A unified method for multivariate mixed-type response regression

Karl Oskar Ekvall
Division of Biostatistics
Institute of Environmental Medicine
Karolinska Institute
karl.oskar.ekvall@ki.se

Aaron J. Molstad
Department of Statistics
Genetics Institute
University of Florida
amolstad@ufl.edu

Abstract

We propose a new method for multivariate response regressions where the elements of the response vector can be of mixed types, for example some continuous and some discrete. Our method is based on a model which assumes the observable mixed-type response vector is connected to a latent multivariate normal response linear regression through a link function. We explore the properties of this model and show its parameters are identifiable under reasonable conditions. We propose an algorithm for approximate maximum likelihood estimation that works "off-the-shelf" with many different combinations of response types, and which scales well in the dimension of the response vector. Our method typically gives better predictions and parameter estimates than fitting separate models for the different response types and allows for approximate likelihood ratio testing of relevant hypotheses such as independence of responses. The usefulness of the proposed method is illustrated using simulations and through three data examples.

1 Introduction

In many regression applications, there are multiple response variables of mixed types: some are continuous and some discrete, some are necessarily positive and some need not be, etc. Joint modeling of the responses can lead to more efficient estimation, better prediction, and allows for the testing of joint hypotheses without the need for multiple testing corrections. Popular regression models, however, typically assume all responses are of the same type. For example, the classical multivariate normal linear regression model assumes responses are conditionally multivariate normally distributed given the predictors. Modeling mixed-type responses is more complicated and over the years a substantial effort has been made to address this. Many methods for specific combinations of response types, such as some Bernoulli and some normal, some Poisson and some normal, and so on, have been proposed (Olkin and Tate, 1961; Poon and Lee, 1987; Catalano and Ryan, 1992; Cox and Wermuth, 1992; Fitzmaurice and Laird, 1995; Catalano, 1997; Fitzmaurice and Laird, 1997; Gueorguieva and Agresti, 2001; de Leon, 2005; Gueorguieva and Sanacora, 2006; Yang et al., 2007; Faes et al., 2008). Models that allow for many different response distributions have also been considered (Sammel et al., 1997; Dunson, 2000; Gueorguieva, 2001; Rabe-Hesketh et al., 2004; de Leon and Carrièger, 2007; Goldstein et al., 2009; Bonat and Jørgensen, 2016; Ekvall and Jones, 2020; Bai et al., 2020; Kang et al., 2020).

Existing methods for mixed-type responses typically either (i) assume dependence between responses can be parsimoniously parameterized, (ii) can be prohibitively time consuming to fit unless there are very few dependent responses, or (iii) are fit using algorithms that require substantial user modification depending on which types of responses are modeled. For example, multivariate generalized linear mixed models assume dependence between mixed-type responses is due to random effects. Usually, the random effects are fewer than the responses and have a distribution depending on a relatively small number of parameters. Such parsimonious structures can sometimes be motivated by subject-specific knowledge or the sampling design, and they are convenient from a computational perspective since estimates can be obtained by unconstrained maximization of a likelihood (approximation) using off-the-shelf solvers. However, they also lead to restrictions on the joint distribution of the responses that can be difficult to fully appreciate and which complicate inference. For example, the strength of dependence between observations is often determined by the same parameters determining marginal means and variances, and it is often not clear which parameters are identifiable (Jiang, 2007; Lele et al., 2010).

To address these issues, we propose a method for multivariate mixed-type response regressions which

1. gives better predictions and estimates than fitting separate models for the different response types in a wide range of settings,
2. allows for the testing of relevant joint hypotheses while avoiding the multiple testing burden,
3. works with many different combinations of mixed-type responses off-the-shelf, and
4. is fast enough to be practically useful.

Our method is based on a model which assumes a latent multivariate linear regression is connected to observable responses through a link function. Specifically, a response vector $Y \in \mathbb{R}^r$, non-stochastic predictor vector $x \in \mathbb{R}^p$, and latent vector $W \in \mathbb{R}^r$ satisfy, for some $\mathcal{B} \in \mathbb{R}^{p \times r}$,

$$g\{\mathbb{E}(Y \mid W)\} = W \text{ and } W = \mathcal{B}^{\mathsf{T}} x + \varepsilon, \tag{1}$$

where $\varepsilon \sim \mathcal{N}(0, \Sigma)$ and $g : \mathbb{R}^r \to \mathbb{R}^r$ is a known, injective link function. The elements of Y are conditionally uncorrelated given W, with conditional variances that can depend on W in a way to be specified. Our interest is in inference on $(\mathcal{B}, \Sigma)$ and the prediction of new responses, using n independent observations $(y_i, x_i) \in \mathbb{R}^r \times \mathbb{R}^p$, $i = 1, \ldots, n$, from (1). We do not specify exactly how n, r, and p compare but have in mind settings where n is substantially larger than r and at least as large as p. Some intuition for the model can be gained by noting that the multivariate normal linear regression model is a special case of (1) where $Y = W$ and, hence, g is the identity. Here, however, interest is primarily in cases where g is not the identity and $\mathrm{cov}(Y \mid W)$ is positive definite; in particular, W is truly latent and not a deterministic function of Y.

In general, our methods only require specifying the first two moments of $Y \mid W$, but we will also discuss fully specified parametric models consistent with (1). For example, a multivariate generalized linear mixed model where each response Y_j has its own random intercept ε_j and linear predictor $W_j = \mathcal{B}_j^{\mathsf{T}} x + \varepsilon_j$, where $\mathcal{B}_j$ is the jth column of $\mathcal{B}$, satisfies (1). McCulloch (2008) studied special cases of that model and noted its potential for modeling mixed-type responses, and similar models have been considered for repeated measures (dependent) data (Jaffa et al., 2016). Despite the mathematical connection, there are fundamental differences between our setting and those where mixed models are typically considered. In particular, in our setting the number of latent variables is the same as the number

of responses (r), their covariance matrix is unstructured, and we have n independent observations of a response vector and its corresponding predictors. A particularly useful property of our parameterization, which we discuss in Section 2, is that off-diagonal elements of Σ affect the covariances of responses but not their means or variances. Moreover, if an off-diagonal element is zero, then the corresponding responses are uncorrelated, and, under regularity conditions, the covariance between two responses Y_j and Y_k is a strictly increasing function of Σ_{jk}. We use these observations to establish identifiability of the parameters and to design a test for whether responses are uncorrelated, without that hypothesis also implying restrictions on means and variances. If the responses are assumed to be conditionally independent given W, then this is also a test of independence.

The following example illustrates the model.

Example 1. Suppose we observe n independently sampled bivariate responses, consisting of one continuous and one (non-negative) count variable, and a single predictor, $x \in \mathbb{R}$. A possible version of (1) for these data takes $\mathcal{B}^\mathsf{T} \in \mathbb{R}^2$, $W = \mathcal{B}^\mathsf{T} x + \varepsilon$, and

$$\mathbb{E}(Y \mid W) = \begin{bmatrix} W_1 \\ \exp(W_2) \end{bmatrix} \quad \text{and} \quad \operatorname{cov}(Y \mid W) = \operatorname{diag}\{1, \exp(W_2)\},$$

which corresponds to $g(t) = [t_1, \log(t_2)]^\mathsf{T}, t = [t_1, t_2]^\mathsf{T}$. One way to obtain this version of (1) from a fully specified parametric model is to assume the conditional distribution $Y_1 \mid W$ is normal with mean W_1 and variance 1; and that of $Y_2 \mid W$ is Poisson with mean (and hence variance) $\exp(W_2)$. Then, assuming Y_1 and Y_2 are conditionally independent given W, the marginal distribution of $Y = [Y_1, Y_2]^\mathsf{T}$ is characterized by

$$f_{\mathcal{B},\Sigma}(y) \propto \int_{\mathbb{R}^2} \exp\left\{-(y_1 - w_1)^2/2 + y_2 w_2 - e^{w_2} - (w - \mathcal{B}^\mathsf{T} x)^\mathsf{T} \Sigma^{-1} (w - \mathcal{B}^\mathsf{T} x)\right\} \mathrm{d}w,$$

where $y = [y_1, y_2]^\mathsf{T}$ and $w = [w_1, w_2]^\mathsf{T}$. The density $f_{\mathcal{B},\Sigma}(y)$ is a model for the distribution of a mixed-type response vector and how it is affected by the predictor. We emphasize that this model can be useful even if W has no practical interpretation. Indeed, we do not assign meaning to W in general but focus on the resulting distribution of Y.

We propose an algorithm for fitting (1) that, loosely speaking, iteratively fits a sequence of multivariate normal models whose moments approximate those implied by (1). This gives an algorithm that on

a high level is similar to penalized quasi-likelihood (Breslow and Clayton, 1993; Breslow, 2004) but for mixed-type responses. In each step of the algorithm, an objective function is minimized by (block) coordinate descent in $\mathcal{B}$ and Σ. The update for $\mathcal{B}$ is a least squares problem and hence has a closed form solution. The update for Σ is a more complicated optimization problem over a set of symmetric and positive (semi-)definite matrices and is fundamentally different from optimization problems solved by common mixed models packages. We solve this problem using an accelerated projected gradient descent algorithm. Because this algorithm does not require the computation of second derivatives, it scales well in the dimension of the response vector. The algorithm natively supports restrictions on Σ that can be expressed as $\Sigma \in \mathbb{M}$ for a set $\mathbb{M}$ such that the projection onto it can be computed, and this is essential for our method: first, some combinations of responses require identifiability restrictions on Σ and this is straightforward to incorporate in the projection step of our algorithm. Secondly, we develop a procedure for approximate likelihood ratio testing which uses the projection step to impose null hypothesis restrictions such as independence between responses. Similarly, if one has subject-specific knowledge suggesting a particular structure for Σ such as, say, a first order autoregressive structure, then one can take $\mathbb{M} = \{\Sigma \in \mathbb{R}^{r\times r} : \Sigma_{i,j} = \rho^{|i-j|}, \rho \in [-1, 1]\}$.

In Section 2 we examine properties of (1) in more detail and give conditions that ensure the parameters are identifiable. Section 3 contains the proposed algorithm and details on its implementation. Section 4 describes an approximate likelihood ratio testing procedure and in Section 5 we examine the prediction, estimation, and testing performance of the proposed method using simulations. Data examples are in Section 6.

2 Model

2.1 Specification

Because it makes our development no more difficult, in what follows we consider a slightly more general version of (1):

$$g\{\mathbb{E}(Y \mid W)\} = W \text{ and } W = X\beta + \varepsilon, \tag{2}$$

where $X \in \mathbb{R}^{r\times q}$ is a non-stochastic design matrix and $\beta \in \mathbb{R}^q$. The classical multivariate response regression setting in (1) which motivates our study, where each response has the same vector of

predictors, is a special case with $X = I_r \otimes x^\mathsf{T}$, $\beta = \mathrm{vec}(\mathcal{B})$, and $q = rp$, where $\otimes$ is the Kronecker product and $\mathrm{vec}(\cdot)$ the vectorization operator. Similarly, a seemingly unrelated regression (Zellner, 1962), where each response has its own predictor vector and coefficient vector, is also a special case. Stochastic predictors whose distribution does not depend on model parameters can be accommodated by conditioning on them in appropriate places. In particular, statements we make about the marginal distribution of Y then instead apply to the distribution of $Y \mid X$, assuming ε is independent of X.

We assume the link function g satisfies

$$g(t) = [g_1(t_1), \ldots, g_r(t_r)]^\mathsf{T},$$

where $t = [t_1, \ldots, t_r]^\mathsf{T}$ and $g_j : \mathbb{R} \to \mathbb{R}$ is injective for every j. Thus, the jth latent variable has a direct effect on the jth response but no other responses. For simplicity, we also assume there are known functions $c_j : \mathbb{R} \to \mathbb{R}$ and dispersion parameters $\psi_j > 0$, $j = 1, \ldots, r$, such that

$$\mathbb{E}(Y \mid W) = [c_1'(W_1), \ldots, c_r'(W_j)]^\mathsf{T} \text{ and } \mathrm{cov}(Y \mid W) = \mathrm{diag}[\psi_1 c_1''(W_1), \ldots, \psi_r c_r''(W_r)],$$

where the primes denote derivatives. This assumption is not crucial to our development but makes notation substantially more convenient and is consistent with assuming generalized linear models (McCullagh and Nelder, 1989) for the distributions of $Y_j \mid W_j$, $j = 1, \ldots, r$. Then, c_j is a conditional cumulant function of the jth response and W_j the corresponding natural "parameter". When $\psi_j = 1$, the generalized linear model distributions are one-parameter exponential family distributions, as in Example 1. Our setting does not assume $\psi_j = 1$ for every j, but we do assume the ψ_j are known in what follows.

2.2 Parameter interpretation and identifiability

It is often difficult to interpret parameters in latent variable models and modeling mixed-type responses makes it more complicated yet. Similarly, it is unclear in many latent variable models which parameters are identifiable. We address some such concerns in this section. We study two examples in detail and state a more general result that can also be useful for analyzing other settings.

The parameters β and Σ are straightforward to interpret in the latent regression, but interpreting them in the marginal distribution of Y implied by (2) requires more work. To that end, note that the

mean vector and covariance matrix of Y are, respectively, by iterated expectations,

$$\mathbb{E}(Y) = \mathbb{E}\{g^{-1}(W)\} \text{ and } \operatorname{cov}(Y) = \operatorname{cov}\{g^{-1}(W)\} + \mathbb{E}\{\operatorname{cov}(Y \mid W)\}. \tag{3}$$

We make a number of observations based on these expressions: first, because $\operatorname{cov}(Y \mid W)$ is diagonal by assumption, the covariance between responses is determined by the covariance matrix of $g^{-1}(W)$. Second, since $\mathbb{E}(Y_j)$ and $\mathbb{E}(Y_j^2)$ are determined by the univariate distribution of Y_j, off-diagonal elements of Σ do not affect means and variances of the responses, but typically do affect covariances of different responses. Third, since g and $\operatorname{cov}(Y \mid W)$ are non-linear and non-constant in general, $\mathbb{E}(Y)$ and $\mathbb{E}\{\operatorname{cov}(Y \mid W)\}$ in general depend on both β and diagonal elements of Σ. Fourth, since $\operatorname{var}(Y_j)$ is increasing in ψ_j and $\operatorname{cov}\{g^{-1}(W)\}$ does not depend on ψ, $\operatorname{cor}(Y_j, Y_k)$ is decreasing in ψ_j and ψ_k. This is intuitive since responses are conditionally uncorrelated and hence, loosely speaking, a large element of ψ indicates substantial variation in the corresponding response is independent of the variation in the other responses.

In some settings, more precise statements are possible by analyzing closed form expressions for the moments in (3), as the next example illustrates.

Example 2 (Normal and quasi-Poisson moments)**.** Suppose there are $r = 4$ responses such that

$$\mathbb{E}(Y_j \mid W) = W_j \text{ and } \operatorname{var}(Y_j \mid W) = \psi_j, \ \ j = 1, 2;$$

$$\mathbb{E}(Y_j \mid W) = \exp(W_j) \text{ and } \operatorname{var}(Y_j \mid W) = \psi_j \exp(W_j), \ \ j = 3, 4.$$

These moments are consistent with assuming $Y_j \mid W \sim \mathcal{N}(W_j, \psi_j)$ for $j = 1, 2$, and, if $\psi_3 = \psi_4 = 1$, $Y_j \mid W \sim \text{Poi}\{\exp(W_j)\}$, $j = 3, 4$. When not assuming $\psi_3 = \psi_4 = 1$, we say these moments are consistent with normal and (conditional) quasi-Poisson distributions. We have picked $r = 4$ so that the covariance matrix includes covariances and variances for all possible combinations of such responses. We examine the effects of these assumptions on the marginal moments of Y, or the moments of $Y \mid X$ if X is stochastic. Some algebra shows (Supplementary Material):

$$\mathbb{E}(Y_j) = \begin{cases} X_j^\mathsf{T}\beta & j = 1, 2 \\ \exp(X_j^\mathsf{T}\beta + \Sigma_{jj}/2) & j = 3, 4 \end{cases}.$$

and

$$\operatorname{cov}(Y) = \begin{bmatrix} \psi_1 + \Sigma_{11} & \cdot & \cdot & \cdot \\ \Sigma_{21} & \cdot & \cdot & \cdot \\ \Sigma_{31} e^{X_3^\mathsf{T}\beta + \Sigma_{33}/2} & \cdot & e^{2X_3^\mathsf{T}\beta + \Sigma_{33}}(e^{\Sigma_{33}} - 1 + \psi_3 e^{-X_3^\mathsf{T}\beta - \Sigma_{33}/2}) & \cdot \\ \cdot & \cdot & e^{X_3^\mathsf{T}\beta + X_4^\mathsf{T}\beta + \Sigma_{33}/2 + \Sigma_{44}/2}(e^{\Sigma_{43}} - 1) & \cdot \end{bmatrix}.$$

Omitted entries in $\operatorname{cov}(Y)$ are the same as those not omitted up to changes in subscripts. Clearly, both $\mathbb{E}(Y)$ and $\operatorname{cov}(Y)$ depend on β and Σ, but regardless of type, the variance of Y_j is increasing in Σ_{jj}, the mean is increasing in $X_j^\mathsf{T}\beta$, and the covariance between Y_j and Y_k is increasing in Σ_{jk}. We will later use these observations to prove a result which implies β and Σ are identifiable in this example.

Consider the linear dependence between mixed-type responses with conditional normal and quasi-Poisson moments, Y_1 and Y_3, say. The sign of their correlation is the sign of Σ_{13} and the squared correlation satisfies, by Cauchy–Schwarz's inequality,

$$\operatorname{cor}(Y_1, Y_3)^2 = \frac{\Sigma_{31}^2}{(\psi_1 + \Sigma_{11})(e^{\Sigma_{33}} - 1 + \psi_3/\mathbb{E}(Y_3))} \leq \frac{\Sigma_{11}\Sigma_{33}}{(\psi_1 + \Sigma_{11})(e^{\Sigma_{33}} - 1 + \psi_3/\mathbb{E}(Y_3))},$$

which is upper bounded by $\Sigma_{33}/(\exp(\Sigma_{33}) - 1)$. Thus, strong linear dependence between Y_1 and Y_3 is possible if Σ_{33} is small, so that $\exp(\Sigma_{33}) - 1$ is not much larger than Σ_{33}.

To gain some intuition for how two responses with quasi-Poisson moments behave, suppose for simplicity that $\Sigma_{33} = \Sigma_{44}$, $\psi_3 = \psi_4$, and $X_3^\mathsf{T}\beta = X_4^\mathsf{T}\beta$. Then the correlation between Y_3 and Y_4 is

$$\frac{e^{\Sigma_{43}} - 1}{e^{\Sigma_{33}} - 1 + \psi_3/\mathbb{E}(Y_3)}.$$

For a small ψ_3, this correlation is approximately $(\exp(\Sigma_{43}) - 1)/(\exp(\Sigma_{33}) - 1)$, which for $|\Sigma_{43}| \leq \Sigma_{3,3}$ is upper bounded by 1 and lower bounded by $\{\exp(-\Sigma_{33}) - 1\}/\{\exp(\Sigma_{33}) - 1\}$. The latter expression tends to -1 if $\Sigma_{33} \to 0$ and 0 if $\Sigma_{33} \to \infty$. Thus, strong negative correlation between Y_3 and Y_4 requires a small Σ_{33}.

Example 2 is convenient to analyze because the moments have closed form expressions. In more complicated settings, the following result can be useful. It implies that the mean of Y_j and covariance of Y_j and Y_k are strictly increasing in, respectively, the mean of W_j and covariance between W_j and W_k.

The result is intuitive but we have not found it stated and proved in the literature.

Lemma 2.1. *Let $\phi_{\mu,\Sigma}$ be a bivariate normal density with marginal densities ϕ_{μ_1,σ_1^2} and ϕ_{μ_2,σ_2^2} and covariance $\sigma = \Sigma_{12} = \Sigma_{21}$; then for any increasing, non-constant $g, h : \mathbb{R} \to \mathbb{R}$, the functions defined by*

$$\mu_1 \mapsto \int g(t)\phi_{\mu_1,\sigma_1^2}(t)\,\mathrm{d}t \;\; \textit{and} \;\; \sigma \mapsto \int g(t_1)h(t_2)\phi_{\mu,\Sigma}(t)\,\mathrm{d}t$$

are, assuming the (Lebesgue) integrals exist, strictly increasing on $\mathbb{R}$ and $(-1, 1)$, respectively.

We illustrate the usefulness of this result in another example.

Example 3 (Normal and Bernoulli responses)**.** Suppose $r = 2$ with $Y_1 \mid W_1 \sim \mathcal{N}(W_1, \psi_1)$ and Y_2 Bernoulli distributed with

$$\mathbb{E}(Y_2 \mid W_2) = \mathrm{logit}^{-1}(W_2) = \frac{1}{1 + \exp(-W_2)}.$$

Suppose also for simplicity there are no predictors but that each response has its own intercept:

$$W = \beta + \varepsilon.$$

The marginal distribution of Y_2 is Bernoulli with

$$\mathbb{E}(Y_2) = \int \frac{1}{1 + \exp(-\beta_2 - \Sigma_{22}^{1/2}t)}\phi(t)\,\mathrm{d}t,$$

where $\phi(\cdot)$ is the standard normal density. Using this expression, one can show that, if Σ_{22} is fixed, $\mathbb{E}(Y_2) \to 0$ if $\beta_2 \to -\infty$ and $\mathbb{E}(Y_2) \to 1$ if $\beta_2 \to \infty$. That is, any success probability is attainable by varying β and, hence, some parameter restrictions are needed for identifiability. One possibility, which has been used in similar settings, is to fix Σ_{22} to some known value, say 1 (Dunson, 2000; Bai et al., 2020). While fixing $\Sigma_{22} = 1$ does not impose any restrictions on the distribution of Y_2 as long as β_2 can vary freely, it may impose restrictions on the joint distribution of $Y = [Y_1, Y_2]^\mathsf{T}$, properties of which we consider next.

Equation (3) implies

$$\mathrm{cov}(Y_1, Y_2) = \mathrm{cov}(W_1, \mathrm{logit}^{-1}(W_2)) = \int_{\mathbb{R}^2} \frac{t_1}{1 + \exp(-t_2)}\phi_{\beta,\Sigma}(t)\,\mathrm{d}t - \beta_1\mathbb{E}(Y_2),$$

where $\phi_{\beta,\Sigma}(\cdot)$ is the bivariate normal density with mean β and covariance matrix Σ. The integral does not admit a closed form expression, but Lemma 2.1 says the covariance is strictly increasing in Σ_{12}, which can be used to show the parameters are identifiable in this example if Σ_{22} is known (Theorem 2.2).

To understand which values $\mathrm{cov}(Y_1, Y_2)$ can take, consider the limiting case as $\Sigma_{12} \to \Sigma_{11}^{1/2}\Sigma_{22}^{1/2}$ and assume for simplicity $\beta_1 = \beta_2 = 0$. In the limit, the covariance matrix is singular and the distribution of W the same as that obtained by letting $W_2 = (\sqrt{\Sigma_{22}}/\sqrt{\Sigma_{11}})W_1$. Then

$$\mathrm{cov}(W_1, \mathrm{logit}^{-1}(W_2)) = \int \frac{\Sigma_{11}^{1/2} t}{1 + \exp(-\Sigma_{22}^{1/2} t)} \phi(t)\, \mathrm{d}t.$$

One can also verify that $\beta_2 = 0$ implies $\mathbb{E}(Y_2) = 1/2$ regardless of Σ, and hence $\mathrm{var}(Y_2) = 1/4$ and

$$\mathrm{cor}(Y_1, Y_2) = \frac{\mathrm{cov}(W_1, \mathrm{logit}^{-1}(W_2))}{\sqrt{\mathrm{var}(Y_1)\,\mathrm{var}(Y_2)}} = \frac{2}{\sqrt{\psi_1 + \Sigma_{11}}} \int \frac{\Sigma_{11}^{1/2} t}{1 + \exp(-\Sigma_{22}^{1/2} t)} \phi(t)\, \mathrm{d}t.$$

By using the dominated convergence theorem as $\psi_1 \to 0$ and $\Sigma_{22} \to \infty$, the last right hand side can be shown to tend to and be upper bounded by $\sqrt{2/\pi} \approx 0.8$. This correlation corresponds to a limiting case and is an upper bound on the attainable correlation between Bernoulli and normal responses. For context, we note $\sqrt{2/\pi}$ is also the correlation between a standard normal random variable, say Z, and a Bernoulli variable that is 1 if Z is greater than zero, and zero otherwise; and it is an upper bound on the correlation between normal and Bernoulli responses sharing a random intercept in a multivariate generalized linear mixed model with a probit link for the Bernoulli responses (McCulloch, 2008, Equation 17).

We conclude this section with two results on identifiability. The results are stated for some common choices of conditional moments of $Y \mid W$ but the proof idea can apply also in other settings. Essentially, Lemma 2.1 ensures that β and off-diagonal elements of Σ are identifiable under quite general conditions but some care is needed to show that the diagonal elements of Σ are identifiable, as the preceding example illustrates. The proof of the following theorem is in the Supplementary Material.

Theorem 2.2. *Suppose* $\{(Y_i, X_i) \in \mathbb{R}^r \times \mathbb{R}^{r\times q}; i = 1, \ldots, n\}$ *is an independent sample from* (2) *and that*

1. g_j, $j = 1, \ldots, r$, *is either the identity, natural logarithm, or logit (log-odds) function,*
2. Σ_{jj} *is fixed and known for every* j *corresponding to logit* g_j,

3. $\sum_{i=1}^{n} X_i^\mathsf{T} X_i$ *is invertible,*

then distinct (β, Σ) *correspond to distinct* $\{\mathbb{E}(\mathcal{Y}), \mathrm{cov}(\mathcal{Y})\} \in \mathbb{R}^{rn} \times \mathbb{R}^{rn \times rn}$*, where* $\mathcal{Y} \in \mathbb{R}^{rn}$ *is a vector of all responses.*

Theorem 2.2 asserts distinct parameters correspond to distinct moments. Identifiability in the usual sense that distinct parameters correspond to distinct distributions is obtained as an immediate corollary by further specifying the family of distributions; we omit the proof.

Corollary 2.3. *If the conditions of Theorem 2.2 hold; the* Y_j *are conditionally independent given* W*; and, for every* $j = 1, \ldots, r$*, the conditional distribution of* $Y_j \mid W$*, is either normal with mean* W_j *and variance* ψ_j*, Poisson with mean* $\exp(W_j)$*, or Bernoulli with mean* $1/\{1 + \exp(-W_j)\}$*; then* β *and* Σ *are indentifiable.*

3 Estimation

3.1 Overview

We propose an algorithm for estimating β and Σ that is based on linearization of the conditional mean function $w \mapsto \nabla c(w) = g^{-1}(w)$. This idea is essentially equivalent to linearization of the link function g, traditionally used to motivate algorithms for generalized linear models (McCullagh and Nelder, 1989, Section 2.5) and generalized linear mixed models (Schall, 1991). This linearization admits an algorithm that is similar to penalized quasi-likelihood (Breslow and Clayton, 1993; Breslow, 2004). Because the motivating ideas are relatively well known, we give only a brief overview and focus more on solving the resulting optimization problem, which is quite different from those typically considered in the mixed models literature.

Consider the elementwise first order Taylor approximation of $g^{-1}(\cdot) = \nabla c(\cdot)$ around an arbitrary $w \in \mathbb{R}^r$,

$$\mathbb{E}(Y \mid W) = \nabla c(W) \approx \nabla c(w) + \nabla^2 c(w)(W - w).$$

Applying expectations and covariances on both sides yields

$$\mathbb{E}(Y) \approx \nabla c(w) + \nabla^2 c(w)(X\beta - w) := m(w, \beta) \quad \text{and} \quad \mathrm{cov}\{\mathbb{E}(Y \mid W)\} \approx \nabla^2 c(w) \Sigma \nabla^2 c(w).$$

Further applying the approximation $\mathbb{E}\{\mathrm{cov}(Y \mid W)\} = \mathbb{E}\{\mathrm{diag}(\psi)\nabla^2 c(W)\} \approx \mathrm{diag}(\psi)\nabla^2 c(w)$ leads to an approximate covariance matrix:

$$\mathrm{cov}(Y) \approx \mathrm{diag}(\psi)\nabla^2 c(w) + \nabla^2 c(w)\Sigma\nabla^2 c(w) := C(w, \Sigma).$$

Intuitively, we expect $m(w, \beta)$ and $C(w, \Sigma)$ to be good approximations if W takes values near w with high probability. Now, consider a working model which says $Y_1, \ldots, Y_n$ are independent with

$$Y_i \sim \mathcal{N}\{m(w_i, \beta), C(w_i, \Sigma)\}, \tag{4}$$

for observation-specific approximation points $w_i \in \mathbb{R}^r$, $i = 1, \ldots, n$. The corresponding negative log-likelihood is, up to scaling and additive constants,

$$h_n(\beta, \Sigma \mid w_1, \ldots, w_n) = \sum_{i=1}^{n} \left[\log\det\{C(w_i, \Sigma)\} + \{y_i - m(w_i, \beta)\}^\mathsf{T} C(w_i, \Sigma)^{-1}\{y_i - m(w_i, \beta)\}\right].$$

Minimizers of h_n are approximate maximum likelihood estimates, whose quality depend on the accuracy of the working model (4). Based on these motivations, a natural algorithm for estimating β and Σ would iterate between updating the pair (β, Σ) by minimizing h_n with the w_i held fixed; and then updating the w_i to get a more accurate working model. In our algorithm, we update the w_i by setting them to equal to the "posterior" prediction of W_i having observed y_i; that is, the maximizer of the conditional density

$$f_{\beta,\Sigma}(w_i \mid y_i) \propto \exp\left\{y_i^\mathsf{T} w_i - c(w_i) - \frac{1}{2}(w_i - X_i\beta)^\mathsf{T}\Sigma^{-1}(w_i - X_i\beta)\right\}.$$

This update for w_i is closely related to Laplace approximation arguments used to motivate common algorithms for mixed models (see e.g. Breslow, 2004). If β and Σ are the true parameters, the working model approximates the moments of the ith response vector around the mode of the distribution of $W_i \mid Y_i$. To summarize, we propose a blockwise iterative algorithm whose $(k+1)$th iterates are obtained

using the updating equations

$$(\beta^{(k+1)}, \Sigma^{(k+1)}) = \arg\min_{\beta,\Sigma} h_n(\beta, \Sigma \mid w_1^{(k)}, \ldots, w_n^{(k)}) \tag{5}$$

$$(w_1^{(k+1)}, \ldots, w_n^{(k+1)}) = \arg\max_{w_1,\ldots,w_n} \sum_{i=1}^{n} \log f_{\beta^{(k+1)},\Sigma^{(k+1)}}(w_i \mid y_i). \tag{6}$$

This algorithm can be run for a pre-specified number of iterations or until convergence of the β and Σ iterates, for example. While the complete algorithm is not designed to minimize a particular objective function, the individual updates, which we discuss in more detail in the following subsections, minimize objective functions that can be tracked to determine convergence within each update. In our experience, the values of Σ and β tend to converge after (at most) tens of iterations of (5) and (6).

In the following subsections, we first discuss the separate optimization problems in (5) and (6) and then state the full algorithm formally. Computing times are discussed in Section 5.

3.2 Updating β and Σ

To solve the optimization problem in (5), we use a blockwise coordinate descent algorithm. Treating $w = \{w_1, \ldots, w_n\}$ as fixed throughout this subsection (and ignoring the iterate counting superscript), this algorithm iterates between updating β with Σ held fixed and vice versa. Specifically, the $(l+1)$th iterates of the algorithm for solving (5) can be expressed

$$\beta^{(l+1)} = \arg\min_{\beta} h_n(\beta, \Sigma^{(l)} \mid w_1, \ldots, w_n) \tag{7}$$

$$\Sigma^{(l+1)} = \arg\min_{\Sigma} h_n(\beta^{(l+1)}, \Sigma \mid w_1, \ldots, w_n) \tag{8}$$

The partial minimization with respect to β, (7), is straightforward. Dropping terms which do not depend on β from h_n, and letting $\tilde{y}_i = y_i - \nabla c(w_i) + \nabla^2 c(w_i) w_i$ and $\tilde{X}_i = \nabla^2 c(w_i) X_i$, the objective function from (7) can be expressed

$$h_n(\beta, \Sigma^{(l)} \mid w_1, \ldots, w_n) \propto \sum_{i=1}^{n} (\tilde{y}_i - \tilde{X}_i \beta)^{\mathsf{T}} C(w_i, \Sigma^{(l)})^{-1} (\tilde{y}_i - \tilde{X}_i \beta),$$

which is simply a (weighted) residual sum-of-squares. Hence, (7) can be computed in closed form:

$$\beta^{(l+1)} = \left\{ \sum_{i=1}^{n} \tilde{X}_i^{\mathsf{T}} C(w_i, \Sigma^{(l)})^{-1} \tilde{X}_i \right\}^{-1} \sum_{i=1}^{n} \tilde{X}_i^{\mathsf{T}} C(w_i, \Sigma^{(l)})^{-1} \tilde{y}_i. \tag{9}$$

Turning to (8), minimizing h_n with respect to Σ is non-trivial owing to the non-convexity of h_n and the constraint that Σ is positive semi-definite. One possibility is to parameterize Σ in a way that lends itself to unconstrained optimization (see e.g. Pinheiro and Bates, 1996) and use a generic solver. However, such parameterizations are inconvenient when imposing additional constraints on Σ. For example, as discussed in Section 2, when some responses are binary we restrict diagonal elements of Σ to be equal to a prespecified constant for identifiability. Similarly, when testing the correlation of response components, we fit our model with certain off-diagonal elements of Σ constrained to be zero. Thus, we need an algorithm that both allows restrictions on the elements of Σ and ensures estimates are positive semi-definite.

By picking an appropriate (convex) $\mathbb{M} \subseteq \mathbb{R}^{r \times r}$, the optimization problem in (8) can be more precisely characterized as

$$\arg\min_{\Sigma} h_n(\beta^{(l+1)}, \Sigma \mid w_1, \ldots, w_n) \text{ subject to } \Sigma \in \mathbb{M}. \tag{10}$$

To handle both the non-convexity and general constraints, we propose to solve (10) using a variation of the inertial proximal algorithm proposed by Ochs et al. (2014). This is an accelerated projected gradient descent algorithm that can be used to minimize an objective function which is the sum of a non-convex smooth function and convex non-smooth function. In our case, h_n (as a function of Σ) is the non-convex smooth function and the convex non-smooth function is the optimization indicator that $\Sigma \in \mathbb{M}$, that is, the function which equals ∞ if $\Sigma \notin \mathbb{M}$ and zero otherwise. This algorithm, like many popular accelerated first order algorithms, e.g., FISTA (Beck and Teboulle, 2009), uses "inertia" in the sense that the standard projected gradient step is modified to account for the direction of change from $\Sigma^{(t-1)}$ to $\Sigma^{(t)}$, where $\Sigma^{(t)}$ denotes the tth iterate of the algorithm used to solve (8). This can lead to faster convergence than the standard projected gradient descent algorithm.

To summarize briefly, with β and w held fixed, our algorithm for solving (10) has $(t+1)$th iterate

$$\Sigma^{(t+1)} = \mathcal{P}_{\mathbb{M}} \left[\Sigma^{(t)} - \alpha \nabla_{\Sigma} h_n(\beta, \Sigma^{(t)} \mid w_1, \ldots, w_n) + \gamma \{ \Sigma^{(t)} - \Sigma^{(t-1)} \} \right],$$

where $\gamma = (0, 1)$, α is determined using backtracking line search (see Algorithm 4 of Ochs et al. (2014) for a precise description of the backtracking line search conditions), and $\mathcal{P}_{\mathbb{M}}$ is the projection onto $\mathbb{M}$. Hence, each update requires computing the projection

$$\mathcal{P}_{\mathbb{M}}(\Sigma) = \underset{M \in \mathbb{M}}{\arg\min} \|\Sigma - M\|_F,$$

where subscript F indicates the Frobenius norm. We assume in what follows that this projection is defined; it suffices, for example, that $\mathbb{M}$ is non-empty, closed, and convex (e.g. Megginson, 1998, Corollary 5.1.19). In our software, we allow $\mathbb{M}$ to be the intersection of a constraint set and the set of positive semi-definite matrices with eigenvalues bounded below by $\epsilon \geq 0$, e.g.,

$$\mathbb{M} = \left\{\Sigma = \Sigma^T : \Sigma \succeq \epsilon I_r, \Sigma_{ij} = \Sigma_{ji} = c_{ij} \text{ for } (i, j) \in \mathcal{C}\right\},$$

where $\mathcal{C}$ denotes the set of indices which are constrained and c_{ij} is the value which the (i, j)th entry of Σ is constrained to equal. To compute projections onto $\mathbb{M}$ of this form, we implement Dykstra's alternating projection algorithm (Boyle and Dykstra, 1986). This algorithm iterates between projections onto each of the two sets whose intersection defines $\mathbb{M}$. Because the projection onto $\{\Sigma = \Sigma^T : \Sigma \succeq \epsilon I\}$ and the projection onto $\{\Sigma = \Sigma^T : \Sigma_{ij} = \Sigma_{ji} = C_{ij} \text{ for } (i, j) \in \mathcal{C}\}$ can both be computed in closed form, this algorithm tends to be very efficient. As a special case, if $\mathcal{C} = \emptyset$, then $\mathcal{P}_{\mathbb{M}}$ can be computed in closed form.

Thus, to apply this algorithm, we need only evaluate the gradient of h_n with respect to Σ. Letting $r_i = \tilde{y}_i - \tilde{X}_i\beta$ and $D_i = \nabla^2 c(w_i)$, we can write

$$h_n(\beta, \Sigma \mid w_1, \ldots, w_n) = \sum_{i=1}^{n} \left[\log\det\{D_i\Sigma D_i + D_i \operatorname{diag}(\psi)\} + r_i^{\mathsf{T}}\{D_i\Sigma D_i + D_i \operatorname{diag}(\psi)\}^{-1} r_i\right]. \tag{11}$$

Letting $C_i(\Sigma) = \{D_i\Sigma D_i + D_i \operatorname{diag}(\psi)\}^{-1}$, for $i = 1, \ldots, n$, routine calculations give

$$\nabla_\Sigma h_n(\beta, \Sigma; w_1, \ldots, w_n) = \sum_{i=1}^{n} D_i \left\{C_i(\Sigma) - C_i(\Sigma) r_i r_i^{\mathsf{T}} C_i(\Sigma)\right\} D_i.$$

This algorithm is terminated when the objective function values converge.

3.3 Updating the approximation points

We use a trust region algorithm for updating $w_i, i = 1, \ldots, n$, a detailed description of which can be found in Nocedal and Wright (2006). Essentially, the trust region algorithm approximates the objective function locally by a quadratic and requires the computation of gradients and Hessians. These are, for $i = 1, \ldots, n$, assuming Σ^{-1} exists,

$$\nabla_{w_i} \log f(w_i \mid y_i; \beta, \Sigma) = y_i - \nabla c(w_i) - \Sigma^{-1}(w_i - X_i^\mathsf{T}\beta)$$

and

$$\nabla^2_{w_i} \log f(w_i \mid y_i; \beta, \Sigma) = -\nabla^2 c(w_i) - \Sigma^{-1}.$$

Since $\nabla^2 c(w_i)$ and Σ^{-1} are positive definite and the latter does not depend on w_i, the objective function is strongly concave. Thus, it has a unique maximizer and stationary point. In practice, however, Σ can be singular or near-singular and our experience suggests that, especially in early iterations and for starting values far from the solution, the Hessian $-\Sigma^{-1} - \nabla^2 c(w)$ can be nearly singular or have a very large condition number even if Σ is non-singular. To improve stability, we suggest regularizing by (i) adding an L_2-penalty on $w_i - X_i\beta$ and (ii) replacing Σ by $\underline{\Sigma} = \Sigma + \epsilon I_r$ for some small $\epsilon > 0$ when updating w_i. Then the optimization problem for updating w_i can then be written

$$\arg\min_w \left\{ -y_i^\mathsf{T} w + c(w) + \frac{1}{2}(w_i - X_i^\mathsf{T})^\mathsf{T} \underline{\Sigma}^{-1}(w_i - X_i\beta) + \tau \|w_i - X_i^\mathsf{T}\beta\|^2 \right\},$$

where $\tau \geq 0$. In practice, we have found that setting τ to the largest eigenvalue of the current iterate of Σ works well. The intuition for shrinking w_i to $X_i\beta$ is that, when β is the true parameter, the latter is the mean of W_i. The penalty and regularization of Σ are only included in the update for w_i, not in the objective function for updating β and Σ. That is, regularization is used to find useful points around which to apply the working model (4), not in the fitting of that model. Since the n objective functions for the w_i are separate, the updates can be done in parallel. We use the implementation in the R package `trust` (Geyer, 2020).

3.4 Summary and discussion

We summarize the steps of our algorithm for estimating β and Σ in Algorithm 1 below.

Algorithm 1: Blockwise iterative algorithm for estimating (β, Σ)

1. Given $\epsilon_\beta > 0$, $\epsilon_\Sigma > 0$, and $\mathbb{M}$, initialize $\Sigma^{(1)} \in \mathbb{M}$, and $\beta^{(1)} \in \mathbb{R}^q$. Set $k = 1$.
2. $w_i^{(k+1)} \leftarrow \arg\max_{w \in \mathbb{R}} \left\{ \log f_{\beta^{(k)}, \Sigma^{(k)}}(w \mid y_i) - \tau \|y_i - X_i \beta^{(k)}\|^2 \right\}$ for $i = 1, \ldots, n$.
3. Set $\tilde{\Sigma}^{(1)} = \Sigma^{(k)}$. For $l = 1, 2, \ldots$ until convergence:
   (a) $\tilde{\beta}^{(l+1)} \leftarrow \arg\min_\beta h_n(\beta, \tilde{\Sigma}^{(l)} \mid w_1^{(k+1)}, \ldots, w_n^{(k+1)})$ using (9)
   (b) Set $\bar{\Sigma}^{(0)} = \bar{\Sigma}^{(1)} = \tilde{\Sigma}^{(t)}$. For $t = 1, 2, \ldots,$ until convergence:
      i. $\bar{\Sigma}^{(t+1)} \leftarrow \mathcal{P}_{\mathbb{M}} \left[\bar{\Sigma}^{(t)} - \alpha \nabla_\Sigma h_n(\tilde{\beta}^{(l+1)}, \bar{\Sigma}^{(t)}, w_1^{(k+1)}, \ldots, w_n^{(k+1)}) + \gamma \{ \bar{\Sigma}^{(t)} - \bar{\Sigma}^{(t-1)} \} \right],$
   (c) $\tilde{\Sigma}^{(l+1)} \leftarrow \bar{\Sigma}^{(t^*)}$ where $\bar{\Sigma}^{(t^*)}$ is the final iterate from 3(b).
4. $(\beta^{(k+1)}, \Sigma^{(k+1)}) \leftarrow (\tilde{\beta}^{(t^*)}, \tilde{\Sigma}^{(t^*)})$ where $(\tilde{\beta}^{(l^*)}, \tilde{\Sigma}^{(l^*)})$ are the final iterates from 3.
5. If $\|\beta^{(k+1)} - \beta^{(k)}\|_F^2 \leq \epsilon_\beta$ and $\|\Sigma^{(k+1)} - \Sigma^{(k)}\|_F^2 \leq \epsilon_\Sigma$, terminate. Otherise, set $k \leftarrow k + 1$ and return to 2.

Initializing values can affect the final estimates of (β, Σ). For this reason, we propose a two-step initialization approach which we find leads to good initial values. In the first step, we run Algorithm 1 after initializing $w_i = 0$, $\beta = 0$, and $\Sigma = I_r$ under the restriction that Σ is diagonal. Once this algorithm has converged, in the second step, we run Algorithm 1 again by initializing (β, Σ) and the w_i at their final iterates from the first step. However, we drop the constraint that Σ is diagonal, and allow Σ to be unrestricted (i.e., Σ need not belong to $\mathbb{M}$). We also replace step 3(b) – (c) by a trust region algorithm which often converges quickly but does not guarantee positive semi-definiteness. Once this algorithm has converged, we use the final iterates of (β, Σ) and the w_i as our initial values for Algorithm 1 under the restriction that $\Sigma \in \mathbb{M}$. In terms of computing time, we found this approach is often faster than running Algorithm 1 directly; and tends to lead to better estimates of (β, Σ). If r is relatively large (e.g., as in Section 6.2), the trust region update of Σ used to get initial values can be slow since it requires repeatedly computing a Hessian of dimension $\{r(r+1)/2\} \times \{r(r+1)/2\}$; the second initialization step can then be skipped.

4 Approximate likelihood ratio testing

In order to perform inference on (β, Σ), we propose a novel procedure for hypothesis testing that is based on the likelihood for the working model (4). We focus on testing hypotheses of the form

$(\beta, \Sigma) \in \mathbb{H}_0$ versus $(\beta, \Sigma) \in \mathbb{H}_1$ and propose the test statistic

$$T_n = h_n(\tilde{\beta}, \tilde{\Sigma} \mid \tilde{w}_1, \ldots, \tilde{w}_n) - h_n(\bar{\beta}, \bar{\Sigma} \mid \tilde{w}_1, \ldots, \tilde{w}_n),$$

where $(\tilde{\beta}, \tilde{\Sigma})$ and the approximation points $\tilde{w} = \{\tilde{w}_1, \ldots, \tilde{w}_n\}$ are obtained by running Algorithm 1 with the restrictions implied by $\mathbb{H}_0$ and

$$(\bar{\beta}, \bar{\Sigma}) = \underset{(\beta,\Sigma)\in\mathbb{H}_1}{\arg\min}\, h_n(\beta, \Sigma \mid \tilde{w}_1, \ldots, \tilde{w}_n).$$

That is, $(\tilde{\beta}, \tilde{\Sigma})$ and $\tilde{w}$ are estimates and expansion points, respectively, from fitting the null model while $\bar{\beta}$ and $\bar{\Sigma}$ are obtained by maximizing the working likelihood from (4) with the expansion points fixed at those obtained by fitting the null model. Thus, the statistics $\bar{\beta}$ and $\bar{\Sigma}$ are not the same as the unconstrained estimates of β and Σ. We fix the expansion points to ensure $(\tilde{\beta}, \tilde{\Sigma})$ and $(\bar{\beta}, \bar{\Sigma})$ are maximizers of the same working likelihood, but under different restrictions. We chose the null model's expansion points to be conservative; that is, to favor the null hypothesis model. If the working model is accurate, then we expect T_n to be, under the null hypothesis, approximately chi-square distributed with degrees of freedom equal to the additional number of restrictions implied by $\mathbb{H}_0$ relative to $\mathbb{H}_1$. For example, a test of independence between all r responses may take $\mathbb{H}_0 = \{(\beta, \Sigma) \in \mathbb{R}^p \times \mathbb{M} : \Sigma_{ij} = \Sigma_{ji} = 0, \forall (i \neq j)\}$, which imposes $(r-1)r/2$ restrictions. Null models corresponding to hypotheses that constrain elements of Σ are straightforward to fit by simply including those constraints in the definition of the set $\mathbb{M}$ in (10).

We summarize the hypothesis testing procedure in Algorithm 2.

Algorithm 2: Hypothesis testing procedure for $(\beta, \Sigma) \in \mathbb{H}_0$ versus $(\beta, \Sigma) \in \mathbb{H}_1$

1. Given $\mathbb{H}_0$ and $\mathbb{H}_1$, initialize $(\beta^{(1)}, \Sigma^{(1)}) \in \mathbb{H}_0$.
2. For $k = 1, 2, \ldots$ until convergence:
   (a) $w_i^{(k+1)} \leftarrow \arg\max_{w\in\mathbb{R}} \left\{\log f_{\beta^{(k)},\Sigma^{(k)}}(w \mid y_i) - \tau \|y_i - X_i\beta^{(k)}\|^2\right\}$ for $i = 1, \ldots, n$
   (b) $(\beta^{(k+1)}, \Sigma^{(k+1)}) \leftarrow \arg\min_{(\beta,\Sigma)\in\mathbb{H}_0} h_n(\beta, \Sigma \mid w_1^{(k+1)}, \ldots, w_n^{(k+1)})$
3. Set $(\tilde{\beta}, \tilde{\Sigma}) = (\beta^{(k^*)}, \Sigma^{(k^*)})$ and $\tilde{w}_i = w_i^{(k^*)}$ for $i = 1, \ldots, n$ where k^* denotes the final iterate from 2.
4. Compute $(\bar{\beta}, \bar{\Sigma}) = \arg\min_{(\beta,\Sigma)\in\mathbb{H}_1} h_n(\beta, \Sigma \mid \tilde{w}_1, \ldots, \tilde{w}_n)$ using algorithm to solve (5)
5. Return $T_n = h_n(\tilde{\beta}, \tilde{\Sigma} \mid \tilde{w}_1, \ldots, \tilde{w}_n) - h_n(\bar{\beta}, \bar{\Sigma} \mid \tilde{w}_1, \ldots, \tilde{w}_n)$.

We investigate the size and power of the proposed procedure through extensive numerical experiments in Section 5.4.

5 Numerical experiments

5.1 Overview

In this section, we study the numerical performance of our method from multiple perspectives. In the first subsection, we compare our method to two generalized linear mixed models which model the responses of different types separately. In the second subsection, we focus on comparing two versions our method: that with Σ constrained to be diagonal and that with the off-diagonal elements of Σ unconstrained. Finally, in the last subsection, we examine the power and size of our proposed hypothesis testing procedure under various data generating models. Our software and code that can be used to reproduce the results are available at `https://github.com/koekvall/lvmmr`.

5.2 Comparison to existing models and methods

We are not aware of a publicly available (or otherwise) software that fits our model outside of special cases. Thus, to evaluate our method, we compare to existing methods which assume related but somewhat different models. We consider $r = 9$ response variables, three conditionally normally distributed, three conditionally Poisson, and three Bernoulli. A reasonable alternative to our method is to use separate generalized linear mixed models for the three response types. However, common software cannot fit these models due to the unusual random effects structure required; see the Supplementary Material for additional details. To get a comparison, we consider two simplifications of the type-specific models: (i) $\Sigma_{jj} = \sigma_j^2$ and $\Sigma_{jk} = 0$ for $j \neq k$ or (ii) $\Sigma = \sigma^2 1_3 1_3^\mathsf{T}$, for some $\sigma^2 > 0$ and $1_3 = [1, 1, 1]^\mathsf{T}$. Option (i) corresponds to assuming all responses are independent and that latent variables corresponding to different responses have their own variance, and option (ii) to using a shared random effect for observations in the same cluster, where a cluster here consists of responses of the same type in the same response vector. For short, we refer to these as, respectively, independent and clustered generalized linear mixed models. With these simplifications, there are several software packages that can fit the separate models. Somewhat arbitrarily, we pick the `glmm` (Knudson et al., 2020) package in R to fit (i) and fit

(ii) using `lme4` (Bates et al., 2015) in R. Briefly, the former uses a Monte Carlo approximation of the likelihood and the latter uses adaptive Gaussian quadrature. We emphasize that both are based on models that are misspecified in general in our setting and, accordingly, our experience from experimenting with several software packages is that differences in performance are due to misspecification rather than any particular implementation. We consider two versions of our method, one which constrains Σ to be diagonal and one where Σ is (nearly) unconstrained. For both versions, Σ is positive semi-definite and, following Theorem 2.2, we constrain $\Sigma_{jj} = 1$ for all j corresponding to Bernoulli responses. The true value of Σ_{jj} used to generate data is in general not equal to 1, however. We do not constrain diagonal elements of Σ when fitting independent or clustered generalized linear mixed models because the software used does not support it.

The comparisons focus on out of sample prediction errors. Our predictions are formed by plugging estimates into the expressions for $\mathbb{E}(Y_i) = \mathbb{E}\{\mathbb{E}(Y_i \mid W_i)\}$ discussed in Section 2. When a closed form expression is unavailable, the expectation is easily obtained by r one-dimensional numerical integrations. As a reference, we compare to (oracle) predictions using the true β and Σ. The responses have different predictors and we partition accordingly: $\beta = [\beta_1^\mathsf{T}, \ldots, \beta_r^\mathsf{T}]^\mathsf{T}$, $\beta_j \in \mathbb{R}^{p_j}$, $q = \sum_{j=1}^r p_j$, and we write $X_{i,j} \in \mathbb{R}^{p_j}$ for the ith observation of the predictors for the jth response. Then, $X_i \in \mathbb{R}^{r \times q}$ has jth row $[0, \ldots, X_{i,j}^\mathsf{T}, 0, \ldots, 0]^\mathsf{T} \in \mathbb{R}^p$, there being $\sum_{k=1}^{j-1} p_k$ leading zeros and $\sum_{k=j+1}^r p_k$ trailing. Thus, $W_{i,j} = X_{i,j}^\mathsf{T}\beta_j + \varepsilon_{i,j}$. In all simulations, each $X_{i,j}$ consists of a one in the first element (an intercept) and, in the remaining $p_j - 1$ elements, independent realizations of a $\mathrm{Uniform}[-1, 1]$ random variable, where for simplicity we take $p_j = p_k$ for all j and k. For $j = 1, \ldots, r$, the true regression coefficient β_j has first element (the intercept) equal to β_{0j} and all other elements chosen as independent realizations of a $\mathrm{Uniform}[-.5, .5]$. We set $\beta_{0j} = 2$ if the response is normal or (quasi-)Poisson, and equal to zero if the response is Bernoulli. Similarly, if the response is normal, we set $\psi_j = .01$; otherwise, we set $\psi_j = 1$.

We consider three different structures for Σ. Namely, for some $\rho \in (0, 1)$: we set $\Sigma = 0.5\tilde{\Sigma}$ where $\tilde{\Sigma}$ is (i) autoregressive, meaning $\tilde{\Sigma}_{jk} = \rho^{|j-k|}$; (ii) compound symmetric, meaning $\tilde{\Sigma}_{jk} = \rho\mathbb{I}(j \neq k) + \mathbb{I}(j = k)$; or (iii) block diagonal, meaning $\tilde{\Sigma}_{jk} = \rho\mathbb{I}(j \neq k) + \mathbb{I}(j = k)$ if $(j, k) \in \{1, 4, 7\} \times \{1, 4, 7\}$, $(j, k) \in \{2, 5, 8\} \times \{2, 5, 8\}$, or $(j, k) \in \{3, 6, 9\} \times \{3, 6, 9\}$ and zero otherwise. The first through third responses are normal, the fourth through sixth Bernoulli, and seventh through ninth quasi-Poisson. Hence, the block-diagonal structure in (iii) consists of block of responses of all unique types. To be clear, these structures are used to generate the data but treated as unknown and, hence, not imposed

when fitting models.

For each of the structures in Σ, we investigate the effects of the sample size (n), the number of predictors (p_j, $j = 1, \ldots, r$), and the correlation parameter (ρ). In Figure 2, we display the average relative squared prediction error, which we define as the ratio of each method's sum of squared prediction error to the sum of squared prediction error of the oracle prediction. Averages are based on 500 independent replications, and for each replication, the out of sample predictions are on an independent test set of 10^4 observations.

Results are displayed in Figure 1. In the top row, we observe that as n increases with $p_j = 5$ and $\rho = 0.90$, each method's performance improves relative to the oracle performance. However, across all settings, our method's performance is best. When the covariance structure is non-sparse (e.g., autoregressive or compound symmetric), we see that the clustered generalized linear mixed models can outperform both our method with the diagonality constraint and the independent generalized linear mixed models. The same relative performances are observed as p increases with $n = 200$ and $\rho = 0.9$ fixed; and with ρ increasing and $(n, p_j) = (200, 5)$. It is notable that when Σ is block diagonal, both versions of our method outperform the competitors. In the case of clustered generalized linear mixed models, this is due to the fact that the specified covariance structure cannot not well approximate the true covariance. For independent generalized linear mixed models, this is likely due to the fact that `glmm` (nor other software) can impose the identifiability condition on Σ for the Bernoulli responses.

Table 1 shows the times to fit our model and the models assumed by `glmm` and `glmer`. Our method, notably, scales well in n, is frequently substantially faster than `glmm`, but somewhat slower than `glmer`. These timings are intuitive as `glmm` uses a Monte Carlo-approximation that is useful for high-dimensional integrals but which is not optimized for our setting with many independent observations. By contrast, `glmer` is fast because it is here fitting a model which incorrectly assumes there is only one variance parameter; this gives a much simpler optimization problem than that our method is solving, and faster numerical integration than what is possible when there are several possibly dependent random effects.

5.3 Performances for different response types

Next, we compare the two versions of our method (off-diagonals of Σ constrained versus unconstrained) from the previous subsection in terms of their performance on responses of the different types. Because

Covariance	Model	Sample size 100	150	200	250	300	350	400	450	500
AR(1)	lvmmr	98.2	79.1	112.1	138.6	163.8	92.6	182.2	166.5	186.1
	glmm	76.0	88.9	206.6	412.9	617.0	442.1	1193.7	1623.0	2213.8
	glmer	14.5	16.4	24.5	31.5	42.4	32.0	56.0	61.1	68.7
BD	lvmmr	24.0	29.1	68.0	67.0	72.2	39.3	64.8	57.8	80.6
	glmm	45.0	81.3	267.6	414.9	612.7	368.8	904.8	778.2	1801.8
	glmer	6.3	9.6	20.3	24.6	31.4	22.9	37.7	30.8	52.8
CS	lvmmr	88.7	130.4	121.2	160.9	165.2	158.4	186.7	84.9	87.6
	glmm	62.7	129.3	183.4	356.8	518.5	712.2	1123.6	666.4	850.2
	glmer	12.4	22.2	25.2	35.9	40.4	44.2	57.6	34.5	40.3

Table 1: Median computing times (in seconds) for our method with off-diagonals of Σ unconstrained (`lvmmr`), independent generalized linear mixed models fit using `glmm`, and clustered generalized linear models fit using `glmer` under the settings considered in the top row of Figure 1. AR(1), BD, and CS correspond to autoregressive, block diagonal, and compound symmetric covariance structures, respectively.

both versions have correctly specified univariate response distributions and are fit using the same algorithm except for the constraints on Σ, these simulations investigate the usefulness of joint modeling of mixed-type responses (i.e., leaving off-diagonals of Σ unconstrained). In the results from Figure 1, averages are taken over all responses types; however, it is also of interest to analyze results stratified by type. We generate data in the same manner as the previous section except we set $\psi_j = .01$ for both normal responses and quasi-Poisson responses; see the Supplementary Material for how we generate quasi-Poisson response with $\psi_j \neq 1$. Notably, neither `glmm` nor `glmer` could accommodate conditionally quasi-Poisson distributed responses.

In Figure 2, we display the average squared prediction errors relative to the oracle prediction. In the first row of Figure 2, we see that as n increases, both methods' relative mean squared prediction error approaches the oracle prediction error. However, the predictions from joint modeling outperforms those based on the model constraining Σ to be diagonal. The smallest differences between the two methods are observed for the Bernoulli response: we investigate this point further later in the section. A similar result is observed in the second row of Figure 2 where the number of predictors per response increases. We see that as p_j approaches 10, both methods' relative performance degrades, although for all three response types, predictions from the joint modeling degrades more gradually. Finally, in the bottom-most row of Figure 2, we display results as ρ increases from 0.5 to 0.95. When $\rho = 0.5$, especially under the block diagonal and AR(1) models, there is a less substantial difference between the

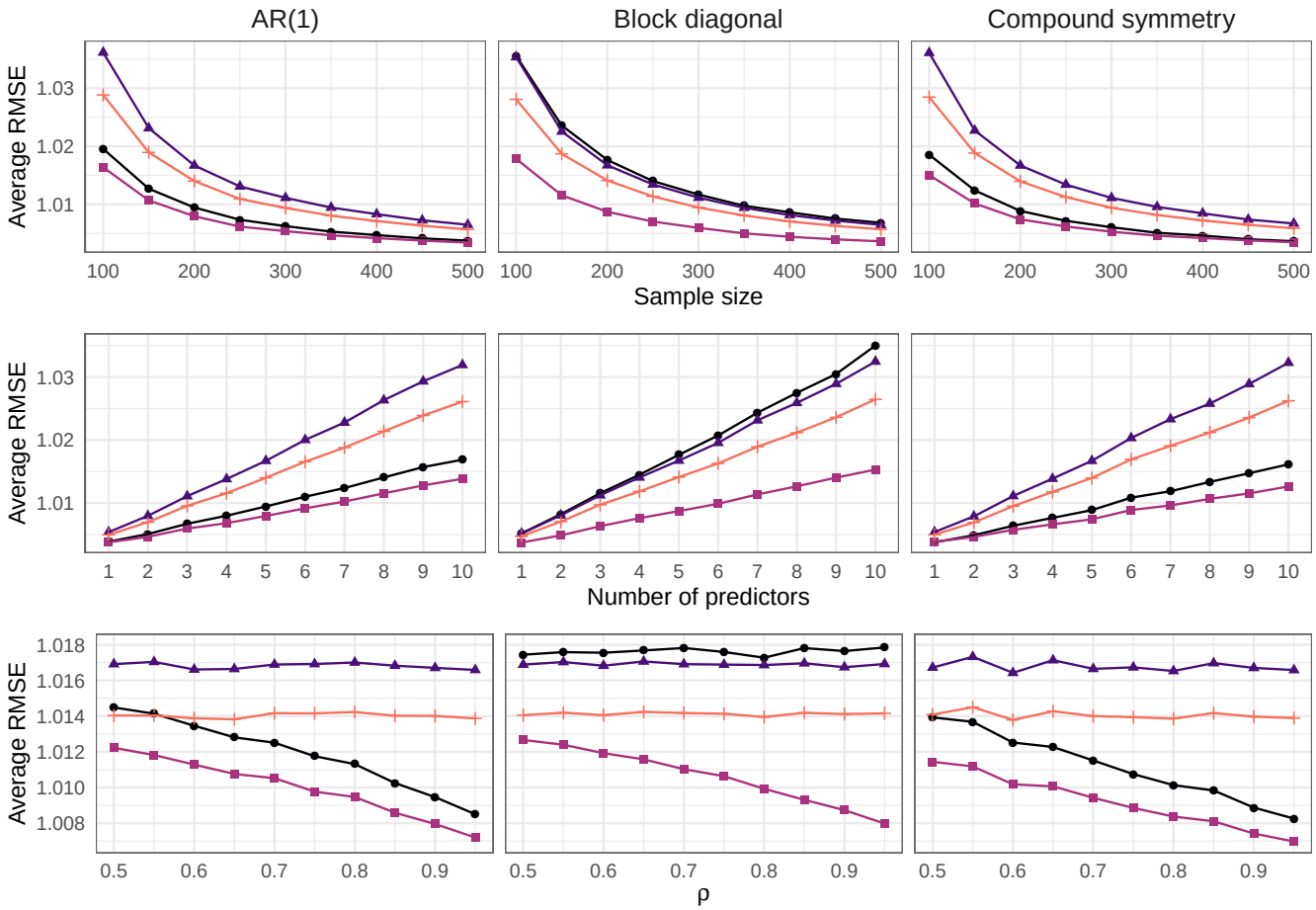

Figure 1: Average relative squared prediction errors over 500 independent replications and $r = 9$ response variables as (top row) the sample size varies with $\rho = 0.9$ and $p_j = 5$ for $j = 1, 2, \ldots, 9$; (middle row) the number of predictors varies with $n = 200$ and $\rho = 0.9$; and (bottom row) the correlation parameter ρ varies with $n = 200$ and $p_j = 5$ for $j = 1, 2, \ldots, 9$. Methods compared are independent generalized linear mixed models using `glmm` (dark purple, triangle); clustered generalized linear mixed models using `glmer` (black, dot); our method with Σ's off-diagonals constrained to be zero (orange, plus sign); and our method with Σ's' off-diagonals unconstrained (light purple, square).

two methods. As ρ approaches 0.95, under each of the covariance structures, the difference between the two methods becomes greater. This result is also observed in the Bernoulli responses, but to a lesser degree than the normal and quasi-Poisson.

In the second part of this simulation study, we focus on modeling many Bernoulli responses and a single normal response. We include these results to highlight the fact that even though the relative

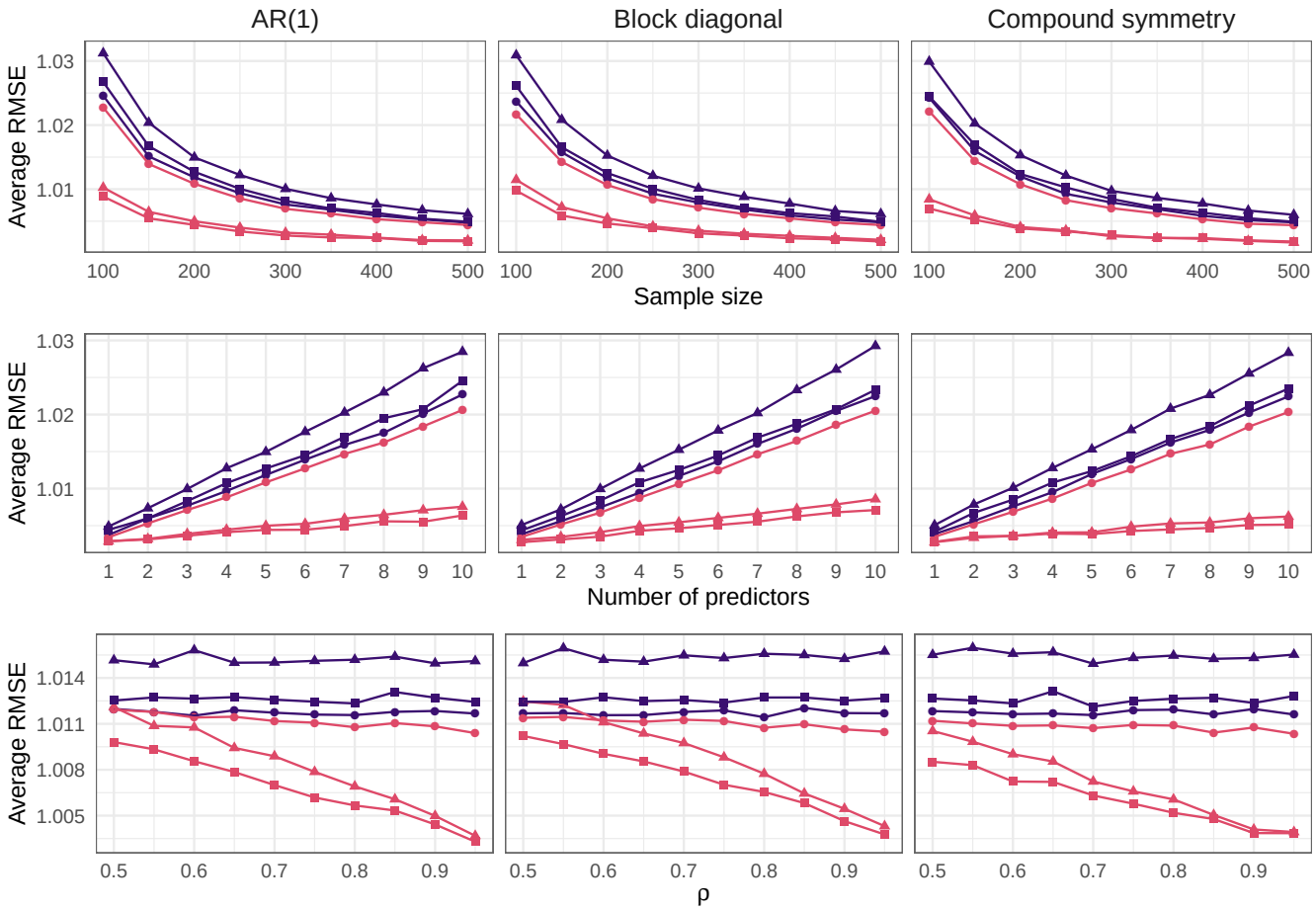

Figure 2: Average relative squared prediction errors over 500 independent replications as (top row) the sample size varies with $\rho = 0.9$ and $p_j = 5$ for $j = 1, 2, \ldots, 9$; (middle row) the number of predictors varies with $n = 200$ and $\rho = 0.9$; and (bottom row) the correlation ρ with $n = 200$ and $p_j = 5$ for $j = 1, 2, \ldots, 9$. Purple lines denote predictions based on our method with Σ's off-diagonals constrained to be zero; magenta lines represent our method with off-diagonals unconstrained. Triangles denote averages over all normal responses and replications; squares over all quasi-Poisson responses and replications, and circles Bernoulli.

square prediction error for the Bernoulli responses is only slightly improved by joint modeling, which was also observed in Figure 2, one can realize substantial prediction accuracy gains for the single normal response variable by exploiting dependence in components of the W_i. Data are generated in the same manner as in Section 5.2, except with a single normal response and eight Bernoulli response variables.

Results are displayed in Figure 3. As before, we observe that as ρ increases from 0.5 to 0.95, the difference between joint and separate modeling becomes more apparent. Notably, the relative mean

squared prediction error for the single normal response variable improves more dramatically under both the autoregressive and compound symmetric covariance structures. Under the block diagonal covariance, the differences are less apparent. This agrees with intuition as under the block diagonal covariance structure, the normal response is only correlated with two of the Bernoulli responses, whereas with the other structures it is correlated with all eight Bernoulli responses. Together with the results in Figure 2, these results suggest that substantial efficiency gains can be achieved using our method for joint modeling of mixed-type responses – even in the case where most response variables are binary.

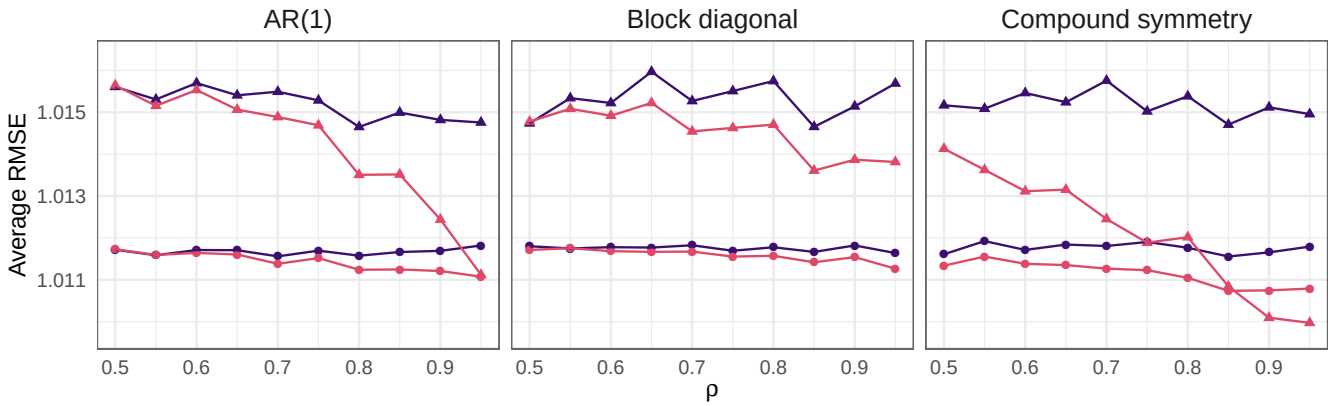

Figure 3: Average relative squared prediction errors over 500 independent replications as the correlation parameter ρ varies with $n = 200$ and $p_j = 5$ for $j = 1, 2, \ldots, 9$. Purple lines represent our algorithm with diagonal Σ and magenta lines represent unstructured Σ. Triangles denote average of the normal response over the replications, and circles denote the average over all Bernoulli responses and replications.

5.4 Approximate likelihood ratio testing

In this section, we examine the approximate likelihood ratio testing procedure described in Algorithm 2. Let $\mathbb{D}^r_{++}$ be the set of $r \times r$ diagonal and positive definite matrices. We study the size and power of the proposed approximate likelihood ratio tests for

$$H_0 : \Sigma \in \mathbb{D}^r_{++} \quad \text{versus} \quad H_A : \Sigma \in \mathbb{S}^r_{++}, \tag{12}$$

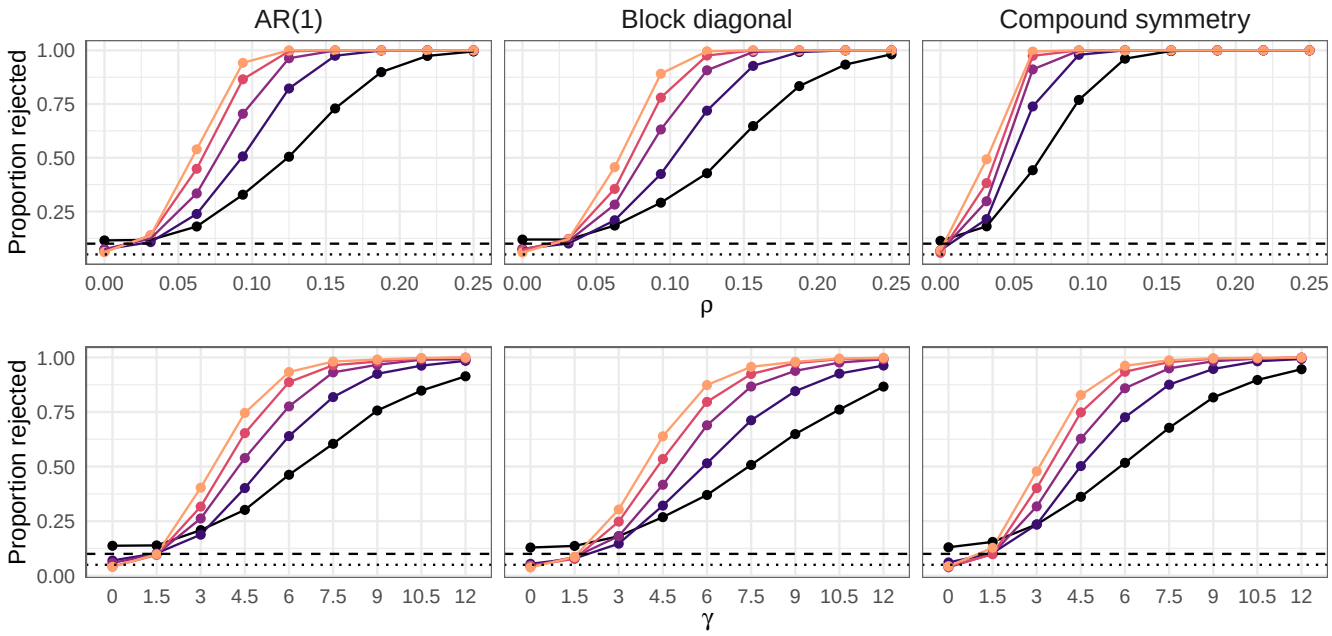

Figure 4: (Top) Proportion of $H_0 : \Sigma \in \mathbb{D}^r_{++}$ rejected (out of 2500 independent replications) at 0.05 level. (Bottom) Proportion of $H_0 : \mathcal{B}_{kj} = 0$ rejected at the 0.05 level. In both, horizontal dashed and dotted black lines correspond to 0.10 and 0.05 respectively. Colors denote sample sizes: the black solid line corresponds to $n = 200$, the dark blue solid line to $n = 400$, the purple solid line to $n = 600$, the magenta solid line to $n = 800$, and the light orange solid line to $n = 1000$.

and, assuming all responses have the same predictors as in (1),

$$H_0 : \mathcal{B}_{kj} = 0, \quad \text{for every } j = 1, \ldots, r, \quad \text{versus} \quad H_A : \mathcal{B} \in \mathbb{R}^{r \times p}, \tag{13}$$

where $\mathcal{B}_{kj}$ denotes the kth predictor's effect on the jth response variable, i.e., the (k, j)th element of $\mathcal{B}$. Without loss of generality, we set $k = 2$. Thus, the null hypothesis in (13) implies that the first predictor (ignoring the intercept), has no effect on any response. This highlights how joint hypotheses are straightforward to test in our model; multiple testing corrections, which are often needed when using separate models for the r responses, are not needed here. Test statistics for are computed using the procedure from Algorithm 2.

Simulations in this section use data generated as in Section 5.3 but with $X_{i,1} = X_{i,2} = \cdots = X_{i,r}$ for all $i = 1, \ldots, n$ and $\mathcal{B} = [\beta_1, \ldots, \beta_r] \in \mathbb{R}^{p \times r}$. Reported averages are based on 2500 independent replications. In the first setting, we consider $n \in \{200, 400, \ldots, 1000\}$ and generate responses with

$\tilde{\Sigma}_{jk} = \rho^{|j-k|}$, $\rho \in \{0.0, 0.05, \ldots, 0.4\}$. In the top row of Figure 4, we display the proportion of rejections at the 0.05 significance level based on the approximate likelihood ratio test statistic (using $\chi^2_{8(8-1)/2}$ quantiles). When $\rho = 0$ (i.e., H_0 is true), approximately 0.10 of tests were incorrectly rejected when $n = 200$ under all three covariance structures. When $n \geq 400$, the proportion of incorrect rejections fall below 0.075 for all three covariance structures. Proprotions of rejections approach 0.05 (the nominal level) as $n = 2000$. As ρ increases, we observe that even with $n = 200$, the proportion of correctly rejected null hypotheses approaches one as $\rho = 0.4$. The power depends on both the magnitude of ρ and the sample size: as one or both increase, the power increases as well.

In the second setting, we fix $\rho = 0.5$ and study how the effect size of the $\mathcal{B}_{kj}$ affects power when testing (13). To this end, after generating $\mathcal{B}$ as in Section 5.3, for $j = 1, \ldots, r$ independently, we replace $\mathcal{B}_{kj}$ with a realization of a $\mathrm{Uniform}[-\gamma 10^{-2}, \gamma 10^{-2}]$ where $\gamma \in [0, 12]$. The second row of Figure 4 shows that when $\gamma = 0$, so that $\mathcal{B}_{kj} = 0$ for all $j = 1, \ldots, r$, the proportion of rejections is slighty above 0.10 for $n = 200$, but close to 0.05 (the nominal size) for all $n \geq 400$. There is also a indication that correlation between responses benefits power. For example, the power curves under compound symmetry tend to be above the corresponding ones under block diagonal structure.

6 Data examples

6.1 Fertility data analysis

We analyze a dataset collected on 333 women who were having difficulty becoming pregnant (Cannon et al., 2013). The goal is to model four mixed-type response variables, all related to the ability to conceive. The predictors are age and three variables related to antral follicles: small antral follicle count, average antral follicle count, and maximum follicle stimulating hormone level. Antral follicle counts can be measured via noninvasive ultrasound and, thus, are often used to model fertility. We standardize predictors before model fitting for ease of interpretation.

The four response variables quantify the ability to conceive in different ways: two are approximately normally distributed (square-root estradiol level and log-total gondadotropin level); and two are counts (number of egg cells and number of embyros). We modeled the latter using our model with conditional quasi-Poisson distributions. Following the simulation studies, we set $\psi_j = 10^{-2}$ for the two normally

distributed responses, and set $\psi_j = 10^{-1}$ for the two count responses. First, we test the hypothesis $H_0 : \Sigma \in \mathbb{D}^4_{++}$ versus $H_A : \Sigma \in \mathbb{S}^4_{++}$ and find strong evidence against the null hypothesis (p-value $< 10^{-16}$) using the test described in Section 5.4. That is, there is strong evidence suggesting the four responses are not independent given the predictors. Fitting the unrestricted model using our software took less than three seconds on a laptop computer with 2.3 GHz 8-Core Intel Core i9 processor. The hypothesis testing procedure took less than six seconds on the same machine.

The estimated correlation matrices for the four observed responses and four latent variables are

$$\widehat{\mathrm{cor}}(Y_i \mid X_i = \bar{X}) = \begin{pmatrix} 1.00 & 0.01 & -0.08 & -0.09 \\ 0.01 & 1.00 & -0.03 & -0.09 \\ -0.08 & -0.03 & 1.00 & 0.69 \\ -0.09 & -0.09 & 0.69 & 1.00 \end{pmatrix}, \quad \widehat{\mathrm{cor}}(W_i \mid X_i) = \begin{pmatrix} 1.00 & 0.02 & -0.09 & -0.10 \\ 0.02 & 1.00 & -0.04 & -0.09 \\ -0.09 & -0.04 & 1.00 & 0.74 \\ -0.10 & -0.09 & 0.74 & 1.00 \end{pmatrix}$$

where the response variable ordering is square-root estradiol level, log-total gondadotropin level, number of egg cells, and number of embyros. Of course, $\mathrm{cor}(Y_i \mid X_i)$ depends on X_i, and its estimate is here evaluated at $\bar{X} = \sum_{i=1}^n X_i/n$. The estimates indicate the number of egg cells and number of embryos are highly dependent whereas estradiol and gondadotropin levels are negatively correlated with these two variables. In general, the estimated linear dependence is slightly stronger for the latent variables than the observed responses.

We also test whether the small antra follicle count is a significant predictor of any of the response variables after accounting for age, average antral follicle count, and maximum follicle stimulating hormone level. The number of small antra follicles (2-5 mm) is highly correlated with the number of total antra follices (2-10 mm), and it has been argued that only total antra follicle count are needed in practice (La Marca and Sunkara, 2013). Fitting our model with $\Sigma \in \mathbb{S}^4_{++}$, we reject the null hypothesis that the four regression coefficients (one for each response) corresponding to antra follicle count is zero (p-value $= 0.0052$).

6.2 Somatic mutations and gene expression in breast cancer

In our second data analysis, we focus on jointly modelling common somatic mutations and gene expression measured on patients with breast cancer collect by the The Cancer Genome Atlas Project (TCGA). A somatic mutation is an alteration in the DNA of a somatic cell. Somatic mutations are

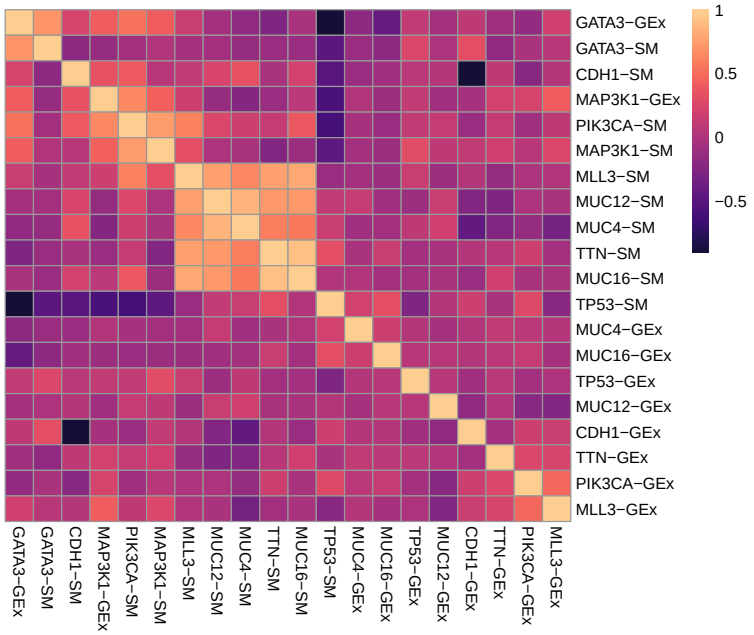

Figure 5: A heatmap of the estimated correlation matrix for the $W_i \mid X_i$ as described in Section 6.2. Names with -SM suffix represent somatic mutations; names with -GEx suffix represent gene expression. Variables were sorted by heirarhical clustering to improve visualization.

believed to play a central role in the development of cancer. Because somatic mutations modify gene expression, either directly and indirectly, it is natural to want to model somatic mutations and gene expression jointly.

The somatic mutations we model are a binary variable indicating the presence or absence of a somatic mutation in the region of a particular gene. We focus on the ten genes where somatic mutations were present in more than 5% of subjects in our dataset. Thus, we have $r = 20$, coming from ten genes each with one response corresponding to gene expression and one to the presence of a somatic mutation. For gene expression, we model log-transformed RPKM measurements as normal random variables. We treat each patients' age as the lone predictor.

First, we test the covariance matrix for block-diagonality. Under H_0, we assume entries of Σ corresponding the correlations between somatic mutations and gene expression measurements are zero (i.e., assuming there is no correlation between somatic mutations and gene expression). We observe test statistic $T_n = 1016.375$ with 100 degrees of freedom for a p-value $< 10^{-16}$.

In Figure 5, we display the estimated correlation matrix for the $W_i \mid X_i$. We observe that the

Coefficient	WOMAC score	Days missed	p-value
Intercept	-0.23822	-4.67790	
BMI	0.02885	0.15758	4.155×10^{-26}
Age	0.00172	-0.03269	2.855×10^{-1}
Sex	0.13314	-0.31994	6.948×10^{-6}

Table 2: Regression coefficient estimates (i.e., $\hat{\mathcal{B}}$) for the three predictors and two response variables in the OAI data analysis. In the rightmost column is the p-value for the test that the corresponding row of $\mathcal{B}$ is entirely zero.

latent variables corresponding to somatic mutations and gene exprssion in CDH1 are highly negatively correlated, whereas for GATA3, somatic mutation and gene expression latent variables have a strong positive correlation. Latent variables for many of the somatic mutations are highly correlated (e.g., TTN, MLL3, MUC4, MUC12, MUC16). However latent variables corresponding to some somatic mutations, e.g., those in the region of TP53, exhibit small or even negative correlations with many others (e.g., GATA3, CDH1, PIK3CA).

6.3 Osteoarthritis initiative data analysis

In this section, we analyze data collected through the Osteoarthritis Initiative (OAI), a prospective observational study of knee osteoarthritis progression (`nda.nih.gov/oai/`). Following McCulloch (2008), we model two outcome variables: Western Ontario and McMaster Univerities disability score (WOMAC), and the number of days of work missed in the three months proceeding data collection. The WOMAC scores are modelled as a normal random variable after adding one and performing a log-transformation; whereas the number of days of work missed are treated as quasi-Poisson random variables. To model these data, we consider BMI, age, and sex as predictors. As in the data analysis in Section 6.1, we set $\psi_j = 10^{-2}$ for the normally distributed response and $\psi_j = 10^{-1}$ for the quasi-Poisson response.

The goal of our analysis was to test for the effect of each of the three predictors on both responses simultaneously. Our analysis included only those subjects who had no missingness in either response variables or predictors, so that $n = 1602$. Fitting the full model to the data, we obtain the coefficient estimates listed in Table 2. Based on the results, we would conclude that both BMI and Sex are significant predictors for both response variables, while Age did not reach the .05 significance cutoff.

Acknowledgements

We are grateful to Galin Jones and Adam Rothman for their comments on an earlier version of the paper. We thank Charles McCulloch for sharing the OAI dataset analyzed in Section 6.3. Substantial parts of the work was done while Ekvall was at the Vienna University of Technology (TU Wien), supported by FWF (Austrian Science Fund, `https://www.fwf.ac.at/en/`) [P30690-N35].

References

Bai, H., Zhong, Y., Gao, X., and Xu, W. (2020). Multivariate mixed response model with pairwise composite-likelihood method. *Stats*, 3(3).

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

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1).

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

Boyle, J. P. and Dykstra, R. L. (1986). A method for finding projections onto the intersection of convex sets in Hilbert spaces. In Dykstra, R., Robertson, T., and Wright, F. T., editors, *Advances in Order Restricted Statistical Inference*, Lecture Notes in Statistics, New York, NY. Springer.

Breslow, N. (2004). Whither PQL? In *Proceedings of the second Seattle symposium in biostatistics*. Springer New York.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421).

Cannon, A. R., Cobb, G. W., Hartlaub, B. A., Legler, J. M., Lock, R. H., Moore, T. L., Rossman, A. J., and Witmer, J. (2013). *Stat2: Building Models for a World of Data*. W. H. Freeman and Company, New York.

Catalano, P. J. (1997). Bivariate modelling of clustered continuous and ordered categorical outcomes. *Statistics in Medicine*, 16(8).

Catalano, P. J. and Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, 87(419).

Cox, D. R. and Wermuth, N. (1992). Response models for mixed binary and quantitative variables. *Biometrika*, 79(3).

de Leon, A. R. (2005). Pairwise likelihood approach to grouped continuous model and its extension. *Statistics & Probability Letters*, 75(1).

de Leon, A. R. and Carriègre, K. C. (2007). General mixed-data model: Extension of general location and grouped continuous models. *Canadian Journal of Statistics*, 35(4).

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(2).

Ekvall, K. O. and Jones, G. L. (2020). Consistent maximum likelihood estimation using subsets with applications to multivariate mixed models. *Annals of Statistics*, 48(2).

Faes, C., Aerts, M., Molenberghs, G., Geys, H., Teuns, G., and Bijnens, L. (2008). A high-dimensional joint model for longitudinal outcomes of different nature. *Statistics in Medicine*, 27(22).

Fitzmaurice, G. M. and Laird, N. M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*, 90(431).

Fitzmaurice, G. M. and Laird, N. M. (1997). Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics*, 53(1).

Geyer, C. J. (2020). *trust: Trust Region Optimization*.

Goldstein, H., Carpenter, J., Kenward, M. G., and Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling*, 9(3).

Gueorguieva, R. V. (2001). A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Statistical Modeling*, 1(3).

Gueorguieva, R. V. and Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*, 96(455).

Gueorguieva, R. V. and Sanacora, G. (2006). Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Statistics in Medicine*, 25(8).

Jaffa, M. A., Gebregziabher, M., Luttrell, D. K., Luttrell, L. M., and Jaffa, A. A. (2016). Multivariate generalized linear mixed models with random intercepts to analyze cardiovascular risk markers in type-1 diabetic patients. *Journal of Applied Statistics*, 43(8).

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer Series in Statistics. Springer-Verlag, New York.

Kang, X., Chen, X., Jin, R., Wu, H., and Deng, X. (2020). Multivariate regression of mixed responses for evaluation of visualization designs. *IISE Transactions*.

Knudson, C., Benson, S., Geyer, C., and Jones, G. (2020). Likelihood-based inference for generalized linear mixed models: inference with the R package glmm. *Stat*.

La Marca, A. and Sunkara, S. K. (2013). Individualization of controlled ovarian stimulation in IVF using ovarian reserve markers: from theory to practice. *Human reproduction update*, 20(1).

Lele, S. R., Nadeem, K., and Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105(492).

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall/CRC, Boca Raton, FL.

McCulloch, C. E. (2008). Joint modelling of mixed outcome types using latent variables. *Statistical Methods in Medical Research*, 17(1).

Megginson, R. E. (1998). *An introduction to Banach space theory*. Number 183 in Graduate texts in mathematics. Springer, New York.

Nocedal, J. and Wright, S. (2006-09-01, 2006). *Numerical Optimization*. Springer-Verlag GmbH.

Ochs, P., Chen, Y., Brox, T., and Pock, T. (2014). iPiano: inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2).

Olkin, I. and Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, 32(2).

Pinheiro, J. C. and Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and computing*, 6(3).

Poon, W.-Y. and Lee, S.-Y. (1987). Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients. *Psychometrika*, 52(3).

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, 69(2).

Sammel, M. D., Ryan, L. M., and Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(3).

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78(4).

Yang, Y., Kang, J., Mao, K., and Zhang, J. (2007). Regression models for mixed Poisson and continuous longitudinal data. *Statistics in Medicine*, 26(20).

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298).