Behavior Genetics
https://doi.org/10.1007/s10519-021-10095-3

ORIGINAL RESEARCH

Multivariate Generalized Linear Models for Twin and Family Data

Wagner Hugo Bonat[1] · Jacob V. B. Hjelmborg[2]

Received: 30 March 2021 / Accepted: 30 November 2021

Abstract

Multivariate twin and family studies are one of the most important tools to assess diseases inheritance as well as to study their genetic and environment interrelationship. The multivariate analysis of twin and family data is in general based on structural equation modelling or linear mixed models that essentially decomposes sources of covariation as originally suggested by Fisher. In this paper, we propose a flexible and unified statistical modelling framework for analysing multivariate Gaussian and non-Gaussian twin and family data. The non-normality is taken into account by actually modelling the mean and variance relationship, while the covariance structure is modelled by means of a linear covariance model including the option to model the dispersion components as functions of known covariates in a regression model fashion. The marginal specification of our models allows us to extend classic models and biometric indices such as the bivariate heritability, genetic, environmental and phenotypic correlations to non-Gaussian data. We illustrate the proposed models through simulation studies and six data analyses and provide computational implementation in `R` through the package `mglm4twin`.

Keywords Estimating functions · Generalized linear models · Multivariate regression · Twin and family data

Introduction

Twin and family studies have played a critical role in understanding diseases aetiology and elucidating the nature of familial effects on specific diseases. By studying twins, siblings and their families, we can estimate how genes and environment interact to influence character, strengths, vulnerabilities and values. van Dongen et al. (2012) highlighted the importance and the wide range of applications of twin and family studies. The authors argue that novel applications of the classical twin design can provide fundamental insights into the biological mechanism underlying complex traits. In special, for cancer research twin studies can provide insights into the genetic aetiology of disease development over time and can aid in the detection of biomarker profiles for medical conditions (Boomsma et al. 2002; van Dongen et al. 2012).

The biometric analysis of twin and family data is in general based on two equivalent statistical modelling frameworks: biometric path analysis which is based on structural equation modelling (SEM) (Prescott 2004) and biometric variance component models, which correspond to a special specification of a linear mixed model (McArdle and Prescott 2005). In spite of being flexible to deal with Gaussian data, such approaches are unsuitable to deal with the combinations of binary, binomial, continuous bounded, count and asymmetric semi-continuous and continuous traits. For the special case of multiple binary traits only, the liability model is a frequent approach (Holst et al. 2016). However, in the context of multiple traits analysis it is unclear how to deal with the mix of Gaussian and binary traits using SEM or linear mixed models.

The main goal of this paper is to propose a novel and flexible statistical modelling framework to deal with multivariate Gaussian and non-Gaussian data in the context of twin and family data. The proposed model class can deal with binomial, continuous bounded, under-, equi-, over-dispersed counts, symmetric and asymmetric semi-continuous and continuous traits as well as combination of them in a unified framework.

As an extension of the standard generalized linear models (Nelder and Wedderburn 1972), the proposed framework

Edited by: Stacey S. Cherny.

✉ Wagner Hugo Bonat
wbonat@ufpr.br

[1] Department of Statistics, Paraná Federal University, Curitiba, PR, Brazil

[2] Department of Epidemiology and Biostatistics, University of Southern Denmark, Odense, Denmark

Published online: 16 January 2022

take into account non-normality by modelling the mean and variance relationship, while the mean structure is modelled by combining a link function and a linear predictor. Moreover, the covariance structure induced by the twin and family designs is modelled by means of a linear covariance model including the option to model the dispersion parameters as functions of known covariates in a regression model fashion. As we shall show in the Sect. Multivariate generalized linear models for twin and family data the linear covariance model provides an intuitive and flexible way to specify the covariance matrix in the context of twin and family studies and a suitable parametrization to assess the genetic and environment influences. Furthermore, the marginal specification of our multivariate models allows us to extend popular indices in genetic studies such as the bivariate heritability, genetic, environmental and phenotypic correlations to non-Gaussian traits. Further, classic multivariate models like the full cholesky model and its submodels, eg., the independent pathway model, the common pathway model, the latent growth curve model are easily specified in this framework and generalized so that they are expressed for a mix of normal and non-normal outcomes. Following, the strategy proposed in Bonat and Jørgensen (2016) the models are fitted by using the efficient two steps `chaser` algorithm based on the quasi-score and Pearson estimating functions, using only second-moment assumptions. As a supplementary material, we provide the `R` package `mglm4twin` which implements the models proposed in this article.

Section Multivariate generalized linear models for twin and family data presents the novel statistical modelling framework with emphasis to the modelling of the covariance structure. Estimation and inference for the proposed models are presented in Sect. Estimation and Inference. Section Simulation studies presents the results of three comprehensive simulation studies carry out to evaluate the finite sample properties of the estimating functions estimators. The application of the proposed models is illustrated in Sect. Data analyses. Finally, discussions and directions for future work are given in Sect. Discussion.

Multivariate generalized linear models for twin and family data

In this section, we shall describe a multivariate extension of the standard generalized linear models and show how to specify its covariance structure in the context of twin and family studies. The adopted approach for model specification resembles Wedderburn's quasi-likelihood (Wedderburn 1974) method and the generalized estimating equations of Liang and Zeger (1986) and relies only on second-moment assumptions.

For simplicity, let us start by discussing twin models for one trait. Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2})^\top$ be the 2×1 response vector of the ith twin pair for $i = 1, \dots, n$. Let $\mathbf{x}_i = \begin{bmatrix} 1 & x_{i11} & \dots & x_{i1k} \\ 1 & x_{i21} & \dots & x_{i2k} \end{bmatrix}$ denote a $2 \times (k+1)$ design matrix with values of k covariates to be associated with the response variable and $\boldsymbol{\beta}$ a $(k+1) \times 1$ regression parameter vector. Consider a cross-section dataset, $(\mathbf{y}_i, \mathbf{x}_i)$, where $\mathbf{y}_i's$ are independent realizations of $\mathbf{Y}_i$ according to an unspecified bivariate distribution, whose expectation and covariance matrix are given by

$$\begin{aligned} \mathrm{E}(\mathbf{Y}_i|\mathbf{x}_i) =& \mu_i = g^{-1}(\mathbf{x}_i\boldsymbol{\beta}) \\ \mathrm{var}(\mathbf{Y}_i|\mathbf{x}_i) =& \Sigma_i = \mathrm{V}(\mu_i;p)^{\frac{1}{2}}\Omega\mathrm{V}(\mu_i;p)^{\frac{1}{2}} \end{aligned} \tag{1}$$

where g is a suitable link function applied element-wise to the linear predictor $\mathbf{x}_i\boldsymbol{\beta}$. $\mathrm{V}(\mu_i;p) = \mathrm{diag}\{\vartheta(\mu_i;p)\}$ is a diagonal matrix whose main entries are given by the variance function $\vartheta(\cdot;p)$ applied element-wise to the vector μ_i. Finally p is a power parameter and Ω is a 2×2 dispersion matrix to be specified later, see subsect. Modelling the covariance matrix in biometrical genetic models for twin and family data.

The model in Eq. (1) extends the standard generalized linear models (GLMs) in two different ways. First, we introduce the extra power parameter in the variance function to bring more flexibility on the modelling of the mean and variance relationship. Second, it introduces the non-diagonal dispersion matrix $\boldsymbol{\Omega}$ to take into account the dependence induced by the twin and family designs. Thus, the model deals with non-Gaussianity by modelling the mean and variance relationship, while the data dependence is handled through the dispersion matrix and both are estimated based on observed data.

The variance function plays an important role in the context of GLMs, since different choices imply different assumptions about the response variable distribution. In this paper, we are following Bonat and Jørgensen (2016) in using three sets of variance/dispersion functions to deal with continuous, count and bounded data.

Continous data For continuous traits, we assume the power variance function $\vartheta(\mu;p) = \mu^p$, because in the univariate case it characterizes the Tweedie family of distributions, whose most important special cases are the Gaussian ($p = 0$), compound Poisson ($1 < p < 2$), gamma ($p = 2$) and inverse Gaussian ($p = 3$) distributions (Jørgensen 1987, 1997). The Tweedie family of distributions can deal with non-negative right-skewed data and can handle continuous data with probability mass at zero (Bonat and Kokonendji 2017).

Count data Similarly, the Poisson–Tweedie family of distributions is adopted to deal with count traits. The Poisson–Tweedie family of distributions is characterized by a dispersion function of the form $\vartheta(\mu;p) = \mu + \tau\mu^p$

(Jørgensen and Kokonendji 2016). In this case, the covariance matrix in Eq. (1) takes the special form $\Sigma_i = \mathrm{diag}(\mu_i) + \mathrm{V}(\mu_i;p)^{\frac{1}{2}}\Omega\mathrm{V}(\mu_i;p)^{\frac{1}{2}}$, because the dispersion parameter τ appears only in the second term. It is important to highlight that the dispersion function is not a variance function in the sense of Jørgensen (1997), however, in the context of models specified by second-moment assumptions both functions are analogous. Some of the most important special cases of the Poisson–Tweedie family are the Hermite ($p = 0$), Neyman Type A ($p = 1$), negative binomial ($p = 2$) and Poisson inverse-Gaussian ($p = 3$) distributions. Bonat et al. (2018) showed that the extended Poisson–Tweedie regression model offers a unified framework to deal with under-, equi-, overdispersed, zero-inflated and heavy-tailed count data.

Continuos bounded data and binomial data Finally, to deal with continuous bounded and binomial data Bonat et al. (2018) proposed a flexible class of regression models by including an extra power parameter on the standard binomial variance function. In this case, the extended binomial variance function takes the form $\vartheta(\mu;p) = \mu^p(1-\mu)^p$, where the power parameter p was included to bring more flexibility. Bonat et al. (2018) showed that the regression models specified using this mean and variance relationship fit well to continuous bounded data generated from the beta and simplex distributions. Furthermore, by fixing the power parameter at $p = 1$, we have the binomial mean and variance relationship.

For all aforementioned cases the power parameter plays an important role, since it is an index that distinguish between different distributions. Thus, in this paper we adapted the `chaser` algorithm presented by Bonat and Jørgensen (2016) in order to estimate the power parameter, which works as an implicit distribution selector. It is important to highlight that the strategy to use this set of mean and variance relationships was introduced in Bonat and Jørgensen (2016) in the context of multivariate covariance generalized linear models.

In order to extend the model in Eq. (1) to deal with multiple traits, let $\mathbf{Y}_{ir}$ be the 2×1 response vector of the i*th* twin pair and r*th* trait for $r = 1, \dots, R$. Similarly, let $\mathbf{x}_{ir}$ be the $2 \times (k_r + 1)$ design matrix with the values of k_r covariates and $\boldsymbol{\beta}_r$ the $(k_r + 1) \times 1$ regression parameter vector associated to the r*th* response variable. To simplify the discussion, let $\mathcal{Y}_i = (\mathbf{Y}_{i1}^\top, \dots, \mathbf{Y}_{iR}^\top)^\top$ be the $(2R \times 1)$ stacked vector of response variables. Similarly, let $\mathbf{X}_i = \mathrm{Bdiag}(\mathbf{x}_{i1}, \dots, \mathbf{x}_{iR})$ denote the $2R \times \sum_{r=1}^{R}(k_r + 1)$ design matrix and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_R^\top)^\top$ the $\sum_{r=1}^{R}(k_r + 1) \times 1$ vector of regression coefficients. Thus, we can easily specify a multivariate generalized linear model for twin and family data by

$$\begin{aligned}\mathrm{E}(\mathcal{Y}_i|\mathbf{X}_i) &= \boldsymbol{\mu}_i = \left(g_1^{-1}(\mathbf{x}_{i1}\boldsymbol{\beta}_1), \dots, g_R^{-1}(\mathbf{x}_{iR}\boldsymbol{\beta}_R)\right)\\ \mathrm{var}(\mathcal{Y}_i|\mathbf{X}_i) &= \boldsymbol{\Sigma}_i = \mathrm{V}(\boldsymbol{\mu}_i;\boldsymbol{p})^{\frac{1}{2}}\boldsymbol{\Omega}\mathrm{V}(\boldsymbol{\mu}_i;\boldsymbol{p})^{\frac{1}{2}},\end{aligned} \tag{2}$$

where $\mathrm{V}(\boldsymbol{\mu}_i;\boldsymbol{p}) = \mathrm{diag}\{\vartheta_1(\boldsymbol{\mu}_1;p_1), \dots, \vartheta_R(\boldsymbol{\mu}_R;p_R)\}$, $\boldsymbol{\Omega}$ is a $2R \times 2R$ dispersion matrix and $\boldsymbol{p} = (p_1, \dots, p_R)$ is an $R \times 1$ vector of power parameters. Note that the model allows us to specify specific link and variance functions for each response variables. Consequently, we can easily deal with the case of mixed types of response variables by the simple choice of suitable link and variance functions.

Modelling the covariance matrix in biometrical genetic models for twin and family data

In his seminal work Fisher introduced the idea to decompose the variability of a phenotype into the genetic and non-genetic components. Indeed, the genetic component is decomposed into the *additive genetic* (A) and *dominance genetic* (D) effects. Similarly, the non-genetic component is decomposed into *common environment* (C) and *unique environment* (E) effects. In general, the genetic effects are interpreted as “nature” effects while the non-genetic are interpreted as “nurture” effects. Such a decomposition is quite convenient for Gaussian data, since it can be easily modelled through the covariance matrix of a multivariate Gaussian distribution. For instance, the covariance matrix of the well-known ACDE model for twin data written as a linear covariance model has the form

$$\Omega = \tau_A \begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix} + \tau_C \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + \tau_D \begin{bmatrix} 1 & d \\ d & 1 \end{bmatrix} + \tau_E \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{3}$$

where $a = 1$ and $d = 1$ for MZ twins and $a = \frac{1}{2}$ and $d = \frac{1}{4}$ for DZ twins. For simplicity, we shall use the shorter notation

$$\Omega = \tau_A \mathrm{A} + \tau_C \mathrm{C} + \tau_D \mathrm{D} + \tau_E \mathrm{E}.$$

It is well-known that in the Gaussian case the covariance matrix does not depend on the expected values. Consequently, the expected values do not appear in Eq. (3), which implies that such a structure is suitable for dealing with continuous and symmetric data only.

Jørgensen (1997) showed that the main difference between Gaussian and non-Gaussian data is that for the latter a mean and variance relationship is present. Thus, for dealing with non-Gaussian data based on second-moment assumptions it is sufficient to model the mean and variance relationship, which can be done by choosing a suitable variance/dispersion function. Thus, we propose to combine the orthodox covariance decomposition (genetic + non-genetic effects) with suitable link and variance functions in a generalized linear models fashion. Thus, by plugging the structure in Eq. (3) in Eq. (1), we provide a very flexible statistical modelling framework to deal with Gaussian and non-Gaussian twin and family data.

Note that, from Eq. (3) is clear that both genetic and non-genetic effects do not depend on the mean structure and it is also clear that the mean and variance relationship does not introduce dependence between observations. Thus, the additional source of variability is introduced only to take the non-Gaussianity into account, however, the genetic and non-genetic structures are the ones modelling the correlation between observations as usual in the Gaussian case. We could say that in this case we have three sources of variability (non-Gaussianity, genetic and non-genetic), but only two sources of correlations (genetic and non-genetic). Such a decomposition provides a straightforward interpretation for the dispersion components $\boldsymbol{\tau} = (\tau_A, \tau_C, \tau_D, \tau_E)^\top$. Furthermore, it is trivial to extend genetic measures such as broad sense multivariate heritability, common environmentality and unique environmentality to non-Gaussian data. For instance, for the ACDE model these measures are respectively defined as

$$h^2 = \frac{\tau_A + \tau_D}{\tau_A + \tau_C + \tau_D + \tau_E}, \quad c^2 = \frac{\tau_C}{\tau_A + \tau_C + \tau_D + \tau_E} \quad \text{and}$$
$$e^2 = \frac{\tau_E}{\tau_A + \tau_C + \tau_D + \tau_E}.$$

This is a direct generalisation in line with Fisher's eminent proposal of comparing familial difference in covariance to the total variance of traits. The key point for specifying multivariate generalized linear models for twin data is the specification of the dispersion matrix $\boldsymbol{\Omega}$. Let $\boldsymbol{\nabla}_{rr'}$ denote an $R \times R$ matrix, whose entries $r = r'$ and $r' = r$ are equal to 1 and 0 elsewhere, for $r = 1, \dots, R$ and $r' \leq r$. Note that, we have $R + R(R-1)/2$ dispersion parameters for each model component. Thus, the dispersion matrix of the multivariate ACDE model for twin and family data is given by

$$\begin{aligned} \boldsymbol{\Omega} =& \tau_{A_{rr\prime}} \left\{ \boldsymbol{\nabla}_{rr\prime} \otimes \mathrm{A} \right\} + \tau_{C_{rr\prime}} \left\{ \boldsymbol{\nabla}_{rr\prime} \otimes \mathrm{C} \right\} \\ &+ \tau_{D_{rr\prime}} \left\{ \boldsymbol{\nabla}_{rr\prime} \otimes \mathrm{D} \right\} + \tau_{E_{rr\prime}} \left\{ \boldsymbol{\nabla}_{rr\prime} \otimes \mathrm{E} \right\}, \end{aligned} \tag{4}$$

where $\tau_{A_{rr\prime}}, \tau_{C_{rr\prime}}, \tau_{D_{rr\prime}}$ and $\tau_{E_{rr\prime}}$ are dispersion parameters associated with the additive genetic, common environment, dominance genetic and unique environment effects. Finally, the operator $\otimes$ denotes the Kronecker product.

For instance, the dispersion matrix associated to the bivariate ACDE model is given by

$$\begin{aligned} \boldsymbol{\Omega} =& \tau_{A_{11}} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \otimes \mathrm{A} + \tau_{A_{21}} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \otimes \mathrm{A} + \tau_{A_{22}} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \otimes \mathrm{A} \\ &+ \tau_{C_{11}} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \otimes \mathrm{C} + \tau_{C_{21}} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \otimes \mathrm{C} + \tau_{C_{22}} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \otimes \mathrm{C} \\ &+ \tau_{D_{11}} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \otimes \mathrm{D} + \tau_{D_{21}} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \otimes \mathrm{D} + \tau_{D_{22}} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \otimes \mathrm{D} \\ &+ \tau_{E_{11}} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \otimes \mathrm{E} + \tau_{E_{21}} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \otimes \mathrm{E} + \tau_{E_{22}} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \otimes \mathrm{E}. \end{aligned}$$

In the adopted parametrization the broad sense bivariate heritability, common environmentality and unique environmentality are easily obtained by

$$\begin{aligned} h_{rr'} &= \frac{\tau_{A_{rr'}} + \tau_{D_{rr'}}}{\tau_{A_{rr'}} + \tau_{C_{rr'}} + \tau_{E_{rr'}}}, \\ c_{rr'} &= \frac{\tau_{C_{rr'}}}{\tau_{A_{rr'}} + \tau_{C_{rr'}} + \tau_{E_{rr'}}} \quad \text{and} \\ e_{rr'} &= \frac{\tau_{E_{rr'}}}{\tau_{A_{rr'}} + \tau_{C_{rr'}} + \tau_{E_{rr'}}}. \end{aligned}$$

Similarly, the genetic, common environmental and unique environmental correlations are given by

$$r_{G_{rr'}} = \frac{\tau_{A_{rr'}} + \tau_{D_{rr'}}}{\sqrt{\tau_{A_{rr}} + \tau_{D_{rr}}}\sqrt{\tau_{A_{r'r'}} + \tau_{D_{r'r'}}}},$$
$$r_{C_{rr'}} = \frac{\tau_{C_{rr'}}}{\sqrt{\tau_{C_{rr}}}\sqrt{\tau_{C_{r'r'}}}} \quad \text{and} \quad r_{E_{rr'}} = \frac{\tau_{E_{rr'}}}{\sqrt{\tau_{E_{rr}}}\sqrt{\tau_{E_{r'r'}}}}.$$

Finally, the phenotypic correlation is given by

$$r_{P_{rr'}} = \frac{\tau_{Prr'}}{\sqrt{\tau_{Prr}}\sqrt{\tau_{Pr'r'}}},$$

where $\tau_{Prr'} = \tau_{Arr'} + \tau_{Crr'} + \tau_{Drr'} + \tau_{Err'}$.

Estimation and inference for all the aforementioned measures are easily obtained by the delta method, since they are estimated as simple functions of the dispersion parameter estimates.

We have focused on the ACDE model for twin data because of its popularity, although of some proposes Ozaki et al. (2011) it is well-known that the full ACDE model is challenging for estimation and inference. Thus, from now one, we consider the special cases ACE, ADE, AE, CE and E models obtained by dropping the respective terms of Eq. (4). Furthermore, many other models for family data can be specified in a similar way. Rabe-Hesketh et al. (2008) described the A and C matrices associated to the *siblings plus cousins* and *nuclear family* models. In order to represent the dispersion structure compactly, they considered two pairs of siblings, both sharing the same grandparents. The additive genetic and common environmental effects are, respectively

$$A = \begin{bmatrix} 1 & 1/2 & 1/8 & 1/8 \\ 1/2 & 1 & 1/8 & 1/8 \\ 1/8 & 1/8 & 1 & 1/2 \\ 1/8 & 1/8 & 1/2 & 1 \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

In a similar way, in the nuclear family case the additive genetic and common environmental effects (for mother, father, child 1 and child 2) are, respectively,

$$A = \begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1/2 & 1 \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

We refer to Khoury et al. (1993) for details and further examples. Note that, these structures are easily extended to multivariate traits using the same strategy employed for the ACDE model, see Eq. (4).

Modelling the dispersion parameters and parametrization cautions

An additional flexibility of the proposed model is the possibility to model any of the dispersion parameter components as a function of known covariates in a linear regression model fashion. For further exploration of this point, consider for instance the one response variable ACE model in Eq. (3). Suppose without loss of generality that is of interest to model the component τ_A as a linear combination of a vector of known covariates. Let z_{ijq} denote the value of the q*th* covariate associated with the *j*th twin of the *i*th twin pair, for $q = 1, \dots, q$, $j = 1, 2$ and $i = 1, \dots, n$. Let z_i be a $(2 \times q)$ design matrix with values of $q - 1$ known covariates associated to the i*th* twin pair, i.e.

$$z_i = \begin{bmatrix} 1 & z_{i11} & \cdots & z_{i1q} \\ 1 & z_{i21} & \cdots & z_{i2q} \end{bmatrix}.$$

Similarly, let $\boldsymbol{\tau_A} = (\tau_A(0), \tau_A(1), \dots, \tau_A(q))$ denote the $(q \times 1)$ vector of dispersion parameters. Thus, the dispersion components associated to the additive genetic effect are given by

$$\tau_{A(0)} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \circ \mathrm{A} + \tau_{A(1)} \begin{bmatrix} z_{i11} \\ z_{i21} \end{bmatrix} \circ \mathrm{A} + \dots \tau_{A(q)} \begin{bmatrix} z_{i1q} \\ z_{i2q} \end{bmatrix} \circ \mathrm{A}, \qquad (5)$$

where $\circ$ denotes the Hadamard product. It is important to highlight that all dispersion parameters can be modelled as in Eq. (5) and the model remains a linear covariance model and consequently easy to specify and interpret. We note by pass that the values of the covariates are not required to be the same for both twins. Furthermore, as we shall discuss in Sect. Estimation and Inference the linear covariance structure allows us to obtain a very simple and unified fitting algorithm.

In all described models, we directly model the covariance matrix of an unspecified multivariate distribution. Thus, the unique restriction we impose to obtain a valid regression model is that the matrix $\boldsymbol{\Sigma}_i$ is positive definite. The parameter space of the dispersion parameters is the set $\Theta = \{\boldsymbol{\tau} : \boldsymbol{\Sigma}_i > 0\}$, where $\boldsymbol{\Sigma}_i > 0$ means that $\boldsymbol{\Sigma}_i$ is a positive definite matrix. Furthermore, recall that a necessary condition for a matrix be positive definite is that the elements of its diagonal are larger than zero. It implies that the dispersion parameter estimates associated to the additive genetic, dominance genetic and common environmental effects at least at some extend can be negative and consequently have no practical interpretation. Of course, these negative values for the dispersion components will impact the computation of all described indices, such as bivariate heritability, genetic and common-environmental correlations that in this case will show negative or larger than |1| values without any practical interpretation.

On the other hand, the proposed parametrization is quite easy to specify and interpret. Moreover, in the context of genetic models, we are frequently interested in the assessment of the genetic and common environmental effects for their significance. In general, such a goal can be formulated as a hypothesis test of the form $H_0 : \tau = 0$ against $H_1 : \tau \neq 0$, where τ can represent any of the dispersion parameters. In our parametrization such a hypothesis test is well-defined, since the null hypothesis is not placed on the boundary of the parameter space and consequently the simple Wald or score tests can be used, without any extra adjustment. Note that, negative values for the dispersion components will appear frequently when the associated effects are not significantly different from zero, i.e. $\tau = 0$. In this case, we have $\mathrm{E}(\hat{\tau}) = 0$, where $\hat{\tau}$ denotes an unbiased estimator of τ. However, in practical data analysis it is perfectly possible to observe realized values of $\hat{\tau}$ smaller than zero which can be due to sample variation, as usual when testing regression coefficients in the mean structure. In the Sect. Simulation studies, as part of our simulation study, we assess the empirical distribution of $\hat{\tau}$ and show that under the null hypothesis $\tau = 0$ a simple Wald test can be used to test the significance of the dispersion coefficients.

Estimation and inference

In this section we shall review and adapted the general fitting algorithm presented in Bonat and Jørgensen (2016) for estimation and inference for models specified by second-moments assumptions. The model presented in Eq. (2) is completely specified by two vector of parameters, thus let $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\lambda}^\top)^\top$. In this notation, $\boldsymbol{\beta}$ denotes a $K \times 1$ vector of all regression coefficients involved in the mean regression structure. Similarly, $\boldsymbol{\lambda}$ denotes a $Q \times 1$ vector of all power and dispersion parameters involved in the covariance structure.

For the estimation of the regression parameters, we adopt the quasi-score function,

$$\psi_{\beta}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \sum_{i=1}^{n} \boldsymbol{D}_i^\top \boldsymbol{\Sigma}_i^{-1} (\mathcal{Y}_i - \boldsymbol{\mu}_i)$$

where $\boldsymbol{D}_i = \nabla_{\beta} \boldsymbol{\mu}_i$ is a $2R \times K$ matrix and ∇_{β} denotes the gradient operator. In the context of estimating functions the sensitivity and variability matrices are two important ingredients. The $K \times K$ sensitivity and variability matrices of $\psi_{\beta}(\boldsymbol{\beta}, \boldsymbol{\lambda})$ are respectively given by

$$\mathrm{S}_{\beta} = \mathrm{E}(\nabla_{\beta}\psi_{\beta}) = -\sum_{i=1}^{n} \boldsymbol{D}_i^{\top}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{D}_i \quad \text{and} \tag{6}$$

$$\mathrm{V}_{\beta} = \mathrm{Var}(\psi_{\beta}) = \sum_{i=1}^{n} \boldsymbol{D}_i^{\top}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{D}_i. \tag{7}$$

Similarly, for the estimation of the power and dispersion parameters, we adopt the Pearson estimating function, defined by the components

$$\psi_{\lambda_q}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \sum_{i=1}^{n} \mathrm{tr}(W_{\lambda_q}(\boldsymbol{r}_i^{\top}\boldsymbol{r}_i - \boldsymbol{\Sigma}_i)) \quad \text{for} \quad q = 1, \ldots, Q, \tag{8}$$

where $\boldsymbol{r}_i = (\mathcal{Y}_i - \boldsymbol{\mu}_i)$ and $W_{\lambda_q} = -\partial\boldsymbol{\Sigma}_i^{-1}/\partial\lambda_q$ are the Crowder's weights to be further discussed later.

The entry (q, q') of the $Q \times Q$ sensitivity matrix of $\psi_{\boldsymbol{\lambda}}$ is given by,

$$\mathrm{S}_{\lambda_{qq'}} = \mathrm{E}\left(\frac{\partial}{\partial\lambda_q}\psi_{\lambda_{q'}}\right) = -\sum_{i=1}^{n} \mathrm{tr}\left(W_{\lambda_q}\boldsymbol{\Sigma}_i W_{\lambda_{q'}}\boldsymbol{\Sigma}_i\right). \tag{9}$$

The entry (q, q') of the $Q \times Q$ variability matrix of $\psi_{\boldsymbol{\lambda}}$ is given by

$$\begin{aligned} \mathrm{V}_{\lambda_{qq'}} &= \mathrm{Cov}(\psi_{\lambda_q}, \psi_{\lambda_{q'}}) \\ &= \sum_{i=1}^{n} \left\{ 2\mathrm{tr}(W_{\lambda_q}\boldsymbol{\Sigma}_i W_{\lambda_{q'}}\boldsymbol{\Sigma}_i) + \sum_{l=1}^{2R} k_l^{(4)}(W_{\lambda_q})_{ll}(W_{\lambda_{q'}})_{ll} \right\} \end{aligned} \tag{10}$$

where $k_l^{(4)}$ denotes the fourth cumulant of $\mathcal{Y}_i$.

To take into account the covariance between the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, Bonat and Jørgensen (2016) provided expressions for the cross-sensitivity $\mathrm{S}_{\lambda\beta}$ and $\mathrm{S}_{\beta\lambda}$ as well as for the cross-variability $\mathrm{V}_{\lambda\beta}$ matrices. Thus, the joint sensitivity and variability matrices of ψ_{β} and ψ_{λ} are respectively given by

$$\mathrm{S}_{\theta} = \begin{pmatrix} \mathrm{S}_{\beta} & \mathrm{S}_{\beta\lambda} \\ \mathrm{S}_{\lambda\beta} & \mathrm{S}_{\lambda} \end{pmatrix} \quad \text{and} \quad \mathrm{V}_{\theta} = \begin{pmatrix} \mathrm{V}_{\beta} & \mathrm{V}_{\lambda\beta}^{\top} \\ \mathrm{V}_{\lambda\beta} & \mathrm{V}_{\lambda} \end{pmatrix}.$$

Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^{\top}, \hat{\boldsymbol{\lambda}}^{\top})^{\top}$ denote the estimating function estimator of $\boldsymbol{\theta}$. The asymptotic distribution of $\hat{\boldsymbol{\theta}}$ is

$$\hat{\boldsymbol{\theta}} \sim \mathrm{N}(\boldsymbol{\theta}, \mathrm{J}_{\boldsymbol{\theta}}^{-1})$$

where $\mathrm{J}_{\boldsymbol{\theta}}^{-1}$ is the inverse of Godambe information matrix,

$$\mathrm{J}_{\boldsymbol{\theta}}^{-1} = \mathrm{S}_{\boldsymbol{\theta}}^{-1}\mathrm{V}_{\boldsymbol{\theta}}\mathrm{S}_{\boldsymbol{\theta}}^{-\top},$$

where $\mathrm{S}_{\boldsymbol{\theta}}^{-\top} = (\mathrm{S}_{\boldsymbol{\theta}}^{-1})^{\top}$.

In order to solve the system of equations $\psi_{\beta} = \boldsymbol{0}$ and $\psi_{\lambda} = \boldsymbol{0}$ Jørgensen and Knudsen (2004), Bonat and Jørgensen (2016) proposed the chaser algorithm, defined by

$$\begin{aligned} \boldsymbol{\beta}^{(i+1)} &= \boldsymbol{\beta}^{(i)} - \mathrm{S}_{\beta}^{-1}\psi_{\beta}(\boldsymbol{\beta}^{(i)}, \boldsymbol{\lambda}^{(i)}) \\ \boldsymbol{\lambda}^{(i+1)} &= \boldsymbol{\lambda}^{(i)} - \alpha\mathrm{S}_{\lambda}^{-1}\psi_{\lambda}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)}), \end{aligned} \tag{11}$$

where α is an extra tuning constant included to control the step length.

The algorithm afore described has been implemented in `R` through the `mcglm` package (Bonat et al. 2018) for the special class of the multivariate covariance generalized linear models (`McGLM`). In the `McGLM` class of models, the joint covariance structure for the multiple response variables is specified using the so called generalized Kronecker product. On the other hand, for twin and family models, as proposed in this paper, the covariance structure proposed in Bonat and Jørgensen (2016) is restrictive, since we have only one parameter to model the correlation between the multiple traits. Consequently, the `McGLM` class does not allow us to decompose the correlation between traits in genetic, common-environmental and unique environmental components and consequently the algorithm should be adapted.

The main difference in terms of model specification between the models proposed in this paper and the `McGLM` class is the specification of the joint covariance matrix. Such difference impact on the computation of the Crowder's weights appearing in the Pearson estimating functions as well as in its variability and sensitivity matrices. By using standard matrix calculus, the Crowder's weights are given by the components

$$W_{\lambda_q} = -\frac{\partial\boldsymbol{\Sigma}_i^{-1}}{\partial\lambda_q} = \boldsymbol{\Sigma}_i^{-1}\frac{\partial\boldsymbol{\Sigma}_i}{\partial\lambda_q}\boldsymbol{\Sigma}_i^{-1}, \tag{12}$$

where the derivatives of $\boldsymbol{\Sigma}_i$ with respect to the power and dispersion parameters are respectively given by

$$\frac{\partial\boldsymbol{\Sigma}_i}{\partial\boldsymbol{p}_q} = \frac{\partial\mathrm{V}(\boldsymbol{\mu}_i;\boldsymbol{p})^{\frac{1}{2}}}{\partial\boldsymbol{p}_q}\boldsymbol{\Omega}\mathrm{V}(\boldsymbol{\mu}_i;\boldsymbol{p})^{\frac{1}{2}} + \mathrm{V}(\boldsymbol{\mu}_i;\boldsymbol{p})^{\frac{1}{2}}\boldsymbol{\Omega}\frac{\partial\mathrm{V}(\boldsymbol{\mu}_i;\boldsymbol{p})^{\frac{1}{2}}}{\partial\boldsymbol{p}_q}, \tag{13}$$

and

$$\frac{\partial\boldsymbol{\Sigma}_i}{\partial\boldsymbol{\tau}_q} = \mathrm{V}(\boldsymbol{\mu}_i;\boldsymbol{p})^{\frac{1}{2}}\frac{\boldsymbol{\Omega}}{\partial\boldsymbol{\tau}_q}\mathrm{V}(\boldsymbol{\mu}_i;\boldsymbol{p})^{\frac{1}{2}}. \tag{14}$$

The derivatives in Eqs. (13) and (14) depend on the derivative of the variance function and dispersion matrix respectively, and it should be evaluated accordingly. However, note that given the linear covariance structure of the dispersion matrix such derivatives are easily obtained.

The `R` package `mglm4twin` implements the described algorithm and dispersion matrix components for a variety of twin and family models as well as facilities for the estimation and inference for all measures of interest described in subsect. Modelling the covariance matrix in biometrical genetic models for twin and family data. The `mglm4twin` package is freely available at `github` as a supplementary material of this article. For details and completely reproducible examples, see http://leg.ufpr.br/~wagner/mglm4twin/.

Simulation studies

In this section, we present the main results of three comprehensive simulation studies that were conducted to investigate the finite sample properties of the estimating functions estimators. For all simulation studies, we simulated 1000 data sets considering four sample sizes (twin pairs) (100, 250, 500 and 1000) and evaluated the bias and consistency of the estimating function estimators. The `R` package `NORTARA` that implements the `NORTA` (Normal To Anything) algorithm was employed for simulating the data sets.

The first simulation study regards the simplest case of our model, i.e. the one trait case. Based on the ACE model, we obtained five simulation scenarios by fixing the dispersion parameters at the values:

- Scenario 1—$\tau_A = 0, \tau_C = 0, \tau_E = 1$,
- Scenario 2—$\tau_A = 0.75, \tau_C = 0, \tau_E = 0.25$.
- Scenario 3—$\tau_A = 0, \tau_C = 0.75, \tau_E = 0.25$,
- Scenario 4—$\tau_A = 0.5, \tau_C = 0.25, \tau_E = 0.25$
- Scenario 5—$\tau_A = 0.25, \tau_C = 0.25, \tau_E = 0.5$.

Furthermore, we adopt four marginal distributions for the trait: binomial, gamma, Gaussian and Poisson. In the binomial case, we specified the logit link function with the parameter $\beta_0 = 0$. Similarly, for the gamma and Poisson cases, we used the logarithm link function and $\beta_0 = \log(10)$. Finally, in the Gaussian case, we used the identity link function with $\beta_0 = 10$. Figure 1 presents the average bias plus and minus the average standard errors for the parameters under each scenario. The scales are standardized for each parameter by dividing the average bias and the limits of the confidence intervals by the standard errors obtained for the sample of size 100.

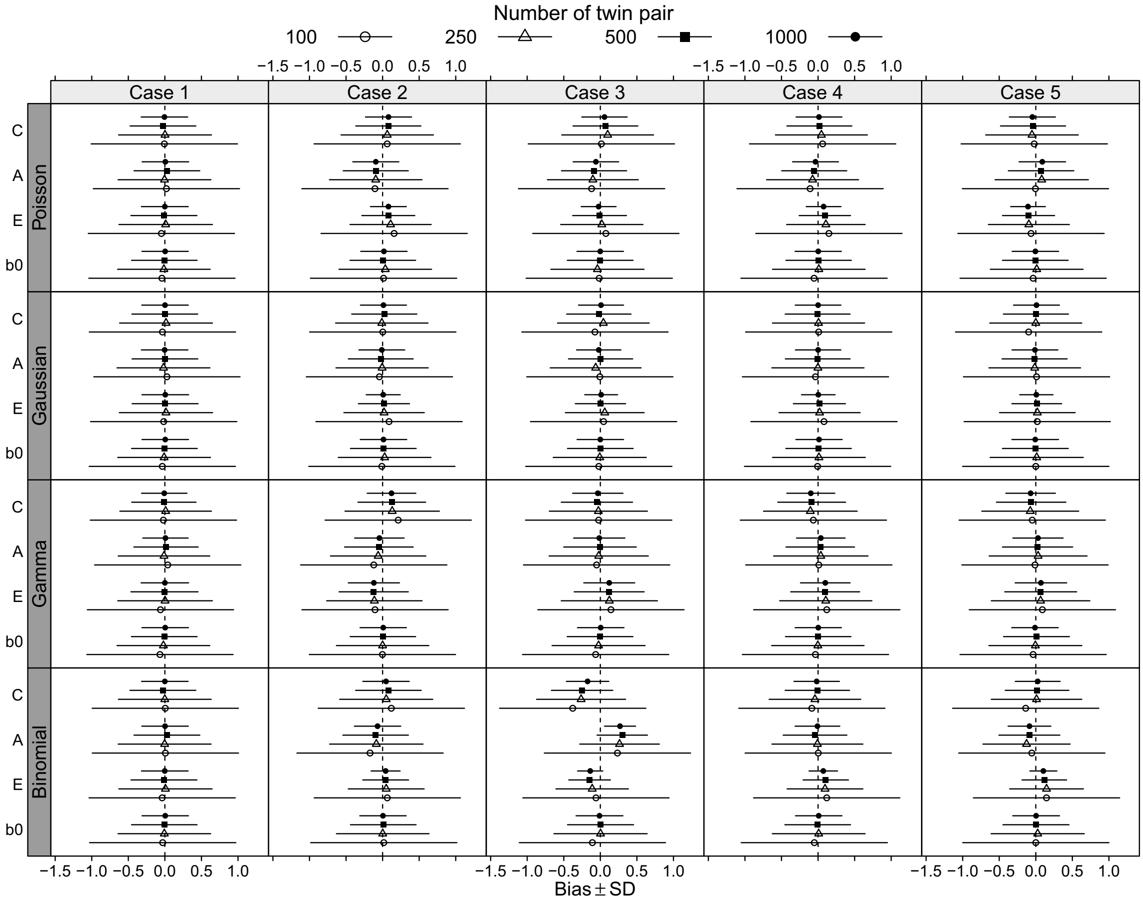

Fig. 1 Average bias and confidence intervals on a standardized scale by sample size and simulation scenario—one trait case

The results in Fig. 1 show that for the simulation scenario 1, where the genetic and common environment effects are not present our fitting algorithm provides unbiased estimates for all marginal distributions considered even for small sample sizes. In general, for the scenarios 2 to 5 the predominant effect is slightly underestimated, while the others effects are slightly overestimated. Such a pattern is markable for the binomial case, where the largest biases appear mainly in the scenarios 2 and 3. However, overall the bias and the confidence limits tend to decrease while the sample size increases, which in turn suggests the consistency of our estimators.

In the second simulation study we considered a bivariate ACE model. We considered four configurations for the dispersion parameters and combine them with four marginal distributions for the response vector, for instance binomial, gamma, Gaussian and Poisson resulting in 16 simulation scenarios. The dispersion parameters were fixed at the values:

- Scenario 1—$\tau_{E1} = 1, \tau_{E2} = 1, \tau_{E12} = 0.75, \tau_{A1} = 0, \tau_{A2} = 0, \tau_{A12} = 0, \tau_{C1} = 0, \tau_{C2} = 0, \tau_{C12} = 0.$
- Scenario 2—$\tau_{E1} = 0.25, \tau_{E2} = 0.5, \tau_{E12} = 0.25, \tau_{A1} = 0.75, \tau_{A2} = 0.5, \tau_{A12} = 0.5, \tau_{C1} = 0, \tau_{C2} = 0, \tau_{C12} = 0$
- Scenario 3—$\tau_{E1} = 0.25, \tau_{E2} = 0.5, \tau_{E12} = 0.25$,,
- Scenario 4—$\tau_{E1} = 0.25, \tau_{E2} = 0.25, \tau_{E12} = 0.1, \tau_{A1} = 0.5, \tau_{A2} = 0.25, \tau_{A12} = 0.25, \tau_{C1} = 0.25, \tau_{C2} = 0.50, \tau_{C12} = 0.25.$

The parameter values were specified in order to explore a wide range of values for the unique environment, genetic and common environment correlation as well as to have special cases as the E model in the scenario 1, the AE and CE models in the scenarios 2 and 3 and the ACE model in the scenario 4. The marginal expectation for each scenario was specified as in the first simulation study. Figure 2 presents the average bias plus and minus the average standard errors for the parameters under each scenario. The scales are standardized as in the simulation study 1.

The results in Fig. 2 show that for all simulation scenarios both the average bias and standard errors tend to 0 as the sample size is increased. Exceptions appear mainly in the scenario 4 of the gamma marginal distribution for the parameters τ_{E12} and τ_{A12} where the bias is going to zero, but slower than in the other simulation scenarios. Overall the results are promising, however the gamma and binomial cases are more challenged than the Gaussian and Poisson cases for estimation and inference, mainly when the additive genetic and common environment effects are present in the model (scenario 4).

Finally, in the third simulation study we considered a mixed types of outcomes scenario where a combination of binomial, Gaussian and Poisson traits is observed. The dispersion parameters were fixed at the values:

- Unique environment components: $\tau_{E1} = 0.15, \tau_{E2} = 0.20, \tau_{E3} = 0.50, \tau_{E12} = 0.125, \tau_{E13} = 0, \tau_{E23} = 0.10.$
- Additive genetic components: $\tau_{A1} = 0.70, \tau_{A2} = 0.60, \tau_{A3} = 0.25, \tau_{A12} = 0.50, \tau_{A13} = 0.25, \tau_{A23} = 0.10.$
- Common environment components: $\tau_{C1} = 0.70, \tau_{C2} = 0.60, \tau_{C3} = 0.25, \tau_{C12} = 0.50, \tau_{C13} = 0.25, \tau_{C23} = 0.10.$

In order to further explore the properties of our estimators, we change the order of the marginal distributions to allow for different combinations of the unique environmental, genetic and common environmental correlations between the different traits. Thus, we have six different scenarios as follow:

- Scenario 1—Binomial, Gaussian and Poisson.
- Scenario 2—Binomial, Poisson and Gaussian.
- Scenario 3—Gaussian, binomial and Poisson.
- Scenario 4—Gaussian, Poisson and binomial.
- Scenario 5—Poisson, binomial and Gaussian.
- Scenario 6—Poisson, Gaussian and binomial.

The results presented in Figs. 3 and 4 suggest that for the most simulation scenarios our estimators are unbiased and consistent. The estimates of the parameter τ_{E12} show a small bias even for large samples, mainly in the scenarios 1 to 3, but overall the relative biases are small in its magnitude.

Data analyses

In this section, we illustrate the application of multivariate generalized linear models through the analysis of six datasets. It is important to highlight that our main goal is to show how our approach adapts to different types of twin data and we do not provide an in deep analysis of each data set.

Dataset 1: dichotomous data

The first dataset concerns psychiatric disorders in 1030 (440 DZ and 590 MZ) Caucasian female twin-pairs sampled from the Virginia Twin Registry. Lifetime psychiatric illness is a binary trait and was diagnosed using an adapted version of the Structured Clinical Interview for DSM-II-R Diagnosis. The dataset was analysed by Neale

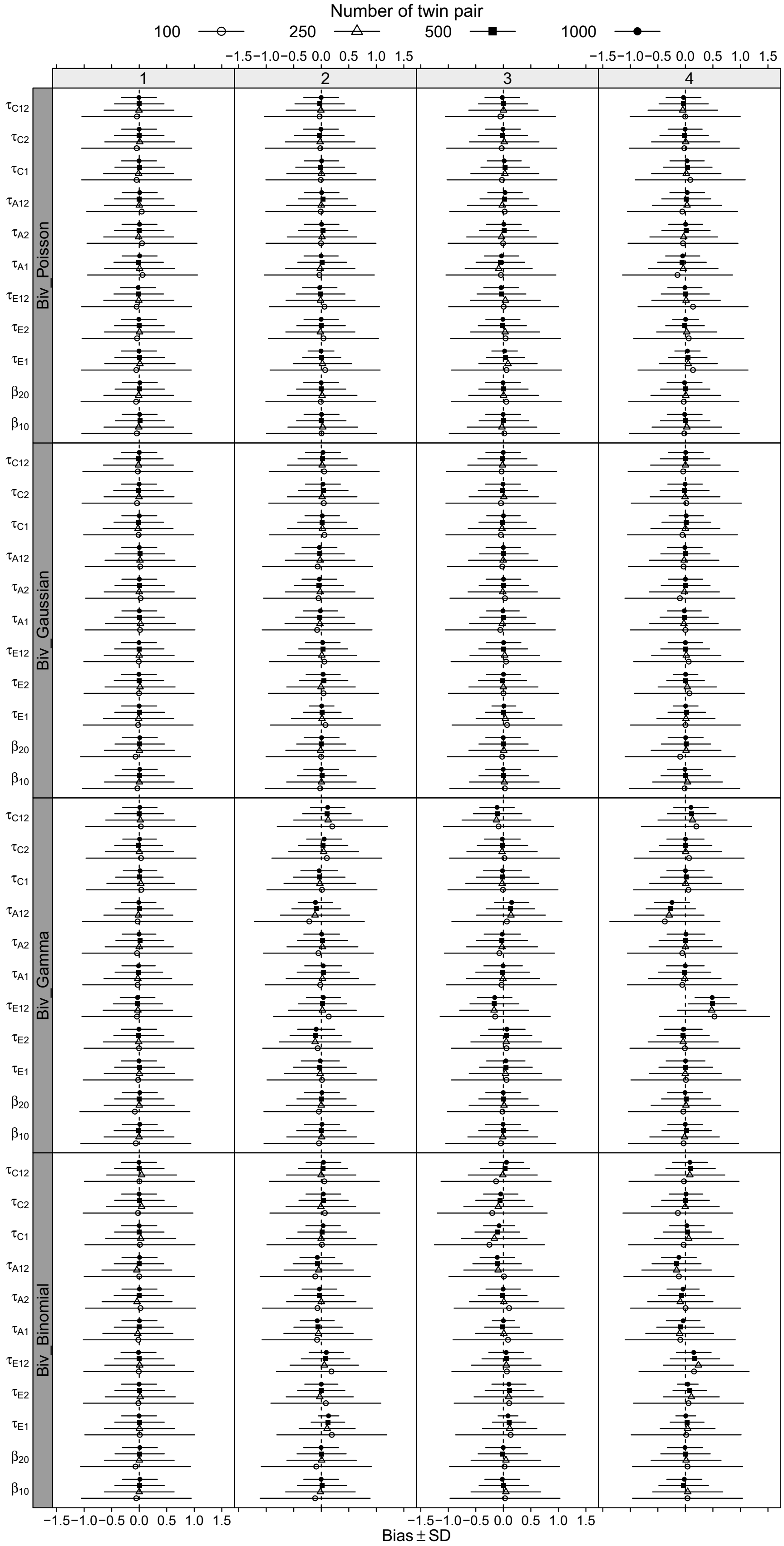

Fig. 2 Average bias and confidence intervals on a standardized scale by sample size and simulation scenario—two traits case

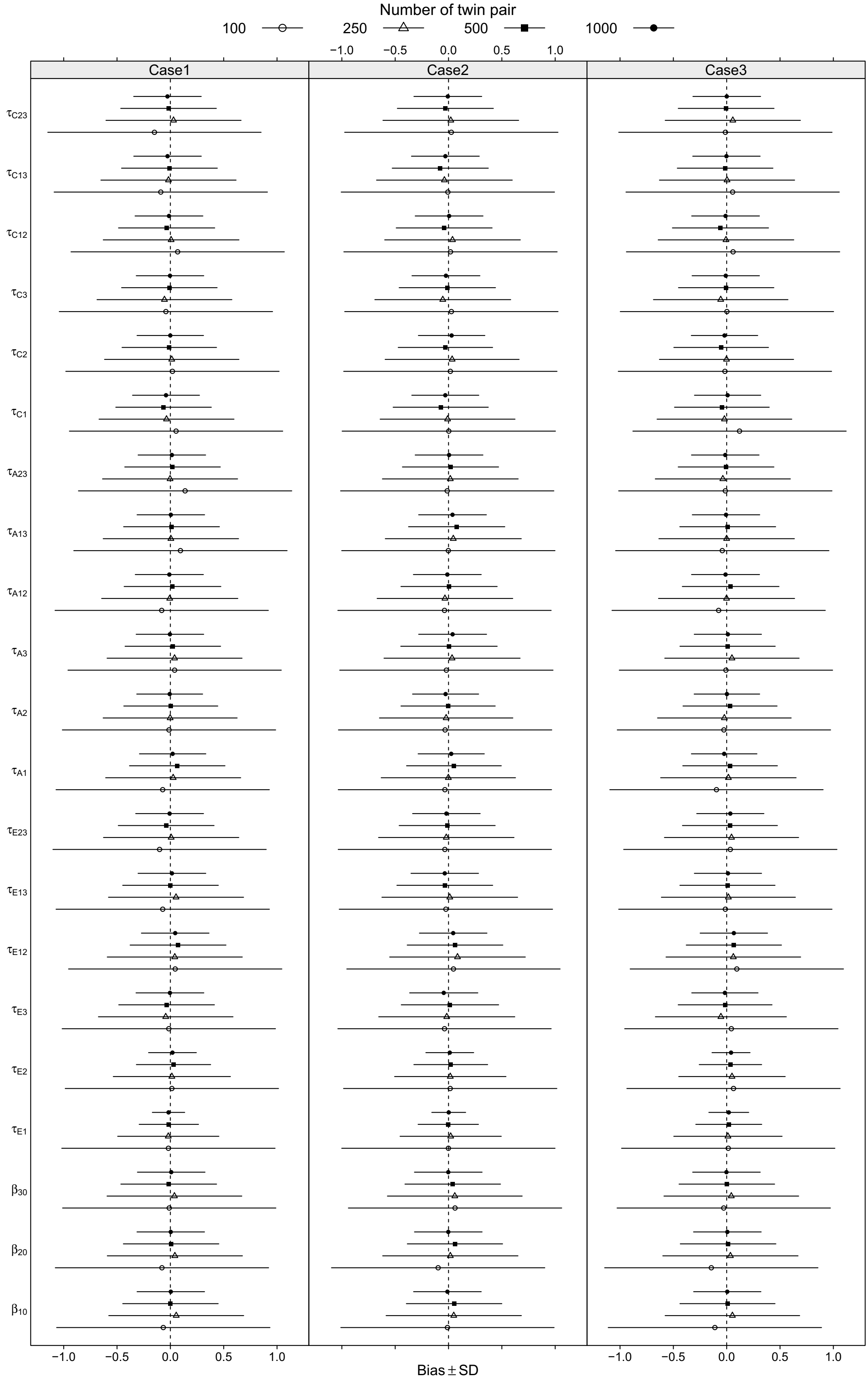

Fig. 3 Average bias and confidence intervals on a standardized scale by sample size and simulation scenario—mixed traits case

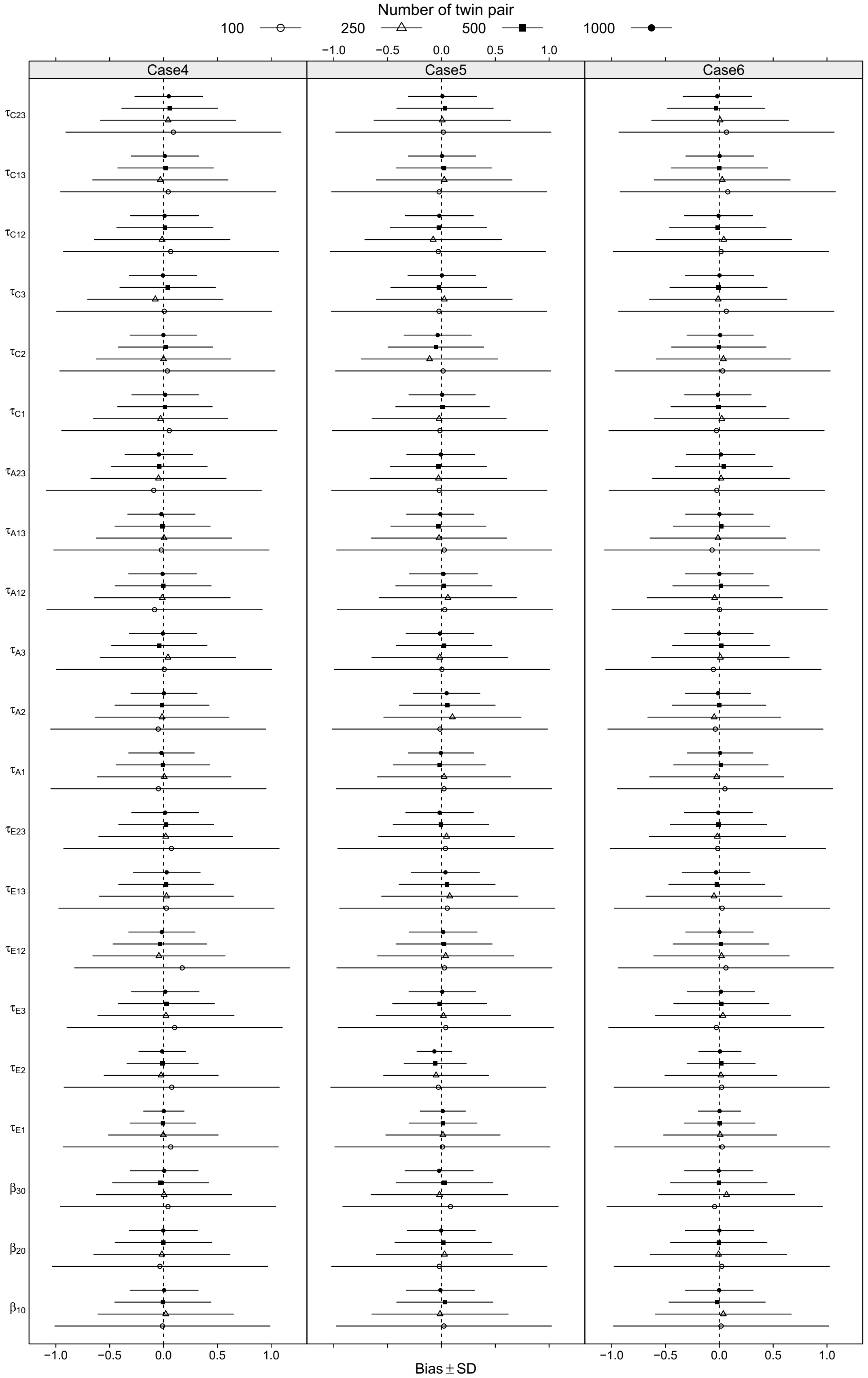

Fig. 4 Average bias and confidence intervals on a standardized scale by sample size and simulation scenario—mixed traits case

Table 1 Parameter estimates, standard errors and pseudo log-likelihood values by fitted models—dichotomous twin data

Parameters	Models			
	E	AE	CE	ACE
τ_E	1.00(0.03)	0.73(0.03)	0.80(0.03)	0.72(0.03)
τ_A	–	0.27(0.04)	–	0.34(0.11)
τ_C	–	–	0.20(0.03)	−0.06(0.10)
h^2	–	0.27(0.03)	–	0.34(0.11)
pll(df)	−1340.99(2)	−1315.58(3)	−1319.61(3)	−1315.39(4)

and Maes (2004) and Rabe-Hesketh et al. (2008) through generalized linear mixed models. The main goal of this study is to measure the genetic influence on the binary trait. We specified the model using the logit link and the binomial variance functions as usual in logistic regression models. In this example, we do not have covariates to compose the linear predictor. We fitted a series of special cases of the ACE model whose estimates and standard errors are presented in Table 1 along with the pseudo log-likelihood (pll) values and degrees of freedom (df) (Bonat et al. 2018).

The first interesting result in Table 1 is the negative value and consequently non-significance of the common environment component. Such a result agrees with previous analysis of Rabe-Hesketh et al. (2008). It is also an indication that the AE model should provide a better fit than the CE model. This conclusion is corroborated by the pseudo log-likelihood values. The heritability of the binary trait is 0.27 (0.03) and highly significant (p-value < 0.001).

Rabe-Hesketh et al. (2008) determined the heritability of the lifetime psychiatric illness trait as 0.43 $(0.35 - 0.50)$. However, it is important to highlight that the two measures of heritability are not comparable because they are based on different model assumptions. The Rabe-Hesketh et al. (2008) approach measures the correlation at the latent (random effects) level. On the other hand, our approach measures the correlation at the trait (marginal) level. Thus, it is expected that our approach provides lower correlation than the one based on the random effects approach. In our approach the total variance of the binary trait is clearly divided into the unique environment, additive genetic and common environment effects, while in the random effects approach the genetic and common environment influences are measured as a kind of overdispersion in the binary trait and modelled by the Gaussian random effects.

Dataset 2: Continuous Data

The second example regards a fairly common continuous trait analysis. We use the `twinbmi` data available in the `mets` package for the statistical software `R`. The analysis goal is to investigate the genetic and common environment influences on the body mass index (BMI). The dataset consists of 11188 observations, however, for this data analysis we considered only paired twin-pairs. The resulting dataset consists of 4271(2788 DZ and 1483 MZ) twin-pairs. Additionally, we have the covariates age and gender to compose the linear and matrix linear predictors. Thus, it is also of interest to verify how these covariates affect the dispersion components as discussed in subsect. Modelling the dispersion parameters and parametrization cautions. The BMI is a continuous trait, thus we specified our model using the identity link function and the constant variance function as usual in Gaussian mixed models. The linear predictor was composed of the interactive effect between zygosity (DZ and MZ) and twin pair code (Twin 1 and Twin 2).

In order to explore the influence of the age and gender on the dispersion components, we fitted the ACE model for twin data modelling each of its components as a linear function of the covariates age and gender. The covariate age was standardized to have mean zero and variance one and the level female was specified as the reference level of the covariate gender. Based on the fit of the ACE model, we verified that the components associated with the common environment effect were not significant. Thus, we dropped all these terms from the model and fit the AE model, but still with both τ_A and τ_E components modelled as a function of the covariates. Finally, analysing the fitted model, we concluded that the covariate gender was significant for both additive genetic and unique environment components. On the other hand, the covariate age was significant only for the unique environment component. The results are summarized in Table 2.

The model is parametrized such that $\tau_{E(0)}$, $\tau_{A(0)}$ and $\tau_{C(0)}$ are the intercepts of the unique environment, additive genetic and common environment effects, respectively. Similarly, $\tau_{E(1)}, \tau_{A(1)}, \tau_{C(1)}$ and $\tau_{E(2)}, \tau_{A(2)}, \tau_{C(2)}$ are the effects of the age

Table 2 Parameter estimates, standard errors, log-likelihood and *Akaike* criterion values by fitted models—continuous twin data

Parameters	Models		
	ACE	AE	AE Simplified
$\tau_{E(0)}$	4.53(0.21)	4.45(0.20)	4.50(0.20)
$\tau_{E(1)}$	0.44(0.15)	0.49(0.14)	0.61(0.12)
$\tau_{E(2)}$	−1.28(0.28)	−1.21(0.26)	−1.23(0.26)
$\tau_{A(0)}$	9.26(0.75)	10.11(0.37)	10.05(0.37)
$\tau_{A(1)}$	0.90(0.51)	0.40(0.25)	–
$\tau_{A(2)}$	−2.11(0.98)	−2.81(0.48)	−2.80(0.48)
$\tau_{C(0)}$	0.84(0.64)	–	–
$\tau_{C(1)}$	−0.48(0.43)	–	–
$\tau_{C(2)}$	−0.69(0.83)	–	–
ll	−22293.07(13)	−22294.71(10)	−22295.88(9)
AIC	44612.14	44609.42	44609.76

and gender on the unique environment, additive genetic and common environment components, respectively.

The results in Table 2 show that the covariate age increases the unique environment component. On the other hand, the covariate gender indicates that the unique environment and additive genetic effects are smaller for males than for females. To better illustrate the effect of modelling the dispersion components Fig. 5 shows the unique environment, additive genetic and heritability index along with their respective 95% confidence intervals as a function of the standardized age.

Results in Fig. 5 show that females are more susceptible to the environment and genetic effects than males. For both genders the age increases the unique environment component resulting in smaller heritability.

Dataset 3: bivariate dichotomous data

The goal of this third example is to show how our approach adapts to deal with bivariate binary traits in the context of twin studies. We use the dataset analysed by Feng et al. (2009) regarding bronchopulmonary dysplasia (BPD) and respiratory distress syndrome (RDS) on preterm infants. Both diseases are lung related and expected to have a genetic component. The dataset consists of 200 twin-pairs being 137 DZ and 63 MZ. Additionally, we considered the covariates: birth weight (BW), gestation age (GA) and gender (1: male and 0: female).

The linear predictor for both traits was specified as a linear combination of the covariates BW, GA, gender and interactive effect between zygosity (DZ and MZ) and twin pair code (Twin 1 and Twin 2). We fitted the bivariate ACE model and its special cases CE, AE and E as well as their univariate counterparts obtained by fixing the cross-trait covariance parameters at zero. Table 3 presents the pseudo

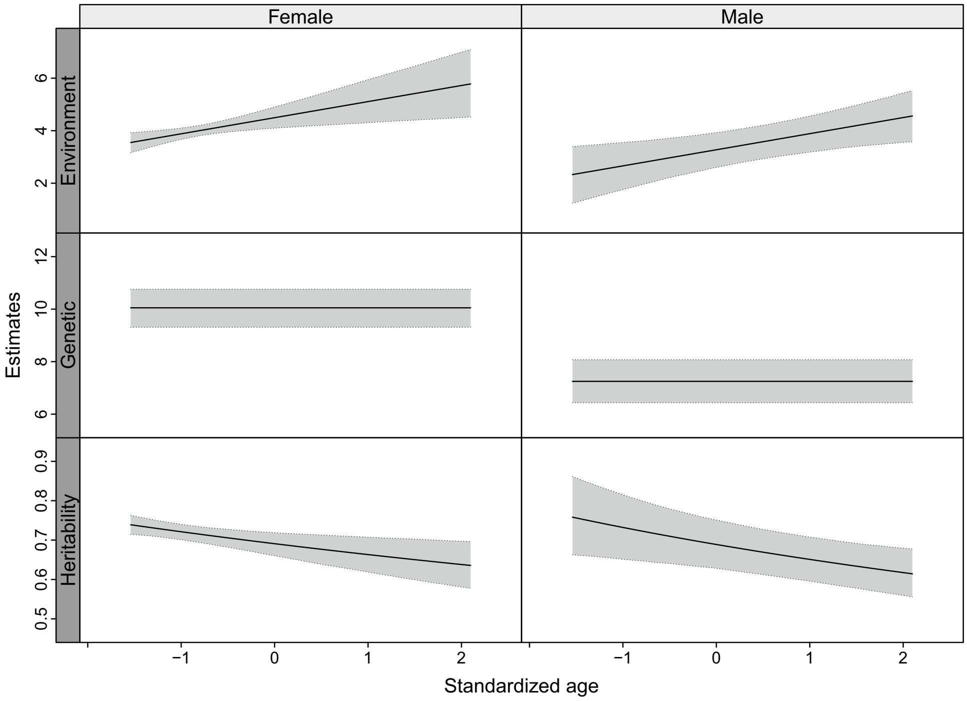

Fig. 5 Unique environment, additive genetic and heritability index as a function of standardized age—continuous twin data

Table 3 Pseudo log-likelihood (`pll`) values along with the pseudo *Akaike* (`pAIC`), Bayesian (*pBIC*) information criterion and degrees of freedom (df) by fitted models—bivariate binary twin data

Criterion	Models							
	Univariate				Multivariate			
	ACE	CE	AE	E	ACE	CE	AE	E
pll	−323.03	−334.97	−324.22	−359.99	−311.86	−323.82	−312.90	−347.59
pAIC	686.06	705.94	684.44	751.98	669.72	687.64	665.80	729.18
pBIC	779.75	790.26	768.76	826.93	777.47	781.33	759.49	808.82
df	20	18	18	16	23	20	20	17

Table 4 Parameter estimates and standard errors for some genetic and environment indices of interest—bivariate binary twin data

Indices	Traits		
	BPD	RDS	BPD × RDS
Heritability	0.70(0.15)	0.43(0.07)	0.98(0.18)
Environmentality	0.30(0.15)	0.57(0.07)	0.02(0.18)
Genetic correlation	–	–	0.45(0.11)
Environment correlation	–	–	0.01(0.11)

log-likelihood (`pll`) values along with the pseudo *Akaike* (`pAIC`), Bayesian (*pBIC*) information criterion and degrees of freedom (df) by fitted model.

The results in Table 3 show that the multivariate models provide a better fit than its univariate counterparts. In the univariate and multivariate cases all goodness-of-fit measures agree that the AE model provides the best balance between goodness-of-fit and complexity. Thus, Table 4 presents some environment and genetic indices of interest when analysing bivariate data.

Results in Table 4 show that both traits BPD and RDS are genetic influenced and highly correlated. We note by passing that the unique environment correlation and bivariate environmentality are not statistically significant. Feng et al. (2009) analysed only the trait BPD using a probit model and got an heritability of 78.18% using SAS and 78.94% using OpenMx, which show an agreement between their and our approaches.

Dataset 4: bivariate continuous data

The main goal of this example is to illustrate how to deal with a fairly common case of bivariate continuous traits in the context of twin studies. Furthermore, we explore the flexibility of our proposed model class and model the dispersion components as a linear combination of a covariate of interest. We defined as the traits of interest the body weight and height measures on 861 (327 DZ and 534 MZ) twin-pairs. These traits are well-known to be highly genetic influenced and correlated. The data set is available as an example in the OpenMx package (Neale et al. 2016). The covariate age was used for modelling both the mean and covariance structures. Both traits and the covariate were standardized to have mean zero and variance one. We used the identity link function and the constant variance function as usual in Gaussian mixed models.

We fitted the bivariate ACE model for twin data and model each of its dispersion parameters as a linear function of the covariate age. Then, we analysed the output of the fitted model and in order to obtain a more parsimonious model we dropped all the non-significant terms. Thus, we got a final model that is a special case of the AE model, where the additive genetic coefficient associated with the trait body weight is modelled by the covariate age. Similarly, the unique environment component of both traits depend significantly on the covariate age. The results are summarized in Table 5.

Table 5 Parameter estimates, standard errors (SE) and Z statistics by fitted models—bivariate continuous twin data

Parameters	Models			
	ACE		Simplified AE	
	Est(SE)	Z-value	Est(SE)	Z-value
$\tau_{E1(0)}$	0.15(0.01)	16.09	0.16(0.01)	16.18
$\tau_{E1(1)}$	0.03(0.01)	3.16	0.03(0.01)	3.78
$\tau_{E2(0)}$	0.12(0.01)	16.17	0.12(0.01)	16.26
$\tau_{E2(1)}$	0.02(0.01)	2.55	0.02(0.01)	3.01
$\tau_{E12(0)}$	0.02(0.01)	3.62	0.02(0.01)	3.97
$\tau_{E12(1)}$	0.00(0.01)	−0.27	–	–
$\tau_{A1(0)}$	0.96(0.10)	9.86	0.85(0.04)	20.53
$\tau_{A1(1)}$	0.23(0.09)	2.49	0.11(0.03)	3.48
$\tau_{A2(0)}$	0.90(0.09)	10.32	0.88(0.04)	21.48
$\tau_{A2(1)}$	0.04(0.09)	0.50	–	–
$\tau_{A12(0)}$	0.47(0.07)	6.61	0.44(0.03)	14.02
$\tau_{A12(1)}$	−0.03(0.07)	−0.42	–	–
$\tau_{C1(0)}$	−0.12(0.09)	−1.34	–	–
$\tau_{C1(1)}$	−0.14(0.09)	−1.64	–	–
$\tau_{C2(0)}$	−0.02(0.09)	−0.21	–	–
$\tau_{C2(1)}$	−0.05(0.08)	−0.55	–	–
$\tau_{C12(0)}$	−0.03(0.07)	−0.41	–	–
$\tau_{C12(1)}$	0.02(0.07)	0.29	–	–
ll(df)	−3914.91(28)			−3920.20(19)
AIC	7885.82			7878.40
BIC	8057.86			7995.14

Results in Table 5 show that the simplified AE model provides the best balance between goodness-of-fit and complexity, as expected. To better explore the model's output, we compute some genetic measures of interest such as the heritability of each trait and the bivariate heritability. Furthermore, we compute the genetic, environment and phenotypic correlations as a function of the covariate age. The results are shown in Fig. 6.

Results in Fig. 6 show that both traits are highly genetic influenced and correlated, as expected. For both traits the heritability index decreases while the age increases. On the other hand, the bivariate heritability is not influenced by the covariate age. Finally, it is interesting to note that the additive genetic coefficient associated with the trait body weight depends significantly on the covariate age. Such a result could indicate that there is an interaction between environment and genetic factors affecting this trait (Table 6).

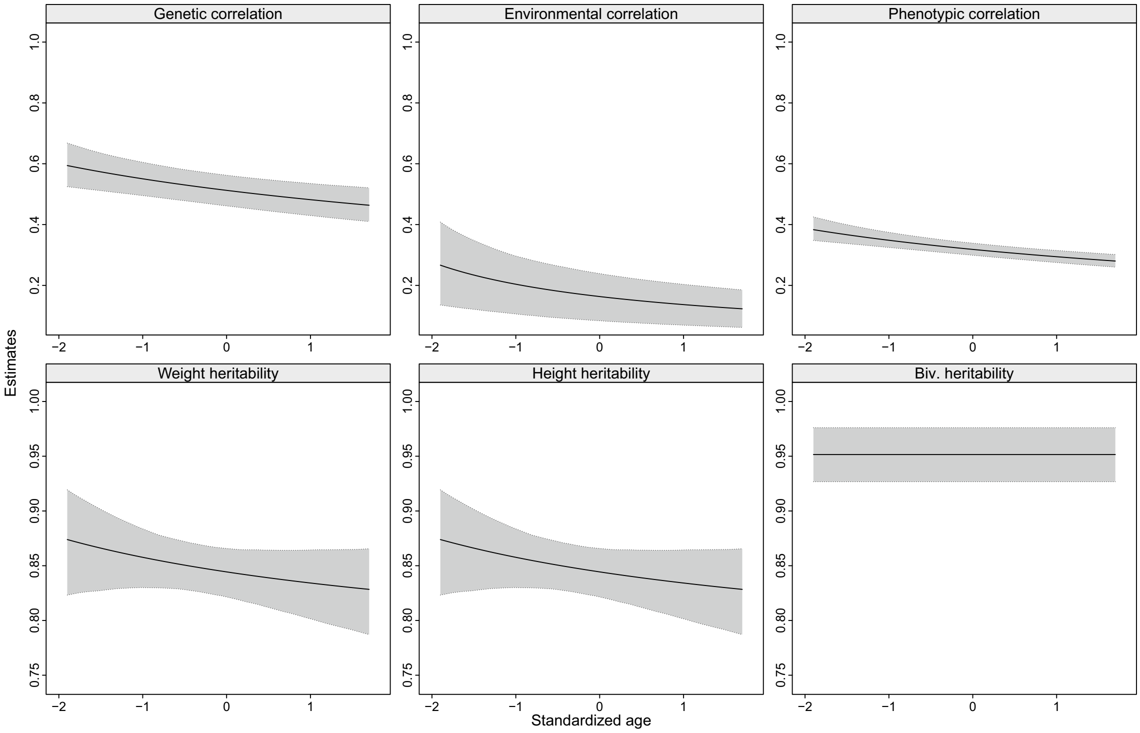

Fig. 6 Genetic, environment and phenotypic correlations (first line). Body weight heritability, height heritability and bivariate heritability (second line)—bivariate continuous twin data

Table 6 Pseudo log-likelihood (pll) values, degrees of freedom (df), pseudo *Akaike* (pAIC) and Bayesian (*pBIC*) information criterion by fitted models—mixed types of traits data

Models	Gaussian			Mixed		
	pll(df)	pAIC	pBIC	pll(df)	pAIC	pBIC
E	−107.28(19)	252.56	345.88	−101.56(21)	245.12	348.26
AE	−95.85(22)	235.70	343.75	−90.74(24)	229.48	347.36
CE	−97.32(22)	238.64	346.69	−92.29(24)	232.58	350.46
ACE	−95.85(25)	241.70	364.49	−90.69(27)	235.38	367.99

Dataset 5: mixed types of traits data

In this example, we analyse a simulated data set based on a real data set concerning sleep's quality in a sample of 250(135 DZ and 116 MZ) Danish twin pairs. We opted to simulate the data set because we do not have permission to circulate it as a supplementary material for our paper. The traits are cortisone levels when waking up (T0) and PSQI (Pittsburgh Sleep Quality Index). The first one is a continuous trait, while the second one is a scale varying from 0 to 21 (slower values better sleep's quality). Consequently, the goal of this example is to show how to deal with mixed types of traits in the context of twin data analysis.

Exploratory analyses showed that the continuous trait T0 is symmetric, however heavier tails than the Gaussian distribution could be observed. Thus, we compared the fit of the standard Gaussian model (identity link and constant variance functions) with the fit of a Tweedie model obtained by using the logarithm link and Tweedie variance functions. The power parameter p was estimated based on the data. The trait PSQI was divided by 21(largest value of the scale) in order to have values in the unit interval. Thus, we treat the PSQI trait as a continuous bounded data. For comparison proposes, we fitted the standard Gaussian model and the flexible quasi-beta regression model obtained by specifying the logit link and extended binomial variance functions. The power p indexing the extended binomial variance function was estimated based on the data. In order to take into account the sources of dependences introduced by the twin design, we fitted the bivariate ACE model as well as its special cases AE, CE and E models. The fitted models are compared through the values of the pseudo log-likelihood function, pseudo *Akaike* and Bayesian information criterion, whose values are presented in Table 7.

Table 7 Parameter estimates and standard errors for some genetic and environment indices of interest—mixed types of traits data

Indices	Traits		
	T0	PSQI	T0 × PSQI
Heritability	0.22(0.08)	0.30(0.08)	1.36(0.70)
Environmentality	0.78(0.08)	0.70(0.08)	−0.36(0.70)
Genetic correlation	–	–	−0.51(0.23)
Environment correlation	–	–	0.05(0.08)

Table 8 Pseudo log-likelihood (`pll`) values, degrees of freedom (df), pseudo *Akaike* (`pAIC`) and Bayesian (*pBIC*) information criterion by fitted models—mixed types of traits data

Models	Goodness of fit measures		
	pll(df)	pAIC	pBIC
E	1294.32(15)	−2558.64	−2471.37
AE	1311.04(21)	−2580.08	−2457.91
CE	1307.94(21)	−2573.88	−2451.71
ACE	1311.69(27)	−2569.38	−2412.30

Results in Table 7 show that the mixed approach provides a better fit than the bivariate Gaussian for all models considered, which in turn highlights the flexibility of our approach. The power parameter associated with the Tweedie variance function was estimated at −1.37 (1.59). In general, negative values for the power parameter in the Tweedie variance function indicates heavier tails than the Gaussian distribution. Such a result agrees with our exploratory analysis, however, the uncertainty associated with this estimate is large and a 95% confidence interval contains the value zero, which shows that the Gaussian distribution is not rejected as a suitable distribution for this trait. On the other hand, for the trait PSQI the power parameter indexing the extended binomial variance function was estimated at 1.52 (0.95). Bonat et al. (2018) argued that power parameter close to one approximates well the mean and variance relationship of the beta distribution.

Concerning the twin structure we have that the AE model provides a fit similar to the ACE model in terms of pseudo log-likelihood values. The pAIC and pBIC indicate the AE model as the best balance betweeen goodnes-of-fit and complexity. Finally, we analysed the fitted ACE model and verified that all the dispersion parameter estimates associated with the common environment effect were not significant. Table 7 presents estimates and standard errors for some genetic and unique environment measures of interest such as heritability, environmentality, bivariate heritability and environmentality as well as the environment and genetic correlations.

The results in Table 7 show that both traits are barely genetic influenced and negative correlated. The negative correlation implies the larger than one value of the bivariate heritability. Finally, the environment correlation is not statistically significant.

Dataset 6: multivariate bounded data

In the last example, we aim to disentagle genetic and environmental influences to three bounded outcomes reflecting mental conditions. The data set consists of 828(524 DZ and 274 MZ) twin pairs and is simulated based on real data. As in the previous example, we opted to simulate the data set based on the results of the real data in order to circulate the data set as a supplementary material for our paper. The cognitive battery includes the Mini Mental State Examination (MMSE) (Folstein et al. 1975) a standard neurological screen that is especially effective for screening at the low end of cognitive functioning. The MMSE takes integer values in the set $[0, \dots, 30]$ and is highly right-skewed in distribution in any population-based sample. The second outcome corresponds to depression symptomatology which was evaluated based on an adaptation of the depression section of the Cambridge Mental Disorders of the Elderly Examination (CAMDEX) (Roth et al. 1986). The depression scale used here is a composite of responses to 17 depression items (McGue and Christensen 1997). The depression score is taking integer values in the interval set $[0, \dots, 30]$ and is highly left-skewed in distribution in any population-based sample. The third outcome is a Social Activity scale based on six items that assess the frequency with which the individual is engaged with others (e.g., how often do you leave your home, how often do you go to a party) and mental pursuits (e.g., how often do you engage in a hobby). Each item is rated on a 1 (Never) to 4 ($5 - 7$ days a week) scale. These are frequently occuring type of outcomes obtained from experiments using questionaires or similar assesment. Such outcomes or transformation of these are usually assumed to be gaussian for the practical analysis and the boundedness of the score is ignored.

In order to take into account the boundness, we assume a flexible quasi-beta model for each outcome (Bonat et al. 2019). As in the previous examples, the sources of dependences introduced by the twin design were modelled using the ACE model as well as its special cases AE, CE, and E models. The fitted models are compared through the values of the pseudo log-likelihood, *Akaike* and Bayesian information criterion, whose values are presented in Table 8.

The results in Table 8 show that the AE offers the best balance between goodness-of-fit and simplicity. Thus, we opted to continue the analysis by reporting some genetic and unique environment indices of interest in Table 9.

Table 9 Parameter estimates and standard errors for some genetic and environment indices of interest—bounded types of traits data

Traits	Indices			
	Heritability	Environmentability	Genetic corr.	Environmental corr.
MMSE cognition state	0.18(0.07)	0.82(0.07)	–	–
Depression index	0.27(0.07)	0.73(0.07)	–	–
Social activity index	0.18(0.07)	0.82(0.07)	–	–
MMSE × Depress	0.25(0.13)	0.75(0.13)	−0.44(0.19)	−0.37(0.06)
MMSE × Social act	0.15(0.15)	0.85(0.15)	0.28(0.24)	0.35(0.06)
Depress × Social act	0.17(0.10)	0.83(0.10)	−0.39(0.18)	−0.54(0.05)

In that case, the results suggest that all traits are weakly genetic influenced and present weak genetic and environment correlation.

Discussion

In this paper we have presented a comprehensive statistical modelling framework for the analysis of twin and family data. Motivated by six real data sets, we have shown that our framework can deal with a wide variety of traits types and correlation structures where existing modelling approaches have difficulties. Our models inherent the structure of the traditional generalized linear models and consequently are specified by using separate pairs of link and variance functions combined with linear and matrix linear predictors for the mean and covariance structure in the style of Bonat and Jørgensen (2016). Thus, we believe that there are some pedagogical advantages on our modular specification of models incentivising the researcher to think constructively about the covariance structure, while drawing on previous experiences from generalized linear models.

Linear mixed models and structural equation models along with the liability model for binary traits are the main statistical modelling frameworks to deal with twin and family data. The advantages of our approach in comparison with the above mentioned approaches are the flexibility to deal with non-Gaussian data by the simple choice of a link and variance functions while keeping a simple interpretation for the dispersion components. The marginal specification based only on second-moments assumptions allows us to extend all the standard measures of interest in genetic studies to non-Gaussian data in a straightforward way. Furthermore, the proposed parametrization of the covariance structure obtained by using a linear combination of known matrices and without imposing any boundary constrains on parameters allows us to extend the models even more by modelling the dispersion components as linear functions of covariates in a regression fashion. Such a flexibility is not available at most currently software packages for the analysis of twin and family data. Additionally, the proposed parametrization allows easy inference for all measures of interest including hypothesis tests for the dispersion components. The scope of our approach is wide as the class of models is very broad covering the classic multivariate generalized linear models and beyond to those satisfying the mean and variance relationship. The relationship which essentially allows the analysts for imposing Taylor's power law providing for instance deeper biological meaning.

The main technical advantage of the proposed framework is the simplicity of the fitting method, which amounts to finding the root for a set of non-linear equations. It is in sharp contrast for instance with Generalized linear mixed models (GLMMs) where the lack of a closed form expression for the likelihood function and marginal distribution of the data vector implies the use of numerical methods for integration and maximization resulting in a more complex and computationally demanding numerical problem. This fact is also related to the interpretation of the model parameters and derived measures of interest such as heritability, genetic, environment and phenotypic correlations and etc. In the GLMMs the correlation induced by the twin or family designs is modelled through Gaussian random effects, thus all the correlation are related to the random effect rather than the trait itself, which in turn does not provide a simple interpretation for these coefficients and does not allow a simple extension of the aforementioned measures to non-Gaussian data.

We conducted a set of simulation studies on the properties of the estimating function estimators. In general, the results suggested that the estimating function estimators are unbiased and consistent for all scenarios considered. Then, we proceed to the analysis of six data sets. Thus, the first and second datasets illustrate the case of a single binary and continuous trait, respectively. Similarly, the third and fourth datasets explore the case of bivariate binary and continuous traits, respectively. Finally, the last two examples are the most challenging where in the fifth we have a mixed of a potentially heavy tail continuous trait combined with a bounded continuous trait and in the sixth we have three continuous bounded traits. For all cases our model approach was quite effective providing a complete tool for the data analysis, generating new insights and opening opportunities for further investigations. For instance, the modelling of the dispersion components in a regression model fashion opens a new avenue for research on the interaction between genetic and environment effects. Hence the proposed model class is novel to twin analysis and may prove valuable for instance in combining outcomes in register studies as seen from the Nordic twin cohorts or in studies of integrating jointly related omics outcomes in twin and family studies. Further, previously conducted studies may be re-analyzed exploiting the dispersion modelling and the opportunity to gain deeper insight to mutual relationships of outcomes at genetic and environmental levels including their interaction.

There are many possible extensions to the basic multivariate model discussed in this article. For instance, we intend to include penalized splines and the use of regularization for high dimensional data, with important applications in genetics. There is also a demand to include facilities to deal with missing and censored data. Finally, the model complexity increases rapidly when the number of traits increases. Thus, tools for visualizing and interpret the model's results are in high demand.

Funding Not applicable.

Data Availability Data is available as a supplementary material.

Declarations

Conflict of interest Wagner Hugo Bonat and Jacob V. B. Hjelmborg declares that they have no conflict of interest.

Ethical Approval Not applicable.

Human and Animal Rights statement and Informed consent Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Code Availability R code is available as a supplementary material.

References

Bonat WH, Jørgensen B (2016) Multivariate covariance generalized linear models. J Royal Statist Soc: Series C 65:649–675

Bonat WH, Kokonendji CC (2017) Flexible tweedie regression models for continuous data. J Statist Comput Simulat 87(11):2138–2152

Bonat WH, Jørgensen B, Kokonendji CC, Hinde J, Demétrio CGB (2018) Extended Poisson–Tweedie: properties and regression models for count data. Stat Modell 18(1):24–49

Bonat WH, Peterle R, Hinde J, Demétrio CGB (2018) Flexible regression models for continuous bounded data. Stat Modell

Bonat WH, Petterle RR, Hinde J, Demétrio CG (2019) Flexible quasi-beta regression models for continuous bounded data. Stat Model 19(6):617–633

Boomsma D, Busjahn A, Peltonen L (2002) Classical twin studies and beyond. Nat Rev Genet 3:872–882

Feng R, Zhou G, Zhang M, Zhang H (2009) Analysis of twin data using sas. Biometrics 65(2):584–589

Folstein MF, Folstein SE, McHugh PR (1975) Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 12(3):189–198

Holst KK, Scheike TH, Hjelmborg JB (2016) The liability threshold model for censored twin data. Comput Stat Data Anal 93:324–335

Jørgensen B (1987) Exponential dispersion models. J Royal Statist Soc Series B 49(2):127–162

Jørgensen B (1997) The theory of dispersion models. Chapman & Hall

Jørgensen B, Knudsen SJ (2004) Parameter orthogonality and bias adjustment for estimating functions. Scand J Stat 31(1):93–114

Jørgensen B, Kokonendji CC (2016) Discrete dispersion models and their tweedie asymptotics. AStA Adv Stat Anal 100(1):43–78

Khoury MJ, Beaty TH, Cohen BH (1993) Fundamentals of genetic epidemiology. Oxford University Press, Fundamentals of Genetic Epidemiology

Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73(1):13–22

McArdle JJ, Prescott CA (2005) Mixed-effects variance components models for biometric family analyses. Behav Genet 35(5):631–652

McGue M, Christensen K (1997) Genetic and environmental contributions to depression symptomatology: evidence from danish twins 75 years of age and older. J Abnorm Psychol 106(3):439–448

Neale MC, Maes HH (2004) Methodology for genetic studies of twins and families. Technical report, Virginia Common wealth University, Department of Psychiatry. http://ibgwww.colorado.edu/workshop2004/cdrom/HTML/book2004a.pdf

Neale MC, Hunter MD, Pritikin JN, Zahery M, Brick TR, Kirkpatrick RM, Estabrook R, Bates TC, Maes HH, Boker SM (2016) OpenMx 2.0: extended structural equation and statistical modeling. Psychometrika 81(2):535–549

Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc Ser A 135(3):370–384

Ozaki K, Toyoda H, Iwama N, Kubo S, Ando J (2011) Using non-normal sem to resolve the acde model in the classical twin design. Behav Genet 41(2):329–339

Prescott CA (2004) Using the mplus computer program to estimate models for continuous and categorical data from twins. Behav Genet 34(1):17–40

Rabe-Hesketh S, Skrondal A, Gjessing HK (2008) Biometrical modeling of twin and family data using standard mixed model software. Biometrics 64(1):280–288

Roth M, Tym E, Mountjoy CQ (1986) Camdex: a standardized instrument for the diagnosis of mental disorder in the elderly with special reference to the elderly detection of dementia. Br J Psychiatry 149:698–709

van Dongen J, Slagboom PE, Draisma HHM, Martin NG, Boomsma DI (2012) The continuing value of twin studies in the omics era. Nat Rev Genet 13:640–653

Wedderburn RWM (1974) Quasi-likelihood functions, generalized linear models, and the gauss-newton method. Biometrika 61(3):439–447