Wagner Hugo Bonat*

Modelling Mixed Types of Outcomes in Additive Genetic Models

DOI 10.1515/ijb-2017-0001

Abstract: We present a general statistical modelling framework for handling multivariate mixed types of outcomes in the context of quantitative genetic analysis. The models are based on the multivariate covariance generalized linear models, where the matrix linear predictor is composed of an identity matrix combined with a relatedness matrix defined by a pedigree, representing the environmental and genetic components, respectively. We also propose a new index of heritability for non-Gaussian data. A case study on house sparrow (*Passer domesticus*) population with continuous, binomial and count outcomes is employed to motivate the new model. Simulation of multivariate marginal models is not trivial, thus we adapt the NORTA (Normal to anything) algorithm for simulation of multivariate covariance generalized linear models in the context of genetic data analysis. A simulation study is presented to assess the asymptotic properties of the estimating function estimators for the correlation between outcomes and the new heritability index parameters. The data set and `R` code are available in the supplementary material.

Keywords: genetic, estimating functions, mixed models, multiple outcomes, non-Gaussian data

1 Introduction

Linear mixed effects models [1, 2] are the main statistical tool in the context of quantitative genetic data analysis. The estimation of the additive genetic variance and thus the heritability of different outcomes is one of the main goals in quantitative genetic [3, 4]. Examples appear frequently in evolutionary biology [5] and animal breeding [6]. Although flexible for Gaussian data, linear mixed effects models are unsuitable for analysing binary, binomial, count and asymmetric continuous outcomes.

For the analysis of non-Gaussian outcomes the generalized linear mixed models (GLMMs) framework is a frequent choice [7, 8]. GLMMs consist of specifying a generalized linear model conditionally on a multivariate latent distribution, often the multivariate Gaussian. In general GLMMs are computationally demanding and many different algorithms have been proposed in the past three decades, see McCulloch [9] and [10] for additional references. A further aspect of GLMMs that gives rise to concern is the general lack of a closed-form expression for the likelihood and the marginal distribution of the data vector. A related question is the special interpretation of parameters inherent from the construction of GLMMs. Thus, the covariate effects are conditional on the latent variables, whereas the correlation structure is marginal for the latent variables rather than for the outcomes.

The literature to deal with mixed types of outcomes in the context of quantitative genetic is sparse. Hadfield and Nakagawa [7], Hadfield [11] proposed to use the GLMMs family relying on MCMC (Markov Chain Monte Carlo) methods for estimation in the Bayesian framework. MCMC methods for GLMMs are challenging in terms of convergence and computational time. Furthermore, these challenges seem to be amplified for mixed types of outcomes and covariance modelling using genetic structures. Moreover, the implementation itself can be difficult, especially for end-users who might not be expert in programming. For a more comprehensive discussion on the computational challenges for MCMC methods, we refer the interested reader to Rue et al. [12] and Fong et al. [10]. Liu et al. [13] presented the penalized multivariate linear mixed model for the analysis of multi-traits in the context of genome-wide association studies. For a recent application of this approach and further discussion about mixed types of outcomes, see Liu et al. [14].

***Corresponding author: Wagner Hugo Bonat,** Department of Statistics, Universidade Federal do Paraná, Curitiba, Brazil, E-mail: wbonat@ufpr.br

The computation of the heritability index in the Gaussian case is trivially defined by the proportion of the variance attributed to the genetic component. For non-Gaussian data, it is not immediately obvious how to define the heritability of an outcome. A comprehensive discussion about the computation of the heritability index for non-Gaussian data based on GLMMs is presented in de Villemereuil et al. [15] along with a proposal of solution. Although, this new development on the computing of the heritability index for non-Gaussian data, it remains a daunting task requiring intricate calculation involving multivariate integrals and specific algorithms.

In this paper, we present an alternative approach for modelling mixed types of outcomes in the context of genetic data based on the multivariate covariance generalized linear models (McGLMs) [16].

In the McGLM framework the marginal covariance matrix is explicitly modelled combining a matrix linear predictor and a covariance link function. Variance function are employed to take into account non-normality and the mean structure is modelled using a link function and a linear predictor. McGLMs are fitted using quasi-likelihood and Pearson estimating functions, based on second-moment assumptions, and implemented in an efficient Newton scoring algorithm. McGLMs are an extension of the linear mixed effects models to deal with multivariate non-Gaussian data and have the last one as a special case. Furthermore, from the marginal specification of McGLMs emerges a marginal measure of heritability for non-Gaussian data.

McGLMs have many in common with GEE (Generalized Estimating Equations) models popular in the analysis of longitudinal data [17]. However, McGLMs were explicitly designed to deal with multiple outcomes and allow for a flexible and interpretable specification of the marginal covariance structure using a covariance link function combined with a linear combination of known matrices. Furthermore, in the McGLM framework the dispersion parameters play an important role in terms of model specification and interpretation. On the other hand, current GEE implementations deal only with one outcome and include a short list of pre-specified covariance structures, such as auto-regression and compound symmetry [18, 19]. In general in the GEE framework the dispersion parameters are considered as nuisance. In terms of fitting algorithm, both methodologies use the quasi-score function for estimation of the regression coefficients. In the McGLM framework the dispersion parameters are estimated using the Pearson estimating function [20], while in the GEE context the method of moments and the GEE2 are frequent choices [21, 22].

The model we shall present in Section 3 is motivated by the data set analysed in Holand et al. [8] consisting of a natural meta population of house sparrow (*Passer domesticus*) on six islands off the coast of Helgeland, Northern Norway. In this study blood samples were used to determine genetic parenthood, and the genetic pedigree for the birds could be established. The outcomes of interest are: (1) bill depth, (2) breeding season success, and (3) average reproductive intensity. The main biological goals are to compute the heritability of the different outcomes, as well as the possible correlation between outcomes that can be due to genetic and environment effects. This example is particularly interesting because the outcomes are of mixed types, i.e. bill depth is a continuous outcome, while breeding season success and average reproductive intensity are examples of binomial and count outcomes.

In view of the recent developments in the McGLMs framework the main contributions of this article are: (i) introducing a suitable specification of the McGLMs to deal with mixed types of outcomes in the context of genetic data analysis. (ii) presenting the extended binomial variance function for handling binomial and restricted outcomes. (iii) proposing a new marginal measure of heritability for non-Gaussian data. (iv) presenting a comparison between the conditional specification of GLMMs and the marginal specification of McGLMs to explain why negative heritability index can appear in practical data analysis and how to use this fact to easily assess the significance of the genetic effects. (v) adapting the NORTA (Normal to anything) [23] algorithm for simulation of McGLMs. (vi) presenting a simulation study to check the asymptotic properties of the estimating function estimators with emphasis to the correlation between outcomes and the new measure of heritability and (vii) analysing the house sparrow data set comparing the results with the ones obtained by Holand et al. [8].

We present the data set in Section 2. Section 3 presents the model and discusses its properties. Section 6 compares the conditional and marginal model specifications and explains why negative heritability index appears naturally in the marginal specification. Section 5 adapts the NORTA algorithm for simulation

of McGLMs. Section 6 presents a simulation study to verify the asymptotic properties of the estimating function estimators. Section 7 presents the application of the model to the data. Finally, the main results are discussed in Section 8, including some directions for future investigations. The data set and R [24] code are available in the supplementary material.

2 Data set

The case study analysed in this paper uses data from a natural meta population of house sparrow (*Passer domesticus*) on six islands off the coast of Helgeland, Northern Norway. For additional description of the field work, study area and genetic parenthood analyses, see Jensen et al. [5], Holand et al. [8], Pärn et al. [25], Ringsby et al. [26].

In this study, blood samples were used to determine genetic parenthood, and a genetic pedigree for the birds could be established. We highlight that the genetic parenthood structure was determined for each bird. The study has three outcomes of interest: (1) bill depth (BD), (2) breeding season success (BS), and (3) average reproductive intensity (AR). There are three covariates, namely sex (sex), hatch year (HY) and hatch island (HI). The data set that we have access corresponds to the study period from 1993 to 2002, which corresponds to 950 observations. Histograms in Figure 1 suggest that the Gaussian and binomial distributions are suitable choices for BD and BS. While hinting at potential problems with excess of zeroes and overdispersion for AR. The scatter plots suggest correlation only between AR and BS. It is important to highlight that the plots presented in Figure 1 can detect only marginal correlations between the traits and do not take into account the origin, genetic or environment as well as the covariate effects.

Holand et al. [8] analysed the house sparrow data set using univariate GLMMs in the Bayesian framework with random effects specified on the bird level. The INLA (Integrated Nested Laplace Approximation) algorithm [12] was employed for fitting the models. For the outcome BD the authors assumed a Gaussian distribution, while for BS and AR their choices were the binomial and zero-inflated Poisson distributions, respectively. The *deviance information criterion* (DIC) was applied for model selection. In the Section 7, we compare the results obtained by Holand et al. [8] with the ones obtained by fitting the multivariate model that we shall present in Section 3.

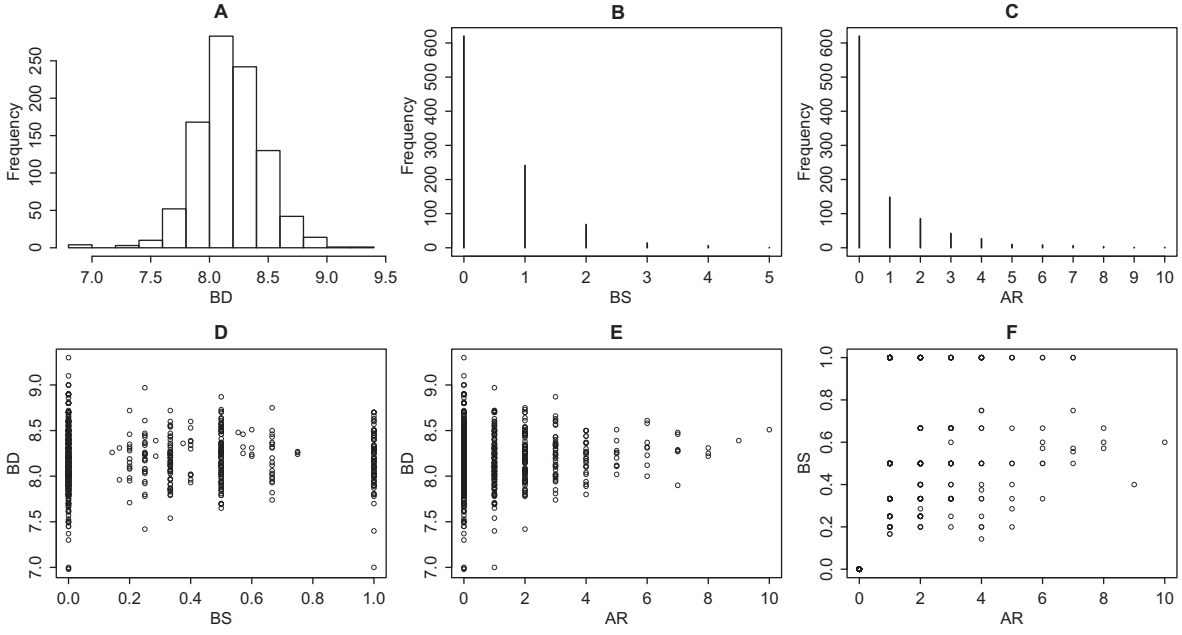

Figure 1: Histograms (A to C) and scatter plots (D to F) for outcomes in the house sparrow data set.

3 McGLMs for mixed types of outcomes

Let $\mathbf{Y}_{N\times R} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_R\}$ be an outcome matrix and let $\mathbf{M}_{N\times R} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_R\}$ denote the corresponding matrix of expected values. Let $\boldsymbol{\Sigma}_r$ denote the $N \times N$ covariance matrix within the outcome r for $r = 1, \ldots, R$. Similarly, let $\boldsymbol{\Sigma}_b$ be the $R \times R$ correlation matrix between outcomes. The McGLM as proposed by Bonat and Jørgensen [16] is given by

$$\begin{aligned} \mathrm{E}(\mathbf{Y}) &= \mathbf{M} = \{g_1^{-1}(\boldsymbol{X}_1\boldsymbol{\beta}_1), \ldots, g_R^{-1}(\boldsymbol{X}_R\boldsymbol{\beta}_R)\} \\ \mathrm{Var}(\mathbf{Y}) &= \boldsymbol{C} = \boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b \end{aligned} \tag{1}$$

where

$$\boldsymbol{\Sigma}_R \overset{G}{\otimes} \boldsymbol{\Sigma}_b = \mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1, \ldots, \tilde{\boldsymbol{\Sigma}}_R)(\boldsymbol{\Sigma}_b \otimes \boldsymbol{I})\mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1^T, \ldots, \tilde{\boldsymbol{\Sigma}}_R^T)$$

is the generalized Kronecker product [27]. The matrix $\tilde{\boldsymbol{\Sigma}}_r$ denotes the lower triangular matrix of the Cholesky decomposition of $\boldsymbol{\Sigma}_r$. The operator **Bdiag** denotes a block diagonal matrix and $\boldsymbol{I}$ denotes an $N \times N$ identity matrix. The functions g_r are orthodox link functions. Let $\boldsymbol{X}_r$ denote an $N \times k_r$ design matrix and $\boldsymbol{\beta}_r$ a $k_r \times 1$ regression parameter vector.

The key point for specifying McGLMs for mixed types of outcomes is the specification of the covariance matrix within outcomes $\boldsymbol{\Sigma}_r$. Following Bonat and Jørgensen [16] we define the covariance within outcomes by

$$\boldsymbol{\Sigma}_r = \mathrm{V}(\boldsymbol{\mu}_r; p_r)^{\frac{1}{2}}(\boldsymbol{\Omega}(\boldsymbol{\tau}_r))\mathrm{V}(\boldsymbol{\mu}_r; p_r)^{\frac{1}{2}}$$

where $\mathrm{V}(\boldsymbol{\mu}_r; p_r) = \mathrm{diag}(\vartheta(\boldsymbol{\mu}_r; p_r))$, denotes a diagonal matrix, whose main entries are given by the variance function $\vartheta(\cdot; p_r)$ applied element wise to the vector $\boldsymbol{\mu}_r$. The variance function plays an important role in McGLMs, since different choices for $\vartheta(\cdot; p_r)$ imply different marginal outcome distributions.

In the house sparrow data set, we have three types of outcomes, namely, a continuous (`BD`), a binomial (`BS`) and a count (`AR`). To take into account the nature of the outcomes, in this paper, we adopted three set of variance functions. To deal with continuous outcomes the power variance function $\vartheta(\cdot; p_r) = \mu_r^{p_r}$ provides a flexible family of models, since it describes the Tweedie family of distributions that has as special cases the Gaussian ($p = 0$), Gamma ($p = 2$) and inverse Gaussian ($p = 3$) distributions [29, 30, 28].

For handling binomial data we propose to use the extended binomial variance function defined by $\vartheta(\cdot; p_r) = \mu_r^{p_{r1}}(1-\mu_r)^{p_{r2}}$ where we introduced two new extra power parameters in order to become the orthodox binomial variance function more flexible. This variance function can also deal with bounded outcomes, as proportions or indexes. An interesting special case of the extended binomial variance function, is $p_1 = p_2 = 3$ that corresponding to the Simplex distribution [31]. However, we highlight that such variance function is not suitable for binary outcomes.

Finally, for modelling count outcomes we following Bonat et al. [32] adopted the Poisson-Tweedie dispersion function [33], i.e. $\vartheta(\cdot; p) = \mu + \tau\mu^p$, where τ is the dispersion parameter. We highlight that the dispersion function introduced by Jørgensen and Kokonendji [33] is not a variance function in the sense of Jørgensen [30], but for practical data analysis both are completely analogous. In that case the covariance within outcomes takes the special form,

$$\boldsymbol{\Sigma}_r = \mathrm{diag}(\boldsymbol{\mu}_r) + \mathrm{V}(\boldsymbol{\mu}_r; p_r)^{\frac{1}{2}}(\boldsymbol{\Omega}(\boldsymbol{\tau}_r))\mathrm{V}(\boldsymbol{\mu}_r; p_r)^{\frac{1}{2}},$$

where $\mathrm{V}(\boldsymbol{\mu}_r; p_r) = \mathrm{diag}(\boldsymbol{\mu}_r^{p_r})$, is a diagonal matrix whose main entries are given by the power variance function. The Poisson-Tweedie family of distributions provide a rich class of models to deal with count outcomes, since many important distributions appear as special cases, examples include the Hermite ($p = 0$), Neyman Type A ($p = 1$), negative binomial ($p = 2$) and Poisson-inverse Gaussian ($p = 3$).

The power parameter p plays an important role in the context of McGLMs for mixed types of outcomes, since for all variance functions discussed it is an index which distinguishes between important distributions. The algorithm proposed by Bonat and Jørgensen [16] and implemented in the `R` package `mcglm` [34] allows us to estimate the power parameter, which works as an automatic distribution selection.

The dispersion matrix $\boldsymbol{\Omega(\tau_r)}$ describes the part of the covariance within outcomes that does not depend on the mean structure. Based on the ideas of Anderson [35] and Pourahmadi [36] Bonat and Jørgensen [16] proposed to model the dispersion matrix using a matrix linear predictor combined with a covariance link function, i.e.

$$h(\boldsymbol{\Omega}(\boldsymbol{\tau}_r)) = \tau_{r0}Z_{r0} + \cdots + \tau_{rD}Z_{rD}, \tag{2}$$

where h is the covariance link function, Z_{rd} with $d = 0, \ldots, D$ are known matrices reflecting the covariance structure within the response variable r, and $\boldsymbol{\tau}_r = (\tau_{r0}, \ldots, \tau_{rD})$ is a $(D+1) \times 1$ vector of dispersion parameters.

McGLMs are fitted using an efficient Newton scoring algorithm based on quasi-likelihood and Pearson estimating functions, using only second-moment assumptions. The algorithm is described in detail by Bonat and Jørgensen [16] and implemented in the `mcglm` [34] package for the `R` statistical software. The data set and `R` code are available in the supplementary material.

3.1 Linear covariance models for genetic data

Hadfield and Nakagawa [7] showed that virtually all models used in the fields of quantitative genetic and phylogenetic in the Gaussian case are special specifications of the matrix linear predictor eq. (2) involving different types of known matrices, such as the additive genetic relatedness matrix [2, 37]. Furthermore, Demidenko [38] showed that the covariance structure induced by the orthodox Gaussian linear mixed effects model is a linear covariance model, using the identity covariance link function. In this sense, the models presented in this paper can be seen as an extension of the Gaussian linear mixed effects models for handling mixed types of outcomes.

Let $\boldsymbol{A}$ denote an additive genetic relatedness matrix. Such a matrix can be obtained for a given pedigree structure using for example the `R` package `nadiv` [39] through the function `makeA()`. Thus, the matrix linear predictor eq. (2) takes the form

$$\boldsymbol{\Omega}(\boldsymbol{\tau}_r) = \tau_{r0}\boldsymbol{I} + \tau_{r1}\boldsymbol{A}, \tag{3}$$

where $\boldsymbol{Z_{r0} = I}$ and $\boldsymbol{Z_{r1} = A}$. The dispersion parameters τ_{r0} and τ_{r1} measure the environmental and additive genetic effects, respectively. It is important to highlight that the relatedness matrix $\boldsymbol{A}$ is completely defined by the pedigree structure, i.e. it does not depend on any parameter. Further examples include the phylogenetic meta-analysis, taxonomic mixed model and the orthodox animal models [7].

In special for the house sparrow data set presented in Section 2, we have three outcomes, a continuous (`BD`), a binomial (`BS`) and a count (`AR`). For the outcome `BD` exploratory analysis showed that symmetry is a reasonable assumption, thus we assume the identity link function and constant variance function, which in turn correspond to assume the Gaussian distribution. Note that, these assumptions also correspond to use the Tweedie variance function with power parameter fixed at zero. Similarly, for the binomial outcome `BS` we adopt the logit link function and the extended binomial variance function. Finally, for the count outcome our exploratory analysis showed a possible excess of zero and overdispersion, consequently we adopt the standard logarithm link function and Poisson-Tweedie dispersion function.

For all outcomes the matrix linear predictor was specified as a linear combination of an identity matrix (environment effect) and a relatedness matrix $\boldsymbol{A}$ (genetic effect), see eq. (3). The identity covariance link function was adopted for all outcomes. In our study case, we have $R = 3$ outcomes thus the general model in eq. (1) takes the form,

$$\begin{aligned} \mathrm{E}(\mathbf{Y}) = \mathbf{M} &= \{\boldsymbol{X}_1\boldsymbol{\beta}_1, (1+\exp(-\boldsymbol{X}_2\boldsymbol{\beta}_2))^{-1}, \exp(\boldsymbol{X}_3\boldsymbol{\beta}_3)\} \\ \mathrm{Var}(\mathbf{Y}) = \boldsymbol{C} &= \mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1, \tilde{\boldsymbol{\Sigma}}_2, \tilde{\boldsymbol{\Sigma}}_3)(\boldsymbol{\Sigma}_b \otimes \boldsymbol{I})\mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1^T, \tilde{\boldsymbol{\Sigma}}_2^T, \tilde{\boldsymbol{\Sigma}}_3^T), \end{aligned} \tag{4}$$

where

$$\boldsymbol{\Sigma}_1 = \tau_{10}\boldsymbol{I} + \tau_{11}\boldsymbol{A},$$

$$\boldsymbol{\Sigma}_2 = \mathrm{diag}(\boldsymbol{\mu}_2^{p_{21}}(1-\boldsymbol{\mu}_2)^{p_{22}})^{\frac{1}{2}}(\tau_{20}\boldsymbol{I} + \tau_{21}\boldsymbol{A})\mathrm{diag}(\boldsymbol{\mu}_2^{p_{21}}(1-\boldsymbol{\mu}_2)^{p_{22}})^{\frac{1}{2}}$$

and

$$\boldsymbol{\Sigma}_3 = \mathrm{diag}(\boldsymbol{\mu}_3) + \mathrm{diag}(\boldsymbol{\mu}_3^{p_{31}})^{\frac{1}{2}}(\tau_{30}\boldsymbol{I} + \tau_{31}\boldsymbol{A})\mathrm{diag}(\boldsymbol{\mu}_3^{p_{31}})^{\frac{1}{2}}.$$

The design matrices $\boldsymbol{X}_1, \boldsymbol{X}_2$ and $\boldsymbol{X}_3$ are composed by the values of three covariates, `sex` a factor with two levels, hatch year (`HY`) that we considered as continuous and hatch island (`HI`) a factor with six levels. Moreover, the 3×3 correlation matrix between outcomes is given by,

$$\boldsymbol{\Sigma}_b = \begin{bmatrix} 1 & & \\ \rho_{12} & 1 & \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix},$$

where ρ_{12} denotes the correlation between `BD` and `BS`. Similarly, ρ_{13} and ρ_{23} denote the correlation between `BD` and `AR` and between `BS` and `AR` respectively.

Finally, from the specification of the matrix linear predictor emerges a natural marginal measure of heritability, defined for the component d by

$$h_{rd} = \tau_{rd} / \sum_{d=0}^{D} \tau_{rd}, \tag{5}$$

that is completely analogous to the definition of heritability in the Gaussian case and does not depend on the choice of the variance function. For the special case eq. (3) the heritability index is given by $h_{r1} = \tau_{r1}/(\tau_{r0} + \tau_{r1})$, which highlights that eq. (5) is a natural extension of the heritability index in the Gaussian case to the non-Gaussian case. Furthermore, the models can also accommodate other sources of dependence as repeated measures and longitudinal structure as well as effects of covariates in a linear mixed effects model fashion [32].

4 Conditional and marginal models: Explaining negative heritability index

McGLMs introduce the additive genetic effects directly in the marginal covariance matrix by using a linear combination of known matrices, see eq. (3). It is in contrast to the more orthodox approach based on hierarchical models, where the additive genetic effects are introduced through random effects using a conditional specification, see for example Sorensen and Gianola [2] and Hadfield [11]. In this section, we discuss conditional and marginal model specifications in the context of additive genetic models and explain why negative heritability index appears naturally in the marginal specification. Furthermore, we use this fact to explain how to assess the significance of the genetic effects. To simplify the discussion and without loss of generality consider the one outcome case, i.e $R = 1$. Thus, the outcome matrix takes the form $\mathbf{Y}_{N\times 1}$. An often approach to model genetic data is to assume a hierarchical model as follows:

$$\begin{aligned} \mathbf{Y}|\mathbf{Z} &\sim f(\cdot; \boldsymbol{\mu}, \tau_0) \\ g(\boldsymbol{\mu}) &= \boldsymbol{X\beta} + \mathbf{Z} \\ \mathbf{Z} &\sim f^*(\mathbf{0}, \tau_1 \boldsymbol{A}). \end{aligned} \tag{6}$$

It is assumed that the components of $\mathbf{Y}$ are conditionally independent given additive genetic effects $\mathbf{Z}$ and distributed as $f(\cdot; \boldsymbol{\mu}, \tau_0)$. The additive genetic effects are assumed f^* distributed, where f^* represents a multivariate distribution with vector of expected values $\mathbf{0}$ and covariance matrix $\tau_1 \boldsymbol{A}$. The parameters $\tau_0 > 0$ and $\tau_1 > 0$ are dispersion parameters associated with the environmental and genetic effects, respectively. The linear predictor is linked to the mean, $\boldsymbol{\mu}$ by a link function g and consists of the sum of fixed effects $\boldsymbol{X\beta}$ and additive genetic effects $\mathbf{Z}$. The $N \times k$ design matrix $\boldsymbol{X}$ contains values of k covariates and $\boldsymbol{\beta}$ is a $k \times 1$ vector of regression parameters. The class of generalized linear mixed models is obtained from eq. (4) by assuming that $f(\cdot; \boldsymbol{\mu}, \tau_0)$ belongs to the exponential family of distributions and $f^*(\mathbf{0}, \tau_1 \boldsymbol{A})$ is a multivariate Gaussian distribution. In this context, the parameters τ_0 and τ_1 are frequently called variance components.

The conditional specification eq. (6) makes clear that the dispersion parameters should be positive and provides a convenient interpretation of the statistical model components for quantitative genetic analysis. In this context, however, we are interested in testing whether or not the genetic additive effects are absent in the model, which in turn is equivalent to testing the dispersion parameter τ_1 equals to zero. Note that, such a null hypothesis places the dispersion parameter on the boundary of the parameter space and commonly used tests, such as the likelihood ratio, Wald and score tests, do not have the traditional chi-squared distribution [40]. In spite of corrections for the test statistics are available [42, 41], it remains an inconvenient feature of the conditional specification. On the other hand, as we shall discuss below such inconvenient feature can be overcome by a marginal specification. In general, the marginal distribution of $\mathbf{Y}$ cannot be obtained in closed-form from the conditional specification. Moreover, the model parameters should be carefully interpreted, since the covariate effects are conditional on the additive genetic effects, whereas the covariance structure is marginal for the random effects rather than for the outcome.

A notorious special case of eq. (6) is the Gaussian linear mixed effects models obtained by assuming that $f(\cdot; \boldsymbol{\mu}, \tau_0)$ is Gaussian distributed and g as the identity link function. In that case is straightforward to show that the marginal distribution of $\mathbf{Y}$ is Gaussian with mean and variance given by

$$\begin{aligned} \mathrm{E}(\mathbf{Y}) &= \boldsymbol{X\beta} \\ \mathrm{Var}(\mathbf{Y}) &= \boldsymbol{\Sigma} = \tau_0 \boldsymbol{I} + \tau_1 \boldsymbol{A}, \end{aligned} \tag{7}$$

where it is clear that the only request to obtain a valid multivariate Gaussian distribution is that the covariance matrix $\boldsymbol{\Sigma}$ be positive definite. The parameter space of the dispersion parameters in the marginal model is the set $\Theta = \{\tau_0, \tau_1 : \boldsymbol{\Sigma} > 0\}$, where $\boldsymbol{\Sigma} > 0$ means that $\boldsymbol{\Sigma}$ is a positive definite matrix. Furthermore, recall that a necessary condition for a matrix be positive definite is that the elements of its diagonal $\boldsymbol{\Sigma}_{ll} > 0, \forall l, l = 1, \ldots, N$. It implies that $\tau_0 \boldsymbol{I}_{ll} > -\tau_1 \boldsymbol{A}_{ll}$, which in turn shows that negative values at least at some extend are allowed for the parameter τ_1. It is important to highlight that the condition $\boldsymbol{\Sigma}_{ll} > 0$ is necessary, but not sufficient to $\boldsymbol{\Sigma} > 0$. The heritability index eq. (5) in that case takes the form $h_1 = \tau_1/(\tau_0 + \tau_1)$, which implies that when $\tau_1 < 0$ the heritability index is negative. In the marginal specification the parameter space is not trivially specified, but it has the advantage that the null hypotheses $\tau_1 = 0$ is not placed on the boundary of the parameter space and commonly used tests, such as the likelihood ratio, Wald and score tests can be applied without any additional corrections. In the cases where the additive genetic effects are non-significant, i.e. $\tau_1 = 0$ we expected that $\mathrm{E}(\hat{\tau}_1) = 0$, where $\hat{\tau}_1$ denotes an unbiased estimator of τ_1. However, in practical data analysis is perfectly possible to observe realized values of $\hat{\tau}_1$ smaller than zero, it is due to the sample variation, as usual when testing regression coefficients significance in the linear predictor. In the Section 6, as part of our simulation study, we assess the empirical distribution of $\hat{\tau}_1$ and show that under the null hypothesis $\tau_1 = 0$ a simple Wald test can be used to test the significance of the additive genetic effects.

The McGLMs extend the Gaussian marginal specification eq. (7) to deal with non-Gaussian data by introducing link, variance and covariance link functions. In the case of one outcome and using the identity covariance link function, the general model eq. (1) customized to take into account the additive genetic effects has the following form

$$\begin{aligned} \mathrm{E}(\mathbf{Y}) &= g^{-1}(\boldsymbol{X\beta}) \\ \mathrm{Var}(\mathbf{Y}) &= \boldsymbol{\Sigma} = \mathrm{V}(\boldsymbol{\mu}; p)^{\frac{1}{2}}(\tau_0 \boldsymbol{I} + \tau_1 \boldsymbol{A})\mathrm{V}(\boldsymbol{\mu}; p)^{\frac{1}{2}}. \end{aligned}$$

Hence, the terms of the diagonal matrix $\mathrm{V}(\boldsymbol{\mu}; p)$ are given by the variance function applied element wise to the vector $\boldsymbol{\mu}$, these values are always positive. Thus, the argument used to explain negative heritability index in the Gaussian case applies directly to the non-Gaussian case. Finally, by the properties of the generalized Kronecker product, if each marginal covariance matrix is positive definite, the joint covariance matrix is also positive definite. Finally, it is important to highlight that the McGLMs are specified based only on second-moments assumptions. Thus, the existence of a conditional model, a part of the Gaussian case, whose moments are given by eq. (1) remains an open question.

5 Simulation of McGLMs

In this section we adapt the NORTA algorithm for simulation of McGLMs with mixed types of outcomes. The NORTA algorithm described in detail by Cario and Nelson [23] is a simple, but still powerful method for simulation of random vectors with arbitrary marginal distributions and pre-specified correlation matrix. The key idea behind the NORTA method is to transform a multivariate Gaussian random vector into the desired random vector.

To simplify the notation, let $\mathcal{Y} = (\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_R^\top)^\top$ denote the $NR \times 1$ stacked vector of the outcome matrix $\mathbf{Y}_{N\times R}$ by columns. The goal is to simulate a random vector $\mathcal{Y}$ whose components $\{\mathcal{Y}_1, \ldots, \mathcal{Y}_{NR}\}$ have arbitrary marginal distributions denoted by $F_{\mathcal{Y}_l}$ for $l = 1, \ldots, NR$ and covariance matrix $\mathrm{Var}(\mathcal{Y}) = \boldsymbol{C}$. Here, $F_{\mathcal{Y}_l}$ denotes an arbitrary cumulative distribution function (cdf). Without loss of generality, suppose that $\boldsymbol{C}$ is a correlation matrix. Let $\mathcal{Y}$ represent a transformation of an NR-dimensional standard multivariate Gaussian vector $\mathcal{Z} = \{\mathcal{Z}_1, \ldots, \mathcal{Z}_{NR}\}^\top$ with correlation matrix $\boldsymbol{C}_{\mathcal{Z}}$. The NORTA vector is given by

$$\mathcal{Y} = \begin{pmatrix} F_{\mathcal{Y}_1}^{-1}(\Phi[\mathcal{Z}_1]) \\ \vdots \\ F_{\mathcal{Y}_{NR}}^{-1}(\Phi[\mathcal{Z}_{NR}]) \end{pmatrix}$$

where $\Phi[\cdot]$ is the univariate standard Gaussian cdf applied element-wise to the vector $\mathcal{Z}$ and $F_{\mathcal{Y}_l}^{-1}(u) \equiv \inf\{y : F_{\mathcal{Y}_l}(y) \geq u\}$ denotes the inverse cdf. As point out by Cario and Nelson [23] the transformation $F_{\mathcal{Y}_l}^{-1}(\Phi[\cdot])$ ensures that $\mathcal{Y}_l$ has the desired marginal distribution $F_{\mathcal{Y}_l}$. Thus, the central problem is to specify the correlation matrix $\boldsymbol{C}_{\mathcal{Z}}$ in a such way that the vector $\mathcal{Y}$ has the desired correlation matrix $\boldsymbol{C}$. Note that, the marginal variance can be controlled by the marginal distribution $F_{\mathcal{Y}_l}$, so the specification in terms of correlation matrix does not imply any loss of generality.

Let $\mathcal{Y}_l$ and $\mathcal{Y}_{l'}$ denote the elements l and l' of the vector $\mathcal{Y}$, respectively. For $l \neq l'$, let $\rho_{\mathcal{Z}}(l, l')$ be the $(l, l')^{th}$ element of $\boldsymbol{C}_{\mathcal{Z}}$, and let $\rho_{\mathcal{Y}}(l, l')$ be the $(l, l')^{th}$ element of $\boldsymbol{C}$. The correlation matrix of $\mathcal{Z}$ directly determines the correlation matrix of $\mathcal{Y}$, since

$$\rho_{\mathcal{Y}}(l, l') = \mathrm{Corr}(\mathcal{Y}_l, \mathcal{Y}_{l'}) = \mathrm{Corr}(F_{\mathcal{Y}_l}^{-1}(\Phi[\mathcal{Z}_l]), F_{\mathcal{Y}_{l'}}^{-1}(\Phi[\mathcal{Z}_{l'}]))$$

for all $l \neq l'$. The correlation is defined by

$$\text{Corr}(\mathcal{Y}_l, \mathcal{Y}_{l'}) = \frac{\text{E}(\mathcal{Y}_l\mathcal{Y}_{l'}) - \text{E}(\mathcal{Y}_l)\text{E}(\mathcal{Y}_{l'})}{\sqrt{\text{Var}(\mathcal{Y}_l)\text{Var}(\mathcal{Y}_{l'})}}, \tag{8}$$

where the marginal quantities $\text{E}(\mathcal{Y}_l)$, $\text{E}(\mathcal{Y}_{l'})$, $\text{Var}(\mathcal{Y}_l)$ and $\text{Var}(\mathcal{Y}_{l'})$ are defined by $F_{\mathcal{Y}_l}$ and $F_{\mathcal{Y}_{l'}}$. The idea is to adjust the correlation matrix $\boldsymbol{C}_{\mathcal{Z}}$ to obtain the desired correlation matrix $\boldsymbol{C}$ of $\mathcal{Y}$. In eq. (8) is clear that we can keep attention on the $\text{E}(\mathcal{Y}_l\mathcal{Y}_{l'})$ term. Recall that $(\mathcal{Z}_l, \mathcal{Z}_{l'})$ has a standard bivariate Gaussian distribution. Thus,

$$\begin{aligned}\text{E}(\mathcal{Y}_l\mathcal{Y}_{l'}) &= E\left(F^{-1}_{\mathcal{Y}_l}(\Phi[\mathcal{Z}_l])F^{-1}_{\mathcal{Y}_{l'}}(\Phi[\mathcal{Z}_{l'}])\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F^{-1}_{\mathcal{Y}_l}(\Phi[z_l])F^{-1}_{\mathcal{Y}_l}(\Phi[z_{l'}])\Phi(z_l, z_{l'})dz_l dz_{l'}\end{aligned} \tag{9}$$

where $\Phi(z_l, z_{l'})$ denotes a standard bivariate Gaussian probability density function with correlation $\rho_{\mathcal{Z}}(l, l')$. We restrict our attention in distributions for which the integral eq. (9) exists. In general the integral eq. (9) has no closed form, thus we adopted the Retrospective Approximation algorithm [43] implemented in the `R` package `NORTARA` [44]. Note that, the problem of determining $\boldsymbol{C}_{\mathcal{Z}}$ that provides the desired correlation matrix $\boldsymbol{C}$ for $\mathcal{Y}$ reduces to $NR(NR-1)/2$ independent problems. For large data sets the problem becomes time consuming. However, we noted that for many situations the correlation $\rho_{\mathcal{Z}}(l, l')$ is not far away of the correlation $\rho_{\mathcal{Y}}(l, l')$. Thus, we can solve the integral eq. (9) for a grid of values of $\rho_{\mathcal{Y}}(l, l')$ and apply a linear regression model to interpolate all values of $\rho_{\mathcal{Y}}(l, l')$ required to build the matrix $\boldsymbol{C}$.

6 Simulation study

In this section we present a simulation study to verify the properties of the estimating function estimators. We considered a mixed types of outcomes scenario where a combination of Gaussian, binomial and Poisson outcomes is observed. The matrix linear predictor for each outcome was specified by a linear combination of an identity matrix and a relatedness matrix defined by a pedigree structure, see eq. (3). The `R` packages `pedigree`[45] and `nadiv` were employed to simulate the pedigree structure and build the relatedness matrix, respectively. The simulated pedigree consists of ten independent families, each one with five generations and an increasing number of individuals. We considered 5, 10 and 20 individuals per family, resulting in sample of sizes 250, 500 and 1000.

In this simulation study, we focus on the estimation of the correlation between outcomes and the new heritability index. We designed five simulation scenarios in order to explore the range of values of the heritability index. The first scenario considered a negative heritabililty index, thus we fixed the dispersion parameter values at ($\tau_{r0} = 1.2, \tau_{r1} = -0.1$) which implies that $h_{r1} = -0.090$. Since, the values of the main diagonal of the relatedness matrix $\boldsymbol{A}$ are always positive and in general close to 1 more extreme negative values will generate improper models, i.e. $\boldsymbol{\Sigma}$ is not positive definite. It is important to highlight that this scenario is challenge, since the dispersion parameter τ_{r1} is close to the boundary of the parameter space. The second scenario considered the case of non-significant additive genetic effects, i.e. ($\tau_{r0} = 1, \tau_{r1} = 0$) which implies that the heritability index is zero. This scenario is interesting, since it allows us to study the empirical distribution of h_{r1} under the null hypotheses $\tau_{r1} = 0$, see the discussion in Section 4. Finally, for the other three scenarios, we considered the parameters ($\tau_{r0} = 0.75, \tau_{r1} = 0.25$), ($\tau_{r0} = 0.5, \tau_{r1} = 0.5$) and ($\tau_{r0} = 0.25, \tau_{r1} = 0.75$), respectively. Thus, the heritability index presents the values 0.25, 0.5 and 0.75, respectively.

For all simulation scenarios, the correlations between outcomes were specified as $\rho_{12} = 0.8$, $\rho_{13} = 0.6$ and $\rho_{23} = 0.4$, where ρ_{12} denotes the correlation between the Gaussian and binomial marginals, similarly ρ_{13} and ρ_{23} denote the correlation between the Gaussian and Poisson marginals and between the binomial and Poisson marginals, respectively.

The Gaussian, binomial and Poisson scenarios consider linear predictors with intercepts and slopes given by: ($\beta_{10} = 10, \beta_{11} = 2$), ($\beta_{20} = 5, \beta_{21} = 0.2$) and ($\beta_{30} = 0.5, \beta_{31} = 0.6$). The covariate is a sequence from −2 to 2 and length equals the sample size. The values of the regression coefficients and the covariate were chosen in order to explore a large part of the support of each marginal distribution.

The Gaussian marginal was modelled using the identity link function and constant variance function. For the binomial marginal we adopted the logit link function and the orthodox binomial variance function. Finally, the Poisson marginal was modelled using the log link function and the Tweedie variance function with power parameter fixed at 1. For all outcomes the identity covariance link function was adopted.

For each simulation scenario, we generated 500 data sets to evaluate the bias, consistency and coverage rate of the estimating function estimators. Figure 2 shows the average bias plus and minus the average standard error for the correlation and heritability parameter estimators for each scenario. The scales are standardized for each parameter dividing the average bias and the limits of the confidence intervals by the standard error obtained on the sample of size 250.

Figure 2 shows that for the scenarios 2 to 5 the estimating function estimators for the correlation parameters between outcomes are unbiased and consistent. For the scenario 1 the correlation between the Gaussian and binomial marginals (ρ_{12}) is underestimated. Although, the bias is small in its magnitude and decreases while the sample size increases.

The heritability index is underestimated in the scenarios 2 to 5 and small sample size. However, in general the bias decreases to zero while the sample size increases, as expected. On the other hand, in the scenario 1 the heritability index is well estimated for small samples, but for large sample the bias increases. The standard errors for all estimators decrease while the sample size increases, showing the consistency of the estimating function estimators.

In order to further investigate the origin of the underestimation of the heritability index Figure 3 presents the average bias plus and minus the average standard error on standardized scale for the dispersion parameter estimators for each simulation scenario.

The results presented in Figure 3 show that in the scenarios 3 to 5 the dispersion parameters associated with the environmental effects are overestimated for small samples, while the dispersion parameters associated with the additive genetic effects are underestimated. These results explain the underestimation of the heritability index detected in the Figure 2. The scenario 1 presents a different situation, where the dispersion parameters associated with the environmental effects are slightly underestimated and the ones associated with the genetic effects are slightly overestimated. Similarly, we have seen in Figure 2 the bias associated with the parameters τ_{11}, τ_{21} and τ_{31} increases for large samples in the scenario 1.

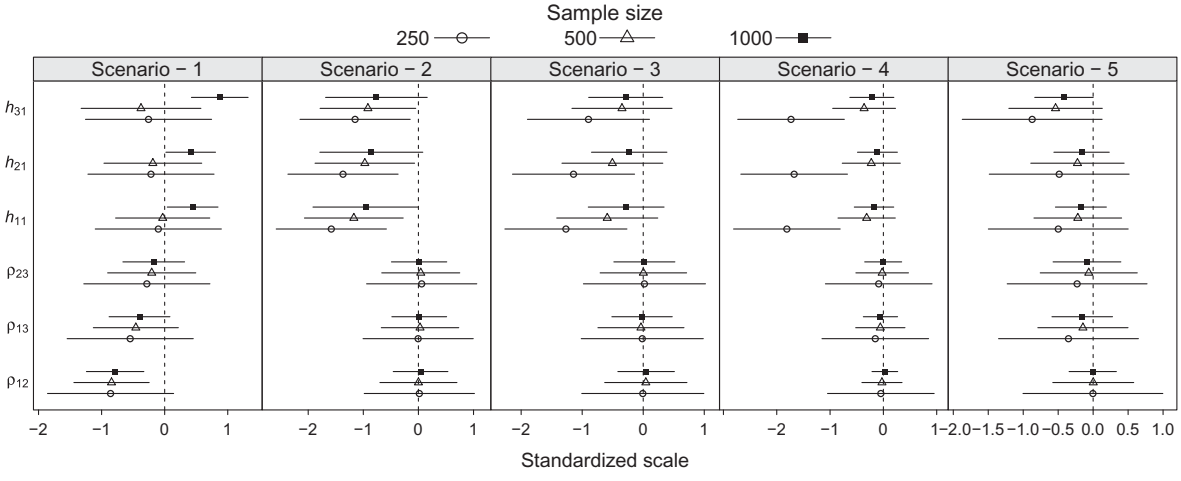

Figure 2: Average bias and confidence interval on a standardized scale by scenarios and sample sizes for the correlation and heritability parameter estimators.

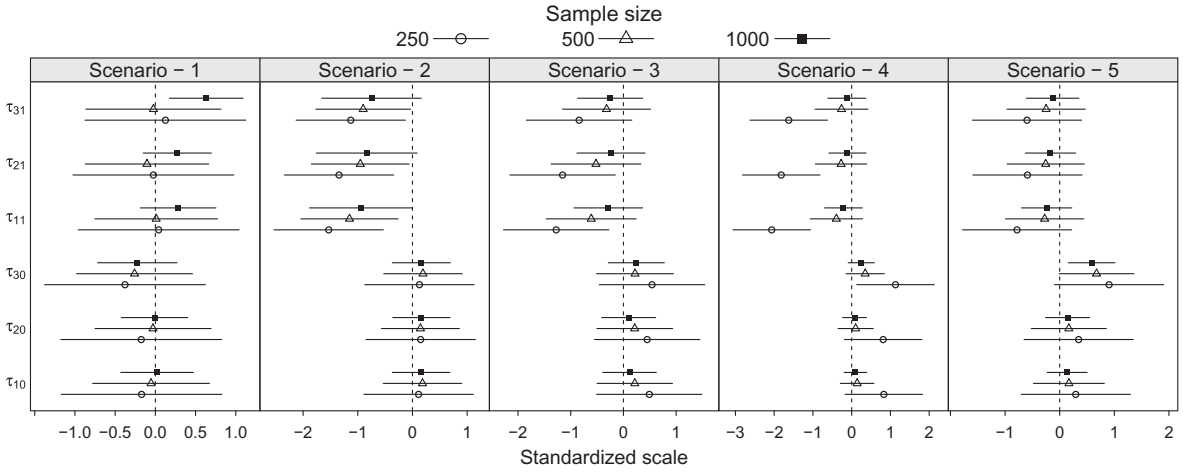

Figure 3: Average bias and confidence interval on a standardized scale by scenarios and sample sizes for the dispersion parameter estimators.

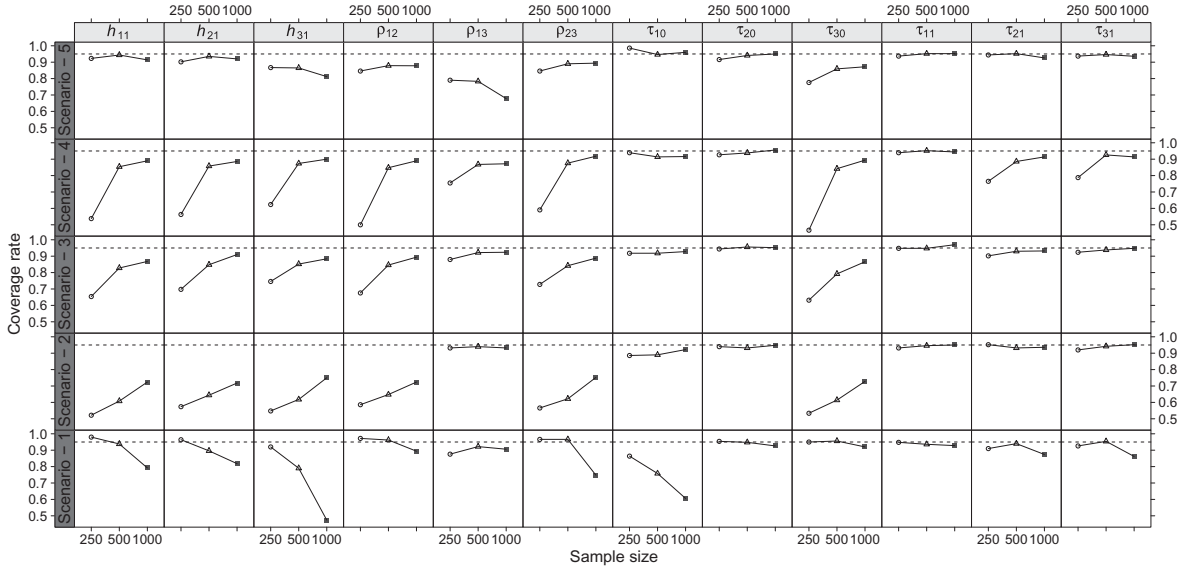

Figure 4: Coverage rate by scenarios and sample sizes for the heritability, correlation and dispersion parameter estimators.

The simulation scenario 2 shows that the dispersion parameters associated with the additive genetic effects are underestimated, even for large samples. This result shows that in practice when the additive genetic effects are non-significant, i.e. $\tau_{r1} = 0$ we still can observe negative values for $\hat{\tau}_{r1}$ and consequently for the heritability index. Such results should be interpreted carefully and accompanied of the standard error associated with both the dispersion and heritability parameter estimators. The estimating function approach used to fit McGLMs provides asymptotic standard errors that can be used to construct confidence intervals as well as Wald statistics test. To check the accuracy of such standard errors Figure 4 presents the coverage rate of the confidence intervals for the heritability, correlation and dispersion parameters. The nominal level of significance was fixed at 95%.

The empirical coverage rate presented values close to the nominal level for scenarios 3 to 5 and large samples for all parameters. In general, for small sample sizes the coverage rate is below the nominal level. The worst results appear for the heritability index in the scenario 2. It is due to the underestimation detected

in the Figure 2. Note, however, that the coverage rate for the dispersion parameters associated with the additive genetic effects is in general close to the nominal level, even for small sample sizes. It indicates that the Wald test can be used to test the null hypothesis $\tau_{r1} = 0$. In the scenario 1 the coverage rate decreases while the sample size increases. This result is compatible with the increase of the bias detected in Figures 2 and 3 and highlights the difficult to estimate when parameter values are close to the parameter space boundary.

7 Data analysis

In this section, we apply the McGLMs for mixed types of outcomes to analyse the house sparrow data set presented in Section 2. The linear predictor for each outcome was specified using the three covariates available. We adopted the identity, logit and log link functions along with constant, extended binomial and Poisson-Tweedie variance/dispersion functions for the outcomes `BD`, `BS` and `AR`, respectively.

The matrix linear predictor for each outcome is composed of an identity matrix combined with a relatedness matrix, representing the environmental and genetic structures respectively, see Section 3.1 for detail. The `R` package `nadiv` was employed for building the relatedness matrix based on the pedigree structure available for the house sparrow data set. As discussed in Section 3.1 we adopted the identity covariance link function.

First, we investigated the estimation of the extended binomial variance function for the outcome `BS`. We considered three special cases: Case 1 corresponds to the orthodox binomial variance function, i.e. $p_1 = p_2 = 1$. Case 2 estimates p_1 with the restriction $p_1 = p_2$. Finally, in the case 3 both p_1 and p_2 are estimated. Table 1 presents estimates and standard errors for the power and dispersion parameters for each case.

The results in Table 1 show that the model in the case 3 is clearly over parametrized. Moreover, in the case 2 the estimate of the power parameter is not different of 1, i.e. the asymptotic confidence interval $\hat{p}_1 \pm Z_{(1-\alpha)/2}SE(\hat{p}_1)$, where $Z_{(1-\alpha)/2}$ denotes the $(1-\alpha)/2$ critical value of the standard normal distribution, contains the value 1. Equivalently, the Wald test can be used as an approximation to the likelihood ratio test. For example, for choosing between the models presented as cases 1 and 2 in Table 1, the hypotheses are $H_0 : p_1 = 1$ against $H_1 : p_1 \neq 1$ whose Z statistics is $\hat{p}_1 - 1/SE(\hat{p}_1)$. In that case the value of the Z statistics is 1.546 which shows that the model 2 does not differ from the model 1. Thus, we choose the simplest one, i.e. model 1. Similar procedure can be used to choose between the models 1 and 3 and the results show that model 1 is the best choice for the house sparrow data set. From now on, by non-significant we means that the asymptotic confidence interval contains the value zero. The confidence level is fixed at 95%.

In all cases presented in Table 1 the dispersion parameters associated with the genetic structure (τ_1) were negative, but non-significant. Table 2 presents estimates and standard errors for the joint McGLMs fitted to the house sparrow data set. The heritability index for each outcome was computed using eq. (5). To verify the effect of consider the correlation between outcomes in the fitted model, we also present estimates and standard errors obtained by fitting univariate McGLMs for each outcome.

Table 2 shows that the genetic structure has a significant effect only for the outcome `BD`. In that case, the heritability index is different from zero, as expected. On the other hand, for the outcomes `BS` and `AR` the heritability index was negative, but non-significant.

Table 1: Power and dispersion parameter estimates and standard errors (SE) for each special case of the extended binomial variance function for the outcome `BS`.

			Estimates (SE)
	Case 1	**Case** 2	**Case** 3
p_1	1	1.439(0.284)	3.195(1.350)
p_2	–	–	7.609(4.471)
τ_0	1.048(0.071)	2.288(1.143)	159.719(508.685)
τ_1	−0.020(0.052)	−0.042(0.115)	−2.880(12.028)

Table 2: Power, dispersion and heritability parameter estimates and standard errors (SE) by models fitted to the house sparrow data set.

				Estimates (SE)
	BD	BS	AR	Joint
τ_{10}	0.035(0.005)	–	–	0.035(0.005)
τ_{11}	0.028(0.005)	–	–	0.028(0.005)
τ_{20}	–	1.048(0.071)	–	1.067(0.064)
τ_{21}	–	−0.020(0.052)	–	−0.032(0.047)
τ_{30}	–	–	0.919(0.186)	0.799(0.139)
τ_{31}	–	–	−0.070(0.101)	−0.099(0.076)
p_{31}	–	–	1.905(0.353)	1.545(0.460)
h_{11}	0.447(0.077)	–	–	0.441(0.078)
h_{21}	–	−0.019(0.050)	–	−0.031(0.045)
h_{31}	–	–	−0.083(0.121)	−0.141(0.122)

In general the univariate and multivariate models provide the same interpretation in terms of genetic effects. Differences appear in the size of the standard errors associated with the dispersion parameters and consequently in the standard errors of the heritability index. On average the standard errors from the joint model are 11.170% and 34.849% smaller than the ones from the univariate models, for the outcomes BS and AR, respectively. Regarding the outcome BD both models provide virtually the same estimates and standard errors.

The differences may be explained by the correlation between outcomes, whose estimates and standard errors were as follows:

$$\hat{\boldsymbol{\Sigma}}_b = \begin{bmatrix} 1 & & \\ -0.005(0.032) & 1 & \\ 0.010(0.032) & 0.872(0.034) & 1 \end{bmatrix}.$$

The correlation between BD and BS (−0.005) and BD and AR (0.010) were non-significant. These results show that the outcomes BS and AR share no information with the outcome BD, thus the estimates from the univariate and multivariate models are virtually the same for the outcome BD. However, the correlation between BS and AR (0.8722) was significant, explaining why the standard errors from the multivariate model were smaller than the ones from the univariate models. As expected the correlation between outcomes improves the efficiency of the dispersion parameter estimators. The estimate of the power parameter suggests that the Pólya Aeppli ($p = 1.5$) distribution is a suitable choice for the outcome AR. In order to compare the regression coefficients Figure 5 shows the regression estimates and confidence intervals obtained from the univariate and multivariate models. The intercept is not shown to avoid scale issues.

Figure 5 shows that for the outcome BD the confidence intervals from the univariate and multivariate models are really similar, in fact the difference on average is 0.015%. Regarding the outcomes AR and BS on average the confidence intervals from the multivariate model are 11.086% and 2.483% smaller than the ones from the univariate models, respectively.

Holand et al. [8] results showed that the heritability for BD, BS and AR were 0.35 (SE= 0.07), 0.04 (SE = 0.01) and 0.03 (SE = 0.01), respectively. Based on the DIC values the authors concluded that the best model for the outcome BS does not include linear additive genetic effects. For the outcome AR the models without and with additive genetic effects were very close in DIC, with preference for the last one.

The results presented in Table 2 show that the heritability index for the outcome BD was 0.447 (SE = 0.077). The difference between the two approaches can be due to the priory distribution requested in the context of Bayesian inference for the variance components. The priory distribution adopted by Holand et al. [8] was an inverse gamma, which in turn can bias downwards the heritability index. Regarding the outcomes BS and AR the estimates of the heritability are not comparable, since the models fitted in Holand et al. [8] were

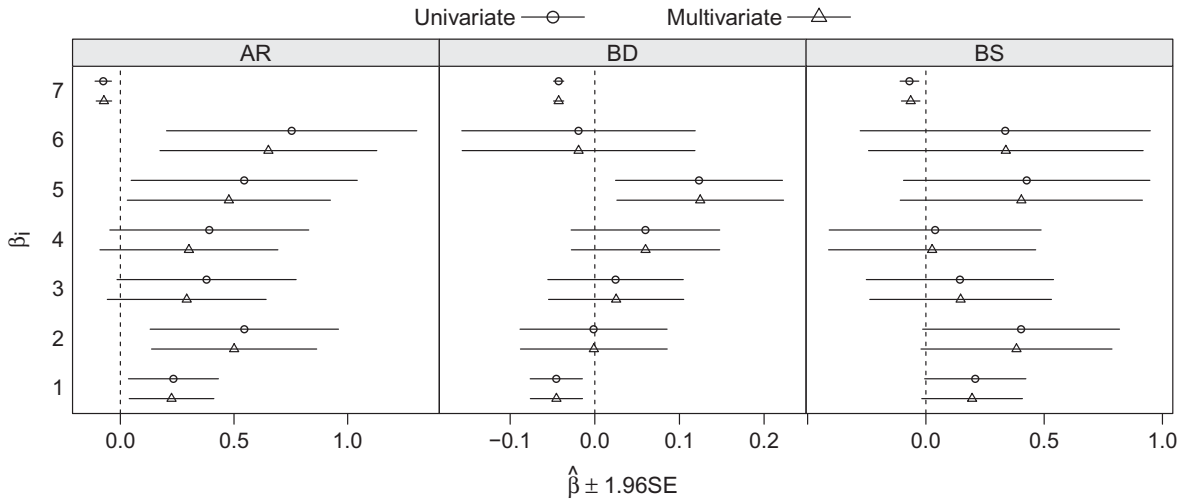

Figure 5: Regression parameter estimates and 95% confidence intervals by outcomes and models for the house sparrow data.

based on a conditional specification. However, the two approaches agree that for the outcome `BS` the additive genetic effects are not significant. Furthermore, for the outcome `AR` the McGLM approach shows clearly that the additive genetic effects are not significant. On the other hand, in Holand et al. [8] the model including the additive genetic effects shows a fit slightly better than the one without the additive genetic effects. However, as point out by Holand et al. [8] for small values of τ_1 they observed systematic errors and priory sensitivity, mainly when using Poisson and binomial likelihood, which can explain the weak importance of the additive genetic effects, detected by their model.

8 Discussion

We presented a general statistical modelling framework for analysis of mixed types of outcomes in the context of genetic data. The models were motivated by the house sparrow data set, where a mixed of continuous, binomial and count outcomes are of interest. Furthermore, we adapted the NORTA algorithm for simulation of McGLMs and proposed a new index of heritability.

The marginal covariance structure within outcomes were modelled by means of a variance function and a matrix linear predictor combined with a covariance link function. In this paper, we focused on the identity covariance link function, since many important models used in quantitative genetic analysis appear as special cases. The variance function plays an important role in our framework, since it allows us to deal with different types of outcomes in a unified way. Furthermore, the estimation of the power parameter provides extra flexibility for our models. The linear structure of the matrix linear predictor allows us to accommodate the genetic structure as well as extend the model to deal with repeated measures, longitudinal, spatial and covariates effects. Moreover, in the context of genetic data the specification of the matrix linear predictor provides an intuitive way to define a measure of heritability for non-Gaussian data. Finally, the joint covariance matrix is specified using the generalized Kronecker product, allowing to compute the correlation between outcomes.

The second-moment assumptions allow us to adopt an estimation function approach for parameter estimation and inference. The advantages of this approach are that the estimation procedure relies on a relatively simple and efficient Newton scoring algorithm and keep the traditional *population average* interpretation for both regression and dispersion parameters. The simulation study presented in Section 5 showed that in general the estimating function estimators are unbiased and consistent for large samples. The simulation study also showed that for small sample size the heritability index can be strongly underestimate mainly

when the parameters are close to the border of the parameter space, as appear in the scenarios 1 and 2 of our simulation study. The negative heritability index in the scenario 1 is unrealistic in terms of genetic interpretation, but in practical data analysis, such negative heritability values can appear as for example in the house sparrow data set. In general, it is associated with the non-significance of the additive genetic effects. It is also important to highlight that our estimation and inference procedure provides a Wald-type test that can be used to assess the significance of the genetic structure.

Regarding the data analysis, our results showed a significant genetic effect for the outcome `BD` and non-significant genetic effect for the outcomes `BS` and `AR`. These results agree with previous results obtained by Holand et al. [8], although these authors found a weak but still significant genetic effect for the outcome `AR`. The main advantages to fit a multivariate model are the possibility to compute the correlation matrix between outcomes and improve the estimation of the regression and dispersion parameters. As shown in our data analysis, in general the correlation between outcomes implies smaller standard errors for the regression and dispersion parameters.

Possible topics for further investigation and extensions include designing simulation studies to explore in detail the effect of different number of generations and families in the pedigree structure. An important aspect of the presented framework is the possibility to obtain negative heritability index. Thus, an interesting topic is to investigate from a genetic viewpoint what such negative values represent in practical data analyses. In the genetic context the genetic and environment correlations are two measures of interest. However, in the framework presented in this paper, we can assess only the correlation between outcomes (phenotypic correlation), i.e. we cannot distinguish between the genetic and environment correlations. Thus, a topic for future investigation is to propose marginal models able to distinguish between these two sources of correlation. In terms of modelling framework is interesting to investigate the performance of the McGLMs for the analysis of multivariate binary traits. Furthermore, theoretical and computational developments are required in order to construct new estimating functions to handle data not missing at random and make prediction using the Best Linear Unbiased Predictor (BLUP).

Acknowledgment: For Bent Jørgensen (1954–2015) in memoriam. The author is supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) - Brazil.

References

1. Henderson C. Estimation of genetic parameters. Ann Math Stat 1950;21:309–210.
2. Sorensen D, Gianola D. Likelihood, Bayesian, and MCMC methods in quantitative genetics. New York: Statistics for Biology and Health, Springer, 2007.
3. Dempster ER, Lerner IM. Heritability of threshold characters. Genetics 1950;35:212–236.
4. Hill WG. Understanding and using quantitative genetic variation. Philos Trans R Soc London, Ser B 2009;365:73–85.
5. Jensen H, Steinsland I, Ringsby TH, Sæther B-E. Evolutionary dynamics of a sexual ornament in the house sparrow *(Passer domesticus)*: the role of indirect selection within and beween sexes. Evolution 2008;62:1275–1293.
6. Gianola D, Fernando RL. Bayesian methods in animal breeding theory. J Anim Sci 1986;63:217–244.
7. Hadfield JD, Nakagawa S. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. J Evol Biol 2010;23:494–508.
8. Holand AM, Steinsland I, Martino S, Jensen H. Animal models and integrated nested Laplace approximations. G3: Genes, Genome, Genetics, 2013;3:1241–1251.
9. McCulloch CE. Maximum likelihood algorithms for generalized linear mixed models. J Am Stat Assoc 1997;92:162–170.
10. Fong Y, Rue H, Wakefield J. Bayesian inference for generalized linear mixed models. Biostatistics 2010;11:397–412.
11. Hadfield JD. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. J Stat Software 2010;33:1–22.
12. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc, Ser B 2009;71:319–392.
13. Liu J, Huang J, Ma S. Penalized multivariate linear mixed model for longitudinal genome-wide association studies. BMC Proc 2014;8:1–4.

14. Liu J, Yang C, Shi X, Li C, Huang J, Zhao H, Ma S. Analyzing association mapping in pedigree-based gwas using a penalized multitrait mixed model. Genet Epidemiol 2016;40:382–393.
15. de Villemereuil P, Schielzeth H, Nakagawa S, Morrissey M. General methods for evolutionary quantitative genetic inference from generalised mixed models. Genetics 2016;204(3):1281–1294.
16. Bonat WH, Jørgensen B. Multivariate covariance generalized linear models. J R Stat Soc, Ser C 2016;65:649–675.
17. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.
18. Carey VJ. gee: generalized estimation equation solver, 2015, http://CRAN.R-project.org/package=gee, r package version 4.13-19.
19. Højsgaard S, Halekoh U, Yan J. The R package geepack for generalized estimating equations. J Stat Software 2006;15:1–11.
20. Jørgensen B, Knudsen SJ. Parameter orthogonality and bias adjustment for estimating functions. Scand J Stat 2004;31, 93–114.
21. Hall DB, Severini TA. Extended generalized estimating equations for clustered data. J Am Stat Assoc 1998;93, 1365–1375.
22. Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. Analysis of longitudinal data. Oxford: Oxford Statistical Science Series, 2002.
23. Cario MC, Nelson BL. Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix, Technical report, Northwestern University, 1997.
24. R Core Team. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2015, ISBN 3-900051-07-0.
25. Pärn H, Jensen H, Ringsby TH, Sæther B-E. Sex-specific fitness correlates of dispersal in a house sparrow metapopulation. J An Ecol 2009;78:1216–1225.
26. Ringsby TH, Sæther B-E, Tufto J, Jensen H, Solberg EJ. Asynchronous spatiotemporal demography of a house sparrow metapopulation in a correlated environment. Ecology 2002;83:561–569.
27. Martinez-Beneito MA. A general modelling framework for multivariate disease mapping. Biometrika 2013;100:539–553.
28. Bonat WH, Kokonendji CC. Flexible Tweedie regression models for continuous data. J Stat Comput Simul 2017;87:2138–2152.
29. Jørgensen B. Exponential dispersion models. J R Stat Soc, Ser B 1987:49:127–162.
30. Jørgensen B. The theory of dispersion models. London: Chapman & Hall, 1997.
31. Barndorff-Nielsen O, Jørgensen B. Some parametric models on the simplex. J Multivariate Anal, 1991;39:106–116.
32. Bonat WH, Olivero J, Grande-Verga M, Fárfan MA, Fa JE. Modelling the covariance structure in marginal multivariate count models. J Agric Biol Environ Stat 2017;1–19.
33. Jørgensen B, Kokonendji C. Discrete dispersion models and their Tweedie asymptotics. AStA Adv Stat Anal 2015;100:43–78.
34. Bonat WH. mcglm: multivariate covariance generalized linear models, 2016. http://git.leg.ufpr.br/wbonat/mcglm, R package version 0.3.0.
35. Anderson TW. Asymptotically efficient estimation of covariance matrices with linear structure. The Ann Stat 1973;1:135–141.
36. Pourahmadi M. Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. Biometrika 2000;87:425–435.
37. Lynch M, Walsh B. Genetics and analysis of quantitative traits. Oxford: Sinauer, 1998.
38. Demidenko E. Mixed models: theory and applications with R. New Jersey: Wiley, 2013.
39. Wolak ME. nadiv: an R package to create relatedness matrices for estimating non-additive genetic variances in animal models. Meth Ecol Evol 2012;3:792–796.
40. Self SG, Liang K-Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc 1987;82:605–610.
41. Crainiceanu CM, Ruppert D. Likelihood ratio tests in linear mixed models with one variance component. J R Stat Soc, Ser B 2004;66:165–185.
42. Stram DO, Lee JW. Variance components testing in the longitudinal mixed effects model. Biometrics 1994;50:1171–1177.
43. Huifen C. Initialization for norta: generation of random vectors with specified marginals and correlations. INFORMS J Comput 2001;13:312–331.
44. Su P. NORTARA: generation of multivariate data with arbitrary marginals, 2014, R package version 1.0.0.
45. Coster A. pedigree: pedigree functions, 2013, R package version 1.4.

Supplemental Material: The R code and data set are available in the supplementary material website. http://www.leg.ufpr.br/doku.php/publications:papercompanions:biometrics