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

Niklas Hagemann [1], Giampiero Marra [2], Frank Bretz [3] and Kathrin Möllenhoff [1]

[1] Mathematical Institute, Heinrich-Heine-University Düsseldorf, Germany

[2] Department of Statistical Science, University College London, London, UK

[3] Statistical Methodology, Novartis Pharma AG, Basel, Switzerland

---

Address for correspondence: Kathrin Möllenhoff, Mathematical Institute, Heinrich-Heine-University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany.

E-mail: kathrin.moellenhoff@hhu.de.

Phone: (+49) 211 81 12876.

Fax: (+49) 211 81 12876.

Abstract: tba

Key words: efficacy–toxicity; equivalence test; distributional copula regression;

mixed outcome; similarity test;

1 Introduction

Testing whether an effect of an explanatory variable on a response variable of interest is equivalent or similar between two groups is a common problem in applied research, especially, but not exclusively, in clinical trials (see e.g. Otto et al., 2008; Jhee et al., 2004). In this context similarity is defined as equivalence up to a threshold value, the *similarity threshold*.

The question of similarity is usually assessed by testing whether the (marginal) effects of the covariates on the response variable are similar enough between the groups, either based on confidence intervals (Liu et al., 2009; Gsteiger et al., 2011; Bretz et al., 2018) or on distance measures (Möllenhoff et al., 2018; Dette et al., 2018). All of these approaches assume a univariate continuous outcome variable. As outlined by Möllenhoff et al. (2021), this might not be appropriate for many applications. On the one hand the outcome might by measured as a binary variable or even on a categorical or ordinal scale, on the other hand, in applied research, multivariate (often bivariate) outcome variables frequently occur (e.g. Jhee et al., 2004). If these variables cannot be assumed to be independent of each other, they have to be modelled jointly.

There are different approaches to jointly model multivariate outcome variables, usually based on copulae (Sklar, 1959). Tao et al. (2013) propose to use Archimedean copulae for such models and Möllenhoff et al. (2021) suggest to use 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 suggest to use elliptical copulae, especially the Gaussian one. de Leon and Wu (2011) use the Gaussian copula for regression models with bivariate mixed outcomes and Chiu and Crump (2012) use the same copula to model bivariate binary outcomes. The Gaussian copula is highly flexible when it comes to modelling the dependence structure (Joe, 2015). It also makes it possible to combine various types of variables (e.g., continuous, binary, categorical, continuous non-negative, ordinal) (de Leon and Wu, 2011). As an example, the Gaussian copula can be used to model jointly a binary variable (with logit, probit or c-log-log link) and continuous response based on the normal, log-normal, exponential or another distribution (e.g., Klein et al., 2019).

This model class is often also referred to as *generalised joint regression models* and the corresponding inference based on penalised maximum likelihood estimation. Radice and Marra (2016) discussed the case of bivariate models with binary margins which was then generalised by Filippou et al. (2017) to trivariate binary models. Marra and Radice (2017) extended this to bivariate copula models with continuous margins and Klein et al. (2019) additionally developed models for mixed responses (binary and continuous).

Möllenhoff et al. (2021) employed, for the first time, such copula-based models to generalise the idea of a distance based similarity test from univariate continuous responses to associated bivariate binary responses, using the Gumbel model. This represented a major improvement over previous similarity tests. In contrast, the proposed approach is considerably more flexible as it allows for several types of (mixed) outcome variables. The proposal is based on the aforementioned generalised joint regression framework, with Gaussian copula. The introduced test can regarded as a generali-

sation of the approach of Möllenhoff et al. (2021) for more than two dimensions and different scales of measures as it also relies on the distance of marginal curves. However, in contrast to Möllenhoff et al. (2021), our testing procedure is not based on the intersection union principle but on a new *maximum of maxima* approach which is less conservative and leads to a higher power of the test.

To sum up, compared to the literature, the proposal has the following advantages:

- It can be applied to an arbitrary number of dimensions of the multivariate response variable, although, in practice, the number of dimensions may be limited by computational costs and/or software availability.
- It captures arbitrary scales of measures (e.g., continuous, binary, categorical, ordinal) of the outcomes.
- Compared to testing based on the intersection union principle, the proposed *maximum of maxima* testing approach leads to a higher power of the test.
- Given the modelling approach adopted here, the issue of increasing type I error with increasing sample size, as experienced in Möllenhoff et al. (2021), is not present here.

The paper is structured as follows: In Section 2, a motivating example in terms of efficacy-toxicity trials is given. Section 3 discusses the modelling framework based on the Gaussian copula. In Section 4, the hypothesis and its associated testing procedure is introduced. Type I and II error rates for the three most relevant cases (bivariate binary, bivariate continuous and bivariate mixed outcome) are studied in Section 5. Finally, an empirical application is given in Section **??**.

2 Motivating example: Efficacy and toxicity of two drugs

As described before, testing whether an effect of an explanatory variable on the response variable is similar between two groups is especially demanded in clinical trials. It is used to decide whether a drug has an equivalent effect for two groups or not (see e.g. Jhee et al., 2004) where the doses is usually the (main) explanatory variable. If this grouping is based on different treatments rather than on characteristics of the participants (e.g gender, age classes, etc.), this is also linked to the topic of bioequivalence (see e.g. Möllenhoff et al., 2022, on how to use such tests in the context of bioequivalence).

In many clinical trials there is a multivariate (mainly bivariate) response variable where not only the efficacy but also the toxicity of the drug is investigated. Efficacy and toxicity are usually correlated and, therefore, need to be modelled jointly.

Most approaches assume a continuous efficacy outcome variable and do not account for toxicity anyway. In contrast, the approach of Möllenhoff et al. (2021) has been developed specifically for bivariate binary efficacy-toxicity outcomes. In many cases one of these models might be appropriate but in general more diverse combinations of outcome variables might occur. This could for example be continuous-continuous, binary-continuous or even combinations including categorical or ordinal outcomes.

All these effects can be included within our modelling approach. Even tough our modelling approach is also able to capture more than two outcome dimensions, for our application we mainly stick to bivariate efficacy-toxicity responses to assure compa-

rability with existing approaches and because it is a practically relevant application.

3 Modelling techniques for multivariate response variables

3.1 Basics and notation

Leaving out the observation index $i = 1, ..., n$ for simplicity of the notation, our modelling approach for the univariate case be written as

$$y = m(x, \theta) + e$$

where y is the univariate response variable with $\mathcal{Y} \subseteq \mathbb{R}$ denoting its set of possible outcomes, $x \in \mathcal{X} \subseteq \mathbb{R}$ is the deterministic, explanatory variable, $m(\cdot)$ is the function modelling the effect of x on y, $\theta \in \mathbb{R}^{\dim(\theta)}$ are the parameters of this function and e is the residual with $\mathbb{E}(e) = 0$. $m(\cdot)$ could for example be a linear model, a generalised linear model (e.g. a probit model), an additive model (e.g. a spline-based model) or any other modelling approach. However, for simplicity we assume $m(\cdot)$ to be continuous and at least once continuously differentiable in x. This modelling approach can be easily generalised to more than one explanatory variable in the manner of an additive model, i.e. $y = m(x, \theta) + m_1(z_1, \vartheta_1) + ... + m_{\tilde{K}}(z_{\tilde{K}}, \vartheta_{\tilde{K}}) + e,\ \tilde{K} \in \mathbb{N}$.

As our goal is to compare the effect of the variable x onto y for two seperate groups, we get

$$y^{(1)} = m^{(1)}(x, \theta^{(1)}) + e^{(1)} \text{ and } y^{(2)} = m^{(2)}(x, \theta^{(2)}) + e^{(2)}.$$

In the multivariate case, this generalises to

$$\boldsymbol{y}^{(l)} = \boldsymbol{m}^{(l)}(x, \boldsymbol{\theta}^{(l)}) + \boldsymbol{e}^{(l)} \Leftrightarrow \begin{pmatrix} y_1^{(l)} \\ \vdots \\ y_K^{(l)} \end{pmatrix} = \begin{pmatrix} m_1^{(l)}(x, \theta_1^{(l)}) \\ \vdots \\ m_K^{(l)}(x, \theta_K^{(l)}) \end{pmatrix} + \begin{pmatrix} e_1^{(l)} \\ \vdots \\ e_K^{(l)} \end{pmatrix}, \quad l = 1, 2$$

with dimension $K \in \mathbb{N}$.

Transforming the example from Section 2 to this notation would lead to a model where

- $K = 2$,
- x is the dose,
- $m_1^{(1)}(x, \theta^{(1)})$ models the efficacy of group 1,
- $m_2^{(1)}(x, \theta^{(1)})$ models the toxicity of group 1,
- $m_1^{(2)}(x, \theta^{(2)})$ models the efficacy of group 2, and
- $m_2^{(2)}(x, \theta^{(2)})$ models the toxicity of group 2.

The models $m_l^{(k)}(\cdot)$, $l = 1, 2$, $k = 1, ..., K$ can (according to its specification) be estimated as a linear, a generalised linear, a non-linear or an additive regression model. As we do not assume the K response variables to be independent from each other, the models cannot just be estimated separately but need to be estimated jointly. This is usually done by multivariate regressions models under the assumption that all response variables come from one multivariate response distribution (see Section 3.2). A more flexible modelling approach are copula-based *generalised joint regression models* (see Section 3.3).

3.2 A rare special case: Direct modelling

An easy to handle but rare special case occurs, if response variables $y_1^{(l)}, ..., y_K^{(l)}$, $l = 1, 2$, follow one known K-dimensional distribution. This could be any multivariate distribution (e.g. the multinomial distribution or the Dirichlet distribution) but the multivariate normal distribution

$$\begin{pmatrix} y_1^{(l)} \\ \vdots \\ y_K^{(l)} \end{pmatrix} \sim N_k \left(\begin{pmatrix} \mathbb{E}(y_1^{(l)}) \\ \vdots \\ \mathbb{E}(y_K^{(l)}) \end{pmatrix}, \mathbf{\Sigma}^{(l)} \right) \quad l = 1, 2$$

with $\mathbf{\Sigma}^{(l)} = \mathbf{Cov}(y_1^{(l)}, ..., y_K^{(l)})$, $l = 1, 2$ might be the most relevant one.

Obviously this implies that all marginal distributions are normal distributions. However, this requirement can be relaxed by not including the variables themselves into the vector but latent variables (see de Leon and Wu, 2011, for an explanation of the estimation of multivariate response models including latent variables). This would imply a probit model for binary variables and a multinomial probit model for categorical variables.

Nevertheless, the assumption of multivariate normality is quite strict which is reasoned by the fact that even if $y_1^{(l)}, ..., y_K^{(l)}$, $l = 1, 2$ are all univariate normally distributed and all pairwise covariances are known, this does not necessarily imply that $\left(y_1^{(l)}, ..., y_K^{(l)}\right)^\top$ is multivariate normal distributed (Melnick and Tenenbein, 1982). The assumption of all response variables coming from one multivariate distribution becomes less restrictive by including mixture distributions (e.g. mixture of normals). A Modelling approach for multivarite responses with Gaussian Mixture distribution is introduced by Lee et al. (2012). However, the practical relevance of this approach

might still be limited to a few special cases.

In the case of such a multivariate response distribution no copula is needed and the model can be estimated directly based on the likelihood associated to the multivariate distribution (see e.g. Wood et al., 2016; Lee et al., 2012, for approaches based on penalised likelihood maximisation).

3.3 Modelling via copulae

A more flexible way to jointly model several response variables are copulae (Sklar, 1959). For a multivariate K-dimensional distribution with cummulative distribution function (cdf) $\boldsymbol{F}$ and $F_1, ..., F_K$ being the cdfs of its univariate marginals, the copula $C : [0,1]^K \rightarrow [0,1]$ associated with $\boldsymbol{F}$ is the distribution function with $U(0,1)$ marginal distributions, such that

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

is fulfilled (Joe, 2015). A comprehensive introduction to copula is for example given by Joe (2015) or Trivedi and Zimmer (2007). There are different classes of copulae, where the Archimedean (including the Gumbel, the Frank and the Clayton copula) and the meta-elliptical copulae (including the Gaussian copula) are the most commonly used ones (Joe, 2015).

The choice of the copula varies a lot between applications and a lot of work has been done on how to choose the most appropriate copula for a specific applications which is referred as *copula model choice problem* (Trivedi and Zimmer, 2007). For some specific applications, there are standard models, e.g. the Gumbel-model (Murtaugh and Fisher, 1990; Heise and Myers, 1996) as bivariate generalisation of a logit-regression

model. Also, for specific applications there is often a specific Archimedean copulae satisfying the individual requierments of this applications (Joe, 2015).

However, finding the best copula for a given modelling problem is not the focus of this paper, especially because our modelling approach should be able to be transferred to a wide range of applications and not tailored to a few specific situations. To ensure this, we will use the Gaussian copula for our application which can flexibly capture a wide range of dependencies between the response variables (Joe, 2015). Hence, using the Gaussian copula might lead to reasonable results in many applications although for a specific application it might be outperformed by another copula.

The Gaussian copula is given by

$$\begin{aligned} \boldsymbol{F}(y_1, ..., y_k) &= C_{Gaussian}(F_1(y_1), ..., F_K(y_K)) \\ &= \Phi_K(\Phi^{-1}(F_1(y_1)), ..., \Phi^{-1}(F_K(y_K)), \boldsymbol{\Gamma}) \end{aligned}$$

or in the bivariate case

$$\boldsymbol{F}(y_1, y_2) = \Phi_2(\Phi^{-1}(F_1(y_1)), \Phi^{-1}(F_2(y_2)), \rho)$$

where Φ_K is the cdf of the multivariate Gaussian distribution of dimension K, Φ^{-1} is the quantile function of the univariate Gaussian distribution, $F_1(y_1), ..., F_K(y_K)$ are the cdfs of the marginal distributions, $\boldsymbol{\Gamma} = \mathbf{Cor}(y_1, ..., y_K)$ and $\rho = \text{Cor}(y_1, y_2)$ (Joe, 2015).

Using the Gaussian copula it is not only possible to combine continuous and binary variables but also any other type of variable, including categoricaland ordinal variables (de Leon and Wu, 2011). Therefore, it is used with several types of effects, including logit, probit, c-log-log and robit for binary data and normal, log-normal, exponential and many other for continuous data (Klein et al., 2019; de Leon and Wu, 2011).

3.4 Estimation

While the estimation of multiple response regression models under the assumption of independence is relatively simple and already common in applied research, it is still quite new without this assumption. Most approaches focus on the special case described in Section 3.2, i.e. on responses coming for one multivariate distribution, including mixture distributions.

The inference for *generalised joint regression models* was developed by Radice and Marra (2016), Filippou et al. (2017), Marra and Radice (2017) and Klein et al. (2019) is implemented in the in the R-package GJRM (Marra and Radice, 2023) and relays on on penalised likelihood inference. The structure of the likelihood functions depends on whether the marginal distributions are continuous, discrete or mixed. As shown by Song et al. (2009) the likelihood for the bivariate continuous case is, leaving out the conditions for a clearer notation, given by

$$L(\theta^{(1)}, \theta^{(2)}, \rho) = \prod_{i=1}^{n} \left(c(F_1(y_{i1}), F_2(y_{i2})) \cdot f_1(y_{i1}) \cdot f_1(y_{i2}) \right)$$

where

$$\begin{aligned} c(F_1(y_{i1}), F_2(y_{i2})) &= \frac{\partial^2}{\partial y_{i1} \partial y_{i2}} C(F_1(y_{i1}), F_2(y_{i2})) \\ &= \frac{\partial^2}{\partial y_{i1} \partial y_{i2}} \Phi_2(\Phi^{-1}(F_1(y_{i1})), \Phi^{-1}(F_2(y_{i2})), \rho). \end{aligned}$$

For the bivariate discrete case it is given by

$$\begin{aligned} L(\theta^{(1)}, \theta^{(2)}, \rho) = \prod_{i=1}^{n} \Big(& \Phi_2(\Phi^{-1}(F_1(y_{i1})), \Phi^{-1}(F_2(y_{i2})), \rho) \\ & - \Phi_2(\Phi^{-1}(F_1(y_{i1})), \Phi^{-1}(F_2(y_{i2}-)), \rho) \\ & - \Phi_2(\Phi^{-1}(F_1(y_{i1}-)), \Phi^{-1}(F_2(y_{i2})), \rho) \\ & + \Phi_2(\Phi^{-1}(F_1(y_{i1}-)), \Phi^{-1}(F_2(y_{i2}-)), \rho) \Big) \end{aligned}$$

where $F_k(y_{ik}-)$ is the left-hand limit of F_k at y_{ik}. Finally, for the bivariate mixed (y_{i1} continuous and y_{i2} discrete) case the likelihood is given by

$$L(\theta^{(1)}, \theta^{(2)}, \rho) = \prod_{i=1}^{n} \left(f_1(y_{i1}) \cdot \left(\frac{\partial}{\partial F_1(y_{i1})} \Phi_2(\Phi^{-1}(F_1(y_{i1})), \Phi^{-1}(F_2(y_{i2})), \rho) \right. \right.$$
$$\left. \left. - \frac{\partial}{\partial F_1(y_{i1})} \Phi_2(\Phi^{-1}(F_1(y_{i1})), \Phi^{-1}(F_2(y_{i2}-)), \rho) \right) \right).$$

For the likelihood functions of the general case $K \geq 2$, the reader is referred to Song et al. (2009).

4 Testing for similarity of multivariate non-independent responses

4.1 Hypothesis

The hypothesis of similarity of two curves $m^{(1)}(x)$, $m^{(2)}(x)$ for the univariate case is given by

$$H_0 : \max_{x \in \mathcal{X}} |m^{(1)}(x) - m^{(2)}(x)| \geq \epsilon \text{ vs. } H_1 : \max_{x \in \mathcal{X}} |m^{(1)}(x) - m^{(2)}(x)| < \epsilon \tag{4.1}$$

where ϵ is the prespecified threshold for similarity (Möllenhoff et al., 2021; Dette et al., 2018). In other words the null hypothesis states that for at least one point the distance between the two curves is larger than the threshold. Therefore, rejecting the null hypothesis leads to the conclusion that on the given significance level the curves are similar since their distance is always lower than the threshold value.

For multivariate responses the null hypothesis can be expressed as

$$\begin{gathered} H_0 : \max_{x\in\mathcal{X}} |m_k^{(1)}(x) - m_k^{(2)}(x)| \geq \epsilon_k \text{ for at least one } k \in \{1, ..., K\} \\ \text{vs.} \\ H_1 : \max_{x\in\mathcal{X}} |m_k^{(1)}(x) - m_k^{(2)}(x)| < \epsilon_k \; \forall \; k \in \{1, ..., K\} \end{gathered} \tag{4.2}$$

where $k = 1, ..., K$ is the index for the dimension, i.e. $K = 2$ for a bivariate response, and ϵ_k is the threshold value for similarity in k-direction. This hypothesis is a generalisation of the hypothesis for the bivariate case stated by Möllenhoff et al. (2021).

For a clearer notation we define

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

such that hypothesis (4.2) can be rewritten as

$$H_0 : d_k \geq \epsilon_k \text{ for at least one } k \in \{1, ..., K\} \text{ vs. } H_1 : d_k < \epsilon_k \; \forall \; k \in \{1, ..., K\}. \tag{4.3}$$

While in the univarite case the hypotheses (4.1) can be tested directly, the global hypotheses (4.2) of the multivariate case can not be tested directly with current state-of-the-art methods (Möllenhoff et al., 2021).

Möllenhoff et al. (2021) suggest to test all K individual hypothesises

$$\begin{gathered} H_0^{(1)} : d_1 \geq \epsilon_1 \text{ vs. } H_1 : d_1 < \epsilon_1 \\ \vdots \\ H_0^{(K)} : d_K \geq \epsilon_K \text{ vs. } H_1 : d_K < \epsilon_K \end{gathered} \tag{4.4}$$

individually and to reject the global hypotheses (4.2) if all of these individual hypothesises are rejected. Due to the intersection union principle (Berger, 1982) the test decision of this testing procedure is fully equivalent to the test decision of the global hypothesis. However, testing based on the intersection union principle is extremely conservative and leads to a relatively low power of the test, especially if $K \geq 3$ (Möllenhoff et al., 2021).

Under the common assumption of $\epsilon_1 = ... = \epsilon_K$ an alternative could be to define

$$d_{max} = \max\{d_1, ..., d_K\}$$

and only test the hypothesis

$$H_0 : d_{max} \geq \epsilon \text{ vs. } H_1 : d_{max} < \epsilon. \tag{4.5}$$

This *maximum of maxima* approach is motivated by the fact that if the largest difference is small enough, all other differences are small enough, too.

4.2 Testing procedure

In order to test the hypothesis of similarity of marginal curves, we will use a parametric bootstrap test. Compared to Möllenhoff et al. (2021) we will use the less conservative maximum of maxima approach and test hypothesis (4.5) instead of testing based on the intersection union principle, i.e. hypothesis (4.4).

Testing hypothesis (4.5) does only involve the test of a single scalar difference against a single threshold value. Therefore, the testing procedure is very close to algorithm 1 of Möllenhoff et al. (2021) and also similar to the testing approach of Dette et al. (2018).

Algorithm 1.

(1) Calculate the MLE $\hat{\theta}_k^{(l)}$, $l = 1, 2$, $k = 1, ..., K$, by maximizing for each group the log-likelihood. The test statistic is obtained by

$$\hat{d}_{max} = \max\{\hat{d}_1, ..., \hat{d}_K\}$$

where

$$\hat{d}_k = \max_{x \in \mathcal{X}} |m_k^{(1)}(x, \hat{\theta}_k^{(1)}) - m_k^{(2)}(x, \hat{\theta}_k^{(2)})|, \quad k = 1, ..., K.$$

(2) Define estimators of the parameters $\theta_k^{(l)}$, $l = 1, 2$, $k = 1, ..., K$, so that the corresponding curves fulfill the null hypothesis (4.5), that is

$$\hat{\bar{\theta}}_k^{(l)} = \begin{cases} \hat{\theta}_k^{(l)} & \text{if} \quad \hat{d}_{max} \geq \epsilon \\ \bar{\theta}_k^{(l)} & \text{if} \quad \hat{d}_{max} < \epsilon \end{cases} \qquad l = 1, 2, \; k = 1, ..., K,$$

where $\bar{\theta}_k^{(l)}$, $l = 1, 2$, $k = 1, ..., K$ maximize the same objective function as $\hat{\theta}_k^{(l)}$, $l = 1, 2$, $k = 1, ..., K$ does, but under the constraint

$$d_{max} = \epsilon. \tag{4.6}$$

We discretize the range $\mathcal{X}$ of the explanatory variable to get a feasible optimization. We solve the constrained optimization problem by using the augmented Lagrangian minimization algorithm as implemented with the auglag() function in the R package alabama (Varadhan, 2022).

(3) Proceed as follows:

(i) Generate bootstrap data under the null hypothesis (**??**) by creating data specified by the parameters $\hat{\bar{\theta}}_k^{(l)}$, $l = 1, 2$, $k = 1, ..., K$.

(ii) From the bootstrap data calculate the MLE $\tilde{\theta}_k^{(l)}$ as in step (1) and the test statistic

$$\tilde{d}_{max} = \max\{\tilde{d}_1, ..., \tilde{d}_K\} \tag{4.7}$$

where

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

(iii) Repeat the steps (i) and (ii) n_{boot} times to generate replicates $\tilde{d}_{max,1}, \ldots, \tilde{d}_{max,(n_{boot})}$ of $\tilde{D}$. Let $\tilde{d}_{max,(1)} \leq \ldots \leq \tilde{d}_{max,(n_{boot})}$ denote the corresponding order statistic. The estimator of the α-quantile of the distribution of $\tilde{d}_{max}$ is defined by $\tilde{d}_{max,(\lfloor n_{boot}\alpha \rfloor)}$. Reject the null hypothesis (**??**) and decide for similarity if

$$\hat{d}_{max} < \tilde{d}_{max,(\lfloor n_{boot}\alpha \rfloor)}. \tag{4.8}$$

Alternatively, calculate the p-value $\hat{F}_{n_{boot}}(\hat{d}_{max}) = \frac{1}{n_{boot}} \sum_{i=1}^{n_{boot}} \mathbb{1}(\tilde{d}_{max,i} \leq \hat{d}_{max})$ and reject the null hypothesis (4.5) if $\hat{F}_{n_{boot}}(\hat{d}_{max}) < \alpha$ for a pre-specified significance level α, where $\hat{F}_{n_{boot}}(\hat{d}_{max})$ denotes the empirical cumulative distribution function of the bootstrap sample.

4.3 Generation of correlated data

The bootstrap step of algorithm 4.2 requires the simulation of multivariate correlated (mixed) outcomes. In general, this data can be generated using the NORTARA algorithm (*NORmal To Anything* with *Retrospective Approximation*, Chen, 2001) with one exception: The NORTARA algorithm is not suitable if all outcome variables are normally distributed. In addition to this exception, we will also not use the NORTARA algorithm for the bivariate binary case but stick to the algorithm of Emrich and Piedmonte (1991) in order to ensure comparability with existing research.

During the simulation we hold the correlation per does level constant.

4.4 Incorporating different threshold levels in the partially continuous case

In practise, unequal thresholds ϵ_k, $k = 1, ..., K$ for the K response variables might occur. If at most one variable is non-continuous, this is easy to incorporate.

Let $y_1^{(l)}$ be the non-continuous variable and $y_2^{(l)}, ..., y_K^{(l)}$ the continuous variables, each with a (potentially) different ϵ_k, $k = 1, ..., K$ and $y_k^{(l)}, k = 2, ..., K$ are measured on some scales, e.g. milligrams, litres, centimetres, etc. We now set $\epsilon = \epsilon_1$ and scale-transform all continuous variables via $\tilde{y}_k^{(l)} = \frac{\epsilon_1}{\epsilon_k} y_k^{(l)}$ and, correspondingly, their units by $\frac{\epsilon_k}{\epsilon_1}$. As the result, the transformed threshold values $\tilde{\epsilon}_k$, $k = 2, ..., K$ are all the same and equal to ϵ and the transformed variables $\tilde{y}_k^{(l)}$, $k = 2, ..., K$ are measured on the transformed scale.

If for example $K = 2$, $y_2^{(l)}$ is measured in millimetres (mm), $\epsilon_1 = 0.1$ and $\epsilon_2 = 0.02$ mm, we can transform $y_2^{(l)}$ to $\tilde{y}_2^{(l)}$ by multiplying its values with $\frac{\epsilon_1}{\epsilon_2} = \frac{0.1}{0.02} = 5$ and its units with $\frac{\epsilon_2}{\epsilon_1} = \frac{0.02}{0.1} = \frac{1}{5}$. As the result, $\epsilon = \epsilon_1 = \tilde{\epsilon}_2 = 0.1$ and $\tilde{y}_2^{(l)}$ is now measured in fifths of millimetres ($\frac{\text{mm}}{5}$).

5 Type I and Type II error simulations

The following simulation study investigates the Type I error as well as the power of our testing approach. In order to ensure comparability with existing work for the bivariate binary case, the simulation is strongly based on the simulation of Möllenhoff

et al. (2021). This applies to the dose levels being $0, 0.1, 0.2, 0.5, 1, 1.5$ and 2 leading to $g = 7$ groups with equal sample sizes $n_g^{(l)}$ as well as the data generation routine. Therefore, the data is generated for each dose level separately since the algorithms presented in Chapter 4.3 can only handle constant marginal probabilities. The data is than combined afterwords. The correlation parameter is hold constant for each dose level which leads to a different global correlation of the combined data (see Dunlap, 1937, for details of the relationship between dose-wise and global correlation). In addition, we also stick to the assumption of $\rho^{(1)} = \rho^{(2)} = \rho$ as well as the two different similarity threshold values of $\epsilon = 0.15$ and 0.2 introduced by Möllenhoff et al. (2021).

5.1 Bivariate binary outcome

In addition to using the same dose levels and data generation routine, for the bivariate binary case we also use the same parameter scenarios as Möllenhoff et al. (2021) which are given in Table 1 where $\theta^{(l)} = (\theta_1^{(l)}, \theta_2^{(l)}, \rho) = (\beta_{01}^{(l)}, \beta_{11}^{(l)}, \beta_{02}^{(l)}, \beta_{12}^{(l)}, \rho)$, $l = 1, 2$. The global correlations resulting from these parameter scenarios and three different dose-wise correlations $\rho = 0.1, 0.2$ and 0.3 are shown in Table 2.

The Type I error rates are shown in Table 3 where we observe, that for those scenarios where both outcomes are on the edge of hypothesis 4.2, i.e. $d_1 = d_2 \approx \epsilon$, the Type I error rates are below the significance level of $\alpha = 0.05$, except for the smallest group size ($n_g^{(l)} = 7$). In contrast, for those scenarios where one of marginals is under the alternative (i.e. $\min\{d_1, d_2\} = 0$), Type I error rates above $\alpha = 0.05$ occur. However, with increasing group size the Type I error rates decrease and seem to converge to 0.05. Interestingly, the correlation parameter seems not to have a strong influence on the Type I error rates.

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

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)

Table 2: Global correlations resulting from different dose-wise correlations and parameter scenarios. The values are calculated using the formulas given by Dunlap (1937).

$\theta^{(2)}$	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$
$(-1, 2, -3, 3, \rho)$	0.46	0.52	0.57
$(-1.2, 2, -3.3, 3.1, \rho)$	0.48	0.53	0.58
$(-1.5, 2.2, -3.6, 3.2, \rho)$	0.51	0.56	0.61
$(-2, 3.4, -2, 2.51, \rho)$	0.55	0.59	0.64
$(-1, 2, -2, 2.51, \rho)$	0.44	0.50	0.56
$(-2.4, 3.4, -1.8, 2.51, \rho)$	0.56	0.61	0.65
$(-1, 2, -1.8, 2.51, \rho)$	0.43	0.50	0.56

Table 3: Type I errors for the binary-binary case.

ϵ	$\theta^{(2)}$	(d_1, d_2)	$n_g^{(l)}$	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$
0.15	(-2, 3.4,-2, 2.51, ρ)	(0.15, 0.15)	7	0.057	0.051	0.058
0.15	(-2, 3.4,-2, 2.51, ρ)	(0.15, 0.15)	14	0.032	0.022	0.026
0.15	(-2, 3.4,-2, 2.51, ρ)	(0.15, 0.15)	21	0.021	0.022	0.02
0.15	(-2, 3.4,-2, 2.51, ρ)	(0.15, 0.15)	28	0.012	0.013	0.007
0.15	(-2, 3.4,-2, 2.51, ρ)	(0.15, 0.15)	50	0.013	0.010	0.008
0.15	(-1, 2,-2, 2.51, ρ)	(0, 0.15)	7	0.089	0.097	0.088
0.15	(-1, 2,-2, 2.51, ρ)	(0, 0.15)	14	0.085	0.077	0.087
0.15	(-1, 2,-2, 2.51, ρ)	(0, 0.15)	21	0.075	0.081	0.082
0.15	(-1, 2,-2, 2.51, ρ)	(0, 0.15)	28	0.062	0.068	0.088
0.15	(-1, 2,-2, 2.51, ρ)	(0, 0.15)	50	0.067	0.083	0.073
0.2	(-2.4, 3.4, -1.8, 2.51, ρ)	(0.2, 0.2)	7	0.031	0.037	0.038
0.2	(-2.4, 3.4, -1.8, 2.51, ρ)	(0.2, 0.2)	14	0.012	0.012	0.018
0.2	(-2.4, 3.4, -1.8, 2.51, ρ)	(0.2, 0.2)	21	0.013	0.006	0.012
0.2	(-2.4, 3.4, -1.8, 2.51, ρ)	(0.2, 0.2)	28	0.007	0.006	0.005
0.2	(-2.4, 3.4, -1.8, 2.51, ρ)	(0.2, 0.2)	50	0.006	0.009	0.004
0.2	(-1, 2, -1.8, 2.51, ρ)	(0, 0.2)	7	0.072	0.106	0.106
0.2	(-1, 2, -1.8, 2.51, ρ)	(0, 0.2)	14	0.084	0.100	0.089
0.2	(-1, 2, -1.8, 2.51, ρ)	(0, 0.2)	21	0.082	0.088	0.078
0.2	(-1, 2, -1.8, 2.51, ρ)	(0, 0.2)	28	0.064	0.070	0.078
0.2	(-1, 2, -1.8, 2.51, ρ)	(0, 0.2)	50	0.054	0.066	0.058

Compared to the study of Möllenhoff et al. (2021), we mostly observe larger Type I error rates. This is not surprising because our approach is less conservative compared to testing based on the intersection union principle as explained in chapter 4.1. However, as long as both outcomes are on the edge of hypothesis 4.2, the error rates are smaller then $\alpha = 5\%$ anyway. Möllenhoff et al. (2021) observed increasing Type I error rates with increasing sample size up to a value of 12.7%. This is contrary to our model where the Type I error rate decreases for increasing sample size. In addition, we do not observe values as high as 12.7%. The fact that we do not observe Type I error rates that high although we use a less conservative approach, might be an indicator that the model based on the Gaussian copula might outperform the Gumbel model. This conclusion is supported by the fact, that our Type I error rates decrease with increasing number of observations while Möllenhoff et al. (2021) observed increasing Type I error rates.

The simulated power is shown in Table 4. Our testing approach achieves reasonable power especially for sufficiently large sample sizes. The highest power of 0.919 is observed for scenario 7 with $\epsilon = 0.2$, $\rho = 0.2$ and $n_g^{(l)} = 50$. For the lower margin of $\epsilon = 0.15$ the maximum power is smaller, but still increasing with increasing sample size, reaching a maximum of 0.639 for scenario 7 with$\rho = 0.3$ and $n_g^{(l)} = 50$. For a medium sample size ($n_g^{(l)} \in \{21, 28\}$) we observe a power between 0.214 and 0.650 for $\epsilon = 0.2$ and between 0.095 and 0.384 for $\epsilon = 0.15$. For small sample size ($n_g^{(l)} \in \{7, 14\}$) our model still achieves reasonable power, reaching from 0.128 to 0.334 for $\epsilon = 0.2$ and from 0.086 and 0.204 for $\epsilon = 0.15$. Again, surprisingly the correlation parameter seems not to have a strong influence.

Compared to the results of Möllenhoff et al. (2021) we achieve a comparably high

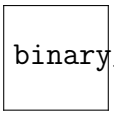

Figure 1: Power for the bivariate binary case.

power for large sample size. However, for small and medium sample sizes our approach strongly outperforms the approach of Möllenhoff et al. (2021) as we achieve a much higher (in some cases even more than 5 times higher) power. This is not surprising as our approach is less conservative compared to testing based on the intersection union principle which usually leads to a much higher power.

5.2 Bivariate continuous outcome

For the simulation of the bivariate continuous case, we use the setting stated by Bretz et al. (2018) who suggested to use a linear dose-response model

$$m_k^{(1)}(x, \theta_k^{(1)}) = \beta_{0k}^{(1)} + \beta_{1k}^{(1)}$$

for one group and a quadratic model

$$m_k^{(2)}(x, \theta_k^{(2)}) = \beta_{0k}^{(2)} + \beta_{1k}^{(2)} x + \beta_{2k}^{(2)} x^2$$

for the other group. For continuity with chapter 5.1 we transformed their model such that it fits to our dose range of $x \in [0, 2]$. This can be achieved by setting $\beta_{0k}^{(1)} = 0, \beta_{1k}^{(1)} = 1, \beta_{0k}^{(2)} = 0, \beta_{1k}^{(2)} = (1 - 2d_k)$ and $\beta_{2k}^{(2)} = d_k$, such that

$$m_k^{(1)}(x, \theta_k^{(1)}) = x,$$

and

$$m_k^{(2)}(x, \theta_k^{(2)}) = (1 - 2d_k)x + d_k x^2,$$

Table 4: Power

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

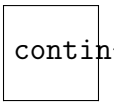

Figure 2: Curves of the two groups $l = 1, 2$ where the maximum distance $d_k = 0.2$ is observed for $x = 1$ and corresponds to the length of the arrow.

$k = 1, 2$, where d_k is the corresponding distance between the curves. This is exemplarily shown for $d_k = 0.2$ in Figure 2. The curves coincide at the boundary doses $x = 0$ and $x = 2$ and the maximum difference d_k occurs at middle dose $x = 1$. As in chapter 5.1, the same correlation for both groups, i.e. $\rho = \rho^{(1)} = \rho^{(2)}$, is assumed. Additionally, we also assume the variances of the continuous variables to be equal between responses and groups, i.e. $\sigma = \sigma_1^{(1)} = \sigma_2^{(1)} = \sigma_1^{(2)} = \sigma_2^{(2)}$.

The parameter combinations are shown in Table 5 where $\theta^{(1)} = (\beta_{01}^{(1)}, \beta_{11}^{(1)}, \beta_{02}^{(1)}, \beta_{12}^{(1)}, \sigma, \rho)$ and $\theta^{(2)} = (\beta_{01}^{(2)}, \beta_{11}^{(2)}, \beta_{21}^{(2)}, \beta_{02}^{(2)}, \beta_{12}^{(2)}, \beta_{22}^{(2)}, \sigma, \rho)$.

Table 5: Parameter scenarios used for the simulation of the bivariate continuous case.

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)

Table 6: Type I error rates for bivariate binary outcomes

				$\sigma^2 = 0.5$			$\sigma^2 = 1$			$\sigma^2 = 2$		
ϵ	θ	(d_1, d_2)	n	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$
0.2	(0, 0.6, 0.2, 0, 0.6, 0.2, σ, ρ)	(0.2, 0.2)	7	0.199	0.186	0.184	0.103	0.115	0.109	?	?	?
0.2	(0, 0.6, 0.2, 0, 0.6, 0.2, σ, ρ)	(0.2, 0.2)	14	0.118	0.124	0.131	0.099	0.089	0.096	?	?	?
0.2	(0, 0.6, 0.2, 0, 0.6, 0.2, σ, ρ)	(0.2, 0.2)	21	0.082	0.103	0.098	0.069	0.085	0.081	?	?	?
0.2	(0, 0.6, 0.2, 0, 0.6, 0.2, σ, ρ)	(0.2, 0.2)	28	0.083	0.064	0.067	0.071	0.060	0.059	?	?	?
0.2	(0, 0.6, 0.2, 0, 0.6, 0.2, σ, ρ)	(0.2, 0.2)	50	0.049	0.043	0.055	0.053	0.057	0.060	?	?	?
0.2	(0, 1, 0, 0, 0.6, 0.2, σ, ρ)	(0, 0.2)	7	0.198	0.214	0.194	0.139	0.129	0.130	?	?	?
0.2	(0, 1, 0, 0, 0.6, 0.2, σ, ρ)	(0, 0.2)	14	0.172	0.182	0.169	0.104	0.128	0.126	?	?	?
0.2	(0, 1, 0, 0, 0.6, 0.2, σ, ρ)	(0, 0.2)	21	0.142	0.133	0.135	0.109	0.109	0.091	?	?	?
0.2	(0, 1, 0, 0, 0.6, 0.2, σ, ρ)	(0, 0.2)	28	0.116	0.119	0.143	0.105	0.115	0.096	?	?	?
0.2	(0, 1, 0, 0, 0.6, 0.2, σ, ρ)	(0, 0.2)	50	0.139	0.115	0.089	0.097	0.097	0.095	?	?	?

Acknowledgements

References

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.

Chen, H. (2001). Initialization for norta: Generation of random vectors with specified marginals and correlations. *INFORMS Journal on Computing*, **13**(4), 312–331.

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**(1), 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**(2), 175–185.

Dette, H., Möllenhoff, K., Volgushev, S., and Bretz, F. (2018). Equivalence of regression curves. *Journal of the American Statistical Association*, **113**(522), 711–729.

Dunlap, J. W. (1937). Combinative properties of correlation coefficients. *The Journal of Experimental Education*, **5**(3), 286–288.

Emrich, L. J. and Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician*, **45**(4), 302–304.

Filippou, P., Marra, G., and Radice, R. (2017). Penalized likelihood estimation of a trivariate additive probit model. *Biostatistics*, **18**(3), 569–585.

Gsteiger, S., Bretz, F., and Liu, W. (2011). Simultaneous confidence bands for non-linear regression models with application to population pharmacokinetic analyses. *Journal of Biopharmaceutical Statistics*, **21**(4), 708–725.

Heise, M. A. and Myers, R. H. (1996). Optimal designs for bivariate logistic regression. *Biometrics*, **52**(2), 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**(3), 258–264.

Joe, H. (2015). *Dependence Modeling with Copulas*. Chapman & Hall/CRC Press, New York/Boca Raton, 1 edition.

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.

Lee, W., Du, Y., Sun, W., Hayes, D. N., and Liu, Y. (2012). Multiple response regression for gaussian mixture models with known labels. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **5**(6), 493–508.

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.

Marra, G. and Radice, R. (2023). *GJRM: generalised joint regression modelling*. R package version 0.2-6.4.

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

Melnick, E. L. and Tenenbein, A. (1982). Misspecifications of the normal distribution. *The American Statistician*, **36**(4), 372–373.

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., Dette, H., and Bretz, F. (2021). Testing for similarity of binary efficacy-toxicity responses. *Biostatistics*, **23**(3), 949–966.

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.

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**(6), 2003–2020.

Otto, C., Fuchs, I., Altmann, H., Klewer, M., Walter, A., Prelle, K., Vonk, R., and Fritzemeier, K.-H. (2008). Comparative Analysis of the Uterine and Mammary Gland Effects of Drospirenone and Medroxyprogesterone Acetate. *Endocrinology*, **149**(8), 3952–3959.

Radice, R. and Marra, G. (2016). Copula regression spline models for binary outcomes. *Statistics and Computing*, **26**(5), 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**(1), 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**(11), 1–6.

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

Varadhan, R. (2022). *alabama: Constrained Nonlinear Optimization*. R package version 2022.4-1.

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.