Mixed-type multivariate response regression with covariance estimation

Karl Oskar Ekvall[*] Aaron J. Molstad[†]

karl.oskar.ekvall@ki.se amolstad@ufl.edu

Division of Biostatistics, Institute of Environmental Medicine, Karolinska Institute[*]

Applied Statistics Research Unit, Institute of Statistics and Mathematical Methods in Economics, TU Wien[*]

Department of Statistics and Genetics Institute, University of Florida[†]

June 2021

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 impose no parametric restrictions on the covariance of the latent normal other than positive definiteness, thereby avoiding assumptions about unobservable variables which can be difficult to verify. To accommodate this generality, we propose a novel 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 in simulations; and one biomedical and one genomic data example.

1 Introduction

In many regression applications, there are multiple response variables of mixed types. For instance, when modeling complex biological processes like fertility (see Section 6.1), the outcome (ability to conceive) is often best characterized through a collection of variables of mixed types: both count (number of egg cells and number of embryos) and continuous (square-root estradiol levels and log gonadotropin levels). Similarly, one may want to model the dependence between various types of -omic data in a particular genomic region (see Section 6.2), some binary (e.g., presence of somatic mutations) and some continuous (e.g., gene expression). Joint modeling can lead to 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. Over the years a substantial effort has been made to address this. Many methods for specific combinations of response types 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; Gueorguieva and Sanacora, 2006; Yang et al., 2007; Faes et al., 2008), but models that allow for several different response distributions have also been considered (Sammel et al., 1997; Dunson, 2000; Gueorguieva, 2001; Rabe-Hesketh et al., 2004; de Leon and Carriègre, 2007; Goldstein et al., 2009; Bonat and Jørgensen, 2016; Ekvall and Jones, 2020; Bai et al., 2020; Kang et al., 2021). 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 modification depending on which types of responses are modeled. For example, multivariate generalized linear mixed models (MGLMMs) 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.

To address these issues, we propose a method for multivariate mixed-type response regressions which (i) gives better predictions and estimates than separate models for different response types in many settings, (ii) allows for the testing of relevant joint hypotheses while avoiding the multiple testing burden, (iii) works with many different combinations of mixed-type responses off-the-shelf, and (iv) 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 $Y \in \mathbb{R}^r$ (of possibly mixed types), non-stochastic predictor vector

$x \in \mathbb{R}^p$, and latent vector $W \in \mathbb{R}^r$ satisfy, for some $\mathcal{B} \in \mathbb{R}^{p\times r}$,

$$g\{\mathbb{E}(Y \mid W)\} = W \text{ and } W \sim \mathrm{N}_r(\mathcal{B}^\mathsf{T} x, \Sigma), \tag{1}$$

where $g : \mathbb{R}^r \to \mathbb{R}^r$ is a known link function and $\Sigma \in \mathbb{S}^r_+$ is an unknown covariance matrix. The elements of Y are conditionally uncorrelated given W, with conditional variances that can depend on W in a way to be specified. 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. We illustrate a less trivial application in Example 1.

Example 1. Suppose we observe n independently sampled bivariate responses, consisting of one continuous and one (non-negative) count variable, and a single predictor, $x \in \mathbb{R}$. Naively, one may fit separate normal and Poisson generalized linear models (implicitly assuming independence). Instead, a version of (1) can account for dependence of the two responses: we could take $\mathbb{E}(Y \mid W) = [W_1, \exp(W_2)]^\mathsf{T}$, and $\mathrm{cov}(Y \mid W) = \mathrm{diag}\{1, \exp(W_2)\}$, which corresponds to $g(t) = [t_1, \log(t_2)]^\mathsf{T}, t = [t_1, t_2]^\mathsf{T}$. Then, as in (1), we assume $\mathcal{B}^\mathsf{T} \in \mathbb{R}^2$, $W = \mathcal{B}^\mathsf{T} x + \varepsilon$. This is effectively the same as assuming that conditionally independently, $Y_1 \mid W$ is normal with mean W_1 and variance 1 and $Y_2 \mid W$ is Poisson with mean (and hence variance) $\exp(W_2)$. We discuss the (unconditional) joint distribution of the two responses as a function of $x, \mathcal{B}$ and Σ in Example 2.

Two particularly useful properties of our parameterization (see Section 2) are, first, that off-diagonal elements of Σ affect the covariances of responses but not their means or variances. Second, if an off-diagonal element of Σ 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}. That is, inference on Σ allows one to draw conclusions about the dependence between, e.g., binary and integer-valued response variables. However, standard likelihood ratio testing requires one to maximize the likelihood corresponding to (1) subject to restrictions like $\Sigma_{j,k} = 0$: our new algorithms elegantly handle such problems. In addition, we use these two observations to prove that the parameters $(\mathcal{B}, \Sigma)$ are identifiable and to design a test for whether responses are uncorrelated, without that hypothesis 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.

In this article, we propose an algorithm for fitting (1) that iteratively fits a sequence of multivariate normal models whose moments approximate those implied by (1). This gives an algorithm that is conceptually similar to penalized quasi-likelihood (Breslow and Clayton, 1993) 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 which we solve using an accelerated projected gradient descent scheme. The algorithm scales well in the dimension of the response vector and 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. 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, then one can define $\mathbb{M}$ accordingly. Software implementing this method, along with a set of auxilary functions, are available for download at `github.com/koekvall/lvmmr`.

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) where $W = X\beta + \varepsilon$ for a non-stochastic design matrix $X \in \mathbb{R}^{r \times q}$, $\beta \in \mathbb{R}^q$. The classical multivariate response regression setting in (1) which motivates our study 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. The specification also allows for distinct predictors for each response.

We assume the link function g satisfies $g(t) = [g_1(t_1), \dots, g_r(t_r)]^\mathsf{T}$, where $t = [t_1, \dots, t_r]^\mathsf{T}$ and $g_j : \mathbb{R} \to \mathbb{R}$ is increasing 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 $\psi_j > 0$, $j = 1, \dots, r$, such that $\mathbb{E}(Y \mid W) = [c_1'(W_1), \dots, c_r'(W_j)]^\mathsf{T}$ and $\mathrm{cov}(Y \mid W) = \mathrm{diag}[\psi_1 c_1''(W_1), \dots, \psi_r c_r''(W_r)]$, where primes denote derivatives. This assumption is not crucial but makes notation substantially more convenient and is consistent with assuming GLMs (McCullagh and Nelder, 1989) for the distributions of $Y_j \mid W_j$, $j = 1, \dots, r$. When $\psi_j = 1$, the GLM distributions are one-parameter exponential family distributions, as in Example 1. We do 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, similarly, it is often unclear which parameters are identifiable – we address some such concerns in this section. The parameters are straightforward to interpret in the latent regression, but interpreting them in the marginal distribution of Y requires more work. To that end, note that the mean vector and covariance matrix of Y are, respectively, by iterated expectations,

$$\mathbb{E}(Y) = \mathbb{E}\{g^{-1}(W)\} \text{ and } \operatorname{cov}(Y) = \operatorname{cov}\{g^{-1}(W)\} + \mathbb{E}\{\operatorname{cov}(Y \mid W)\}. \tag{2}$$

We make a number of observations based on (2): first, because $\operatorname{cov}(Y \mid W)$ is assumed diagonal, the covariance between responses is determined $\operatorname{cov}\{g^{-1}(W)\}$. 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. 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 as 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 (2), as the next example illustrates.

Example 2. (Normal and Poisson responses) Suppose there are $r = 4$ responses such that $\mathbb{E}(Y_j \mid W) = W_j$ and $\operatorname{var}(Y_j \mid W) = \psi_j$ for $j = 1, 2$, and $\mathbb{E}(Y_j \mid W) = \exp(W_j)$ and $\operatorname{var}(Y_j \mid W) = \psi_j \exp(W_j)$ for $j = 3, 4$. These moments are consistent with assuming $Y_j \mid W \sim \mathrm{N}(W_j, \psi_j)$ for $j = 1, 2$, and, if $\psi_3 = \psi_4 = 1$, $Y_j \mid W \sim \mathrm{Poi}\{\exp(W_j)\}$, $j = 3, 4$. When not assuming $\psi_3 = \psi_4 = 1$, we say these moments are consistent with normal and (conditional) quasi-Poisson distributions. We examine the effects of these assumptions on the marginal moments of Y. Some algebra gives the following moments (Supplementary Materials): $\mathbb{E}(Y_1) = X_1^{\mathsf{T}}\beta$, $\mathbb{E}(Y_3) = \exp(X_3^{\mathsf{T}}\beta + \Sigma_{33}/2)$, $\operatorname{var}(Y_1) = \psi_1 + \Sigma_{11}$, $\operatorname{var}(Y_3) = \exp(2X_3^{\mathsf{T}}\beta + \Sigma_{33})\{\exp(\Sigma_{33} - 1 + \psi_3 \exp(-X_3^{\mathsf{T}}\beta - \Sigma_{33}/2)\}$, $\operatorname{cov}(Y_1, Y_2) = \Sigma_{21}$, $\operatorname{cov}(Y_1, Y_3) = \Sigma_{31}\exp(X_3^{\mathsf{T}}\beta + \Sigma_{33}/2)$, and $\operatorname{cov}(Y_3, Y_4) = \exp(X_3^{\mathsf{T}}\beta + X_4^{\mathsf{T}}\beta + \Sigma_{33}/2 + \Sigma_{44}/2)\{\exp(\Sigma_{43} - 1)\}$; the remaining entries of $\operatorname{cov}(Y)$ are the same as those given up to obvious 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 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, $\operatorname{cor}(Y_1, Y_3)^2 \leq \Sigma_{11}\Sigma_{33}/[(\psi_1 + \Sigma_{11})\{\exp(\Sigma_{33}) - 1 +$

$\psi_3/\mathbb{E}(Y_3)\}] \leq \Sigma_{33}/\{\exp(\Sigma_{33}) - 1\}$. Thus, strong linear dependence between Y_1 and Y_3 requires a small Σ_{33}.

To gain 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 $\operatorname{cor}(Y_3, Y_4) = \{\exp(\Sigma_{43}) - 1\}\{\exp(\Sigma_{33}) - 1 + \psi_3/\mathbb{E}(Y_3)\}$. For 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 because the moments have closed form expressions. In more complicated settings, the following result can be useful. It implies 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.

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 $\mu_1 \mapsto \int g(t)\phi_{\mu_1,\sigma_1^2}(t)\,\mathrm{d}t$ 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 $Y_1 \mid W_1 \sim \mathrm{N}(W_1, \psi_1)$ and Y_2 Bernoulli distributed with $\mathbb{E}(Y_2 \mid W_2) = \operatorname{logit}^{-1}(W_2) = 1/\{1 + \exp(-W_2)\}$. Suppose also for simplicity $W = \beta + \varepsilon$, $\beta \in \mathbb{R}^2$. The marginal distribution of Y_2 is Bernoulli with $\mathbb{E}(Y_2) = \int[\phi(t)/\{1 + \exp(-\beta_2 - \sqrt{\Sigma_{22}}t)\}]\,\mathrm{d}t$, where $\phi(\cdot)$ is the standard normal density. 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 restrictions are needed for identifiability. One possibility, which has been used in similar settings, is to fix Σ_{22} to some value, say 1 (Dunson, 2000; Bai et al., 2020). While fixing $\Sigma_{22} = 1$ does not impose 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 (2) implies $\operatorname{cov}(Y_1, Y_2) = \int[\{t_1\phi_{\beta,\Sigma}(t)\}/\{1 + \exp(-t_2)\}]\,\mathrm{d}t - \beta_1\mathbb{E}(Y_2)$. 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 $\operatorname{cov}(Y_1, Y_2)$ can take, consider the limiting case as $\Sigma_{12} \to \sqrt{\Sigma_{11}}\sqrt{\Sigma_{22}}$ 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 one can show $\operatorname{cor}(Y_1, Y_2) = 2\int[\{\sqrt{\Sigma_{11}}t\phi(t)\}/\{1 + \exp(-\sqrt{\Sigma_{22}}t)\}]\,\mathrm{d}t/\sqrt{\psi_1 + \Sigma_{11}}$. By using the dominated convergence theorem as $\psi_1 \to 0$ and $\Sigma_{22} \to \infty$, this 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.

We conclude with a result on identifiability. The result is stated for some common choices of conditional moments of $Y \mid W$ but the proof idea applies to other settings. The proof is in the Supplementary Materials.

Theorem 2.2. *Suppose $\{(Y_i, X_i) \in \mathbb{R}^r \times \mathbb{R}^{r\times q}; i = 1, \dots, n\}$ is an independent sample from our model and that (i) g_j, $j = 1, \dots, r$, is either the identity, natural logarithm, or logit (log-odds) function; (ii) Σ_{jj} is fixed and known for every j corresponding to logit g_j; and (ii) $\sum_{i=1}^n X_i^\mathsf{T} X_i$ is invertible, then distinct (β, Σ) correspond to distinct $\{\mathbb{E}(\mathcal{Y}), \operatorname{cov}(\mathcal{Y})\} \in \mathbb{R}^{rn} \times \mathbb{R}^{rn\times rn}$, where $\mathcal{Y} \in \mathbb{R}^{rn}$ is a vector of all responses.*

Identifiability in the usual sense that distinct parameters correspond to distinct distributions is a corollary if distinct distributions in the family considered do not have the same first two moments (e.g. GLM distributions for $Y_j \mid W$). From a computational perspective, non-identifiability can lead to likelihoods with infinitely many maximizers or ridges along which the likelihood is constant. In light of this result, we constrain the diagonals of Σ corresponding to binary response variables: we found this leads to faster computation and improved estimation accuracy.

3 Estimation

3.1 Overview

We propose an algorithm 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 (McCullagh and Nelder, 1989; Schall, 1991). 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 different from those typically considered in mixed models.

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 $\mathbb{E}(Y) \approx \nabla c(w) + \nabla^2 c(w)(X\beta - w) := m(w, \beta)$ and $\operatorname{cov}\{\mathbb{E}(Y \mid W)\} \approx \nabla^2 c(w)\Sigma\nabla^2 c(w)$. Approximating $\mathbb{E}\{\operatorname{cov}(Y \mid W)\} = \mathbb{E}\{\operatorname{diag}(\psi)\nabla^2 c(W)\} \approx \operatorname{diag}(\psi)\nabla^2 c(w)$ leads to $\operatorname{cov}(Y) \approx \operatorname{diag}(\psi)\nabla^2 c(w) + \nabla^2 c(w)\Sigma\nabla^2 c(w) := C(w, \Sigma)$. 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 $Y_1, \dots, Y_n$ are independent with

$$Y_i \sim \mathrm{N}\{m(w_i, \beta), C(w_i, \Sigma)\}, \tag{3}$$

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

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

If all responses are normal, then the working model is exact and minimizers of h_n are maximum likelihood estimates (MLEs). More generally, minimizers of h_n are approximate MLEs whose quality depend on the accuracy of the working model (3). A natural algorithm for estimating β and Σ would iterate between updating (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. 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 $f_{\beta,\Sigma}(w_i \mid y_i)$. This update for w_i is closely related to Laplace approximation arguments used to motivate common algorithms for mixed models. 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

$$(\beta^{(k+1)}, \Sigma^{(k+1)}) = \arg\min_{\beta,\Sigma} h_n(\beta, \Sigma \mid w_1^{(k)}, \ldots, w_n^{(k)}); \tag{4}$$

$$(w_1^{(k+1)}, \ldots, w_n^{(k+1)}) = \arg\max_{w_1,\ldots,w_n} \sum_{i=1}^{n} \log f_{\beta^{(k+1)},\Sigma^{(k+1)}}(w_i \mid y_i). \tag{5}$$

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 shortly, 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 (4) and (5). We provide a formal statement of the complete algorithm in Algorithm 1.

3.2 Updating β and Σ

To solve (4), we use a blockwise coordinate descent algorithm. Treating $w = \{w_1, \ldots, w_n\}$ as fixed throughout and ignoring the iterate superscript, this algorithm iterates between updating β and Σ.

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

```
1. Given $\epsilon_\beta > 0$, $\epsilon_\Sigma > 0$, 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 $i = 1, \dots, n$.
3. Set $\tilde{\Sigma}^{(1)} = \Sigma^{(k)}$. For $l = 1, 2, \dots$ until convergence:
    (a) $\tilde{\beta}^{(l+1)} \leftarrow \arg\min_\beta h_n(\beta, \tilde{\Sigma}^{(l)} \mid w_1^{(k+1)}, \dots, w_n^{(k+1)})$
    (b) Set $\bar{\Sigma}^{(0)} = \bar{\Sigma}^{(1)} = \tilde{\Sigma}^{(t)}$. For $t = 1, 2, \dots,$ until convergence:
            $\bar{\Sigma}^{(t+1)} \leftarrow \mathcal{P}_{\mathbb{M}} \left[ \bar{\Sigma}^{(t)} - \alpha \nabla_\Sigma h_n(\tilde{\beta}^{(l+1)}, \bar{\Sigma}^{(t)}, w_1^{(k+1)}, \dots, w_n^{(k+1)}) + \gamma\{\bar{\Sigma}^{(t)} - \bar{\Sigma}^{(t-1)}\} \right]$,
    (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. Otherwise, set $k \leftarrow k + 1$
   and return to 2.
```

Specifically, the $(l+1)$th iterates of the algorithm for solving (4) can be expressed

$$\beta^{(l+1)} = \arg\min_{\beta} h_n(\beta, \Sigma^{(l)} \mid w_1, \dots, w_n); \tag{6}$$

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

Update (6) can be shown to be a weighted residual sum-of-squares with solution

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

where $\tilde{X}_i = \nabla^2 c(w_i) X_i$ and $\tilde{y}_i = y_i - \nabla c(w_i) + \nabla^2 c(w_i) w_i$. Minimizing h_n with respect to Σ is non-trivial owing to non-convexity 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 since we, as discussed in Section 2, sometimes restrict diagonal elements of Σ to be equal to a prespecified constant for identifiability. Similarly, testing correlation of responses requires constraining some off-diagonal elements to equal 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) $\mathbb{M} \subseteq \mathbb{R}^{r \times r}$, (7) can be characterized as an optimization problem over $\mathbb{R}^{r\times r}$ with the constraint that $\Sigma \in \mathbb{M}$. To handle both the non-convexity and general constraints, we propose to solve this problem 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 $\Sigma \in \mathbb{M}$ 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 gradient step is modified to account for the direction of change from the previous iteration, which can lead to faster convergence.

To summarize briefly, our algorithm for (7) has $(t+1)$th iterate $\Sigma^{(t+1)} = \mathcal{P}_{\mathbb{M}}[\Sigma^{(t)} - \alpha\nabla_\Sigma h_n(\beta, \Sigma^{(t)} \mid w_1, \dots, w_n) + \gamma\{\Sigma^{(t)} - \Sigma^{(t-1)}\}]$, where $\gamma = (0, 1)$, α is determined using backtracking line search (see Ochs et al., 2014, Algorithm 4), and $\mathcal{P}_{\mathbb{M}}$ is the projection onto $\mathbb{M}$. We assume the 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, $\mathbb{M}$ is the intersection of a set of matrices with constrained elements and the set of symmetric matrices with eigenvalues bounded below by $\epsilon \geq 0$. 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 $\mathbb{M}$. Both projections can be computed in closed form, so this algorithm tends to be very efficient. The gradient of h_n with respect to Σ needed for implementing the algorithm can be found in the Supplementary Materials. This algorithm is terminated when the objective function values converge.

3.3 Updating the approximation points

We use a trust region algorithm for updating $w_i, i = 1, \dots, n$ (e.g. Nocedal and Wright, 2006, Chapter 4). Essentially, the trust region algorithm approximates the objective function locally by a quadratic and requires the computation of gradients and Hessians. The gradient is given in the Supplementary Materials and the Hessian is, assuming Σ^{-1} is positive definite, for $i = 1, \dots, n$, $\nabla^2_{w_i} \log f_\theta(w_i \mid y_i) = -\nabla^2 c(w_i) - \Sigma^{-1}$. 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 and therefore has a unique maximizer and stationary point. In practice, however, Σ can be singular or near-singular and the Hessian $-\Sigma^{-1} - \nabla^2 c(w)$ can have a large condition number. To improve stability, we regularize 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 $\epsilon > 0$.

Then the optimization problem for updating w_i is

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

where $\tau \geq 0$. The intuition for shrinking w_i to $X_i\beta$ is that the latter is the mean of W_i when β is the true parameter. 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. In the Supporting Materials, we outline a procedure for obtaining starting values that can improve computing times and the quality of the resulting estimates relative to naive starting values.

4 Approximate likelihood ratio testing

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 $(\bar{\beta}, \bar{\Sigma}) = \arg\min_{(\beta,\Sigma)\in\mathbb{H}_1} h_n(\beta, \Sigma \mid \tilde{w}_1, \ldots, \tilde{w}_n)$. 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 (3) with the expansion points fixed at those obtained by fitting the null model. 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, 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 $\mathbb{H}_1$. Null models corresponding to hypotheses that constrain elements of Σ are straightforward to fit by including those constraints in the definition of the set $\mathbb{M}$ in the update for Σ. A formal statement of the full algorithm for hypothesis testing is given in Algorithm 2. We investigate the size and power of the proposed procedure in Section 5.3.

5 Numerical experiments

5.1 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. We therefore compare to existing methods which assume related but somewhat different models. We consider $r = 9$ response variables, three normally distributed, three Poisson, and three Bernoulli. A reasonable alternative to our method is to use separate GLMMs for the three response

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 $k = 1, 2, \dots$ 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 $i = 1, \dots, n$
    (b) $(\beta^{(k+1)}, \Sigma^{(k+1)}) \leftarrow \arg\min_{(\beta,\Sigma)\in\mathbb{H}_0} h_n(\beta, \Sigma \mid w_1^{(k+1)}, \dots, w_n^{(k+1)})$
3. Set $(\tilde{\beta}, \tilde{\Sigma}) = (\beta^{(k^*)}, \Sigma^{(k^*)})$ and $\tilde{w}_i = w_i^{(k^*)}$ for $i = 1, \dots, n$ where $k^*$ denotes the final iterate
   from 2.
4. Compute $(\bar{\beta}, \bar{\Sigma}) = \arg\min_{(\beta,\Sigma)\in\mathbb{H}_1} h_n(\beta, \Sigma \mid \tilde{w}_1, \dots, \tilde{w}_n)$
5. Return $T_n = h_n(\tilde{\beta}, \tilde{\Sigma} \mid \tilde{w}_1, \dots, \tilde{w}_n) - h_n(\bar{\beta}, \bar{\Sigma} \mid \tilde{w}_1, \dots, \tilde{w}_n)$.
```

types. However, common software cannot fit these models due to the unusual random effects structure; see the Supplementary Materials. 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) assumes all responses are independent and option (ii) corresponds to using a shared random effect for observations of the same type in the same response vector. We refer to these as independent and clustered GLMMs. There are many software packages that can fit these. We pick the `glmm` (Knudson et al., 2021) package to fit (i) and fit (ii) using `lme4` (Bates et al., 2015). 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 our setting and, accordingly, in our experience differences in performance are due to misspecification rather than particular implementations. We consider two versions of our method, one which constrains Σ to be diagonal and one where Σ is (nearly) unconstrained; for both versions, we constrain $\Sigma_{jj} = 1$ for all j corresponding to Bernoulli responses. Diagonal Σ corresponds to ignoring dependence between mixed-type responses.

The comparisons focus on out of sample prediction errors. Predictions are formed by plugging estimates into the expressions for $\mathbb{E}(Y_i) = \mathbb{E}\{\mathbb{E}(Y_i \mid W_i)\}$ in Section 2. When a closed form expression is unavailable, the expectation is obtained by r one-dimensional numerical integrations. We compare to (oracle) predictions using the true β and Σ. The responses have different predictors and we partition accordingly: $\beta = [\beta_1^\mathsf{T}, \dots, \beta_r^\mathsf{T}]^\mathsf{T}$, $\beta_j \in \mathbb{R}^{p_j}$, $q = \sum_{j=1}^r p_j$, and we write $X_{i,j} \in \mathbb{R}^{p_j}$ for the ith observation of the predictors for the jth response. 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 $\mathrm{U}[-1, 1]$ random variable, where $p_j = p_k$ for all j and k. For $j = 1, \dots, r$, the true regression coefficient β_j has first element equal to β_{0j} and all other elements chosen as independent realizations of a $\mathrm{U}[-.5, .5]$. 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: for some $\rho \in (0, 1)$ we set $\Sigma = 0.5\tilde{\Sigma}$ where $\tilde{\Sigma}$ is (i) autoregressive ($\tilde{\Sigma}_{jk} = \rho^{|j-k|}$); (ii) compound symmetric ($\tilde{\Sigma}_{jk} = \rho\mathbb{I}(j \neq k) + \mathbb{I}(j = k)$); 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, each of the blocks given by the structure in (iii) includes one of each response type. These structures are used to generate the data but are not imposed when fitting models.

For each structure of Σ, we investigate the effects of the sample size (n), the number of predictors (p_j, $j = 1, \dots, r$), and the correlation parameter (ρ). We present relative squared prediction errors, defined as the ratio of a 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, out of sample predictions are on an independent test set of 10^4 observations.

In the top row of Figure 1, as n increases, each method's performance improves relative to oracle predictions. However, across all settings, our method performs best. When the covariance structure is non-sparse (e.g., autoregressive or compound symmetric), the clustered GLMMs can outperform both our method with the diagonality constraint and the independent GLMMs. The same relative performances are observed as p increases in the middle row; and when ρ increases in the bottom row. When Σ is block diagonal, both versions of our method outperform the competitors. In the case of clustered GLMMs, this is likely due to the the specified covariance structure being a poor approximation to the true covariance. For independent GLMMs, this is likely due to the fact that `glmm` (nor other software) can impose the identifiability condition on Σ for the Bernoulli responses.

The Supplementary Materials include a comparison of computing times. Our method scales well in n, is frequently faster than `glmm` but slower than `glmer`. These timings are intuitive as `glmm` uses a Monte Carlo-approximation, which is not optimized for our setting with many independent observations, and `glmer` is fitting a model with only one variance parameter.

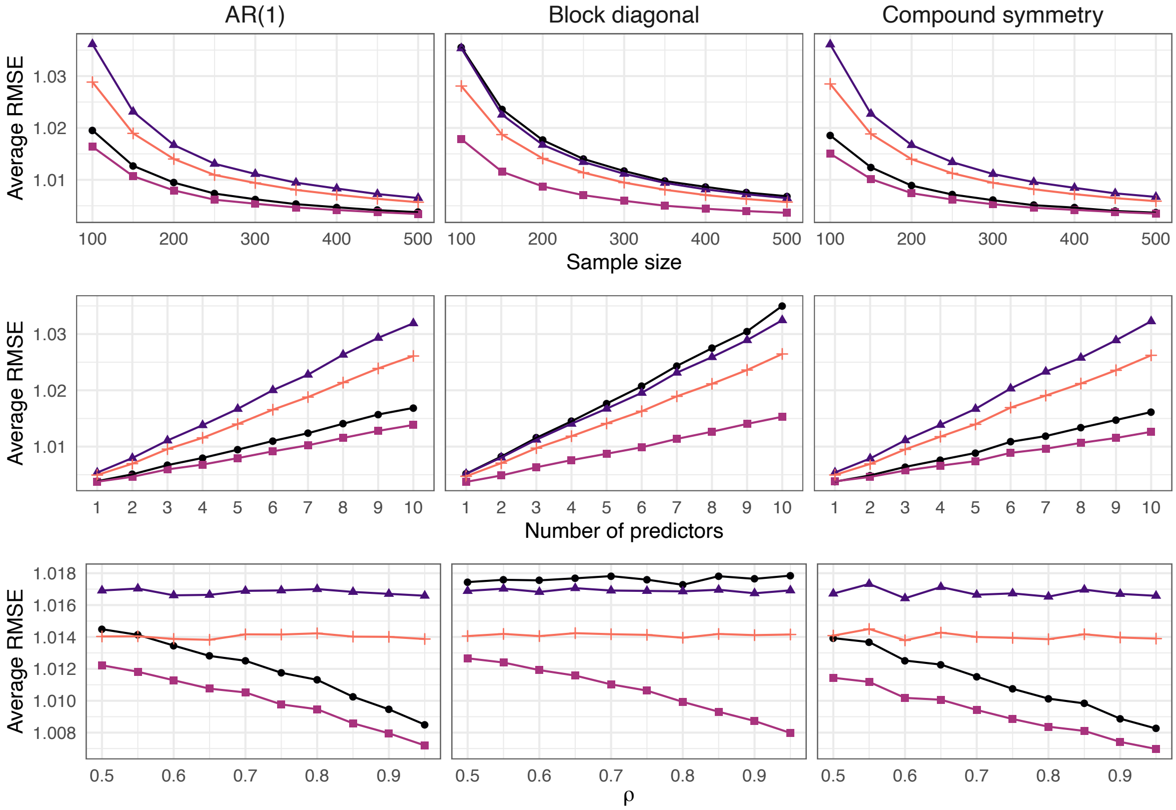

Figure 1: Average relative squared prediction errors. Top: $\rho = 0.9$ and $p_j = 5$ for $j = 1, \dots, 9$. Middle: $n = 200$ and $\rho = 0.9$. Bottom: $n = 200$ and $p_j = 5$ for $j = 1, \dots, 9$. Squares are our method, plus signs are our method with diagonal Σ, triangles are independent GLMMs, and dots clustered GLMMs.

5.2 Performances for different response types

In Figure 1 averages were taken over all responses types. To see if the benefits of joint modeling are greater for some response types, it is of interest to stratify results by type. We compare the two versions of our method (diagonal versus non-diagonal Σ). Because 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. Data are generated as in the previous section but with $\psi_j = .01$ for both normal responses and quasi-Poisson responses.

In the first row of Figure 2, as n increases, both methods' relative mean squared prediction error approaches the oracle prediction error. However, the predictions from joint modeling outperforms those using a diagonal Σ. The differences between the two methods are smallest for Bernoulli

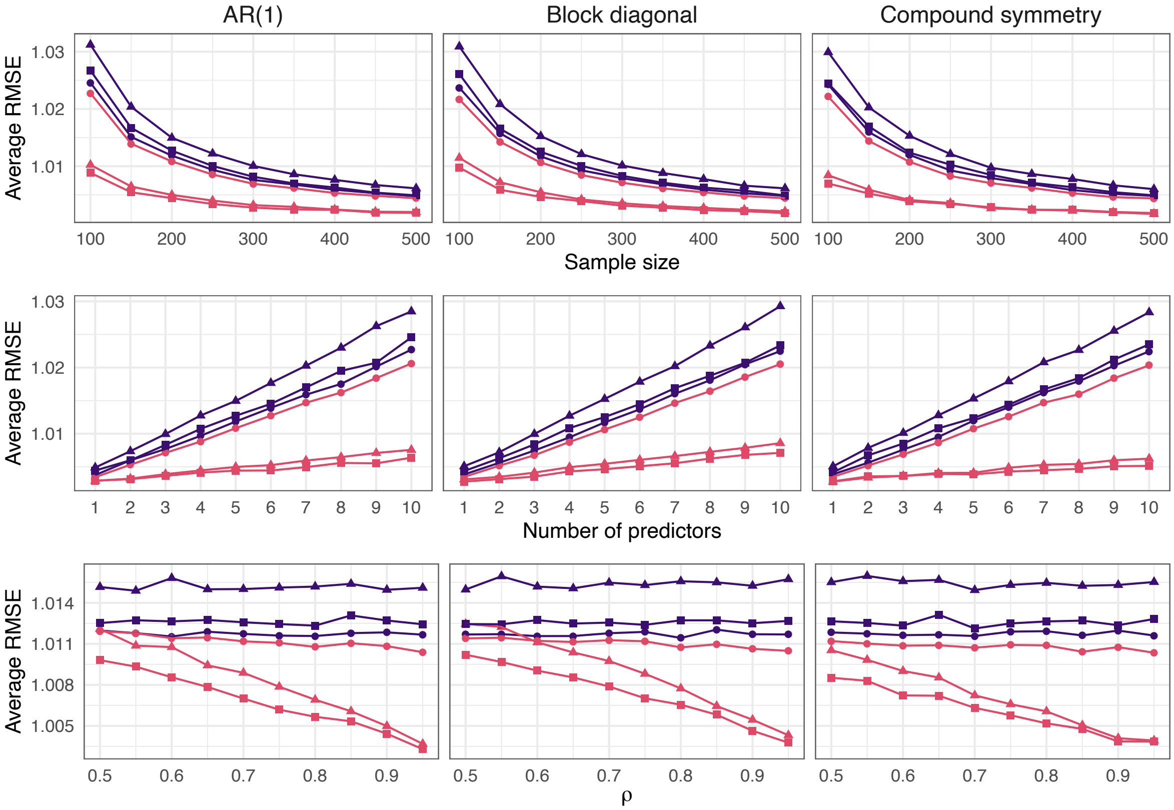

Figure 2: Average relative squared prediction errors. Top: $\rho = 0.9$ and $p_j = 5$ for $j = 1, \dots, 9$. Middle: $n = 200$ and $\rho = 0.9$. Bottom $n = 200$ and $p_j = 5$ for $j = 1, \dots, 9$. Purple: our method with diagonal Σ. Magenta: our method. Triangles are averages over normal responses, squares over quasi-Poisson, and circles over Bernoulli.

responses. A similar result is observed in the second row: 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, we display results as ρ varies. When $\rho = 0.5$ there is a less substantial difference between the two methods. As ρ approaches 0.95, 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 Supplementary Materials, we present a simulation study which focuses on modeling many Bernoulli responses and a single normal response. Those results highlight that even though the relative squared prediction error for the Bernoulli responses is only slightly improved by joint modeling, one can realize substantial prediction accuracy gains for the single normal response variable by exploiting dependence between it and the Bernoulli responses.

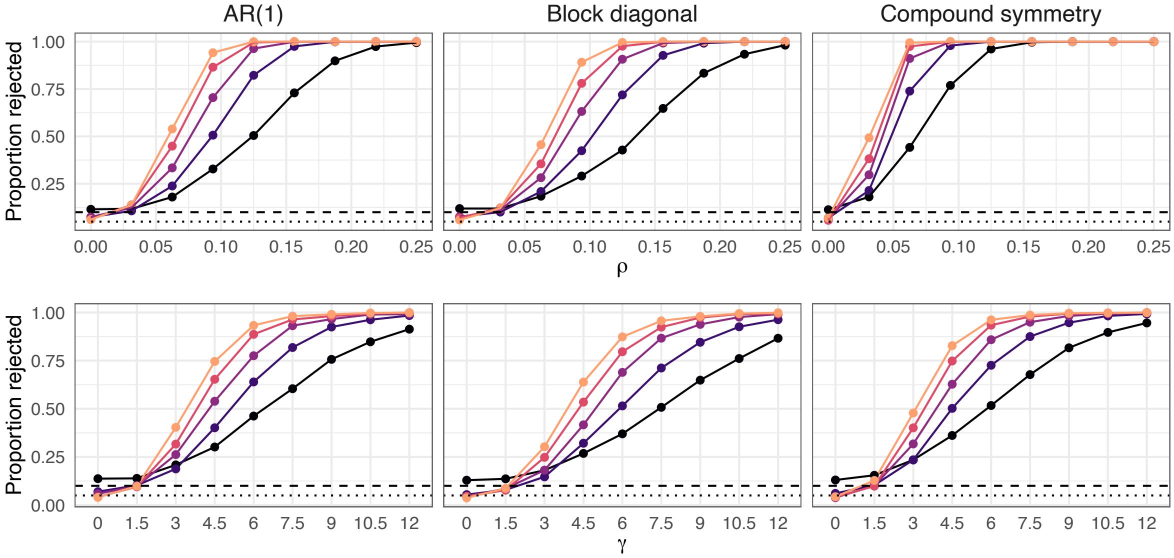

Figure 3: Top: Proportion of $H_0 : \Sigma \in \mathbb{D}^r_{++}$ rejected at 0.05 level. Bottom: Proportion of $H_0 : \mathcal{B}_{kj} = 0$ rejected at the 0.05 level. Horizontal dashed and dotted black lines indicate 0.10 and 0.05. Colors denote sample sizes: black solid line is $n = 200$, dark blue is $n = 400$, the purple is $n = 600$, the magenta is $n = 800$, and the light orangeis $n = 1000$.

5.3 Approximate likelihood ratio testing

We examine the approximate likelihood ratio testing procedure described in Section 4. 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 tests for $H_0 : \Sigma \in \mathbb{D}^r_{++}$ versus $H_A : \Sigma \in \mathbb{S}^r_{++}$, and, assuming all responses have the same predictors as in (1), $H_0 : \mathcal{B}_{kj} = 0$ for every $j = 1, \dots, r$ versus $H_A : \mathcal{B} \in \mathbb{R}^{p\times r}$, where $\mathcal{B}_{kj}$ denotes the kth predictor's effect on the jth response variable, i.e., the (k, j)th element of $\mathcal{B}$. We set $k = 2$. Thus, the null hypothesis implies that the first predictor (ignoring the intercept) has no effect on any response. Multiple testing corrections, which are often needed when using separate models for the r responses, are not needed here.

Data are generated as in Section 5.2 but with $X_{i,1} = X_{i,2} = \cdots = X_{i,r}$ for all $i = 1, \dots, n$ and $\mathcal{B} = [\beta_1, \dots, \beta_r] \in \mathbb{R}^{p\times r}$. In the first setting, $n \in \{200, 400, \dots, 1000\}$ and $\tilde{\Sigma}_{jk} = \rho^{|j-k|}$, $\rho \in \{0.0, 0.05, \dots, 0.4\}$. The top row of Figure 3 displays the proportion of rejections at the 0.05 significance level. When $\rho = 0$ (H_0 is true), the proportion of rejections is approximately 0.10 when $n = 200$, below 0.075 when $n \geq 400$, and near 0.05 (the nominal level) when $n = 2000$. As ρ increases, even with $n = 200$, the proportion of correctly rejected null hypotheses is near one when $\rho = 0.4$. The power depends positively on both the magnitude of ρ and the sample size.

In the second setting, we fix $\rho = 0.5$ and study how the effect size of the $\mathcal{B}_{kj}$ affects power. After generating $\mathcal{B}$ as in Section 5.2, for $j = 1, \ldots, r$ independently, we replace $\mathcal{B}_{kj}$ with a realization of a $\mathrm{U}[-\gamma 10^{-2}, \gamma 10^{-2}]$ where $\gamma \in [0, 12]$. The second row of Figure 3 shows that when $\gamma = 0$, so that $\mathcal{B}_{kj} = 0$ for all $j = 1, \ldots, r$, the proportion of rejections is slightly 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 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.

The response variables quantify the ability to conceive in different ways. Two are approximately normally distributed (square-root estradiol level and log-total gonadotropin level); and two are counts (number of egg cells and number of embryos). We modeled the latter using our model with conditional quasi-Poisson distributions. We set $\psi_j = 10^{-2}$ for continuous responses and $\psi_j = 10^{-1}$ for counts. First, we test the hypothesis $H_0 : \Sigma \in \mathbb{D}^4_{++}$ versus $H_A : \Sigma \in \mathbb{S}^4_{++}$ and find evidence against the null hypothesis (p-value $< 10^{-16}$) using the test described in Section 5.3. That is, there is 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, $\widehat{\mathrm{cor}}(Y_i \mid X_i = \bar{X})$, and the latent variables, $\widehat{\mathrm{cor}}(W_i \mid X_i)$, are, respectively,

$$\begin{pmatrix} 1.00 & 0.01 & -0.08 & -0.09 \\ 0.01 & 1.00 & -0.03 & -0.09 \\ -0.08 & -0.03 & 1.00 & 0.69 \\ -0.09 & -0.09 & 0.69 & 1.00 \end{pmatrix} \text{ and } \begin{pmatrix} 1.00 & 0.02 & -0.09 & -0.10 \\ 0.02 & 1.00 & -0.04 & -0.09 \\ -0.09 & -0.04 & 1.00 & 0.74 \\ -0.10 & -0.09 & 0.74 & 1.00 \end{pmatrix},$$

where the variable ordering is square-root estradiol level, log-total gonadotropin level, number of egg cells, and number of embyros. The estimate of $\mathrm{cor}(Y_i \mid X_i)$ is here evaluated at $\bar{X} = \sum_{i=1}^{n} X_i/n$. The estimates indicate substantial positive correlation between the number of egg cells and number

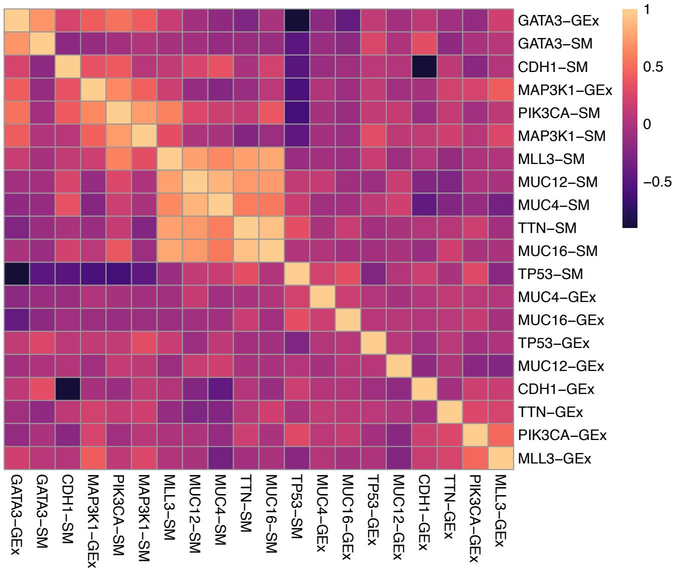

Figure 4: Heatmap of the estimated correlation matrix for the $W_i \mid X_i$. Suffix -SM for somatic mutations; -GEx for gene expression.

of embryos, whereas estradiol and gonadotropin levels appear weakly negatively correlated with these two variables.

We also test whether the small antra follicle count is a significant predictor of any of the responses after accounting for age, average antral follicle count, and maximum follicle stimulating hormone level. The number of small antra follicles (2-5 mm) is 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 modeling common somatic mutations and gene expression measured on patients with breast cancer collected by the The Cancer Genome Atlas Project (TCGA). A somatic mutation is an alteration in the DNA of a somatic cell. Somatic mutations are believed to play a central role in the development of cancer. Because somatic mutations modify gene expression, directly and indirectly, it is natural to model somatic mutations and gene expression jointly.

The somatic mutations we model are binary variables for 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. 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 a predictor.

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 4, we display the estimated correlation matrix for the $W_i \mid X_i$. We observe that the latent variables corresponding to somatic mutations and gene expression 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).

Supplementary Materials

The Supplementary Materials include proofs, additional details, and an osteoarthritis initiative data analysis.

Acknowledgements

We are grateful to Galin Jones and Adam Rothman for their comments on an earlier version. We thank Charles McCulloch for sharing the OAI dataset analyzed in the Supplementary Materials. Ekvall and Molstad gratefully acknowledge support from FWF (Austrian Science Fund) [P30690-N35] and NIH R01HL152439, respectively.

References

Bai, H., Zhong, Y., Gao, X., and Xu, W. (2020). Multivariate mixed response model with pairwise composite-likelihood method. *Stats*, 3(3):203–220.

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):183–202.

Bonat, W. H. and Jørgensen, B. (2016). Multivariate covariance generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(5):649–675.

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, pages 28–47, New York, NY. Springer.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

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):883–900.

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):651–658.

Cox, D. R. and Wermuth, N. (1992). Response models for mixed binary and quantitative variables. *Biometrika*, 79(3):441–461.

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):533–548.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(2):355–366.

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):932–952.

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):4408–4427.

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):845–852.

Fitzmaurice, G. M. and Laird, N. M. (1997). Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics*, 53(1):110–122.

Goldstein, H., Carpenter, J., Kenward, M. G., and Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling*, 9(3):173–197.

Gueorguieva, R. V. (2001). A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Statistical Modeling*, 1(3):177–193.

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):1102–1112.

Gueorguieva, R. V. and Sanacora, G. (2006). Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Statistics in Medicine*, 25(8):1307–1322.

Kang, X., Chen, X., Jin, R., Wu, H., and Deng, X. (2021). Multivariate regression of mixed responses for evaluation of visualization designs. *IISE Transactions*, 53(3):313–325.

Knudson, C., Benson, S., Geyer, C., and Jones, G. (2021). Likelihood-based inference for generalized linear mixed models: Inference with the R package glmm. *Stat*, 10(1):e339.

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):124–140.

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

Megginson, R. E. (1998). *An Introduction to Banach Space Theory*. Springer, New York, NY.

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer-Verlag GmbH, New York, NY.

Ochs, P., Chen, Y., Brox, T., and Pock, T. (2014). iPiano: inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419.

Olkin, I. and Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, 32(2):448–465.

Pinheiro, J. C. and Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and computing*, 6(3):289–296.

Poon, W.-Y. and Lee, S.-Y. (1987). Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients. *Psychometrika*, 52(3):409–430.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, 69(2):167–190.

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):667–678.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78(4):719–727.

Yang, Y., Kang, J., Mao, K., and Zhang, J. (2007). Regression models for mixed Poisson and continuous longitudinal data. *Statistics in Medicine*, 26(20):3782–3800.