-}2, 2.51, \rho)$	(0.15, 0.15)	14	0.032	0.022	0.026
0.15	$(-2, 3.4, -2, 2.51, \rho)$	(0.15, 0.15)	21	0.021	0.022	0.020
0.15	$(-2, 3.4, -2, 2.51, \rho)$	(0.15, 0.15)	28	0.012	0.013	0.007
0.15	$(-2, 3.4, -2, 2.51, \rho)$	(0.15, 0.15)	50	0.013	0.010	0.008
0.15	$(-1, 2, -2, 2.51, \rho)$	(0, 0.15)	7	0.089	0.097	0.088
0.15	$(-1, 2, -2, 2.51, \rho)$	(0, 0.15)	14	0.085	0.077	0.087
0.15	$(-1, 2, -2, 2.51, \rho)$	(0, 0.15)	21	0.075	0.081	0.082
0.15	$(-1, 2, -2, 2.51, \rho)$	(0, 0.15)	28	0.062	0.068	0.088
0.15	$(-1, 2, -2, 2.51, \rho)$	(0, 0.15)	50	0.067	0.083	0.073

The nominal level is $\alpha = 0.05$.

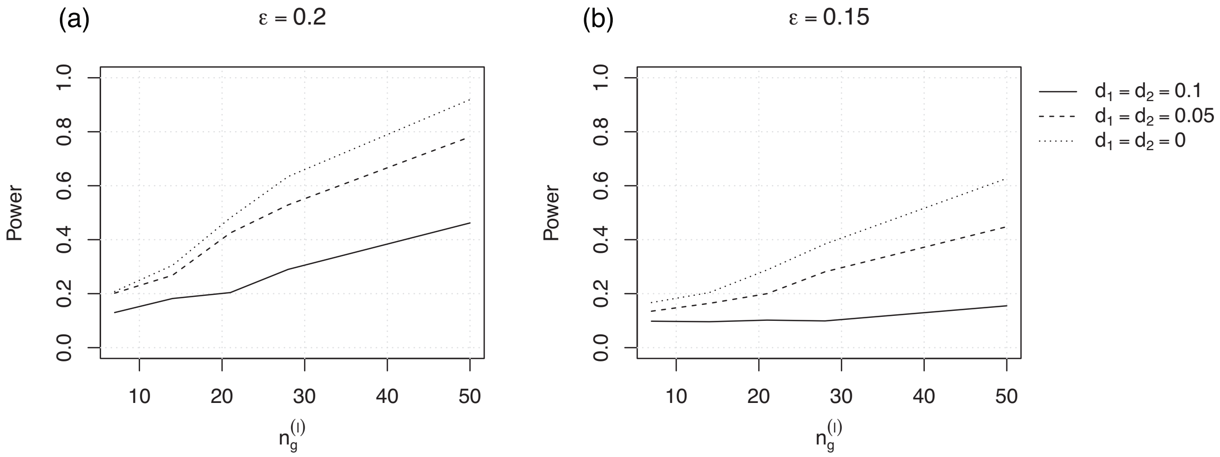


FIGURE 1 Simulated power of the test proposed in Algorithm 1 for bivariate binary outcomes for different sample sizes and $\rho = 0.2$. The three different scenarios are shown in terms of different line types. The nominal level is $\alpha = 0.05$.

et al. (2021), although we use a less conservative approach. This might suggests that the model based on the Gaussian copula outperforms the Gumbel model in this setting.

Figure 1 displays the simulated power as function of sample size for the different scenarios. As for the type I errors, the level of correlation has little effect on power, shown in detail for $\rho = 0.2$ in Figure 1 (complete results in Web Table 2). Our testing approach shows increasing power with larger sample sizes, converging to one in all scenarios. The highest power of 0.919 is observed for $(d_1, d_2) = (0, 0)$ with $\varepsilon = 0.2$, $\rho = 0.2$ and $n_g^{(l)} = 50$ (see Figure 1). For a medium sample size ($n_g^{(l)} \in \{21, 28\}$), the power is between 0.214 and 0.650 for $\varepsilon = 0.2$, and between 0.095 and 0.384 for $\varepsilon = 0.15$, respectively. Finally, when considering small sample sizes ($n_g^{(l)} \in \{7, 14\}$), our model still achieves reasonable power, with values from 0.128 to 0.334 for $\varepsilon = 0.2$, and from 0.086 to 0.204 for $\varepsilon = 0.15$. Compared to Möllenhoff et al. (2021), our method demonstrates similar high power for large samples but considerably higher power for small and medium samples (exceeding in some cases by over 5-fold). This, again, highlights that the proposed approach is less conservative relative to approaches based on the intersection union principle. Of note, this power gain can probably be explained, at least in part, by the better approximation of the significance level, which is not properly calibrated in the test proposed by Möllenhoff et al. (2021), as the simulated type I errors are predominantly close to zero.

4.2 Bivariate continuous outcome

In the case of bivariate continuous outcomes, we adopt the set up of Bretz et al. (2018), which is a linear dose-response model

$$m_k^{(1)}(x_i^{(1)}, \boldsymbol{\theta}_k^{(1)}) = \beta_{0k}^{(1)} + \beta_{1k}^{(1)} x_i^{(1)}$$

for the first group, and a quadratic model

$$m_k^{(2)}(x_i^{(2)}, \boldsymbol{\theta}_k^{(2)}) = \beta_{0k}^{(2)} + \beta_{1k}^{(2)} x_i^{(2)} + \beta_{2k}^{(2)} \left(x_i^{(2)}\right)^2$$

for the second group, $i = 1, \dots n_l$, $l = 1, 2$, $k = 1, 2$. The related full parameter vectors are given by

$$\boldsymbol{\theta}^{(1)} = (\boldsymbol{\theta}_1^{(1)}, \boldsymbol{\theta}_2^{(1)}, \sigma, \rho) = (\beta_{01}^{(1)}, \beta_{11}^{(1)}, \beta_{02}^{(1)}, \beta_{12}^{(1)}, \sigma, \rho)$$

and

$$\boldsymbol{\theta}^{(2)} = (\boldsymbol{\theta}_1^{(2)}, \boldsymbol{\theta}_2^{(2)}, \sigma, \rho) = (\beta_{01}^{(2)}, \beta_{11}^{(2)}, \beta_{21}^{(2)}, \beta_{02}^{(2)}, \beta_{12}^{(2)}, \beta_{22}^{(2)}, \sigma, \rho).$$

For consistency with Section 4.1, we transform the model such that it applies to the dose range of $x_i \in [0, 2]$; this is achieved by setting $\beta_{0k}^{(1)} = 0$, $\beta_{1k}^{(1)} = 1$, $\beta_{0k}^{(2)} = 0$, $\beta_{1k}^{(2)} = (1 - 2d_k)$ and $\beta_{2k}^{(2)} = d_k$, so that

$$m_k^{(1)}(x_i^{(1)}, \boldsymbol{\theta}_k^{(1)}) = x_i^{(1)} \quad \text{and}$$

$$m_k^{(2)}(x_i^{(2)}, \boldsymbol{\theta}_k^{(2)}) = (1 - 2d_k)x^{(2)} + d_k \left(x_i^{(2)}\right)^2, \quad k = 1, 2, \tag{8}$$

where d_k is the corresponding distance between the curves. This leads to $\boldsymbol{\theta}^{(1)} = (0, 1, 0, 1, \sigma, \rho)$ for all scenarios. The curves coincide at the boundary doses $x_i^{(l)} = 0$ and $x_i^{(l)} = 2$, and the maximum difference d_k occurs at a middle dose of $x_i^{(l)} = 1$ (an example of this is visualized in Web Figure 1).

As in Section 4.1, we assume equal correlation for both groups ($\rho = \rho^{(1)} = \rho^{(2)}$) and equal variances for the continuous variables across responses and groups, that is $\sigma = \sigma_1^{(1)} = \sigma_2^{(1)} = \sigma_1^{(2)} = \sigma_2^{(2)}$. The variance levels are chosen to be $\sigma^2 = 0.05, 0.1, 0.2$ such that the ratios ε/σ are similar to the ones chosen by Bretz et al. (2018). We investigate the same seven scenarios as in Section 4.1, that is $(d_1, d_2) = (\varepsilon, \varepsilon)$ and $(0, \varepsilon)$ with $\varepsilon = 0.15, 0.2$ for the type I error rate simulation, and $(d_1, d_2) = (0.1, 0.1)$, $(0.05, 0.05)$ and $(0, 0)$ for the power simulations. The complete parameter combinations are in Web Table 3.

Table 2 shows the simulated type I error rates. As already observed for the bivariate binary case, the level of correlation has little impact on the results; hence, we only report the type I er-

TABLE 2 Simulated type I error rates of the test proposed in Algorithm 1 for bivariate continuous outcomes specified in Equation 8 with $\rho = 0.2$ and two different similarity thresholds ε.

$\boldsymbol{\varepsilon}$	$\boldsymbol{\theta}^{(2)}$	$(\boldsymbol{d_1}, \boldsymbol{d_2})$	$\boldsymbol{n_g^{(l)}}$	$\boldsymbol{\sigma^2 = 0.05}$	$\boldsymbol{\sigma^2 = 0.1}$	$\boldsymbol{\sigma^2 = 0.2}$
0.2	$(0, 0.6, 0.2, 0, 0.6, 0.2, \sigma, \rho)$	$(0.2, 0.2)$	7	0.044	0.037	0.045
0.2	$(0, 0.6, 0.2, 0, 0.6, 0.2, \sigma, \rho)$	$(0.2, 0.2)$	14	0.029	0.037	0.038
0.2	$(0, 0.6, 0.2, 0, 0.6, 0.2, \sigma, \rho)$	$(0.2, 0.2)$	21	0.016	0.021	0.039
0.2	$(0, 0.6, 0.2, 0, 0.6, 0.2, \sigma, \rho)$	$(0.2, 0.2)$	28	0.005	0.012	0.022
0.2	$(0, 0.6, 0.2, 0, 0.6, 0.2, \sigma, \rho)$	$(0.2, 0.2)$	50	0.013	0.012	0.018
0.2	$(0, 1, 0, 0, 0.6, 0.2, \sigma, \rho)$	$(0, 0.2)$	7	0.090	0.104	0.095
0.2	$(0, 1, 0, 0, 0.6, 0.2, \sigma, \rho)$	$(0, 0.2)$	14	0.064	0.087	0.089
0.2	$(0, 1, 0, 0, 0.6, 0.2, \sigma, \rho)$	$(0, 0.2)$	21	0.075	0.073	0.080
0.2	$(0, 1, 0, 0, 0.6, 0.2, \sigma, \rho)$	$(0, 0.2)$	28	0.054	0.072	0.086
0.2	$(0, 1, 0, 0, 0.6, 0.2, \sigma, \rho)$	$(0, 0.2)$	50	0.056	0.077	0.079
0.15	$(0, 0.7, 0.15, 0, 0.7, 0.15, \sigma, \rho)$	$(0.15, 0.15)$	7	0.044	0.055	0.07
0.15	$(0, 0.7, 0.15, 0, 0.7, 0.15, \sigma, \rho)$	$(0.15, 0.15)$	14	0.035	0.045	0.06
0.15	$(0, 0.7, 0.15, 0, 0.7, 0.15, \sigma, \rho)$	$(0.15, 0.15)$	21	0.018	0.044	0.043
0.15	$(0, 0.7, 0.15, 0, 0.7, 0.15, \sigma, \rho)$	$(0.15, 0.15)$	28	0.017	0.033	0.040
0.15	$(0, 0.7, 0.15, 0, 0.7, 0.15, \sigma, \rho)$	$(0.15, 0.15)$	50	0.014	0.021	0.034
0.15	$(0, 1, 0, 0, 0.7, 0.15, \sigma, \rho)$	$(0, 0.15)$	7	0.096	0.079	0.106
0.15	$(0, 1, 0, 0, 0.7, 0.15, \sigma, \rho)$	$(0, 0.15)$	14	0.092	0.099	0.096
0.15	$(0, 1, 0, 0, 0.7, 0.15, \sigma, \rho)$	$(0, 0.15)$	21	0.081	0.084	0.089
0.15	$(0, 1, 0, 0, 0.7, 0.15, \sigma, \rho)$	$(0, 0.15)$	28	0.065	0.083	0.079
0.15	$(0, 1, 0, 0, 0.7, 0.15, \sigma, \rho)$	$(0, 0.15)$	50	0.054	0.089	0.096

The nominal level is $\alpha = 0.05$.

ror rates for the medium correlation level of $\rho = 0.2$ and refer the reader to Web Tables 4–5 for the complete set of results. For $d_1 = d_2 \approx \varepsilon$, type I error rates closely align with the 5% level across all configurations. Similar to the findings in Section 4.1, we note a slight inflation in type I errors for $(d_1, d_2) = (0, \varepsilon)$, consistent in magnitude with the binary outcomes. Notably, with lower variance, type I error rates tend to decrease and align with the desired 5% level as sample sizes increase. We partially observe a similar trend for higher variance levels.

Figure 2 displays the simulated power. Again, the level of correlation has little influence; hence, we focus on the medium correlation level of $\rho = 0.2$ (full results are in Web Tables 6–8). The proposal achieves reasonable power. For instance, for a medium sample size $(n_g^{(l)} \in \{21, 28\})$, we find a power between 0.205 and 0.968 for $\varepsilon = 0.2$, and between 0.102 and 0.790 for $\varepsilon = 0.15$. At small sample sizes, $n_g^{(l)} \in \{7, 14\}$, the approach still achieves satisfying power, reaching values from 0.098 to 0.710 for $\varepsilon = 0.2$, and from 0.107 to 0.494 for $\varepsilon = 0.15$. Finally, the power converges to one for decreasing variance and increasing sample size.

4.3 Bivariate mixed outcome

For the case of bivariate mixed outcomes, we focus on binary and continuous responses and combine the scenarios considered in Sections 4.1–4.2, corresponding to a bivariate efficacy-toxicity outcome.

As in Section 4.2, we assume $\rho = \rho^{(1)} = \rho^{(2)}$, $\sigma = \sigma^{(1)} = \sigma^{(2)}$, and the same variance levels ($\sigma^2 = 0.05$, 0.1 and 0.2). Different to the bivariate binary and continuous cases, $(d_1, d_2) = (0, \varepsilon)$ and $(d_1, d_2) = (\varepsilon, 0)$ are not equivalent for bivariate mixed outcomes; hence, we have to investigate them separately. As a consequence, we observe nine different configurations of

$$\boldsymbol{\theta}^{(1)} = (\boldsymbol{\theta}_1^{(1)}, \boldsymbol{\theta}_2^{(1)}, \sigma, \rho) = (\beta_{01}^{(1)}, \beta_{11}^{(1)}, \beta_{02}^{(1)}, \beta_{12}^{(1)}, \sigma, \rho)$$

and

$$\boldsymbol{\theta}^{(2)} = (\boldsymbol{\theta}_1^{(2)}, \boldsymbol{\theta}_2^{(2)}, \sigma, \rho) = (\beta_{01}^{(2)}, \beta_{11}^{(2)}, \beta_{21}^{(2)}, \beta_{02}^{(2)}, \beta_{12}^{(2)}, \sigma, \rho),$$

investigating $(d_1, d_2) = (\varepsilon, \varepsilon), (0, \varepsilon)$ and $(\varepsilon, 0)$ for $\varepsilon = 0.15, 0.2$ for the type I error simulations and $(d_1, d_2) = (0.1, 0.1), (0.05, 0.05)$ and $(0, 0)$ for the power simulations, with $\boldsymbol{\theta}^{(1)}$ constantly held as $(0, 1, -1, 2, \sigma, \rho)$. The complete parameter combinations are in Web Table 9.

Table 3 shows the simulated type I error rates for $\rho = 0.2$ and mirror findings from bivariate continuous outcomes, with $d_1 = d_2 \approx \varepsilon$ showing alignment with the 5% error level. Slight inflation is observed for $(d_1, d_2) = (0, \varepsilon)$ and $(\varepsilon, 0)$. Variance, affecting only one outcome, seems to have a reduced impact on the type I error rates. Interestingly, $d_1 = 0$ tends to produce slightly higher type I error rates, aligning with observations from Section 4.2. Results for $\rho = 0.1$ and 0.2 are in Web Tables 10–11.

Figure 3 shows the simulated power values for $\rho = 0.2$ (full results are in Web Tables 12–14) with power generally increasing for smaller variances and larger sample sizes. We observe the maximum power of 0.984 for $(d_1, d_2) = (0, 0)$, $\varepsilon = 0.2$, and $n_g^{(l)} = 50$. Medium sample sizes lead to power values ranging from 0.218 to 0.813 for $\varepsilon = 0.2$, decreasing slightly for $\varepsilon = 0.15$. Even with smaller sample sizes, our approach maintains reasonable power, reinforcing the robustness of our approach.

5 CASE STUDY

We illustrate the proposed methodology through a case study, inspired by Möllenhoff et al. (2021). The goal of this study was to investigate dental pain reduction of a nonsteroidal anti-inflammatory drug after the removal of two or more impacted third molar teeth. Specifically, interest was in assessing similarity with an already available marketed product with regard to a bivariate efficacy-toxicity outcome. For the purpose of the follow-

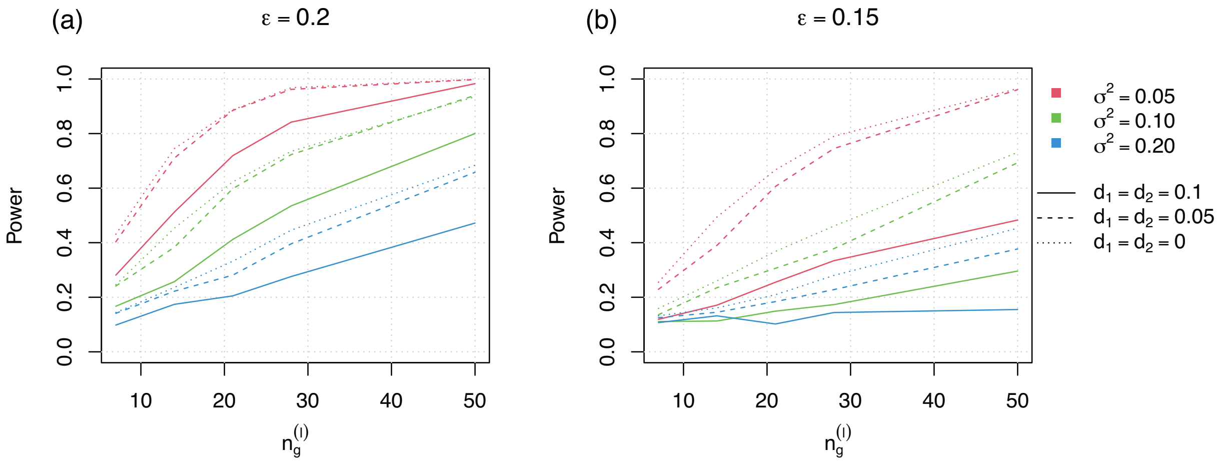


FIGURE 2 Simulated power of the test proposed in Algorithm 1 for bivariate continuous outcomes for different sample sizes and $\rho = 0.2$. The different variance levels are shown in terms of colors and the three different scenarios are shown in terms of different line types. The nominal level is $\alpha = 0.05$.

TABLE 3 Simulated type I error rates of the test proposed in Algorithm 1 for bivariate mixed outcomes with $\rho = 0.2$ and two different similarity thresholds ε.

$\boldsymbol{\varepsilon}$	$\boldsymbol{\theta}^{(2)}$	$(\boldsymbol{d_1}, \boldsymbol{d_2})$	$\boldsymbol{n_g^{(l)}}$	$\boldsymbol{\sigma^2 = 0.05}$	$\boldsymbol{\sigma^2 = 0.1}$	$\boldsymbol{\sigma^2 = 0.2}$
0.2	$(0, 0.6, 0.2, -2.4, 3.4\sigma, \rho)$	$(0.2, 0.2)$	7	0.032	0.035	0.038
0.2	$(0, 0.6, 0.2, -2.4, 3.4\sigma, \rho)$	$(0.2, 0.2)$	14	0.022	0.031	0.03
0.2	$(0, 0.6, 0.2, -2.4, 3.4\sigma, \rho)$	$(0.2, 0.2)$	21	0.018	0.009	0.017
0.2	$(0, 0.6, 0.2, -2.4, 3.4\sigma, \rho)$	$(0.2, 0.2)$	28	0.014	0.016	0.019
0.2	$(0, 0.6, 0.2, -2.4, 3.4\sigma, \rho)$	$(0.2, 0.2)$	50	0.013	0.014	0.016
0.2	$(0, 0.6, 0.2, -2, 2, \sigma, \rho)$	$(0.2, 0)$	7	0.067	0.076	0.083
0.2	$(0, 0.6, 0.2, -2, 2, \sigma, \rho)$	$(0.2, 0)$	14	0.077	0.078	0.079
0.2	$(0, 0.6, 0.2, -2, 2, \sigma, \rho)$	$(0.2, 0)$	21	0.070	0.071	0.103
0.2	$(0, 0.6, 0.2, -2, 2, \sigma, \rho)$	$(0.2, 0)$	28	0.073	0.072	0.084
0.2	$(0, 0.6, 0.2, -2, 2, \sigma, \rho)$	$(0.2, 0)$	50	0.049	0.057	0.069
0.2	$(0, 1, 0, -2.4, 3.4, \sigma, \rho)$	$(0, 0.2)$	7	0.114	0.117	0.076
0.2	$(0, 1, 0, -2.4, 3.4, \sigma, \rho)$	$(0, 0.2)$	14	0.082	0.075	0.070
0.2	$(0, 1, 0, -2.4, 3.4, \sigma, \rho)$	$(0, 0.2)$	21	0.061	0.072	0.068
0.2	$(0, 1, 0, -2.4, 3.4, \sigma, \rho)$	$(0, 0.2)$	28	0.059	0.055	0.062
0.2	$(0, 1, 0, -2.4, 3.4, \sigma, \rho)$	$(0, 0.2)$	50	0.059	0.045	0.055
0.15	$(0.7, 0.15, -2, 3.4, 0, \sigma, \rho)$	$(0.15, 0.15)$	7	0.037	0.036	0.051
0.15	$(0.7, 0.15, -2, 3.4, 0, \sigma, \rho)$	$(0.15, 0.15)$	14	0.029	0.022	0.038
0.15	$(0.7, 0.15, -2, 3.4, 0, \sigma, \rho)$	$(0.15, 0.15)$	21	0.023	0.024	0.034
0.15	$(0.7, 0.15, -2, 3.4, 0, \sigma, \rho)$	$(0.15, 0.15)$	28	0.013	0.02	0.031
0.15	$(0.7, 0.15, -2, 3.4, 0, \sigma, \rho)$	$(0.15, 0.15)$	50	0.018	0.009	0.019
0.15	$(0.7, 0.15, -2, 2, 0, \sigma, \rho)$	$(0.15, 0)$	7	0.073	0.077	0.090
0.15	$(0.7, 0.15, -2, 2, 0, \sigma, \rho)$	$(0.15, 0)$	14	0.075	0.077	0.082
0.15	$(0.7, 0.15, -2, 2, 0, \sigma, \rho)$	$(0.15, 0)$	21	0.075	0.084	0.089
0.15	$(0.7, 0.15, -2, 2, 0, \sigma, \rho)$	$(0.15, 0)$	28	0.048	0.069	0.084
0.15	$(0.7, 0.15, -2, 2, 0, \sigma, \rho)$	$(0.15, 0)$	50	0.070	0.071	0.080
0.15	$(1, 0, -2, 3.4, 0, \sigma, \rho)$	$(0, 0.15)$	7	0.097	0.091	0.092
0.15	$(1, 0, -2, 3.4, 0, \sigma, \rho)$	$(0, 0.15)$	14	0.092	0.086	0.075
0.15	$(1, 0, -2, 3.4, 0, \sigma, \rho)$	$(0, 0.15)$	21	0.076	0.099	0.076
0.15	$(1, 0, -2, 3.4, 0, \sigma, \rho)$	$(0, 0.15)$	28	0.056	0.068	0.067
0.15	$(1, 0, -2, 3.4, 0, \sigma, \rho)$	$(0, 0.15)$	50	0.045	0.051	0.073

The nominal level is $\alpha = 0.05$.

ing analysis, we used a hypothetical data set, simulated according to real data, due to confidentiality reasons.

Pain intensity is measured on an ordinal scale at baseline, and several times after the administration of a single dose. Even though the original scale is ordinal, the average over the repeated measurements can be modeled as a continuous variable. Besides the placebo, there are 4 dose levels for each drug ($g = 1, \ldots, 5$), where the levels are 0.05, 0.20, 0.50, and 1 for the investigational drug and 0.10, 0.30, 0.60, and 1 for the marketed product, respectively. The actual doses are scaled to lie within the $[0, 1]$

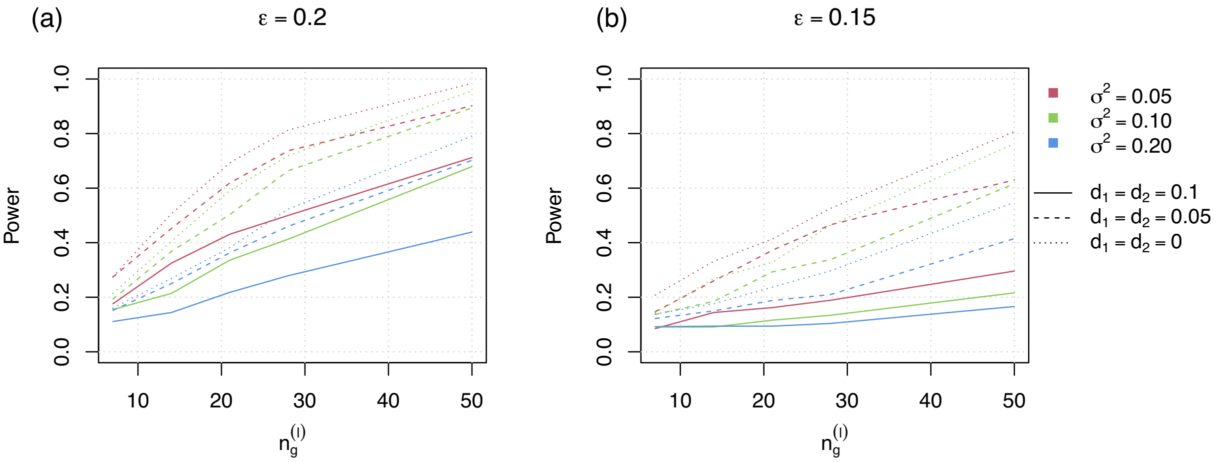


FIGURE 3 Simulated power of the test proposed in Algorithm 1 for bivariate mixed outcomes for different sample sizes and $\rho = 0.2$. The different variance levels are shown in terms of colors and the three different scenarios are shown in terms of different line types. The nominal level is $\alpha = 0.05$.

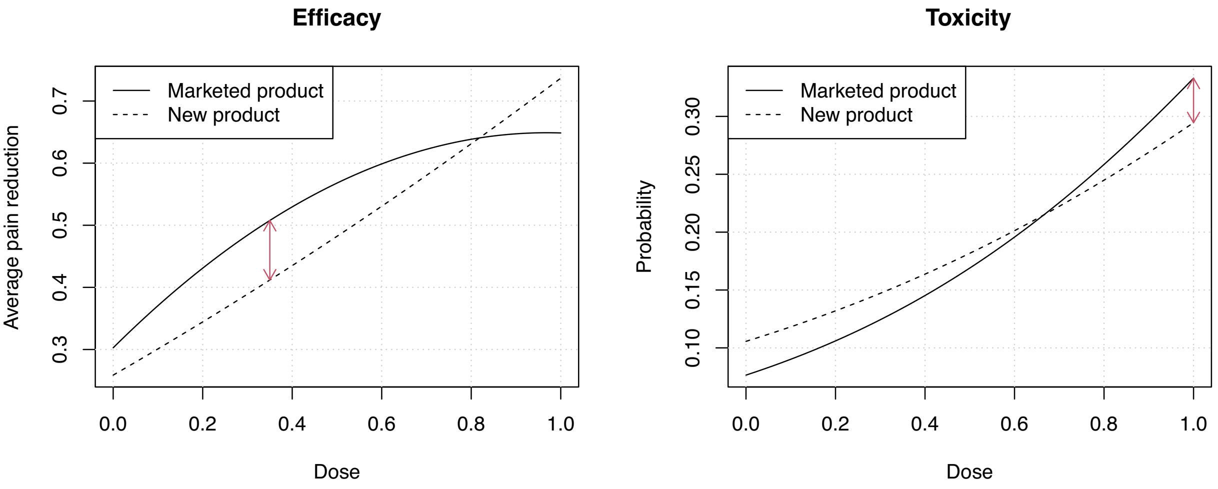


FIGURE 4 Estimated dose response curves for the efficacy and toxicity of the marketed product and the new product, respectively. The arrows indicate the maximum distance between the curves.

interval to maintain confidentiality. A total of $n = 300$ patients are evenly allocated to the five dose levels of the two drugs, resulting in $n_g^{(l)} = 30$ patients per group. In order to incorporate different similarity thresholds $\varepsilon_1 \neq \varepsilon_2$ into the analysis, we linearly rescale the average pain reduction as suggested in Section 3.1.

A binary toxicity variable indicates whether or not side effects (eg, nausea and sedation after dosing) occur. Thus, the outcome of interest is a bivariate mixed outcome (continuous-binary). The approach of Möllenhoff et al. (2021) is limited to binary variables; hence, the authors created a binary success variable for efficacy by comparing the average pain reduction to a clinical relevance threshold. Our approach is not restricted to binary efficacy outcomes so that we can consider the original average pain reduction on the continuous scale, allowing us to better exploit the available information in the analysis. Specifically, one can construct the dataset considered in Möllenhoff et al. (2021) from our dataset by using 0.5 as the clinical relevance threshold for the rescaled efficacy variable.

We fit two bivariate mixed outcome models based on the Gaussian copula (as introduced in Section 2.2), one for the new product and the other for the marketed product. We assume a quadratic dose-response curve for the continuous efficacy variable and a logit model for the binary toxicity outcome. Figure 4 shows the estimated dose-response curves with the corresponding coefficients estimates given in Web Table 15. For the toxicity outcome, our findings align closely with those of Möllenhoff et al. (2021), suggesting that our model does not depend sensitively on the copula used or the continuous modeling of efficacy. This leads to a maximum absolute distance of ~0.0385 for toxicity, observed at the highest dose of 1. For the efficacy outcome, we observe the maximum absolute distance between the curves of ~0.0958 at dose 0.35. Accordingly, the maximum of maxima distance is given by $\widehat{d}_{max} = \widehat{d}_{\text{Efficacy}} \approx 0.096$.

To assess similarity, we apply Algorithm 1 to test the hypotheses in Equation 4 for three different choices of ε, namely 0.2, 0.15 and 0.1. This leads to $\widehat{d}^*_{max,(\lfloor n_{boot}\alpha \rfloor)}$ being 0.144, 0.105, and 0.078, respectively, and corresponds to P-values given by

0.003, 0.023, and 0.136. Hence, for $\varepsilon = 0.2, 0.15$, we reject the null hypothesis, suggesting similarity. However, for $\varepsilon = 0.1$, we cannot reject the null hypothesis. These findings mostly align with those of Möllenhoff et al. (2021) for $\varepsilon = 0.2$ and $\varepsilon = 0.1$. Unlike Möllenhoff et al. (2021), our approach allows rejecting H_0 even at a more liberal threshold ($\varepsilon = 0.15$), indicating the benefit of using a continuous efficacy variable over a binary one. This aligns with our simulation study, which showed higher power for bivariate mixed outcomes as compared to binary ones. The increased power of the proposal is evidenced by the ability to conclude treatment similarity at a 5% level for $\varepsilon = 0.15$; hence, highlighting its potential impact in clinical research.

6 CONCLUSION

We introduced a novel model-based equivalence testing approach for multivariate responses, leveraging the flexibility of *generalized joined regression models*. This method stands out for its versatility across various modeling problems, particularly benefiting from the Gaussian copula's capacity to generalize to multi-dimensional settings. In contrast to existing approaches, our proposal is not limited to univariate or bivariate outcomes allowing for multivariate responses of arbitrary dimension. It accommodates different scales of measures of the outcome variables (eg, continuous, binary, categorical, or ordinal). Additionally, we propose an alternative, less conservative testing procedure that contrasts with the intersection union principle.

The simulation study demonstrates that our method effectively maintains the type I error rate at or below the 5% nominal significance level as sample sizes increase, despite some inflation at smaller sizes for any of the investigated types of outcomes. This effect particularly occurs in scenarios, which are on the boundary of the null hypothesis space but only because of one outcome, that is, if $K = 2, d_1 < \varepsilon$, or $d_2 < \varepsilon$. The reason for this is that for these scenarios in finite samples, the upper bound of the type I error probability

$$\mathbb{P}(\text{type I error}) \leq \max\Big\{\mathbb{P}_{H_0}\big(\widehat{d_1} < \widehat{d}^*_{max,(\lfloor n_{boot}\alpha\rfloor)} \big| \widehat{d_1} \geq \widehat{d_2}\big), \mathbb{P}_{H_0}\big(\widehat{d_2} < \widehat{d}^*_{max,(\lfloor n_{boot}\alpha\rfloor)} \big| \widehat{d_2} \geq \widehat{d_1}\big)\Big\}$$

becomes less strict. However, this effect disappears with increasing sample sizes, which is in line with the theoretical arguments given in Section 3.2. Additionally, we achieve reasonably power values that converge to one as sample sizes increase. Note that we do not observe type I error rates as large as Möllenhoff et al. (2021) do, even though we use a less conservative testing procedure. For large sample sizes, both approaches achieve reasonable power. However, at small sample sizes, our new approach outperforms the procedure of Möllenhoff et al. (2021).

Future possible research includes extending the generalized joint regression models to more than three dimensions, a limitation of the current implementation of the proposed approach. The sensitivity of results to the assumption of Gaussian copula in various contexts, as well as alternative copula options, merits further exploration. Additionally, we aim to adapt the testing procedure for less standard distributions and explore spline-based regression curve specifications.

Finally, a much-needed extension is the derivation of a power formula that allows practitioners to perform sample size calculations at the design stage of a trial. Such a formula could be derived from the asymptotic distribution of the test statistic or from simulations. We leave a validation and comparison of these two proposed approaches for future research.

ACKNOWLEDGMENTS

The authors would like to thank Holger Dette for fruitful discussions during the early phase of developing the method proposed in this paper.

SUPPLEMENTARY MATERIALS

Supplementary material is available at *Biometrics* online.

Web Tables and Figures referenced in Sections 4.1, 4.2, 4.3, and 5, along with codes, are available with this paper at the Biometrics website on Oxford Academic. The codes used for this paper are also available at https://github.com/Niklas191/Testing_for_similarity_of_multivariate_mixed_outcomes.

FUNDING

This work has been supported by the Research Training Group "Biostatistical Methods for High-Dimensional Data in Toxicology" (RTG 2624, P7) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, Project Number 427806116).

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY

The data that support the findings in this paper are available with this paper at the Biometrics website on Oxford Academic. The data are also available at https://github.com/Niklas191/Testing_for_similarity_of_multivariate_mixed_outcomes.

REFERENCES

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24, 295–300.

Bretz, F., Möllenhoff, K., Dette, H., Liu, W. and Trampisch, M. (2018). Assessing the similarity of dose response and target doses in two non-overlapping subgroups. *Statistics in Medicine*, 37, 722–738.

Chiu, W. A. and Crump, K. S. (2012). Using copulas to introduce dependence in dose-response modeling of multiple binary endpoints. *Journal of Agricultural, Biological, and Environmental Statistics*, 17, 107–127.

de Leon, A. R. and Wu, B. (2011). Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine*, 30, 175–185.

Demirtas, H. and Doganay, B. (2012). Simultaneous generation of binary and normal data with specified marginal and association structures. *Journal of Biopharmaceutical Statistics*, 22, 223–236.

Dette, H., Möllenhoff, K., Volgushev, S. and Bretz, F. (2018). Equivalence of regression curves. *Journal of the American Statistical Association*, 113, 711–729.

Dunlap, J. W. (1937). Combinative properties of correlation coefficients. *The Journal of Experimental Education*, 5, 286–288.

Emrich, L. J. and Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician*, 45, 302–304.

Filippou, P., Marra, G. and Radice, R. (2017). Penalized likelihood estimation of a trivariate additive probit model. *Biostatistics*, 18, 569–585.

Gsteiger, S., Bretz, F. and Liu, W. (2011). Simultaneous confidence bands for nonlinear regression models with application to population pharmacokinetic analyses. *Journal of Biopharmaceutical Statistics*, 21, 708–725.

Heise, M. A. and Myers, R. H. (1996). Optimal designs for bivariate logistic regression. *Biometrics*, 52, 613–624.

Jhee, S. S., Lyness, W. H., Rojas, P. B., Leibowitz, M. T., Zarotsky, V. and Jacobsen, L. V. (2004). Similarity of insulin detemir pharmacokinetics, safety, and tolerability profiles in healthy Caucasian and Japanese American subjects. *The Journal of Clinical Pharmacology*, 44, 258–264.

Joe, H. (2015). *Dependence Modeling with Copulas*. 1st edn. New York and Boca Raton: Chapman & Hall and CRC Press.

Klein, N., Kneib, T., Marra, G., Radice, R., Rokicki, S. and McGovern, M. E. (2019). Mixed binary-continuous copula regression models with application to adverse birth outcomes. *Statistics in Medicine*, 38, 413–436.

Liu, W., Bretz, F., Hayter, A. J. and Wynn, H. P. (2009). Assessing nonsuperiority, noninferiority, or equivalence when comparing two regression models over a restricted covariate region. *Biometrics*, 65, 1279–1287.

Marra, G. and Radice, R. (2017). Bivariate copula additive models for location, scale and shape. *Computational Statistics and Data Analysis*, 112, 99–113.

Marra, G. and Radice, R. (2020). Copula link-based additive models for right-censored event time data. *Journal of the American Statistical Association*, 115, 886–895.

Marra, G. and Radice, R. (2023). GJRM: Generalised Joint Regression Modelling. CRAN: Package GJRM (r-project.org). R package version 0.2-6.1. www.cran.r-project.org/web/packages/GJRM. [21 December 2023].

Marra, G., Radice, R. and Zimmer, D. (2020). Estimating the binary endogenous effect of insurance on doctor visits by copula-based regression additive models. *Journal of the Royal Statistical Society Series C*, 69, 953–971.

Möllenhoff, K., Binder, N. and Dette, H. (2024). Testing similarity of parametric competing risks models for identifying potentially similar pathways in healthcare. *arXiv: 2401.04490 [stat.ME]*.

Möllenhoff, K., Dette, H. and Bretz, F. (2021). Testing for similarity of binary efficacy-toxicity responses. *Biostatistics*, 23, 949–966.

Möllenhoff, K., Dette, H., Kotzagiorgis, E., Volgushev, S. and Collignon, O. (2018). Regulatory assessment of drug dissolution profiles comparability via maximum deviation. *Statistics in Medicine*, 37, 2968–2981.

Möllenhoff, K., Loingeville, F., Bertrand, J., Nguyen, T. T., Sharan, S., Zhao, L. et al. (2022). Efficient model-based bioequivalence testing. *Biostatistics*, 23, 314–327.

Murtaugh, P. A. and Fisher, L. D. (1990). Bivariate binary models of efficacy and toxicity in dose-ranging trials. *Communications in Statistics—Theory and Methods*, 19, 2003–2020.

Otto, C., Fuchs, I., Altmann, H., Klewer, M., Walter, A., Prelle, K. et al. (2008). Comparative analysis of the uterine and mammary gland effects of drospirenone and medroxyprogesterone acetate. *Endocrinology*, 149, 3952–3959.

Radice, R. and Marra, G. (2016). Copula regression spline models for binary outcomes. *Statistics and Computing*, 26, 981–995.

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

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

Tao, Y., Liu, J., Li, Z., Lin, J., Lu, T. and Yan, F. (2013). Dose-finding based on bivariate efficacy-toxicity outcome using Archimedean copula. *PLoS One*, 8, 1–6.

Trivedi, P. K. and Zimmer, D. M. (2007). Copula modeling: an introduction for practitioners. *Foundations and Trends in Econometrics*, 1, 1–111.

Varadhan, R. (2022). alabama: Constrained Nonlinear Optimization. R package version 2022.4-1. https://cran.r-project.org/web/packages/alabama/. [Accessed 10 July 2023].

Received: January 18, 2024; Revised: June 5, 2024; Accepted: July 24, 2024

Supplementary material for "Testing for similarity of multivariate mixed outcomes using generalised joint regression models with application to efficacy-toxicity responses“ by Niklas Hagemann, Giampiero Marra, Frank Bretz and Kathrin Möllenhoff

Web Table 1: Parameter scenarios used for the simulation of the bivariate binary outcomes.

No.		$\theta^{(1)}$	$\theta^{(2)}$	(d_1, d_2)
1	Null hypothesis	$(-1, 2, -3, 3, \rho)$	$(-2.4, 3.4, -1.8, 2.51, \rho)$	(0.2,0.2)
2	Null hypothesis	$(-1, 2, -3, 3, \rho)$	$(-1, 2, -1.8, 2.51, \rho)$	(0,0.2)
3	Null hypothesis	$(-1, 2, -3, 3, \rho)$	$(-2, 3.4, -2, 2.51, \rho)$	(0.15,0.15)
4	Null hypothesis	$(-1, 2, -3, 3, \rho)$	$(-1, 2, -2, 2.51, \rho)$	(0,0.15)
5	Alternative	$(-1, 2, -3, 3, \rho)$	$(-1.5, 2.2, -3.6, 3.2, \rho)$	(0.1,0.1)
6	Alternative	$(-1, 2, -3, 3, \rho)$	$(-1.2, 2, -3.3, 3.1, \rho)$	(0.05,0.05)
7	Alternative	$(-1, 2, -3, 3, \rho)$	$(-1, 2, -3, 3, \rho)$	(0,0)

Web Table 2: Simulated power for bivariate binary outcomes and two different similarity thresholds ε.

ε	$\boldsymbol{\theta}^{(2)}$	(d_1, d_2)	$n_g^{(l)}$	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$
0.2	$(-1.5, 2.2, -3.6, 3.2, \rho)$	$(0.1, 0.1)$	7	0.151	0.130	0.128
0.2	$(-1.5, 2.2, -3.6, 3.2, \rho)$	$(0.1, 0.1)$	14	0.178	0.182	0.177
0.2	$(-1.5, 2.2, -3.6, 3.2, \rho)$	$(0.1, 0.1)$	21	0.202	0.204	0.214
0.2	$(-1.5, 2.2, -3.6, 3.2, \rho)$	$(0.1, 0.1)$	28	0.295	0.290	0.299
0.2	$(-1.5, 2.2, -3.6, 3.2, \rho)$	$(0.1, 0.1)$	50	0.426	0.462	0.473
0.2	$(-1.2, 2, -3.3, 3.1, \rho)$	$(0.05, 0.05)$	7	0.166	0.202	0.179
0.2	$(-1.2, 2, -3.3, 3.1, \rho)$	$(0.05, 0.05)$	14	0.293	0.268	0.291
0.2	$(-1.2, 2, -3.3, 3.1, \rho)$	$(0.05, 0.05)$	21	0.424	0.425	0.417
0.2	$(-1.2, 2, -3.3, 3.1, \rho)$	$(0.05, 0.05)$	28	0.522	0.529	0.526
0.2	$(-1.2, 2, -3.3, 3.1, \rho)$	$(0.05, 0.05)$	50	0.777	0.781	0.808
0.2	$(-1, 2, -3, 3, \rho)$	$(0, 0)$	7	0.219	0.208	0.222
0.2	$(-1, 2, -3, 3, \rho)$	$(0, 0)$	14	0.334	0.305	0.306
0.2	$(-1, 2, -3, 3, \rho)$	$(0, 0)$	21	0.495	0.481	0.502
0.2	$(-1, 2, -3, 3, \rho)$	$(0, 0)$	28	0.621	0.634	0.650
0.2	$(-1, 2, -3, 3, \rho)$	$(0, 0)$	50	0.906	0.919	0.914
0.15	$(-1.5, 2.2, -3.6, 3.2, \rho)$	$(0.1, 0.1)$	7	0.100	0.098	0.107
0.15	$(-1.5, 2.2, -3.6, 3.2, \rho)$	$(0.1, 0.1)$	14	0.095	0.096	0.086
0.15	$(-1.5, 2.2, -3.6, 3.2, \rho)$	$(0.1, 0.1)$	21	0.095	0.102	0.096
0.15	$(-1.5, 2.2, -3.6, 3.2, \rho)$	$(0.1, 0.1)$	28	0.098	0.099	0.129
0.15	$(-1.5, 2.2, -3.6, 3.2, \rho)$	$(0.1, 0.1)$	50	0.141	0.155	0.148
0.15	$(-1.2, 2, -3.3, 3.1, \rho)$	$(0.05, 0.05)$	7	0.117	0.135	0.140
0.15	$(-1.2, 2, -3.3, 3.1, \rho)$	$(0.05, 0.05)$	14	0.161	0.164	0.151
0.15	$(-1.2, 2, -3.3, 3.1, \rho)$	$(0.05, 0.05)$	21	0.212	0.200	0.196
0.15	$(-1.2, 2, -3.3, 3.1, \rho)$	$(0.05, 0.05)$	28	0.261	0.281	0.267
0.15	$(-1.2, 2, -3.3, 3.1, \rho)$	$(0.05, 0.05)$	50	0.434	0.448	0.439
0.15	$(-1, 2, -3, 3, \rho)$	$(0, 0)$	7	0.131	0.167	0.159
0.15	$(-1, 2, -3, 3, \rho)$	$(0, 0)$	14	0.168	0.204	0.173
0.15	$(-1, 2, -3, 3, \rho)$	$(0, 0)$	21	0.296	0.288	0.271
0.15	$(-1, 2, -3, 3, \rho)$	$(0, 0)$	28	0.375	0.384	0.382
0.15	$(-1, 2, -3, 3, \rho)$	$(0, 0)$	50	0.636	0.627	0.639

Web Table 3: Parameter scenarios used for the simulation of the bivariate continuous outcomes.

No.		$\theta^{(1)}$	$\theta^{(2)}$	(d_1, d_2)
1	Null hypothesis	$(0, 1, 0, 1, \sigma, \rho)$	$(0, 0.6, 0.2, 0, 0.6, 0.2, \sigma, \rho)$	(0.2,0.2)
2	Null hypothesis	$(0, 1, 0, 1, \sigma, \rho)$	$(0, 1, 0, 0, 0.6, 0.2, \sigma, \rho)$	(0,0.2)
3	Null hypothesis	$(0, 1, 0, 1, \sigma, \rho)$	$(0, 0.7, 0.15, 0, 0.7, 0.15, \sigma, \rho)$	(0.15,0.15)
4	Null hypothesis	$(0, 1, 0, 1, \sigma, \rho)$	$(0, 1, 0, 0, 0.7, 0.15, \sigma, \rho)$	(0,0.15)
5	Alternative	$(0, 1, 0, 1, \sigma, \rho)$	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)
6	Alternative	$(0, 1, 0, 1, \sigma, \rho)$	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)
7	Alternative	$(0, 1, 0, 1, \sigma, \rho)$	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)

Web Table 4: Simulated type I error rates for bivariate continuous outcomes with $\rho = 0.1$ and two different similarity thresholds ε.

ε	$\boldsymbol{\theta}^{(2)}$	(d_1, d_2)	$n_g^{(l)}$	$\sigma^2 = 0.05$	$\sigma^2 = 0.1$	$\sigma^2 = 0.2$
0.2	$(0, 0.6, 0.2, 0, 0.6, 0.2, \sigma, \rho)$	(0.2,0.2)	7	0.026	0.046	0.063
0.2	$(0, 0.6, 0.2, 0, 0.6, 0.2, \sigma, \rho)$	(0.2,0.2)	14	0.019	0.036	0.042
0.2	$(0, 0.6, 0.2, 0, 0.6, 0.2, \sigma, \rho)$	(0.2,0.2)	21	0.013	0.029	0.029
0.2	$(0, 0.6, 0.2, 0, 0.6, 0.2, \sigma, \rho)$	(0.2,0.2)	28	0.013	0.018	0.026
0.2	$(0, 0.6, 0.2, 0, 0.6, 0.2, \sigma, \rho)$	(0.2,0.2)	50	0.013	0.011	0.012
0.2	$(0, 1, 0, 0, 0.6, 0.2, \sigma, \rho)$	(0,0.2)	7	0.082	0.096	0.077
0.2	$(0, 1, 0, 0, 0.6, 0.2, \sigma, \rho)$	(0,0.2)	14	0.079	0.088	0.077
0.2	$(0, 1, 0, 0, 0.6, 0.2, \sigma, \rho)$	(0,0.2)	21	0.061	0.086	0.095
0.2	$(0, 1, 0, 0, 0.6, 0.2, \sigma, \rho)$	(0,0.2)	28	0.067	0.069	0.078
0.2	$(0, 1, 0, 0, 0.6, 0.2, \sigma, \rho)$	(0,0.2)	50	0.045	0.060	0.079
0.15	$(0, 0.7, 0.15, 0, 0.7, 0.15, \sigma, \rho)$	(0.15,0.15)	7	0.037	0.043	0.093
0.15	$(0, 0.7, 0.15, 0, 0.7, 0.15, \sigma, \rho)$	(0.15,0.15)	14	0.030	0.040	0.062
0.15	$(0, 0.7, 0.15, 0, 0.7, 0.15, \sigma, \rho)$	(0.15,0.15)	21	0.022	0.034	0.035
0.15	$(0, 0.7, 0.15, 0, 0.7, 0.15, \sigma, \rho)$	(0.15,0.15)	28	0.008	0.028	0.036
0.15	$(0, 0.7, 0.15, 0, 0.7, 0.15, \sigma, \rho)$	(0.15,0.15)	50	0.012	0.004	0.029
0.15	$(0, 1, 0, 0, 0.7, 0.15, \sigma, \rho)$	(0,0.15)	7	0.090	0.091	0.112
0.15	$(0, 1, 0, 0, 0.7, 0.15, \sigma, \rho)$	(0,0.15)	14	0.076	0.087	0.086
0.15	$(0, 1, 0, 0, 0.7, 0.15, \sigma, \rho)$	(0,0.15)	21	0.080	0.089	0.068
0.15	$(0, 1, 0, 0, 0.7, 0.15, \sigma, \rho)$	(0,0.15)	28	0.079	0.092	0.074
0.15	$(0, 1, 0, 0, 0.7, 0.15, \sigma, \rho)$	(0,0.15)	50	0.063	0.066	0.087

Web Table 5: Simulated type I error rates for bivariate continuous outcomes with $\rho = 0.3$ and two different similarity thresholds ε.

ε	$\boldsymbol{\theta}^{(2)}$	(d_1, d_2)	$n_g^{(l)}$	$\sigma^2 = 0.05$	$\sigma^2 = 0.1$	$\sigma^2 = 0.2$
0.2	$(0, 0.6, 0.2, 0, 0.6, 0.2, \sigma, \rho)$	(0.2,0.2)	7	0.033	0.047	0.055
0.2	$(0, 0.6, 0.2, 0, 0.6, 0.2, \sigma, \rho)$	(0.2,0.2)	14	0.018	0.041	0.050
0.2	$(0, 0.6, 0.2, 0, 0.6, 0.2, \sigma, \rho)$	(0.2,0.2)	21	0.024	0.030	0.040
0.2	$(0, 0.6, 0.2, 0, 0.6, 0.2, \sigma, \rho)$	(0.2,0.2)	28	0.014	0.030	0.029
0.2	$(0, 0.6, 0.2, 0, 0.6, 0.2, \sigma, \rho)$	(0.2,0.2)	50	0.017	0.018	0.030
0.2	$(0, 1, 0, 0, 0.6, 0.2, \sigma, \rho)$	(0,0.2)	7	0.090	0.092	0.098
0.2	$(0, 1, 0, 0, 0.6, 0.2, \sigma, \rho)$	(0,0.2)	14	0.069	0.085	0.088
0.2	$(0, 1, 0, 0, 0.6, 0.2, \sigma, \rho)$	(0,0.2)	21	0.059	0.072	0.096
0.2	$(0, 1, 0, 0, 0.6, 0.2, \sigma, \rho)$	(0,0.2)	28	0.049	0.081	0.077
0.2	$(0, 1, 0, 0, 0.6, 0.2, \sigma, \rho)$	(0,0.2)	50	0.059	0.069	0.082
0.15	$(0, 0.7, 0.15, 0, 0.7, 0.15, \sigma, \rho)$	(0.15,0.15)	7	0.051	0.050	0.083
0.15	$(0, 0.7, 0.15, 0, 0.7, 0.15, \sigma, \rho)$	(0.15,0.15)	14	0.033	0.051	0.058
0.15	$(0, 0.7, 0.15, 0, 0.7, 0.15, \sigma, \rho)$	(0.15,0.15)	21	0.034	0.044	0.049
0.15	$(0, 0.7, 0.15, 0, 0.7, 0.15, \sigma, \rho)$	(0.15,0.15)	28	0.017	0.021	0.042
0.15	$(0, 0.7, 0.15, 0, 0.7, 0.15, \sigma, \rho)$	(0.15,0.15)	50	0.016	0.023	0.044
0.15	$(0, 1, 0, 0, 0.7, 0.15, \sigma, \rho)$	(0,0.15)	7	0.084	0.090	0.091
0.15	$(0, 1, 0, 0, 0.7, 0.15, \sigma, \rho)$	(0,0.15)	14	0.070	0.094	0.090
0.15	$(0, 1, 0, 0, 0.7, 0.15, \sigma, \rho)$	(0,0.15)	21	0.076	0.085	0.072
0.15	$(0, 1, 0, 0, 0.7, 0.15, \sigma, \rho)$	(0,0.15)	28	0.061	0.077	0.088
0.15	$(0, 1, 0, 0, 0.7, 0.15, \sigma, \rho)$	(0,0.15)	50	0.069	0.077	0.090

Web Table 6: Simulated power for bivariate continuous outcomes with $\rho = 0.1$ and two different similarity thresholds ε.

ε	$\boldsymbol{\theta}^{(2)}$	(d_1, d_2)	$n_g^{(l)}$	$\sigma^2 = 0.05$	$\sigma^2 = 0.1$	$\sigma^2 = 0.2$
0.2	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	7	0.257	0.126	0.118
0.2	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	14	0.526	0.237	0.159
0.2	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	21	0.689	0.395	0.215
0.2	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	28	0.851	0.515	0.280
0.2	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	50	0.982	0.792	0.452
0.2	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	7	0.412	0.207	0.140
0.2	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	14	0.712	0.399	0.208
0.2	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	21	0.887	0.579	0.340
0.2	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	28	0.974	0.694	0.387
0.2	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	50	0.999	0.938	0.668
0.2	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	7	0.406	0.242	0.155
0.2	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	14	0.749	0.457	0.232
0.2	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	21	0.907	0.636	0.333
0.2	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	28	0.959	0.739	0.444
0.2	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	50	0.999	0.930	0.670
0.15	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	7	0.119	0.111	0.110
0.15	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	14	0.180	0.123	0.098
0.15	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	21	0.244	0.139	0.108
0.15	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	28	0.315	0.186	0.125
0.15	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	50	0.506	0.28	0.182
0.15	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	7	0.232	0.146	0.112
0.15	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	14	0.373	0.241	0.162
0.15	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	21	0.590	0.318	0.161
0.15	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	28	0.729	0.394	0.249
0.15	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	50	0.952	0.666	0.382
0.15	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	7	0.277	0.168	0.161
0.15	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	14	0.488	0.263	0.169
0.15	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	21	0.671	0.377	0.208
0.15	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	28	0.775	0.487	0.267
0.15	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	50	0.964	0.736	0.477

Web Table 7: Simulated power for bivariate continuous outcomes with $\rho = 0.2$ and two different similarity thresholds ε.

ε	$\boldsymbol{\theta}^{(2)}$	(d_1, d_2)	$n_g^{(l)}$	$\sigma^2 = 0.05$	$\sigma^2 = 0.1$	$\sigma^2 = 0.2$
0.2	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	7	0.281	0.167	0.098
0.2	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	14	0.512	0.257	0.174
0.2	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	21	0.719	0.412	0.205
0.2	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	28	0.842	0.535	0.276
0.2	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	50	0.983	0.800	0.472
0.2	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	7	0.403	0.241	0.141
0.2	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	14	0.710	0.383	0.222
0.2	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	21	0.885	0.598	0.281
0.2	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	28	0.962	0.723	0.396
0.2	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	50	0.998	0.940	0.658
0.2	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	7	0.431	0.246	0.144
0.2	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	14	0.745	0.453	0.236
0.2	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	21	0.887	0.623	0.332
0.2	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	28	0.968	0.735	0.446
0.2	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	50	0.999	0.935	0.684
0.15	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	7	0.119	0.111	0.107
0.15	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	14	0.171	0.113	0.132
0.15	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	21	0.255	0.149	0.102
0.15	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	28	0.334	0.173	0.144
0.15	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	50	0.483	0.296	0.155
0.15	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	7	0.229	0.136	0.125
0.15	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	14	0.390	0.235	0.145
0.15	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	21	0.606	0.306	0.184
0.15	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	28	0.745	0.378	0.228
0.15	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	50	0.961	0.693	0.377
0.15	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	7	0.255	0.159	0.132
0.15	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	14	0.494	0.260	0.161
0.15	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	21	0.667	0.368	0.209
0.15	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	28	0.790	0.461	0.280
0.15	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	50	0.965	0.731	0.453

Web Table 8: Simulated power for bivariate continuous outcomes with $\rho = 0.3$ and two different similarity thresholds ε.

ε	$\boldsymbol{\theta}^{(2)}$	(d_1, d_2)	$n_g^{(l)}$	$\sigma^2 = 0.05$	$\sigma^2 = 0.1$	$\sigma^2 = 0.2$
0.2	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	7	0.288	0.152	0.132
0.2	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	14	0.553	0.278	0.155
0.2	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	21	0.722	0.400	0.220
0.2	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	28	0.853	0.527	0.297
0.2	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	50	0.982	0.786	0.493
0.2	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	7	0.390	0.221	0.141
0.2	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	14	0.730	0.367	0.237
0.2	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	21	0.883	0.609	0.299
0.2	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	28	0.970	0.699	0.372
0.2	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	50	0.999	0.948	0.671
0.2	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	7	0.413	0.209	0.149
0.2	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	14	0.713	0.451	0.242
0.2	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	21	0.897	0.631	0.355
0.2	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	28	0.962	0.732	0.451
0.2	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	50	1.000	0.928	0.690
0.15	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	7	0.125	0.103	0.095
0.15	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	14	0.193	0.129	0.112
0.15	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	21	0.244	0.171	0.116
0.15	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	28	0.332	0.209	0.147
0.15	$(0, 0.8, 0.1, 0, 0.8, 0.1, \sigma, \rho)$	(0.1,0.1)	50	0.497	0.286	0.197
0.15	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	7	0.234	0.139	0.12
0.15	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	14	0.424	0.207	0.133
0.15	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	21	0.591	0.319	0.174
0.15	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	28	0.726	0.433	0.220
0.15	$(0, 0.9, 0.05, 0, 0.9, 0.05, \sigma, \rho)$	(0.05,0.05)	50	0.953	0.693	0.393
0.15	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	7	0.257	0.175	0.155
0.15	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	14	0.500	0.266	0.163
0.15	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	21	0.680	0.375	0.238
0.15	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	28	0.806	0.488	0.282
0.15	$(0, 1, 0, 0, 1, 0, \sigma, \rho)$	(0,0)	50	0.967	0.748	0.429

Web Table 9: Parameter scenarios used for the simulation of bivariate mixed outcomes.

No.		$\theta^{(1)}$	$\theta^{(2)}$	(d_1, d_2)
1	Null hypothesis	$(0, 1, -1, 2, \sigma, \rho)$	$(0, 0.6, 0.2, -2.4, 3.4, \sigma, \rho)$	$(0.2, 0.2)$
2	Null hypothesis	$(0, 1, -1, 2, \sigma, \rho)$	$(0, 0.6, 0.2, -1, 2, \sigma, \rho)$	$(0.2, 0)$
3	Null hypothesis	$(0, 1, -1, 2, \sigma, \rho)$	$(0, 1, 0, -2.4, 3.4, \sigma, \rho)$	$(0, 0.2)$
4	Null hypothesis	$(0, 1, -1, 2, \sigma, \rho)$	$(0, 0.7, 0.15, -2, 3.4, \sigma, \rho)$	$(0.15, 0.15)$
5	Null hypothesis	$(0, 1, -1, 2, \sigma, \rho)$	$(0, 0.7, 0.15, -1, 2, \sigma, \rho)$	$(0.15, 0)$
6	Null hypothesis	$(0, 1, -1, 2, \sigma, \rho)$	$(0, 1, 0, -2, 3.4, \sigma, \rho)$	$(0, 0.15)$
7	Alternative	$(0, 1, -1, 2, \sigma, \rho)$	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$
8	Alternative	$(0, 1, -1, 2, \sigma, \rho)$	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$
9	Alternative	$(0, 1, -1, 2, \sigma, \rho)$	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$

Web Table 10: Simulated type I error rates for bivariate mixed outcomes with $\rho = 0.1$ and two different similarity thresholds ε.

ε	$\boldsymbol{\theta}^{(2)}$	(d_1, d_2)	$n_g^{(l)}$	$\sigma^2 = 0.05$	$\sigma^2 = 0.1$	$\sigma^2 = 0.2$
0.2	$(0, 0.6, 0.2, -2.4, 3.4, \sigma, \rho)$	$(0.2, 0.2)$	7	0.039	0.036	0.041
0.2	$(0, 0.6, 0.2, -2.4, 3.4, \sigma, \rho)$	$(0.2, 0.2)$	14	0.021	0.022	0.027
0.2	$(0, 0.6, 0.2, -2.4, 3.4, \sigma, \rho)$	$(0.2, 0.2)$	21	0.016	0.021	0.017
0.2	$(0, 0.6, 0.2, -2.4, 3.4, \sigma, \rho)$	$(0.2, 0.2)$	28	0.012	0.013	0.014
0.2	$(0, 0.6, 0.2, -2.4, 3.4, \sigma, \rho)$	$(0.2, 0.2)$	50	0.009	0.011	0.008
0.2	$(0, 0.6, 0.2, -1, 2, \sigma, \rho)$	$(0.2, 0)$	7	0.084	0.081	0.083
0.2	$(0, 0.6, 0.2, -1, 2, \sigma, \rho)$	$(0.2, 0)$	14	0.097	0.081	0.098
0.2	$(0, 0.6, 0.2, -1, 2, \sigma, \rho)$	$(0.2, 0)$	21	0.058	0.074	0.069
0.2	$(0, 0.6, 0.2, -1, 2, \sigma, \rho)$	$(0.2, 0)$	28	0.079	0.087	0.088
0.2	$(0, 0.6, 0.2, -1, 2, \sigma, \rho)$	$(0.2, 0)$	50	0.065	0.056	0.068
0.2	$(0, 1, 0, -2.4, 3.4, \sigma, \rho)$	$(0, 0.2)$	7	0.104	0.097	0.089
0.2	$(0, 1, 0, -2.4, 3.4, \sigma, \rho)$	$(0, 0.2)$	14	0.085	0.078	0.067
0.2	$(0, 1, 0, -2.4, 3.4, \sigma, \rho)$	$(0, 0.2)$	21	0.053	0.079	0.077
0.2	$(0, 1, 0, -2.4, 3.4, \sigma, \rho)$	$(0, 0.2)$	28	0.051	0.079	0.075
0.2	$(0, 1, 0, -2.4, 3.4, \sigma, \rho)$	$(0, 0.2)$	50	0.044	0.066	0.060
0.15	$(0, 0.7, 0.15, -2, 3.4, \sigma, \rho)$	$(0.15, 0.15)$	7	0.050	0.037	0.065
0.15	$(0, 0.7, 0.15, -2, 3.4, \sigma, \rho)$	$(0.15, 0.15)$	14	0.017	0.025	0.038
0.15	$(0, 0.7, 0.15, -2, 3.4, \sigma, \rho)$	$(0.15, 0.15)$	21	0.024	0.027	0.025
0.15	$(0, 0.7, 0.15, -2, 3.4, \sigma, \rho)$	$(0.15, 0.15)$	28	0.011	0.016	0.036
0.15	$(0, 0.7, 0.15, -2, 3.4, \sigma, \rho)$	$(0.15, 0.15)$	50	0.017	0.007	0.026
0.15	$(0, 0.7, 0.15, -1, 2, \sigma, \rho)$	$(0.15, 0)$	7	0.078	0.074	0.072
0.15	$(0, 0.7, 0.15, -1, 2, \sigma, \rho)$	$(0.15, 0)$	14	0.070	0.090	0.071
0.15	$(0, 0.7, 0.15, -1, 2, \sigma, \rho)$	$(0.15, 0)$	21	0.079	0.091	0.087
0.15	$(0, 0.7, 0.15, -1, 2, \sigma, \rho)$	$(0.15, 0)$	28	0.061	0.094	0.066
0.15	$(0, 0.7, 0.15, -1, 2, \sigma, \rho)$	$(0.15, 0)$	50	0.068	0.057	0.071
0.15	$(0, 1, 0, -2, 3.4, \sigma, \rho)$	$(0, 0.15)$	7	0.103	0.092	0.088
0.15	$(0, 1, 0, -2, 3.4, \sigma, \rho)$	$(0, 0.15)$	14	0.085	0.067	0.061
0.15	$(0, 1, 0, -2, 3.4, \sigma, \rho)$	$(0, 0.15)$	21	0.070	0.069	0.062
0.15	$(0, 1, 0, -2, 3.4, \sigma, \rho)$	$(0, 0.15)$	28	0.065	0.059	0.065
0.15	$(0, 1, 0, -2, 3.4, \sigma, \rho)$	$(0, 0.15)$	50	0.048	0.058	0.055

Web Table 11: Simulated type I error rates for bivariate mixed outcomes with $\rho = 0.3$ and two different similarity thresholds ε. For some of the given scenarios data with $\rho = 0.3$ does not exist because $\rho = 0.3$ lies outside of the range of feasible correlations resulting from the marginal distributions (see Demirtas and Doganay, 2012, for details on these feasiblity ranges). This is indicated by a asterisk in the corresponding cells.

ε	$\boldsymbol{\theta}^{(2)}$	(d_1, d_2)	$n_g^{(l)}$	$\sigma^2 = 0.05$	$\sigma^2 = 0.1$	$\sigma^2 = 0.2$
0.2	$(0, 0.6, 0.2, -2.4, 3.4, \sigma, \rho)$	$(0.2, 0.2)$	all	*	*	*
0.2	$(0, 0.6, 0.2, -1, 2, \sigma, \rho)$	$(0.2, 0)$	7	0.087	0.094	0.093
0.2	$(0, 0.6, 0.2, -1, 2, \sigma, \rho)$	$(0.2, 0)$	14	0.068	0.088	0.071
0.2	$(0, 0.6, 0.2, -1, 2, \sigma, \rho)$	$(0.2, 0)$	21	0.077	0.070	0.071
0.2	$(0, 0.6, 0.2, -1, 2, \sigma, \rho)$	$(0.2, 0)$	28	0.066	0.070	0.093
0.2	$(0, 0.6, 0.2, -1, 2, \sigma, \rho)$	$(0.2, 0)$	50	0.049	0.053	0.065
0.2	$(0, 1, 0, -2.4, 3.4, \sigma, \rho)$	$(0, 0.2)$	all	*	*	*
0.15	$(0, 0.7, 0.15, -2, 3.4, \sigma, \rho)$	$(0.15, 0.15)$	all	*	*	*
0.15	$(0, 0.7, 0.15, -1, 2, \sigma, \rho)$	$(0.15, 0)$	7	0.066	0.084	0.080
0.15	$(0, 0.7, 0.15, -1, 2, \sigma, \rho)$	$(0.15, 0)$	14	0.078	0.087	0.077
0.15	$(0, 0.7, 0.15, -1, 2, \sigma, \rho)$	$(0.15, 0)$	21	0.060	0.086	0.089
0.15	$(0, 0.7, 0.15, -1, 2, \sigma, \rho)$	$(0.15, 0)$	28	0.078	0.085	0.086
0.15	$(0, 0.7, 0.15, -1, 2, \sigma, \rho)$	$(0.15, 0)$	50	0.052	0.080	0.074
0.15	$(0, 1, 0, -2, 3.4, \sigma, \rho)$	$(0, 0.15)$	all	*	*	*

Web Table 12: Simulated power for bivariate mixed outcomes with $\rho = 0.1$ and two different similarity thresholds ε.

ε	$\boldsymbol{\theta}^{(2)}$	(d_1, d_2)	$n_g^{(l)}$	$\sigma^2 = 0.05$	$\sigma^2 = 0.1$	$\sigma^2 = 0.2$
0.2	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	7	0.186	0.148	0.113
0.2	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	14	0.327	0.216	0.147
0.2	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	21	0.398	0.321	0.217
0.2	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	28	0.514	0.407	0.296
0.2	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	50	0.709	0.657	0.494
0.2	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	7	0.269	0.213	0.162
0.2	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	14	0.460	0.366	0.238
0.2	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	21	0.611	0.536	0.343
0.2	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	28	0.712	0.655	0.448
0.2	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	50	0.925	0.911	0.752
0.2	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	7	0.33	0.215	0.164
0.2	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	14	0.482	0.409	0.275
0.2	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	21	0.700	0.581	0.389
0.2	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	28	0.827	0.718	0.534
0.2	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	50	0.971	0.961	0.783
0.15	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	7	0.098	0.088	0.095
0.15	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	14	0.118	0.100	0.085
0.15	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	21	0.149	0.105	0.092
0.15	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	28	0.190	0.131	0.088
0.15	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	50	0.273	0.191	0.136
0.15	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	7	0.181	0.124	0.099
0.15	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	14	0.258	0.208	0.158
0.15	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	21	0.354	0.252	0.205
0.15	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	28	0.469	0.353	0.231
0.15	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	50	0.673	0.599	0.396
0.15	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	7	0.177	0.152	0.114
0.15	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	14	0.313	0.245	0.18
0.15	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	21	0.426	0.327	0.199
0.15	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	28	0.526	0.447	0.295
0.15	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	50	0.808	0.741	0.535

Web Table 13: Simulated power for bivariate mixed outcomes with $\rho = 0.2$ and two different similarity thresholds ε.

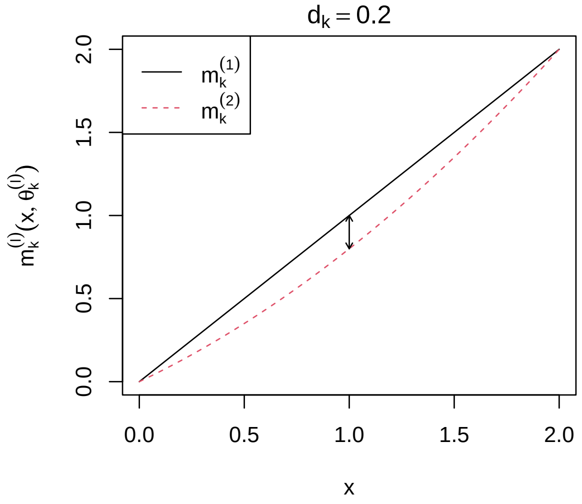
ε	$\boldsymbol{\theta}^{(2)}$	(d_1, d_2)	$n_g^{(l)}$	$\sigma^2 = 0.05$	$\sigma^2 = 0.1$	$\sigma^2 = 0.2$
0.2	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	7	0.176	0.155	0.111
0.2	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	14	0.325	0.214	0.144
0.2	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	21	0.431	0.336	0.218
0.2	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	28	0.500	0.413	0.279
0.2	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	50	0.712	0.679	0.439
0.2	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	7	0.274	0.193	0.152
0.2	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	14	0.452	0.368	0.250
0.2	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	21	0.619	0.503	0.363
0.2	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	28	0.737	0.664	0.460
0.2	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	50	0.902	0.894	0.702
0.2	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	7	0.275	0.214	0.160
0.2	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	14	0.506	0.405	0.269
0.2	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	21	0.693	0.591	0.382
0.2	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	28	0.813	0.719	0.523
0.2	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	50	0.984	0.956	0.791
0.15	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	7	0.085	0.091	0.092
0.15	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	14	0.144	0.091	0.094
0.15	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	21	0.162	0.116	0.094
0.15	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	28	0.189	0.134	0.104
0.15	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	50	0.296	0.216	0.166
0.15	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	7	0.147	0.137	0.122
0.15	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	14	0.260	0.185	0.150
0.15	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	21	0.372	0.293	0.188
0.15	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	28	0.466	0.338	0.210
0.15	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	50	0.630	0.616	0.415
0.15	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	7	0.208	0.143	0.138
0.15	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	14	0.332	0.268	0.176
0.15	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	21	0.413	0.328	0.237
0.15	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	28	0.526	0.464	0.297
0.15	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	50	0.807	0.763	0.548

Web Table 14: Simulated power for bivariate mixed outcomes with $\rho = 0.3$ and two different similarity thresholds ε.

ε	$\boldsymbol{\theta}^{(2)}$	(d_1, d_2)	$n_g^{(l)}$	$\sigma^2 = 0.05$	$\sigma^2 = 0.1$	$\sigma^2 = 0.2$
0.2	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	7	0.207	0.149	0.108
0.2	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	14	0.356	0.240	0.166
0.2	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	21	0.411	0.345	0.231
0.2	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	28	0.538	0.411	0.279
0.2	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	50	0.727	0.657	0.514
0.2	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	7	0.270	0.213	0.171
0.2	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	14	0.441	0.356	0.251
0.2	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	21	0.625	0.525	0.342
0.2	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	28	0.766	0.673	0.456
0.2	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	50	0.919	0.918	0.763
0.2	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	7	0.281	0.231	0.161
0.2	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	14	0.512	0.436	0.262
0.2	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	21	0.679	0.617	0.41
0.2	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	28	0.815	0.746	0.536
0.2	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	50	0.982	0.960	0.769
0.15	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	7	0.105	0.106	0.088
0.15	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	14	0.157	0.112	0.095
0.15	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	21	0.179	0.136	0.091
0.15	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	28	0.214	0.135	0.097
0.15	$(0, 0.8, 0.1, -1.5, 2.2, \sigma, \rho)$	$(0.1, 0.1)$	50	0.298	0.199	0.159
0.15	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	7	0.177	0.136	0.127
0.15	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	14	0.278	0.205	0.144
0.15	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	21	0.378	0.284	0.202
0.15	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	28	0.468	0.396	0.247
0.15	$(0, 0.9, 0.05, -1.2, 2, \sigma, \rho)$	$(0.05, 0.05)$	50	0.667	0.595	0.402
0.15	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	7	0.196	0.151	0.119
0.15	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	14	0.318	0.237	0.185
0.15	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	21	0.437	0.327	0.252
0.15	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	28	0.555	0.482	0.323
0.15	$(0, 1, 0, -1, 2, \sigma, \rho)$	$(0, 0)$	50	0.821	0.769	0.514

Web Table 15: Coefficient estimates for the case study.

		Marketed product	New product
Efficacy	Intercept	0.303	0.259
	Dose	0.715	0.416
	Dose^2	-0.369	0.062
Toxicity	Intercept	-2.492	-2.136
	Dose	1.797	1.263



Web Figure 1: Visualisation of the curves for the two groups $l = 1, 2$, where the maximum distance $d_k = 0.2$ is observed for $x = 1$ and corresponds to the length of the arrow.

References

Demirtas, H. and Doganay, B. (2012). Simultaneous generation of binary and normal data with specified marginal and association structures. *Journal of Biopharmaceutical Statistics*, 22(2):223–236.

3.2 Overcoming model uncertainty – how equivalence tests can benefit from model averaging

This article discusses how to adapt model-based equivalence tests to be capable of the second direction of multidimensionality, i.e. model uncertainty. A solution to this problem is proposed by flexibly extending model-based equivalence tests using model averaging in order to enable its applicability under model uncertainty. Precisely, model averaging is based on smooth BIC weights and a testing procedure that makes use of the duality between confidence intervals and hypothesis testing is introduced. The validity of the approach is demonstrated by means of a simulation study and its practical relevance is illustrated by a time-response case study with toxicological gene expression data.

Authorship:	First author
Coauthors:	Kathrin Möllenhoff
Contribution statement:	My contribution to this project was developing the method, programming it in R, implementing the simulation study, analyzing the case study data, preparing the figures and tables, and writing the first draft of the manuscript. The initial idea for this work came from Kathrin Möllenhoff, who also supported the development of the method. Discussions and revising were done together with Kathrin Möllenhoff.
Status:	Published
Journal:	Statistics in Medicine
DOI:	10.1002/sim.10309

Statistics in Medicine

WILEY

Statistics
in Medicine

RESEARCH ARTICLE OPEN ACCESS

Overcoming Model Uncertainty — How Equivalence Tests Can Benefit From Model Averaging

Niklas Hagemann | Kathrin Möllenhoff

Institute of Medical Statistics and Computational Biology (IMSB), Faculty of Medicine, University of Cologne, Cologne, Germany

Correspondence: Kathrin Möllenhoff (kathrin.moellenhoff@uni-koeln.de)

Received: 23 May 2024 | **Revised:** 24 October 2024 | **Accepted:** 2 December 2024

Funding: This work has been supported by the Research Training Group "Biostatistical Methods for High-Dimensional Data in Toxicology" (RTG 2624, P7) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, Project Number 427806116).

Keywords: bootstrap | dose–response models | gene expression | model averaging | model-based equivalence tests | time-response models

ABSTRACT

A common problem in numerous research areas, particularly in clinical trials, is to test whether the effect of an explanatory variable on an outcome variable is equivalent across different groups. In practice, these tests are frequently used to compare the effect between patient groups, for example, based on gender, age, or treatments. Equivalence is usually assessed by testing whether the difference between the groups does not exceed a pre-specified equivalence threshold. Classical approaches are based on testing the equivalence of single quantities, for example, the mean, the area under the curve or other values of interest. However, when differences depending on a particular covariate are observed, these approaches can turn out to be not very accurate. Instead, whole regression curves over the entire covariate range, describing for instance the time window or a dose range, are considered and tests are based on a suitable distance measure of two such curves, as, for example, the maximum absolute distance between them. In this regard, a key assumption is that the true underlying regression models are known, which is rarely the case in practice. However, misspecification can lead to severe problems as inflated type I errors or, on the other hand, conservative test procedures. In this paper, we propose a solution to this problem by introducing a flexible extension of such an equivalence test using model averaging in order to overcome this assumption and making the test applicable under model uncertainty. Precisely, we introduce model averaging based on smooth Bayesian information criterion weights and we propose a testing procedure which makes use of the duality between confidence intervals and hypothesis testing. We demonstrate the validity of our approach by means of a simulation study and illustrate its practical relevance considering a time-response case study with toxicological gene expression data.

1 | Introduction

In numerous research areas, particularly in clinical trials [1, 2], a common problem is to test whether the effect of an explanatory variable on an outcome variable is equivalent across different groups. Equivalence is usually assessed by testing whether the difference between the groups does not exceed a pre-specified equivalence threshold. The choice of this threshold is crucial as it resembles the maximal amount of deviation for which equivalence can still be concluded. One usually chooses the threshold based on prior knowledge, as a percentile of the range of the outcome variable or resulting from regulatory guidelines. Equivalence tests provide a flexible tool for plenty of research questions. For instance, they can be used to test for equivalence across patient groups, for example, based on gender or age, or between treatments. Moreover, they are a key ingredient of bioequivalence studies [3, 4], investigating whether two formulations of a drug have nearly the same effect and are hence considered to be interchangeable.

Classical approaches [5, 6] are based on testing the equivalence of single quantities, for example, the mean, the area under the curve (AUC) or other values of interest. However, when differences depending on a particular covariate are observed, these approaches can turn out to be not very accurate. Instead, considering the entire covariate range, describing for instance the time window or a dose range, has recently been proposed by testing equivalence of whole regression curves. Such tests [7–9] are typically based on the principle of confidence interval inclusion. However, a more direct approach applying various distance measures has been introduced by Dette et al. [10] which turned out to be particularly more powerful. Based on this, many further developments [11–14], for example, for different outcome distributions, specific model structures and/or responses of higher dimensions, have been introduced.

All these approaches have one thing in common: they base on the assumption that the true underlying regression model is known. In practice this usually implies that the models need to be chosen manually, either based on prior knowledge or visually. Hence, these approaches [15, 16] might not be robust with regard to model misspecification and, consequently, suffer from problems like inflated type I errors or reduced power. One idea to tackle this problem is implementing a testing procedure which explicitly incorporates the model uncertainty. This can be based on a formal model selection procedure, see, for example, Möllenhoff et al. [17] who propose conducting a classical model choice procedure prior to preforming the equivalence test.

An alternative to this is the incorporation of a model averaging approach into the test procedure. As outlined by Bornkamp [18] model selection has some disadvantages compared to model averaging. Particularly, model selection is not stable in the sense that minor changes in the data can lead to major changes in the results [19]. This also implies that model selection is non-robust with regard to outliers. In addition, the estimation of the distribution of post model selection parameter estimators is usually biased [20, 21]. Model averaging is omnipresent whenever model uncertainty is present, which is, besides other applications, often the case in parametric dose response analysis. Besides practical applications, there are also several methodological studies regarding model averaging in dose–response studies [22–24] and Bornkamp et al. [25] incorporated model averaging as an alternative to model selection in their widely used dose-finding method MCPMod.

Therefore, in this paper, we propose an approach utilizing model averaging rather than model selection. There are frequentist as well as Bayesian model averaging approaches. The former almost always use the smooth weights structure introduced by Buckland et al. [26] These weights depend on the values of an information criterion of the fitted models. Predominantly, the Akaike information criterion (AIC) [27] is used but other information criteria can be used as well. While only few of the Bayesian approaches perform fully Bayesian inference (see, e.g., Ley and Steel [28]), the majority makes use of the fact that the posterior model probabilities can be approximated by weights based on the Bayesian information criterion (BIC) [29] that have the same smooth weights structure as the frequentist weights [30]. Despite the prevalence of the AIC and BIC, other information criteria are sometimes used as well: Price et al. [31] suggested to use the deviance information criterion (DIC) [32], which is the Bayesian analog to the AIC. Hence, it bases on the samples of a Markov chain Monte Carlo simulation rather than on the log-likelihood. Hjort and Claeskens [33] introduced model averaging based on the focused information criterion (FIC) [34]. In contrast to other information criteria, the FIC does not aim for the best overall fit but focuses directly on a parameter of primary interest (e.g., the mean, the median or a specific quantile). Therefore, it favors models which lead to the best estimated precision with regard to this focus parameter. Occasionally, model averaging also bases on cross-validation, for example, jackknife model averaging [35], or machine learning methods, for example, random forests or boosting [36]. Alternatively, rather simple model averaging approaches with fixed model weights exist as well, for example, using equal weights. However, the performance of such approaches strongly depends on prior knowledge and can easily lead to (partial) model missspecification. For a more general introduction to model averaging techniques the reader is referred to, for example, Fletcher [37] or Claeskens and Hjort [38] and an overview specifically focusing on dose–response models is given by Schorning et al. [22]

In this paper, we propose an equivalence test incorporating model-averaging and hence overcoming the problems caused by model uncertainty. Precisely, we first make use of the duality between confidence intervals and hypothesis testing and propose a test based on the derivation of a confidence interval. By doing so, we both guarantee numerical stability of the procedure and provide confidence intervals for the measure of interest.

We demonstrate the usefulness of our method with the example of toxicological gene expression data. In this application, using model averaging enables us to analyze the equivalence of time–response curves between two groups for 1000 genes of interest without the necessity of specifying all 2000 correct models separately, thus avoiding both a time-consuming model selection step and potential model misspecifications.

The paper is structured as follows: In Section 2, dose–response models and the concept of model averaging are succinctly discussed. In Section 3, the testing approach is introduced, proposing three different variations. Finite sample properties in terms of Type I and II error rates are studied in Section 4. Section 5 illustrates the method using the toxicological gene expression example before Section 6 closes with a discussion.

2 | Model Averaging for Dose–Response Models

2.1 | Dose–Response Models

We consider two different groups, indicated by an index $l = 1, 2$, with corresponding response variables y_{lij} with $Y \subseteq R$ denoting the set of all possible outcomes. There are $i = 1, \ldots, I_l$ dose levels and $j = 1, \ldots, n_{li}$ denotes the observation index within each dose level. For each group the total number of observations is n_l and n

is the overall number of observations, that is, $n_l = \sum_{i=1}^{I_l} n_{li}$ and $n = n_1 + n_2$. For each group we introduce a flexible dose–response model

$$y_{lij} = m_l\left(x_{li}, \theta_l\right) + e_{lij}, \quad j = 1, \dots, n_{li}, \quad i = 1, \dots, I_l, \qquad l = 1, 2,$$

where $x_{li} \in \mathcal{X} \subseteq \mathbb{R}$ is the dose level, that is, the deterministic explanatory variable. We assume the error terms e_{lij} to be independent, have expectation zero and finite variance σ_l^2. The function $m_l(\cdot)$ models the effect of x_{li} on y_{lij} via a regression curve with $\theta_l \in R^{\dim(\theta_l)}$ being its parameter vector. We assume $m_l(\cdot)$ to be twice continuously differentiable. In dose–response studies, as well as in time-response studies, often either a linear model

$$m_l\left(x, \theta_l\right) = \beta_{l0} + \beta_{l1} x \tag{1}$$

a quadratic model

$$m_l\left(x, \theta_l\right) = \beta_{l0} + \beta_{l1} x + \beta_{l2} x^2 \tag{2}$$

an emax model

$$m_l\left(x, \theta_l\right) = \beta_{l0} + \beta_{l1} \frac{x}{\beta_{l2} + x} \tag{3}$$

an exponential (exp) model

$$m_l\left(x, \theta_l\right) = \beta_{l0} + \beta_{l1} \left(exp\left(\frac{x}{\beta_{l2}} \right) - 1 \right) \tag{4}$$

a sigmoid emax (sigEmax) model

$$m_l\left(x, \theta_l\right) = \beta_{l0} + \beta_{l1} \frac{x^{\beta_{l3}}}{\left(\beta_{l2}\right)^{\beta_{l3}} + x^{\beta_{l3}}} \tag{5}$$

also known as Hill model or 4pLL-model, or a beta model

$$m_l\left(x, \theta_l\right) = \beta_{l0} + \beta_{l1} \left(\frac{\left(\beta_{l2} + \beta_{l3}\right)^{\beta_{l2} + \beta_{l3}}}{\left(\beta_{l2}\right)^{\beta_{l2}} + \left(\beta_{l3}\right)^{\beta_{l3}}} \right) \left(\frac{x}{s} \right)^{\beta_{l2}} \left(1 - \frac{x}{s} \right)^{\beta_{l3}} \tag{6}$$

where s is a fixed scaling parameter, is deployed [39–42]. These models strongly vary in the assumed underlying dose–response relation, for example, in terms of monotonicity, and consequently in the shape of their curves. Therefore, choosing a suitable dose–response model is crucial for all subsequent analyses.

However, in practical applications the true underlying model shape is in general unknown. Thus, it might not always be clear which functional form of (1–6) should be imployed. A possible answer to this is implementing model averaging which, as outlined in Section 1, has several advantages over the simpler alternative of model selection.

2.2 | Model Averaging

As outlined before, frequentist as well as Bayesian model averaging approaches usually both use the same smooth weights structure introduced by Buckland et al. [26] and Wasserman [30], respectively. Accordingly, by leaving out the group index $l = 1, 2$ for better readability the averaged model is given by

$$m(x, \hat{\theta}) := \sum_{k=1}^{K} w_k m_k\left(x, \hat{\theta}_k\right) \tag{7}$$

where the $m_k\left(x, \hat{\theta}_k\right)$, $k = 1, .., K$, correspond to the K candidate models,

$$w_k = \frac{\exp(-0.5 I\left(m_k\left(x, \hat{\theta}_k\right)\right)}{\sum_{\tilde{k}=1}^{K} \exp\left(-0.5 I\left(m_{\tilde{k}}\left(x, \hat{\theta}_{\tilde{k}}\right)\right)\right)} \tag{8}$$

are the corresponding weights and $I(\cdot)$ is an information criterion with smaller values corresponding to better model fit. All information criteria considered here are based on the calculation of a penalized log-likelihood for each candidate model. Usually the AIC is used for frequentist model averaging, while the BIC is usually deployed for Bayesian model averaging [22]. A notable special case arises when the number of parameters is the same for all candidate models: the smooth AIC and smooth BIC weights are exactly equal in this situation, as the penalty term vanishes from Equation (8). In this case the weights only depend on the value of the respective log-likelihood.

2.3 | Inference

As the parameter estimation is conducted for each of the candidate models separately, it is not influenced by the subsequent model averaging. Therefore, here the index k is left out. Inference can be based on an ordinary least squares (OLS) estimator, that is, minimizing

$$\sum_{i=1}^{I_l} \sum_{j=1}^{n_{li}} \left(y_{lij} - m_l\left(x_{li}, \theta_l\right)\right)^2, \qquad l = 1, 2$$

In general, no distributional assumption is needed regarding the error terms, they just need to be independent, have expectation zero and finite variance σ_l^2 as outlined in Section 2.1. However, by making a distributional assumption, a maximum likelihood estimator can also be deployed. Usually, normality of the error terms, that is

$$e_{lij} \overset{iid}{\sim} N\left(0, \sigma_l^2\right), \quad l = 1, 2$$

with log-likelihood

$$\ell\left(\theta_l, \sigma_l^2\right) = -\frac{n_l}{2} \ln\left(2\pi\sigma_l^2\right) - \frac{1}{2\sigma_l^2} \sum_{i=1}^{I_l} \sum_{j=1}^{n_{li}} \left(y_{lij} - m_l\left(x_{li}, \theta_l\right)\right)^2, \quad l = 1, 2 \tag{9}$$

is assumed but other distributions can be considered as well. Under normality both approaches are identical and, hence, lead to the same parameter estimates $\hat{\theta}_l$. From (9) a maximum likelihood estimator for the variance

$$\hat{\sigma}_l^2 = \frac{1}{n_l} \sum_{i=1}^{I_l} \sum_{j=1}^{n_{li}} \left(y_{lij} - m_l\left(x_{li}, \hat{\theta}_l\right)\right)^2, \quad l = 1, 2 \tag{10}$$

can be derived as well. In R inference is performed with the function fitMod from the package DoseFinding [25, 41] which performs OLS estimation. The value of the log-likelihood needed for the AIC or BIC is then calculated by plugging the OLS estimator into the log-likelihood (9).

Even though smooth AIC and smooth BIC weights share the same structure, there are differences regarding their asymptotic properties. As outlined by several authors [33, 38, 43], the asymptotic distribution of model average estimators is in general no longer a normal distribution due to being a non-linear transformation of normal distributions. An exception to this is model averaging with fixed weights: due to the weights being non-random, the asymptotic distribution of the model average estimator is a linear combination of normal distributions and, hence, also a normal distribution.

Wang et al. [43] investigated asymptotic properties for smooth AIC and smooth BIC weights explicitly. Under regulatory assumptions, the asymptotic distribution of the model average estimator using smooth AIC weights is in general non-normal. In contrast, using smooth BIC weights leads to the asymptotic distribution of the model average estimator being a normal distribution. With regard to our study, it will turn out that this is an important advantage of smooth BIC weights, as asymptotic normality is used in order to show the asymptotic validity of the testing approach introduced in Section 3. In addition, the BIC-based weights provide additional interpretability due to being approximately equal to the posterior model probabilities.

Comparing the different information criteria, we can summarize that smooth BIC weights are advantageous in terms of their ability to be integrated into the framework of model-based equivalence testing. For smooth BIC weights, there is sufficient asymptotic theory to justify the asymptotic validity of the test. In contrast, smooth AIC weights do not provide the necessary asymptotic properties to guarantee a theoretical justification. Smooth DIC weights are not applicable because frequentist inference is performed. Smooth FIC weights are problematic because there is no focus parameter and one searches for the overall best fitting curves. Asymptotic theory is not (yet) available for most cross-validation or machine learning methods. Finally, fixed weights depend on prior knowledge and can lead to (partial) misspecification of the model. Therefore, we will use smooth BIC weights for the rest of this paper.

3 | Model-Based Equivalence Tests Under Model Uncertainty

3.1 | Equivalence Testing Based on Confidence Intervals

Model-based equivalence tests [10, 44] have been introduced in terms of the L^2-distance, the L^1-distance or the maximal absolute deviation (also called L^∞-distance) of the model curves. Although all of these approaches have their specific advantages and disadvantages as well as specific applications, subsequent research [11, 12, 14, 17] is predominately based on the maximal absolute deviation due to its easy interpretability. Accordingly, we state the hypotheses

$$H_0 : d \geq \varepsilon \text{ vs. } H_1 : d < \varepsilon \tag{11}$$

of equivalence of regression curves with respect to the maximal absolute deviation, that is

$$d = \max_{x \in \mathcal{X}} \left| m_1(x, \theta_1) - m_2(x, \theta_2) \right|$$

is the maximal absolute deviation of the curves and ε is the pre-specified equivalence threshold, meaning that a difference of ε is believed not to be clinically relevant. The test statistic is given as the estimated maximal deviation between the curves

$$\hat{d} = \max_{x \in \mathcal{X}} \left| m_1\left(x, \hat{\theta}_1\right) - m_2\left(x, \hat{\theta}_2\right) \right|. \tag{12}$$

As the distribution of $\hat{d}$ under the null hypothesis is in general unknown, it is usually either approximated based on a parametric bootstrap procedure or by asymptotic theory. In Dette et al. [10], the asymptotic validity of both approaches is proven, but the corresponding simulation study shows that the bootstrap test outperforms the asymptotic test in finite samples. For the bootstrap test several studies [10, 11, 17] show reasonable results for finite samples across applications.

However, in light of practical application, this approach can have two disadvantages: First, it does not directly provide confidence intervals (CI) which provide useful information about the precision of the test statistic. Further, they would have an important interpretation analogously to their interpretation in classical equivalence testing known as TOST [5] (two one-sided tests), where the bounds of the confidence interval are typically compared to the confidence region of $[-\varepsilon, \varepsilon]$.

Second, it requires the estimation of the models under the constraint of being on the edge of the null hypothesis, that is, the maximal absolute deviation being equal to ε (see Algorithm 1 in Dette et al. [10]). Technically, this is usually conducted using augmented Lagrangian optimisation. However, with increasing model complexity, this becomes numerically challenging. In the context of model averaging, these numerical issues are particularly relevant since all models would need to be estimated jointly as they need to jointly fulfill the constraint. This leads to a potentially high dimensional optimisation problem with a large number of parameters. In addition, for model averaging the side constraint has a highly complex structure because with every parameter update not only the model curves change but also the model weights do.

As an alternative to approximating the distribution under the null hypothesis, we propose to test hypotheses (11) based on the well-known duality between confidence intervals and hypothesis testing [45]. This testing approach is similar to what Bastian et al. [44] introduced for the L^1-distance of regression models. Therefore, let $(-\infty, u]$ be a one-sided lower $(1-\alpha)$-CI for d which we can rewrite as $[0, u]$ due to the non-negativity of d, that is

$$\mathbb{P}(d \leq u) = \mathbb{P}(d \in (-\infty, u]) = \mathbb{P}(d \in [0, u]) \geq 1 - \alpha$$

According to the duality between CI and hypothesis testing, we reject the null hypothesis and conclude equivalence if

$$\varepsilon > u \tag{13}$$

ALGORITHM 1 |

1. Obtain parameter estimates $\hat{\theta}_{lk}, k = 1, \dots, K_l, l = 1, 2$, for the candidate models, either via OLS or maximum likelihood optimisation (see Section 2.3). Determine the averaged models from the candidate models using Equation (7), that is, by calculating

$$m_l\left(x, \hat{\theta}_l\right) = \sum_{k=1}^{K_l} w_{lk} m_{lk}\left(x, \hat{\theta}_{lk}\right), \quad l = 1, 2,$$

with weights (8) as well as the variance estimator $\hat{\sigma}_l^2, l = 1, 2$ from Equation (10). Alternatively, use fixed weights instead of weighting scheme (8).

2. Calculate the test statistic (12).

3. Execute the following steps:
   a. Obtain bootstrap samples by generating data according to the model parameters $\hat{\theta}_l = \left(\hat{\theta}_{l1}, \dots, \hat{\theta}_{lK}\right), l = 1, 2$, and the weights $w_{l1}, \dots, w_{lK}, l = 1, 2$, obtained in step 1. Under the assumption of normality, that is

$$y^*_{lij} \sim N\left(\hat{\mu}_{li}, \hat{\sigma}_l^2\right), \quad j = 1, \dots, n_{li}, i = 1, \dots, I_l, l = 1, 2$$

   where

$$\hat{\mu}_{li} = m_l\left(x_{li}, \hat{\theta}_l\right) = \sum_{k=1}^{K_l} w_{lk} m_{lk}\left(x_{li}, \hat{\theta}_{lk}\right), \quad i = 1, \dots, I_l, l = 1, 2$$

   Alternative distributions with corresponding mean and variance can be used as well.

   b. From the bootstrap samples, estimate the models $m_l\left(x_{li}, \hat{\theta}^*_l\right), l = 1, 2$ as in step (1) and the test statistic

$$\hat{d}^* = \max_{x \in \mathcal{X}} \left| m_1\left(x, \hat{\theta}^*_1\right) - m_2\left(x, \hat{\theta}^*_2\right) \right|. \tag{15}$$

   c. Repeat steps (3.a) and (3.b) n_{boot} times to generate replicates $\hat{d}^*_1, \dots, \hat{d}^*_{n_{boot}}$ of $\hat{d}^*$. Let $\hat{d}^*_{(1)} \leq \dots \leq \hat{d}^*_{(n_{boot})}$ denote the corresponding order statistic.

4. Calculate the CI using one of the following approaches:
   a. *Percentile CI*: Obtain the estimated right bound of the percentile bootstrap CI as the $(1 - \alpha)$ -quantile of the bootstrap sample

$$\hat{u} = \hat{q}^*(1 - \alpha) = \hat{d}^*_{(\lfloor n_{boot}(1-\alpha) \rfloor)}.$$

   b. *Hybrid CI*: Obtain the estimator for the standard error of $\hat{d}$ as $\widehat{\mathrm{se}}(\hat{d}) = \sqrt{\widehat{\mathrm{Var}}\left(\hat{d}^*_1, \dots, \hat{d}^*_{n_{boot}}\right)}$ and the estimated right bound of the hybrid CI as

$$\hat{u} = \hat{d} + \widehat{\mathrm{se}}(\hat{d}) z$$

5. Reject the null hypothesis in (11) and assess equivalence if

$$\varepsilon > \hat{u}$$

This testing procedure is an α-level test as

$$\mathbb{P}_{H_0}(\varepsilon > u) \leq \mathbb{P}(d > u) = 1 - \mathbb{P}(d \leq u) \leq 1 - (1 - \alpha) = \alpha$$

However, as the distribution of $\hat{d}$ is in general unknown, obtaining u is again a challenging problem. It is obvious from (13) that the quality of the testing procedure crucially depends on the quality of the estimator for u. If the CI is too wide the test procedure is conservative and lacks power. In contrast, a too narrow CI can lead to type I error inflation due to not reaching the desired coverage probability $1 - \alpha$. We propose three different possibilities to calculate the CI, namely

1. CI based on a parametric percentile bootstrap,
2. asymptotic CI based on the asymptotic distribution of $\hat{d}$ derived by Dette et al. [10], and
3. a hybrid approach using the asymptotic normality of $\hat{d}$ but estimating its standard error based on a parametric bootstrap.

One-sided CIs based on a parametric percentile bootstrap can be constructed in the same way Möllenhoff et al. [17] proposed for two-sided CIs. In order to do so, they obtain parameter estimates (either via OLS or maximum likelihood optimisation), generate bootstrap data from these estimates and calculate the percentiles from the ordered bootstrap sample. The resulting test is similar to what Bastian et al. [44] derived for the L^1-distance of regression models. That is

$$\left[0, \hat{q}^*(1-\alpha)\right],$$

where $\hat{q}^*(1-\alpha)$ denotes the $(1-\alpha)$-quantile of the ordered bootstrap sample.

Asymptotic CIs can be derived directly from test (5.4) of Dette et al. [10] and are given by

$$\left[0, \hat{d} + \sqrt{\frac{\widehat{\mathrm{Var}}(d)}{n}}z\right] \tag{14}$$

where z is the $(1-\alpha)$-quantile of the standard normal distribution and $\widehat{\mathrm{Var}}(d)$ is the closed-form estimator for the variance of d given by equation (4.7) of Dette et al. [10] However, the asymptotic validity of this variance estimator is only given under the assumption that within $\mathcal{X}$ there is only one unique value x_0 where the absolute difference curve attains its maximum, that is, $x_0 = \mathrm{argmax}_{x\in\mathcal{X}}\left|m_1\left(x,\theta_1\right) - m_2\left(x,\theta_2\right)\right|$ and, moreover, that this value x_0 is known. This does not hold in general as Dette et al. [10] give two explicit counterexamples in terms of two shifted emax or exponential models. In addition, in practical applications x_0 is in general not known and needs to be estimated. If the absolute deviation along x is small, the estimation of x_0 can become unstable leading to an unstable variance estimator. Moreover, as mentioned before, the simulation study of Dette et al. [10] shows that for finite samples the bootstrap test is superior to the asymptotic test.

Given the disadvantages of the asymptotic CI and especially of the underlying variance estimator, we introduce a hybrid approach which is a combination of both approaches. It is based on the asymptotic normality of $\hat{d}$ but estimates the standard error of $\hat{d}$ based on a parametric bootstrap leading to

$$[0, \hat{d} + \widehat{\mathrm{se}}(\hat{d})z]$$

where the estimator $\widehat{\mathrm{se}}(\hat{d})$ of the standard error of $\hat{d}$ is the empirical standard deviation of the bootstrap sample.

Under the assumptions introduced by Dette et al. [10] all three approaches are asymptotically valid. For the test based on the asymptotic CI, this follows directly from Dette et al. [10] This also applies to the hybrid CI-based test due to $\widehat{\mathrm{se}}(\hat{d})$ being an asymptotically unbiased estimator for the standard error of $\hat{d}$ as outlined by Efron and Tibshirani [46]. The asymptotic validity of the percentile approach follows from Dette et al. [10] (Appendix: proof of Theorem 4). The finite sample properties of the three methods are compared in Section 4.1.

3.2 | Model-Based Equivalence Tests Incorporating Model Averaging

We now combine the model averaging approach presented in Section 2.2 with the CI-based test introduced in Section 3.1. For the asymptotic test, that is, estimating $m_l\left(x,\hat{\theta}_l\right), l = 1,2$ using Equation (7) with model weights (8) and then calculating the test statistic (12). Subsequently, the asymptotic CI (14) can be determined using the closed form variance estimator given by Dette et al. [10]. Using this CI, the test decision is based on decision rule (13).

The testing procedure of the percentile and hybrid approach is shown in Algorithm 1, where the first two steps are essentially the same as for the asymptotic test. The percentile test is conducted by performing Algorithm step 4a, while conducting step 4b instead leads to the hybrid test. In the following we will refer to this as Algorithm 1a and Algorithm 1b, respectively.

The asymptotic validity discussed at the end of Section 3.1 only transfers to averaged models if the asymptotic distribution of the model average estimator is normal. As outlined in Section 2.3, this is given for smooth BIC weights as well as for fixed weights.

4 | Finite Sample Properties

In the following we investigate the finite sample properties of the proposed tests by a simulation study. In order to ensure comparability, we reanalyze the simulation scenarios given by Dette et al. [10] The dose range is given by $\mathcal{X} = [0,4]$ and data is observed for dose levels $x = 0,1,2,3$ and 4 with equal number of observations $n_{li} = \frac{n_l}{5}$ for each dose level. All three simulation scenarios use the same three variance configurations $\left(\sigma_1^2, \sigma_2^2\right) \in \{(0.25, 0.25), (0.25, 0.5), (0.5, 0.5)\}$ as well as the same four different sample sizes $\left(n_1, n_2\right) \in \{(10,10), (10,20), (20,20), (50,50)\}$ and the same significance level of $\alpha = 0.05$.

In the first simulation scenario the equivalence of an emax model and an exponential model is investigated. The other two simulation scenarios consist of testing for equivalence of two shifted models, either both being emax models or both being exponential models. In contrast to the first scenario where the absolute deviation of the models is observed at one unique x_0, this is not the case for the latter two scenarios. Here, the deviation of both models is constant across the whole dose range $\mathcal{X} = [0,4]$, that is

$$\left|m_1\left(x,\theta_1\right) - m_2\left(x,\theta_2\right)\right| = d \;\forall\; x \in \mathcal{X}$$

as m_1, m_2 are just shifted.

Hence, for these two scenarios the asymptotic test is not applicable as its close form variance estimator bases on the uniqueness of x_0. Therefore, only simulation Scenario 1 is used to compare the three CI-based tests to each other as well as to the results observed by Dette et al. [10] Subsequently, all three scenarios are used to compare the performance of the test using model averaging to the one based on the correct specification of the models as well as under model misspecification.

4.1 | Finite Sample Properties of Confidence Interval-Based Equivalence Testing

Prior to the investigation of the effect of model averaging onto the finite sample properties, we first inspect the performance of the

CI-based testing approach, denoted by (13), itself. For the asymptotic test, the CI is defined by (14). For the percentile as well as the hybrid approach, the tests are conducted as explained in Section 3.1 which is formally defined by setting $K_1 = K_2 = 1$ in Algorithm 1.

As outlined before, simulation scenario 1 of Dette et al. [10] is given by tesing for the equivalence of an emax model (3) with $\theta_1 = \left(\beta_{10}, \beta_{11}, \beta_{12}\right) = (1, 2, 1)$ and an exponential model (4) with $\theta_2 = \left(\beta_{20}, \beta_{21}, \beta_{22}\right) = \left(\beta_{20}, 2.2, 8\right)$. It consists of 60 sub-scenarios resulting from the three different variance configurations each being combined with the four different sample sizes and five different choices of $\beta_{20} \in \{0.25, 0.5, 0.75, 1, 1.5\}$, leading to the corresponding deviations of the regression curves being $d \in \{1.5, 1.25, 1, 0.75, 0.5\}$. The test is conducted for $\epsilon = 1$ such that the first three deviations are under the null hypothesis and, therefore, used to investigate the type I error rates. The latter two deviations correspond to the alternative and are used to estimate the power of the tests.

As the type I error rates are always smaller than the nominal level of $\alpha = 0.05$ for all three approaches (see Table S1 of the Supporting Information for exact values), that is, all testing approaches always hold the nominal level, the following analysis focuses on the power of the tests. Figure 1 shows the power for all three tests for all sub-scenarios under the alternative as well as the corresponding power of the tests of Dette et al. [10] In each sub-scenario we observe that the hybrid test has superior power compared to the other two CI-based tests. In addition, one can observe that the power achieved by the hybrid test is quite similar to the one Dette et al. [10] observed for their bootstrap test and, therefore, is also superior to the power of their asymptotic test. The power of the test based on the percentile CI is considerably smaller which indicates that the test might be overly conservative in finite samples. The test based on the asymptotic CI leads to nearly the same results as the asymptotic test of Dette et al. [10] which is not surprising as it is directly derived from it. Consequentially, the lack of power that Dette et al. [10] observed for their asymptotic test in comparison to their bootstrap test is also present for the test based on the asymptotic CI.

In conclusion, the hybrid approach which provides numerical advantages compared to the bootstrap test of Dette et al. [10] and also leads to additional interpretability due to providing CIs, achieves nearly the same power as the bootstrap test while holding the nominal level.

4.2 | Finite Sample Properties Under Model Uncertainty

We now investigate the finite sample properties under model uncertainty. Due to the clear superiority of the hybrid approach observed in Section 4.1, only the hybrid test is used for this analysis. We compare the performance of the test using model averaging to the one based on the correct specification of the models as well as under model misspecification. We use Bayesian model averaging with smooth BIC weights due to its theoretical advantages outlined in Section 3.2. However, as both candidate models have the same number of parameters, smooth BIC and smooth AIC weights are exactly equal as discussed in Section 2.2. For comparison, we additionally conduct the test based on model averaging with fixed equal weights, that is, $w_{l1} = w_{l2} = 0.5, l = 1, 2$. The corresponding equivalence tests are conducted using Algorithm 1b.

4.2.1 | Comparison of an Emax With an Exponential Model

First, we again investigate the first simulation scenario introduced in Section 4.1 but now under model uncertainty where it is unclear if an emax or an exponential model applies for each of the groups implying $K_1 = K_2 = 2$ and leading to one correct

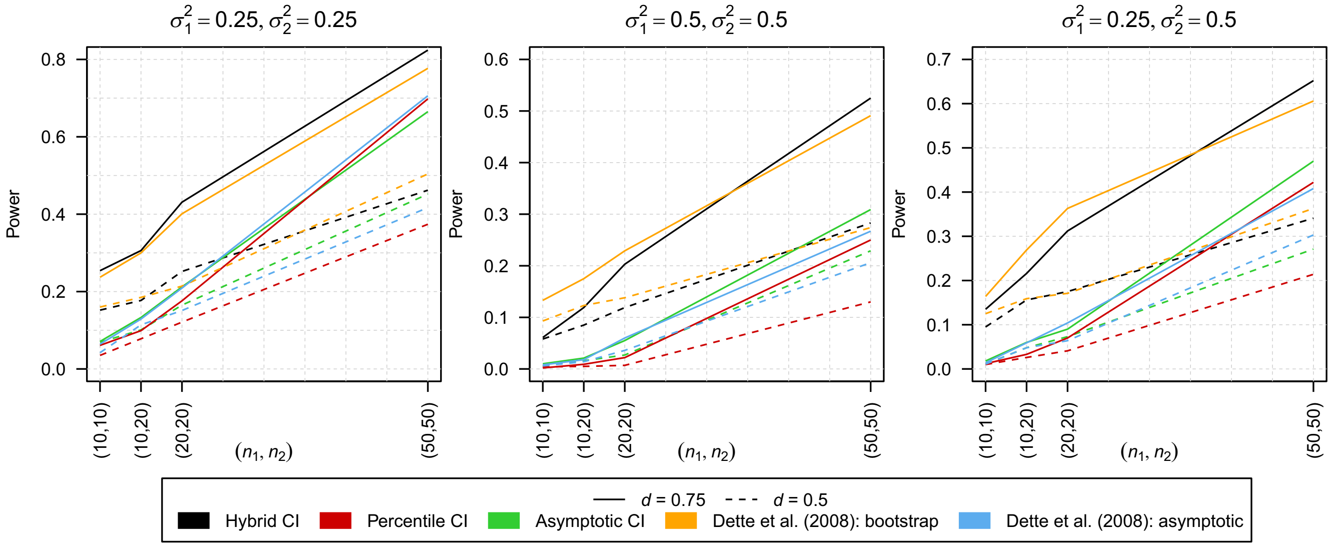


FIGURE 1 | Comparison of the power of the CI-based testing approaches to the testing approaches proposed by Dette et al. [10] with $\epsilon = 1$. The results are shown for two distances of the regression curves $d \in \{0.5, 0.75\}$ and three different combinations of variances $\left(\sigma_1^2, \sigma_2^2\right) \in \{(0.25, 0.25), (0.5, 0.5), (0.25, 0.5)\}$.

specification, as well as three misspecifications. Figure 2 shows the corresponding type I error rates for all sub-scenarios under the null hypothesis, that is, for $d \in \{1, 1.25, 1.5\}$ and $\left(\sigma_1^2, \sigma_2^2\right) \in \{(0.25, 0.25), (0.25, 0.5), (0.5, 0.5)\}$ for the correct specification, the three misspecifications as well as under model averaging.

One can observe that falsely specifying the same model for both responses leads to highly inflated type I error rates which, in addition, even increase for increasing sample size. The highest type I errors are present if an exponential model is specified for both responses leading to type I error rates being as large as 0.410 which is observed for $\sigma_1^2 = \sigma_2^2 = 0.25$ and $n_1 = n_2 = 50$. If an emax model is specified for both responses, the type I error inflation is smaller but still present and reaches up to 0.187 which is observed for $\sigma_1^2 = \sigma_2^2 = 0.25$ and $n_1 = n_2 = 50$. The third misspecification under investigation is that specifying the models the wrong way round, that is, an exponential model for the first group and an emax model for the second one. In comparison to the other two misspecifications, this leads to less extreme results but type I error inflation is still observable. This results from the fact that using one convex (exponential) and one concave (emax) model usually leads to a larger maximal absolute deviation than using two convex or two concave models and, therefore, in general to fewer rejections of the null hypothesis.

Compared to these results, the type I errors resulting from model averaging with smooth BIC weights are closer to the nominal level of the test. However, for two out of the 36 investigated sub-scenarios ($\sigma_1^2 = \sigma_2^2 = 0.25$ and $n_1 = n_2 = 20$ as well as $n_1 = n_2 = 50$) the type I errors still exceeds the nominal level but to a much lesser extent compared to model misspecification, reaching a maximum of 0.061. In contrast, using model averaging with equal weights leads to a high type I error inflation similar to the one observed under model misspecification. As expected, when using the true underlying model, the test holds the nominal level of $\alpha = 0.05$. Comparison of the power of the tests is not meaningful as some of them are not holding the nominal level. However, the estimated power is shown in Table S2 of the Supporting Information.

4.2.2 | Comparison of Two Shifted Emax Models

We continue by investigating the fine sample properties for the case of two shifted emax models, that is, model (3) now applies for both groups, where $\theta_1 = \left(\beta_{10}, \beta_{11}, \beta_{12}\right) = \left(\beta_{10}, 5, 1\right)$ and $\theta_2 = \left(\beta_{20}, \beta_{21}, \beta_{22}\right) = (0, 5, 1)$, which implies $d = \beta_{10}$. The levels of d under investigation are 1, 0.75, 0.5, 0.25, 0.1 and 0. The test is conducted for $\epsilon = 0.5$ such that the first three deviations are under the null hypothesis and, therefore, used to investigate the type I error rates. The latter three deviations are under the alternative and used to estimate the power of the tests.

We only observe few type I error rates which are non-zero and these are still much smaller than the nominal level of $\alpha = 0.05$, reaching a maximum of only 0.003 (all values can be found in Table S3 of the Supporting Information). Hence, the analysis focuses on the power of the tests which is shown in Figure 3.

One can observe falsely specifying one of the models to be an exponential model leads to the power being constantly equal to zero even for sub-scenarios which are quite far under the alternative. For misspecification in terms of using an exponential model for both responses, the power loss is not that extensive but still occurs for smaller sample sizes, which is especially visible for $\sigma_1^2 = \sigma_2^2 = 0.25$ due to the estimation uncertainty being the smallest. In contrast, model averaging with smooth BIC weights results in nearly the same power as using the true model. This also leads to the fact that in some cases the black line is even hardly visual as it is nearly perfectly overlapped by the green one. The use of model averaging with equal weights leads to high power, even exceeding the power obtained by using the true models. Therefore, this high number of rejections of the null hypothesis may

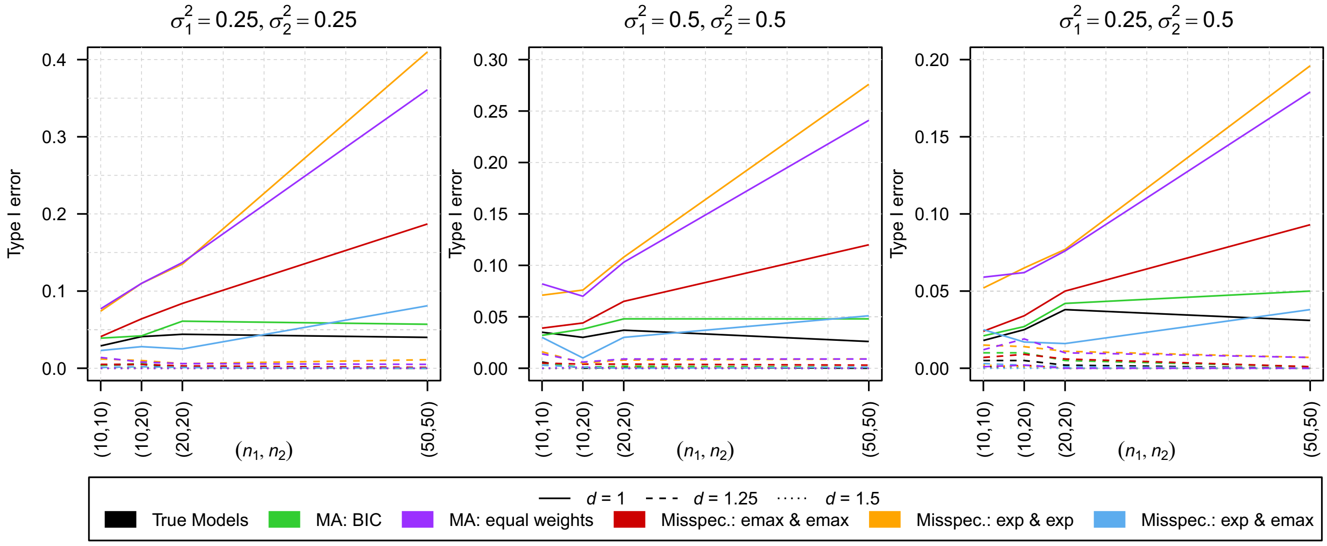


FIGURE 2 | Comparison of the type I error rates of the test using the true model, the model averaging-based test with smooth BIC weights (MA: BIC), the model averaging-based test with equal weights (MA: equal weights) and the tests under model misspecification in scenario 1. The results are shown for $\epsilon = 1$, three distances of the regression curves $d \in \{1,1.25,1.5\}$ and three different combinations of variances $\left(\sigma_1^2, \sigma_2^2\right) \in \{(0.25, 0.25), (0.25, 0.5), (0.5, 0.5)\}$.

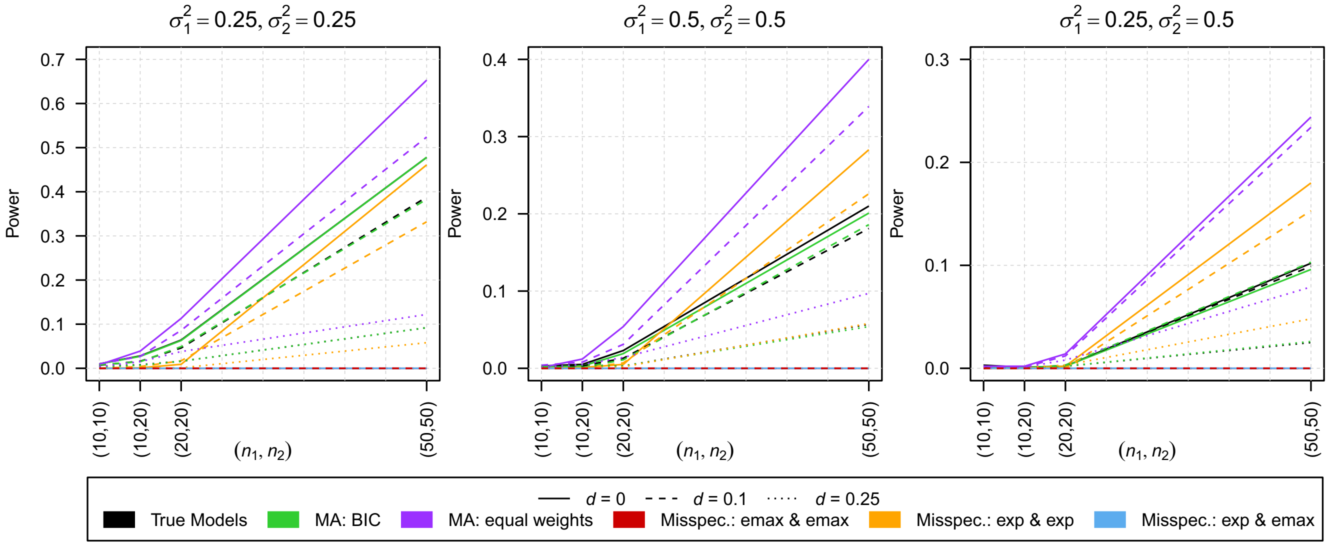


FIGURE 3 | Comparison of the power of the test using the true model, the model averaging-based test with smooth BIC weights (MA: BIC), the model averaging-based test with equal weights (MA: equal weights) and the tests under model misspecification in scenario 2. The results are shown for $\varepsilon = 0.5$, three distances of the regression curves $d \in \{1, 1.25, 1.5\}$ and three different combinations of variances $\left(\sigma_1^2, \sigma_2^2\right) \in \{(0.25, 0.25), (0.25, 0.5), (0.5, 0.5)\}$.

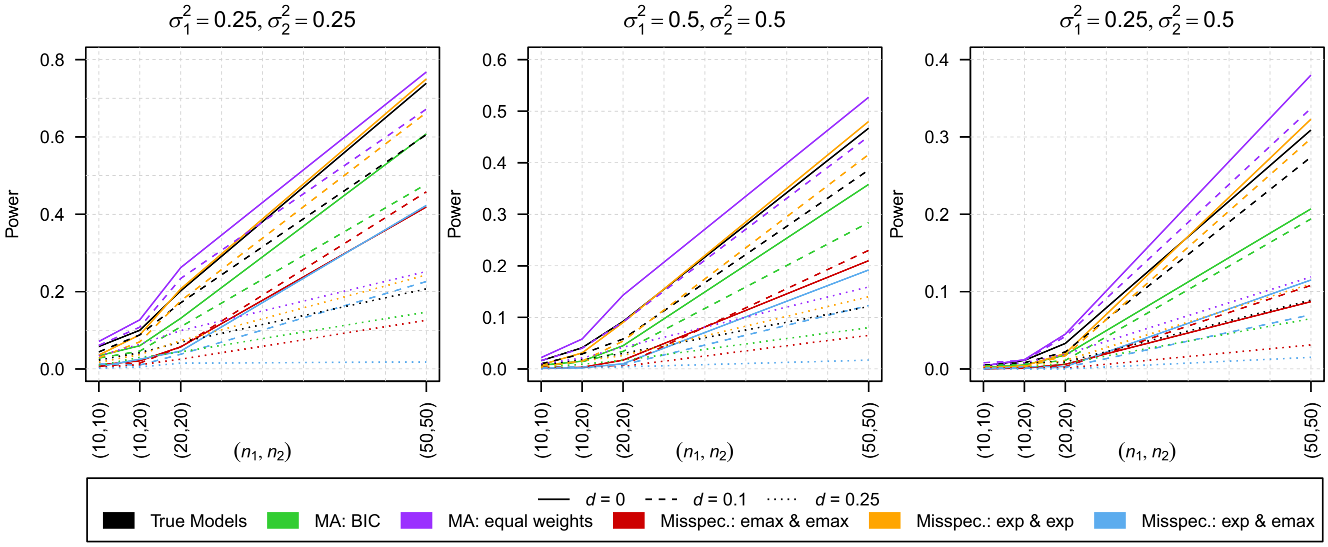


FIGURE 4 | Comparison of the power of the test using the true model, the model averaging-based test with smooth BIC weights (MA: BIC), the model averaging-based test with equal weights (MA: equal weights) and the tests under model misspecification in scenario 3. The results are shown for $\varepsilon = 0.5$, three distances of the regression curves $d \in \{1, 1.25, 1.5\}$ and three different combinations of variances $\left(\sigma_1^2, \sigma_2^2\right) \in \{(0.25, 0.25), (0.25, 0.5), (0.5, 0.5)\}$.

not be caused by the model fitting the data well, but by forcing the data into a (partly) wrong model, similar to what can be observed in some scenarios when using two exponential models.

4.2.3 | Comparison of Two Shifted Exponential Models

The third simulation scenario is given by two shifted exponential models, that is, model (4) now applies for both groups, where $\theta_1 = \left(\beta_{10}, \beta_{11}, \beta_{12}\right) = \left(\beta_{10}, 2.2, 8\right)$ and $\theta_2 = \left(\beta_{20}, \beta_{21}, \beta_{22}\right) = (0, 2.2, 8)$ which implies $d = \beta_{10}$, resulting in the same values for d as in Section 4.2.2. The test is conducted for $\varepsilon = 0.5$ such that the first three deviations are under the null hypothesis and, therefore, used to investigate the type I error rates. The latter three deviations are under the alternative and used to estimate the power of the tests.

As previously observed in Section 4.2.2 only few type I error rates are non-zero and these exceptions are still much smaller than the nominal level of $\alpha = 0.05$, reaching a maximum of only 0.009 (all values can be found in Table S4 of the Supporting Information). Hence, the analysis focuses on the power of the tests which is shown in Figure 4. The loss of power resulting from model misspecification is not as large as in Section 4.2.2 but still present. Especially if one of the models is falsely specified to be an emax model but the other one specified correctly, we observe a notable loss of power not only compared to using the true model but also compared to using model averaging. Moreover, this effect is increasing with increasing sample size. In contrast to Section 4.2.2, the power resulting from using model averaging with smooth BIC weights is notably smaller than the one observed when using the true model. However, compared to two out of the three misspecifications, the loss in power is extensively reduced. Similar to Section 4.2.2, using model averaging with equal weights leads to a high power even exceeding the power which results from using the true models. As in Section 4.2.2, this might be caused by forcing the data into a (partly) wrong model rather than by the model fitting the data well. In conclusion, if the models are misspecified we observe either type I error inflation or a lack of power, both often of substantial extend, in all three scenarios. Using model averaging with smooth BIC weights considerably reduces these problems, often leading to similar results as knowing and using the true underlying model.

5 | Case Study

We illustrate the proposed methodology through a case study analyzing the equivalence of time-response curves (also known as exposure duration-response curves) using data which was originally published by Ghallab et al. [47] The study aims to investigate dietary effects onto the gene expression. The dataset consists of two groups of mice which were fed with two different diets and then sacrificed at different time points. The first one is a high-fat or "Western" diet (WD) while the other one is a standard diet (SD). As no data has been collected in the first 3 weeks, they are not included into our analysis. Consequentially, the beginning of the study ($t = 0$) resembles week 3 of the actual experiment. Data is then observed at $t = 0, 3, 9, 15, 21, 27, 33, 39$, and 45 for the Western diet and at $t = 0, 3, 27, 33, 39$, and 45 for the standard diet with sample sizes 5, 5, 5, 5, 5, 5, 5, 4, 8 and 7, 5, 5, 7, 3, 5, respectively. For each group, the gene expression of 20 733 genes is measured in terms of gene counts. For our analysis, we focus on the 1000 genes Ghallab et al. [47] classified as especially interesting due to high activity. Although gene expression is measured as count data, it is treated as continuous due to the very high number of counts. The raw count data is preprocessed in terms of the gene count normalization conducted by Ghallab et al. [47] and subsequent $\log_2$-transformation of the normalized counts as suggested by Duda et al. [42, 48]

Using this data, we aim to investigate the equivalence of the time-gene expressions curves between the two diets at a 5% significance level. From Ghallab et al. [47] it is known that there are quite large differences between the diets, such that for the majority of the genes we expect not to conclude equivalence. However, precisely for this reason it is of interest for which genes equivalence can be concluded nevertheless. As we are interested in the results for each gene separately and are not aiming for a global conclusion, we do not adjust for multiple testing.

As time-response studies are relatively rare, no specific time-response models have been developed. Hence, dose–response models are deployed for time-response relations as well. Methodological review studies [49] do also not distinguish between dose–response and time-response studies. In addition, it seems intuitive that the effects of the high-fat diet accumulate with increasing time of consumption in a similar manner as the effects in dose response-studies accumulate with increasing dose.

As outlined by Ghallab et al. [47] the dose–response relations vary across genes such that there is no single model which fits to all of them and, hence, model uncertainty is present. In addition, the models cannot be chosen manually due to the high number of genes. We introduce model averaging using BIC-based weights and the equivalence tests are performed using hybrid CI, that is, by conducting Algorithm 1b. We deploy the set of candidate models suggested by Duda et al. [42], that is, a linear model (1), a quadratic model (2), an emax model (3), an exponential model (4), a sigmoid emax model (5), and a beta model (6). This set of candidate models can capture quite diverse effects, as it includes linear and nonlinear, increasing and decreasing, monotone and non-monotone as well as convex, concave and sigmoid curves.

The ranges of the response variables, that is, the ranges of $\log_2$ (normalized counts), are not comparable across different genes. Hence, different equivalence thresholds are needed for each of the genes. As such thresholds can not be chosen manually due to the high number of genes, we determine the thresholds as a percentile of the range of the response variable. For a gene $g \in \{1, \dots, 1000\}$, that is

$$\epsilon_g = \tilde{\epsilon} \left(\max_{l,i} \left(\hat{y}_{gli} \right) - \min_{l,i} \left(\hat{y}_{gli} \right) \right)$$

where $\tilde{\epsilon} \in (0, 1)$ is the corresponding percentile and $\tilde{\epsilon} = 0.2$ or 0.25 would be typical choices. Alternatively, one can proceed the other way around, calculate

$$\tilde{u}_g = \frac{u_g}{\max_{l,i} \left(\hat{y}_{gli} \right) - \min_{l,i} \left(\hat{y}_{gli} \right)}$$

and directly compare $\tilde{u}_g$ to $\tilde{\epsilon}$, that is, the decision rules $\epsilon_g > u_g$ and $\tilde{\epsilon} > \tilde{u}_g$ are equivalent.

Figure 5a,b show boxplots of the model weights for both diets. It can be observed that less complex models, the linear and quadratic model, have higher weights for the standard diet compared to the Western diet. In contrast, for the two most complex models, the beta and sigEmax model, the opposite can be observed: they have higher weights for the Western diet compared to the standard diet.

In addition, Figure 5c,d show histograms of the highest model weight per gene. It can be observed that for the Western diet the model weights tend to be larger compared to the standard diet. In addition, for the Western diet there are notably many genes for which the highest model weight is very close to one, that is, the averaged model consists nearly fully of only one of the candidate models.

Figure 5e shows the number of genes for which equivalence of the time-gene expression curves of both diets can be concluded, that is, the number of genes for which H_0 can be rejected depending on the choice of $\tilde{\epsilon}$. For very small choices of $\tilde{\epsilon}$ (e.g., 0.05, 0.06,

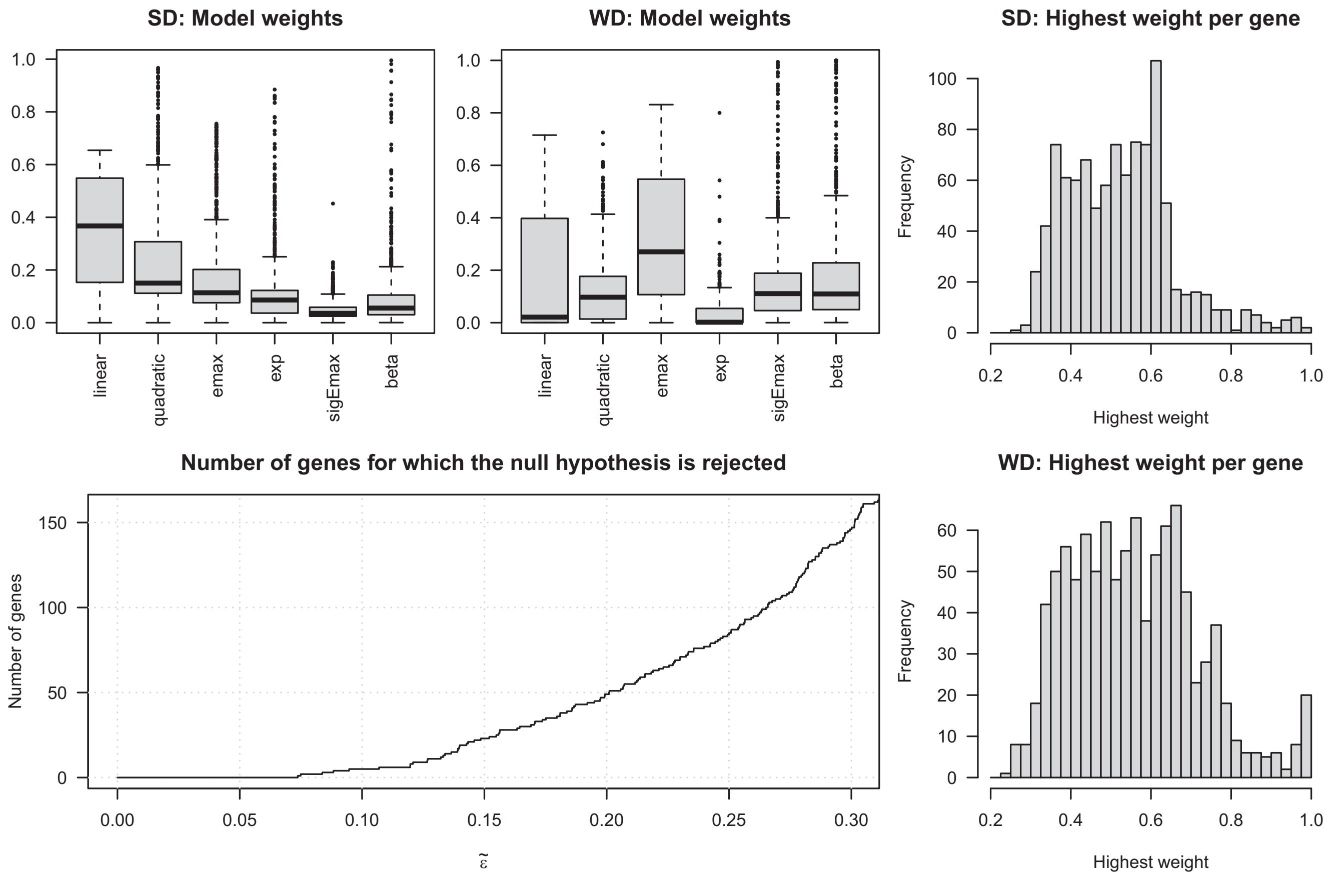


FIGURE 5 | Subfigures (a) and (b) show boxplots of the model weights for each of the two diets. Subfigures (c) and (d) show histograms of the highest model weight per gene for both genes. Subfigure (e) shows the number of genes for which equivalence between the time-gene expression curves of the two diets can be concluded in dependence of the equivalence threshold $\tilde{\varepsilon}$.

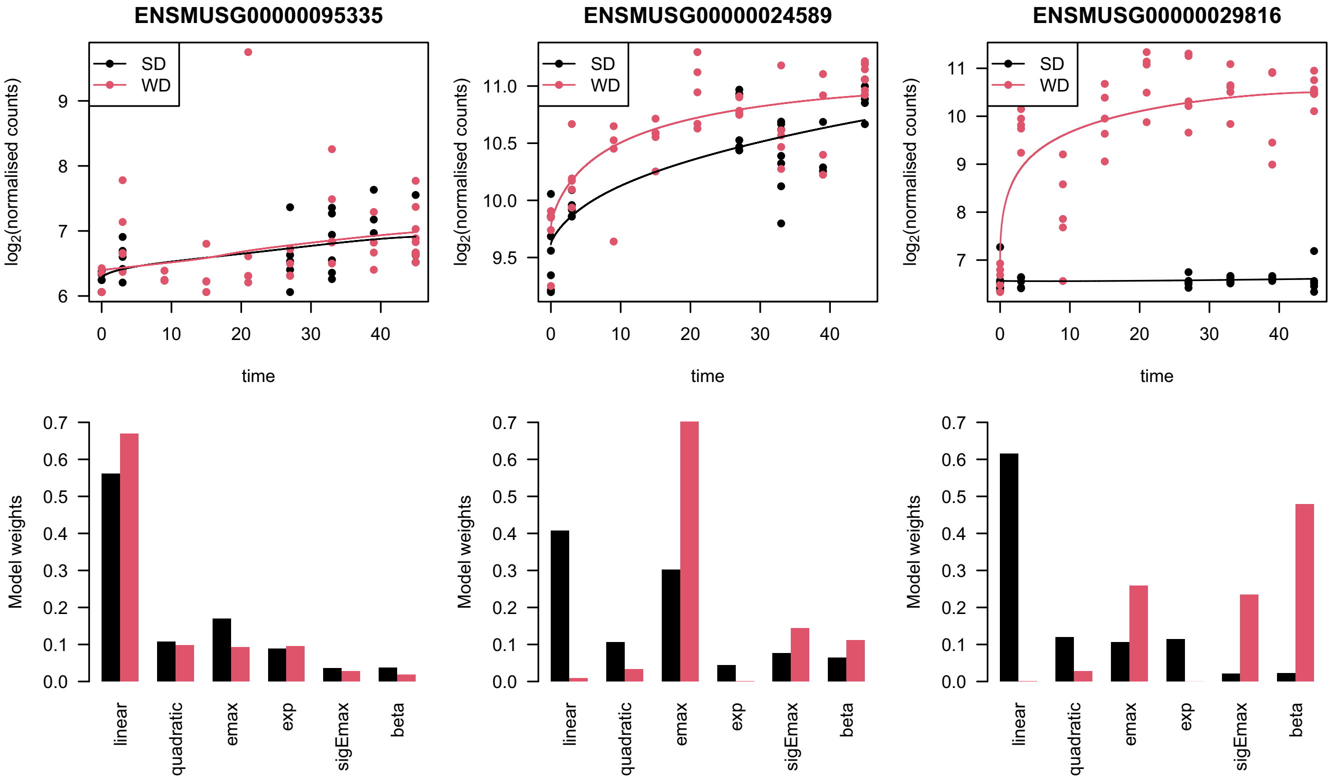


FIGURE 6 | Results for three exemplary genes. The first row of figures shows the data for both diets as well as the fitted models. The second row of figures shows the corresponding model weights.

or 0.07) equivalence cannot be concluded for any gene and for $\tilde{\epsilon} = 0.1$ only five genes would be assessed as equivalent. For more typical choices of $\tilde{\epsilon}$ being 0.2, 0.25, or 0.3, equivalence could be concluded for 50, 85, and 147 genes, respectively. With further increasing $\tilde{\epsilon}$ the number of rejections also further increases and approaches 1000. However, this is not shown for $\tilde{\epsilon} > 0.3$ as performing an equivalence test with a threshold larger than 30% of the range of the response variable might not have practical relevance.

Figure 6 shows the results for three exemplary genes. For ENSMUSG00000095335 it can be observed that both time-response curves are extremely close to each other and that the maximum absolute deviation of the curves is quite small. This leads to $\tilde{u} \approx 0.084$. Regarding the model weights it can be observed that both time-response curves consist essentially of the same models. For ENSMUSG00000024589 we observe that both time-response curves have a similar shape both being emax-like, although their model weights are not as similar as before. However, their distance is larger than for ENSMUSG00000095335 which leads to $\tilde{u} \approx 0.303$. Hence, the curves are not equivalent for typical choices of $\tilde{\epsilon}$ being, for example, 0.2 or 0.3 but only for extremely liberal choices of $\tilde{\epsilon}$, for example, for $\tilde{\epsilon} = 0.35$. For the last example ENSMUSG00000029816 we observe that the two curves are completely different with regard to both, shape and location. For the standard diet an almost constant curve is present while for the Western diet a typical emax shape is observable. This is also reflected by the model weights where models which have high weights for one curve, have small ones for the other one and vise versa, the only exempt to this is the emax model which has a medium large weight for both of the groups. Due to the large maximum absolute deviation between the curves given by $\hat{d} \approx 3.891$, similarity cannot be concluded for any reasonable equivalence threshold.

6 | Conclusion

In this paper, we introduced a new approach for model-based equivalence testing which can also be applied in the presence of model uncertainty — a problem which is usually faced in practical applications. Our approach is based on a flexible model averaging method which relies on information criteria and a testing procedure which makes use of the duality of tests and confidence intervals rather than simulating the distribution under the null hypothesis, providing a numerically stable procedure. Due to the advantages of theoretical validity based on asymptotic theory, we chose to use the BIC as the information criterion. Moreover, our approach leads to additional interpretability due to the provided confidence intervals while retaining the asymptotic validity and a similar performance in finite samples as the bootstrap based test proposed by Dette et al. [10].

Precisely, we investigated the finite sample properties of the proposed method by reanalysing the simulation study of Dette et al. [10] and observed similar results for the CI-based test compared to their test. In the presence of model uncertainty, model misspecification frequently led to either type I error inflation or a lack of power, both often of substantial extend. In contrast, our approach considerably reduced these problems and in many cases even achieved similar results as knowing and using the true underlying model. In direct comparison, a simpler model averaging method, that is, using fixed equal weights, was not able to prevent high type I error inflation. Therefore, we strongly recommend to use information criteria-based model averaging. The presented case study outlines the practical usefulness of the proposed method based on a large data application where choosing the models manually would be time-consuming and could easily lead to many model misspecifications. Hence, introducing model averaging here is essential to test for the equivalence of time-gene expression curves for such large numbers of genes, typically occurring in practice.

Future possible research includes extending the presented method for other model averaging techniques, for example, cross validation-based model averaging. In addition, transferring this approach to other model classes (e.g., survival models) as well as to multidimensional responses, that is, multiple endpoints, merits further exploration.

Acknowledgments

The authors would like to thank two referees and the Associate Editor for their useful suggestions which greatly improved the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Software in the form of R code available at https://github.com/Niklas191/equivalence_tests_with_model_averaging.git. The case study data set is publicly available at the SRA database with reference number PRJNA953810.

References

1. C. Otto, I. Fuchs, H. Altmann, et al., "Comparative Analysis of the Uterine and Mammary Gland Effects of Drospirenone and Medroxyprogesterone Acetate," *Endocrinology* 149, no. 8 (2008): 3952–3959.

2. S. S. Jhee, W. H. Lyness, P. B. Rojas, M. T. Leibowitz, V. Zarotsky, and L. V. Jacobsen, "Similarity of Insulin Detemir Pharmacokinetics, Safety, and Tolerability Profiles in Healthy Caucasian and Japanese American Subjects," *Journal of Clinical Pharmacology* 44, no. 3 (2004): 258–264.

3. K. Möllenhoff, F. Loingeville, J. Bertrand, et al., "Efficient Model-Based Bioequivalence Testing," *Biostatistics* 23, no. 1 (2022): 314–327.

4. D. Hauschke, V. Steinijans, and I. Pigeot, *Bioequivalence Studies in Drug Development* (West Sussex, UK: John Wiley & Sons, Ltd, 2007).

5. D. J. Schuirmann, "A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability," *Journal of Pharmacokinetics and Biopharmaceutics* 15 (1987): 657–680.

6. D. Lakens, "Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses," *Social Psychological and Personality Science* 8, no. 4 (2017): 355–362.

7. W. Liu, F. Bretz, A. J. Hayter, and H. P. Wynn, "Assessing Nonsuperiority, Noninferiority, or Equivalence When Comparing Two Regression Models Over a Restricted Covariate Region," *Biometrics* 65, no. 4 (2009): 1279–1287.

8. S. Gsteiger, F. Bretz, and W. Liu, "Simultaneous Confidence Bands for Nonlinear Regression Models With Application to Population

Pharmacokinetic Analyses," *Journal of Biopharmaceutical Statistics* 21, no. 4 (2011): 708–725.

9. F. Bretz, K. Möllenhoff, H. Dette, W. Liu, and M. Trampisch, "Assessing the Similarity of Dose Response and Target Doses in Two Non-Overlapping Subgroups," *Statistics in Medicine* 37, no. 5 (2018): 722–738.

10. H. Dette, K. Möllenhoff, S. Volgushev, and F. Bretz, "Equivalence of Regression Curves," *Journal of the American Statistical Association* 113, no. 522 (2018): 711–729.

11. K. Möllenhoff, F. Bretz, and H. Dette, "Equivalence of Regression Curves Sharing Common Parameters," *Biometrics* 76, no. 2 (2020): 518–529.

12. K. Möllenhoff, H. Dette, and F. Bretz, "Testing for Similarity of Binary Efficacy-Toxicity Responses," *Biostatistics* 23, no. 3 (2021): 949–966.

13. K. Möllenhoff, N. Binder, and H. Dette, "Testing Similarity of Parametric Competing Risks Models for Identifying Potentially Similar Pathways in Healthcare," *Statistics in Medicine* 43, no. 28 (2024): 5316–5330.

14. N. Hagemann, G. Marra, F. Bretz, and K. Möllenhoff, "Testing for Similarity of Multivariate Mixed Outcomes Using Generalized Joint Regression Models With Application to Efficacy-Toxicity Responses," *Biometrics* 80, no. 3 (2024): 1–11.

15. M. Guhl, F. Mercier, C. Hofmann, et al., "Impact of Model Misspecification on Model-Based Tests in PK Studies With Parallel Design: Real Case and Simulation Studies," *Journal of Pharmacokinetics and Pharmacodynamics* 49, no. 5 (2022): 557–577.

16. B. Dennis, J. M. Ponciano, M. L. Taper, and S. R. Lele, "Errors in Statistical Inference Under Model Misspecification: Evidence, Hypothesis Testing, and AIC," *Frontiers in Ecology and Evolution* 7 (2019): 372.

17. K. Möllenhoff, H. Dette, E. Kotzagiorgis, S. Volgushev, and O. Collignon, "Regulatory Assessment of Drug Dissolution Profiles Comparability via Maximum Deviation," *Statistics in Medicine* 37, no. 20 (2018): 2968–2981.

18. B. Bornkamp, "Viewpoint: Model Selection Uncertainty, Pre-Specification, and Model Averaging," *Pharmaceutical Statistics* 14, no. 2 (2015): 79–81.

19. L. Breiman, "Heuristics of Instability and Stabilization in Model Selection," *Annals of Statistics* 24 (1996): 2350–2383.

20. H. Leeb and B. M. Pötscher, "Model Selection and Inference: Facts and Fiction," *Econometric Theory* 21, no. 1 (2005): 21–59.

21. H. Leeb and B. M. Pötscher, "Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators?," *Econometric Theory* 24, no. 2 (2008): 338–376.

22. K. Schorning, B. Bornkamp, F. Bretz, and H. Dette, "Model Selection Versus Model Averaging in Dose Finding Studies," *Statistics in Medicine* 35, no. 22 (2016): 4021–4040.

23. Y. Aoki, D. Röshammar, B. Hamrén, and A. C. Hooker, "Model Selection and Averaging of Nonlinear Mixed-Effect Models for Robust Phase III Dose Selection," *Journal of Pharmacokinetics and Pharmacodynamics* 44 (2017): 581–597.

24. S. Buatois, S. Ueckert, N. Frey, S. Retout, and F. Mentré, "Comparison of Model Averaging and Model Selection in Dose Finding Trials Analyzed by Nonlinear Mixed Effect Models," *AAPS Journal* 20 (2018): 1–9.

25. B. Bornkamp, J. Pinheiro, and F. Bretz, "MCPMod: An R Package for the Design and Analysis of Dose-Finding Studies," *Journal of Statistical Software* 29, no. 7 (2009): 1–23.

26. S. T. Buckland, K. P. Burnham, and N. H. Augustin, "Model Selection: An Integral Part of Inference," *Biometrics* 53, no. 2 (1997): 603–618.

27. H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control* 19, no. 6 (1974): 716–723.

28. E. Ley and M. F. Steel, "On the Effect of Prior Assumptions in Bayesian Model Averaging With Applications to Growth Regression," *Journal of Applied Econometrics* 24, no. 4 (2009): 651–674.

29. G. Schwarz, "Estimating the Dimension of a Model," *Annals of Statistics* 6, no. 2 (1978): 461–464.

30. L. Wasserman, "Bayesian Model Selection and Model Averaging," *Journal of Mathematical Psychology* 44, no. 1 (2000): 92–107.

31. M. J. Price, N. J. Welton, A. H. Briggs, and A. Ades, "Model Averaging in the Presence of Structural Uncertainty About Treatment Effects: Influence on Treatment Decision and Expected Value of Information," *Value in Health* 14, no. 2 (2011): 205–218.

32. D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64, no. 4 (2002): 583–639.

33. N. L. Hjort and G. Claeskens, "Frequentist Model Average Estimators," *Journal of the American Statistical Association* 98, no. 464 (2003): 879–899.

34. G. Claeskens and N. L. Hjort, "The Focused Information Criterion," *Journal of the American Statistical Association* 98, no. 464 (2003): 900–916.

35. B. E. Hansen and J. S. Racine, "Jackknife Model Averaging," *Journal of Econometrics* 167, no. 1 (2012): 38–46.

36. T. M. Le and B. S. Clarke, "Model Averaging Is Asymptotically Better Than Model Selection for Prediction," *Journal of Machine Learning Research* 23, no. 33 (2022): 1–53.

37. D. Fletcher, *Model Averaging (Springer Briefs in Statistics)* (Berlin, Germany: Springer, 2018).

38. G. Claeskens and N. L. Hjort, *Model Selection and Model Averaging*. Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge, UK: Cambridge University Press, 2008).

39. F. Bretz, J. C. Pinheiro, and M. Branson, "Combining Multiple Comparisons and Modeling Techniques in Dose-Response Studies," *Biometrics* 61, no. 3 (2005): 738–748.

40. J. C. Pinheiro, F. Bretz, and M. Branson, "Analysis of Dose–Response Studies–Modeling Approaches," in *Dose Finding in Drug Development*, ed. N. Ting (New York: Springer, 2006).

41. J. Pinheiro, B. Bornkamp, E. Glimm, and F. Bretz, "Model-Based Dose Finding Under Model Uncertainty Using General Parametric Models," *Statistics in Medicine* 33, no. 10 (2014): 1646–1661.

42. J. Duda, F. Kappenberg, and J. Rahnenführer, "Model Selection Characteristics When Using MCP-Mod for Dose-Response Gene Expression Data," *Biometrical Journal* 64, no. 5 (2022): 883–897.

43. M. Wang, X. Zhang, A. Wan, and G. Zou, "On the Asymptotic Distribution of Model Averaging Based on Information Criterion." arXiv: 1910.12208 [stat.ME], 2019.

44. P. Bastian, H. Dette, L. Koletzko, and K. Möllenhoff, "Comparing Regression Curves—An L^1-Point of View," *Annals of the Institute of Statistical Mathematics* 76, no. 1 (2024): 159–183.

45. J. Aitchison, "Confidence-Region Tests," *Journal of the Royal Statistical Society: Series B (Methodological)* 26, no. 3 (1964): 462–476.

46. B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Monographs on Statistics & Applied Probability (New York: Chapman and Hall, 1994).

47. A. Ghallab, M. Myllys, A. Friebel, et al., "Spatio-Temporal Multiscale Analysis of Western Diet-Fed Mice Reveals a Translationally Relevant Sequence of Events During NAFLD Progression," *Cells* 10, no. 10 (2021): 1–29.

48. J. Duda, C. Drenda, H. Kästel, J. Rahnenführer, and F. Kappenberg, “Benefit of Using Interaction Effects for the Analysis of High-Dimensional Time-Response or Dose-Response Data for Two-Group Comparisons,” *Scientific Reports* 13 (2023): 20804.

49. F. Kappenberg, J. Duda, L. Schürmeyer, et al., “Guidance for Statistical Design and Analysis of Toxicological Dose-Response Experiments, Based on a Comprehensive Literature Review,” *Archives of Toxicology* 97, no. 10 (2023): 2741–2761.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Overcoming model uncertainty – how equivalence tests can benefit from model averaging

Supporting information

Niklas Hagemann[1] and Kathrin Möllenhoff[1]

[1] Institute of Medical Statistics and Computational Biology (IMSB),

Faculty of Medicine, University of Cologne, Germany

October 21, 2024

S 1: Comparison of the type I error rates of the CI-based testing approaches to the testing approaches proposed by Dette et al. (2018) with $\varepsilon = 1$. The results are shown for two distances of the regression curves $d \in \{0.5, 0.75\}$ and three different combinations of variances $(\sigma_1^2, \sigma_2^2) \in \{(0.25, 0.25), (0.5, 0.5), (0.25, 0.5)\}$.

						hybrid	percentile	asymptotic	Dette et al. (2018)	Dette et al. (2018)
β_{20}	d	σ_1^2	σ_2^2	n_1	n_2	CI	CI	CI	bootstrap	asymptotic
0.25	1.50	0.25	0.25	10	10	0.000	0.000	0.000	0.001	0.000
0.25	1.50	0.25	0.25	10	20	0.000	0.000	0.000	0.000	0.000
0.25	1.50	0.25	0.25	20	20	0.000	0.000	0.000	0.000	0.000
0.25	1.50	0.25	0.25	50	50	0.000	0.000	0.000	0.000	0.000
0.25	1.50	0.25	0.50	10	10	0.000	0.000	0.000	0.000	0.000
0.25	1.50	0.25	0.50	10	20	0.000	0.000	0.000	0.000	0.000
0.25	1.50	0.25	0.50	20	20	0.000	0.000	0.000	0.000	0.000
0.25	1.50	0.25	0.50	50	50	0.000	0.000	0.000	0.000	0.000
0.25	1.50	0.50	0.50	10	10	0.001	0.000	0.000	0.001	0.000
0.25	1.50	0.50	0.50	10	20	0.001	0.000	0.000	0.002	0.000
0.25	1.50	0.50	0.50	20	20	0.000	0.000	0.000	0.000	0.000
0.25	1.50	0.50	0.50	50	50	0.000	0.000	0.000	0.000	0.000
0.50	1.25	0.25	0.25	10	10	0.004	0.000	0.000	0.005	0.001
0.50	1.25	0.25	0.25	10	20	0.005	0.004	0.001	0.004	0.000
0.50	1.25	0.25	0.25	20	20	0.001	0.000	0.000	0.001	0.000
0.50	1.25	0.25	0.25	50	50	0.000	0.000	0.000	0.000	0.000
0.50	1.25	0.25	0.50	10	10	0.006	0.001	0.002	0.006	0.000
0.50	1.25	0.25	0.50	10	20	0.000	0.000	0.002	0.005	0.001
0.50	1.25	0.25	0.50	20	20	0.001	0.000	0.000	0.000	0.001
0.50	1.25	0.25	0.50	50	50	0.000	0.000	0.000	0.000	0.000
0.50	1.25	0.50	0.50	10	10	0.005	0.001	0.000	0.011	0.001
0.50	1.25	0.50	0.50	10	20	0.005	0.000	0.001	0.013	0.005
0.50	1.25	0.50	0.50	20	20	0.002	0.000	0.000	0.004	0.000
0.50	1.25	0.50	0.50	50	50	0.000	0.000	0.000	0.000	0.000
0.75	1.00	0.25	0.25	10	10	0.029	0.007	0.005	0.045	0.012
0.75	1.00	0.25	0.25	10	20	0.041	0.008	0.015	0.045	0.019
0.75	1.00	0.25	0.25	20	20	0.044	0.016	0.015	0.034	0.011
0.75	1.00	0.25	0.25	50	50	0.040	0.016	0.027	0.051	0.016
0.75	1.00	0.25	0.50	10	10	0.035	0.004	0.005	0.036	0.003
0.75	1.00	0.25	0.50	10	20	0.030	0.005	0.009	0.028	0.009
0.75	1.00	0.25	0.50	20	20	0.037	0.009	0.010	0.048	0.009
0.75	1.00	0.25	0.50	50	50	0.026	0.008	0.014	0.058	0.012
0.75	1.00	0.50	0.50	10	10	0.018	0.001	0.002	0.037	0.005
0.75	1.00	0.50	0.50	10	20	0.025	0.002	0.001	0.046	0.006
0.75	1.00	0.50	0.50	20	20	0.038	0.003	0.005	0.038	0.036
0.75	1.00	0.50	0.50	50	50	0.031	0.010	0.010	0.059	0.015

S 2: Comparison of the power of the test using the true model, the model averaging-based tests and the tests under model misspecification in scenario 1. The results are shown for $\varepsilon = 1$, three distances of the regression curves $d \in \{1, 1.25, 1.5\}$ and three different combinations of variances $(\sigma_1^2, \sigma_2^2) \in \{(0.25, 0.25), (0.25, 0.5), (0.5, 0.5)\}$.

β_{20}	d	σ_1^2	σ_2^2	n_1	n_2	true model	MA: BIC	MA: EW	emax & emax	exp & exp	exp & emax
1.0	0.75	0.25	0.25	10	10	0.152	0.173	0.296	0.190	0.293	0.133
1.0	0.75	0.25	0.25	10	20	0.177	0.213	0.385	0.239	0.383	0.173
1.0	0.75	0.25	0.25	20	20	0.252	0.283	0.539	0.359	0.553	0.246
1.0	0.75	0.25	0.25	50	50	0.462	0.508	0.937	0.693	0.932	0.659
1.0	0.75	0.25	0.50	10	10	0.095	0.107	0.204	0.112	0.193	0.096
1.0	0.75	0.25	0.50	10	20	0.156	0.168	0.305	0.190	0.284	0.135
1.0	0.75	0.25	0.50	20	20	0.175	0.215	0.386	0.260	0.379	0.152
1.0	0.75	0.25	0.50	50	50	0.341	0.408	0.800	0.593	0.784	0.462
1.0	0.75	0.50	0.50	10	10	0.058	0.071	0.143	0.078	0.129	0.071
1.0	0.75	0.50	0.50	10	20	0.085	0.099	0.204	0.105	0.204	0.107
1.0	0.75	0.50	0.50	20	20	0.119	0.143	0.283	0.157	0.290	0.127
1.0	0.75	0.50	0.50	50	50	0.283	0.338	0.662	0.445	0.689	0.336
1.5	0.50	0.25	0.25	10	10	0.254	0.316	0.575	0.363	0.589	0.576
1.5	0.50	0.25	0.25	10	20	0.306	0.387	0.658	0.435	0.700	0.727
1.5	0.50	0.25	0.25	20	20	0.431	0.506	0.850	0.597	0.889	0.917
1.5	0.50	0.25	0.25	50	50	0.824	0.847	0.995	0.930	0.997	0.999
1.5	0.50	0.25	0.50	10	10	0.135	0.177	0.371	0.214	0.345	0.323
1.5	0.50	0.25	0.50	10	20	0.216	0.293	0.530	0.340	0.536	0.544
1.5	0.50	0.25	0.50	20	20	0.312	0.375	0.699	0.462	0.724	0.725
1.5	0.50	0.25	0.50	50	50	0.652	0.690	0.974	0.839	0.974	0.994
1.5	0.50	0.50	0.50	10	10	0.061	0.097	0.226	0.108	0.214	0.183
1.5	0.50	0.50	0.50	10	20	0.119	0.177	0.327	0.201	0.349	0.332
1.5	0.50	0.50	0.50	20	20	0.203	0.289	0.536	0.326	0.571	0.551
1.5	0.50	0.50	0.50	50	50	0.525	0.552	0.912	0.698	0.930	0.961

S 3: Comparison of the type I error rates of the test using the true model, the model averaging-based tests and the tests under model misspecification in scenario 2. The results are shown for $\varepsilon = 0.5$, three distances of the regression curves $d \in \{1, 1.25, 1.5\}$ and three different combinations of variances $(\sigma_1^2, \sigma_2^2) \in \{(0.25, 0.25), (0.25, 0.5), (0.5, 0.5)\}$.

$\beta_{10} = d$	σ_1^2	σ_2^2	n_1	n_2	true model	MA: BIC	MA: EW	emax & exp	exp & exp	exp & emax
0.50	0.25	0.25	10	10	0.001	0.000	0.000	0.000	0.000	0.000
0.50	0.25	0.25	10	20	0.001	0.001	0.000	0.000	0.000	0.000
0.50	0.25	0.25	20	20	0.003	0.003	0.005	0.000	0.000	0.000
0.50	0.25	0.25	50	50	0.001	0.001	0.002	0.000	0.000	0.000
0.50	0.25	0.50	10	10	0.001	0.000	0.000	0.000	0.000	0.000
0.50	0.25	0.50	10	20	0.001	0.000	0.005	0.000	0.000	0.000
0.50	0.25	0.50	20	20	0.001	0.001	0.003	0.000	0.000	0.000
0.50	0.25	0.50	50	50	0.001	0.001	0.001	0.000	0.000	0.000
0.50	0.50	0.50	10	10	0.000	0.000	0.000	0.000	0.000	0.000
0.50	0.50	0.50	10	20	0.000	0.000	0.000	0.000	0.000	0.000
0.50	0.50	0.50	20	20	0.000	0.000	0.002	0.000	0.000	0.000
0.50	0.50	0.50	50	50	0.000	0.000	0.002	0.000	0.000	0.000
0.75	0.25	0.25	10	10	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.25	0.25	10	20	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.25	0.25	20	20	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.25	0.25	50	50	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.25	0.50	10	10	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.25	0.50	10	20	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.25	0.50	20	20	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.25	0.50	50	50	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.50	0.50	10	10	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.50	0.50	10	20	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.50	0.50	20	20	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.50	0.50	50	50	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.25	0.25	10	10	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.25	0.25	10	20	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.25	0.25	20	20	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.25	0.25	50	50	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.25	0.50	10	10	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.25	0.50	10	20	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.25	0.50	20	20	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.25	0.50	50	50	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.50	0.50	10	10	0.000	0.000	0.001	0.000	0.000	0.000
1.00	0.50	0.50	10	20	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.50	0.50	20	20	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.50	0.50	50	50	0.000	0.000	0.000	0.000	0.000	0.000

S 4: Comparison of the type I error rates of the test using the true model, the model averaging-based tests and the tests under model misspecification in scenario 3. The results are shown for $\varepsilon = 0.5$, three distances of the regression curves $d \in \{1, 1.25, 1.5\}$ and three different combinations of variances $(\sigma_1^2, \sigma_2^2) \in \{(0.25, 0.25), (0.25, 0.5), (0.5, 0.5)\}$.

$\beta_{10} = d$	σ_1^2	σ_2^2	n_1	n_2	true model	MA: BIC	MA: EW	emax & exp	emax & emax	exp & emax
0.50	0.25	0.25	10	10	0.005	0.002	0.007	0.000	0.000	0.000
0.50	0.25	0.25	10	20	0.005	0.002	0.006	0.001	0.002	0.000
0.50	0.25	0.25	20	20	0.005	0.006	0.010	0.002	0.006	0.000
0.50	0.25	0.25	50	50	0.001	0.001	0.001	0.000	0.002	0.000
0.50	0.25	0.50	10	10	0.002	0.000	0.003	0.002	0.001	0.000
0.50	0.25	0.50	10	20	0.009	0.003	0.011	0.001	0.003	0.000
0.50	0.25	0.50	20	20	0.002	0.001	0.006	0.000	0.003	0.000
0.50	0.25	0.50	50	50	0.000	0.000	0.001	0.001	0.001	0.000
0.50	0.50	0.50	10	10	0.002	0.001	0.002	0.000	0.001	0.001
0.50	0.50	0.50	10	20	0.001	0.001	0.002	0.000	0.000	0.000
0.50	0.50	0.50	20	20	0.003	0.002	0.004	0.000	0.001	0.000
0.50	0.50	0.50	50	50	0.003	0.001	0.004	0.002	0.003	0.000
0.75	0.25	0.25	10	10	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.25	0.25	10	20	0.000	0.000	0.001	0.000	0.001	0.000
0.75	0.25	0.25	20	20	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.25	0.25	50	50	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.25	0.50	10	10	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.25	0.50	10	20	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.25	0.50	20	20	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.25	0.50	50	50	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.50	0.50	10	10	0.000	0.000	0.000	0.000	0.000	0.000
0.75	0.50	0.50	10	20	0.001	0.000	0.001	0.000	0.000	0.000
0.75	0.50	0.50	20	20	0.001	0.000	0.001	0.000	0.000	0.000
0.75	0.50	0.50	50	50	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.25	0.25	10	10	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.25	0.25	10	20	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.25	0.25	20	20	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.25	0.25	50	50	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.25	0.50	10	10	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.25	0.50	10	20	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.25	0.50	20	20	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.25	0.50	50	50	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.50	0.50	10	10	0.000	0.000	0.001	0.000	0.001	0.000
1.00	0.50	0.50	10	20	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.50	0.50	20	20	0.000	0.000	0.000	0.000	0.000	0.000
1.00	0.50	0.50	50	50	0.000	0.000	0.000	0.000	0.000	0.000

3.3 Dynamic Heterogeneity in Discrete Choice Experiments

This article discusses the third direction of multidimensionality, i.e. multidimensional covariate effects. Here, heterogeneously time-varying covariable effects are addressed by developing functional random coefficients. Such effects occur when a covariate influences the response variable not only in a heterogeneous and time-varying manner but its time-variation is also heterogeneous. This article can also be seen as preliminary work for Section 3.4, which is based on the methodologies developed in this section. In addition, this article introduces Bayesian estimation of such effects as an alternative to the frequentist inference which will be used in Section 3.4. The proposed model is applied to conditional logit models, a class of models commonly used in marketing research, but it directly generalizes to other GAMs. A case study, also from the field of marketing research, outlines the demand for such models in applied research. The flexibility of the approach as well as its superiority with regard to benchmark models is demonstrated through a simulation study.

Authorship:	First author
Coauthors:	Daniel Guhl, Thomas Kneib, Kathrin Möllenhoff and Winfried Steiner
Contribution statement:	My contribution to this project was developing the method, programming it in R, implementing the simulation study, analyzing the case study data, preparing the first draft of the figures and tables, and writing the first draft of the manuscript. The initial idea for this work came from Thomas Kneib, who also supported the development of the method. The figures were revised by Daniel Guhl who also contributed to writing the empirical application section. Discussions and revising were done together with all coauthors.
Status:	Under review, preprint available.
DOI:	10.2139/ssrn.4957076
Note according to § 7 of the Promotionsordnung:	Major parts of this article were developed as part of my Master's thesis. During the time of the doctoral studies, these results were revised, complemented by a simulation study and condensed into an article.

Dynamic Heterogeneity in Discrete Choice Experiments

Niklas Hagemann*
Institute of Medical Statistics and Computational Biology (IMSB),
Faculty of Medicine, University of Cologne, Germany

Daniel Guhl
Institute of Marketing, School of Business and Economics,
Humboldt University of Berlin, Germany

Thomas Kneib
Chair of Statistics and Campus Institute Data Science,
Georg-August-Universität Göttingen, Germany

Kathrin Möllenhoff
Institute of Medical Statistics and Computational Biology (IMSB),
Faculty of Medicine, University of Cologne, Germany

Winfried J. Steiner
Department of Marketing, Institute of Management and Economics,
Clausthal University of Technology, Germany

October 23, 2024

Abstract

In choice-based conjoint experiments, a special type of choice experiment, respondents' choice decisions are studied based on repeated tasks in an experimental setting. Especially when conducting longer sequences of choice tasks to increase the overall amount of information, effects such as learning or fatigue may come into play that affect the identification of the effects of choice task attributes. These effects may be exacerbated by customer-specific heterogeneity. We introduce Bayesian multinomial logit models with heterogeneously time-varying coefficients constructed as tensor products of random effects and penalized splines to capture both time variation (nonlinear dynamics) and customer-specific heterogeneity (cross-sectional heterogeneity). In an empirical application on a local public good, we use the developed method and find evidence for the presence of heterogeneously time-varying effects. The proposed approach further outperforms competing benchmark models that account for only cross-sectional heterogeneity or dynamics in fit and especially predictive accuracy, which is also demonstrated through a simulation study.

Keywords: Choice experiments; Functional random effects; Multinomial logit model; Penalized splines; Tensor products

[*]The authors thank Craig Broadbent (Brigham Young University, Rexburg, Idaho, USA) for the permission to use the data he originally collected for Broadbent et al. (2010).

1 Introduction

When analyzing repeated measurements on the same observational units in choice data, one challenge is disentangling different potential sources of heterogeneity. While it is nowadays standard to include random effects that reflect static, i.e., time-invariant cross-sectional heterogeneity of the observational units (e.g. Jain et al. 1994, Keane 1997, Elshiewy et al. 2017, for an overview of such models in the context of marketing applications), heterogeneity along the time dimension usually receives less attention. For longer series of repeated events, purely dynamic heterogeneity can, for example, be accounted for by including a vector autoregressive process (see, e.g. Kim et al. 2005) or a penalized spline estimate (see, e.g., Guhl et al. 2018) reflecting an overall time trend. Both approaches can also easily be combined in an additive model specification, but this crucially relies on the assumption of no interaction between cross-sectional heterogeneity of the observational units and dynamic heterogeneity. Finally, random intercepts, as well as temporal trends, are usually not interacted with covariate effects of interest.

In this paper, we focus on exactly this situation where covariate effects (as well as the overall intercept) in a regression model can be heterogeneously time-varying. More precisely, our research is motivated by choice-based conjoint (CBC) experiments (see, e.g. Rao 2014) in which each of the $i = 1, \dots, n$ respondents face $t = 1, \dots, T$ choice tasks. In each of these choice tasks, the respondent chooses one option out of a given set of alternatives $r = 1, \dots, c + 1$ that are characterized by different attributes (e.g., price, product attributes, etc. in case of a marketing-related experiment). An *outside option* (alternative $c + 1$) is typically included in each choice task to account for primary demand effects. CBC experiments are particularly popular in marketing research when analyzing preference structures for a product or service and the determinants of these preferences, but there are also other fields of application (e.g., transportation, psychology, health, and economics) as we will see later when analyzing preferences for a public good in this paper.

Choices in CBC are commonly modeled by a *multinomial logit (MNL) model* (McFadden 1973) with predictors $\eta_{itr} = \boldsymbol{x}'_{itr}\boldsymbol{\beta}$, $r = 1, \ldots, c$ where $\boldsymbol{x}_{itr}$ comprises information on the alternative-specific attributes while the regression coefficients $\boldsymbol{\beta}$ are constant across the alternatives. The latter is the default setting for unlabeled alternatives with generic attributes, implying that the position of alternatives within a given choice task is arbitrary and differences in positions have no meaning. Only the outside option as the reference category $c + 1$ has a constant position, serving as a baseline for the analysis (Rao 2014).

Just as in other empirical studies, the accuracy of the results crucially depends on the overall amount of information represented by the total number of observations $n \cdot T$, and a cost-efficient way to increase the sample size is to increase the number of choice tasks T per respondent. This also facilitates modeling heterogeneity across respondents using random effect specifications or hierarchical models. Indeed, as mentioned above, accounting for cross-sectional heterogeneity is nowadays state-of-the-art in academic research or industry applications of CBC (Allenby & Ginter 1995, Baumgartner & Steiner 2007, Kamakura & Wedel 2004, Elshiewy et al. 2017), and literature on constructing choice experiments also includes heterogeneity in optimal choice designs (Kessels et al. 2009). However, increasing T (i.e., presenting more choice tasks to the respondent) can have the disadvantage that respondents get fatigued or bored (e.g. Day et al. 2012, Savage & Waldman 2008), which may bias the results. In addition, learning effects (e.g. Day et al. 2012, Hess et al. 2012) are usually present in conjoint experiments, especially for more complex tasks. Recently, Li et al. (2022) showed that asking “too many” questions can even decrease external validity.

The literature on preference dynamics typically assumes that only the means of the heterogeneous parameter distributions vary over time (e.g. Liechty et al. 2005, Kim et al. 2005, Lachaab et al. 2006, Guhl et al. 2018). While this seems reasonable in applications of discrete choice models for market data, where seasonal effects or marketing-related efforts of firms at the aggregate level (e.g., advertising or price changes) affect all customers, individual-level dynamics could also be relevant in choice experiments. In particular, the marketing literature argues that there is a lack of methods to account for preference dynamics in conjoint models and calls for more work on this topic (Netzer et al. 2008). Only a few papers have addressed this issue so far. E.g., DeSarbo et al. (2005) analyze preference evolution in "traditional" conjoint analysis using a Bayesian dynamic linear model (e.g. Frühwirth-Schnatter et al. 2004), where the dependent variable is a continuous measure of preference. Hence, the data contains more information than our discrete choice data and can be analyzed using a linear regression framework. Dew et al. (2020) introduced a model using Gaussian processes to account for individual-level dynamic heterogeneity. The authors apply their model to choice data from a consumer panel that spans six years. The average number of purchases per individual is much larger than the number of choice tasks T in typical CBC applications; therefore, it is unclear if this approach would work in an application with choice experiments. Including flexible utility functions in choice models to address non-linear effects of variables on utility using semi- and non-parametric approaches is well established (see, e.g. Abe 1999, Briesch et al. 2010, Kim et al. 2007). Baumgartner et al. (2018) and Guhl et al. (2018) introduce penalized splines for modeling time-varying (average) effects in choice models for panel data.

We combine both strands of the literature and introduce a Bayesian MNL model with heterogeneously time-varying coefficients that takes respondent (i.e., cross-sectional) heterogeneity into account by including respondent-specific random effects and also allows for (potentially non-linear) time-varying effects (representing, for example, fatigue, learning, or task-adaptation effects), as well as their interaction, such that dynamic effects may vary heterogeneously across respondents. In contrast to Dew et al. (2020), who use Gaussian processes for a similar aim, we rely on penalized splines (Eilers & Marx 1996, Lang & Brezger 2004) for the time-varying effects while the interaction of random and time-varying effects will be cast into the general framework of tensor product interactions (Kneib et al. 2019). Inference will be conducted in a Bayesian framework based on MCMC simulation techniques, and the penalized spline specification allows us to take advantage of sparse matrix structures in the involved precision matrices and work with a moderately large number of basis coefficients. In addition, overfitting of the splines is implicitly countered by imposing roughness penalty terms (hence P-splines), and the amount of smoothness of a spline is determined simultaneously with all other parameters estimates in the Bayesian estimation framework (Lang & Brezger 2004, Aschersleben & Steiner 2022). To the best of our knowledge, this is the first CBC study introducing a fully Bayesian approach to estimate individual-level dynamic heterogeneity using penalized splines.

The rest of this paper is structured as follows: Section 2 introduces the modeling theory, including the modeling problem in Section 2.1, the underlying theory in Section 2.2, the adaptation of the theory to the modeling problem in Sections 2.3 and 2.4 and fit and predictive performance measures for model comparison in Section 2.5. Section 3 outlines the Bayesian inference used for model estimation. Section 4 provides an empirical application example. In Section 5, a simulation study is conducted to better understand the results from our empirical application and to ensure that the superior performance of the proposed approach is not due to overfitting. Section 6 summarizes the previous chapters, highlights the main conclusions, and gives an outlook on topics for future research.

2 Dynamic Heterogeneity in MNL Models

2.1 Multinomial Choice Model

As a foundation for the remainder of this section, we first introduce the notation for the CBC MNL model in more detail. Let $i = 1, \dots, n$ index the respondents facing $t = 1, \dots, T$ choice tasks with $r = 1, \dots, c{+}1$ alternatives, including the outside option $c{+}1$ as reference category. Alternatives are characterized by attributes $\boldsymbol{x}_{itr}$ for alternative r presented to customer i in choice task t while the actual choice made by the respondent is denoted as $y_{it} \in \{1, \dots, c+1\}$. Assuming i.i.d. Extreme Value (EV) type I distributed error terms and utility maximizing respondents, the resulting MNL model with predictors η_{itr} is

$$\mathbb{P}(y_{it} = r) = \frac{\exp(\eta_{itr})}{\sum_s \exp(\eta_{its})}. \tag{1}$$

In their simplest parametric form with homogeneous and time-constant effects, we have

$$\eta_{itr} = \boldsymbol{x}'_{itr}\boldsymbol{\beta}, \quad r = 1, \dots, c \tag{2}$$

and $\eta_{it,c+1} = 0$ to ensure identifiability. To account for time variation and respondent heterogeneity, we are interested in a model where

$$\eta_{itr} = \boldsymbol{x}'_{itr}\boldsymbol{\beta}_i(t) \tag{3}$$

i.e., the covariate effects are heterogeneous both concerning the respondent and over time. As special cases, this model also comprises time-invariant (static) but respondent-specific heterogeneity for $\boldsymbol{\beta}_i(t) \equiv \boldsymbol{\beta}_i$ and purely dynamic heterogeneity for $\boldsymbol{\beta}_i(t) \equiv \boldsymbol{\beta}(t)$.

2.2 Anisotropic Tensor Product Interactions

To represent respondent-specific time-varying parameters $\boldsymbol{\beta}_i(t)$, we rely on the framework of tensor product interactions as a general means of constructing interaction effects in generalized additive models (Kneib et al. 2019, Fahrmeir et al. 2013, Wood 2017). Consider a regression predictor $\eta = \ldots + x\beta(z_1, z_2) + \ldots$ where, in a varying-coefficient type fashion, the effect of covariate x varies according to two interaction variables z_1 and z_2. We can then interpret $\beta(z_1, z_2)$ as a bivariate surface and utilize tensor product basis function approaches for representing it. Let therefore $\beta_1(z_1)$ and $\beta_2(z_2)$ be "main effects" which are represented in terms of basis function expansions as

$$\beta_1(z_1) = \sum_{d_1=1}^{D_1} \gamma_{1d_1} B_{1d_1}(z_1), \quad \beta_2(z_2) = \sum_{d_2=1}^{D_2} \gamma_{2d_2} B_{2d_2}(z_2).$$

A tensor-product interaction is then obtained as

$$\beta(z_1, z_2) = \sum_{d_1=1}^{D_1} \sum_{d_2=1}^{D_2} \gamma_{d_1 d_2} B_{d_1 d_2}(z_1, z_2), \text{ where}$$

$$B_{d_1 d_2}(z_1, z_2) = B_{1d_1}(z_1) B_{2d_2}(z_2)$$

are the *tensor product basis functions* resulting from pairwise interactions of the main effect basis functions. While this approach is mostly used in the context of estimating bivariate interaction surfaces for two continuous covariates, often as a building block in smoothing spline analysis of variance-type models, each of the two main effects can be one of the model terms of structured additive regression models, including spatial and random effects (Kneib et al. 2019). We will use the tensor product interaction framework for interacting random effects and penalized splines to represent respondent-specific heterogeneity and time variation, respectively.

Since not only the main effects themselves but, in particular, the resulting tensor product interaction comprise a larger number of basis coefficients, some form of regularized estimation is necessary. We implement a Bayesian form of regularization by constructing informative priors for the tensor product. For this, we again start from the main effects, where (partially improper) multivariate Gaussian priors:

$$p(\boldsymbol{\gamma}_j | \tau_j^2) \propto \left(\frac{1}{\tau_j^2}\right)^{\frac{\mathrm{rk}(\boldsymbol{K}_j)}{2}} \exp\left(-\frac{1}{2\tau_j^2} \boldsymbol{\gamma}_j' \boldsymbol{K}_j \boldsymbol{\gamma}_j\right), \quad j = 1, 2$$

are applied to the vectors of basis coefficients $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jD_j})'$ with precision matrix $\boldsymbol{K}_j$ and variance parameter τ_j^2 that determine the exact form and strength of regularization, respectively. Specific choices for functional random effects will be discussed in the next section.

The resulting prior for the tensor product parameters $\boldsymbol{\gamma} = (\gamma_{11}, \ldots, \gamma_{1D_2}, \ldots, \gamma_{D_1 1}, \ldots, \gamma_{D_1 D_2})'$ is then of the same multivariate normal form but with precision matrix

$$\frac{1}{\tau_1^2}(\boldsymbol{K_1} \otimes \boldsymbol{I}_{D_2}) + \frac{1}{\tau_2^2}(\boldsymbol{K_2} \otimes \boldsymbol{I}_{D_1}) \tag{4}$$

and two variance parameters τ_j^2, $j = 1, 2$ that determine the regularization along the two axes defined by the main effect covariates. Effectively, this prior precision matrix implies that $\frac{1}{\tau_1^2}\boldsymbol{K}_1$ is applied in z_1-direction while $\frac{1}{\tau_2^2}\boldsymbol{K}_2$ is applied in z_2-direction. The ratio of τ_1^2 and τ_2^2 implies the relative importance of the two priors. Their overall magnitude relative to the error variance specifies the absolute impact of the prior on the estimated interaction effect. Including two separate variance parameters enables anisotropic forms of regularization, which is particularly important when interaction effects of very different nature, as in our application on functional random effects, are included.

2.3 Functional Random Effects

In contrast to common models with random intercepts and random slopes, in functional random effects models the whole nonlinear curves of continuous covariate effects are group-specific. The resulting *functional random effect* is then of the form $\beta_i(z)$, where $i \in \{1, \ldots, n\}$ denotes the grouping index while z is the continuous covariate of interest. In our application, we will focus on time as the continuous covariate of interest and will also consider functional random coefficients of the form $x\beta_i(t)$, where the effect of covariate x varies over time in a respondent-specific manner. In the following, we will show how functional random effects can be cast into the generic framework of tensor product interactions introduced in the previous section.

For heterogeneous, group- or, in our case, respondent-specific effects, we consider i.i.d. random effects for the first main effect. Here, the basis function representation is given by

$$\beta_i = \sum_{d_1=1}^{D_1} \gamma_{1d_1} B_{1d_1}(i) = \sum_{d_1=1}^{D_1} \gamma_{1d_1} \mathbb{1}(i = d_1) = \gamma_{1i}$$

i.e., the $D_1 = n$ basis functions are indicator functions for the group membership. For standard assumption of an i.i.d. Gaussian prior for the random effects, i.e. $\gamma_{1i} \overset{iid}{\sim} N(0, \tau_1^2)$, we set $\boldsymbol{K}_1 = \boldsymbol{I}_{D_1}$ resulting in a proper, multivariate normal distribution.

For time-varying effects $\beta(t)$, we rely on Bayesian forms of P-splines (Lang & Brezger 2004) where B-spline basis functions are employed in combination with first- or second-order *random walk* priors. The precision matrix is then of the form $\boldsymbol{K}_2 = \boldsymbol{D}'\boldsymbol{D}$ where $\boldsymbol{D}$ is a first or second order difference matrix of dimension D_2. The resulting prior can be considered as an approximation to the integrated squared first or second-order derivative penalty commonly assumed in smoothing spline approaches (Fahrmeir et al. 2013). Unlike with random effects, the resulting prior is partially improper since the precision matrix $\boldsymbol{K}_2$ is rank-deficient with the rank deficiency given by the order of the random walk. An in-depth introduction to (P-)splines is available in Eilers & Marx (2021).

2.4 Heterogeneously Time-Varying Coefficients

We now integrate the tensor product-based functional random coefficients framework with the MNL model (1). Let therefore $\beta_{il}(t)$ be one of the $l = 1, \ldots, L$ heterogeneously time-varying effects in (3) associated with covariate x_{itrl} for respondent i in choice task t and characterizing alternative r. We then specify a functional random coefficient for each $\beta_{il}(t)$. For interpretation, it is helpful to decompose the overall effect of x_{itrl} as

$$x_{itrl}\beta_l + x_{itrl}\beta_{il}(t),$$

i.e. to remove the overall population- and time-constant effect from the functional random effect, which then only comprises deviations from this overall effect. Using this decomposition, we can express our model with L covariates as

$$\begin{aligned} \eta_{itr} &= \beta_0 + \beta_{i0}(t) + (\beta_1 + \beta_{i1}(t))\, x_{itr1} + \ldots + (\beta_L + \beta_{iL}(t))\, x_{itrL} \\ &= \boldsymbol{x}'_{itr}\boldsymbol{\beta} + \boldsymbol{x}'_{itr}\boldsymbol{\beta}_i(t) \quad r = 1, \ldots, c \end{aligned} \tag{5}$$

and, as before, $\eta_{it,c+1} = 0$ for identification.

As competitors to our fully flexible functional random coefficient specification (5) (referred to as model M0), we will also consider the following benchmark models M1–M4:

- *Homogeneous model* (M1): A model without any form of heterogeneity in the effects.

- *Time-constant heterogeneity only* (M2): Removing the time-dependency in the functional random coefficients yields $x_{itrl}\beta_l + x_{itrl}\beta_{il}$ with i.i.d. respondent-specific random effects β_{il} reflecting time-constant heterogeneity of the respondents. This benchmark model in CBC analyses enables us to compare our results to a model that only captures cross-sectional heterogeneity while fully ignoring time effects.

- *Heterogeneously varying effects with linear time trends* (M3): In addition to respondent-specific heterogeneity, a heterogeneously varying linear time trend is added to the model, leading to $x_{itrl}\beta_l + x_{itrl}\beta_{0il} + x_{itrl}\beta_{1il}t$ where both β_{0il} and β_{1il} are i.i.d. random effects. In this model specification, dynamic effects are restricted to linear shapes. This model can capture cross-sectional heterogeneity and linear heterogeneous time variation, i.e., individual linear time trends. Thus, this benchmark model allows for investigation of the difference in model performance between our flexible semiparametric approach and the less complex but also less flexible linear approach.

- *Additive heterogeneity and homogeneous time variation* (M4): Rather than adding a heterogeneous yet linear time trend, this model specification features nonlinear yet homogeneous time variation, i.e. $x_{itrl}\beta_l + x_{itrl}\beta_{0il} + x_{itrl}\beta_l(t)$ with $\beta_l(t)$ being specified as a penalized spline. This model was developed by Guhl et al. (2018), and its time-varying coefficients are based on the approach of Biller & Fahrmeir (2001). It can be used to investigate whether or not modeling time-variation heterogeneously with our proposed approach can further improve the model performance compared to modeling the time-variation as well flexibly but globally (i.e., as the same across respondents).

While we discussed the specification of functional random coefficients for covariate effects, the same can, of course, also be applied to the intercept. If the reference category of an MNL model is chosen meaningfully, the corresponding intercept also has an interesting interpretation: The intercept parameter is added for all categories except for the reference category and, therefore, is the same as if it would be interacted with "not-the-reference-category" dummy $\mathbb{1}(r \neq c+1)$. This is a direct consequence of restricting $\eta_{it,c+1} = 0$.

In analogy, this also applies to the heterogeneously time-varying intercept. Therefore, higher values of the functional random intercept imply a higher probability of not choosing this outside option, given all covariates and their effects. This means that the time-varying intercept can capture general time-dependent tendencies unrelated to the covariates (e.g., learning or fatigue effects of respondents during an experiment).

2.5 Performance Measures

Several fit measures are used to evaluate the model performance following the arguments of Guhl et al. (2018) and Kneib et al. (2007). These are the Brier score as well as the spherical score, and the log-likelihood (log-Lik). Due to the different properties of these measures, there is no single best measure, hence it is reasonable to include multiple measures to assess the model performance reliably (Guhl et al. 2018). The measures are defined as

$$\text{log-Lik} = \sum_{i=1}^{n}\sum_{t=1}^{T} \log\left(\hat{\mathbb{P}}(y_{it} = r^*)\right),$$

$$\text{Brier score} = -\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{s=1}^{c+1}\left(\mathbb{1}(y_{it} = s) - \hat{\mathbb{P}}(y_{it} = s)\right)^2,$$

$$\text{Spherical score} = \sum_{i=1}^{n}\sum_{t=1}^{T}\frac{\hat{\mathbb{P}}(y_{it} = r^*)}{\sqrt{\sum_{s=1}^{c+1}\left(\hat{\mathbb{P}}(y_{it} = s)\right)^2}},$$

where r^* denotes the alternative that is chosen by person i at time t (Kneib et al. 2007).

In addition, the percentage of correct in-sample predictions (Correct-%)

$$\text{Correct-\%} = \frac{100\%}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbb{1}(y_{it} = \hat{y}_{it})$$

is used as an easy-to-understand fit measure.

In contrast to the other three fit measures, the percentage of correct predictions only considers the predicted choice and not the (un-)certainty with which this prediction is made. Therefore, it contains less information than the other ones but has a simple interpretation.

The out-of-sample predictive accuracy is estimated based on an information criterion. As explained by Vehtari et al. (2017) and Gelman et al. (2014) the *Watanabe-Akaike information criterion* (WAIC; Watanabe 2010) has advantages (e.g., averaging over the posterior distribution instead of using a point estimate) over other information criteria, especially the *Akaike information criterion* (AIC; Akaike 1974) and the *deviance information criterion* (DIC; Spiegelhalter et al. 2002). We use WAIC to estimate the out-of-sample predictive accuracy.

3 Bayesian Inference using Hamiltonian Monte Carlo

Recall the prior density

$$p(\boldsymbol{\gamma}|\tau_1^2, \tau_2^2) \propto \left(\frac{1}{\tau^2}\right)^{\frac{\mathrm{effdim}(\boldsymbol{\gamma})}{2}} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\gamma}'\boldsymbol{K}\boldsymbol{\gamma}\right) \mathbb{1}(\boldsymbol{A}\boldsymbol{\gamma} = \boldsymbol{0})$$

for the tensor product basis coefficients where

$$\begin{aligned}\frac{1}{\tau^2}\boldsymbol{K} &= \frac{1}{\tau_1^2}(\boldsymbol{K_1} \otimes \boldsymbol{I}_{D_2}) + \frac{1}{\tau_2^2}(\boldsymbol{K_2} \otimes \boldsymbol{I}_{D_1})\\ &= \frac{1}{\tau_1^2}(\boldsymbol{I}_{D_1} \otimes \boldsymbol{I}_{D_2}) + \frac{1}{\tau_2^2}\left((\boldsymbol{D}'\boldsymbol{D}) \otimes \boldsymbol{I}_{D_1}\right)\end{aligned}$$

and $\boldsymbol{A}$ implies constraints needed to make the model identifiable (Kneib et al. 2019). To avoid conditioning on an event with probability zero, this should not be seen as a condition in the probabilistic sense but as a projection. The effective dimension, effdim($\boldsymbol{\gamma}$) changes due to the constraint and fulfills $0 < \mathrm{effdim}(\boldsymbol{\gamma}) < \dim(\boldsymbol{\gamma})$ if at least one nontrivial constraint is applied (Kneib et al. 2019).

Since there is not much prior information about τ_1, τ_2, a weakly-informative prior distribution will be used. As shown by Gelman (2006) the (positive) half-normal distribution

$$\tau_1, \tau_2 \overset{iid}{\sim} \mathrm{HN}_+(\sigma_\tau)$$

is a weakly-informative prior if the variance $\sigma_\tau^2 \left(1 - \frac{2}{\pi}\right)$ of the half-normal is large enough.

For a clearer notation, we will relabel τ_1 as $\tau^{(i)}$ and τ_2 as $\tau^{(t)}$ not meaning that these variances vary along the corresponding index variable but just indicating to which direction they belong. Since all variables are modeled with time-varying coefficients, we will use a double index, e.g., $\tau_0^{(i)}$ for the standard deviation in i direction of the intercept.

Bayesian inference is then conducted by using the *No-U-Turn sampler* (NUTS, Hoffman & Gelman 2014), which is a variant of *Hamiltonian Monte Carlo* (Duane et al. 1987) also known as *hybrid Monte Carlo* or HMC. Using HMC, a single proposal is computationally more costly than using the Metropolis algorithm. However, since proposals are more efficient, the acceptance rate is much higher, and fewer samples are needed to describe the posterior distribution (McElreath 2020). The superiority of the effectiveness of HMC applies especially to highly complex models (Hoffman & Gelman 2014).

Estimation is implemented using the `R` package `brms` (Bürkner 2017). P-splines and tensor products are built using the `R` package `mgcv` (Wood 2021), and the Bayesian inference is based on the mixed model representation from the package `gamm4` (Wood & Scheipl 2020). `brms` uses `RStan` (Stan Development Team 2020) to perform statistical inference, and `RStan` is an `R` interface to the Stan programming language, using HMC and NUTS.

4 Empirical Application

4.1 Data and Research Question

In the following, the proposed flexible model with time-varying coefficients will be applied to the CBC setting described in Section 1. The choice experiment was originally conducted by Broadbent et al. (2010). In contrast to classical conjoint experiments in marketing, this study does not focus on consumer products or services but on a local public good. This local good are forest restoration activities of post-wildfire areas, which are performed by removing non-native highly flammable trees and planting native vegetation in exchange. Although the public good is local and all respondents were given an instructional period, this type of product/good is still relatively complex.

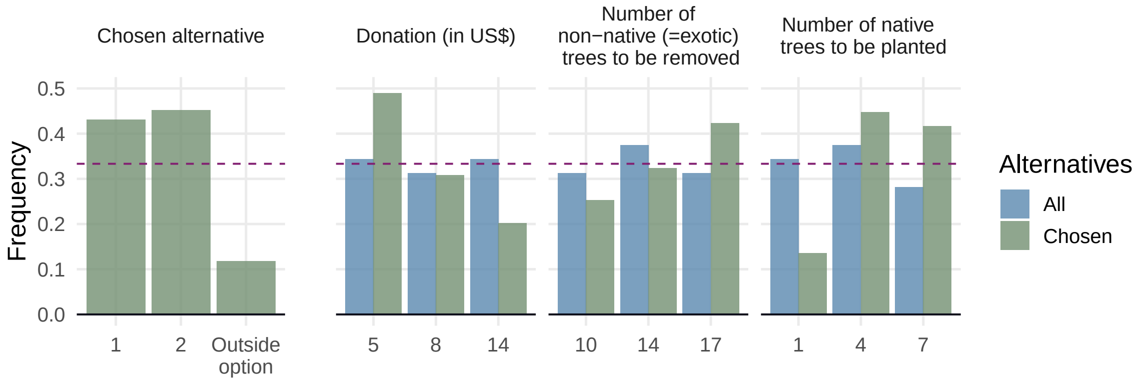
Each choice set contains three alternatives: Two real restoration alternatives and the *outside option*. There are three (metric) explanatory variables:

- *non-native*: the number of non-native trees to be removed (levels: 10, 14, and 17),
- *native*: the number of native trees to be planted (levels: 1, 4, and 7).
- *donation*: the voluntary donation for supporting the restoration activities (levels (in USD): \$5, \$8, and \$14),

where the donation variable can be interpreted as a price variable. Since the donation variable is not expressed in terms of the total amount needed for the restoration activity but as the average donation needed per person, the number of non-native and native trees is also measured per person. This ensures that the numbers are reasonable for a respondent.

All $n = 35$ respondents faced the same 20 choice sets in the same order, however four choice sets (choice sets 15, 17, 18 and 20) were only included in the original experiment in order to test for transitivity and stability of preferences. Like Broadbent et al. (2010), we exclude these four "control-questions" such that our dataset used for model estimation includes 16 choice sets with $T = 19$. Note that the flexibility of our proposed approach allows us to easily incorporate varying distances in time between observations, which arise after excluding the four choice sets in the last quarter of the conjoint exercise.

Having the same choice sets in the same order across all respondents further ensures that dynamic heterogeneity is not caused by differences in the sequence of choice sets but only by cross-sectional and temporal differences between the respondents, i.e., by individual preferences, learning, and fatigue. For more details about the sample and data collection process, we refer to Broadbent et al. (2010). We chose this application as it seems reasonable to expect respondent-level dynamics here, as learning effects especially occur for highly complex and unusual products, and fatigue occurs especially for large T (e.g., $T \geq 16$).



Notes: Actual choice shares for the two "real" alternatives and the outside option (left-hand subplot). Shares for the number of times an attribute level was chosen (in green) versus it was presented (in blue) to respondents across choice sets (middle and right-hand subplots). In a perfectly balanced choice task, all attribute levels would appear equally often (dashed line).

Figure 1: Descriptive statistics.

Figure 1 shows that the choice task design was quite well (yet not perfectly) balanced (subplots 2-4, blue bars). Considering this, it can be concluded that the respondents tended to choose higher numbers of trees and lower donations. This coincides with the expectation that the demand function should be downward-sloping in donations and upward-sloping in the number of trees. In addition, it can be seen from the first subplot that both "real" alternatives (positions 1 and 2) were chosen nearly equally often, which can be expected in the case of a well-balanced design. In contrast, the outside option was chosen less often (with a share of only about 10%), indicating that the study's design (i.e., the levels for the attributes) was well-chosen.

4.2 Model Specification

The model in which all effects are modeled with heterogeneously time-varying coefficients, i.e., model (5), can be expressed for our empirical application as

$$\begin{aligned}\eta_{itr} =& \beta_0 + \beta_{0i}(t) + (\beta_1 + \beta_{1i}(t))\ \textit{non-native}_{itr} \\ & + (\beta_2 + \beta_{2i}(t))\ \textit{native}_{itr} + (\beta_3 + \beta_{3i}(t))\ \textit{donation}_{itr}, \quad r = 1, \ldots, c. \end{aligned} \tag{6}$$

The time effects, i.e., the main effect in t direction, are modeled with B-spline basis functions of degree 2 with 5 knots and a *random walk* prior of order 1. The number of knots is low for a penalized model, but it ensures that there are data points between any two knots. The hyperprior for the smoothness parameters is a half-normal distribution

$$\tau_0^{(i)}, \ldots, \tau_3^{(i)}, \tau_0^{(t)}, \ldots, \tau_3^{(t)} \overset{iid}{\sim} \mathrm{HN}_+(\sigma_\tau = 100)$$

such that the variance equals $100^2(1 - \frac{2}{\pi}) \approx 3634$. Due to this large variance, the half-normal hyperprior is weakly informative as explained in Section 3. A Gaussian prior

$$\beta_0, \ldots, \beta_3 \overset{iid}{\sim} \mathrm{N}(0, 5)$$

is used for the global effects.

The reference models are fitted using the same prior distributions. Also, the same weakly informative half-normal distribution, which we use as hyperprior for the smoothing variances, is used as the prior distribution for the parametric random effects variances.

Convergence of the MCMC estimation is investigated visually using trace plots and based on convergence diagnostics, namely $\widehat{R}$ and the effective sample size (ESS). The corresponding results are included in the supplementary materials (Appendix A). Neither the visual analysis nor the convergence diagnostics indicate any issues.

4.3 Results

Estimates: The estimation results of our proposed model M0, i.e., model eq. (6), are given in Table 1 (Appendix C shows the full estimation result for all models with heterogeneity). The posterior mean point estimates for the global parameters (intercept, non-native, native, and donation) show the expected signs and are all significantly different from zero (in the sense that their 95% credible intervals do not contain zero). This result aligns with the general results of the MNL model in Broadbent et al. (2010) (see also Appendix B for further details regarding comparing model M1 with the original model specification). The lower boundaries of the 95% credible intervals for the smoothness/variance parameters $\hat{\tau}_l^{(i)}$, $l = 0, \ldots, 3$, representing cross-sectional heterogeneity between respondents, are clearly positive for both the three covariate effects and the intercept which indicates that cross-sectional heterogeneity is an issue in the data. The lower boundary of the credible interval for the smoothness parameter of the donation variable in t-direction $\hat{\tau}_3^{(t)}$ is clearly different from zero as well, suggesting quite a lot of time variation in the effect of the donation variable ($\hat{\tau}_3^{(t)}$). In contrast, the lower boundaries of the corresponding credible intervals for the native variable and the intercept ($\hat{\tau}_0^{(t)}$, $\hat{\tau}_2^{(t)}$) are rather close to zero and extremely close to zero for the non-native variable ($\hat{\tau}_1^{(t)}$). Therefore, it is at least doubtful whether time variation in the effect of the number of non-native trees to be removed is actually an issue. In addition, the time variation in the effects of the intercept and the native variable seems to be quite small. Since a weakly informative prior is used, it is ensured that these outcomes are not a result of the prior but the data.

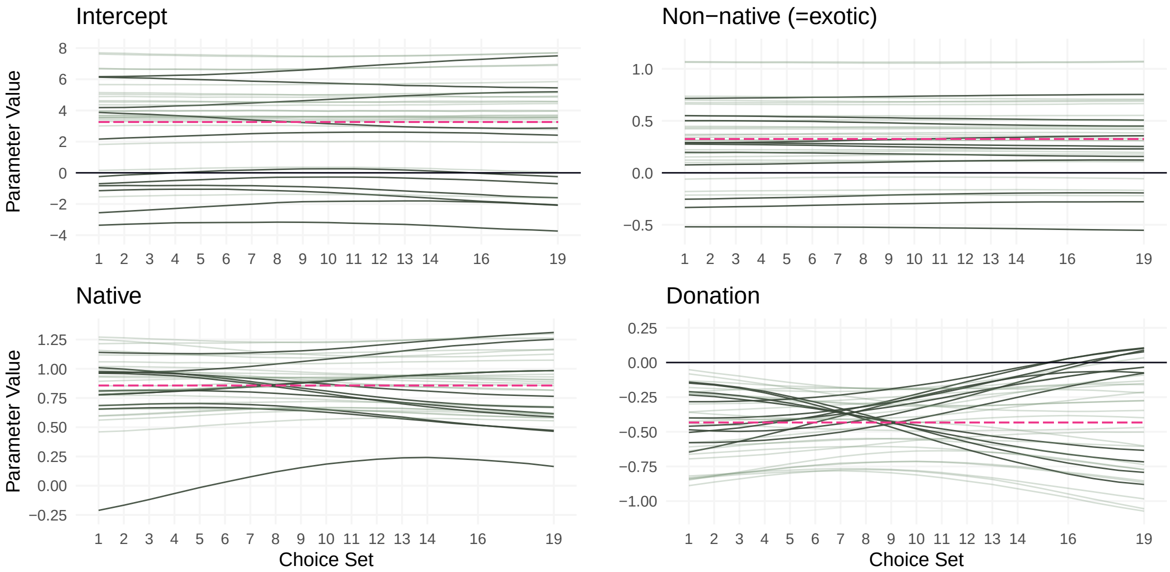
Individual Curves: The estimated individual curves for the intercept and the three covariate coefficients are visualized in Figure 2. We included the fixed effects (i.e., time-constant effects at the population level, eq. (5)) in the plots to simplify the interpretation of (individual) deviations in the time and respondent directions (red dashed horizontal lines).

Table 1: Estimation results of the proposed model M0.

	Mean	SD	95%-CI
Fixed Effects:			
Intercept (β_0)	3.250	1.573	[0.427, 6.683]
Non-native (β_1)	0.324	0.115	[0.117, 0.565]
Native (β_2)	0.859	0.158	[0.597, 1.208]
Donation (β_3)	-0.432	0.099	[-0.652, -0.263]
Smoothness Parameters:			
Heterogeneity:			
$\tau_0^{(i)}$	29.704	9.845	[13.945, 52.427]
$\tau_1^{(i)}$	2.772	0.653	[1.715, 4.273]
$\tau_2^{(i)}$	2.254	0.654	[1.132, 3.693]
$\tau_3^{(i)}$	1.859	0.593	[0.722, 3.107]
Dynamic:			
$\tau_0^{(t)}$	2.566	2.036	[0.088, 7.496]
$\tau_1^{(t)}$	0.146	0.116	[0.005, 0.443]
$\tau_2^{(t)}$	0.554	0.332	[0.043, 1.279]
$\tau_3^{(t)}$	0.646	0.218	[0.262, 1.118]

Notes: Point estimates based on the posterior means, posterior standard deviations (SD), and the 95% credible intervals (CI).

A considerable amount of cross-sectional heterogeneity between respondents is visible for all four coefficients, as represented by the different anchorings of the splines on the y-axis. The comparably large estimate for $\tau_0^{(i)}$ further explains the much larger cross-sectional heterogeneity for the intercept compared to the three covariates (see the scaling on the y-axis), implying very different individual (status-quo) utilities and choice shares for the outside-good across the respondents. Interestingly, we observe both positive and negative signs for the non-native coefficient across respondents, meaning some respondents prefer removing non-native trees while others do not. In contrast, coefficients (except for one respondent in some choice sets) are positive and less heterogeneous across respondents for the native effect, implying a quite clear positive number of respondents in favor of planting native trees. Finally, the negative sign for the donation effects makes sense since we could expect a lower preference for higher donations. Respondents are nevertheless very heterogeneous, i.e., differently donation-sensitive.



Notes: Dark lines indicate curves of the respondents with the highest variation over choice sets (top 50%).

Figure 2: Individual marginal curves of model M0.

As could already be expected from the estimation results shown in Table 1, there is almost no variation over time for the non-native covariate effect; respondents' preferences are highly stable over choice sets here, as the splines appear as almost constant (horizontal) lines for the majority of respondents. The absence of a significant within-respondent variation along the time dimension implies the absence of an interaction effect because there cannot be an interaction if one of both main effects is missing. This coincides with the result that the lower bound of the 95% credible interval of $\hat{\tau}_2^{(t)}$ is extremely close to zero. A similar pattern with missing dynamics is observed for many respondents with regard to the intercept; however, we here see noticeable exceptions for other respondents, with different shaped trends for these respondents. This indicates that the interaction effect between cross-sectional and time heterogeneity is still present for the intercept.

More dynamic heterogeneity can be observed for the native effect, where we find evidence for both crossover and diverging patterns for a number of respondents (Dew et al.

2020). The curves for these respondents are highlighted curves in the lower left subplot of Figure 2. Crossover means that a respondent's curve crosses over the time-constant population mean (red dashed line), indicating that the respondent started out with a higher preference for planting native trees compared to other respondents and then moved to a below-average valuation towards the end of the choice experiment (or vice versa). Crossover coincides with a strong change in a respondent's preferences over time and implies dynamic heterogeneity at the individual respondent level but relatively stable trajectories across individuals. Diverging means that an individual's preference moves away from near the population mean initially to the extremes of the preference distribution for this attribute the longer the choice experiment lasts (cf. Dew et al. 2020, p. 66). Diverging also implies a relatively strong change of preferences at the individual level. It would be very difficult to capture these patterns in a model without individual-level dynamics.

The splines have the most diverse shapes for the donation variable, suggesting different non-linear dynamics at the individual respondent level. In other words, one can observe a distinct cross-sectional heterogeneity between respondents (as the anchorings of the splines are very different), non-linear dynamic heterogeneity (non-linear preference evolution for respondents), and their interaction (as the shapes of the splines vary a lot between different respondents). As a result, we see that some respondents stay relatively stable in the donation effect, while others reveal strongly non-linear increasing or decreasing donation sensitivities along the choice experiment. The fact that all respondents faced the same choice tasks in the same order excludes the possibility that differences between respondents may result from differences in the sequence of presented choice sets.

Different explanations for the evidence of dynamic heterogeneity (i.e., the evolution of a respondent's preferences) in the choice experiment are possible. Increasing curves for the native and non-native effects indicate an increasing personal valuation of the utility contribution for planting and removing trees over the choice tasks. Decreasing curves, on the other hand, imply the opposite effect. Increasing (decreasing) curves could result from an initial undervaluation (overvaluation) of the respective effect due to a lack of initial experience and knowledge and a subsequent correction of this underestimation (overestimation) due to learning effects from being confronted with repeated choice situations. An increasing curve for the intercept parameter may mean that holding all other conditions constant (ceteris paribus), the preference not to choose the outside option increases. This might most likely be a learning effect. The opposite effect, i.e., an increasing preference for the outside option, is likely to be caused by fatigue or boredom or, more generally, by simplifying one's answering behavior. For the donation variable, a decreasing curve indicates an increasing price sensitivity, while an increasing curve indicates a decreasing price sensitivity. It can be observed that at the end of the first half of the choice sets (approx. between $t = 7$ and $t = 10$), the heterogeneity in the donation effect between respondents is comparably small. A plausible reason for this could be that, at this point, learning effects may already have almost completely occurred, but fatigue and boredom have not yet. Fatigue or boredom reduces respondents' attention in later choice tasks and can lead to biased effect estimates.

Model Comparison: The statistical performance of our proposed model M0 in comparison to the four benchmark models (M1–M4) is evaluated along the performance measures introduced in Section 2.5. It can be observed from the results summarized in Table 2, the largest improvement across all measures results from accommodating cross-sectional heterogeneity (model M2) compared to the simplest MNL model that excludes any form of heterogeneity (model M1). This coincides with our visual impression from Figure 2 that revealed cross-sectional heterogeneity as the supposedly stronger dimension in our data. Compared to model M2, accounting additionally for dynamic heterogeneity via individual linear time trends (model M3) or via a global (homogeneous across respondents) but possibly non-linear time trend (model M4), both come with a noticeable further improvement regarding all performance measures. According to the spherical score, Brier score, and log-Lik, heterogeneous linear trends (M3) lead to a slightly better model fit than the more flexible but global dynamic model (M4). Since more complex models tend to have a better model fit, we further computed the WAIC as a penalized fit measure and also as an estimator for the out-of-sample predictive accuracy. While the Correct-% measure favors the flexible model (M4), the WAIC confirms the slightly better performance for the linear trend model (M3). However, the differences between all five performance measures are very close for these two models.

Table 2: Comparison of models M0–M4.

Model	Correct-%	Spherical score	Brier score	log-Lik	WAIC
M1	70.0	420	−241	−424	857
M2	86.4	498	−109	−193	539
M3	88.4	508	−93	−165	510
M4	88.6	505	−98	−174	516
M0	90.5	516	−79	−147	489

The proposed model with cross-sectional and individual non-linear dynamic heterogeneity (M0) comes with an additional improvement in model performance for all five performance measures. Note that the better WAIC indicates that the improvements in fit (in-sample) are not due to overfitting. Except for Correct-%, the improvement compared to model M4 is even larger than from model M2 to model M4. This means allowing for an individual and a flexible non-linear preference evolution, i.e., decoupling cross-sectional and non-linear dynamic heterogeneity, improves the model compared to one where dynamic heterogeneity is also accommodated non-linearly (via splines) but assumed to be homogeneous across respondents (M4). This coincides with observing very differently shaped individual splines for the donation variable in Figure 2. The improvement in model performance of model M0 compared to the model with heterogeneous but linear time trends (M3) is also noticeable. Again, referring to Figure 2, this might also be attributed primarily to the donation variable, where complex and different non-linear shapes are observed across the respondents. Overall, the higher complexity of our proposed model, in which the dynamic heterogeneity is decoupled from the cross-sectional heterogeneity and modeled non-linearly, seems to pay off, at least for the data at hand.

Summary: Our new modeling approach can capture individually different nonlinear preference evolutions over time, hence decoupling cross-sectional and non-linear dynamic heterogeneity. In addition, by using penalized splines as a nonparametric technique, no assumptions about the functional form of the individual preference evolution patterns are necessary; compared to parametric modeling, the functional shapes can be extracted directly from the data. Note that all benchmark models are nested in M0. For instance, if dynamic effects were completely absent, the model degenerates to model M2 with cross-sectional heterogeneity only; similarly, if only linear time trends would exist for all or only some covariates, decreasing or increasing linear effects can be obtained as special cases of splines for these covariates. Finally, the different dynamics for the three covariate effects and the intercept rule out that the dynamics are due to time-varying scale heterogeneity.

4.4 Practical Relevance

So far, this section has outlined that our modeling approach can represent more complex effects in the empirical data than competing models (i.e., better model performance) without overfitting (i.e., robust results based on penalized splines). However, a remaining question is whether this better model performance has practical relevance.

To investigate the practical relevance, we analyze 1) the acceptance probability, i.e., the probability that a person decides to donate, as well as the expected donations of the different models for a given forest restoration policy, and 2) the willingness-to-pay (WTP) for planting native trees and removing non-native trees.

Outcomes of a Policy: We chose as an example a relatively controversial policy in which many of the highly flammable non-native trees are removed (non-native = 17) while only a few native trees are planted (native = 1) in exchange. Such a policy may be able to prevent wildfires quite well in the short term but might not have a large effect in the long term. The donation is set to the middle of its range, i.e., \$9.50.

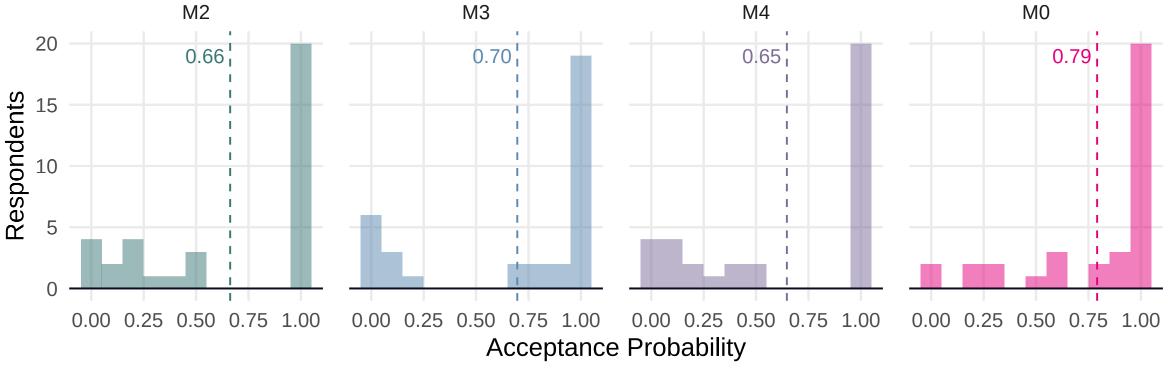


Figure 3: Acceptance probabilities for the heterogeneous models.

We compute the acceptance probability of this policy, i.e., the choice probability of the policy vs. the outside option. While for model M1, the acceptance probability is constant across respondents and estimated to be 0.61, for the other models, the estimated distribution of the acceptance probability is shown in Figure 3 (we use the posterior mean of the avg. parameter value for each reposed and model). Given that our model can represent the true effects most precisely, it follows that the models M2–M4 overestimate the number of respondents with low acceptance probabilities (≤ 0.2) by 17.1, 20.0, and 22.9%-points, respectively. In addition, the models M2–M4 underestimate the number of respondents with large acceptance probabilities (≥ 0.8) by 14.3, 5.7, and 14.3%-points, respectively. Consequently, all competing models also have lower average acceptance probabilities of 0.66 (M2), 0.70 (M3), and 0.65 (M4) (see dashed lines in Figure 3), compared to the value of 0.80 of the M0 model. These differences also lead to relatively large differences in the expected donations per capita. Specifically, while models M1–M4 have values of \$5.83,

$6.29, $6.62, and $6.15, the expected donation based on model M0 is with $7.56 higher (on a relative scale, between 12% and 18%). This relatively large difference in the expected donation can potentially change the decision of whether to start the donation campaign. Therefore, the better performance of M0 is of theoretical interest and practical relevance.

Willingness-To-Pay: Next, we calculate WTP values by dividing the model intercept and the slope parameters for removing non-native trees and planting native trees by the donation parameter. Representing utilities in monetary units is more intuitive for practitioners and decision-makers. However, it also allows us to better understand the dynamic heterogeneity as we now can easily compare WTP results across respondents or time, which can be problematic using utility parameters as they are not scale-invariant. Please note Figure 2 revealed that the donation parameter varied the most over choice sets at the respondent level. Hence, we expect all WTP values to also vary over the choice set. Before analyzing WTP dynamics, we first present aggregate values for each model.

Table 3: Median WTP of the heterogeneous models.

Model	Intercept	Non-native	Native
M2	$8.46	$0.52	$2.14
M3	$9.54	$0.67	$2.38
M4	$10.88	$0.66	$2.56
M0	$9.82	$0.76	$2.35

Table 3 summarizes median WTP values across (the mean value of) the respondents as a robust measure for the potential large values (Sonnier et al. 2007). The results show that there are notable differences across models. First, the model with heterogeneity but ignoring any dynamics (M2) has the lowest median WTP values. Second, M0 shows the highest WTP for removing (a) non-native tree ($0.76), but a value between the lowest value (M2, $2.14) and the highest value (M4, $2.56) of $2.35. Third, all the intercept values (i.e., the baseline WTP for the inside good) are close to the avg. shown donation values of $9, reflecting a reasonable tradeoff between utility and donations in the experiment.[1] Third, models M0 and M3 appear to have more similar WTPs compared to the other models, which is reasonable as those models both account for dynamics heterogeneity.

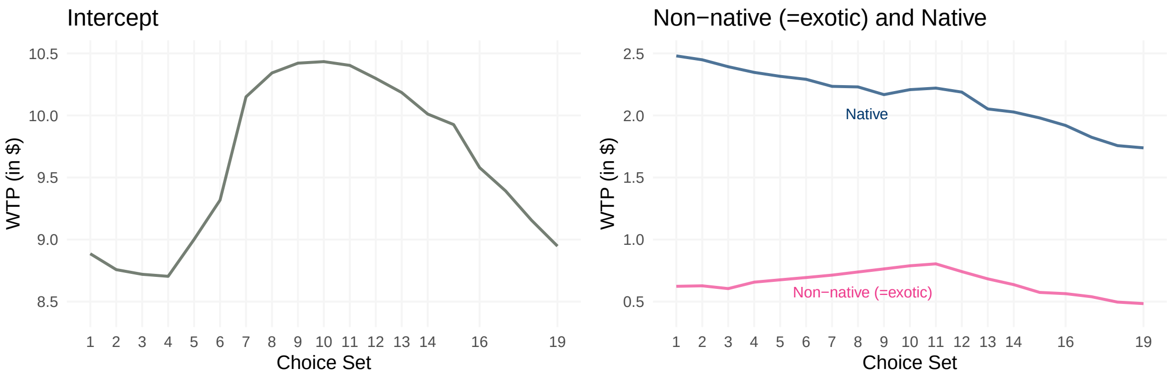


Figure 4: Median WTP values for model M0 per choice task.

Next, we focus on the dynamic aspects of WTP. As in Table 3, Figure 4 also show the median WTP values for model M0, but now for each choice task separately. The figure reveals interesting patterns. After a slight decrease for the first four choice tasks, the WTP for the intercept has an inverted U-shape, with a minimum of \$8.70 (similar to M2) and a maximum close to \$10.50 (similar to M4). The WTP for removing non-native trees also has an inverted U-shape, with a maximum value of \$0.80 (choice set 11) and a minimum below \$0.50 at the end of the experiment. The WTP for planting new native trees is mostly declining across the 19 choice tasks, with values from \$2.48 at the start to \$1.73 at the end. These differences in WTP over choice sets are economically relevant as the length of the experiments clearly affects the results at the aggregate level. The results raise the question of whether the dynamics of the median WTP values reflect the respondent-level results. Our M0 model allows such an analysis as we can easily compute respondent-level WTPs

[1] This is not the case for the model without heterogeneity (M1), with a negative baseline WTP value of \$−0.27. However, the WTP value for removing a non-native tree appears to be comparable (\$0.63) to the other models, but the WTP value for planting a native tree is much lower (\$1.88). See Appendix B for a detailed WTP analysis of model M1, incl. a comparison with the results in Broadbent et al. (2010).

for each choice set. We do not show the individual results for all respondents, but a subset of respondents that are typical for groups of respondents in the sample in Figure 5.

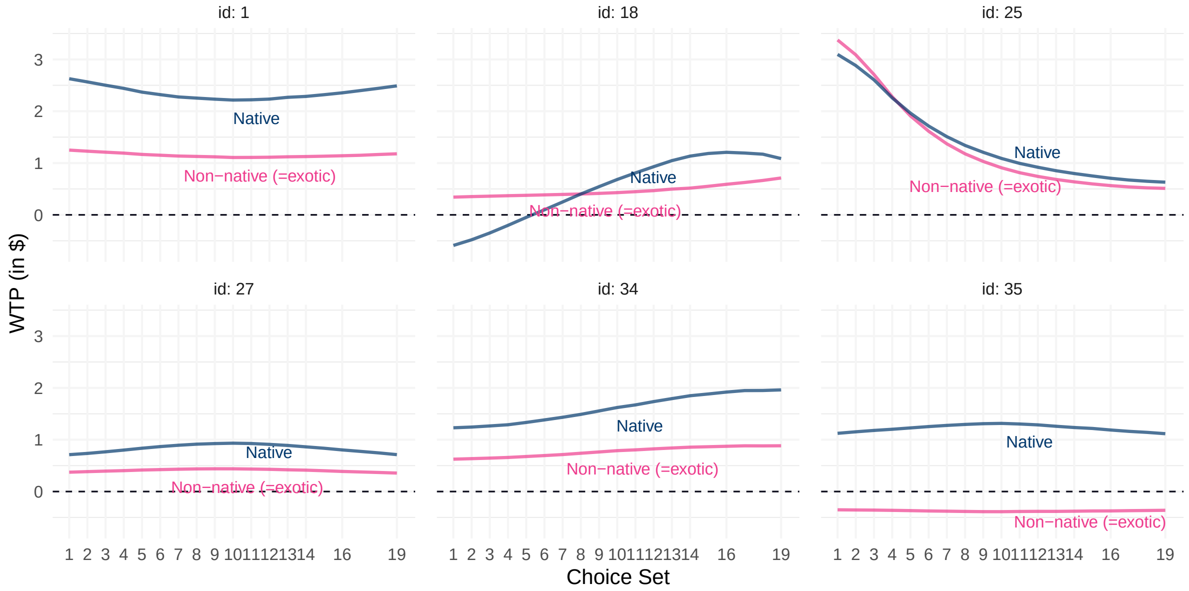


Figure 5: WTP values for model M0 per choice task of selected respondents.

In addition to different levels for the WTP values that reflect purely cross-section heterogeneity, we also see different shapes across choice sets. For example, respondents 1, 27, and 35 have (inverted) U-shaped WTP values for planting native trees. On the other hand, respondents 18 and 34 have an increasing pattern, whereas the WTP of respondent 25 declines over choice sets. We also see very heterogeneous patterns in WTP for removing non-native trees, but given their lower magnitudes, the dynamics appear less prominent. Respondent 18 is also interesting, as the WTP values cross each other over time. At the beginning of the experiment, the person shows a lower (even negative) WTP for planting native trees compared to removing non-native trees. However, after choice set 8, this order is reversed, and at the end of the experiment, the WTP for planting native trees is almost twice as large as the one for removing non-native trees. Note that this outcome would not be possible in models without dynamic heterogeneity (e.g., model M3). The variation in WTP trajectories also questions whether dynamic patterns only occur because of learning or fatigue. Most likely, such causes exist simultaneously across respondents, showing the importance of our M0 model, which can account for this.

To summarize, the results in this section have shown that the outcomes of the models differ. Even at the aggregate level (e.g., acceptance probabilities, median WTP), it appears important to account for preference heterogeneity and dynamics. Furthermore, WTP patterns across choice tasks vary considerably at the aggregate and respondent levels, highlighting the practical value of model M0.

5 Simulation Study

We conduct a simulation study to reveal whether or not the heterogeneously time-varying coefficient model captures complex time effects better than the other models. Specifically, we focus on the heterogeneous models M2 (no dynamics), M4 (agg. dynamics), and M0 (dynamic heterogeneity). We used each model as a generating process (DGP), sampled new observations from the corresponding multinomial distribution, fitted the models to each new dataset, and evaluated the results using measures introduced in Section 2.5. As the Bayesian model estimation has high computational costs, 100 simulation repetitions are used as a trade-off between the accuracy and feasibility of the simulation.

The boxplots in Figure 6 show the log-Lik and WAIC results for each DGP and model combination. As in the empirical application, our model (M0) is superior to the other models when the true DGP is also M0 (higher median log-Lik and smaller median WAIC). As the empirical results of model M0 reported in Figure 2 do not clearly show any aggregate level dynamics, it is unsurprising that the WAIC of models M2 and M4 are almost the same. However, there is an indication of an overfitting of the M4 model as the log-Lik value is slightly higher than that of M2. Similarly, when M2 is the true DGP (i.e., no dynamics at all), both models with dynamics fit the data better in-sample, but this is not true for the

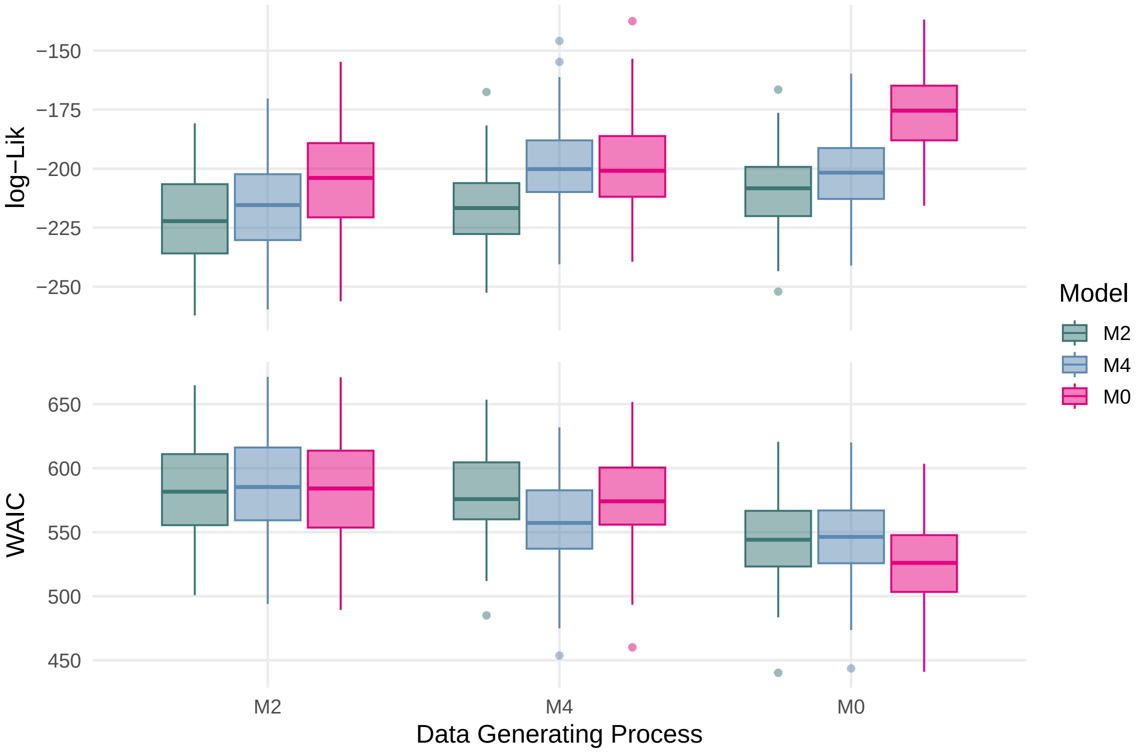


Figure 6: Simulation results.

out-of-sample predictive accuracy in terms of WAIC. Lastly, when M4 is the true DGP, both models with dynamics fit comparably well in-sample, with almost identical median log-Lik values, but M4 is clearly better out-of-sample. Note that this makes intuitive sense, as M0 is more flexible and can also deal with aggregate level preference dynamics, but at the price of an unnecessarily complex model.

To summarize, our new model (M0) handles individual and aggregate-level dynamics well while avoiding overfitting due to the penalization. The WAIC correctly identifies the correct model for each DGP, even in an application with a conservative amount of dynamic heterogeneity. Thus, we suggest using our model for preference measurement as a default and comparing it against simpler alternatives (M2 or M4) using fit measures and the WAIC.

6 Conclusion

We presented how the concept of anisotropic tensor product interactions can be used to construct highly flexible group-specific spline curves and, based on this, heterogeneously time-varying coefficients. The choice of priors and hyperpriors introduces regularization and, therefore, ensures that the flexibility of the approach does not lead to overfitting.

In Section 4, an exemplary application is presented, which outlines that heterogeneous time variation exists in the empirical data for at least some of the variables. In addition, it highlights that in cases where heterogeneous time variation is absent, the model does not lead to overfitting due to the penalized estimation. Investigating several fit measures shows that our approach captures the effects in the data better than any of the four competitive models M1-M4. Precisely, the WAIC implies that this better model fit does not result from overfitting. In addition, Section 4.4 outlines that this better model fit is not only of theoretical interest but has practical relevance. Finally, the simulation study conducted in Section 5 underlines the findings from the empirical application.

In conclusion, the presented modeling approach can capture highly complex time variations in the data. As demanded in the literature (especially by Guhl et al. (2018)) and shown in the application, this is helpful in the context of CBC studies, especially for a large number of repetitions T (20 choice sets in our application) and/or for complex products (e.g., public goods). In addition, as the model results in one curve per respondent for each variable, it also comes with good interpretability.

Some topics remain for future research: Further research is needed to investigate the transferability of this approach to less structured applications, e.g., customer purchase datasets. In such applications, the number of observations per respondent, as well as the time between observations, usually vary between respondents, and for some respondents, the data might be quite scarce between observations. In addition, (considerably) larger datasets might occur in other applications than CBC, e.g., scanner data where thousands of respondents/households are included. For analyzing these datasets, it might be useful to transfer this modeling approach into a frequentist framework. Even though the Bayesian approach has theoretical advantages, it might not be feasible for much larger datasets due to the high computational costs.

References

Abe, M. (1999), 'A generalized additive model for discrete-choice data', *Journal of Business & Economic Statistics* **17**(3), 271–284.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**(6), 716–723.

Allenby, G. M. & Ginter, J. L. (1995), 'Using extremes to design products and segment markets', *Journal of Marketing Research* **32**(4), 392–403.

Aschersleben, P. & Steiner, W. J. (2022), 'A semiparametric approach to estimating reference price effects in sales response models', *Journal of Business Economics* **92**(4), 591–643.

Baumgartner, B., Guhl, D., Kneib, T. & Steiner, W. J. (2018), 'Flexible estimation of time-varying effects for frequently purchased retail goods: a modeling approach based on household panel data', *OR Spectrum* **40**, 837–873.

Baumgartner, B. & Steiner, W. J. (2007), 'Are consumers heterogeneous in their preferences for odd and even prices? findings from a choice-based conjoint study', *International Journal of Research in Marketing* **24**(4), 312–323.

Biller, C. & Fahrmeir, L. (2001), 'Bayesian varying-coefficient models using adaptive regression splines', *Statistical Modelling* **1**(3), 195–211.

Briesch, R. A., Chintagunta, P. K. & Matzkin, R. L. (2010), 'Nonparametric discrete choice models with unobserved heterogeneity', *Journal of Business & Economic Statistics* **28**(2), 291–307.

Broadbent, C., Grandy, J. & Berrens, R. (2010), 'Testing for hypothetical bias in a choice experiment using a local public good: Riparian forest restoration', *International Journal of Ecological Economics and Statistics* **19**(F10), 1–19.

Bürkner, P.-C. (2017), '`brms`: An R package for Bayesian multilevel models using Stan', *Journal of Statistical Software* **80**(1), 1–28.

Day, B., Bateman, I., Carson, R., Dupont, D., Louviere, J., Morimoto, S., Scarpa, R. & Wang, P. (2012), 'Ordering effects and choice set awareness in repeat-response stated preference studies', *Journal of Environmental Economics and Management* **63**(1), 73–91.

DeSarbo, W. S., Fong, D. K., Liechty, J. & Coupland, J. C. (2005), 'Evolutionary preference/utility functions: A dynamic perspective', *Psychometrika* **70**(1), 179–202.

Dew, R., Ansari, A. & Li, Y. (2020), 'Modeling dynamic heterogeneity using Gaussian processes', *Journal of Marketing Research* **57**(1), 55–77.

Duane, S., Kennedy, A., Pendleton, B. J. & Roweth, D. (1987), 'Hybrid Monte Carlo', *Physics Letters B* **195**(2), 216–222.

Eilers, P. H. C. & Marx, B. D. (1996), 'Flexible smoothing with B-splines and penalties', *Statistical Science* **11**(2), 89–121.

Eilers, P. H. C. & Marx, B. D. (2021), *Practical Smoothing: The Joys of P-splines*, Cambridge University Press.

Elshiewy, O., Guhl, D. & Boztug, Y. (2017), 'Multinomial logit models in marketing–from fundamentals to state-of-the-art', *Marketing ZFP* **39**, 32–49.

Fahrmeir, L., Kneib, T., Lang, S. & Marx, B. (2013), *Regression: Models, Methods and Applications*, Springer-Verlag, Berlin/Heidelberg.

Frühwirth-Schnatter, S., Tüchler, R. & Otter, T. (2004), 'Bayesian analysis of the heterogeneity model', *Journal of Business & Economic Statistics* **22**(1), 2–15.

Gelman, A. (2006), 'Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)', *Bayesian Analysis* **1**(3), 515–534.

Gelman, A., Hwang, J. & Vehtari, A. (2014), ‘Understanding predictive information criteria for Bayesian models’, *Statistics and Computing* **24**(6), 997–1016.

Guhl, D., Baumgartner, B., Kneib, T. & Steiner, W. J. (2018), ‘Estimating time-varying parameters in brand choice models: A semiparametric approach’, *International Journal of Research in Marketing* **35**(3), 394–414.

Hess, S., Hensher, D. & Daly, A. (2012), ‘Not bored yet – revisiting respondent fatigue in stated choice experiments’, *Transportation Research Part A: Policy and Practice* **46**(3), 626–644.

Hoffman, M. D. & Gelman, A. (2014), ‘The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo’, *Journal of Machine Learning Research* **15**(1), 1593–1623.

Jain, D. C., Vilcassim, N. J. & Chintagunta, P. K. (1994), ‘A random-coefficients logit brand-choice model applied to panel data’, *Journal of Business & Economic Statistics* **12**(3), 317–328.

Kamakura, W. A. & Wedel, M. (2004), ‘An empirical Bayes procedure for improving individual-level estimates and predictions from finite mixtures of multinomial logit models’, *Journal of Business & Economic Statistics* **22**(1), 121–125.

Keane, M. P. (1997), ‘Modeling heterogeneity and state dependence in consumer choice behavior’, *Journal of Business & Economic Statistics* **15**(3), 310–327.

Kessels, R., Jones, B., Goos, P. & Vandebroek, M. (2009), ‘An efficient algorithm for constructing Bayesian optimal choice designs’, *Journal of Business & Economic Statistics* **27**(2), 279–291.

Kim, J. G., Menzefricke, U. & Feinberg, F. M. (2005), ‘Modeling parametric evolution in a random utility framework’, *Journal of Business & Economic Statistics* **23**(3), 282–294.

Kim, J. G., Menzefricke, U. & Feinberg, F. M. (2007), ‘Capturing flexible heterogeneous utility curves: A Bayesian spline approach’, *Management Science* **53**(2), 340–354.

Kneib, T., Baumgartner, B. & Steiner, W. J. (2007), ‘Semiparametric multinomial logit models for analysing consumer choice behaviour’, *AStA Advances in Statistical Analysis* **91**, 225–244.

Kneib, T., Klein, N., Lang, S. & Umlauf, N. (2019), ‘Modular regression–a Lego system for building structured additive distributional regression models with tensor product interactions’, *TEST* **28**, 1–39.

Lachaab, M., Ansari, A., Jedidi, K. & Trabelsi, A. (2006), ‘Modeling preference evolution in discrete choice models: A Bayesian state-space approach’, *Quantitative Marketing and Economics* **4**(1), 57–81.

Lang, S. & Brezger, A. (2004), ‘Bayesian P-splines’, *Journal of Computational and Graphical Statistics* **13**, 183–212.

Li, Y., Krefeld-Schwalb, A., Wall, D. G., Johnson, E. J., Toubia, O. & Bartels, D. M. (2022), ‘The more you ask, the less you get: When additional questions hurt external validity’, *Journal of Marketing Research* **59**(5), 963–982.

Liechty, J. C., Fong, D. K. & DeSarbo, W. S. (2005), ‘Dynamic models incorporating individual heterogeneity: Utility evolution in conjoint analysis’, *Marketing Science* **24**(2), 285–293.

McElreath, R. (2020), *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, Chapman & Hall/CRC Press, New York/Boca Raton.

McFadden, D. (1973), ‘Conditional logit analysis of qualitative choice behavior’, *Frontiers in Econometrics* pp. 105–142.

Netzer, O., Toubia, O., Bradlow, E. T., Dahan, E., Evgeniou, T., Feinberg, F. M., Feit, E. M., Hui, S. K., Johnson, J., Liechty, J. C., Orlin, J. B. & Rao, V. R. (2008), ‘Beyond conjoint analysis: Advances in preference measurement’, *Marketing Letters* **19**(3), 337–354.

Rao, V. (2014), *Applied Conjoint Analysis*, Springer.

Savage, S. & Waldman, D. (2008), 'Learning and fatigue during choice experiments: A comparison of online and mail survey modes', *Journal of Applied Econometrics* **23**(3), 351–371.

Sonnier, G., Ainslie, A. & Otter, T. (2007), 'Heterogeneity distributions of willingness-to-pay in choice models', *Quantitative Marketing and Economics* **5**, 313–331.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A. (2002), 'Bayesian measures of model complexity and fit', *Journal of the Royal Statistical Society Series B* **64**(4), 583–639.

Stan Development Team (2020), '`RStan`: the `R` interface to Stan'. R package version 2.21.2.

Vehtari, A., Gelman, A. & Gabry, J. (2017), 'Practical bayesian model evaluation using leave-one-out cross-validation and WAIC', *Statistics and Computing* **27**(5), 1413–1432.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. & Bürkner, P.-C. (2021), 'Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC (with Discussion)', *Bayesian Analysis* **16**(2), 667–718.

Watanabe, S. (2010), 'Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory', *Journal of Machine Learning Research* **11**(116), 3571–3594.

Wood, S. N. (2017), *Generalized Additive Models: An Introduction With R*, 2 edn, Chapman & Hall/CRC Press, New York/Boca Raton.

Wood, S. N. (2021), '`mgcv`: Mixed GAM computation vehicle with automatic smoothness estimation'. R package version 1.8-34.

Wood, S. N. & Scheipl, F. (2020), '`gamm4`: Generalized additive mixed models using `mgcv` and `lme4`'. R package version 0.2-6.

SUPPLEMENTARY MATERIAL

Appendix A

Table 4 summarizes diagnostic information about the sampler for the estimation of model M0 (see, e.g., Vehtari et al. 2021, for details about the computation and interpretation). All $\widehat{R}$ are clearly smaller than 1.05, indicating that the Markov chains have mixed well and convergence has been achieved. Furthermore, the effective sample size values (bulk and tail) are reasonably large for reliable posterior inference.

Table 4: $\widehat{R}$ and effective sample size (ESS) of the estimation.

	$\widehat{R}$	Bulk ESS	Tail ESS
Fixed Effects:			
Intercept (β_0)	1.00	3342	4399
Non-native (β_1)	1.00	1828	2564
Native (β_2)	1.00	1888	3097
Donation (β_3)	1.00	2298	3181
Smoothness Parameters:			
Heterogeneity:			
$\tau_0^{(i)}$	1.00	1692	3171
$\tau_1^{(i)}$	1.00	1778	2666
$\tau_2^{(i)}$	1.00	2248	2856
$\tau_3^{(i)}$	1.00	781	512
Dynamic:			
$\tau_0^{(t)}$	1.00	1899	3304
$\tau_1^{(t)}$	1.00	1855	3119
$\tau_2^{(t)}$	1.00	1178	2170
$\tau_3^{(t)}$	1.00	981	1631

Figure 7 shows trace plots for all the parameters and chains of model M0. Visual inspection of the trace plots confirms convergence and good mixing of the sampler.

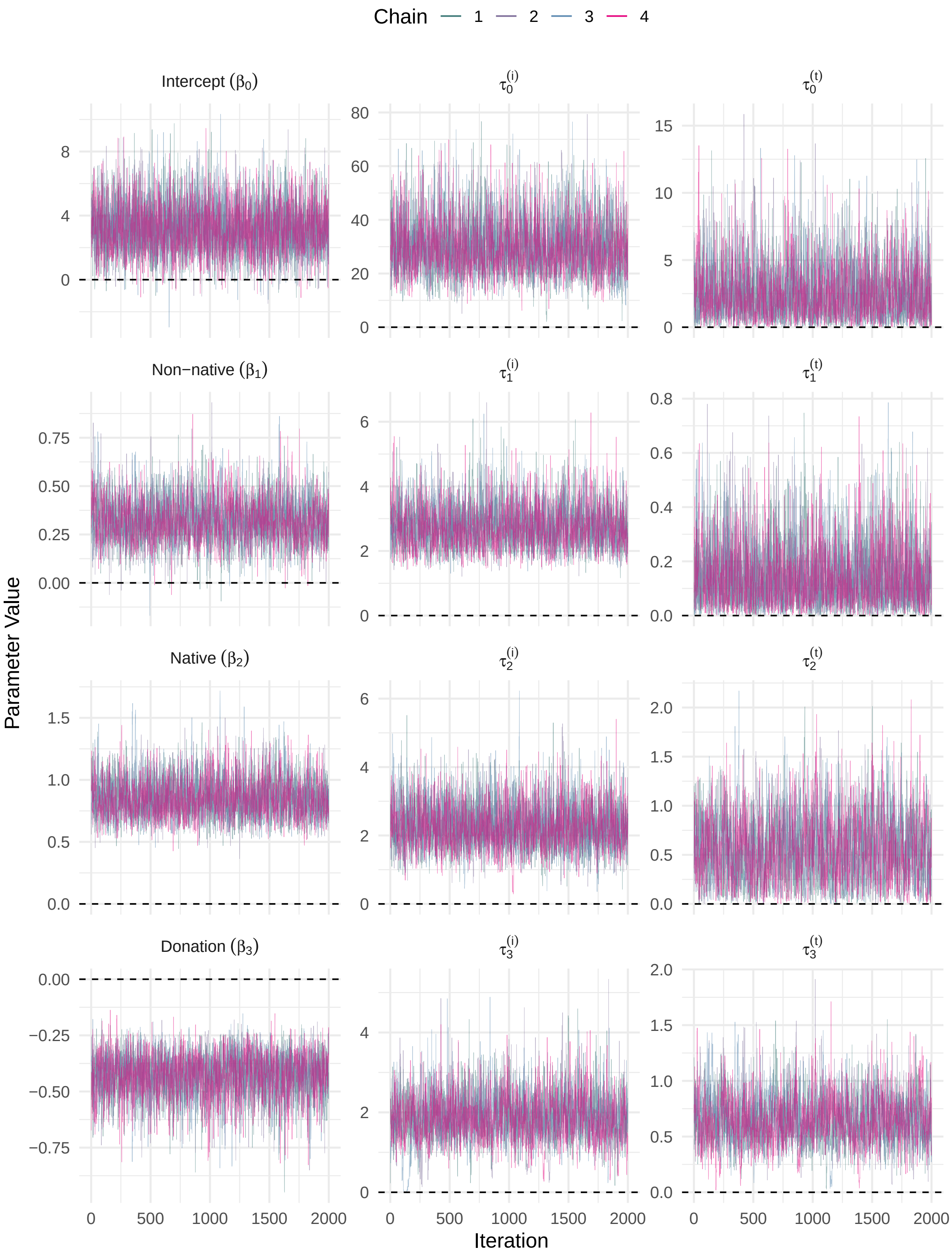


Figure 7: Trace plots of the estimation.

Appendix B

Table 5 shows the estimation results of model M1. While the last column (labeled *Bayes*) shows the results used in the paper's main part, the first two columns summarize the results using MLE. The model in the first column also differs w.r.t. the number of intercepts. We present these results to show a) that our results closely replicate the original results in Broadbent et al. (2010, Table 3) and b) that our preferred specification using one intercept instead of two alternative-specific intercepts leads to almost identical results.

Table 5: Comparison of model M1 with and without alternative-specific intercepts.

	MLE		Bayes
	Two intercepts	One intercept	One intercept
Intercept A (β_{0A})	−0.177 [−0.947, 0.594]		
Intercept B (β_{0B})	−0.022 [−0.760, 0.716]		
Intercept AB (β_0)		−0.033 [−0.774, 0.708]	−0.046 [−0.781, 0.682]
Non-native (β_1)	0.112 [0.064, 0.159]	0.105 [0.058, 0.152]	0.107 [0.061, 0.152]
Native (β_2)	0.312 [0.253, 0.371]	0.313 [0.255, 0.372]	0.316 [0.258, 0.377]
Donation (β_3)	−0.172 [−0.210, −0.133]	−0.169 [−0.207, −0.130]	−0.171 [−0.212, −0.132]
log_Lik	−423.52	−424.42	−424.43

Notes: Brackets report 95 %-confidence or -credible intervals for the frequentist or Bayesian estimation, respectively. For the model estimated using Bayesian estimation, the log-likelihood value is evaluated at the posterior means of the estimates to facilitate comparability.

All estimated intercepts are not statistically different from zero, and, therefore, the difference between the intercepts of the model in column one (as used in Broadbent et al. 2010) is not significant. Indeed, a likelihood-ratio test between the models in the first two columns shows that the model with two intercepts does not fit significantly better

($\chi^2 = 1.789$, $df = 1$, $p = 0.1811$). The small differences between the models' estimates also do not affect the substantive results. As in Broadbent et al. (2010), we computed the WTP for removing non-native trees or planting native trees by dividing the respective parameters by $-\beta_3$ and used the delta method to obtain standard errors. The WTP value for the model M1 with 2 intercepts is \$0.66 (0.15) and \$1.84 (0.22) for non-native and native trees, respectively. The corresponding values for the model with one combined intercept for both alternatives are \$0.63 (0.15) and \$1.88 (0.23). Hence, both variants of model M1 yield almost the same results. The estimation method also does not affect the WTP results, as a comparison with the reported values in the paper shows. We conclude that we can closely replicate the estimates in Broadbent et al. (2010) in the case without heterogeneity. Furthermore, using the more parsimonious version of the model with one intercept neither affects model fit nor substantive results.

Appendix C

Table 6 shows the full estimation results for all heterogeneous models (i.e., M2, M3, M4, and M0). The results for the fixed effects are quite similar across models in terms of the sign and magnitude of the effects. However, we need to be careful when interpreting utility parameters from different models (for the same data set), as the estimates are also affected by the scale of the model. Indeed, the magnitude of model M0 is slightly larger, which can be explained by the superior fit of the model (and hence a smaller error variance).

Table 6: Estimation results.

	M2	M3	M4	M0
Fixed Effects:				
Intercept	2.328	2.723	2.509	3.250
	[0.224, 4.824]	[0.293, 5.795]	[0.031, 5.568]	[0.427, 6.683]
Non-native	0.242	0.303	0.256	0.324
	[0.087, 0.415]	[0.118, 0.516]	[0.062, 0.463]	[0.117, 0.565]
Native	0.631	0.741	0.776	0.859
	[0.459, 0.835]	[0.53, 1.004]	[0.566, 1.031]	[0.597, 1.208]
Donation	-0.342	-0.418	-0.357	-0.432
	[-0.502, -0.209]	[-0.613, -0.249]	[-0.524, -0.217]	[-0.652, -0.263]
Smoothness Parameters:				
Heterogeneity:				
$\tau_0^{(i)}$	4.117	5.204	4.763	29.704
	[1.751, 7.28]	[2.298, 9.177]	[2.213, 8.255]	[13.945, 52.427]
$\tau_1^{(i)}$	0.397	0.435	0.472	2.772
	[0.257, 0.582]	[0.253, 0.691]	[0.309, 0.701]	[1.715, 4.273]
$\tau_2^{(i)}$	0.360	0.347	0.404	2.254
	[0.211, 0.561]	[0.056, 0.678]	[0.238, 0.628]	[1.132, 3.693]
$\tau_3^{(i)}$	0.341	0.435	0.337	1.859
	[0.213, 0.516]	[0.312, 0.835]	[0.203, 0.520]	[0.722, 3.107]
Dynamic:				
$\tau_0^{(t)}$		0.079	3.477	2.566
		[0.003, 0.238]	[0.068, 18.205]	[0.088, 7.496]
$\tau_1^{(t)}$		0.011	0.171	0.146
		[0.001, 0.027]	[0.003, 0.852]	[0.005, 0.443]
$\tau_2^{(t)}$		0.023	1.981	0.554
		[0.002, 0.053]	[0.294, 6.885]	[0.043, 1.279]
$\tau_3^{(t)}$		0.019	0.442	0.646
		[0.002, 0.04]	[0.021, 1.835]	[0.262, 1.118]

Notes: Posterior means and the corresponding 95 %-credible intervals. For each model, $\tau_l^{(i)}$ and $\tau_l^{(t)}$ measure for variable l the amount heterogeneity and dynamic, respectively. However, the values are not always directly comparable across models, particularly for the dynamic component.

3.4 Capturing heterogeneous time-variation in covariate effects in non-proportional hazard regression models

This article discusses multidimensional covariate effects, i.e. the third direction of multidimensionality, within the context of survival data. Here, the functional random coefficients introduced in Section 3.3 are applied to hazard regression models to capture the effect of covariates that influence the survival time in a heterogeneously time-varying manner. In contrast to Section 3.3, frequentist inference is used for numerical reasons. The superiority of this approach in comparison to competitors is demonstrated by means of a simulation study. Finally, the practical relevance of the proposed method is outlined by presenting a brain tumor case study.

Authorship: First author

Coauthors: Thomas Kneib and Kathrin Möllenhoff

Contribution statement: My contribution to this project was developing the initial idea, developing the method, programming it in `R`, implementing the simulation study, analyzing the case study data, preparing the figures and tables, and writing the first draft of the manuscript. Discussions and revising were done together with all coauthors.

Status: Under review, preprint available.

DOI: 10.48550/arXiv.2501.13525

Capturing heterogeneous time-variation in covariate effects in non-proportional hazard regression models

Niklas Hagemann[1,2], Thomas Kneib[3] and Kathrin Möllenhoff[1,2]

[1] Institute of Medical Statistics and Computational Biology, Faculty of Medicine, University of Cologne, Germany

[2] Division of Mathematics, Department of Mathematics and Computer Science, University of Cologne, Germany

[3] Chair of Statistics and Campus Institute Data Science, Georg August University Göttingen, Germany

January 24, 2025

Abstract

A central focus in survival analysis is examining how covariates influence survival time. These covariate effects are often found to be either time-varying, heterogeneous – such as being specific to patients, treatments, or subgroups – or exhibit both characteristics simultaneously. While the standard model, the Cox proportional hazards model, allows neither time-varying nor heterogeneous effects, several extensions to the Cox model as well as alternative modeling frameworks have been introduced. However, no unified framework for incorporating heterogeneously time-varying effects of covariates has been proposed. Such effects occur when a covariate influences survival not only in a heterogeneous and time-varying manner, but when the time-variation is also heterogeneous.

We propose to model such effects by introducing heterogeneously time-varying coefficients to piecewise exponential additive mixed models. We deploy functional random effects, also known as factor smooths, to model such coefficients as the interaction effect of heterogeneity and time-variation. Our approach allows for non-linear time-effects due to being based on penalized splines and uses an efficient random effects basis to model the heterogeneity. Using a penalized basis prevents overfitting in case of absence of such effects. In addition, the penalization mostly solves the problem of choosing the number of intervals which is usually present in unregularized piecewise exponential approaches. We demonstrate the superiority of our approach in comparison to competitors by means of a simulation study. Finally, the practical application and relevance are outlined by presenting a brain tumor case study.

1 Introduction

One of the major topics in survival analysis is analyzing the effect covariates have on the survival time. Frequently, the effects of these covariates can be observed to be either time-varying, heterogeneous, i.e. patient-, treatment- or subgroup-specific, or even both. If the goal is to analyze the effects of covariates on the survival time, hazard regression models play a critical role. They estimate the hazard function, which represents the instantaneous rate of occurrence of the event at a given time, conditional on survival up to that time. By incorporating covariates, hazard regression models allow to assess the impact of various factors on the hazard rate and, hence, to identify significant effects. The widely used standard model is the *Cox proportional hazards model* (Cox, 1972) where the hazard rate of an observation $i \in \{1, ..., n\}$ with corresponding covariate vector $\boldsymbol{x}_i$ is given by

$$\lambda_i(t) = \lambda_0(t) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}),$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients. The assumption of proportionality of the hazards results from the model being strictly split into the time-dependent baseline hazard $\lambda_0(t)$ and the time-constant covariate effects $\exp(x_i^\top \beta)$. In addition, the Cox model assumes the covariate effects to be (exp-transformed) linear effects.

These strict assumptions are often not fulfilled in practice (Li et al., 2015; Jachno et al., 2019). Frequently, this is caused by non-proportional hazards, i.e. the covariates or their effects are time-dependent. In addition, in many cases the effects might not be linear but of a more complex form. In particular, heterogeneity in terms of treatment-specific, subgroup-specific (e.g. gender-specific) or individual effects might be present which cannot be captured properly by linear effects. Therefore, several flexible extensions to the Cox model have been introduced. On the one hand, such extensions relax the proportional hazards assumption: The most commonly used may be the stratified Cox model in which for each level of a categorical variable a separate baseline hazard is fitted. Alternatively, several studies (see, e.g., Zucker and Karr, 1990; Murphy and Sen, 1991) propose to include time-varying coefficients in order to capture time-dependent covariate effects. In addition, Andersen and Gill (1982) introduced an approach to include time-dependent covariates.

On the other hand, extensions of the Cox model have been introduced to allow for more complex effects: Gray (1992) added non-linear smooth spline-based covariate effects and Hess (1994) uses such effects to express covariate effects as a function of time. Regarding the heterogeneity, again the stratified Cox model can be mentioned, where the baseline hazard can be group-specific. However, a more natural way to account for heterogeneous effects, which also allows for more general types of heterogeneity, is to introduce random effects leading to *frailty models* (Vaupel et al., 1979; Ripatti and Palmgren, 2000; Therneau et al., 2003).

In contrast to adding specific extensions to the Cox model, several recent studies have aimed to introduce a new flexible hazard regression framework: Kneib and Fahrmeir (2007) introduce *Cox-type structured hazard regression models*

$$\lambda_i(t|\boldsymbol{x}_i) = \lambda_0(t) \exp\left(\sum_{k=1}^{K} f_k(\boldsymbol{x}_i, t)\right) = \exp\left(\tilde{\lambda}_0(t) + \sum_{k=1}^{K} f_k(\boldsymbol{x}_i, t)\right), \tag{1}$$

where $\tilde{\lambda}_0(t)$ is the log-baseline hazard and f_k can resemble different types of effects, e.g. linear effects, smooth (spline-based) effects, time-varying effects or random effects/frailty. The inclusion of time-varying effects allows for explicit modeling of non-proportional hazards. This approach was further investigated by Hofner et al. (2011, 2013). The corresponding inference is conducted by mixed model-based penalized likelihood estimation. However, the log-likelihood involves an integral over the hazard rate. Hence, the estimation relies on numerical integration which is computationally costly and can be subject to impreciseness. Alternatively, Hennerfeind et al. (2006) proposed a Bayesian estimation scheme for such models. However, this approach is based on the same log-likelihood and, therefore, shares these disadvantages.

Another approach to introduce models of the form (1) is given by *piecewise exponential additive mixed models* (PAMM; Bender et al., 2018; Bender and Scheipl, 2018) which generalizes the concept of *piecewise exponential models* (PEM; Friedman, 1982) from linear to additive predictor terms. The underlying idea is to divide the time axis into a finite number of intervals and assume the hazard rate to be piecewise constant within these intervals. While manually choosing the interval cut-off points is a challenging task and a frequent source of criticism for PEMs, PAMMs avoid the arbitrary choice of cut-off points by using a penalized approach providing a sufficiently good fit while preventing overfitting. Under the assumption of piecewise constant hazard rates, restructuring the data leads to the likelihood of the survival model being proportional to the one of a Poisson regression model (see

Section 2 of Bender et al. (2018) for details). Hence, both models are equivalent with respect to their maximum likelihood estimators and the model parameters are estimated based on the Poisson model. Therefore, estimation can make use of existing methods and implementations for generalized linear models (GLMs) in the case of PEMs and generalized additive (mixed) models (GAMs/GAMMs) in the case of PAMMs.

While these approaches propose several flexible effects such as non-linear, time-varying and random effects, none of them include heterogeneously time-varying effects of covariates. Such effects occur if a covariate influences the survival time not only in a heterogeneous and time-varying manner, but the time-varying effect is heterogeneous, too. A typical example would be that the effect of a covariate is treatment specific, time-varying and that its time-variation is also treatment-specific, e.g., decreasing for an intervention but increasing for a placebo. To the best of our knowledge, this study is the first to propose heterogeneously time-varying covariate effects in hazard regression models.

Based on the framework of PAMMs, we introduce these heterogeneously time-varying coefficients as

$$f_{kg}(t) \cdot x_{ik},$$

where $g \in \{1, ..., G\}$ is the grouping variable. Besides treatments, the grouping can also correspond to characteristics of the participants (e.g. gender or subdiagnoses), characteristics of the study (e.g. centers in multicenter studies) or even individual heterogeneity. We propose to model such heterogeneously time-varying coefficients based on *functional random effects* (FRE; Kneib et al., 2019). This leads to *functional random coefficients*, which Hagemann et al. (2024) recently proposed to use to capture heterogeneous time-variation in covariate effects. Their study focuses on conditional logit models, a class of models commonly used in marketing research, but the approach is directly generalizable to other GAMs and can therefore be applied to PAMMs as well. Functional random effects are also known as *factor smooth interactions* or *random wiggly curves* (Wood, 2017) and are essentially tensor product interactions of smooth effects and random effects.

This paper is structured as follows: In Section 2, piecewise exponential hazard regression models including the corresponding inference are succinctly discussed. In Section 3, we introduce subgroup-specific time variation in covariate effects using functional random coefficients. A simulation study is conducted in Section 4 to show the ability of our approach to capture these effects as proposed. In addition, the simulations demonstrate that the penalized approach prevents overfitting in the absence of such effects. Section 5 illustrates the method and outlines its practical relevance by investigating the effect of fraction genome altered as a predictor of survival time in patients with brain tumors. Finally, Section 6 closes with a discussion.

2 Piecewise exponential hazard regression models

2.1 Piecewise exponential models

Piecewise exponential models (PEM; Friedman, 1982) are an alternative to classical approaches in survival regression, especially to the Cox model. Their main advantage is that the corresponding inference can be based on a Poisson model and, hence, can make use of existing tools for generalized linear models (GLMs) and generalized additive models (GAMs). They require the partition of the time axis into a finite number of intervals and assume the hazard rate to be constant within each interval. The piecewise exponential model is defined as

$$\lambda_i(t|\boldsymbol{x}_i) = \lambda_0(t_j) \exp(\eta(\boldsymbol{x}_i, t_j)) \quad \forall\, t \in (\kappa_{j-1}, \kappa_j], \tag{2}$$

where $\eta(\boldsymbol{x}_i, t_j)$ is the predictor term, $(\kappa_{j-1}, \kappa_j]$, $j = 1, ..., J$ are the intervals for which the hazard rate is assumed to be constant, $\kappa_0 = 0$ and $\kappa_J = \max(t)$. There are different ways of choosing $t_j, j = 1, ..., J$, i.e. the time values at which the hazard function 2 is evaluated. The two most frequently used approaches are interval end-points $t_j = \kappa_j \,\forall\, t \in (\kappa_{j-1}, \kappa_j]$ and interval mid-points $t_j = 0.5(\kappa_j + \kappa_{j-1}) \,\forall\, t \in (\kappa_{j-1}, \kappa_j]$.

In order to make use of the piecewise exponential approach, it is convenient to restructure the data as outlined by Friedman (1982) and Bender et al. (2018). Therefore, let T_i denote the true survival time and C_i the (non-informative) censoring time for subject $i \in \{1, ..., n\}$ such that $t_i := \min(T_i, C_i)$ is its observed right-censored time under risk. The data is then restructured, such that for each subject i there is a row for each interval j in which it was under risk. These rows contain t_{ij}, an interval-specific event indicator δ_{ij}, formally being defined by

$$\delta_{ij} = \begin{cases} 1 & \text{if } t_i \in (\kappa_{j-1}, \kappa_j] \text{ and } t_i = T_i, \\ 0 & \text{else,} \end{cases}$$

as well as an offset value $o_{ij} = \log(t_{ij})$ that gives the log-transformed time under risk and will be needed for the model estimation.

Friedman (1982) proposed a linear time-constant predictor $\eta(\boldsymbol{x}_i, t_j) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ implying a proportional hazards model. However, this easily generalizes to more complex effects, including time-varying effects, which implies a non-proportional hazards model. This leads to the class of piecewise exponential additive mixed models (PAMM; Bender et al., 2018; Bender and Scheipl, 2018).

2.2 Piecewise exponential additive mixed models

Using PAMMs, a structured additive hazard regression model of the form (1) can be constructed as

$$\lambda_i(t|\boldsymbol{x}_i) = \exp\left(\tilde{\lambda}_0(t_j) + \sum_{k=1}^{K} f_k(\boldsymbol{x}_i, t_j)\right), \quad \forall\, t \in (\kappa_{j-1}, \kappa_j].$$

where $\boldsymbol{x}_i$ denotes the covariate vector for subject i. The above notation slightly deviates from the one of Bender et al. (2018) as we do not include different effect types explicitly but implicitly as special cases of $f_k(\boldsymbol{x}_i, t_j)$. Typical examples are, among others, linear effects $f_k(\boldsymbol{x}_i, t_j) = \beta_p \cdot x_{ip}$, time-constant non-linear effects $f_k(\boldsymbol{x}_i, t_j) = f_k(\boldsymbol{x}_i)$, linearly time-varying effects $f_k(\boldsymbol{x}_i, t_j) = \beta_p \cdot x_{ip} \cdot t_j$ as well as frailty/random effects. In addition, this can be easily generalized to time-varying covariates x_{ipt}. However, since time-varying covariates are not our focus, we omit them here for notational simplicity and refer the reader to section 3.4 of Bender et al. (2018).

In contrast to classical PEMs, for which Friedman (1982) suggested a step function as baseline hazard, Bender et al. (2018) propose to use a penalized regression spline as baseline hazard for PAMMs. This eliminates the problem of manually selecting the interval cutoff points, which is a challenging task and a common criticism of PEMs. By deploying a penalized approach, the number of cut-off points just needs to be large enough to provide a sufficiently good fit while overfitting is prevented due to the penalization. Hence, the standard choice of cut-off points for PAMMs is using all unique observed survival times. For sufficiently large datasets with relatively dense and precisely measured survival times (leading to only few ties) this usually leads to very narrow intervals. Therefore, the assumption of constant hazard rates is not very strict in practice, as it only applies for quite short intervals. In addition, in many applications this makes the choice of t_j within the interval mostly irrelevant as there are no large differences. Hence, interval end points, i.e. $t_j = \kappa_j$, are often just chosen for simplicity.

2.3 Poisson-likelihood based inference and software implementation

As outlined by Bender et al. (2018) the main advantage of PAMMs is that their likelihood is proportional to the one of a Poisson regression model

$$\mathbb{E}(\delta_{ij}|x_i) = \exp(\tilde{\lambda}_0(t_j) + \eta(\boldsymbol{x}_i, t_j) + o_{ij}). \tag{3}$$

Hence, both models are equivalent with respect to their parameters and the corresponding estimation can be conducted based on the Poisson model. Therefore, the inference can be based on existing methods and one can make use of the methodological and algorithmic advances in the estimation of GAMs. This includes both, frequentist (e.g. Wood, 2011) as well as Bayesian methods (e.g. Kneib et al., 2019). While the Bayesian approach may have theoretical and interpretive advantages in many cases, it is often not numerically feasible because the data transformation discussed in Section 2.1 can strongly enlarge the datasets leading to very high computational costs.

Estimating the smoothing parameters is a challenging task in frequentist inference. Besides other alternatives (see, e.g. Fahrmeir et al. (2022) or Wood (2017) for an overview), Wood (2011) proposes a method utilizing a random effects perspective while avoiding the formal mixed model framework. The smoothing parameters are estimated directly from the restricted likelihood function without requiring the specification of a full mixed model structure. This is achieved by using a direct Laplace approximation that integrates out the random effects, i.e. the spline coefficients. Hence, this method optimizes a well-defined likelihood function directly with respect to the smoothing parameters. Therefore, it bypasses the need to solve mixed-model equations.

As discussed by Wood (2017), this method is advantageous compared to smoothness selection criterion-based and full mixed model-based approaches in terms of convergence, precision and numerical stability.

3 Capturing heterogeneous time-variation in covariate effects using functional random coefficients

Within the framework of PAMMs several effect types have already been introduced, including smooth, linearly time-varying effects $t_j \cdot f_k(x_{ik})$, linear, smoothly time-varying effects $f_k(t_j) \cdot x_{ik}$, smooth, smoothly time-varying effects $f_k(x_{ik}, t_j)$ (see table 3 of Bender et al. (2018) and table 1 of Bender and Scheipl (2018) for a complete overview) as well as random effects (in terms of log-normal frailty). However, heterogeneously time-varying effects of covariates have not yet been considered. They can be denoted as

$$f_{kg}(t) \cdot x_{ik},$$

where $g \in \{1, ..., G\}$ is the grouping variable. Formally, g also has an index i but we omit that here to avoid double indices. These effects go one step further as they are not only heterogeneous and time-varying but the time-variation is potentially heterogeneous, too.

We model these effects by *functional random coefficients*, which have been recently proposed by Hagemann et al. (2024). They are constructed by using *functional random effects* (FRE; Kneib et al., 2019) as varying coefficients. For better readability, we will leave out the effect index k for the remainder of this section. By using such FREs, we can model whole nonlinear time curves of continuous covariate effects group-specifically. FREs are essentially two-dimensional anisotropic tensor product interactions, that is $f_g(t) := f(g, t)$, of a random effect and a smooth time effect.

In order to introduce anisotropic tensor product interactions formally, we first express the two main effects $f_1(g)$ and $f_2(t)$ in terms of basis function expansion as

$$f_1(g) = \sum_{d_1=1}^{D_1} \gamma_{1d_1} B_{1d_1}(g), \quad f_2(t) = \sum_{d_2=1}^{D_2} \gamma_{2d_2} B_{2d_2}(t),$$

where $B_{1d_1}(g)$ and $B_{2d_2}(t)$ are the basis functions, γ_{1d_1} and γ_{2d_2} are the basis coefficients and D_1 and D_2 are corresponding dimensions. For an introduction to univariate basis function expansions the reader is refereed to Fahrmeir et al. (2022) or Wood (2017). Their tensor product interaction is then given by

$$f(g,t) = \sum_{d_1=1}^{D_1} \sum_{d_2=1}^{D_2} \gamma_{d_1 d_2} B_{d_1 d_2}(g,t),$$

where the *tensor product basis functions*

$$B_{d_1 d_2}(g,t) = B_{1d_1}(g) B_{2d_2}(t)$$

result from pairwise interactions of the main effect basis functions.

While such tensor product interactions are mainly used to construct interaction surfaces of continuous variables, we can also use them to interact smooth effects with random effects by choosing the basis functions correspondingly. Hence, we consider i.i.d. random effects, i.e. log-normal frailty, for the first main effect. As outlined by Kneib et al. (2019) and Hagemann et al. (2024), the corresponding basis function representation is given by

$$f(g) = \sum_{d_1=1}^{D_1} \gamma_{1d_1} B_{1d_1}(g) = \sum_{d_1=1}^{G} \gamma_{1d_1} \mathbb{1}(g = d_1) = \gamma_{1g} \tag{4}$$

with the $D_1 = G$ basis functions being indicator functions for the group membership. The time-varying effect, i.e. the second main effect, can be modeled using P-splines (Eilers and Marx, 1996). That is, using B-spline basis functions in combination with a discrete, usually a first- or second-order, penalty. Accordingly, the univariate penalty matrix of the random effect is given as a unit matrix $\boldsymbol{I}_G$ of dimension G and that of the P-spline as $\boldsymbol{D}_{D_2}^\top \boldsymbol{D}_{D_2}$ where $\boldsymbol{D}_{D_2}$ is a first or second order difference matrix of dimension D_2. Alternatively, other forms of penalized splines can be used as well, e.g. thin plate splines (Wood, 2003).

There are different ways of implementing penalties for tensor product interactions: it can be based on a straight forward combination of univariate penalty matrices (see Kneib et al. (2019) for such an approach). Alternatively, Wood et al. (2013) developed an approach which is not such a straight forward combination of univariate penalty matrices but is numerically advantageous, especially when conducting frequentist inference. This construction is based on reparameterizing the univariate smooths into fixed and random effects using an eigendecomposition of the penalty matrix. This leads to splitting the smooths into components that are not penalized, e.g. constant or linear terms, and components that are subject to penalization. The model matrix for the tensor product smooth is then constructed by calculating row-wise Kronecker products of these components. This results in each component being subject to at most one penalty which makes estimation numerically stable. For the detailed step-wise construction procedure see section 3 of Wood et al. (2013).

4 Simulations

In this simulation study, we investigate the performance of the proposed approach with regard to two critical aspects: achieving superior fit when heterogeneous time-variation is present and preventing overfitting in its absence.

4.1 Software implementation

The data transformation discussed in Section 2.1 can be conducted in `R` by using the function `as_ped` from the package `pammtools` (Bender and Scheipl, 2018).
Frequentist estimation of the Poisson model is implemented in in the `R` package `mgcv`: the method of Wood (2011) can be used via the function `gam` with `method="REML"`.
The tensor product constructor proposed by Wood et al. (2013), which is discussed in Section 3, is implemented in the `R` package `mgcv` as function `t2`. In addition, based on `t2`, a numerically optimized implementation of the FRE is given by `s(bs = "fs")`. The FRE can then be deployed as varying coefficient by linearly interacting it via the `by` argument leading to a functional random coefficient. Using this numerically optimized version, we can implement a functional random coefficient for a variable `x` as

```
s(g, t, by = x, bs = "fs", xt = list(bs = "ps"), m = c(3, 1)),
```

where `g` is the grouping variable encoded as factor, `t` is the time variable and a cubic P-spline with first order penalty is used.

4.2 Data generating processes and models

In order to generate survival data, both parts of the simulation study use the hazard function

$$\lambda_i(t|x_{1i}, x_{2i}) = \exp\left(3t + 3x_{i1} + f(x_{i2}, t, g)\right), \tag{5}$$

which depends on three explanatory variables: x_1, x_2 and g, the grouping variable with 4 levels. In the first part, referred to as scenario (I), heterogeneous time variation is deployed for the effect of x_2, i.e. $f(x_{2i}, t, g) = f_g(t) \cdot x_{2i}$. We then compare the fit of the model using the functional random coefficient for x_2, referred to as model (i), to the one of competing models, which are given by model (ii) including heterogeneity (as random effect) and time-variation but not their interaction, model (iii) including only heterogeneity, and model (iv) including only time-variation. Hence, the three competitors contain less flexible nested effects and, therefore, are suspected to have an inferior model fit. In order to investigate the prevention of overfitting, the second part of the simulation study uses these three cases as data generating processes (DGP) and again fits the four models, leading to simulation scenarios (II)-(IV). The resulting four simulation scenarios are summarized in Table 1.

Table 1: Overview over the four simulation scenarios, their DGPs and the models corresponding to these DGPs.

Scenario	Effect of x_2 in the DGP	Model
(I)	Heterogeneous time-variation	(i)
(II)	Heterogeneity & time-variation but no interaction	(ii)
(III)	Heterogeneity only	(iii)
(IV)	Time-variation only	(iv)

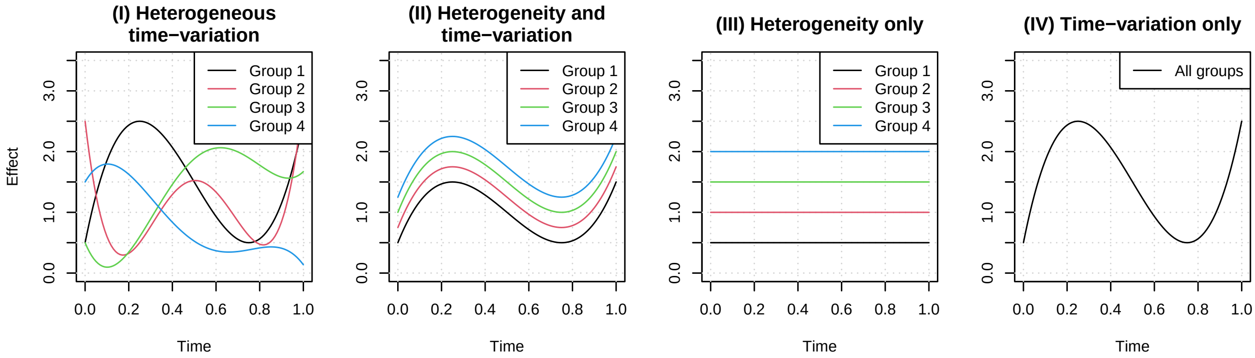


Figure 1: Effect of x_2 in the DGPs of the four scenarios. The first subplot (from the left) shows scenario (I), i.e. heterogeneous time-variation, the second one scenario (II), i.e. the combination of heterogeneity and time-variation (without interaction), the third one scenario (III), i.e. heterogeneity only, and the last one scenario (IV), i.e. time variation only.

Figure 1 shows the effect of x_2 that is deployed in the DGP of each of the four scenarios. It can be observed that in scenario (IV) all subjects follow the curve that applies for group 1 in scenario (I). The effect of scenario (II) results from adding up $\frac{1}{2}$ of the effect of scenario (III) and $\frac{1}{2}$ of the effect of scenario (IV).

The simulation is conducted with three different sample sizes, $n = 200, 400$ and 800 observations, equally distributed among the four groups. This leads to $50, 100$ and 200 observations per group, respectively. The explanatory variables x_1 and x_2 are sampled from a uniform distribution, i.e.

$$x_1, x_2 \overset{i.i.d.}{\sim} U[0, 1]$$

and the survival times are then sampled from 5 using the algorithm of Bender et al. (2005). Censoring is introduced with exponentially distributed censoring times, leading to an average censoring rate of 10.5%. The in-sample model fit is evaluated based on the log-likelihood and the integrated Brier score (IBS; Graf et al., 1999), also known as the cumulative in-sample prediction error. Hence, a smaller IBS is associated with a better model fit. The out-of-sample predictive accuracy is approximated based on an information criterion, namely the Akaike information criterion (AIC). The effective degrees of freedom, which is needed to compute the AIC of penalized models, is calculated according to Wood et al. (2016).

4.3 Results

The simulation is carried out with 1000 repetitions and it can be observed that the results are almost the same across the different sample sizes. Therefore, only the results for the medium sample size, which are shown in Figure 2, are discussed here. The outcomes for the small and large sample sizes are shown in Figures S1 and S2 of the supplementary materials.

It can be observed that, in the presence of heterogeneous time-variation, model (i) leads to a better model fit than the less flexible approaches (ii)-(iv). In addition, the difference between model (i) and the still quite flexible model (ii) is considerably larger than the difference between model (ii) and models (iii) and (iv). The fact that this also applies with regard to the AIC implies that this might not be caused by overfitting, but by modeling the underlying DGP more accurately.

Regarding the fit of the models in scenarios (II) and (IV), we can observe the desired behavior. Here, the fit of model (i) and the model resembling the DGP, i.e. model (ii) and (iv) respectively, is almost

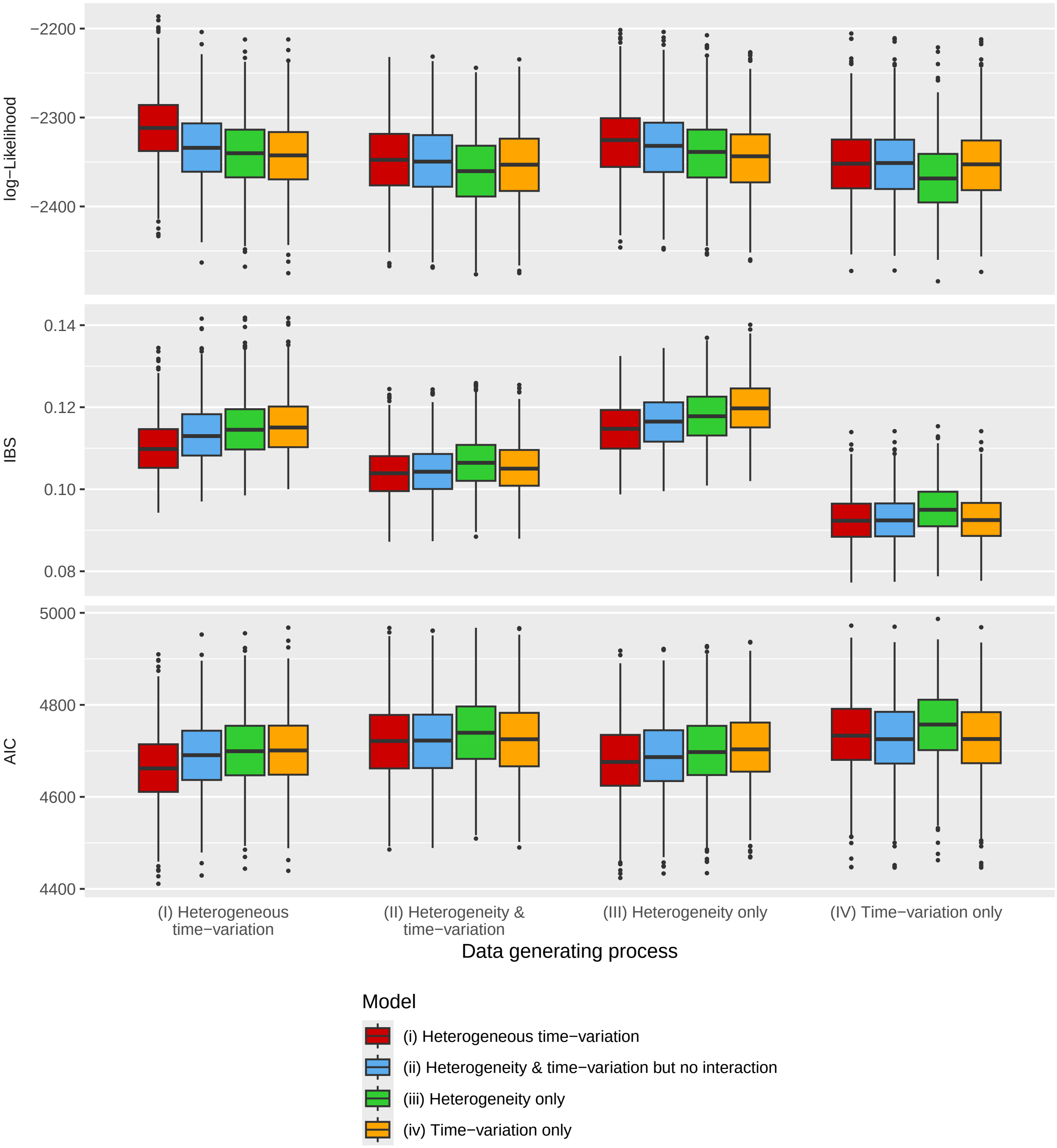


Figure 2: Results of the simulation study with $n = 400$ in terms of the three fit measures. For each of the scenarios (I) - (IV) there is one block of consisting of four boxplots, one for each model (i) - (iv).

equal for all three fit measures. This strongly indicates that model (i) is penalized towards the true (less complex) model, as desired. In contrast, in scenario (III) we can observe model (i) to fit slightly better than model (iii), which resembles the underlying DGP. However, when looking at the estimated models visually (visualizations for the first 100 simulation repetitions are uploaded together with the `R` code), it is, with few exceptions, clearly observable that there might most likely be no time-varying effect. In addition, it should be noted that the problem of slight overfitting also applies for model (ii). In conclusion, the simulation study shows that the proposed functional random coefficient approach can flexibly capture heterogeneous time-variation within the covariable effects. In addition, if the time-variation is homogeneous, our proposed model does not lead to overfitting due being penalized towards the true model. Only if time-variation is fully absent, slight overfitting can be observed. However, in these cases, it is usually easy to visually recognize that there might not be any time variation in the data. Therefore, model selection should always involve a visual inspection and should not be based solely on fit measures. These results are mostly unaffected by the sample size.

5 Brain tumor case study

We apply the proposed approach to a brain tumor survival example based on data from Ceccarelli et al. (2016). This dataset includes patients with a glioma divided into five different diagnoses: anaplastic astrocytoma, astrocytoma (other), anaplastic oligodendroglioma, oligodendroglioma (other), glioblastoma and mixed glioma. After removing 37 patients due to missing values, the total number of participants is $n = 1094$, of which 593 are non-censored and 501 are censored. Survival times were recorded exact to the day and for 3 persons, who died already on their admission day, the survival time is set to half a day. We introduce an end-of-study, i.e. an administrative censoring, after 8 years because only 28 patients remain at this time. In addition to the five diagnoses, the age and sex of the patients and *fraction genome altered* (FGA) are recorded.

The FGA is commonly used in cancer research and represents the proportion of a tumor's genome that is affected by gains or losses of DNA segments, e.g. amplifications or deletions. Previous studies (see, e.g. Mehta et al., 2005) show that a higher FGA can be an indicator for aggressive tumor behavior and Dhital and Rodriguez-Bravo (2023) even state that a high FGA is an independent predictor for a reduced overall survival.

We suspect that the effect of the FGA might be time-varying and diagnosis-specific and that the time-variation might also be heterogeneous between diagnoses. Therefore, we model this effect by a functional random coefficient. This leads to the final model being a PAMM with a P-spline based baseline hazard, the diagnosis-specifically time-varying effect of FGA, a linear effect for sex and age and a fixed effect for the diagnosis. For both, the log-baseline hazard and the functional random effect, we choose cubic P-splines with first order penalty and 9 inner knots, such that each of the intervals corresponds to one year.

The estimated regression coefficients are shown in Table 2. A higher age significantly increases the hazard rate, which is an expected result as age usually increases the risk of death. In contrast, the effect of sex is not significant at a 5%-level. Compared to the anaplastic astrocytoma, which is the reference category, all other diagnoses significantly influence the hazard function. While the risk of death is increased for the glioblastoma, it is reduced for the other three diagnoses.

The non-linear baseline hazard as well as the effect of the FGA is shown in Figure 3. Both effects are significant with regard to the test of Wood (2012). The baseline hazard increases nearly linearly for the first year and a half, then decreases slightly for another year and a half, and then remains nearly constant for the next three years before increasing again. With regard to the effect of FGA, our main

Table 2: Results for the fitted model: for the linear effects the estimated coefficients (Coef.), the standard error (SE) and the p-value, resulting from a z-test, is shown. For the smooth terms the p-value corresponds to the test of Wood (2012).

Linear effects:	Coef.	SE	p-value
Age	0.0401	0.0033	< 0.0001
Sex: male	0.1481	0.0849	0.0811
Diagnosis: astrocytoma (other)	-1.4830	0.4424	0.0008
Diagnosis: mixed glioma	-0.8081	0.2617	0.0020
Diagnosis: anaplastic oligodendroglioma	-0.8842	0.2868	0.0020
Diagnosis: oligodendroglioma (other)	-1.5816	0.3347	< 0.0001
Diagnosis: glioblastoma	0.9002	0.1670	< 0.0001
Smooth terms:			p-value
log-baseline: $f(t)$			< 0.0001
FGA: $f_{\text{Diagnosis}}(t) \cdot$ FGA			0.0002

Table 3: Fit measures for the proposed functional random coefficient and its competitors including a simple linear effect. The fit is measured in terms of the log-likelihood (logLik), the integrated Brier score (IBS) and the AIC.

	logLik	IBS	AIC
Heterogeneous time-variation	-4099.36	0.1162	8238.87
Heterogeneity and time-variation	-4108.27	0.1174	8250.03
Heterogeneity only	-4111.04	0.1185	8250.50
Time-variation only	-4111.00	0.1184	8251.44
Linear effect	-4111.01	0.1184	8251.20

focus, we can indeed observe that the effect of FGA strongly varies over time as well as between the diagnoses and that the time variation is also quite different between the diagnoses. While there are no major differences in the first year and a half, the curves differ considerably thereafter. While the effect decreases over time for glioblastoma, it increases for the other four diagnoses. Hence, these effects might cancel out if the time-variation is not modeled diagnosis-specific. This outlines the practical relevance of our approach.

We compare the model fit to the competitors introduced in Section 4 as well as a model deploying a simple linear effect of FGA, which is shown in Table 3. It can be observed that modeling the time-varying effect of FGA diagnoses specifically leads to the best model fit. This also agrees with the visual impression from Figure 3. Including only random effects or only time variation does not lead to a notable increase in model fit compared to the linear effect. In addition, even the model including heterogeneity as well as time-variation only slightly increases the model fit compared to the increase that is achieved by using the functional random coefficient. This coincides with the impression from Figure 3, indicating that the time-variation might cancel out if it is not modeled diagnoses specifically. The fact that this also applies to the AIC indicates that this does not result from overfitting.

In conclusion, in this application, the use of a functional random coefficient allowed us to capture diagnosis-specific time variation in the effect of FGA on survival. In addition, the individual curves shown in Figure 3 can improve the understanding of FGA as a predictor of survival.

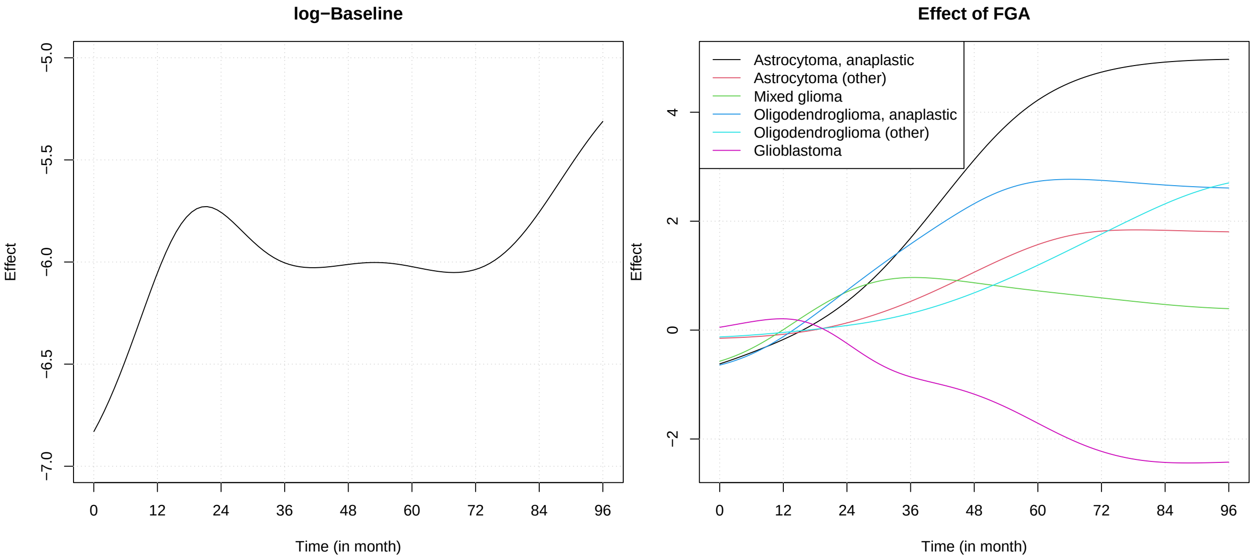


Figure 3: Estimated Smooth effects for the log-baseline function and the functional random coefficient of FGA.

6 Conclusion

In this paper, we introduced heterogeneously time-varying covariable effects to hazard regression models. This provides an appropriate model for cases in which the effect of a covariable is not only time-varying and subgroup-specific but its time-variation is subgroup specific, too.

The proposed method makes use of the existing framework of PAMMs which enables us to deploy an efficient Poisson model-based inference. Our approach allows for non-linear time-effects due to being based on penalized splines and uses an efficient random effects basis to model the heterogeneity. In addition, the penalization mostly prevents our method from overfitting in absence of heterogeneous time-variation. The corresponding simulation study only shows slight overfitting if time-effects are fully absent. However, it is easy to visually assess such cases. On the other hand, in presence of heterogeneous time-variation, the simulation study outlines the superior fit of your approach.

We apply this model to a brain tumor case study. Here, the effect of the FGA varies over time and this time-variation is highly diagnosis-specific. Therefore, modeling FGA with a diagnosis-specific time-varying effect not only greatly improves the model fit, but also prevents the effects from canceling each other out. This provides additional interpretability and may lead to a better understanding of FGA as a risk predictor. Thus, this case study outlines the practical relevance of the proposed method.

Future possible research includes introducing this type of effect to Bayesian survival models. This is mainly a matter of computational efficiency since in a Bayesian setting, for both – the piecewise exponential approach and the direct approach involving an integral over the hazard rate – estimating a FRE might lead to very high computational costs. In addition, other applications, such as use in multicenter studies, should be explored.

Supplementary Material

Supplementary material is available online.

Software and data availability

Software in the form of R code is available at https://github.com/Niklas191/heterogeneous_time-variation. The case study data set is publicly available at the cBioPortal database with IDs gbm_tcga_gdc and difg_tcga_gdc.

Funding

This work has been supported by the Research Training Group "Biostatistical Methods for High-Dimensional Data in Toxicology“ (RTG 2624, P7) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, Project Number 427806116).

Competing interests

The authors declare no competing interests.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120.

Bender, A., Groll, A., and Scheipl, F. (2018). A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, 18(3-4):299–321.

Bender, A. and Scheipl, F. (2018). pammtools: Piece-wise exponential additive mixed modeling tools. *arXiv:1806.01042 [stat]*.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723.

Ceccarelli, M., Barthel, F. P., Malta, T. M., et al. (2016). Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell*, 164(3):550–563.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Dhital, B. and Rodriguez-Bravo, V. (2023). Mechanisms of chromosomal instability (cin) tolerance in aggressive tumors: surviving the genomic chaos. *Chromosome Research*, 31(2):15.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. D. (2022). *Regression: Models, Methods and Applications*. Springer, Berlin, Heidelberg.

Friedman, M. (1982). Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, 10(1):101–113.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420):942–951.

Hagemann, N., Guhl, D., Kneib, T., Möllenhoff, K., and Steiner, W. (2024). Dynamic heterogeneity in discrete choice experiments. *Preprint available at SSRN: 4957076.*

Hennerfeind, A., Brezger, A., and Fahrmeir, L. (2006). Geoadditive survival models. *Journal of the American Statistical Association*, 101(475):1065–1075.

Hess, K. R. (1994). Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Statistics in Medicine*, 13(10):1045–1062.

Hofner, B., Hothorn, T., and Kneib, T. (2013). Variable selection and model choice in structured survival models. *Computational Statistics*, 28:1079–1101.

Hofner, B., Kneib, T., Hartl, W., and Küchenhoff, H. (2011). Building cox-type structured hazard regression models with time-varying effects. *Statistical Modelling*, 11(1):3–24.

Jachno, K., Heritier, S., and Wolfe, R. (2019). Are non-constant rates and non-proportional treatment effects accounted for in the design and analysis of randomised controlled trials? a review of current practice. *BMC Medical Research Methodology*, 19(1):103.

Kneib, T. and Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, 34(1):207–228.

Kneib, T., Klein, N., Lang, S., and Umlauf, N. (2019). Modular regression - a lego system for building structured additive distributional regression models with tensor product interactions. *TEST*, 28:1–39.

Li, H., Han, D., Hou, Y., Chen, H., and Chen, Z. (2015). Statistical inference methods for two crossing survival curves: a comparison of methods. *PLoS One*, 10(1):e0116774.

Mehta, K. R., Nakao, K., Zuraek, M. B., et al. (2005). Fractional genomic alteration detected by array-based comparative genomic hybridization independently predicts survival after hepatic resection for metastatic colorectal cancer. *Clinical Cancer Research*, 11(5):1791–1797.

Murphy, S. A. and Sen, P. K. (1991). Time-dependent coefficients in a cox-type regression model. *Stochastic Processes and their Applications*, 39(1):153–180.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56(4):1016–1022.

Therneau, T. M., Grambsch, P. M., and Pankratz, V. S. (2003). Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*, 12(1):156–175.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(1):95–114.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.

Wood, S. N. (2012). On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1):221–228.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563.

Wood, S. N., Scheipl, F., and Faraway, J. J. (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing*, 23(3):341–360.

Zucker, D. M. and Karr, A. F. (1990). Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *The Annals of Statistics*, 18(1):329–353.

Capturing heterogeneous time-variation in covariate effects in non-proportional hazard regression models

Supplementary Material

Niklas Hagemann[1,2], Thomas Kneib[3] and Kathrin Möllenhoff[1,2]

[1] Institute of Medical Statistics and Computational Biology, Faculty of Medicine, University of Cologne, Germany

[2] Division of Mathematics, Department of Mathematics and Computer Science, University of Cologne, Germany

[3] Chair of Statistics and Campus Institute Data Science, Georg August University Göttingen, Germany

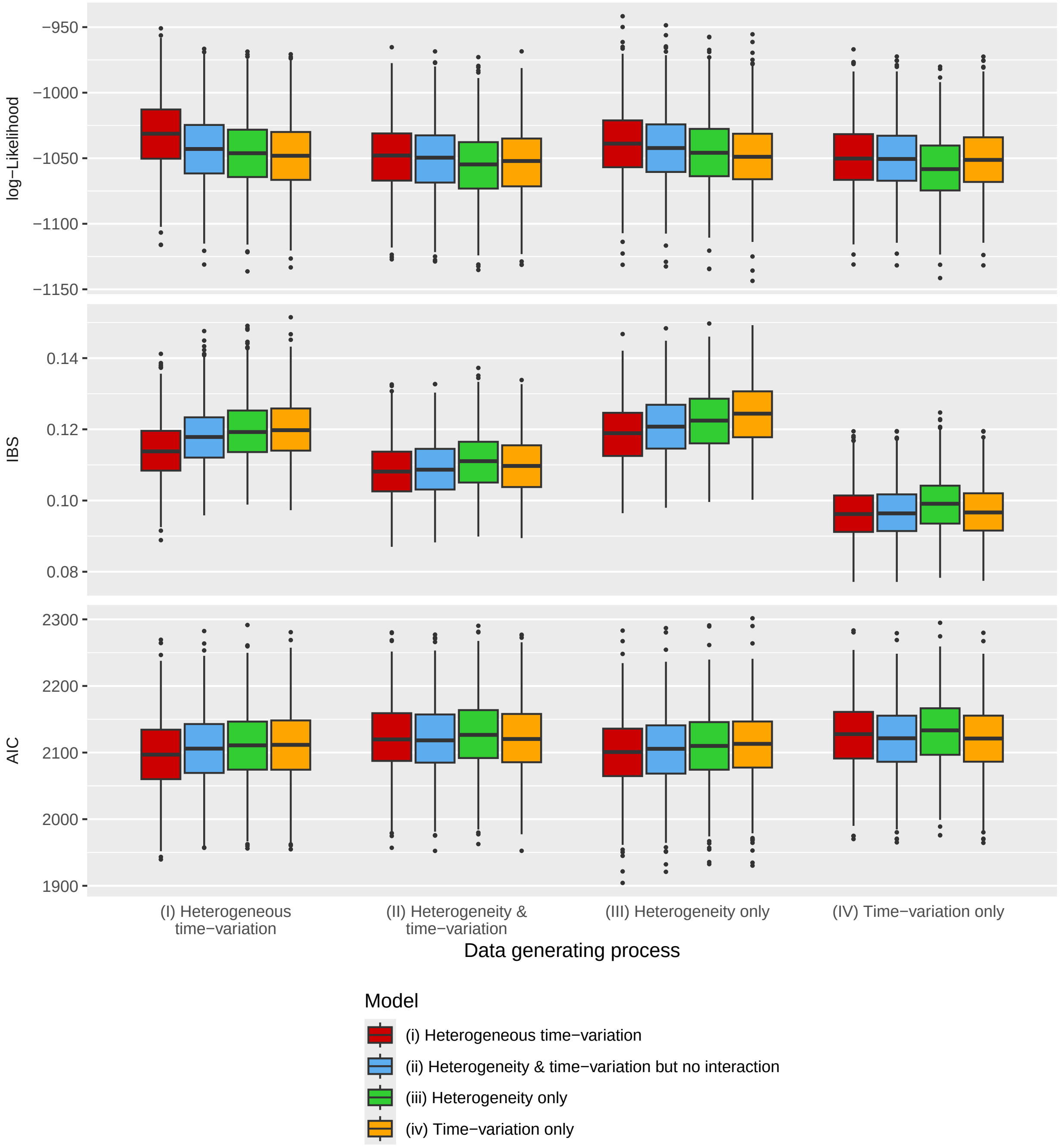


Figure S 1: Results of the simulation study with $n = 200$ in terms of the three fit measures. For each of the scenarios (I) - (IV) there is one block of consisting of four boxplots, one for each model (i) - (iv).

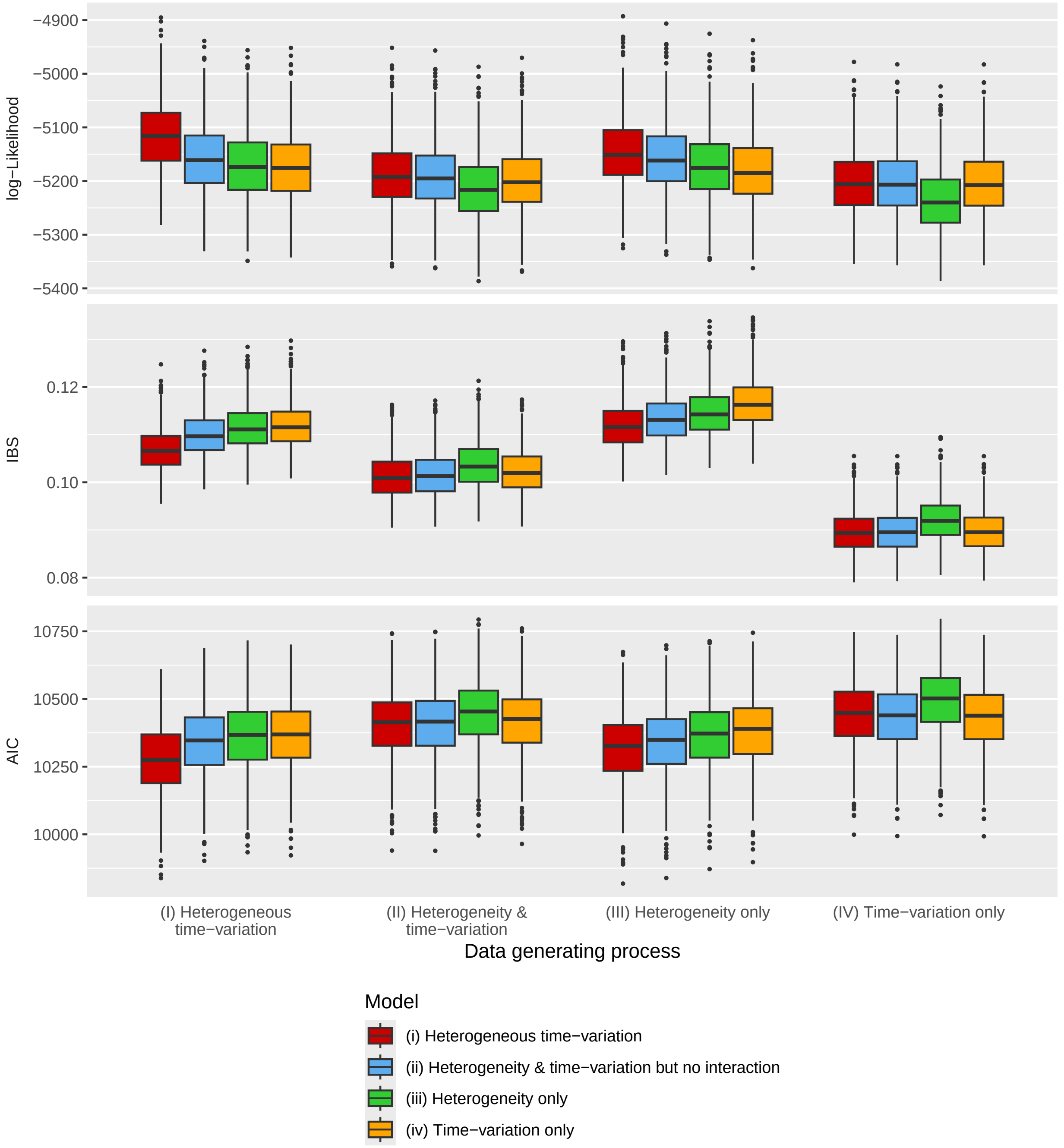


Figure S 2: Results of the simulation study with $n = 800$ in terms of the three fit measures. For each of the scenarios (I) - (IV) there is one block of consisting of four boxplots, one for each model (i) - (iv).

4 Discussion

In this thesis, four articles discuss three different aspects of multidimensionality in biostatistics. While the former two articles investigate the effects of multidimensionality on model-based equivalence tests, the latter two examine multidimensionality in hazard regression models as well as the underlying generalized additive regression models.

The first contribution focuses on multivariate dependent outcome variables in model-based equivalence tests. This is motivated by a dose-response study where the effects on efficacy and toxicity are jointly investigated. It discusses the previous work of Möllenhoff et al. (2021) and its limitations: their approach is only applicable to bivariate binary responses. In contrast, the approach suggested in this article is flexible concerning the dimensions of the outcome as well as its marginal distributions, including mixed outcomes. This method is based on generalized joint regression models as a flexible framework to model such outcomes. It deploys the Gaussian copula due to being flexible for practical modeling which is also supported by previous studies. Based on this model, a testing algorithm is developed. It is similar to the one of Möllenhoff et al. (2021) but proposes to use the maximum of maxima test statistic rather than the one based on the intersection-union principle, which was observed to be overly conservative for smaller sample sizes. This aims for an increased power for small sample sizes while retaining the asymptotic properties. In fact, the extensive simulation study shows a considerable power increase for small and medium sample sizes, exceeding in some cases by over 5-fold. The case study reanalyzes the dataset from Möllenhoff et al. (2021) but without the need to transform one of the responses from a continuous to a binary variable due to the more flexible method. In contrast to Möllenhoff et al. (2021), equivalence can be concluded for a threshold value of $\varepsilon = 0.15$ which can be reasoned by both, not losing information by avoiding the data transformation or the more powerful test. Therefore, the case study outlines the practical relevance of the proposed approach.

The second contribution discusses the issue of model uncertainty: the test of Dette et al. (2018) as well as all further developments based thereon rely on the assumption of knowing the true underlying model. However, this is usually not the case in applied research. To overcome the model uncertainty, this article proposes a flexible model averaging method which relies on the BIC. This ensures that the asymptotic properties of Dette et al. (2018) are retained. Using model averaging increases the estimation complexity leading to the problem that in many cases the testing algorithm of Dette et al. (2018) is no longer numerically feasible. Therefore, an alternative testing procedure is used that utilizes the duality of tests and confidence intervals rather than simulating the distribution under the null hypothesis and provides a numerically stable procedure. Moreover, this approach leads to additional interpretability due to the provided confidence intervals. The simulation study outlines that model misspecification can lead to either type I error inflation or a lack of power, both often to a substantial extent. Model averaging considerably reduced these problems and in many cases achieved results similar to those obtained using the true underlying model. The case study shows that this approach is essential in order to test for the equivalence of time-gene expression curves for a large number of genes. This results from the fact that in this application there is no strong prior knowledge about the underlying models and choosing the models manually would be time-consuming and could easily lead to many model misspecifications.

In the third contribution, the necessity of developing heterogeneously time-varying covariable effects for generalized additive models is discussed. This is motivated by discrete choice experiments, a popular study type in marketing research, in which respondent-specific time-variation is suspected. Such effects are captured by functional random coefficients, which are constructed as anisotropic tensor product interactions of the main effects, i.e. time-variation and heterogeneity. While random effects are used as the main effect for the heterogeneity, a penalized spline is deployed for the time effect. Bayesian estimation using Hamiltonian Monte Carlo is suggested, where the choice of priors and hyperpriors introduces regularization and thus ensures that the flexibility of the approach does not lead to overfitting. The presented case study outlines the practical relevance in terms of the presence of the suspected effect as well as an increased model fit in comparison to competing models. In addition, the WAIC implies that the increased model fit does not result from overfitting. A simulation study underlines both – the model's capability to flexibly capture heterogeneous time-variation whenever it is present and its ability to prevent overfitting in case of the absence of the supposed effect due to the penalized estimation.

In contrast to the third article, the fourth contribution considers time-to-event data and investigates the adaptation of heterogeneously time-varying covariable effects for this type of data. Therefore, the concept of functional random coefficients is transferred to hazard regression models. The proposed method makes use of the existing framework of PAMMs and hence deploys an efficient Poisson model-based inference. Unlike the third article, frequentist inference is conducted due to numerical reasons. The simulation study shows that the penalization mostly prevents overfitting in the absence of heterogeneous time-variation. Only if time effects are fully absent, which is easy to visually assess, slight overfitting can be observed. On the other hand, in the presence of heterogeneous time-variation, the simulation study outlines the superior fit of the proposed approach. The brain tumor case study outlines the practical relevance of the proposed method: the effect of FGA, one of the covariables, is modeled with a diagnosis-specific time-varying effect, leading to a considerable improvement of the model fit and preventing the effects from canceling each other out. This also provides additional interpretability and thus may allow a better understanding of FGA as a risk predictor.

In conclusion, one method is developed for each of these three different aspects of multidimensionality. The new methods introduce additional flexibility while either retaining the asymptotic properties of the model-based equivalence test or the prevention of overfitting of the regression models. In each of the four articles, a simulation study shows that the issue under consideration was successfully resolved and the four case studies outline the practical relevance of the proposed methods.

Future possible research includes the individual extensions mentioned in the four articles. For the first article, this includes implementing generalized joint regression models with more than three dimensions, investigating alternative copula options, adapting the testing procedure for less standard distributions as well as the derivation of a power formula. With regard to the second article, further research is needed in order to develop methods for other model averaging techniques, e.g. cross validation-based model averaging. Investigating the transferability of the approach to less structured applications is the most relevant extension of the third contribution. Transferring the approach to other applications, particularly multicenter studies, is also a remaining topic of the fourth article. In addition, introducing this type of effect to Bayesian survival models also merits further research. Additional topics for future research also result from the combination of the methods, e.g. the introduction of model averaging to the model-based equivalence test for multivariate responses. In addition, with regard to multidimensionality in biostatistics in general, there are also remaining research questions that are either related to another direction of multidimensionality, e.g. equivalence tests with high-dimensional explanatory variables like images, or to other statistical methods, e.g. multivariate responses in non-parametric tests.

References

Altman, D. G. and Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, 311(7003):485.

Bastian, P., Dette, H., Koletzko, L., and Möllenhoff, K. (2024). Comparing regression curves – an L^1-point of view. *Annals of the Institute of Statistical Mathematics*, 76(1):159–183.

Bender, A., Groll, A., and Scheipl, F. (2018). A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, 18(3-4):299–321.

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4):295–300.

Bretz, F., Möllenhoff, K., Dette, H., Liu, W., and Trampisch, M. (2018). Assessing the similarity of dose response and target doses in two non-overlapping subgroups. *Statistics in Medicine*, 37(5):722–738.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, 53(2):603–618.

Cade, B. S. (2011). Estimating equivalence with quantile regression. *Ecological Applications*, 21(1):281–289.

Collett, D. (2015). *Modelling Survival Data in Medical Research*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Dette, H., Möllenhoff, K., Volgushev, S., and Bretz, F. (2018). Equivalence of regression curves. *Journal of the American Statistical Association*, 113(522):711–729.

Dixon, P. M. and Pechmann, J. H. K. (2005). A statistical test to show negligible trend. *Ecology*, 86(7):1751–1756.

Duane, S., Kennedy, A., Pendleton, B. J., and Roweth, D. (1987). Hybrid monte carlo. *Physics Letters B*, 195(2):216–222.

Duda, J., Kappenberg, F., and Rahnenführer, J. (2022). Model selection characteristics when using MCP-Mod for dose-response gene expression data. *Biometrical Journal*, 64(5):883–897.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. D. (2022). *Regression: Models, Methods and Applications*. Springer.

Fang, H.-B., Fang, K.-T., and Kotz, S. (2002). The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis*, 82(1):1–16.

Filippou, P., Marra, G., and Radice, R. (2017). Penalized likelihood estimation of a trivariate additive probit model. *Biostatistics*, 18(3):569–585.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420):942–951.

Gsteiger, S., Bretz, F., and Liu, W. (2011). Simultaneous confidence bands for nonlinear regression models with application to population pharmacokinetic analyses. *Journal of Biopharmaceutical Statistics*, 21(4):708–725.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical science*, 1(3):297–310.

Hauschke, D., Steinijans, V., and Pigeot, I. (2007). *Bioequivalence Studies in Drug Development*. John Wiley & Sons, Ltd.

Hennerfeind, A., Brezger, A., and Fahrmeir, L. (2006). Geoadditive survival models. *Journal of the American Statistical Association*, 101(475):1065–1075.

Hess, K. R. (1994). Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Statistics in Medicine*, 13(10):1045–1062.

Hestenes, M. R. (1969). Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4:303–320.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(47):1593–1623.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2014). *Bayesian Survival Analysis*. John Wiley & Sons, Ltd.

Jachno, K., Heritier, S., and Wolfe, R. (2019). Are non-constant rates and non-proportional treatment effects accounted for in the design and analysis of randomised controlled trials? A review of current practice. *BMC Medical Research Methodology*, 19(1):103.

Joe, H. (2015). *Dependence Modeling with Copulas*. Chapman & Hall/CRC Press, 1 edition.

Kalbfleisch, J. and Prentice, R. (2011). *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. Wiley.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Klein, J. and Moeschberger, M. (2005). *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health. Springer.

Klein, N., Kneib, T., Marra, G., Radice, R., Rokicki, S., and McGovern, M. E. (2019). Mixed binary-continuous copula regression models with application to adverse birth outcomes. *Statistics in Medicine*, 38(3):413–436.

Kneib, T. and Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, 34(1):207–228.

Kneib, T., Klein, N., Lang, S., and Umlauf, N. (2019). Modular regression - a lego system for building structured additive distributional regression models with tensor product interactions. *TEST*, 28:1–39.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212.

Li, H., Han, D., Hou, Y., Chen, H., and Chen, Z. (2015). Statistical inference methods for two crossing survival curves: a comparison of methods. *PLoS One*, 10(1):e0116774.

Liu, W., Bretz, F., Hayter, A. J., and Wynn, H. P. (2009). Assessing nonsuperiority, noninferiority, or equivalence when comparing two regression models over a restricted covariate region. *Biometrics*, 65(4):1279–1287.

MacKenzie, D. I. and Kendall, W. L. (2002). How should detection probability be incorporated into estimates of relative abundance? *Ecology*, 83(9):2387–2393.

Mai, J.-F. and Scherer, M. (2017). *Simulating Copulas: Stochastic Models, Sampling Algorithms, and Applications*. Series In Quantitative Finance. World Scientific Publishing Company, 2 edition.

Marra, G. and Radice, R. (2017). Bivariate copula additive models for location, scale and shape. *Computational Statistics and Data Analysis*, 112:99–113.

McNeil, A. J. and Nešlehová, J. (2009). Multivariate Archimedean copulas, d-monotone functions and ℓ_1-norm symmetric distributions. *The Annals of Statistics*, 37(5B):3059–3097.

Möllenhoff, K., Binder, N., and Dette, H. (2024). Testing similarity of parametric competing risks models for identifying potentially similar pathways in healthcare. *Statistics in Medicine*, 43(28):5316–5330.

Möllenhoff, K., Bretz, F., and Dette, H. (2020). Equivalence of regression curves sharing common parameters. *Biometrics*, 76(2):518–529.

Möllenhoff, K., Dette, H., and Bretz, F. (2021). Testing for similarity of binary efficacy-toxicity responses. *Biostatistics*, 23(3):949–966.

Möllenhoff, K., Dette, H., Kotzagiorgis, E., Volgushev, S., and Collignon, O. (2018). Regulatory assessment of drug dissolution profiles comparability via maximum deviation. *Statistics in Medicine*, 37(20):2968–2981.

Möllenhoff, K., Loingeville, F., Bertrand, J., Nguyen, T. T., Sharan, S., Zhao, L., Fang, L., Sun, G., Grosser, S., Mentré, F., and Dette, H. (2022). Efficient model-based bioequivalence testing. *Biostatistics*, 23(1):314–327.

Murphy, S. A. and Sen, P. K. (1991). Time-dependent coefficients in a cox-type regression model. *Stochastic Processes and their Applications*, 39(1):153–180.

Newey, W. K. and McFadden, D. (1994). Chapter 36 large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier.

Pinheiro, J. C., Bretz, F., and Branson, M. (2006). *Analysis of Dose–Response Studies–Modeling Approaches*. Dose Finding in Drug Development. Springer, New York.

Radice, R. and Marra, G. (2016). Copula regression spline models for binary outcomes. *Statistics and Computing*, 26(5):981–995.

Rahnenführer, J., De Bin, R., Benner, A., Ambrogi, F., Lusa, L., Boulesteix, A.-L., Migliavacca, E., Binder, H., Michiels, S., Sauerbrei, W., McShane, L., and topic group High-dimensional data (TG9) of the STRATOS initiative (2023). Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges. *BMC medicine*, 21(1):182.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56(4):1016–1022.

Schorning, K., Bornkamp, B., Bretz, F., and Dette, H. (2016). Model selection versus model averaging in dose finding studies. *Statistics in Medicine*, 35(22):4021–4040.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15:657–680.

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

Sun, J. (2006). *The Statistical Analysis of Interval-censored Failure Time Data*. Statistics for Biology and Health. Springer, New York.

Therneau, T. M., Grambsch, P. M., and Pankratz, V. S. (2003). Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*, 12(1):156–175.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1):92–107.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

Wood, S. N., Scheipl, F., and Faraway, J. J. (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing*, 23(3):341–360.

Zucker, D. M. and Karr, A. F. (1990). Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *The Annals of Statistics*, 18(1):329–353.

Erklärung zur Dissertation

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen und eingebundenen Artikeln und Manuskripten - noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung des Promotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universität zu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten und der schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Teilpublikationen:

- Hagemann, N., Marra, G., Bretz, F., and Möllenhoff, K. (2024). Testing for similarity of multivariate mixed outcomes using generalized joint regression models with application to efficacy-toxicity responses. *Biometrics*, 80(30). DOI: 10.1093/biomtc/ujae077

- Hagemann, N. and Möllenhoff, K. (2025). Overcoming model uncertainty – how equivalence tests can benefit from model averaging. *Statistics in Medicine*, 44(6). DOI: 10.1002/sim.10309

- Hagemann, N., Guhl, D., Kneib, T., Möllenhoff, K., and Steiner, W. (2024). Dynamic heterogeneity in discrete choice experiments. *Preprint available at SSRN: 4957076*. DOI: 10.2139/ssrn.4957076
  Hinweis nach § 7 Abs. 7 der Promotionsordnung: Erhebliche Teile des Inhalts dieser Teilpublikation wurden im Rahmen meiner Masterarbeit erarbeitet. Während der Promotionszeit wurden diese Ergebnisse überarbeitet, um eine Simulationsstudie ergänz und zu einem Artikel zusammengefasst.

- Hagemann, N., Kneib, T., and Möllenhoff, K. (2025). Capturing heterogeneous time-variation in covariate effects in non-proportional hazard regression models. *Preprint available at arXiv: 2501.13525 [stat.ME]*. DOI: 10.48550/arXiv.2501.13525